ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 5 , May 2018

Sentiment Analysis on Textual Data

Mirsa karim, Smija Das

P.G. Student, Department of Computer Science and Engineering, St.Josephs College of Engineering and Technology ,Pala,Kottayam,India

Assistant Professor, Department of Computer Science and Engineering ,St.Josephs College of Engineering and Technology ,Pala,Kottayam,India

ABSTRACT:. Using Natural Language Processing, there is need to identify sentiment of content or document. Here Sentiment Analysis is done in view of Rule based mechanism , machine learning and deep learning approach.All of these strategies are analyzed and discovered that deep learning is most appropriate for Sentiment Analysis in light of the Accuracy measurement. Sentiment Vader and Senti word net are the Rule based algorithms utilized , LDA analysis on Naive bayes is the machine learning strategy used and recurrent neural network on tensor flow is deep learning strategy used for Sentiment Analysis.

KEYWORDS: Sentiment Vader, Sentiword net, LDA

I.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 alloted 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.

A. 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 long 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,utliize 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.

II. LITERATURE SURVEY

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. RajatBhat, SwapnilGaonkar[9] utilized Hadoop Distributed File framework (HDFS) to store information set and keep running on MapReduce outline for performing Sentiment Analysis. ArunaSathish[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], XuemingQian 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.

XuemingQian, He Feng[1]considered three necessary factors, personal attraction, interpersonal similary consideration, and interpersonal impact, and combined them into a systemic accurate recommendation model probabilistic matrix factorization technique was used for recommendation model. ShengxiangGao, 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 problablistic matrix factorisation for review experts.

Evangelos PsomakelisKonstantinos Tserpes[10] 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 C4.5as 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[11] 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[12] 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[13] 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[15] suggested that for conclusion investigation positive and negative word scores can be checked utilizing sentword net by assessment introduction.

III. METHODOLOGY

In this Paper Sentiment Analysis is done based on Rule based mechanism, machine learning and deep learning approach. All of these methods are compared and found that deep learning is best suited for Sentiment Analysis based on the accuracy measurement. Sentiment Vader and Senti word net are the Rule based mechanisms used , LDA analysis on naive bayes is the machine learning technique used and Recurrent Neural Network is deep Learning approach used

A. DATA SET

Moview review dataset (http://www.cs.cornell.edu/people/pabo/movie-review-data/) from cornell university labelled as positive and negative is taken.

B.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.

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.

norm score = score/math.sqrt((score*score) + alpha)

The review is considered as positive, if compound score is >= 0.5,is considered as neutral if score > -0.5 and< 0.5 and is considered as negative if score is <=-0.5

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 ie,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 negative count is
odd number then sentence is negative ,else positive

C.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.

LDA ANALYSIS ON NAIVE BAYES:

Data Preprocessing for LDA:

All the unique words are constructed in the vocabulary V , each word has a label wi 1, 2, . .. , Nd .For each document dj , we choose a dimensional Dirichlet random variable *θmDirichlet*
(a)F oreachtopiczk, wherek[1, γ], wechooseθkDirichlet
(b)F oreachtopiczk, theinferenceschemeisbasedupontheobservationthat :

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

Extracting Product Features:

Tags are added ie:, 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) wi 1, 2, . . .,Nd : unique word in review
3) dm: document associated with user
4) *T* : total number of topics
5) *θ*m: the multinomial distribution of topics
6) φk : the component for each topic
7) a,b : Dirrichlet priors associated with the multinomial distributionm and k

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 /Nc (Q *Rw *Dw)

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

The Nave bayes classifier is a simple classifier that relies on Bayesian probability and the nave assumption that feature probabilities are independent of one another. Training data give information about conditional and priori probabilities which feed to naive bays classifcation 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 pos or neg reviews.

Extract reviews:

Extract reviews get 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.

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 vocabualary.If a word is Present in vocabulary dictionary hold true for value else show false

Training:

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

D.DEEP LEARNING

In the past few years, deep learning has seen incredible progress and has largely removed the requirement of strong domain knowledge. As a result of the lower barrier to entry, applications to NLP tasks have been one of the biggest areas of deep learning research.

Recurrent Neural Network:

The recurrent neural network structure is a little different from the traditional feed forward NN you may be accostumed to seeing. The feedforward network consists of input nodes, hidden units, and output nodes.

Associated with each time step is also a new component called a hidden state vector ht. From a high level, this vector seeks to encapsulate and summarize all of the information that was seen in the previous time steps. Just like xt is a vector that encapsulates all the information of a specific word, ht is a vector that summarizes information from previous time steps.

The hidden state is a function of both the current word vector and the hidden state vector at the previous time step. The sigma indicates that the sum of the two terms will be put through an activation function

$h(t)=\sigma(W^{H} \ast ht-1+W^{X} \ast xt)$

The weight matrices are updated through an optimization process called back propagation through time.

The hidden state vector at the final time step is fed into a binary softmax classifier where it is multiplied by another weight matrix and put through a softmax function that outputs values between 0 and 1, effectively giving us the probabilities of positive and negative sentiment

Long Short Term Memory Units (LSTMs)

Long Short Term Memory Units are modules that you can place inside of reucrrent neural entworks. At a high level, they make sure that the hidden state vector h is able to encapsulate information about long term dependencies in the text. As we saw in the previous section, the formulation for h in traditional RNNs is relatively simple.

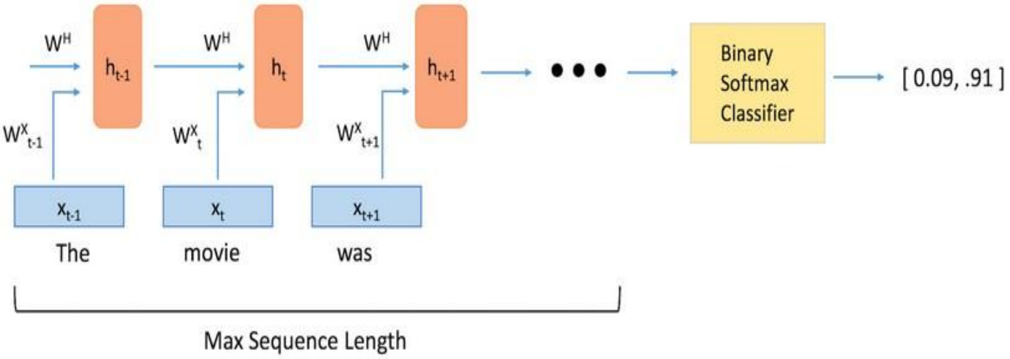


Figure 1. Softmax Layer Probability

sentiment analysis involves taking in an input sequence of words and determining whether the sentiment is positive, negative, or neutral.

1)Loading Data 2) Creating an ID's 3) Training 4) Testing

Loading Data:

Download file on local machine then specify url were file exist,look whether the file already present .Then extract and clean up the reviews Regular expression Can be used to preserve alphabets and numbers from reviews and exclude special character

Creating an ID's:

Map every word in dataset to unique numeric identifier .Maximum sequence length should be set ie, all reviews has same length.Every review must be equal length so max sequence length given one word represent one time instance .For obtaining max sequence length pad short reviews and truncate long reviews.

Training:

Generate word embeddibgs during trainingprocess. Shuffle dataset when given to training process. Placeholders are intialized ie x placeholder for reviews and y for labels. Embedding matrix during traing each word has embeddings Embedding matrix is size of vocabulary. Embeddings for current batch of data shape(batch ,nsteps,n inputs)ie n inputs is dimentinality of single inputs, n steps is of max seqquence length and rnn will be unrolled every time intialise long shrt term memomory . Wrap memory cell in drop out wrapper to prevent overfitting model of input text.Final state is fed to softmax prediction layer to get final otput positive or negative.

Testing:

Soft max layer has a liner layer called logicts .Softmax activation is part of loss function Calcute cross entropy using loss function and pass output to logics layer calculate loss function using reduce mean on cross entropy then set a compute nod which classify reviews .Output with highest probability in softmax is used for prediction. Prediction is accurate if predicted label equals actual label

1V PERFORMANCE EVALUATION

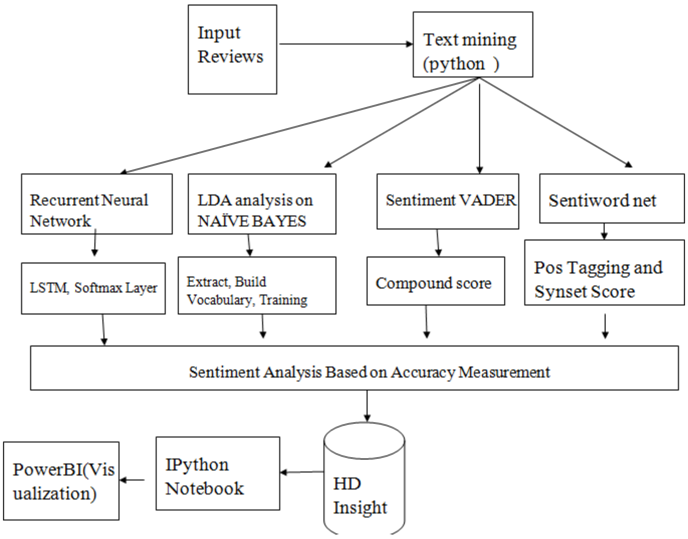
SentiWord Net,Sentiment Vader , LDA Analysis on naive Bayes and RNN was used in order to predict sentiment associates with review and found that machine learning technique is best suited for sentiment analysis.In Sentiword net and Vader label1 is assigned to positive review and 0 for negative review.Summing all values of positive and negative reviews with the total length of reviews is used for performance measurement.While in Naïve baves label 1 for positive review and -1 for negative review is used.

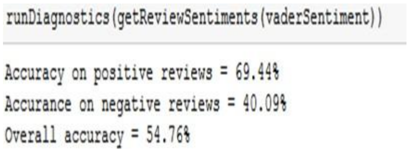
Based on the Accuracy measurement obtained from the same data set given it is shown that Sentiment Vader has 54.7 percent Accuracy which is low .RNN has Accuracy of 81 which is high as compared to rule based mechanisms which shows that Deep learning techniques is more accurate in calculating Sentiment of reviews.

Algorithm	SentiWord Net	Sentiment Vader	Naive Bayes	RNN
Accuracy	59.17	54.76	75.2	81

V ARCHITECTURE

In the proposed system Architecture Text Reviews ie, sequential data is given as input.Then Sentiment Analysis is done by Sentiwordnet,Sentiment Vader,Naive Bayes and RNN.Sentiword net used 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 Vader uses compound score to calculate sentiment associatd with review.For training Naive Bayes build a Vocabualary and extract features from the review.RNN uses softmax Prediction layer to calculate Output.HDInsight is a completely overseen cloud benefit that makes it simple, quick, and savvy to process monstrous measures of information. Utilize famous open-source systems, for example, Hadoop, Spark, Hive, LLAP, Kafka, Storm, R and more. Purplish 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 examination instruments that convey bits of knowledge all through your association. Associate with several information sources, rearrange information prepare, and drive impromptu investigation. Deliver lovely reports, at that point distribute them for your association to devour on the web and crosswise over cell phones.

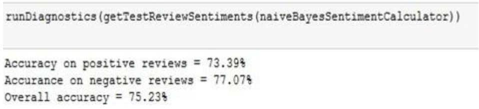




```
runDiagnostics(getReviewSentiments(vaderSentiment))

Accuracy on positive reviews = 69.44%
Accurance on negative reviews = 40.09%
Overall accuracy = 54.76%
```

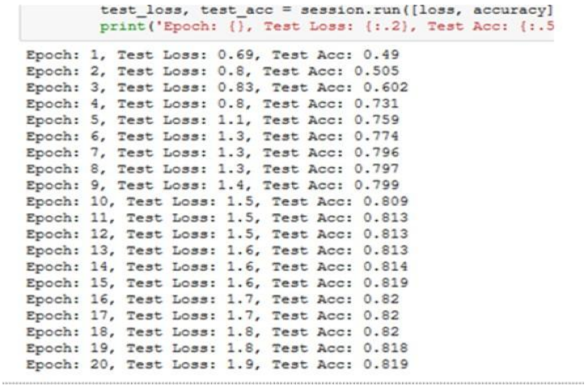
Figure 3. Accuracy of vader



```
runDiagnostics(getTestReviewSentiments(naiveBayesSentimentCalculator))

Accuracy on positive reviews = 73.39%
Accurance on negative reviews = 77.07%
Overall accuracy = 75.23%
```

Figure 4. Accuracy of Naive Bayes



```
        test_loss, test_acc = session.run([loss, accuracy]
        print('Epoch: {}, Test Loss: {:.2}, Test Acc: {:.5

Epoch: 1, Test Loss: 0.69, Test Acc: 0.49
Epoch: 2, Test Loss: 0.8, Test Acc: 0.505
Epoch: 3, Test Loss: 0.83, Test Acc: 0.602
Epoch: 4, Test Loss: 0.8, Test Acc: 0.731
Epoch: 5, Test Loss: 1.1, Test Acc: 0.759
Epoch: 6, Test Loss: 1.3, Test Acc: 0.774
Epoch: 7, Test Loss: 1.3, Test Acc: 0.796
Epoch: 8, Test Loss: 1.3, Test Acc: 0.797
Epoch: 9, Test Loss: 1.4, Test Acc: 0.799
Epoch: 10, Test Loss: 1.5, Test Acc: 0.809
Epoch: 11, Test Loss: 1.5, Test Acc: 0.813
Epoch: 12, Test Loss: 1.5, Test Acc: 0.813
Epoch: 13, Test Loss: 1.6, Test Acc: 0.813
Epoch: 14, Test Loss: 1.6, Test Acc: 0.814
Epoch: 15, Test Loss: 1.6, Test Acc: 0.819
Epoch: 16, Test Loss: 1.7, Test Acc: 0.82
Epoch: 17, Test Loss: 1.7, Test Acc: 0.82
Epoch: 18, Test Loss: 1.8, Test Acc: 0.82
Epoch: 19, Test Loss: 1.8, Test Acc: 0.818
Epoch: 20, Test Loss: 1.9, Test Acc: 0.819
```

Figure 5. Accuracy of RNN

Experiment Results compares accuracy of Vader,Naive Bayes and RNN and shows that RNN is more accurate.

VI.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 , Machine learning based systems , Deep Learning based approaches and found that Deep 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 Min- ing People Attributes and Travel Group Types From Community ContributedPhotos,IEEETRANSACTIONSONMULTIMEDIA,VOL.15,NO.6,OCTOBER2013

[3] Wenjuan Luo1,,FuzhenZhuang, 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 AnalysisInternational Journal of Computer Applications 2015

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

[6] ShengxiangGao, Zhengtao Yu, Linbin Shi, Xin Yan, and Haixia Song Review Expert Collaborative Recommendation Algorithm Based on Topic RelationshipIEEE/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] ArunaSathish,Sentiment Analysis in E-Commerce and Information Security,International Journal of Innovative Research in Computer and Communication Engineering,2016

[9] RajatBhat, SwapnilGaonkar,Sentiment Analysis of Product Reviews using Hadoop,IJSRD - International Journal for Scientific Research Development— Vol. 4, Issue 12, 2017

[10] Evangelos Psomakelis,Konstantinos Tserpes, COMPARING METHODS FOR TWITTER SENTIMENT ANALYSISConference: Conference:International Conference on Knowledge Discovery,2017

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

[12] Anas Collomb, A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation,IEEETRANSACTIONSONMULTIMEDIA,2016

[13]] 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

[14]] Bruno Ohana , Brendan Tierney, Sentiment Classification of Reviews Using SentiWordNet2016

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