ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Design and Implementation of Improved Naive Bayes Algorithm for Sentiment Analysis on Movies Review

Shivangi Sharma

M.Tech Student, Department of C.S.E, S.S.C.E.T Badhani., Pathankot, India

ABSTRACT*:* Social movie review monitoring has been growing day by day so analyzing of social data plays an important role in knowing customer behavior. So we are analyzing Social data such as forms for movie reviews using sentiment analysis which checks the attitude of customer for particular movie brand. This has created a good sensible use of Sentiment Analysis and there has been a lot of innovation during this area in recent days. Sentiment analysis refers to a broad range of fields of text mining, natural language processing and computational linguistics. It historically emphasizes on classification of text document into positive and negative classes. Sentiment analysis of any text document has emerged as the most useful application in the area of sentiment analysis .So our aim is to develop a dictionary based on social media keywords and find hidden relationship pattern from these keywords. The Proposed system is used to find out adjective and nouns forms of various keywords used in the social media standard data available online. This provides hidden relation between different keywords and a dictionary of the keywords on the basis of categories of different comments &tweets. In this, the development of sentiwords dictionary is done using mining algorithm. The dictionary can be further used for classification of words of text documents.

KEYWORDS: data mining, Sentiment analysis, subjectivity, objectivity, sentiword.net, Naïve bayes

I. INTRODUCTION

These days, Social media is turning out to be increasingly famous since cell phones can get to interpersonal organization effortlessly from anyplace. In this way, Social media is turning into a vital theme for research in many fields. As number of individuals utilizing informal organization are developing step by step, to speak with their associates so they can share their own inclination consistently and perspectives are made on extensive scale. Online networking Monitoring or following is most vital subject in today's present situation. In today many organizations have been utilizing Social Media Marketing to publicize their items or brands, so it gets to be distinctly basic for them that they can have the capacity to ascertain the achievement and helpfulness of every item .[2]For Constructing a Social Media Monitoring, different device has been required which includes two segments: one to assess what number of client of their image are pulled in because of their advancement and second to discover what individuals contemplates the specific brand Humors, that have been produced can be assessed normally by performing different Key execution components, for example, the quantity of devotees or companions, the quantity of preferences or shares or remark for every post and more troublesome one like engagement rate, reaction time to assess them and other composite measures. Measuring the Large dataset is typically immediate and should be possible by utilizing some factual method .On the other hand, to assess the supposition of the clients is not as simple as it appears to all clients. For assessing their state of mind may requires to perform Sentiment Analysis, which is characterized as to recognize the extremity of client conduct, the subjective and the feelings of specific record or sentence.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

II. RELATED WORK

BasantAgarwal, Narmita Mittal, Erik Cambria, (2013),this paper represents bi-tagged phrases has been used as features extraction in combination with unigram features for sentiment. Main objective is of designing a machine learning model which can classify a given movie review as positive and negative correctly. Here two types of features are extracted i.e. (i) unigram and (ii) bi-tagged phrase. A Bi-tagged phrase has been extracted using part of speech tag. Here, Dataset has been collected from Cornell Movie Review sites i.e. contain 1000 positive reviews and 1000 negative reviews for movie. Here SVM (support vector machine) and NB (naïve bayes) has been used for classifying the dataset into positive and negative sentiment polarity. Weka tool is used to implement these classifiers. Evaluate these classifier using 10 fold cross validation. Results shows that unigram feature performing individually can give better result compare to bi-gram and bi-tagged feature for both SVM and NB. Further, Bi-tagged phrases consider as features individually not performing well for sentiment classification. But if, bi-tagged phrases combined with the unigram feature can improve the performance of sentiment classification. The main drawback here will be that it is highly computationally expensive [1].

BogdonBatrinca, Philip C. Tr.eleaven, (2014),states an overview of software tool for social media, blogs, chats, newsfeeds etc. and how to use them for scraping, cleansing and analyzing. For scraping the social media it suggests the challenges such as Data cleansing, Data protection, Data analysis and Visualization and analytics Dashboard. This paper presents a survey on methodology of social media, data, providers and analytics techniques such as stream processing, sentimental analysis. An overview of different tools needed for social analysis purpose is also presented. There has been easy availability of APIs provided by Twitter, Facebook and News services which led to explosion of data services for the purpose of scraping and sentiment analysis [2].

Mathew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, (2013), stated how the capabilities of mobile devices are affecting user's privacy. It also presents threat analysis which is classified into two categories i.e. home grown problem in which user upload without sufficient protection which affect user's own privacy. Second, someone is uploading the damage content of other people.It also include privacy analysis of different sites such as flicker, Face book, Picasa web and Google+, Locr and Instagram and PicPlz Instagram and PicPlz. It also presents an analysis of privacy related metadata, particularly location data contain in social media. As it concludes that 10% of all the photos taken by camera devices harm other people's privacy without knowing them. It also represents handling of the location based big data. It includes a concept of watchdog client a server side watchdog service. In it, concept to stay in control from social media uploaded by others has three types of services. Through the regular user account, Operated by the social networks and last one operated by third party i.e. stand alone service [3].

Augustyaniak, Kajdanowicz, Kazienko, (2014), examines two ways to deal with sentiment analysis: lexicon based versus supervised learning in the space of reviews of movies. In assessment, the methodologies were looked at by utilizing a test collection of standard movie review. The outcomes demonstrate that approach based on lexicon is effortlessly outperformed by approach of classification. [28]

III. METHODOLOGY

1. Generating Dataset

Two dataset were collected firstly, from Twitter tweets and secondly, from Online review Dataset. The online review dataset consists of around 800 user's review archived on theIMDB (Internet Movie Database) portal. And for, Twitter dataset around 1000 review were collected and each review were formatted according to .arff file where review text and class label are only two attributes. Class label represent the overall user opinion. Here, we set simple rules for scaling the user review. For dataset, a user rating greater than 6 is considered as positive, between 4 to 6 considered as neutral and less than 4 considered as negative.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

2. Preprocessing.

Collection of raw data and then apply filtering techniques to make that raw data into structured format. For doing the classification, Text preprocessing and feature extraction is a preliminary phase. Preprocessing involves 3 steps:

a) Word parsing and tokenization: In this phase, each user review splits into words of any natural processing language. As movie review contains block of character which are referred to as token.

b) Removal of stop words: Stop words are the words that contain little information so needed to be removed. As by removing them, performance increases. Here, we made a list of around 320 words and created a text file for it. So, at the time of preprocessing we have concluded this stop word so all the words are removed from our dataset i.e. filtered.

c) Stemming: It is defined as a process to reduce the derived words to their original word stem. For example, "talked", "talking", "talks" as based on the root word "talk". We have used Snowball stemmer to reduce the derived word to their origin.

3. Classification

i. Combining naïve bayes with Decision table using Decision tree as Meta classifier.

ii. Meta Learner is a learner scheme that combines the output of the naive bayes and decision table i.e. the base learner. The base learners' level-0 models and the meta-learner is a level-1 model. The predictions of the base learners are input to the meta-learner

Pseudocode:-

Procedure 1: Selecting base and meta classification layers
Input: original-dataset: Dataset, folds: Integer
{
Dataset←original-dataset
define Array-Of-Classifiers as array of classifiers that contains Naïve bayes and Decision Table
For layer = 0 to 2 do:
{
For each fold in folds do:
{
If layer ≠2 do:
Array-Of-Classifiers←Train_SingleLayer (layer, train-set, folds) // Here we are calling classification algos
New-Instances←Classify (test-set, Array-Of-Classifiers)
Add New-Instances to dataset [layer+1]
Else
Train_SingleLayer (layer, train-set, folds)
}
Layer = layer + 1
}
Train-Single-Layer (layer0, original-dataset) //Rebuild Base classifiers using // the original dataset
}
Procedure 2: Classifying a Single layer
Train-Single-Layer Input: Layer
Number: Integer, dataset: Dataset, folds: Integer
Output: Successor-Dataset: Dataset
{
Successor Dataset ← empty Group
For each fold in folds do:
{
Build Classifiers (Layer Number, train-set)
For each instance in test-set

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

{
Produce probabilities-vector by applying instance on current layer's classifiers.
Generate a new Instance from probabilities-vector
Add the new Instance to Successor-Dataset
} }
Return Successor-Dataset
}

4. Analyze the performance

Analyze the performance parameters like FP rate, TP ate, Recall, Precision of Naïve Bayes and new proposed hybridized algorithm and Compare the results of both.

Now for evaluating the result, different parameter are to be calculated. True positive, True negative, False positive and False negative are used for comparing the class label that have been assigned to a document by the classifier with the classes the item actually belongs [18].

Accuracy: It is measured as the proportion of correctly classified instances to the total number of instances being evaluated. Classification performance being evaluated by using this parameter. where True positive – that are truly classified as positive.False positive- not labeled by the classifier as positive but should be True negative- that are truly classified as negative

False negative- not labeled by the classifier as negative but should be positive.

Precision: It is widely used in evaluating the performance in different field such as text mining, information retrieval. Precision is also referred to measure the exactness. It is defined as ratio of the number of correctly labeled as positive to the total numberthat has been classified as positive.

Recall: It is also used in evaluating the performance for text mining and information retrieval. It is also used to measure the completeness of the model. It is defined as the ratio of the number of correctly labeled as positive to the total number that are truly positive.

F-measure: It is referred as the harmonic mean of precision and recall. It helps to give score needed to balance between precision and recall. It combines two of them into one for the convenience as it might optimize the system so that it can favor one of them.

5. Combined dictionary

Combined word of twitter dataset and online review dataset forms a dictionary. As after classifying each word probability as positive, negative and neutral. Compare the probability for each word and categorize each word into three different dictionaries based on highest polarity i.e. positive, negative and neutral of each word. Dataset is used for further evolution of words depending on their uses in daily life as adjectives or nouns in the social media data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

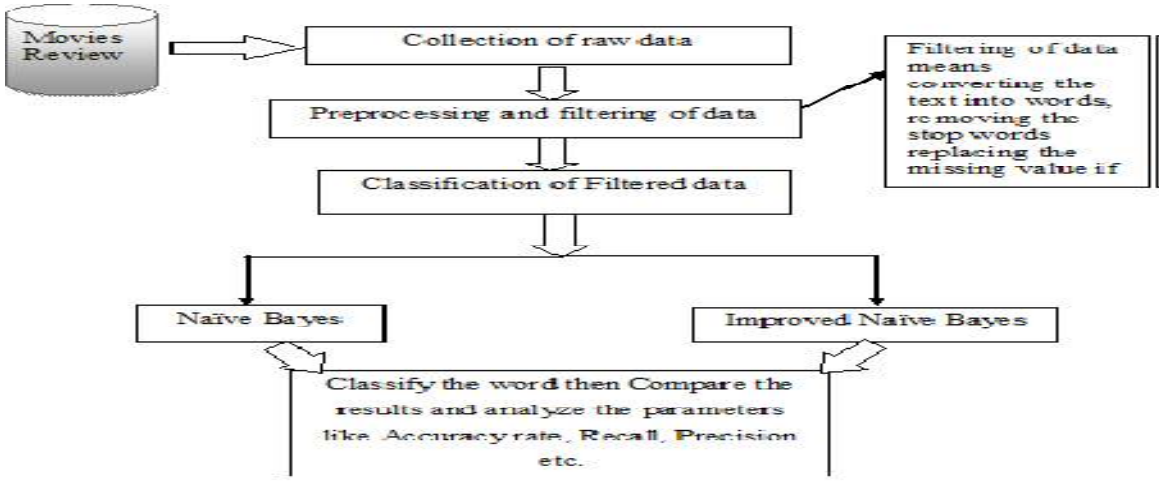
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Flowchart of Proposed Work



IV. RESULTS

The proposed methodology is implemented with the help of Weka and Net beansIDE8.0.Weka is the library that provides the simulation environment of data mining and also provide primary classes for evaluating the classification model.

Table 1: showing the Class Detail parameters comparison.

Algorithm	Tp rate	Fp rate	Precision	Recall	F-measure	ROC area
Naïve Bayes	0.901	0.64	0.901	0.901	0.901	0.765
Improved Naïve Bayes	0.972	0.175	0.97	0.972	0.97	0.863

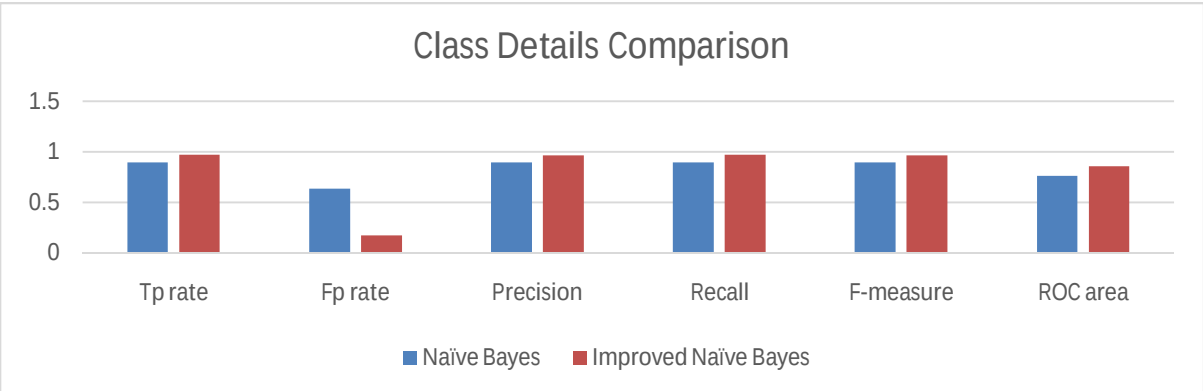


Figure 1: Showing the Class details parameters comparison of naïve bayes and improved naïve bayes

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Algorithms	Correctly Classified	Incorrectly classified
Naïve Bayes	254	28
Improved Naïve Bayes	247	8

Table 2: showing the Correctly and incorrectly classified instances comparison

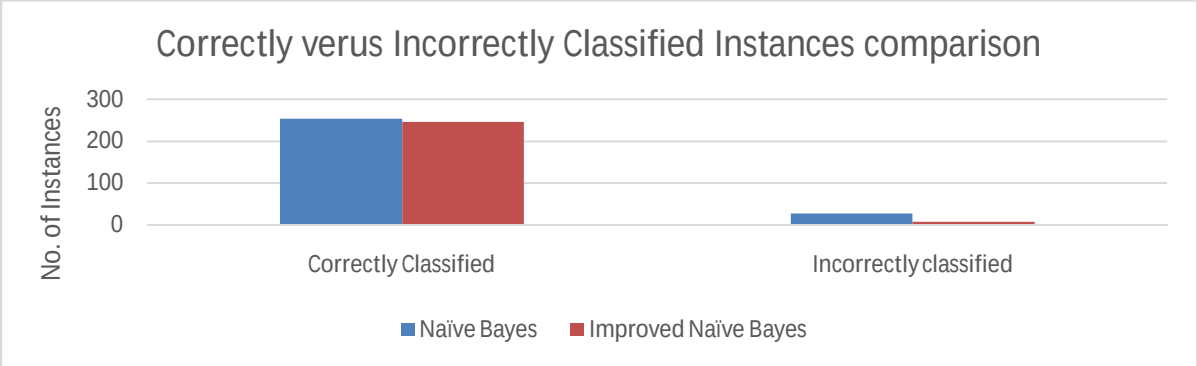


Figure2: Showing the Correctly and incorrectly classified instances comparision.

V. CONCLUSION AND FUTURE WORK

Social media monitoring has been growing very rapidly so there is a need for various organizations to analyze customer behavior or attitude of particular product or any movie review. So, the concepts of sentiment analysis have been introduced. Text analytics and sentiment analysis can help organization to derive valuable business insights. Attitude can be calculated based on polarity check. Sentiment analysis refers to a broad range of fields of natural language processing, computational linguistics, and text mining. Sentiment classification of reviews and comments has merged as the most useful application in the area of sentiment analysis Bag of words and feature based sentiment are the most popular approaches used by researchers to deal with sentiment analysis of opinions about products such as movies etc. Sentiment analysis on movies review are done by forming dictionary using Naïve Bayes Multinomial Algorithm which shows that it is easier to build dictionary on phrases of movies reviews.In this,level sentiment analysis is considering three classes for sentiment polarity of each sentence (positive, neutral and negative). Each class prediction and classification is done by algorithm in terms of accuracy, precision, recall, TP rate, FP rate, F-measure, ROC area.Also, the comparison of Naïve bayes with naïve bayes multinomial is done on the basis of accuracy or the correctly classified instances. Improved Naïve bayes performs better than naïve bayes with accuracy 97.1631% . The testing of the data using the proposed dictionary is done in which the text is classified based on the words in the dictionary.
Future work include to determine features for the movie in detail i.e. make polarity check on different features such as actors, directors, scripts, music etc. and make the dictionary for them. Also consider the case of Twitter as tweets consist of short hands as online review were written in more clear way as compared to Tweets. So, form hidden relationship between different keywords and a dictionary of the words on the basis of different categories of comments & tweets.

REFERENCES

1. F. F. Moghaddam, M. Vala, M. Ahmadi, T. Khodadadi, and K. Madadipouya, "A reliable data protection model based on re-encryption concepts in cloud environments," 2015 IEEE 6th Control and System Graduate Research Colloquium (ICSGRC), pp. 11–16, 2015
2. A. Singh and H. Singh, "An improved LSB based image steganography technique for RGB images," *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–4, 2015.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

3. S. M. Gurav, L. S. Gawade, P. K. Rane, and N. R. Khochare, "Graphical password authentication: Cloud securing scheme," *2014 International Conference on Electronic Systems, Signal Processing and Computing Technologies*, pp. 479–483, 2014.
4. Mukundan, R.; Madria, S.; Linderman, M. (2014) "Efficient integrity verification of replicated data in cloud using homomorphic encryption", *Springer Distributed and Parallel Database*, vol. 32, Issue 4, pp. 507-534, 24 June 2014
5. Abdullah, A., Hashim, F., & Al-Haddad, S. (2014) " A review of cloud security based on cryptographic mechanisms",*IEEE, Kuala Lumpur*, pp. 106-111.
6. Banirostam h., & Hedayati, A. (2013) "A Trust Based Approach for increasing Security in Cloud Computing Infrastructure" *International Conference on computer modeling and simulation, IEEE*, Cambridge, pp. 717-721.
7. Kang, A.N.; Barolli, L.; Park, J.H.; Jeong, Y.S. (2013) "A strengthening plan for enterprise information security based on cloud computing", *Springer cluster computing*, vol. 17, Issue 3, pp. 703-710, September 2013
8. Du, Y.; Zhang, R.; Li, M. (2013) "Research on security mechanism for cloud computing based on virtualization" *Springer Telecommunication systems*, Vol. 53, Issue 1, pp. 19-24, 2013
9. Chen, D., & Zhao, H. (2012). "Data Security and Privacy Protection in cloud computing." *2012 International Conference on Computer Science and Electronics Engineering (ICCSEE)*, pp. 647-651,2012.
10. Abuhussein, A., Bedi, H., & Shiva, S. (2012) "Evaluating Security and Privacy in Cloud Computing Services: A Stakeholder's Perspective", *2012 International Conference for Internet Technology and Secured Transactions*, pp.388-395,2012.
11. M.-H. M. Guo, H.-T. H. Liaw, L.-L. Hsiao, C.-Y. Huang, and C.-T. Yen, "Authentication using graphical password in cloud," *2012 15th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, pp. 177–181, 2012.
12. Pengfai dai and Chaokun Wang , "Software Watermark Approach Based Architecture for cloud Security", *14th Asia-Pacific Web Conference, APWeb 2012, Kunming, China*, pp. 270-281, 2012.
13. Zhan Xin, Research on cloud computing data security Model based on multi-dimension, IEEE, 2012
14. Zhao, G., Rong, C., &Jaatun, M. G. "Reference deployment models for eliminating user concerns on cloud security" Journal of supercomputing, Vol 61, Issue 2, pp 337-352, 2010.
15. Usha, S., Kumar, G. A. S., and Boopathy bagan, K., A secure triple level encryption method using cryptography and steganography, *Computer Science and Network Technology(ICCSNT), International Conference*, pp. 1017-1020,2011.
16. Mohemed Almorsy Collaboration-Based Cloud Computing Security Management Framework, *IEEE International Conference on Cloud Computing (CLOUD)*,2011
17. Josel, G. A., & Sajeev. "Implementation of Data Security in cloud Computing", *International Journal of P2P Networks Trends and Technology*,2011.
18. Shaikh, F. B., & Haider, S. "Security Threats in Cloud Computing", *IEEE*, pp 11-14, 2011.
19. B. S. Park, A. J. Choudhury, T. Y. Kim, and H. J. Lee, "A study on Password Input method using authentication Pattern and Puzzle," *2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, pp. 698–701, 2011.
20. Marwaha, P., Visual cryptographic steganography in images, *Communication and Networking Technologies(ICCCNT), International Conference*, pp 1-6, 2010