ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

Sentiment analysis of data using Naive Bayes Classifier with n-gram approach in Hadoop

Priyanshu Jadon, Miss Rupali Dave

P.G. Student, Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India
Assistant Professor, Department of Computer Science, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

ABSTRACT: Today, we are living in the world where the data is present in enormous amount. There are many web applications and social networking sites where users will provide their sentiments on anything. It becomes crucial to find out the user's sentiment or opinion toward something. Sentiment analysis is most versatile research areas in field of real time knowledge extraction. Real time data analysis can plays very crucial role to observe the thinking and view point of people and users. Analysis of social networking data can help a lot to observe trend of society. It can also help to derive user interest and hidden activities. Sentiment analysis is the approach to determine whether piece of writing is positive, negative or neutral. It also helps to derive user opinion and attitude of writer. Sentiment analysis provides the great practical value on user's view point. Social networking sites are a platform where users share their different view points on certain products. Every day lot of viewpoints of users is generated on an internet about any product, place or a person. Sentiment analysis is a research area which extracts the proper meaning of the user's viewpoint that includes text analytics and classifies the polarity of the user's opinion. Sentiment analysis is also known as appraisal extraction and subjectivity analysis. This work has proposed sentiment analysis model to observe positive and negative view point of different user based on sentiment analysis approach. Here we are using an n-gram approach with a naïve bayes classifier and finding the overall opinion of the sentence regarding positive and negative. We are performing the naïve bayes classifier for the complete sentence. We validate the n-gram approach up to 3-gram. Here, we are also checking the accuracy concerning precision and recall factors and the complete work will be implemented using Hadoop Ecosystem to perform parallel processing on large data. The benefits of our work is that it may gives the overall opinion of the users viewpoint and as Naïve Bayes give a relevant and accurate result and is simple to use.

KEYWORDS: Sentiment Analysis, Naive Bayes Algorithm, n-gram, Hadoop Ecosystem.

I.INTRODUCTION

Sentiment analysis is a novel technique that facilitates to investigate the thinking and thoughts of user. It can be defined as: "Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes." Sentiment analysis refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

In general, the aim of sentiment analysis is to determine the attitude of a user with respect to some topic. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader). A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature level—whether the expressed opinion in a document, a sentence or an entity which shows the given feature is positive, negative, or neutral. In advance, sentiment analysis also classifies the sentence on the basis of emotional states such as "happy", "sad", and "angry".

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

A.CLASSIFICATION OF SENTIMENTAL ANALYSIS:

Sentimental Analysis is classified using two approaches: Lexicon Based Approach and Machine Learning Approach.

1. Lexicon-based Approach:
Lexicon based approach works on supporting sentiment counts and weight. With consideration of labeling, on the basis of corpus based, dictionary based and manual approach, viewpoints are integrated.
Lexicon-based approach is classified as: Corpus based approach and Dictionary approach.

2. Machine Learning Approach:
Artificial Intelligence is sub-sectioned to form Machine learning. Machine learning is studied in the form of algorithm and can be understand easily.

1.5.1 Sentimental Analysis on the basis of Machine Learning Approach:
On the basis of machine learning sentimental analysis is classified as:

1. **Unsupervised Learning:**
   From hidden labels, unlabeled data are categorized using unsupervised learning. Although, learner have unlabeled so no error and no potential solution is evaluated.
1. **Supervised Learning:**
   On the basis of common patterns, labelled data are categorized using supervised approach. In testing phase data is classified and data patterns are examined using training data set. These approaches are used to observe and predict data movement. Trained data are analysed and produces functions for the generation and examining of data class for complete data sample

B. HADOOP ECOSYSTEM

Hadoop Overview and Architecture:

In 1990's GOOGLE suffers to store the huge amount of data, so in 2003 they built a Google File System [GFS] and in 2004 Google comes with a concept of Map reduce and gives a white paper related to GFS and MR where they mainly give some idea that how it works and suggests some techniques but not implement it. Later around in 2006 yahoo have taken white paper and start working on that idea and the same time another person dough cutting also working on storing huge amount of data. Later dough cutting start working with yahoo and gives a concept of hadoop. Later it goes to the Apache Foundation who makes Hadoop a open source tool.
Hadoop is a open source framework given by Apache software foundation for storing and processing huge data sets. It mainly have two basic components Distributed File System [HDFS] and MapReduce [MR]. HDFS is mainly used to store the data that is coming from commodity hardware and MR is used to process that data which is stored in HDFS in Key Value [K, V] pairs.
Hadoop is a framework that is consists of various other programming tools that is helpful for processing the complex datasets like PIG, HIVE, SPARK, OOZIE, HBASE, FLUME, SQOOP, MAHOUT etc. Every programming tool has some unique functionality that is used depending upon its need. The programming tools available in framework are mainly used for big data analytics while transferring the flow of data to the different tools, depends on need, may cause security and privacy challenges. Different version of are available, previous one is 0.1 and latest one is 0.2. In this paper, we study about the 0.1 and the security and privacy challenges present in it. As shown in fig 1 HDFS mainly have three components Name Node, Data Node, and Secondary Name Node while Map Reduce have two components Job Tracker, and Task Tracker. Name Node is a namespace which have all the information of or we can say that Meta data is stored in Name Node and Name Node is available in RAM. Data Node is actually used to store the data. Job Tracker is a job assign by Name Node to Data Node. Task Tracker is a task of Data Node which is executing by itself. If any of the data nodes fails, then here by default 3 copies of data are available at different data nodes. By-default size of block in 0.1 version is 64 block sizes. In first the data is stored in HDFS and after storing the large datasets we start to process it and further it is used to perform analytics and getting some valuable assets.

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

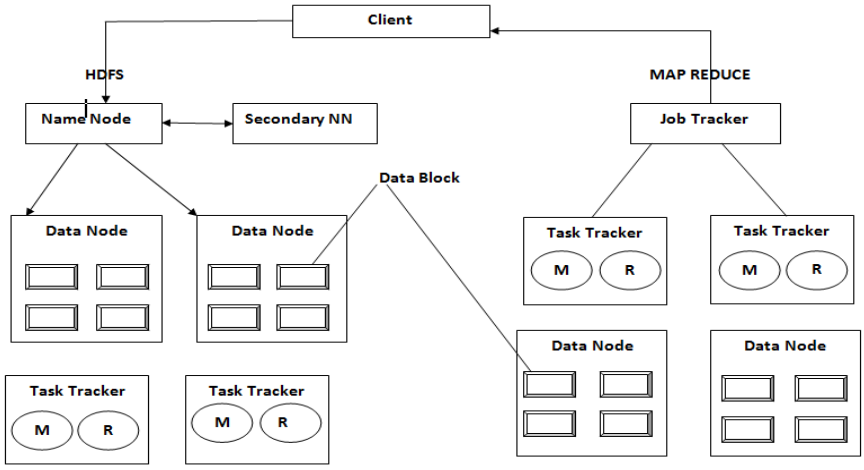

Fig 1: 0.1 Architecture

Name Node, Secondary Name Node, and Job Tracker act as Master services and Data Node and Task Tracker act as Slave services, hence also called as Master-Slave service. Every Master Service will talk with each other and every slave service will talk with each other. Name Node is a master service and its corresponding slave service is data node and they talk with each other. Similarly, Job Tracker is a master service and its corresponding slave service is Data Node and they talk with each other.

II. SIGNIFICANCE OF THE SYSTEM

The paper mainly focuses on how Naive Bayes classifier with n-gram approach is useful in finding the sentiment of the user sentence. The study of literature survey is presented in section III, Methodology is explained in section IV, section V covers the experimental results of the study, and section VI discusses Conclusion.

III. LITERATURE SURVEY

Trupthi et. al. [1] explore that sentiment analysis and opinion mining can be used to know about user thinking and sentiments for eventual way point. They consider twitter data as major source for real data examination and uses Naive Bayes theorem with unigram approach for feature extraction and data classification. The complete proposed architecture is shown below;
This work considers positive words, degree of positive words, positive tweets, and overall tweets to observe the trend of user opening. Afterwards they accumulate the overall trend and conclude with opinion of overall country. Although they proposed a good solution for sentiment analysis still it lack with few limitation which can be considered as scope of improvement for next version of research.
In research done by Walaa Medhat, Ahmed Hassan, Hada Korashy on paper entitled “Sentiment analysis algorithms and applications: A survey” that gives the overview of Sentiment analysis and the various techniques used in it [14].
M. Mazhar Rathore et al. In [7] introduces about Geosocial Networks which is government liability in terms of safety from any kind of disasters. Providing with proper facility from management and decrease in risk of the spread of infection. Common citizens are recommended by system and also citizens are provided with recommended system, healthcare system etc. And also new products can be launched in different fields by monitoring Geosocial data of specific location. For better analysis benefits are provided for employing significant data generation from Geosocial network. A high computing capabilities with better analysis and advanced technology is possible. That is why; a system is proposed by author with better planning, provided with proper management and safety from disasters. This system provides with high speed data in Geosocial networks and also process and analyse and make decisions. Author used twitter data and worked it on Hadoop using spark.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

Cambria E, Olsher D, Schuller B et al. In [8][9][10] introduces an effective progress in opinion mining which is dealed better earlier. Many different scenarios are developed to deal with it in recent year with achieving document level sentimental analysis. Sentimental analysis faces many issues for subjective detection and classification.
Turney et al. In [11] worked on method called as bag-of-word method, where relationships among words in a sentence are not considered for sentimental analysis. Sentence is a collection of words and sentiment of the complete sentence is described as determining sentiment of every single word and using some functions values are aggregated.
B. Liu et al. In [13] introduces Opinion Lexicon consisting of annotations which are positive and these annotations are opinion lexicon which are manually chosen. Opinion lexicon words for positive and negative words are consisted in dictionary with manual annotations. Moreover, it consists of frequency of words that are misspelled
Suchita V Wawre et al. In [18] proposed comparison between SVM and Naive Bayes classifier. These classifiers are the supervised machine learning algorithm for sentimental analysis. Both the approaches does not compares with each other if they have less reviews.

IV. METHODOLOGY

Methodology can help to draw original blueprint of proposed solution. This work proposed the use of Naïve Bayes classifier along with n-gram approach to find out the true sentiment of the user while twitting anything in twitter. The main objective of our work is to analysis the large dataset to extract the user information. We are trying to evaluate our work on the basis of computation time and accuracy. The accuracy is going to be evaluated by a precision and recall factor.

Naïve Bayes Formula :
P(a/b) = P(b/a) * P(a) \ P(b)
Where, P(a) = No. of positive sentence \ Total number of sentences (evidence probability)
P(b/a) = Positive score \ Total number of positive words (prior probability)
P(b) = Score of positive word + score of negative word \ Total no. of words. (probability of evidence)
P(a/b) = Probability of hypothesis given that evidence is there.

n-gram approach :
n gram = the consequence sequence of words
Eg : The cow jumps over the moon
First lemmatize it.
#, cow, jumps, over, moon (In n-gram # is the first character)
Bi-gram = (#, cow), (cow, jump), (jump, over), (over, moon)

The below is proposed algorithm of work:
Dataset
Input: Sentence from dataset, n-gram
Output: Positive and negative Sentiment

1. Start
2. Load file in hadoop HDFS.
3. Enter Dataset D1,D2,D3,…Dn.
4. Train and Test the data
5. Lemmatize the dataset.
6. Apply n-gram
7. Analysis the sentence by naïve-bayes approach.
8. Result as positive and negative sentiment.
9. Find the performance of 1-gram, 2-gram, and 3-gram by precision and accuracy factor.
10. Precision = [relevant document] intersection [retrieved document] / [relevant document]
11. Exit

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

V. EXPERIMENTAL RESULTS

First table shows the analysis of the above sentences on the basis of uni-gram approach. Here, we are doing the previous work. Here we see that from 7 sentences we get only 3 as a relevant result. So we say that, uni-gram is not giving the proper analysis of sentence.
Precision = {Relevant Document} intersection {Retrieved Document} / {Retrieved Document}
Precision = 0.428571429

SENTENCES	EXPECTED OUTCOME	PREDICTED OUTCOME	RELEVANT
S1	P	N	NO
S2	P	P	YES
S3	P	P	YES
S4	P	P	YES
S5	N	P	NO
S6	N	P	NO
S7	N	P	NO
TOTAL		7	3

Fig 6.1: Table for naive bayes Uni-gram

Second table shows the analysis of the above sentences on the basis of bi-gram approach. Here, we are doing our proposed work. Here we see that from 7 sentences we get 5 as a relevant result. So we say that, bi-gram is giving the proper analysis of sentence and is more accurate then the uni-gram approach.
Precision = {Relevant Document} intersection {Retrieved Document} / {Retrieved Document}
Precision = 0.714285714

SENTENCES	EXPECTED OUTCOME	PREDICTED OUTCOME	RELEVANT
S1	P	P	YES
S2	P	P	YES
S3	P	P	YES
S4	P	P	YES
S5	N	N	YES
S6	N	P	NO
S7	N	P	NO
TOTAL		7	5

Fig 6.2: Table for naive bayes Bi-gram

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 5, Issue 4 , April 2018

Third table shows the analysis of the above sentences on the basis of tri-gram approach. Here, we are doing our proposed work. Here we see that from 7 sentences we get 6 as a relevant result. So we say that, tri-gram is giving the proper analysis of sentence and is more accurate then the uni-gram and tri-gram approach.
Precision = {Relevant Document} intersection {Retrieved Document} / {Retrieved Document}
Precision = 0.857142857

SENTENCES	EXPECTED OUTCOME	PREDICTED OUTCOME	RELEVANT
S1	P	P	YES
S2	P	P	YES
S3	P	P	YES
S4	P	P	YES
S5	N	N	YES
S6	N	N	YES
S7	N	P	NO
TOTAL		7	6

Fig 6.3: Table for naive bayes Tri-gram

A. PRECISION GRAPH:

The below graph shows that the previous method (uni-gram approach), is giving the bad result as compare to the proposed approach. Among all tri-gram gives the maximum accuracy.

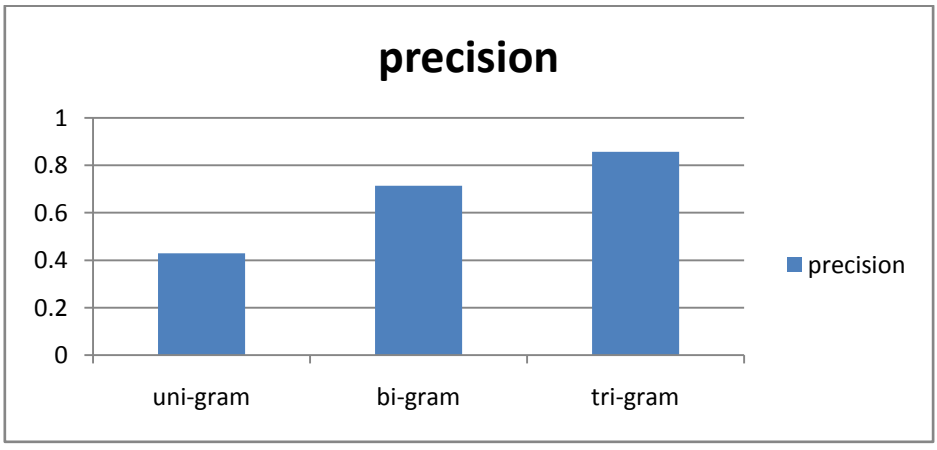

Fig 6.4: Precision graph of uni-gram, bi-gram, tri-gram

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 5, Issue 4 , April 2018

VI. CONCLUSION

This research work will gives brief background of sentiment analysis. It will carry out with an approach to observing user view point and citizen opening to observe current trend of user thinking. This work will be implemented using hadoop ecosystem and evaluated for single and multi node cluster. The complete work will end with the performance observation of sentiment analysis using data mining for traditional system and proposed solution based on computation time, memory consumption, accuracy and relevant factor for adoption of theory. We conclude from above that naive bayes bigram and trigram approach gives a proper analysis of the user's sentiment instead of the naivebayes unigram approach.

REFERENCES

1. Sunil Ray "Understanding Support Vector Machine algorithm from examples" sep 13, 2017.
2. Sunil Ray, " 6 easy steps to learn Nave Bayes algorithm" sep 11,2017.
3. M.Trupthi, Suresh Pabboju, G.Narasimha, "Sentiment analysis on twitter using streaming API", 2017 IEEE 7th International Advance Computing Conference, pp. 915-919.
4. Divya Sehgal, Dr. Ambuj Kumar Agrawal, "Sentiment Analysis of Big Data Applications using Twitter Data with the help of Hadoop Framework", International Conference on System Modeling and Advancement in research Trends, 25th-27th November, 2016, pp.251-255.
5. Walaa Medhat, Ahmed Hassan, Hada Korashy, "Sentiment analysis algorithms and applications : A survey". Ain Shams Engineering Journal, vol 5, issue 4, dec-2014, pp. 1093-1113.
6. A Kowcika and Aditi Guptha "*Sentiment Analysis for Social Media"*,International Journal of Advanced Research in Computer Science and Software Engineering, 216-221,Volume 3,Issue 7, July 2013.
7. [3] M. Mazhar Rathore, Anand Paul, Awais Ahmad, "Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions". IEEE ICC 2017 SAC Symposium Big Data Networking Track.
8. [4] Cambria E, Hussain A. Sentic computing: techniques, tools, and applications. Dordrecht: Springer; 2012.
9. [5] Cambria E, Olsher D, Rajagopal D. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: AAAI. Quebec City; 2014. p. 1515–21.
10. [6] Cambria E, Schuller B, Xia Y, Havasi C. New avenues in opinion mining and sentiment analysis. IEEE Intell Syst. 2013;28(2):15–21.
11. [7] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417–424, Association for Computational Linguistics, 2002.
12. B. Liu, "Sentiment analysis and subjectivity," Handbook of Natural Language Processing,, pp. 627–666, 2010.
13. Suchita V Wawre1, Sachin N Deshmukh2 , "Sentiment Classification using Machine Learning Techniques", International Journal of Science and Research (IJSR) Volume 5 Issue 4, April 2016.
14. Walaa Medhat, Ahmed Hassan, Hada Korashy, "Sentiment analysis algorithms and applications : A survey". Ain Shams Engineering Journal, vol 5, issue 4, dec-2014, pp. 1093-1113

-