Comparison between the colleges on basis of their ratings obtained by Machine Learning Algorithm and Web Mining

College Rating System

[1]Saee Sunil Jadhav, [2] Divya Anil Vispute, [3]Sanjay S. Kadam
[1,2]Bachelor of Engineering Student, [3]Assistant Professor
[1,2,3]Department of Information Technology,
[1,2,3]Bharati Vidyapeeth College of Engineering, Kharghar, Navi Mumbai, India

Abstract: In today's world, a large amount of data is collected over the web with the help of comments, surveys, and reviews. Tons of reviews on the internet for various services are available, so for us, it gets difficult to analyze and track the reviews of the users. Opinion mining also is known as sentiment analysis or emotion AI which uses natural language processing, text analysis, computational linguistics to systematically identify, extract, quantify, and study effective states and subjective information in the source material, and helping a business or organization to understand the user's sentiment towards their brand, product or service. This project is a thorough effort to seek the application and elongation of present work in the field of opinion mining on the fetched data, related to reviews of colleges from other websites with the help of web mining technique, by using Naive Bayes and decision list classifiers, given reviews can be marked as positive or negative. Such reviews have been analyzed in recent studies and a lot of useful information about colleges like user needs, suggestions for improvements, user opinion about some specific features and along with the description of their experience. This paper highlights a comparison between the results as positive and negative obtained by exploiting the Naive Bayes machine learning algorithm to give ratings.

IndexTerms - **Machine Learning, Opinion Mining, Web Mining, Natural Language Processing, Naive-Bayes Algorithm.**

I. INTRODUCTION

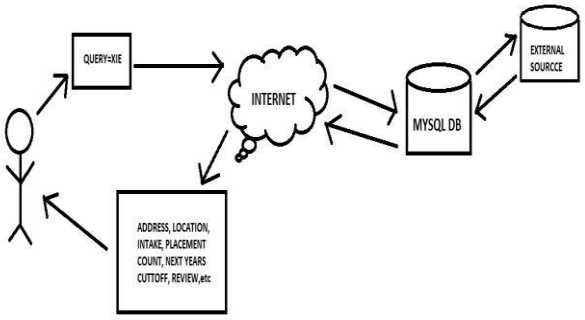
Social Media has gained the attention of the entire world From a user's perspective, people are able to post their own content through various social media platforms. Reviews of people have been one of the most important sources for various services in ever-growing popular networks to manage their firms, businesses, industries, colleges, and their reputation.[1] with the help of reviews or opinions given by the user on many websites are useful to the analyzed experience of users, their requirements, some specific issue regarding the system features.[3] these reviews can be used by the colleges to improve their system and students can choose the colleges wisely.

II. PROBLEM DEFINITION

In the real world, there is no such system wherein we get details of all the engineering colleges in Mumbai under one web application.
To save on the users time of hopping from one web application to another just to seek information our system will provide details such as an address of the college, contact, number, location, photos of the infrastructure, and review of the college as well along with the ratings.

III. PROPOSED SYSTEM

To overcome the drawbacks of the existing system, we can propose a system that extracts reviews of the college from other websites using web mining technique and analyzes these reviews using sentiment analysis.
Sentiment analysis or web mining is the process of automatically extracting knowledge from reviews of others about some topic. Reviews can be identified in a large unstructured/structured data and analyze polarity of reviews.[1]
To tag a given review as positive or negative we can use Naive Bayes algorithm for analysis of college reviews in the system we have proposed.[5] To improve the college system, these results can be used for various purposes such as guiding decisions [5].



III. RELATED WORK

Alekh Agarwal et al., proposed a machine learning method incorporating linguistic knowledge gathered through synonymy graphs, for effective opinion classification. The degree of influence among relationships of documents have on their sentiment analysis is shown in this approach. This is brought about by the use of opinion words and graph-cut technique got through synonymy graphs of Wordnet. An improvement in the accuracy of predictions in the classification task is achieved in the proposed approach. Experiments results with an accuracy of over 90% have been given by this system, with an added advantage of the reduction in processing time, with minimal difference in final accuracies.[8]

Lina Zhou et al., investigated movie review mining using semantic orientation and machine learning. Text classification and supervised classification techniques to classify the movie review are used in the proposed machine learning approach. A collection of text is formed to represent the data in the documents and all the classifiers are trained using this collected data. Thus, more efficiency is shown in the proposed technique. The machine learning approach uses supervised learning, the proposed semantic orientation approach uses "unsupervised learning" because it does not require prior training in order to mine the data. Supervised approach achieved 84.49% accuracy in three-fold cross-validation and 66.27% accuracy on hold-out samples are shown with the help of experimental results. 77% accuracy of movie reviews has been achieved by the proposed semantic orientation approach. Thus, the study concludes that the supervised machine learning requires a considerable amount of time to train the model but is more efficient. On the other hand, the semantic orientation approach is more efficient but is slightly less accurate to use in real-time applications. It is practicable to automatically mine opinions from unstructured data is confirmed from the results.[10]

Bo Pang et al., used machine learning techniques to investigate the effectiveness of classification of documents by overall sentiment. Experiments have demonstrated that the machine learning techniques are better than human produced baseline for sentiment analysis on movie review data. Movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews consist of the experimental setup. Features based on bigrams and unigrams are used for classification. Learning methods maximum entropy classification, Naïve Bayes and support vector machines were employed. Inferences made by Pang et al., for sentiment classification machine learning techniques are better than human baselines. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic-based categorization.[9]

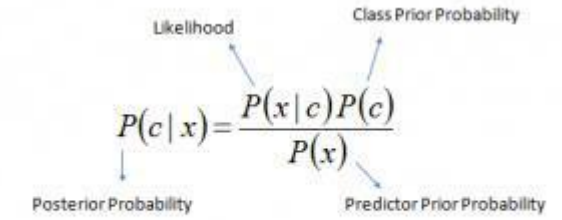
IV. NAIVE-BAYES CLASSIFICATION ALGORITHM

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification.[6] Assumes an underlying probabilistic model and uncertainty about the model is captured in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.[7]

Practical learning algorithms are provided by Bayesian classification and prior knowledge and the data observed can be combined. A useful perspective for understanding is provided by Bayesian Classification and to evaluate many learning algorithms. Explicit probabilities for hypothesis are calculated and is robust to noise in data which is given as input.

Naive Bayes is a model which is easy to build and is particularly useful for data sets which are very large. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.[11]

A way of calculating the posterior probability is provided by Bayes theorem, which is P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

[11]

Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*[11].

V. ALGORITHM WORKING

Let's understand it using an example. Below we have a training data set of college reviews and corresponding target variable 'Positive and Negative' (suggesting possibilities of positive and negative rating).

Doc	Text	Class
1	I loved the college	+
2	I hated the college	-
3	A great college. Good staff	+
4	Poor staff	-
5	Great college. a good canteen	+

Now, we need to classify whether colleges will have a positive rating or not based on reviews. Let's follow the below steps to perform it.
Step 1: Dataset is to be converted into a frequency table.

Step 2: Create a Likelihood table by finding the probabilities like Positive probability = 0.6.

Doc	Occurrence	Probability	Class
I	2	0.2	+
loved	1	0.1	+
hated	1	0.1	-
good	2	0.2	+
poor	1	0.1	-

Likehood Table				
Doc	p(+)	p(-)		
I	2		2/10	0.2
Loved	1		1/10	0.1
Hated		3	3/10	0.3
good	2		2/10	0.2
poor		2	1/10	0.1
All	5	5		
	5/10	2/10		
	0.5	0.2		

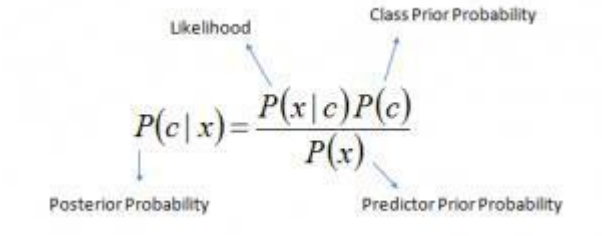
Step 3: Now, To calculate the posterior probability for each class using a Naive Bayesian equation.[11]

The outcome of prediction about college,

Dictionary Generation - Count occurrence of all word in our whole data set and make a dictionary of some most frequent words. The feature set Generation - All the documents are represented as a feature vector over the space of dictionary words. - For each document, keep track of dictionary words along with their number of occurrence in that document. Calculate Probability of occurrence of each label. Here label is negative and positive.
Training -
In this phase, we have to generate training data (words with the probability of occurrence in positive/negative train data files). Calculate for each label. Calculate for each dictionary words and store the result (Here: label will be negative and positive). For each of the defined label now we have a word and corresponding probability.
Document= {+,-}
1. Represent each document by a vector of words
   -one attribute per word position in the document
2. Learning: Use training example to estimate
   -P(+)
   -P(-)

-P(doc|+)
-P(doc|-)

FORMULA :



10 Unique words:
<I, loved, the, college, hated, a, great, poor, acting, good>

p(+)= 3/5=0.6
p(-)= 2/5=0.4

FORMULA : $P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$

VI. MODULE DESCRIPTION

A. Login

In this module, we use a username and password for a user to login into the system. In this login system authentication of the user is done so only valid person can login into the system.

B. Data Collection

In this module, the user selects one college name from a college list and clicks on submit. After submitting our system display the reviews of this college using web mining technique. In web mining technique the system gets data from other websites where college reviews are present related to the college selected by the user.

C. Sentiment Analysis

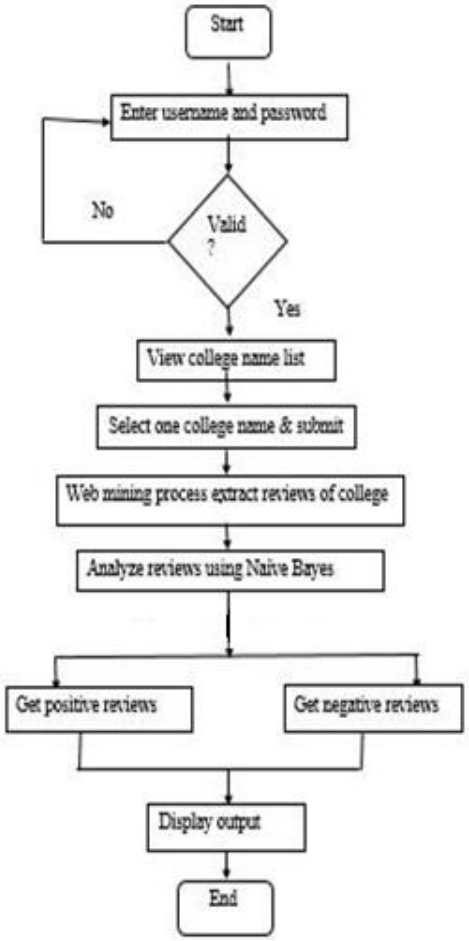
The reviews retrieved using web mining technique from other websites which can be analyzed using Machine Learning Naive Bayes algorithm and get the result as positive and negative reviews.

D. Output

In this module, an output is displayed to the user. The output display college information, Placement, Teaching or faculty, Crowd and a display pie chart of positive and negative percentage and maximum 10 comments

VII. FLOW CHART

1. The user will log into the system by entering his valid UserID and Password.
2. The user then gets to view a huge list of engineering colleges in Mumbai.
3. Then the user clicks on one of the colleges to acquire data.
4. Web mining is performed to extract reviews from other websites.
5. Analysis of those reviews is done using Machine Learning Naive Bayes Algorithm.
6. Reviews are segregated into two parts, i.e. positive and negative and are displayed to the user.



Flowchart

VIII. CONCLUSION

Data can be mined and useful information can be analyzed through the process of sentiment analysis. Various methods which show the impact and applications of sentiment analysis using Twitter were discussed in this paper. We can combine different techniques to overcome their individual drawbacks and enhance the performance of sentiment analysis. Extraction of web page content is extremely useful and essential as it is the basis of many other technologies about data mining. Its main aim is to extract the information which is most relevant and worthy from data-intensive web pages which are filled with noise.

IX. REFERENCES

[1] Heema Krishna, M.Sudheep Elayidom, T.Santhanakrishna, "Impact and Application of Sentiment Analysis using Twitter: A Survey", June 2015.

[2] Badr Hssina, Abdelkarim Merbouha, Hanane Ezzikouri Mohammed Erritali, Belaid Bouikhalene, "An Implementation Of Web Content Extraction Using Mining Techniques", Dec 2013.

[3] Renata Maria, Abrantes Baracho, Gabriel Caires Silva, Luiz G F Ferreira, "Sentiment analysis in social networks: a study on vehicle"

[4] Emitza Guzman, Walid Maalej, "How do users like this feature? A fine grained sentiment analysis of app reviews".

[5] Omkar Borade, Kaushik Gosavi, Ajay Shinde, Avinash Gowda 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939

[6] G Angulakshmi, Dr.R.Manicka Chezian, "Three-level Feature Extraction for Sentiment Classification", August 2014

[7] S.ChandraKala, C.Sindhu, "Opinion Mining And Sentiment Classification: ASurvey".

[8] Alekh Agarwal and Pushpak Bhattacharyya, "Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified", In Proceedings of the International Conference on Natural Language Processing (ICON), 2005.

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan,
"Thumbs up? Sentiment classification using machine learning techniques", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79– 86, 2002.

[10] Lina Zhou, Pimwadee Chaovalit**,** "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", Proceedings of the 38th Hawaii
International Conference on system sciences, 2005.

[11] Naive Bayes

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/