Engineering College Decision Making Assistant

Prof.Suvarna Bhoir, Pushkaraj Prasad Arekar [1], Aadarsh Sanjay Mishra[2], Cleon Ozzie Rodrigues [3]

[1,2,3]Dept. of Information Technology, Xavier Institute Of Engineering, Mumbai

suvarnabhoir@gmail.com pushkar.arekar@gmail.com
aadarshmishra506@gmail.com
cleonrodriguez21@gmail.com

***Abstract*— People can voice their opinion which has allowed them to write a review and comment about how they feel about something. There can be millions of reviews on the internet for services which makes it difficult to track and understand the opinions of the customer. An emerging area of research to extract the subjective information to track and understand opinions of the customer is by using Sentiment Analysis. Accessible and plentiful data is been provided by the reviews for relatively easing the analysis for a wide range of applications.[1]**

This system seeks application and extension of the current work in the field sentiment analysis on data retrieved about college reviews from other websites using web mining technique.[2] A given review can be tagged as positive or negative by using Naive Bayes and decision list classifiers. Features such as bag-of-words and bigrams are compared to one another in their effectiveness in correctly tagging reviews. Recent studies analyzed these reviews and found that it includes information useful for colleges, such as user requirements, ideas for improvements, user sentiments about specific features and descriptions of experiences with these features. In this project, prediction of next yearscutoff on the basis of the last five years. College suggestions are also given based on percentage.

.

***Keywords*— Sentiment Analysis, Naive-Bayes, Opinion Mining, Natural Language Processing, Stupidsid**

I. INTRODUCTION

Opinion of people has become one of the most important sources for various services in ever-growing popular networks. In explicit, online opinions have changed into a sort of virtual currency for businesses wanting to promote their product, establish new opportunities, and manage their faculty reputations[1].

Users are allowed to share their opinion about the system in text reviews on review platforms, where they can express their satisfaction with a specific system feature or request a new feature.[3] Studies have recently shown that reviews stored by many websites include information that is useful to analysts, such as user requirements, bug reports, feature requests, and documentation of user experiences with specific system features.[4] This feedback can represent the "voice of the users" and be used to improve the college system.

II. PROBLEM STATEMENT

In real world, there is no such system wherein we get details of all the engineering colleges in Mumbai under one web application. Every time we had to visit individual websites to obtain details of that particular college. Details such as necessary college details with contact numbers and address, photos of infrastructure, etc.

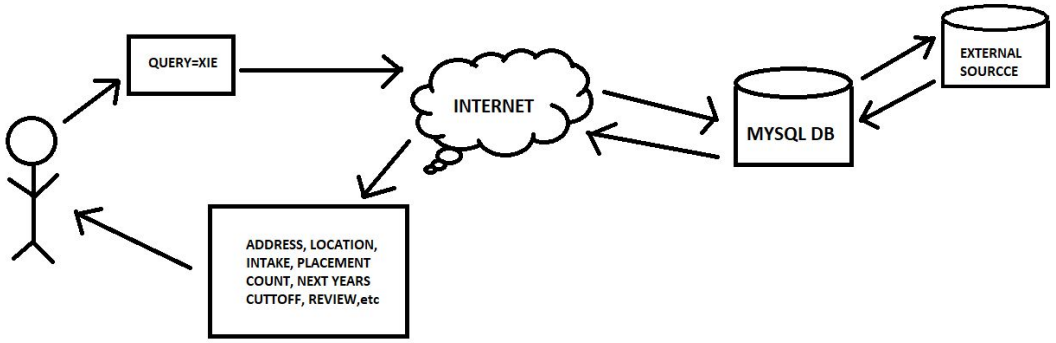
To save on the users time of hopping from one web application to another just to seek information our system will provide details such as an address of the college, contact, number, location, photos of the infrastructure, cut-off percentage(predicting the next years cut-off as well) and review of the college as well.

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 analyze this 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] In this project, we can use Naive Bayes to predict next years cutoff on the basis of last three years and college suggestion based on percentage.



IV. RELATED WORK

Alekh Agarwal et al., planned a machine learning technique incorporating linguistic information gathered through synonymousness 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 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 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 planned linguistics orientation approach uses "unsupervised learning" as a result of it doesn't need previous coaching so as to mine the information. 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 more 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 are consisted in 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]

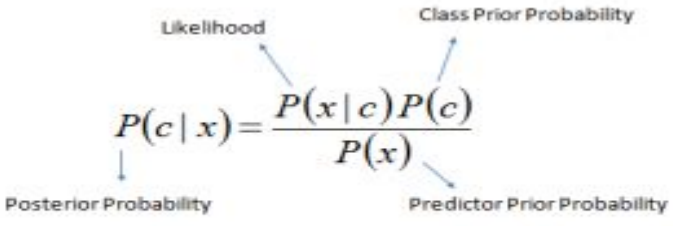
V. NAIVE-BAYES CLASSIFICATION ALGORITHM

The theorem Classification represents a supervised learning methodology in addition as a statistical procedure for classification.[6] Assumes associate underlying probabilistic model and uncertainty regarding the model is captured during 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 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 that the posterior chance of sophistication (c, target) given predictor (x, attributes).*
- *P(c) is the prior probability of class.*
- *P(x|c) is that the probability that is that the chance of predictor given category.*
- *P(x) is the prior probability of predictor.[11]*

A. Algorithm Working

Let's understand it using an example. Below we've got a coaching information set of weather and corresponding target variable 'Play' (suggesting prospects of playing). Now, we'd like to classify whether or not players can play or not supported weather. Let's follow the below steps to perform it.

Step 1: Dataset is to beconverted into a frequency table

Step 2: Create a Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

[11]

Step 3: Now,To calculate the posterior probability for each classuse a Naive Bayesian equation. The outcome of prediction is the class with the highest posterior probability [11]

Dictionary Generation:-

Count prevalence of all word in our whole knowledge set and create a wordbook of some most frequent words.

Feature set Generation

- All the documents are represented as a feature vector over the space of dictionary words.

- For every 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 every wordbook words and store the result (Here: label are negative and positive).

For each of the defined label now we have a word and corresponding probability.

B. Pros and Cons of Naive Bayes

Pros:

- Prediction class of test data set is fast and easy. Multi-class prediction is also an area where it will perform well.
- A Naive Bayes classifier performs better when the assumption of independence holds as compared to other models like logistic regression and you need less training data.
- In case of categorical input variables compared to a numerical variable(s) it will perform well.A normal distribution is assumed for a numerical variable (bell curve, which is a strong assumption).[11]

Cons:

- If the specific variable contains a class (in take a look at knowledge set), that wasn't determined in coaching dataset, then the model can assign a zero (zero) likelihood and can be unable to make a prediction. This is often known as "Zero Frequency". We can use the smoothing technique to solve this. One of the simplest smoothing techniques is called Laplace estimation.[11]
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict_proba are not to be taken too seriously[11].
- An assumption of independent predictors is another limitation of Naive Bayes. In real life, it is almost impossible that we get a set of predictors which are entirely independent.[11]

C. Applications of Naive Bayes Algorithms

- **Real-time Prediction:** Naive Bayes is sure fast and is an eager learning classifier. Thus, prediction in real time could be made.
- **Multi-class Prediction:** Multiclass prediction is also a feature for which this algorithm is well known. The probability of multiple classes of target variable can be predicted here.
- **Text classification/ Sentiment Analysis/ Spam Filtering:** Naive Bayes classifiers mostly used in text classification (due to a better result in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments) and Spam filtering (identify spam e-mail)
- **Recommendation System:** Collaborative Filtering and Naive Bayes Classifier together builds a Recommendation System that uses data mining and machine learning techniques to filter the unseen data and predict whether a user would like a given resource or not[11]

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 user so only valid person login into the system.

B. Data Collection

In this user select one college name from a college list and click on submit after submitting our system get 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 this college.

C. Sentiment Analysis

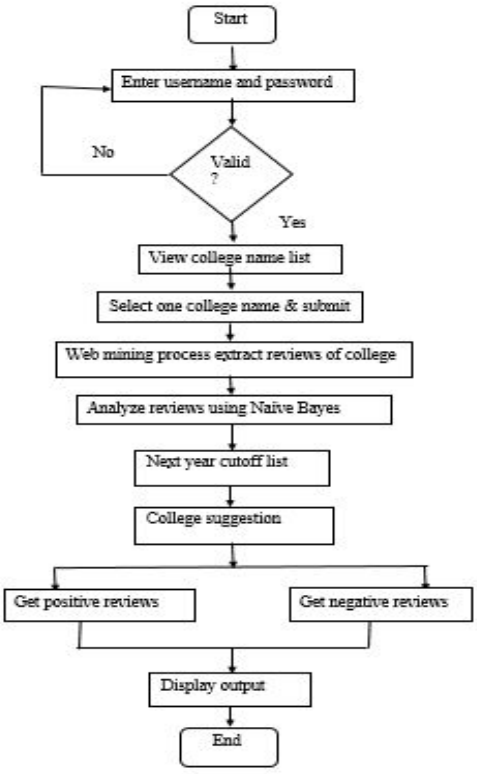
The reviews retrieved using web mining technique from other websites which can be analyzed using Naive Bayes algorithm and get the result as positive and negative reviews.

D. Output

In this module, an output is displayed to the user. The output display college information, Placement, Teaching or faculty, Crowd and 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. He then clicks on one of the college to acquire data.
4. Web mining is performed to extract reviews from other websites.
5. Analysis of those reviews is done using Naive Bayes Algorithm.
6. Next yearscutoff is also predicted.
7. Reviews are segregated into two parts, i.e. positive and negative reviews and are displayed to the user

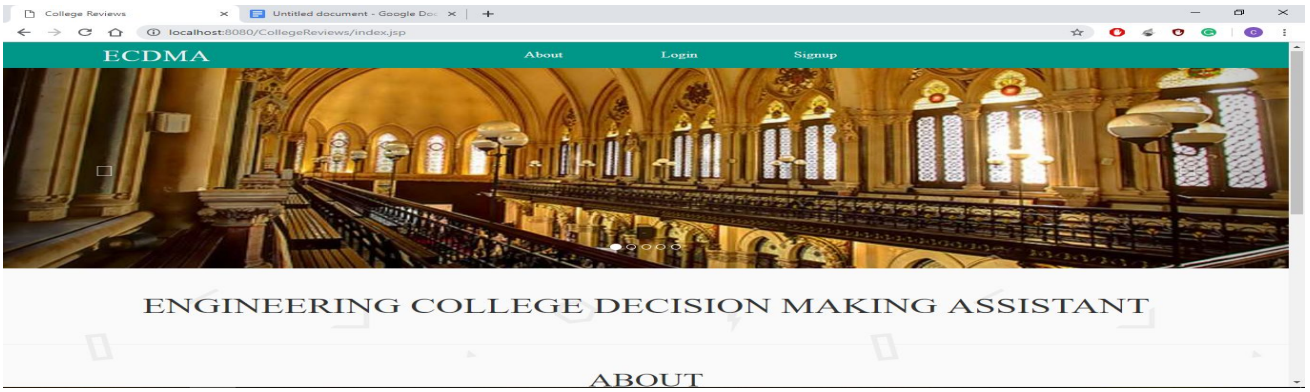


VIII. WORKING

- The user needs to sign up or login into the system.
- A list of colleges will be displayed in front of him from which he can click on any desired college to obtain the necessary information.
- Every college has information such as location, infrastructure, fees, placements, canteen, crowd, etc. This information is pre stored in the database and is retrieved once user clicks on a college.
- A user can even view the intake of every college branch wise. This data is also pre stored in the database and retrieved once user clicks on the college.
- Reviews of a particular college are fetched from www.stupidsid.com. The fetching is done using parameters such as faculty, infrastructure, other, placements, canteen, location and crowd.
- Once a review is fetched it is stored in the database and displayed to the user when he wants to view those reviews.
- Ratings for every review is generated based on the sentiments of particular words. Words present in the word classifier dictionary are analysed to find similar words in the reviews and rate it accordingly.
- Based on these ratings, the average is taken and the college is given an overall rating.
- MH-CET cut off of past 3 years is stored in the database. Using Naive Bayes algorithm, we predict the following years cutoff.

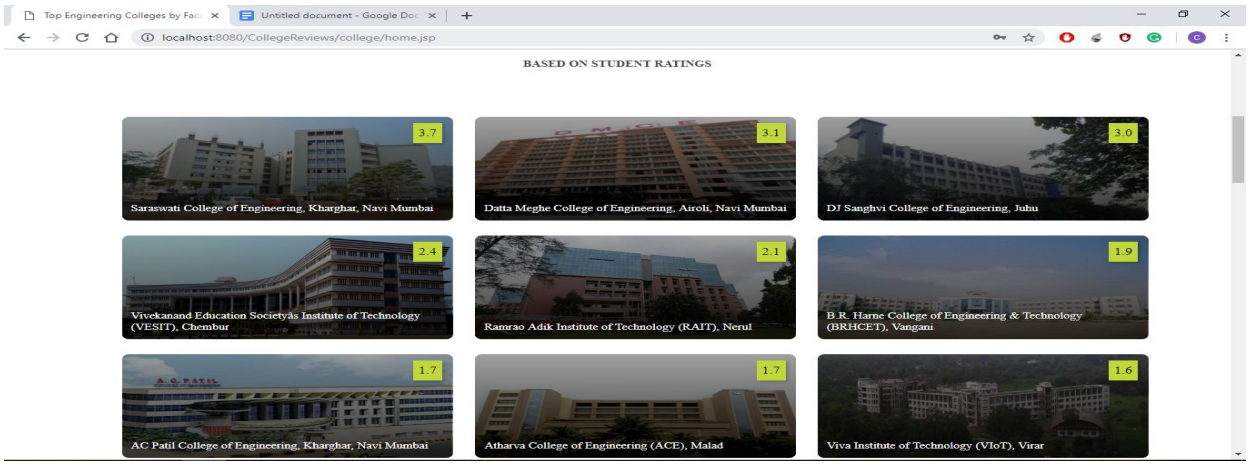
IX. RESULTS

1. INDEX PAGE



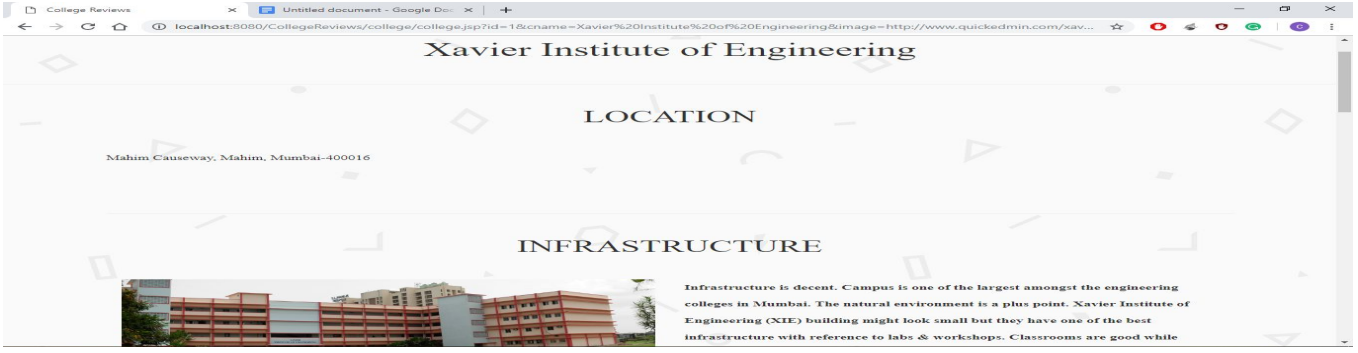
In this index page, the user can sign up onto the website or log in to his existing account. The user can even read some information about the website on this page in the about section.

2 . HOME PAGE



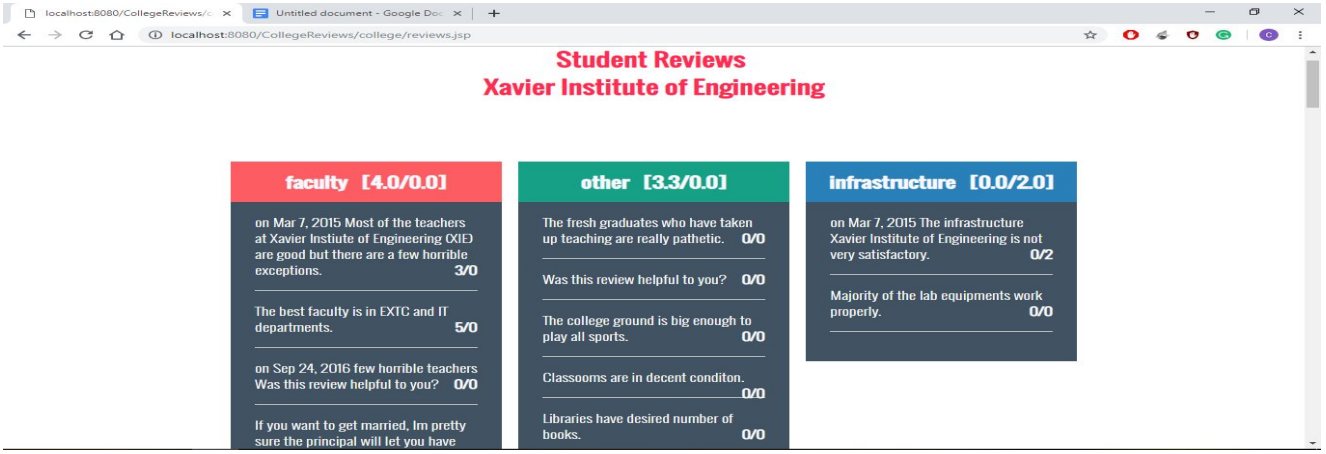
A list of colleges will be displayed to the user. The user will click on the desired college and obtain some basic information about the college. The college ratings based on user reviews are displayed on the top right corner of each college box.

3. COLLEGE DETAILS



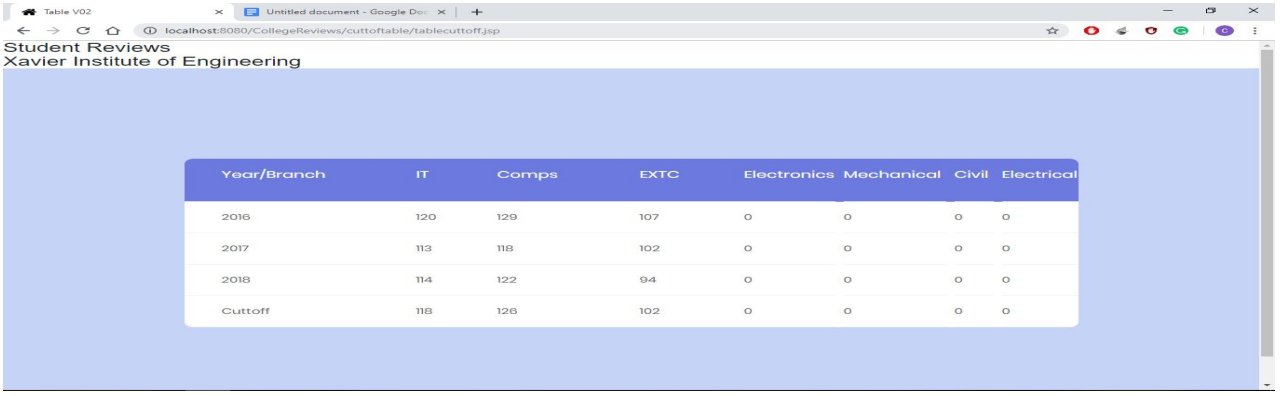
Basic details such as location, infrastructure, faculty, crowd, placement, fees and canteen about the college are shown to the user. The user can go through the details and acquire the required information.

4. COLLEGE REVIEWS



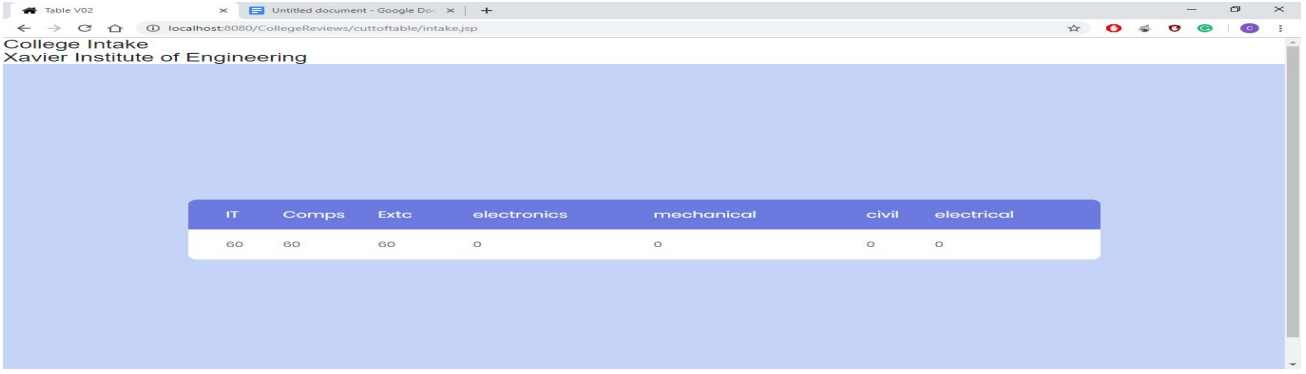
Reviews which are written by many people can be read by the user by clicking on the "View Reviews" button. Reviews are categorized based on their area such as faculty, infrastructure, etc.

5. CUTOFF



The cutoff of the years 2016,2017 & 2018 is given and the cutoff of the following year is being predicted and displayed using Naive Bayes algorithm. The cutoff is provided department wise.

6. INTAKE



The number of students that the college gives admission to is also mentioned here in the intake tab. Intake tab specifies the number of students that will get admission branch wise.

X. FUTURE SCOPE

You can add the following changes in the future:

-Only those users can log in who have an account in any bank system.

-Admin can put various advertisements

- Chatbox feature can be added

-FAQ section can be introduced

XI. 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 is filled with noise. The experimental evaluation suggests that satisfactory sentiment classification can be achieved using this approach. Naïve Bayesian summarizes review depending on features and technical feature values which are extracted from the reviews. In this project, it predicts next year cutoff on the basis of last five years records and college suggestion based on percentage.

XII. REFERENCES

[1]Heema Krishna, M.SudheepElayidom, T.Santhanakrishna, "Impact and Application of Sentiment Analysis using Twitter: A Survey", June 2015.
[2] BadrHssina, AbdelkarimMerbouha, HananeEzzikouri Mohammed Erritali, BelaidBouikhalene, "An Implementation Of Web Content Extraction Using Mining Techniques", Dec 2013
[3] Renata Maria, AbrantesBaracho, Gabriel Caires Silva, Luiz G F Ferreira, "Sentiment analysis in social networks: a study on vehicle"
[4] Emitza Guzman, WalidMaalej, "How do users like this feature? A fine grained sentiment analysis of app reviews".
[5] OmkarBorade, Kaushik Gosavi, Ajay Shinde, AvinashGowda 2017 IJEDR | Volume 5, Issue 2 | ISSN: 2321-9939
[6] G Angulakshmi, Dr.R.ManickaChezian, "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 language process (ICON), 2005.
[9]Bo Pang, Lillian Lee, and ShivakumarVaithyanathan, “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 Chou dynasty, PimwadeeChaovalit, “Movie Review Mining: a Comparison between supervised and unsupervised Classification Approaches”, Proceedings of the thirty eighth Hawaii International Conference on system sciences, 2005.
[11] Naive Bayes ”https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/”