**Text classification and Sentimental Analysis of College Reviews.**

**Minor Project I**

**Submitted by:**

Aditi Saxena (9917103197)

Meenal (9917103176)

Sanchit Goel (9917103175)

**Under the supervision of:**

**DR. CHETNA GUPTA**



**Department of CSE/IT**

**Jaypee Institute of Information Technology University, Noida**

October 2019

**Acknowledgements**

We would like to place on record our deep sense of gratitude to Dr. Chetna Gupta, Associate Professors, Dept. of Computer Science, Jaypee Institute of Information Technology, India for her generous guidance, help and useful suggestions.

We express our sincere gratitude to Prof. Rupesh Koshariya and Prof. Nitin Shukla, Dept. of Computer Science, Jaypee Institute of Information Technology India, for their stimulating guidance, continuous encouragement and supervision throughout the course of present work.

I also wish to extend my thanks to our classmates for their insightful comments and constructive suggestions to improve the quality of this project work.

Aditi Saxena (9917103197)

Meenal (9917103176)

Sanchit Goel (9917103175)

**Abstract**

Today extensive datasets are accessible on-line, holding text data or numerical. It has been the major focus for many practitioners and researchers to apply reasonable approaches and techniques and extract useful information from those datasets. Wide range of techniques have been used and tested to retrieve information [1].

Sentiment analysis or opinion mining is the computational study of people’s opinions, sentiments, emotions, appraisals, and attitudes towards college. Since early 2000, sentiment analysis has grown to be one of the most active research areas in Natural Language Processing (NLP). It is also widely studied in data mining, Web mining, text mining, and information retrieval. In fact, it has spread from computer science to management sciences and social sciences such as marketing, finance, political science, communications, health science, and even history, due to its importance to business and society as a whole. This proliferation is due to the fact that opinions are central to almost all human activities. To beliefs and perceptions of reality, and the choices we make to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, whenever need to make a decision we often seek out the opinions of others. This is not only true for individuals but also true for organizations [1].

Using the reviews we can take opinion of the user of our system. A user review refers to a review written by a user for a product or a service based on her experience as a user of the reviewed service or product. Most of college website gets reviews from users so the college system see view services is better and which services is require changes [2]. We are working on reviews written by users on different educational websites and processing them to generate an overall rating.

**Table of Contents**

**Page No.**

Acknowledgements 2

Abstract 3

Table of Contents 4

Introduction 5

Background Study 6

Requirement Analysis 7

Detailed Design 10

Implementation 11

Experimental Results and Analysis 14

Conclusion 17

References *18*

**Chapter-1**

# Introduction

Reviews are the integral part of our lives. Opinions of other people who have an experience always play an important role in mind of humans to take any decision. We always ask people about their review before trying something new. In today’s world of internet and technology, people Google everything before trying, and taking admission in a college is one of the most important decision one take in their life. This project is made by keeping human mindset in mind and directly targets human mindset. There are various Indian sites available for students to analyze different colleges and take an informative decision. Opinions and reviews which are accessible to us play an important role in taking decision and at a level they also decide success of a college. With an exponential growth of social media these days’ people often take their views to popular educational websites. Our project gathers reviews and opinions from four top educational websites on a single platform and gives us a rating based on different parameters and features. It is like we are collecting and merging different opinions and views of large number of people from different educational websites onto a single platform and presenting the crux of their opinion.

It is said that time is money and saving time equals earning. The benefit of this project is that users don’t have to visit those four sites and read large no of reviews written by people separately. Instead they can give name of college on our application and it will automatically gather reviews from four websites and by performing text classification and sentimental analysis on the backend it will generate a rating based on reviews written by people of that college on those four different websites. This project is one of a kind as there are platforms that gives you a rating of a particular college but collecting reviews from different websites and analyzing them by applying sentimental analysis on reviews and then generating a rating is something which is needed at this point.

Nothing can be made without help. We too for this project have taken help from various past related works and papers based on sentimental analysis of text so that we can analyze reviews efficiently and give a correct rating. The related research papers have helped us to perform sentimental analysis correctly. This project can also be used by college authorities to know what people and students think about them and where they are lacking in performance and where they need to improve Thus this project is based on sentimental analysis of college reviews and will help students to choose the right college by and make an informed decision based on ratings.

**Chapter-2**

# Background Study

Existing research has produced numerous techniques for various tasks of sentiment analysis, which include both supervised and unsupervised methods. In the supervised setting, early papers used all types of supervised machine learning methods (such as Support Vector Machines (SVM), Maximum Entropy, Naïve Bayes, etc.) and feature combinations. Unsupervised methods include various methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. Several survey books and papers have been published, which cover those early methods and applications extensively [1].

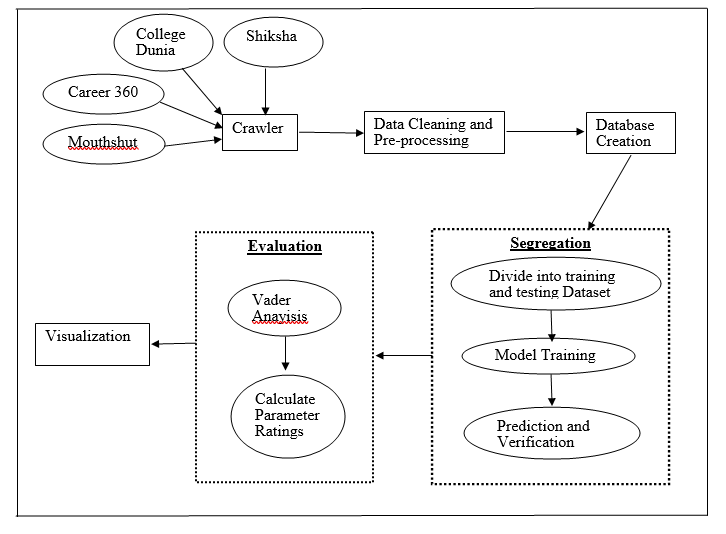
As mentioned, various research papers, articles and blogs were referred for this project.

|  |  |  |  |
| --- | --- | --- | --- |
| **Title of Research Paper** | **Sentiment Analysis of Google Reviews of a College** | **Sentiment Analysis of Top Colleges in India using Twitter Data** | **Sentiment Analysis of College Reviews** |
| **Data collection method** | Data is extracted from google reviews | Used data of top universities (IITs, NITs, AIIMS) from twitter | Data is extracted from various websites using web mining techniques |
| **Method used:** | Data Reprocessing: word Embedding  Sentences are split to find the sentiment of a particular word.  Values were given to the words and thus the sentences.  Maximum entropy was used to find the sentiment of sentences. | * Lexicon based sentiment analysis * Implementing naïve Bayes algorithm for sentiment analysis * Implemented SVM for sentiment analysis * Implemented neural Network for sentiment analysis   All the four results are compared for more accuracy. | **Bag of words**   * Find a word in 2500 files * All words are weighted same * Sequence is neglected   **Naïve-Bayes Classification** is used for classifying. |
| **Advantages** | * High accuracy * (96%) | * Mis-spelled worlds are corrected. * Have high accuracies (87%,90%,91%,92.6%) | * Reviews from multiple websites are used |
| **Disadvantages** | * Reviews from a single platform are used. | * Reviews from a single platform are used. | * Efficiency is low(40-50%) * Sequence of words is neglected |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Website** | **Shiksha** | **Career 360** | **College Dunia** | **Mouthshut** |
| Component Ratings | Overall, Placements, Infrastructure, Faculty & Course Curriculum, Crowd & Campus Life, Value for Money | Placements, Infrastructure, College life, Industry Exposure, Hostel, Faculty, Student crowd, Affordability | Overall student satisfaction, Academic, Food and Hostel, Faculty, Infrastructure, Placements, Social | Student, survey, alumni, tips to get into |
| Features for searching colleges | Degree, Subject, Cut off | College, course, Exams | Degree, Subject | Region within distance of city, name |
| Visualization | Filled Bar | Stars | Ratings as fractions | Pie Charts |
| Provides Specific categories for ratings | Yes | Yes | Yes | Yes |

**Chapter-3**

# Requirement Analysis



**Architecture of System**

# System Features

## DataExtraction

|  |  |
| --- | --- |
| System Feature | Data Extraction |
| Priority | Min |
| Description | To retrieve the data from various websites for all colleges |
| Action | This module is activated after the user provides a college name after selecting other option in Dropdown menu. |
| Result | The data curated from the 4 websites is then combined to form a file for each college. Data prepossessing is then performed on the extracted data. |
| Functional requirements | A focused crawler to extract relevant and verified data from websites. |

## Data Cleaning and Pre-Processing

|  |  |
| --- | --- |
| System Feature | Data cleaning and Pre-Processing |
| Priority | Low |
| Description | It helps reduce the noise of text, reduce dimensionality, and assist in the improvement of classification effectiveness |
| Action | This module is activated after extraction and it Removes numbers, removes punctuation, Lowercase, Removes stopwords. |
| Result | The system shows the results of the search of a query or the output of the hot topic module associated with the sentiment polarity of each item retrieved together with the percentage of Positive, Negative and Neutral sentiment of the whole result. |
| Functional requirements | A preprocessing module to preprocess your text simply means to bring your text into a form that’s predictable and analyzable for your task. |

## Text Classification/Opinion mining

|  |  |
| --- | --- |
| System Feature | Text Classification/Opinion mining |
| Priority | High |
| Description | The process of assigning tags or categories to text according to its content. It’s one of the fundamental tasks in NLP with broad applications such as sentiment analysis, topic labeling etc. |
| Action | Apply Logical Regression, Naïve Bayes, Linear SVC, SVM, Random Forest to the Reviews |
| Result | The system shows the accuracies of each Model and then Linear SVC with the maximum accuracy is applied. |
| Functional requirements | A module that classifies reviews into four features namely faculty, placements, infrastructure, social. |

## Sentiment analysis

|  |  |
| --- | --- |
| System Feature | Sentiment analysis |
| Priority | High |
| Description | Identifying the sentiment polarity (positive, negative or neutral) of reviews from websites. |
| Action | This module is activated After separating /classifying reviews into various parameters, we have performed sentimental analysis on each column of each feature to find out the average rating of college. |
| Result | The system shows the results of the sentiment polarity of each review together with the percentage of Positive, Negative and Neutral sentiment of the whole result on a scale of -1 to 1. |
| Functional requirements | A sentiment classifier module to find ratings of college in each review. |

## Visualization

|  |  |
| --- | --- |
| System Feature | Visualization |
| Priority | Mid |
| Description | Displaying the results in forms of graphs and charts. |
| Action | This module is activated after all the ratings have been calculated and the results need to be displayed to the user. |
| Result | The system shows the results of the search of college and displays result from cumulation of several websites. |
| Functional requirements | A sentiment visualization module which displays results to the user after anaylsis. |

**Chapter-4**

# Detailed Design

**Working of Project**

Project is divided into various steps to make a better understanding:

## Web Scrapping:

Our first task is to scrap review data from various educational websites and collect data to create dataset for analyzing it. During their Extraction we used Beautiful soup library to parse their HTML file and we also used Selenium and Chrome driver to help extract HTML data from College Dunia website. Data of 98 engineering colleges was extracted from Shiksha.com, career.com, collegedunia.com, mouthshut.com and to create database and stored into MongoDB. The reviews of only verified users was collected.

## Text Classification:

The dataset collected have whole reviews of colleges along with website names. We have classified those reviews into four features namely faculty, placements, infrastructure, social to find out the average rating based on these parameters. Five machine learning models were tried for this task and Linear SVC is used for final project. Whole review is also kept for finding overall rating of college.

## Sentimental Analysis:

After separating /classifying reviews into various parameters, we have performed sentimental analysis on each column of each feature to find out the average rating of college. We have used VADER (Valence Aware Dictionary and sentiment reasoner) for finding score of each review. It is toolkit provided by python to calculate sentiment of a text. It uses combination of words which are generally labeled according to their semantic orientation as either positive or negative. It works well on reviews, tweets and it doesn’t require any training data. It is the main purpose of using this in our project. It is fast enough to be used and does not suffer from speed degradation [1].

## Visualization:

**Chapter-5**

# Implementation

## Data Extraction

Data extraction is to retrieve the data from various websites. In our project we have used the top 100 colleges ranked by MHRD (NIRF rating) from the four websites:

1. Shiksha
2. College Duniya
3. Mouth shut
4. Career360

In addition to being a versatile communications platform to users around the globe, these websites are also an excellent source of current information. The data is scraped from the four websites and then it is combined to form a file for each college. Data prepossessing is then performed on the extracted data.

## Data Pre-processing and Data Cleaning

Preprocess your text simply means to bring your text into a form that’s predictableand analyzablefor our task. An initial step in text and sentiment classification is pre-processing. A significant amount of techniques is applied to data in order to reduce the noise of text, reduce dimensionality, and assist in the improvement of classification effectiveness [2].The collected data is preprocessed and various symbols, html tags are removed and the words are converted into vectors, this data set contains many sentences labeled with 0 or 1 depending on its polarity

## Topic classification

Topic classification is the process of assigning tags or categories to text according to its content. It’s one of the fundamental tasks in [Natural Language Processing](https://monkeylearn.com/blog/definitive-guide-natural-language-processing/) (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection .Topic classification with machine learning is usually much more accurate than human-crafted rule systems, especially on complex classification tasks. Also, classifiers with machine learning are easier to maintain and you can always tag new examples to learn new tasks. We experimented with different machine learning models, evaluated their accuracy and find the source of any potential issues. We checked for the following four models:

### Logistic Regression

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [4].

### Naive Bayes

It is a family of statistical algorithms we can make use of when doing text classification. Naive Bayes is based on Bayes Theorem, which helps us compute the conditional probabilities of occurrence of two events based on the probabilities of occurrence of each individual event. [5]

### Support Vector Machine

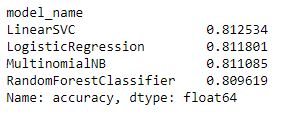
A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyper plane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyper plane which categorizes new examples [6].

### Random Forest

It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction [7].

### Linear Svc

It is a type of support vector machine. The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyper plane that divides, or categorizes, your data. From there, after getting the hyper plane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations [8].



Accuracies of different models

## Sentiment Analysis

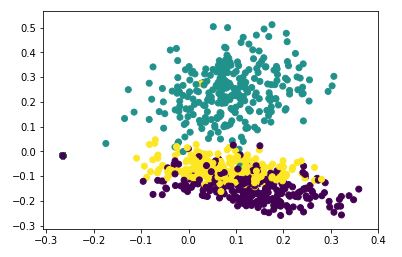
Sentiment analysis is a process of analyzing the given text in order to find out the emotions in it. In simple words, opinion mining is a process of detecting the sentiment of the writer concerning a particular topic. It is a blend of techniques and strategies about distinguishing and detecting subjective information from a text such as opinions and attitudes. Usually, it has been about opinion polarity to find out whether someone has negative, positive or neutral opinion about college. This proliferation is due to the fact that opinions are central to almost all human activities. To beliefs and perceptions of reality, and the choices we make to a considerable degree, conditioned upon how others see and evaluate the world [9]. We have used VADER for finding rating of each review in this project.

**Visualization**

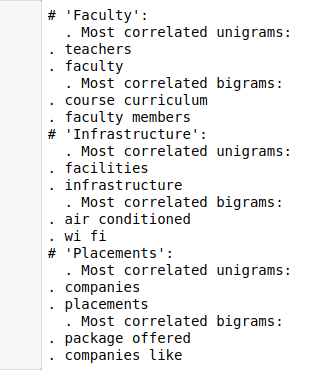
**Chapter-6**

# Experimental Results and Analysis

## Testing reports



TI-IDF feature set plot .We can see in this image that our data can be classified into 3 clusters for classifying on the basis of TI-IDF.



### Naive Bayes:

It is used for classification problem and it has performed multi class text classification on our dataset. There are three labels in our data and it has segregated the reviews with lowest accuracy of 81.10%.

Confusion Matrix:

[[331 125 1]

[ 1 455 1]

[ 1 127 329]]

### Random Forest:

Random Forest has been proven effective in many sentiment tasks and runs efficiently in larger databases but over fitting can easily occurs. This classifier has the accuracy in text classification and classifying data with accuracy of 81.11%.

Confusion Matrix:

[[456 1 0]

[128 329 0]

[130 1 326]]

### Logistic Regression

We have used this for multi class text classification and it has given results with accuracy of 81.18%

Confusion Matrix:

[[455 1 1]

[125 331 1]

[128 1 328]]

### Linear Svc

It is a type of support vector machine. It has the accuracy of 81.25%. We will be using linear svc for multi text classification for our whole project.

Confusion Matrix:

[[455 1 1]

[125 331 1]

[128 1 328]]

Final Rating of particular college:

**Chapter-7**

# Conclusion

This project gives ratings of a particular college that is given as input by the user and its ratings are shown based on various features which are necessary for college selection. We hope this project if implemented on large scale will help college students in time of their admission process. We have collected data from highly trusted educational websites and have applied machine learning algorithms of high accuracy that will give genuine rating of that college. This application will help students and their parents for choosing a college based on their features and will help them in making informed decision. We have collaborate reviews for the college from various college interacts websites and combined the reviews onto a single platform and used Sentiment Analysis to find sentiment rating of review and give proper result of positive or negative reviews and calculate overall rating of the college along with overall ratings of various parameters like Placement, Faculty etc. Models, libraries used in this application are chosen accordingly to give ratings of high accuracy and we will be using linear svc for classification of reviews.

A simple frontend is developed to use this application easily and comfortably with minimum complications. This project would be a great help to students if implemented widely. This application, at this current stage deals only with Indian engineering colleges and those colleges which are available on chosen educational websites. This project can also be implemented on such problem statements which deals with reviews of a product

**Chapter-8**

# References