ISSN XXXX XXXX © 2018 IJESC

Research Article Volume 8 Issue No.8

College Recommendation System using Data Mining and Natural Language Processing

Akshar Panchal[1], KushalGosar[2], HomilParmar[3], Rohini Nair[4]
Assistant Professor[4]
Department of Computer Engineering
K.J.Somaiya College of Engineering, Mumbai, India

Abstract:
One of the main problems faced by students today is selecting proper colleges based on their choice and preferences. This paper explores data mining and natural language processing techniques to recommend colleges to students based on their preferences. It makes use of online source to fetch data from comments given by students to various colleges according to their experiences. The system uses Natural Language Processing (NLP) to analyse comments from sources to generate a dataset which is further used for recommendation. This model predicts comments of students based on criteria like faculty, infrastructure, placements, canteen, library and score. The model aims at offering Recommendation which combines Naïve Bayes and Multi criteria filtering algorithm to produce efficient and high quality recommendations.

Keywords: College Recommendation System; Opinion mining; Natural Language Processing; Text Analytics; Multi-criteria filtering

I. INTRODUCTION

On the Internet, lot of information is created and used by people who make it difficult for everyone to choose different options, this is termed as information overload. Recommender Systems (RS) are most commonly used tools to help users' in decision making process by applying data filtering and data mining algorithms. Therefore, RS is the solution to the problem of inconvenience caused and it plays vital role in providing useful results to everyday problems. RS have the effect of guiding the user in a personalized way to choose own preferences and analyse the same.

The total data that any organization uses is huge. Most of this data is unstructured and in various forms like email, images, notes, web pages and so on. Most of the decision making is done on this unstructured data leaving analytics. In case of educational colleges the same case is valid. Presence of huge amount of unstructured data like student opinions expressed through web logs, emails and other surveys it is time consuming and practically impossible to analyse and summarize the information and useful data which leads to some conclusion and decisions.

The data which is usually analysed is always the structured data which includes attendance, marks, login frequency and so on. Hence it is very important to understand trends in education system by using data like student feedback to efficiently improve overall performance of colleges and enhance learning experience for all. There are many ways of collecting student reviews and ratings for different colleges and one amongst them is using surveys. Also data regarding the colleges can be collected from online sources like websites which will help classify unstructured data from comments of students having some experience in those colleges. However, collecting data using surveys is difficult and time consuming which may lead to lot of difficulties.

The recent developments in the field of data mining, natural language processing and analytics has been brought together in order to solve this issues which has helped unleash great prospects for different colleges. Using above methods this proposed model is implemented to overcome this difficulties by implementing new models for data extraction and analysis. Opinion mining is one such advancement in this field which helps to evaluate the polarity of opinion from entire dataset involving unstructured data. This method is being used by researchers in classifying data which is unstructured and obtaining reviews for different products, movies and so on.

This paper focuses on using various data mining techniques for classifying students' feedback obtained from online sources like websites with respect to various features of teaching and learning like faculty, canteen, placements an so on. The extracted data from online sources was treated using natural language processing algorithm like naïve bayes and also the dataset was sorted and classified. After extraction of useful data from unstructured data multi criteria filtering techniques are applied from data mining in order to recommend colleges to students based on the filters as applied according to personalized preferences. These two algorithms are implemented for extracting and obtaining results for recommendation which is accurate and efficient as well.

II. LITERATURE REVIEW

A. *Opinion mining*

Opinions are views of different people representing their beliefs,views or judgements in some form with respect to a particular subject or matter. This reviews or opinions are considered subjective as they may change with user perspective. More than general facts the opinions of stakeholders play a major role in determining the usefulness of product or system. Opinion mining and sentiment analysis are terms used interchangeably and they involve analysing data obtained and classifying the same based on opinion polarity being positive or negative. The analysis is done based on orientation of the text and data following computational treatment of opinions expressed. Since, data and opinions are expressed in natural human languages it becomes necessary to understand the same and analyse it. Hence for mining data from online sources through comments we use Natural Language Processing (NLP) techniques. Along with NLP

knowledge discovery database methods are mostly employed for various stages of opinion mining like statement detection, features identification, determining the polarity of opinions and also opinion summarization. From among lexical based approaches and supervised machine learning approaches the algorithms used are Naive Bayes (NB), K Nearest Neighbour (KNN) and Support Vector Machine (SVM) which uses large number of labelled data and are used for classification and determination of polarity of the opinions.

SVM works best while using sparse text data by defining definite partitions in the dataset and by dividing it into various classes. The best partition space is found by maximum normal distance between datasets. The NB classifier is the most commonly used technique for classification of unstructured data which uses Bayes theorem to calculate the possibility of given data belonging to a feature, P(l/f) using below formula

$$P(l/f) = (P(l)*P(f/l)/P(f)) \quad (1)$$

Where P(l) is possibility of occurrence of data in the dataset used and P(f/l) is the possibility that a given feature belongs to a particular data. P(f) is the occurrence of particular feature in the dataset used. If the features f1, f2, f3 ….fn are not dependent on one another then the equation (1) becomes

$$P(l/f) = (P(l)* P(f1/l)* P(f2/l).....P(fn/l) / P(f))$$

KNN algorithm deploys an indexing mechanism for datasets. To classify the data it compares the similarity of the data with the training set index and uses the k nearest by measuring similarity using Euclidean distance.

Neural Network classifier algorithm deploys many layers of neurons as a medium of classification where each neuron is used to take a word frequencies of the dataset as input. These are also associated with the weight for calculating the input function. The output of each layer is then propagated to other layers. The algorithm predictions are based purely on the weight, input set and the neurons.

Feature based opinion mining is yet another rule that has been analysed to a great extent. Opinion mining is being used in many industries and online services as well as recommendation systems, public health care, tourism and government sectors like public opinion on taxes, various policies and financial decisions. Work mentioned also shows how Sentiment analysis has been performed on student feedback and comments highlighting the growing popularity of these techniques in educational sector as well. The work basically combines data mining with natural language processing on a particular volume of dataset which is extracted using online sources like weblogs and websites.

B. Abbreviations and Acronyms

Abbreviations used in the paper are Natural Language Processing (NLP), Multi Criteria Filtering (MCF) and Naïve Bayes (NB) classifier algorithm.

III. COLLEGE RECOMMENDATION

A. System Architecture

This is proposed architecture of College Recommendation System. There are various data mining techniques available but not meant for getting recommendations for educational purpose. Main components of the system include Apache Tomcat Server, Online sources for data collection and extraction, MySQL database and Recommendation engine.

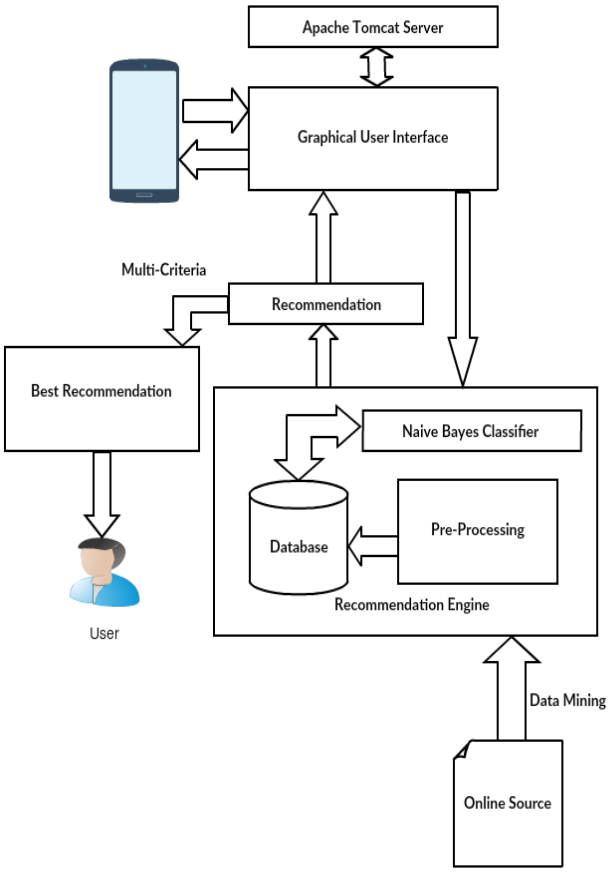

Fig.1. College Recommendation System Architecture

From Online sources like weblogs and websites the data is extracted from student comments and feedback. This data is unstructured data and needs to be processed and transformed into useful and structured data to be used for recommendation. The comments are stored in database in string format for processing.The unstructured data i.e. comments is transformedinto ratings by using Naïve Bayes classifier algorithmand stores the ratings of each college in a separate database table format.

Student using graphical user interface setsrequired filters according to preferences and sends the request for best college recommendation. The recommendation engine according to user query selects the best fit colleges from database which contains ratings and recommends it to user.The list of colleges will contain the best colleges according to overall ratings in proper best match order.

B. Proposed Methodology

To convert the online source data available to useful form and to implement the algorithm following methodology is adopted:

1. Select online sources from where data needs to be extracted and used. Data might be unstructured or in proper format.
2. Process the data using various formatting techniques.
3. Extract the entities and features to find their individual polarity i.e. polarity of comments that whether they are positive or negative.
4. Train the model using Naïve Bayes algorithm to classify into positive and negative classes.
5. Take the average of classes for individual entity.

C. Data Extraction and Processing

Data is extracted first from online sources like weblogs or websites.Data consist of comments and feedback posted by

students on the website.The data which is fetched is unstructured and not in a useful form.This data is properly extracted from web pages and stored in database.

TABLE 1 COLLEGE DETAILS

id	Location		Infrastructure		Faculty
1	Sardar Patel College of ...	535 b...	Infrastructure is decent...	438 b...	The faculty at Sardar Pa...
2	Located conveniently clo...	307 b...	One massive plus point a...	316 b...	The faculty at Sardar Pa...
3	Datta Meghe College of E...	235 b...	This is where Datta Megh...	439 b...	The faculty at Sardar Pa...
4	Bandstand, mate. You hav...	542 b...		1 b...	Recently Fr. Conceicao R...
5	It takes about 5 minutes...	439 b...	Fr. Conceicao Rodrigues ...	781 b...	Excellent. Faculty is go...
6	KC College of Engineerin...	137 b...	The infrastructure at KC...	408 b...	The faculty at KC Colleg...
7	Saraswati College of Eng...	223 b...	Excellent! Saraswati Col...	378 b...	EXTC faculty is supposed...
8	Ramrao Adik Institute of...	161 b...	Ramrao Adik Institute of...	487 b...	RAIT has some really ded...
9	AC Patil College of Engi...	178 b...	The infrastructure at AC...	359 b...	This is probably the one...
10	Anjuman-I-Islam Kalsekar...	415 b...	Although it is still und...	609 b...	Most of the teachers at ...
11	It takes about 15 to 20 ...	389 b...	Infrastructure at Atharv...	781 b...	There is a mixture of ex...
12	Bharati Vidyapeeth Colle...	312 b...	Bharati Vidyapeeth Colle...	590 b...	About 70% of the teacher...
13	B.R. Harne College of En...	530 b...	Currently thereâ€™s only...	654 b...	While majority of teache...
14	â€œReaching Dilkap Colle...	477 b...	Dilkap Research Institut...	374 b...	"Nothing exceptional, bu...
15	â€œReaching Dilkap Colle...	477 b...	The classrooms are roomy...	251 b...	Faculty at Don Bosco Ins...

These comments are stored according tocolleges. The comments are stored in string format and are processed and broken into tokens to analyse each word in the string.This tokens are then sorted to remove repetitive words and prepositions which are not useful for determining polarity of comments. The positive and negative responses are grouped into separate database which were used for deciding the comments to be positive or negative.

TABLE 2 WORD COLLECTION

positivecollection	prate	negativecollection	nrate
good	3	not good	3
not bad	1	bad	2
nice	2	not nice	2
excellent	5	poor	1
like	2	dont like	1
g8	4	hate	4
greate	4	boring	3
intresting	3	dump	4
best	5	dull	5
better	4	rude	2
top	5	not	2
enjoyed	4	worst	5
better	2	harassment	3
superb	4	dont	1
awesome	4	poor	1
*	0		0

The useful tokens are stored in a list to determine their polarity i.e. positive or negative comments.The tokens are then used to determine for which feature the comment is made and the accordingly the ratings are assigned to that feature. The comment is related to a particular feature or not is determined by the words used in the comments provided. Following table illustrates some example

TABLE 3 CRITERIAS/PREFERENCES

Infrastructure	Classroom, infrastructure, lab
Location	Location, distance
Canteen	Canteen, food
Event	Event, festival, days
Attendance	Attendance, defaulter, proxy
Library	Library, novels
Crowd	crowd
Faculty	Faculty, teacher, sir, principal, staff
placement	Placement, company, campus

D. Recommendation of colleges

Recommendation is done using multi criteria filtering; hence data needs to be structured for processing recommended colleges.Table 1 demonstrates the way data is stored for various colleges. Table 5 has ratings related to particular entity for same college from multiple users.Using Naïve Bayes classification, this data is processed as per equation (1).

TABLE 4 COLLEGE RATINGS

id	subject	comment		prate	nrate
8	other	RAIT is the best college ...	34 b...	5	0
8	events	The way students celebrat...	64 b...	0	0
8	faculty	The faculties are a bit r...	116 b...	0	2
8	other	Other than great placemen...	136 b...	2	0
8	other	The footfall is increasin...	125 b...	3	0
8	placements	So if u expect to have an...	234 b...	4	0
8	other	This college is awesome.	24 b...	4	0
8	other	Currently I am studying C...	67 b...	0	0
8	other	The hod is a bit strict.	24 b...	0	0
8	other	The office are too much a...	33 b...	0	0
8	events	All types of festivals an...	53 b...	0	0
8	other	And the placement records...	36 b...	0	0
8	other	If you want to complete e...	95 b...	0	0
8	other	RAIT haters go home you a...	37 b...	0	0
8	placements	RAIT people say it is the...	275 b...	3	0
8	other	though there are many goo...	60 b...	3	0

Positive and negative comments are clubbed together in database for evaluation. This ratingis evaluated by taking average and considering overall probable rating of that college related to that particular field.

TABLE 5 STRUCTURED DATA

id	location	infrastructure	attendence	events	library	crowd	faculty	placements	cantee
8	0,0	0,0	0,0	0.0,0.0	0,0	0,0	0.0,2.0	3.5,0.0	0,0
1	0,0	0,0	0,0	0,0	0,0	0,0	0.0,5.0	3.0,0.0	0,0
5	0,0	0,0	0,0	0,0	0,0	0,0	3.0,5.0	0,0	0,0
15	0,0	5.0,0.0	0,0	0,0	0,0	0,0	0.0,0.0	3.0,1.0	2.0,0.
33	3.0,0.0	0,0	0,0	0,0	0,0	3.0,0.0	3.3333333333333335,0.0	3.0,0.0	2.0,0.
7	0,0	3.0,0.0	0.0,0.0	0,0	0,0	0,0	3.0,1.5	5.0,0.0	0,0
9	0,0	0,0	0,0	3.0,0.0	0,0	0,0	0.0,7.0	0,0	0,0
17	0,0	0.0,5.0	0,0	0,0	0,0	3.0,0.0	0,0	0,0	3.0,0.
10	0,0	4.0,0.0	0,0	0,0	0,0	0,0	0.0,2.0	0,0	0,0
31	0,0	0.0,0.0	0,0	0,0	0,0	0.0,0.0	0.0,0.0	0.0,0.0	0,0
30	0,0	0.0,0.0	0,0	0,0	0,0	0,0	3.0,0.0	0,0	0,0
32	0,0	0,0	0,0	0,0	0,0	0,0	0.0,9.0	3.0,0.0	0,0
28	0,0	0,0	0,0	0,0	0,0	4.0,0.0	5.0,4.0	0,0	0,0
19	0,0	0,0	0,0	0,0	0,0	0,0	3.5,0.0	0.0,0.0	0,0
2	0,0	2.0,0.0	0,0	3.0,0.0	0,0	0,0	0.0,0.0	0,0	4.0,0.

Student can apply filters based on his preferences like placements, canteen, faculty, library, events, infrastructure and others. Recommendations will be based on multi criteria filtering. User will get recommendations based on his filters and ratings on the application.

IV. CONCLUSION AND FUTURE SCOPE

In this research, we have gathered student's comments from online websitesand implemented Naïve Bayes algorithm with Multi-criteria algorithm to classify them as it outperforms other in terms of accuracy. Key criteria's like Infrastructure, placement, faculty, canteen and many other are extracted from the students' comment and classified to be positive and negative based on their polarity to help students to select colleges which are best suited for them. In future research, we intend to perform opinion mining of student feedback gathered using social media and also to comparatively analyse how the student opinion varies using various demographics like age gender and so on.

REFERENCES

[1] Dheerajkumar Bokde, Sheetal Girase, Debajyoti Mukhopadhyay "An Approach to A University Recommendation by Multi-Criteria Collaborative Filtering and Dimensionality Reduction Techniques" 2015 IEEE International Symposium on Nanoelectronic and Information Systems.

[2] Dhanalaxmi V., DhivyaBino, Saravanan A.M. "Opinion mining from student feedback data using supervised learning algorithms" 2016 3rd MEC International Conference on Big Data and Smart City.

[3] Pang, B. and Lillian L. s.l, “Opinion mining and sentiment analysis: Foundations and trends in information retrieval”, Vol. 2, 2008.

[4] Kumar Ravi, Vadlamani Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications”, Published in Knowledge-BasedSystems Vol. 89, 14–46, 2015.

[5] Sunghwan Mac Kim and Rafael, “Sentiment Analysis in StudentExperiences of Learning”, Available at ResearchGate.com

[6] G. Adomavicius and Y. Kwon, “New recommendation techniques formulti-criteria rating system”, IEEE Intelligent Systems, Vol. 22, No.3, pp. 48-55, 2007

[7] Jyotsna TalrejaWassan, “Discovering Big Data Modelling for EducationalWorld”, IETC Procedia - Social and Behavioral Sciences, pp:642 – 649, 2015