COMMENTS CLASSIFICATION USING SENTIMENT ANALYSIS WITH KNN

[1]Vaishnavi Nagose, [2]Anjali Raut
[1]Student, [2]Professor
Computer Science And Engineering,
H.V.P.M's College of Engineering And Technology,
Sant Gadge Baba Amravati University, Amravati, India

Abstract : The numbers of colleges are increasing day-by-day. All they have their counseling centers where new candidate can get information about college. But it requires time to understand which the best deserving college is according to own expectation. College recommendation system is most popular way to get an appropriate college from number of colleges. There are number of sites for college recommendation all having their own strengths, weaknesses and aspects. The proposed system will use KNN algorithm for comment classification which works for better efficiency and accuracy. The proposed system will recommend college according to candidates' requirement and interest on the basis of comments given by college students. The candidate can see the performance graph to decide which college to choose.

***Index Terms* – Sentiment Analysis, Comments Mining, Semantic Analysis, Text Mining, Opinion Mining**

I. INTRODUCTION

The recommendation systems use different techniques and algorithms for recommendation. The algorithms are choosing on the basis of their efficiency, performance, accuracy and other factors. The algorithms used in existing system give some advantages and disadvantages. But with the advancement in existing system the college recommendation is made better for those fresher's who is searching for college. The proposed system works on the comments given by stake holders on their respective college. That comments will be considered for recommendation. The major part in college recommendation is classification of comments which can be done by using KNN algorithm. The comparative study shows that the KNN algorithm gives better performance graph than classification algorithm used in existing system. The sentiment analysis also known as opinion mining will be considered while classification of comments which identify the positive and negative appearance towards comments. The stake holders write comments on their respective college performance which leads fresher's to understand the performance graph easily. The TF-IDF technique will be calculates the score of the comments which will decide the polarity for the graph.

II. LITERATURE REVIEW

College recommendation in existing works on different aspects and one of the aspects is based on college performance with student interest. This classifies into following parts.

2.1 Comments Mining:

[1] Inbal Yahav, Onn Shehory, and David Schwartz discuss the comments mining in social networking. The technique TF-IDF is used with adjustment for the bias and its removal. Where the observation is taken from various Facebook fan pages includes different domains. But using only this technique cannot give more accuracy in education domain.

[2] Leena Deshpande, Nilesh Dikhale, Himanshu Srivastava, Apurva Dudhane, Umesh Gholap discuss the recommendation by using various aspects. The algorithms used are having strength as well as weakness but those algorithms are not working for classification purpose.

[3] A.Kousar Nikhath, K.Subrahmanyam, R.Vasavi discussed about Building a K-Nearest Neighbor Classifier for Text Categorization which used in classification methodology.

2.2 Sentiment Analysis:

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan discuss the topic of sentiment analysis which uses wordnet English dataset for sentiment analysis and polarity detection on movie reviews. This researcher works on machine learning techniques.

[5] Ahmad Ashari, Iman Paryudi, A Min Tjoa discussed about Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool which gives performance graph of the algorithms. There are different training set used for getting experiment result.

[7] The sentiment analysis is used to find out what people think about news, topics, articles etc. sentiment analysis can be achieved from various phases and classification techniques.

III. PROPOSED WORK

The proposed system includes KNN classification algorithm for comments classification. The performance of KNN algorithm is better in comments classification domain.

3.1 The proposed system has four users they are given as

3.1.1 Admin: Admin is a user who will approve college registration request. Admin also view performance of all approved college.

3.1.2 College admin: College Admin will register college with relevant information and approve student’s registration request. College admin will make publicity of own college and can view other college comparative performance.

3.1.3 Students: The students register themselves with respective college and after approval of request they can give feedback on the college posts.

3.1.4 End users: The end user is a fresher candidate who will search for different college and will get recommended with respect to requirement.

3.2 The proposed system consists of five modules

3.2.1 Admin panel: The admin panel approve college registration request and can view comments and comparative performance of college.

3.2.2 College admin panel: The college admin panel register itself and approve own students registration request. College admin panel will register objectives and principles and also upload photos and information about college and events. College admin can also view own performance and comparative performance of other colleges.

3.2.3 Students: Students also known as stakeholders will register itself with their respective college and after approval of college admin student can submit feedback on their own college advertisement. Stakeholders can also view college performance reports after calculation of performance graph.

3.2.4 Comment classification: The comments can be classified using KNN classification algorithm and for polarity TF-IDF technique will be used. For comments classification the comments are taken from submitted feedback of stakeholders. The comments classification can be done through the sentiment analysis using wordnet dataset.

3.2.5 DSS Reports generation: DSS reports will be generated on the basis of feedback analysis. These feedbacks are given by stakeholders. On the basis of feedback analysis, system will automatically generate graphical as well as textual comparative analysis of colleges.

3.3 Working Diagram

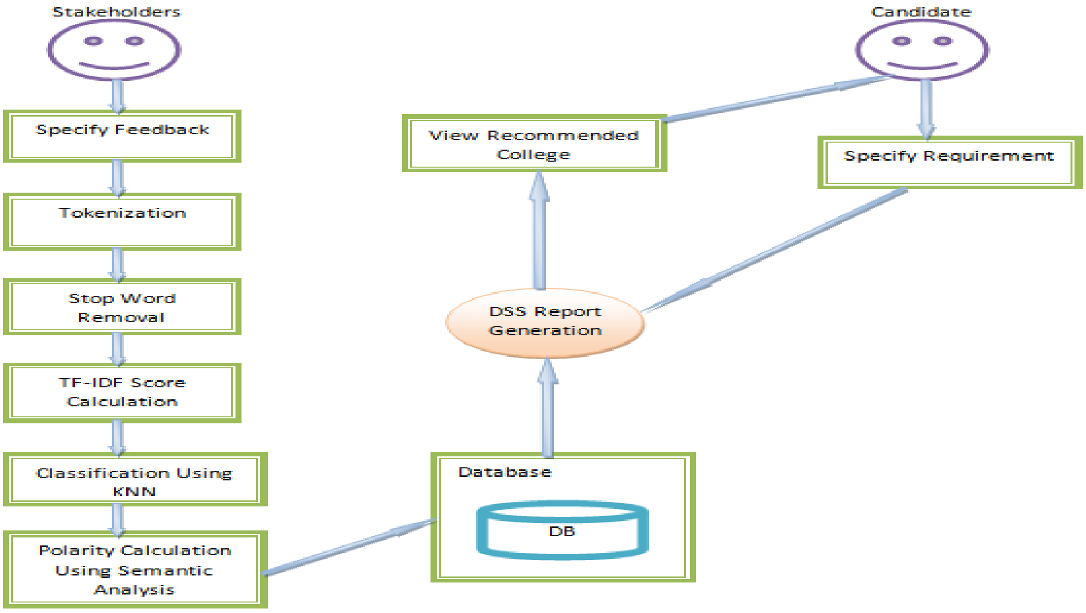


Fig.1 System Flow Diagram

3.3.1 Working of System

The flow of a system begins with the feedback given by stakeholders. This feedback will further processed for tokenization. Whereas tokenization is the process of replacing information into unique identification symbols called as tokens. After generating tokens stop word gets removed from the feedback. The stop word removal removes all the stop words from the feedback. The remaining words in feedback will be considered for TF-IDF score calculation. That calculated score and the remaining words in feedback will be classified using KNN algorithm. The classification result will be used for polarity calculation using semantic or sentiment analysis. With the help of TF-IDF score and classification result the polarity will be calculated which will show performance of every college in graphical as well as textual form. The whole result will be stored in database. On the other side when a candidate will search for college by specifying requirement. According to the requirement the DSS will automatically generate report on the basis of feedback analysis and candidate requirements. The DSS report will recommend college that will be viewed by candidate as a result of specified requirements.

3.4 Sentiment Analysis

Sentiment analysis is used to identify what people think. In general the sentiment analysis is used for polarity detection which will show the performance of college on the basis of specified feedback. The polarity can be positive, negative, neutral or together with strength. Sentiment analysis can be done with various classification techniques. For that there are different datasets are used for polarity detection. The WordNet is also a dataset with most commonly used word which automatically annotated for degree of polarity.

3.4.1 EXAMPLE

Uber: A deep dive analysis

Uber, the highest valued start-up in the world, has been a pioneer in the sharing economy. Being operational in additional than five hundred cities worldwide and serving a huge user base, Uber gets a lot of feedback, suggestions, and complaints by users. Often, social media is that the most popular medium to register such problems. The huge amount of incoming data makes analyzing, categorizing, and generating insights challenging undertaking. After analyzing the online conversations happening on digital media about a few product themes: ***Cancel, Payment, Price, Safety and Service.***

For a wide coverage of data sources, we took data from latest comments on Uber's official Facebook page, Tweets mentioning Uber and latest news articles around Uber. Here's a distribution of knowledge points across all the channels:

3.4.1.1Facebook: **34,173** Comments

3.4.1.2Twitter: **21,603** Tweets

3.4.1.3 News: **4,245** Articles

Analyzing sentiments of user conversations will offer you an inspiration regarding overall whole perceptions. But, to dig deeper, it's necessary to additional classify the information with the assistance of discourse linguistics Search. We ran the discourse linguistics Search rule on an equivalent dataset, taking the same classes in account (Cancel, Payment, Price, Safety, and Service).

3.4.1.1 FACEBOOK:

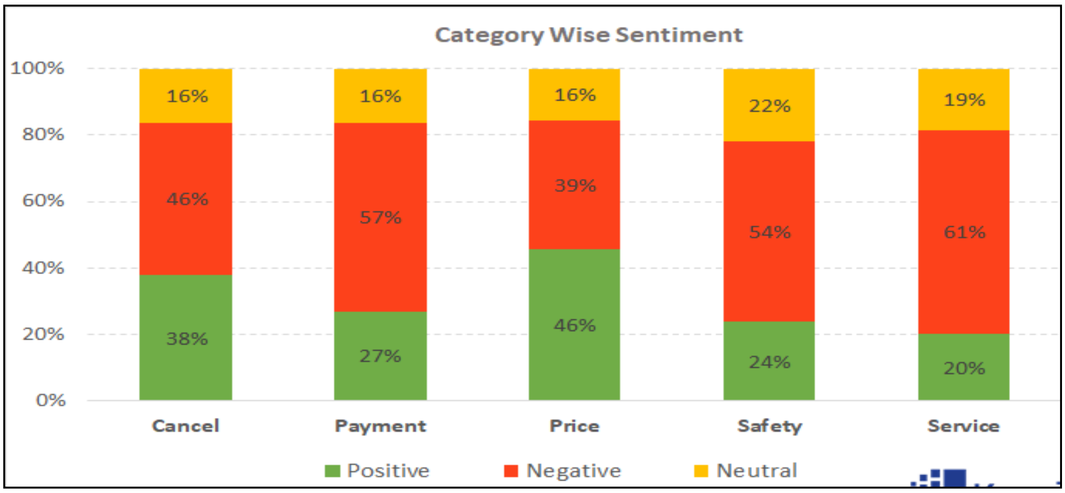


Fig. 2 Sentiment Analysis of Facebook Comments

Noticeably, comments associated with all the classes have a negative sentiment majorly, bar one. The number of positive comments associated with value has outnumbered the negative ones. To dig deeper, we analyzed intent of these comments. Facebook being a social platform, the comments are thronged random content, news shares, marketing and promotional content and spam/junk/unrelated content.

3.4.1.2 TWITTER:

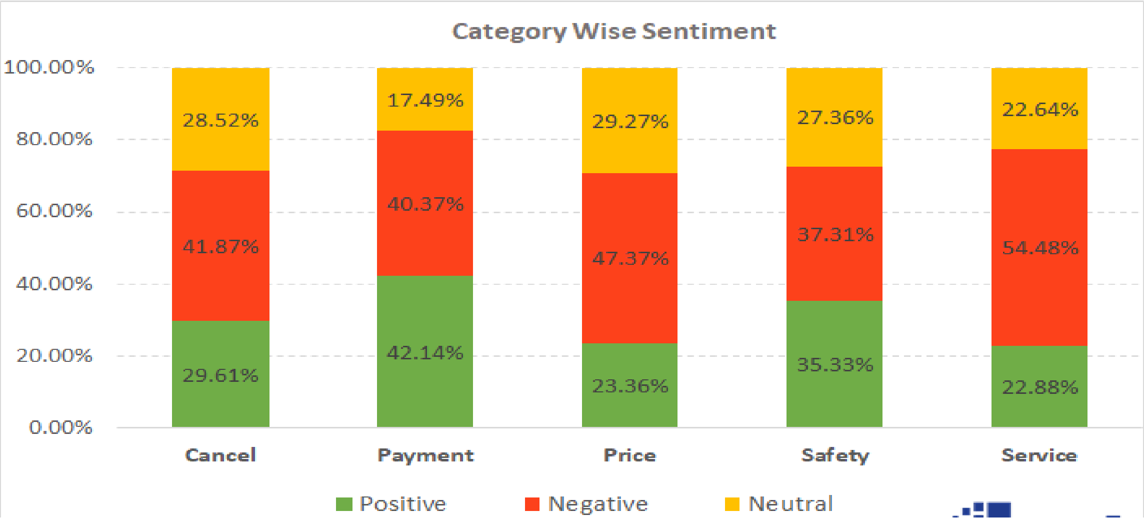
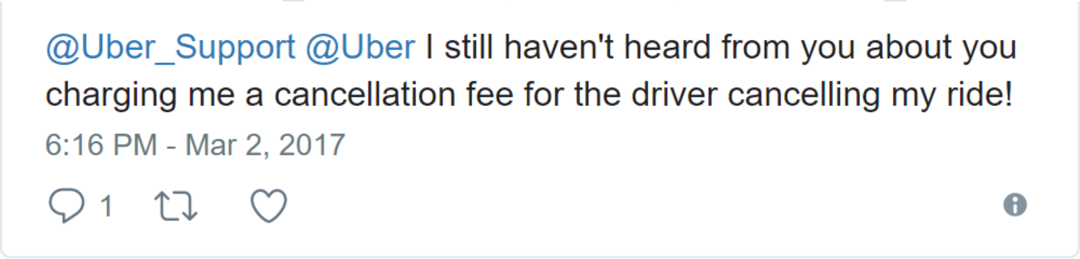


Fig. 3 Sentiment Analysis of Twitter Comments

Cancel, Payment and **Service** (and related words) are the most talked about topics in the comments on Twitter. It seems that people talked most about drivers cancelling their ride and the cancellation fee charged to them. Have a look at this Tweet:



@Uber_Support @Uber I still haven't heard from you about you charging me a cancellation fee for the driver cancelling my ride!

6:16 PM - Mar 2, 2017

Fig. 4 Comment on Twitter

Brand like Uber can rely on such insights and act upon the most critical topics. For example, **Service** related Tweets carried the lowest percentage of positive Tweets and highest percentage of Negative ones. Uber can thus analyze such Tweets and act upon them to improve the service quality.

3.4.1.3 NEWS:

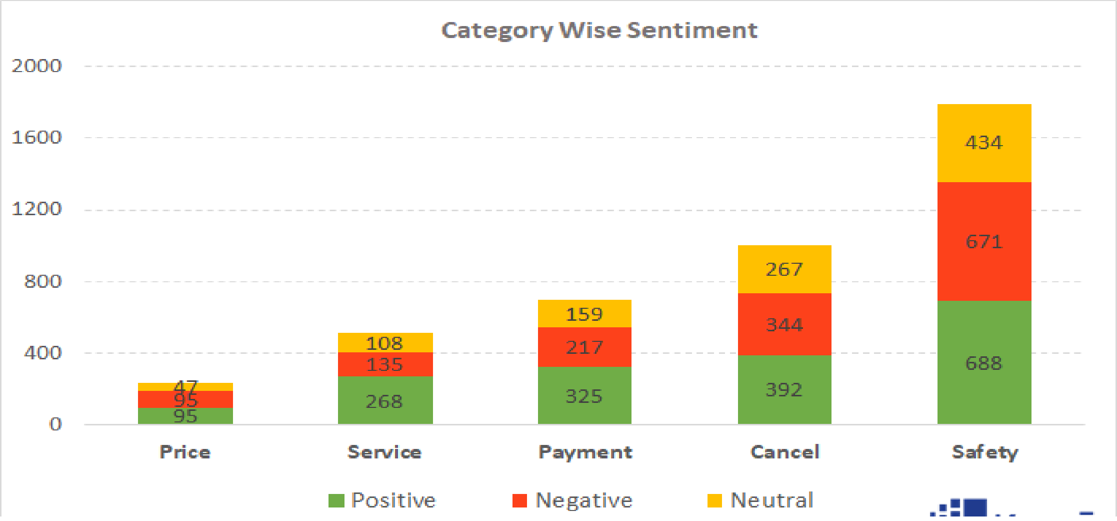


Fig. 5 Sentiment Analysis of News

News are classified based on their popularity score as well. The popularity score is attributed to the share count of the article on different social media channels. Here's a list of top news articles:

1. Uber C.E.O. to Leave Trump Advisory Council After Criticism
2. #DeleteUber: Users angry at Trump Muslim ban scrap app
3. Uber Employees Hate Their Own Corporate Culture, Too
4. Every time we take an Uber we're spreading its social poison
5. Furious customers are deleting the Uber app after drivers went to JFK airport during a protest and strike

3.5 ALGORITHM

K- NEAREST NEIGHBOR ALGORITHM

```
BEGIN
Input: D = {(x1,c1), . . . , (xN, cN)}
x = (x1, . . . ,xn) new instance to be classified
FOR each labeled instance (xi, ci) calculate
distance(xi, x)
Order d(xi, x) from lowest to highest, (i = 1, . . . ,N)
Select the K nearest instances to x: $D^k_x$
Assign to x the most frequent class in  $D^k_x$
END
```

3.5.1 Working Of Algorithm:

KNN is one of the simplest classification method used in data mining. Due to its practical efficiency it is mostly used classification method. It has superior performance for classifying data and doesn't demand fitting a model. The classification performance of k-NN is dependent on the metric used for computing distance between data points. Practically, to calculate KNN data points of interest, Euclidean distances are often considered as similarity metric.

The training samples are taken for creating classification rules of KNN without additional data. In a sophisticated approach, the groups of k objects in training set that are nearest to the test object, and base the assignment of a label on the predominance of a particular class in this neighborhood. The k-Nearest Neighbor formula may be a methodology for classifying objects supported nearest coaching examples within the feature house. KNN may be a form of instance-based learning, or lazy learning wherever the operate is merely approximated domestically and every one computation is postponed till classification.

IV. CONCLUSION

The KNN algorithm used for classification of comments gives more efficiency than other classification algorithm as it fits in any environment. In more general the comments classification by using sentiment analysis with WordNet dataset helps for polarity detection of comments. The overall performance of system will be increase and will be accurate due to KNN classifier.

REFERENCE

[1] Comments Mining With TF-IDF: The Inherent Bias and Its Removal Inbal Yahav, Onn Shehory, and David Schwartz. VOL. 14, NO. 8, AUGUST 2015

[2] Leena Deshpande, Nilesh Dikhale, Himanshu Srivastava, Apurva Dudhane, Umesh Gholap, College Recommendation System (National Conference "NCPCI-2016", 19 March 2016)

[3] Building a K-Nearest Neighbor Classifier for Text Categorization A.Kousar Nikhath, K.Subrahmanyam, R.Vasavi Vol. 7 (1) , 2016, 254-256

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

[5] Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool Ahmad Ashari, Iman Paryudi, A Min Tjoa Vol. 4, No. 11, 2013

[6] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. ACL, 271-278

[7] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability mistreatment tongue process," in Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003, pp. 70– 77.

[8] Shengyi Jiang, Guansong Pang, Meiling Wu, LiminKuang, An improved K-nearest-neighbor algorithm for text categorization.

[9] Dr. Riyad Al-Shalabi, Dr. GhassanKanaan, Manaf H. Gharaibeh, Arabic Text Categorization Using kNN Algorithm.

[10] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010 SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. LREC-2010 –http://sentiwordnet.isti.cnr.it/

[11] Xiaogang Han, Junfa Liu, ZhiqiShen, and Chunyan Miao,An Optimized K-Nearest Neighbor Algorithm for Large Scale Hierarchical Text Classi_cation

[12]Sentiment analysis for facebook on different aspects https: // towards data science.com/ sentiment-analysis-concept-analysis-ssssand-applications- 6c94d6f58c17