

**FINDING PINPOINT OF INTEREST FROM RESTAURANT REVIEWS BY USING
LSA TOPIC MODELING**

BY

SHEIKH MARUF HOSSAIN

ID: 133-15-3051

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. SAMIA NAWSHIN

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

07, MAY, 2018

APPROVAL

This Project titled “**Finding Pinpoint From Restaurant Reviews Using LSA Topic Modeling**”, submitted by **Sheikh Maruf Hossain**, ID No: 133-15-3051 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 7th May 2018.

BOARD OF EXAMINERS



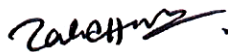
Dr. Syed Akhter Hossain
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman



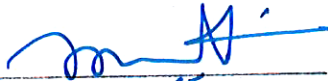
Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Md. Zahid Hasan
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Jahangirnagar University

Internal Examiner



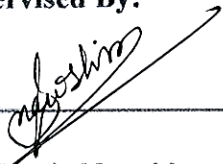
Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

External Examiner

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Ms. Samia Nawshin, Lecturer, Department of CSE Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised By:



Ms. Samia Nawshin
Lecturer
Department of CSE
Daffodil International University

Submitted By:



Sheikh Maruf Hossain
ID: 133-15-3051
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year Thesis successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Ms. Samia Nawshin, Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Data Mining*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

In recent years it is noticeable that sharing text reviews on various businesses specially restaurants through website and social media is a very common phenomenon. Online reviews reflect user's opinion. This huge collection of user data in terms of text reviews can be analyzed to identify user's sentiment and their demand also. Here users are the primary sources. Text reviews are the complete reflection of user's sentiment and also owned by them.

Measuring user's sentiment will also be able to find out the market position of a Restaurant. By making the machine learned about the total reviews, it will be able to categorize the unknown text.

we collect the necessary data for our research work from a verified source. we took a step forward by combining user review texts which were collected from that website to build a model that can give some idea about the topics about what users think the most while writing a review on any restaurant.

Key benefit of our approach is that, by using our proposed Topic model, Owners can identify the main focused term from the review of customers and also can take future step to work on that. As this model is based on text document, it will be very perfect work in all terms and condition.

TABLE OF CONTENTS

CONTENS	Page no
Board of examiners	II
Declaration	III
Acknowledgements	IV
Abstract	V
Chapter 1: Introduction	1
1.1 Introduction	
1.2 Motivation	
1.3 Rationale of the Study	
1.4 Research Questions	
1.5 Expected Output	
1.6 Report Layout	
Chapter 2: Background	4
2.1 Introduction	
2.2 Related Works	
2.3 Research Summary	
2.4 Scope of the Problem	
2.5 Challenges	

Chapter 3: Research Methodology

7

3.1 Introduction

3.2 Research Subject and Instrumentation

3.3 Data Collection Procedure

3.4 Research Procedure

3.5 Knowledge Discovery in Database

3.6 Data Pre-Processing

3.6.1 Review Selection

3.6.2 Tokenizing

3.6.3 Feature Extraction

3.5 Applying Algorithm

Chapter 4: Experimental Results and Discussion

15

4.1 Introduction

4.2 Experimental Results

Chapter 5: Summary, Conclusion, Recommendation and Implication for Future Research

17

5.1 Summary of the Study

5.2 Conclusions

5.3 Future Scope

LIST OF FIGURES

Figure 3.4.1- Research Procedure

Figure 3.5.1- KDD process is used here for basic preprocessing

LIST OF TABLES

Table 4.2.1- Topic Model Table

LIST OF ABBREVIATION

DIU – Daffodil International University

CSE – Computer Science and Engineering

NLP – Natural Language Processing

POS – Parts Of Speech

NLTK – Natural Language Toolkit

KDD – Knowledge Discovery in Database

LSA – Latent Semantic Analysis

LDA – Latent Dirichlet Allocation

TF-IDF – Term Frequency Inverse Term Frequency

SVD – Singular Value Decomposition

CHAPTER 1

Introduction

1.1 Introduction

Nowadays it's been noticed that several businesses offer their consumers for expressing their opinions via reviews in respective website or social media. In this observance, Restaurants are quite ahead. This reviews by customers, mirrors their demands through disclosing personal sentiment on items, services, and overall restaurant itself. So using them as source researchers do the analysis of sentiment. In last few years in Bangladesh, the users of social media, specially has been very popular to share their personal experience or feedback of restaurants.

By dint of that owners get helped by catching where the interest of customers is. It needs a huge collection of data as text reviews. Those measurements and analysis of sentiments results the position of any particular restaurant in the market. Using machine learning techniques, I will work on unsupervised learning. The outcome of Supervised learning is predetermined by both the machine and the applicator of the process. But In Unsupervised Learning there is no predetermined output. Machine calculates this for us by using Unsupervised learn. This is how our work will be done by using Topic model as we work on Unsupervised learn.

On those reviews, I will be able to find topics from those texts. In order to get those topics, reviews were collected from “Priyo review (Beta)”, a Bangladesh based website.

In this work, latent subtopics observed from “Priyo” review has been described through applying quantities of high dimensional review data. These topics can provide significant acuteness to an algorithm, online Latent semantic analysis. The aim was to point out client’s demand from huge restaurants about what clients care approximately with a view to increase their ratings in market which instantly will impact on their revenue. We followed online Latent semantic analysis in order to get latent subtopic from reviews. LSA is a natural language processing technic in order to extract and represent the contextual-usage meaning of words by statistical computations applied to a large corpus of text. I am the first who is working on Bangladeshi Dataset.

1.2 Motivation

This research is enormously motivated by the yelp dataset challenges and various methodology to obtain a goal. Considering text documentation as an interesting field, I can get useful tools to fulfill our intended target.

One of the fact that really motivated us that there is rarely any work has been done that utilized any dataset of Bangladeshi website that features Bangladeshi Restaurants. These are the reasons we expect this to be an uncommon work. I am the premier to work with Bangladeshi dataset.

1.3 Research Question

Some questions have been defined for this work which to be answered serially.

1. What is the reason behind this research?

2. How the machine will generate topics from the reviews?
3. How the method to be implemented
4. How the accuracy will be obtained from the algorithm
5. Which method is faster and most accurate?

1.4 Expected Output

The model I have built, by applying different types of algorithms of unsupervised machine learning we will accurately test the model's performance. In that way we can know, how friendly the response of the algorithm with our data set.

We also get the total word's number of dataset and number of restaurant reviews. A good result of the reviews of a particular restaurant is expected from this research through answering user's priority, topics in discussion by making high dimensional data into lower dimensional using LSA model.

1.5 Report Layout

Chapter 2 discusses with the background of the project. Related works with this research and the challenges also informed in this chapter.

Chapter 3 emphasizes on the method that used in this project work. Explicit works of Natural Language Toolkit technics, data mining, machine learning and data collection procedures.

Chapter 4 speaks about result in details with experiment

Chapter 5 discusses about the future opportunities in our project research that we can attain and thus the conclusion of the thesis

CHAPTER 2

Background

2.1 Introduction

Finding subtopics is a system that frankly basis of user's sentiment of a particular business. Various social medias like wiki, fb, twitter etc. have huge text documents. Before consuming any service of any particular business, such as purchasing goods, foods or enjoying movies.

Therefore, these reviews from online are the key to analyze user's sentiment. That is what makes easy to work with these on topic modeling.

2.2 Related works

To factor models of discrete data there has been some ways. LDA or Latent Dirichlet Allocation is one of them which is almost very basic and very popular that works as factor model to deal with factor and topics. This model has been used for Yelp dataset. Probability distribution of documents on topics as K -parameter is the principal thing whereas K is the hidden topic number. [1]

As an another way, an algorithm called Laplacian Probabilistic Latent Semantic Indexing LapPLSI,

used for modeling the document space by dealing with closest neighbors [2]. There have been also a work of term and text clustering such as email clustering of 90,000 terms into 50 clusters which used non-negative matrix factorization techniques. The corresponding ways to factor analysis are means clustering, Spectral Clustering [3], nnProbabilistic Latent Semantic Analysis [4].

However, The importance of sentiment analysis for understanding the choice of consumer has been highly focused [5]. Therefore, we intend to make our machine learned by a system that we can consider a machine learning system.

An algorithm LSI or Latent Semantic Indexing used as dimensionality omission process which is an information recovery technic that reduce data to a latent space representation by using singular value decomposition [2,6]. However later on a similar algorithm called PLSI or Probabilistic LSI was introduced in order to generate a probabilistic model, from a mixture model for modeling every word in the documents as sample [7].

The other works related this project are, restaurant category prediction from a text document
Generating reviews automatically with Markov chain review generators; The most positive and negative corpus in reviews.

Star ratings prediction using sentiment analysis and using clustering, business categories prediction have also been happened previously.

I worked on Priyo dataset that contains the info about reviews given by users, check-ins, businesses and users themselves. Here we specified our work on restaurant data regarding all type of info.

2.3 Scope of the problem

It is a very wide performing scope with our research . In this process, the capability of number of topics are kind of unknown. So in order to solve this problem we will utilize LSA or Latent Semantic Analysis. Since the dataset we are working with is too new that less than very few task has been done with that till now.

2.4 Challenges

To find out the word selection and model selection is a big challenge. It's not easy to find out. We needed some valid dataset that is fully accurate.

Feature selection was also so much important for this research. Algorithm applying was also critical task. For this research which algorithm is very suitable, finding this was not an easy task.

CHAPTER 3

Research Methodology

3.1 Introduction

The procedure of mining any text is not as same as numeric data, since we can't directly utilize any text document and any numeric result is unable to be generated.

Finding out as usual feature that is found in numeric data, it's quite impossible for text document since we are not having those text document as tabular form.

So text data needed to be pre-processed and was converted to numeric format.

3.2 Research Subject and Instrumentation

The goal for the research has been set very specific. We have worked on users' opinion or review for the restaurants. So about restaurant, what is the sentiment of users is very important to identify. In order to do that, we applied a way to analyze the users' review and made the machine learned about each word and also the relation of each word of a review.

Then the topic model was built. That's why our research perspective is to build a good topic model by using review text data.

But here we have mostly used the Sci-Kit learning library which is developed by python programming language. Now it's becoming a very popular and useful tool for analyzing data and also solving machine learning task.

The majesty of sci-kit learning is that we can import different kinds of libraries which include different algorithm visualization tools. Here in our Research all Machine learning procedures and

data visualizing task have been done by sci-kit learning library.

3.3 Data Collection Procedure

It's a highly challenging to collect data. The dataset we collected for the purpose of our work, is Priyo review dataset, a website of Bangladesh where various types of reviews or ratings that users give to any business are collected. Login is a must thing to do in order to rate or give review on that site.

Since we deal with only Restaurant reviews; so collected those data about only reviews of restaurants from the targeted website.

Over there, fifty restaurants are available but most of them have only a few review. So, we have selected top five restaurants reviews for our research purpose which has more than thirty reviews. Collection of different reviews will provide better result.

3.4 Research Procedure

For achieving our goal we have maintained some steps. These all are related with each other. As we have dealt with text data, the method is isolated. Figure 3.1 shows the processing step of research procedure.

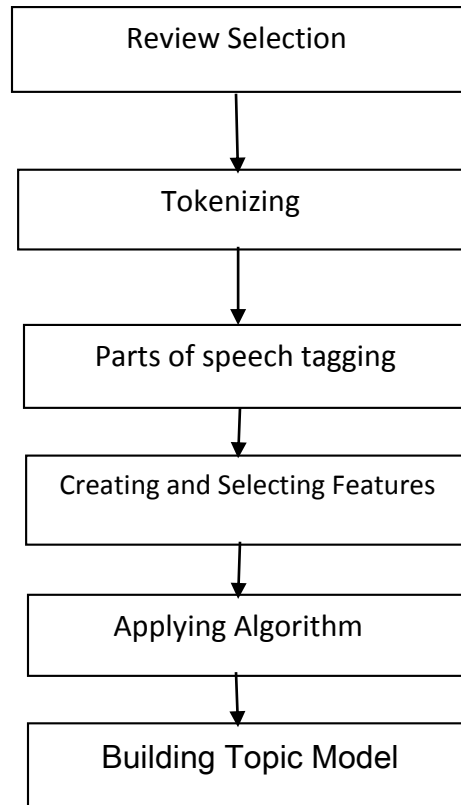


Figure 3.4.1: Research Procedure

3.5 Knowledge Discovery in Database

KDD process is the basic structure for extracting knowledge from raw data. For finding knowledge from data this process is being used globally. For any kinds of machine learning problem we also followed KDD process to complete our task.

The process by how we can extract knowledge from known or unknown data is given be in figure 3.2 .

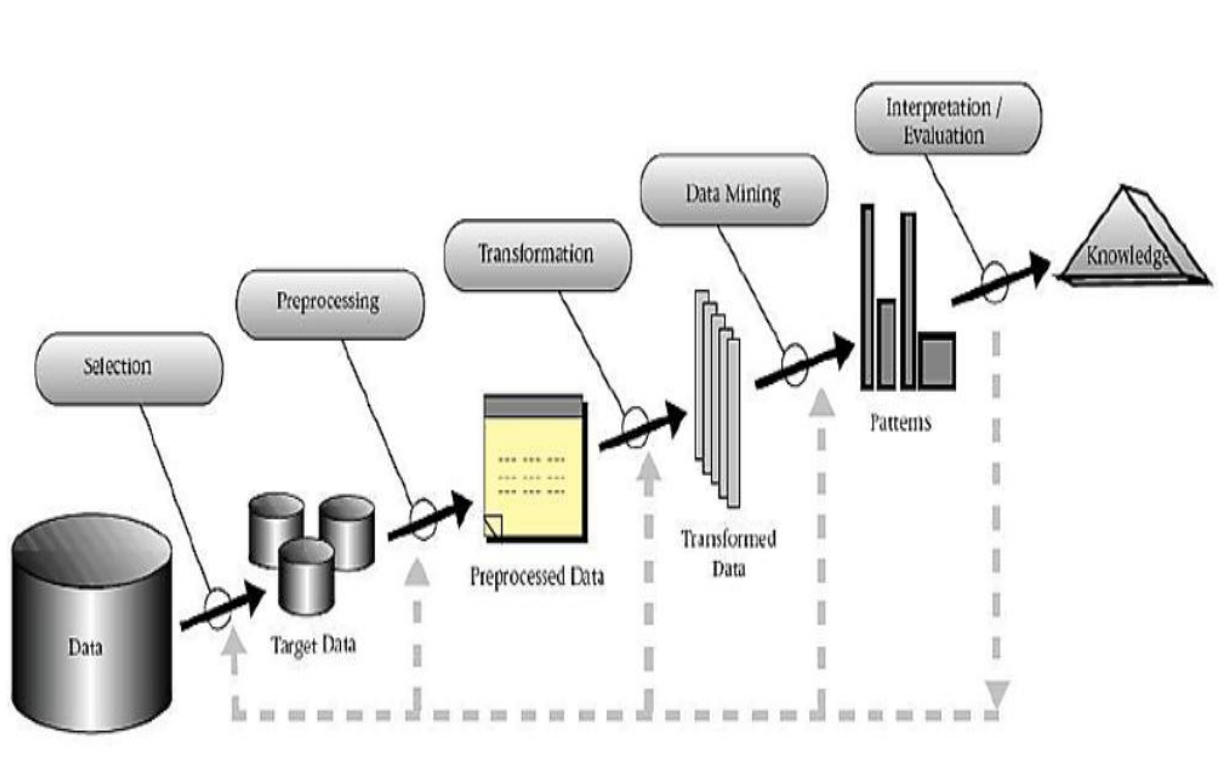


Figure 3.5.1: KDD process is used here for basic preprocessing

3.6 Data Pre-Processing

3.6.1 Review Selection

On the basis of the quality of the reviews, the reviews has been selected. Priyo review collects reviews of local business of Bangladesh. Therefore, I got a lot of reviews which was written in Bangla in English character, those are meaningless like “Naan ruti khaite onek moja lagse, Onek Shundor jayga”

These types of reviews couldn't be used in the training data set. As I have used the sci-kit learning libraries, the tool can't identify the word which is not in English like following “ Food quality “Prochondo baaje, Uff, faaltoo, bichchiri” etc.

In these, users gave also mixed language reviews what we have removed as well.

3.6.2 Tokenizing

For knowing the meaning of the sentences in our data set, we have used sentence tokenizer. To identify the word quantity we have used word tokenizer.

“we always visit there family and also friends sometimes alone.the place is really very large .the foods are delicious all the time . i always enjoy that .the tea and legroast are the best but i can never find chicken jhal fry whenever go there after 7:30.but all foods are delicious”

After tokenizing it look like : ['we always visit there family and also friends sometimes alone.the place is really very large' , 'the foods are delicious all the time .' , 'i always enjoy that .the tea and legroast are the best but i can never find chicken jhal fry whenever go there after.' , 'but all foods are delicious']

and after tokenizing it look like ['we', 'always', 'visit', 'there', 'family', 'and', 'also', 'friends', 'sometimes', 'alone.the', 'place', 'is', 'really', 'very', 'large', '.the', 'foods', 'are', 'delicious', 'all', 'the', 'time', '.', 'i', 'always', 'enjoy', 'that', '.the', 'tea', 'and', 'legroast', 'are',

3.6.3 Parts of Speech Tagging

It is also a special feature because it can declare each word of a sentence of what parts of speech it is. Here we have shown an example of it.["The place is really very large. The foods are delicious all the time. I always enjoy that. The tea and leg roast are the best but I can never find chicken jhal fry whenever go there after 7.30 but all foods are delicious"]

[('the', 'NNP'), ('foods', 'NNS'), ('are', 'VBP'), ('delicious', 'JJ'), ('all', 'PDT'), ('the', 'DT'), ('time', 'NN'), ('.', '.') ('I', 'NN'), ('always', 'RB'), ('enjoy', 'VBP'), ('that', 'IN'), ('.the', 'NNP'), ('tea', 'NN'), ('and', 'CC'), ('leg roast', 'NN'), ('are', 'VBP'), ('the', 'DT'), ('best', 'JJS'), ('but', 'CC'), ('i', 'NN'), ('can', 'MD'), ('never', 'RB'), ('find', 'VB'), ('chicken', 'JJ'), ('jhal', 'NN'), ('fry', 'NN'), ('whenever', 'WRB'), ('go', 'VBP'), ('there', 'RB'), ('after', 'IN'), ('7:30.but', 'CD'), ('all', 'DT'), ('foods', 'NNS'), ('are', 'VBP'), ('delicious', 'JJ')]

We focused mostly on adjective (tagged as "JJ") and noun (tagged as "NN") for finding positive and negative identifier keywords because other words generally not contain any logical information or mean any sentiment of a user.

3.7 Applying Algorithm

My research is based on unsupervised learning. Topic model is a unsupervised learning model. That's why I have used LSA algorithm. As there are different types of tools available for using the algorithm, the procedures are also different. For my research I have used sk-learn library

since it is quite exceptional from other tools.

Latent Semantic Analysis is a dimension reduction method or technique in text mining. LSA and LDA are the topic modeling algorithms, Whereas LSA was first introduced in 1988 ,an information retrieval process as latent semantic structure. But in spite of that, no notable work was done in that period until early 2000s. LDA or Latent Dirichlet Allocation was discovered first time in 2003 by blei at al [1]. Since LSA is an up to date algorithm, an effective outcome is very much expected from this algorithm. A term document matrix also called as TF-IDF or term frequency–inverse document frequency is used as Input to LSA. Here, each document has count of a bag-of-words, which can be called as “vector” of the different terms that appears in that. The different term vectors can also be clustered using a clustering algorithm such as k-means.Now The main difference between Clustering and Latent Semantic Algorithm is, each document to a specific cluster is assigned by Clustering algorithms whereas, a set of topic loadings to each document is assigned by LSA algorithm.

3.8 Implementation Requirements

There was 500 reviews were used for our training and testing data set. I used a bag of words to extract features from the reviews. These words are a sparse vector of occurrence counts of words. I have got 2000 words in my total dataset and the sci-kit learning library considers each word as a feature.

I have taken 8 documents for my models, And from those 8 docs not every documents has meaningful topics. But from the others we can off course generate idea about those topics.

From the achieved table from topic model, I got 6 relevant topics that is really described by its subtopics . For example, Topic 1 has the words like good, place .etc. so it is talking about the place of the restaurant. Topic 2 has words like Burgers, coffee, peri. That means it's saying about fast food or quick served food. Topic 3 talks about Chinese food etc. so it can say about dinner. Apparently topic 3 also has words like nightmare , saltz. That means it describes bad services. Topic 4 has KFC, chicken, ten , overrated. That means it talks about price. Topic 5 talks about Lunch . from the words leg, roast, naan. In the same way topic 5 talks about Environment. Where the words are welcome, music quiet, environment.

CHAPTER 4

Experimental Result and Discussions

4.1 Introduction

To achieve a good output, my research depends on good data selection as well as creating the perfect model. For creating a model by using text data and finding better result the language processing part is also very important.

4.2 Experimental Result

In this field, from the dataset I'm working with, I'll elicit the topics out of those reviews of the dataset. There will be subtopics of each topic extracted. Those subtopics in a particular topic will be related to each other to describe the topic these belong to.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
PLACE	FAST FOOD	DINNER	PRICE	LUNCH	Environment
is	Burgers	place	KFC	Leg	Naga
The	one	is	so	roast	and
Good	peri	you	chicken	naan	nice
And	coffee	of	was		were
food	fair	sub	fried	interior	burger
Place	in	food	original	reasonable	welcome
for	tried	chinese	ten	chicken	music
was	to	nightmare	to	cave	drink
of	of	offers	they	satisfactory	quet
not	the	saltz	overrated	tea	environment

Table 4.2.1 Topic Model Table

CHAPTER 5

Summary, Conclusion and Implication of Future Scope

5.1 Summary of the Study

In order to accomplish my research, I have studied how the techniques of machine learning can be used to get the solution of different kinds of problems about machine learning.

Two different types of machine learning method those are supervised and unsupervised learning has been learnt. That made us known about that our project is all Unsupervised learning and it is a topic model problem.

So we learned deeply about topic model and related algorithms. As our dataset is all about text documentation so we studied about how natural language processing is used for the process of text documentation.

There are many tools for this field. We have learned which tool is better for this research. We have learned different kinds of tools to complete our research. We deeply learned about NLTK and Sci-kit learning libraries and general python programming language.

.

5.2 Conclusion

For beginners, machine Learning is a good area for researching. In Bangladesh related work like user review data of a website has not been done so much. I tried to build here a good model so that the restaurant owners may have acknowledge about the users' choice. By this they can take some effective step in future for the betterment of their service. We have used the review from online that was live because everyone is giving review daily. so lot of reviews are adding every day. I tried to describe all the working methods, models and also the procedure with figure and table.

I have a plan to make this research completed in additional requirements as well.

We have also faced a lot problem to complete this research. There was so noisy data. All kinds of procedure were vast in sum so it took a lot time to understand and implement this in our research.

For data collection we have to work hard. Though we mentioned some related works but actually it is few in number related to our work. But we tried our best. So, for making this kind of project it's needed a tremendous work for guiding us through the right path of research.

My supervisor teacher Ms. Samia Nawshin madam helped me a lot throughout the research.

We have experienced some other problems that were in the beginning of our research. We have also stuck with the learning of the huge field of data mining and also machine learning.

Everything was done by consultation which made it success indeed. In this process if the restaurant owner works on their own dataset, they can easily assume the customers' demand by seeing all those topics and relevant words. They can make a good decision to improve their restaurant business.

5.3 Future Scope

From this it can be done LDA topic modeling by which restaurant owners can identify what the customers focus on mainly. We can rank the restaurant by using boosting algorithm.

In future it will also be possible if we want the users identity. We want to develop a website in where restaurant owners can check their own restaurant's review. By this they will know how many positive and negative reviews have been given by their users. By this project we will be able to show that which restaurant is weekly topper that means who owns the maximum positive reviews.

REFERENCES

- [1] D. Blei, A. Ng, and M. Jordan. "LatentDirichlet Allocation." Journal of Machine Learning Research, 3:9931022, January 2003

- [2] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. "Indexing by latent semantic analysis." Journal of the American Society of Information Science, 41(6):391407, 1990.

- [3] R. Zass, A. Shashua "A Unifying Approach to Hard and Probabilistic Clustering". International Conference on Computer Vision (ICCV) 2005.

- [4] E.Gaussier C. Goutte "Relation between PLSA and NMF and Implications" ACM SIGIR conference on Research and development in information retrieval. SIGIR 2005

- [5] Mike Thelall, Kevan Buckley, GeorgiousPaltou, et al . "Sentiment strength detection in short informal text." Journal of the Association for information Science and Technology 61.12(2010): 2544-2558.

- [6] C. Papadimitriou, P. Raghavan, et all. "Latent Semantic Indexing: A Probabilistic Analysis." Journal of Computer and System Sciences. October 2000.

- [7] M. Hoffman and D. Blei. "Online Learning for Latent Dirichlet Allocation." Neural Information Processing Systems, 2010.

PLAGIARISM

The screenshot displays the Plagiarism Checker interface. The top navigation bar includes the 'plagamme' logo, a notification bell with a '1' indicator, and links for 'FAQ', 'Student', and '1\$'. A search bar and a 'Listed view' toggle are also present. The left sidebar contains an 'Upload' button and menu items for 'Papers', 'Payments', 'Free', and 'Earn money', along with 'RATE US' (5 stars) and 'CONTACT US' (chat icon). The main content area shows a document titled 'paper (2).docx' uploaded '12 minutes ago'. A circular progress indicator shows a '13%' similarity score. Below this, three categories are listed: 'Paraphrase' (2%), 'Improper Citations' (0%), and 'Matches' (19). At the bottom, a warning banner displays three red stars and the text 'HIGHEST PLAGIARISM RISK'.