International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 08 | Aug -2017 www.irjet.net p-ISSN: 2395-0072

Knowledge Discovery of Small Business Domain Using Web Crawling and Data Mining

Latha M[1], Shivanand R D[2]

[1]*M.Tech. Department of Computer Science and Engineering, Bapuji Institute of Technology, Davanagere, Karnataka, India.*
[2]*Associate Professor, Department of Computer Science and Engineering, Bapuji Institute of Technology, Davanagere, Karnataka, India.*

Abstract - Now a days come to be where everything information is obtainable on the web. If also there are many complications to make use of the web professionally. Due to more information, the users does not get the information related to their requirement. . Among the businesses, a lots of small business are generally not noticed by the people. The static features of small business are business name, area, contact and website address of business. Related to small business these static features are easily cached by people. The dynamic features of the business includes reputations. But it is very difficult to get the reputation of small business. Because of the shortage of means of boosting users to write remarks for their services. In this project focused on developing a knowledge base of small business and to give the valuable information about the business to user. To develop a knowledge base of small business static and dynamic data is considered. The web crawling applied on the static data. Twitter considered as the dynamic data source. The user posting the tweets about the business. The posted tweets are analyzed to know whether the tweets contains positive or negative responses about the business.

***Key Words*: Small business, data analysis, web crawling, data mining, knowledge base, knowledge discovery**

1. INTRODUCTION

Currently more information is available on the web related to any subject. But these more information is to be sort out. Because of the big data [1] defines the three main features such as capacity, speed and diversity. Due to this many complications to utilize the source of web in well-organized manner. Most of the user does not get their required information from the web. Related to this most the web content written in web standard markup language that is HTML. It is intended for web programs to parse so they can draw graphical designs. The graphical designs the most part containing writings, connections, images, and some of the time it also contains different media content. Given the outline reason for HTML to show data on screen is the main purpose of HTML. But it is nearly difficult by the computer programs to retrieve helpful data through the HTML composed content without additional language processing.

With respect to above background where information on the web was quickly developing. To increase the performance of the small business is get increasing if the there is an efficient small business exist. To develop a well-organized small business it requires a good knowledge.

Small Businesses [2] are privately or independently owned organizations. The size of the small business is limited in size and only few employees are working in that organization. The annual returns of the small business is less compared to the other business. The very less amount is enough to start small this is one of the advantage of small business. Individuality is another advantage of small business. The small business owners have the capacity to manage by themselves it makes a great thing.

The small business routines mainly connected to meet the user requirements. Those organizations that embrace this promoting idea will probably succeed when they are consolidated with development and differential procedures .The market-orientated vital stances of effective private companies are directed by the requirements of client drove everyday exercises. Small business successful is mainly concentrating on the handling of customer drive and market positioning. If these two perspectives are well handled by the small business means it leads to a successful, the present with upcoming.

With respect to above, developing a knowledge base [3] of small business and to deliver the information about the business to users. Now a days a lots of business are developing rapidly. But the small businesses are not easily noticed by the people. Related to small business the static data once stored it's not change over time. The static data of the small business are business name, area, contact and website address. These data are easily available to the public. But the dynamic data it will change over time. The dynamic features of the business are user opinion about the services or administration. These features are very difficult to retrieve. Because they have a lack of methods to support user to write remarks about for their administration services. The Small business examples are hotels, stores, restaurant, bakery and shops. In this project restaurant is taken as the small business example.

Primary needs of human beings are food and beverages. These primary needs are provided by restaurant. Because of this the restaurant business act as the lifespan business. And it's not a certainly losing attraction in customer's encouragement. If the owner of the restaurant is good capacity to achieve and run it well means the restaurant becomes fast growing. Some individuals thought that if good quality of food and well atmosphere services provided to customer's means to become a successful business. But many functions that are considered for the successful of restaurant not only good taste of food.

To retrieve the reputation of the small business it is very difficult so for that social media considered for the reputation. Now a days the common thing by the people is posting the status about their daily life on social media. Usually the status of the people in the form of comments. These comments involves where they went to spend their valuable time, in which hotel they had food and what is the taste of the food and how was environment of that hotel. The people added status about their daily life usually it is just positive or negative opinions about the service taken by particular organization. The user posted comments on social media are in the form of natural languages and these comments are does not indicating any rating. So some data mining techniques are requires to increase the accuracy of the information.

In this project focused on developing a knowledge discovery [4] of small business. First business static data that is website address is considered. And dynamic data as the Twitter [5]. The web crawling applied on the static data. Then extraction of information from the corresponding webpage. Then allow the user to post the comments about the particular business. Then the user posted comments are analyzed. The analyzing of user comments to know whether the tweets [6] contains positive or negative response about the business. If the comments are more positive responses about the business then it gives the total number of positive tweets with accuracy about the particular business. If the comments are more negative responses about the business then it gives the total number of negative tweets with accuracy about the particular business. So some kind of data analysis [7] is needed to figure out how much positive or negative each of the tweets are about the business.

1.1 Web Crawling

Web crawling [7] is the process of search engines combing through web pages in order to properly index them. Web crawling makes it easier for search engine to return the most relevant results to user after they enter a search query.

1.2 Data Mining

Data mining [8] is the automatic or semiautomatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records, unusual records and dependencies.

2. SYSTEM ARCHITECTURE

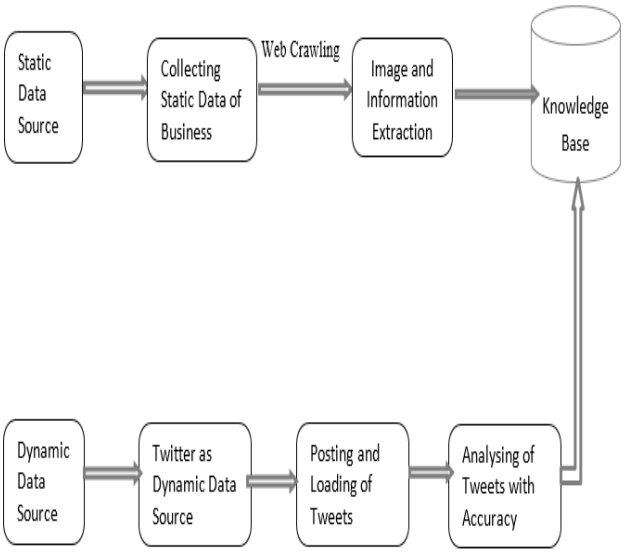


Fig -1: Knowledge Discovery Process

The overall processes for the knowledge discovery are overviewed in the Figure 1.It involves the web crawling on static data source. Image and information extracted from the retrieved page is stored in knowledge base. Twitter as the dynamic data source. Users posted tweets are analyzed with the accuracy.

3. RELATED WORK

The following sections gives the related work for this paper. Related to knowledge discovery of small business both static and dynamic data is considered.

3.1 Web Crawling On Static Data

The initial stage involves data collection of the form static. The defined static data that are not produced by users and likely to remain unchanged for a long time as static data. These data include their names, locations, and contacts, types of businesses. In this phase web crawling applied on the static data of business to produce knowledge base of small business. From the crawled page image and information is extracted is stored in the knowledge base. The website address of Apoorva resorts is located in davanagere is http://www.apoorvaresortsdavangere.in.The following figure shows the web crawling on static data.

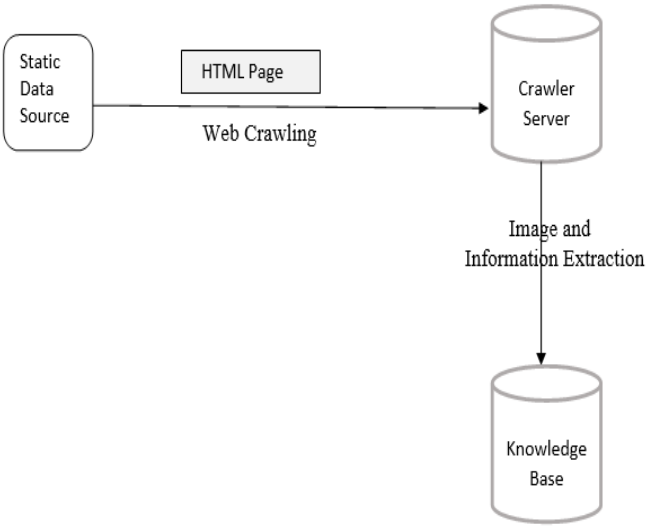


Fig -2: Web Crawling Process

3.2 Tweets Analysis

In this phase first user Twitter account created. To generate Twitter API keys, Access Token and Secret Keys the Twitter application is created. To create Twitter Application which is mandatory to access Twitter. It includes the following steps.

1. Visit the website https://dev.twitter.com/apps/new and logging into Twitter account then click Create New App.It is shown in the figure 3 and figure 4.

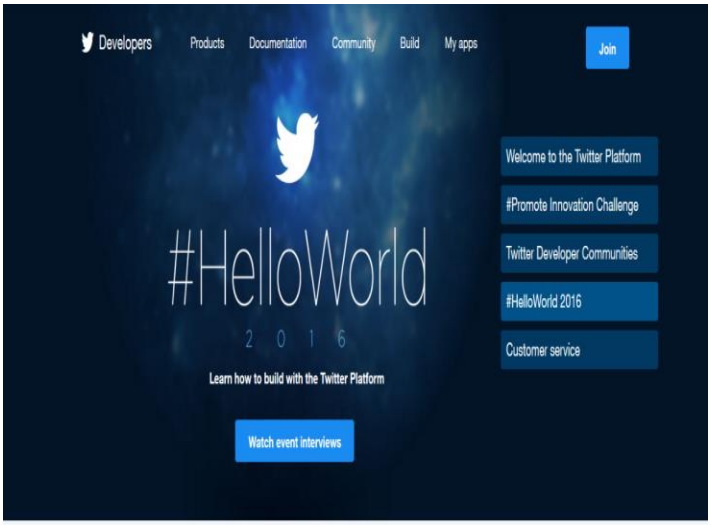


Fig-3: Home Page of Twitter Developer Platform

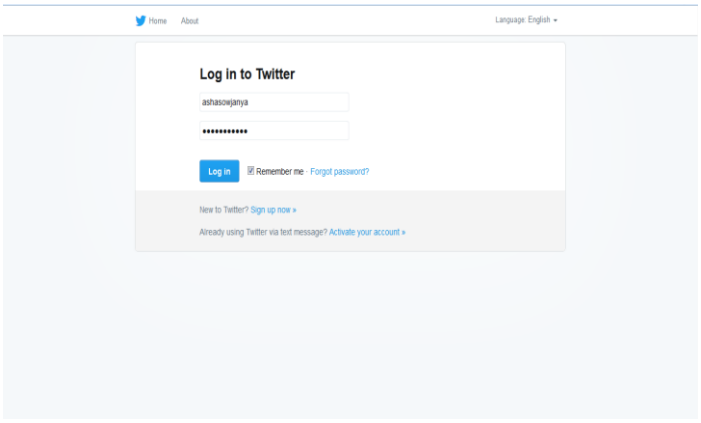


Fig-4: Login Page of Twitter

2. On the Create an application page, enter the requested information for the new twitter application. Enter the Application Name, Description and user website address. Choose an application name it must not be already taken by another user. It is shown in the following figure 5.

3. Twitter Application is created. On the main page, click Keys and Access Tokens. This Application Settings section contains the Consumer Key and Consumer Secret. It is shown in the following figure 6.

4. Under the Your Access Token section, click Create my access token. The access tokens are generated and ready to be used. It is shown in the following figure 7.

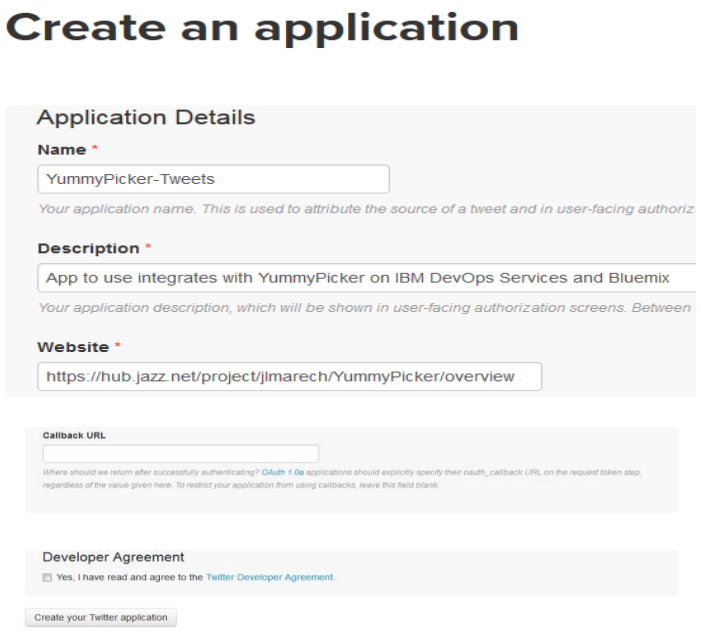


Fig-5: Creation of Twitter Application Details

Fig-6: Consumer Key and Consumer Secret

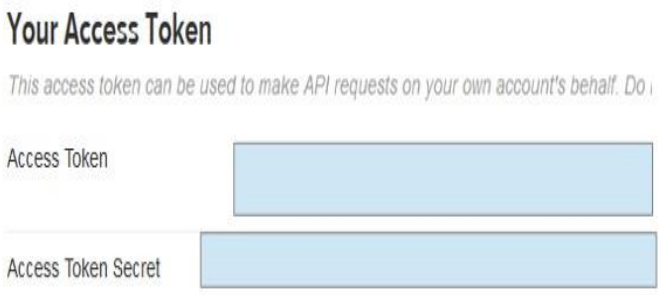


Fig-7: Access Token

5. Twitter authentication information includes the following elements: Consumer Key, Consumer Secret, Access Token and Access Token Secret. All these need for programmatically to Twitter.

Then finally related to tweet analysis the user will tweets are posted and loaded. And tweets are analyzed either positive or negative with the accuracy. The entire tweets analysis is shown in the figure 8.

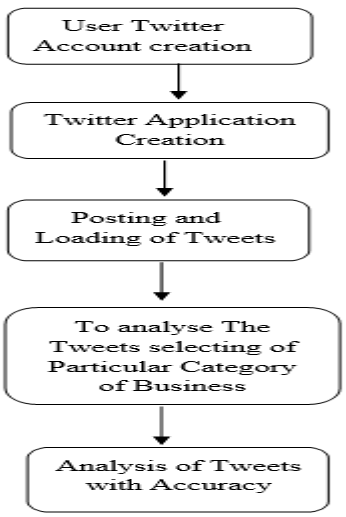


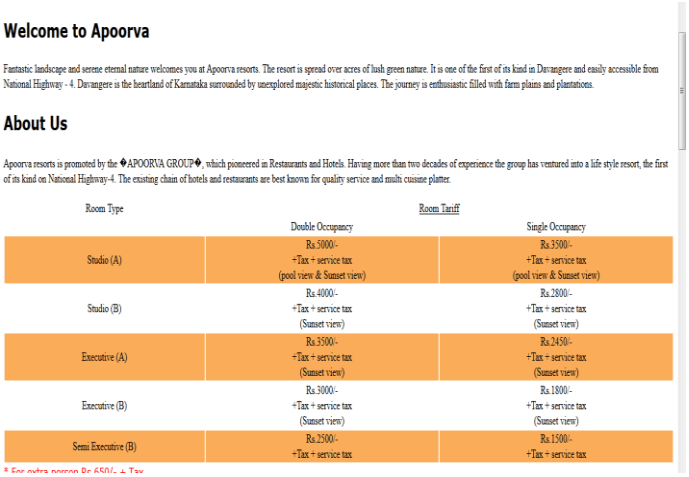
Fig-8-: Tweets Analysis

4. RESULTS

The following sections gives the discussions on experimental results.

4.1 web crawling

Web crawling is applied on the Website of Apoorva resorts Davangere. And the html page is crawled. It is shown in the figure 9.



Welcome to Apoorva

Fantastic landscape and serene eternal nature welcomes you at Apoorva resorts. The resort is spread over acres of lush green nature. It is one of the first of its kind in Davangere and easily accessible from National Highway - 4. Davangere is the heartland of Karnataka surrounded by unexplored majestic historical places. The journey is enthusiastic filled with farm plains and plantations.

About Us

Apoorva resorts is promoted by the ◆APOORVA GROUP◆, which pioneered in Restaurants and Hotels. Having more than two decades of experience the group has ventured into a life style resort, the first of its kind on National Highway-4. The existing chain of hotels and restaurants are best known for quality service and multi cuisine platter.

Room Type	Room Tariff	
	Double Occupancy	Single Occupancy
Studio (A)	Rs.5000/- +Tax + service tax (pool view & Sunset view)	Rs.3500/- +Tax + service tax (pool view & Sunset view)
Studio (B)	Rs.4000/- +Tax + service tax (Sunset view)	Rs.2800/- +Tax + service tax (Sunset view)
Executive (A)	Rs.3500/- +Tax + service tax (Sunset view)	Rs.2450/- +Tax + service tax (Sunset view)
Executive (B)	Rs.3000/- +Tax + service tax (Sunset view)	Rs.1800/- +Tax + service tax (Sunset view)
Semi Executive (B)	Rs.2500/- +Tax + service tax	Rs.1500/- +Tax + service tax

Fig-9: Crawled HTML Page

4.2 Extraction of All Images:

Images of Apoorva resorts are extracted from the crawled html page. It is shown in the figure 10.

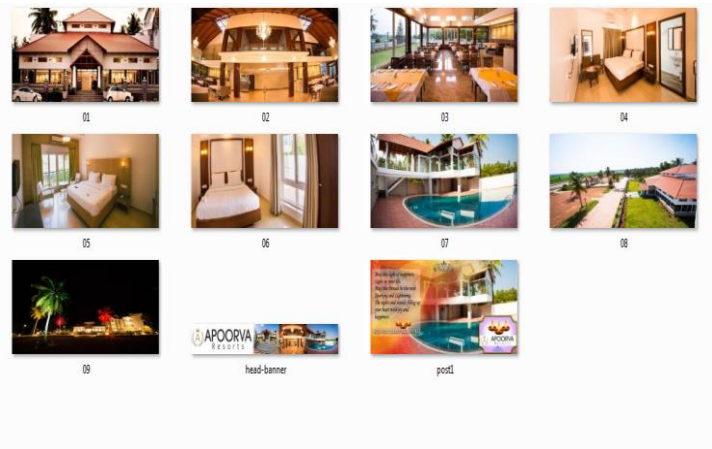


Fig-10: Image Extraction from the Crawled HTML Page

4.3 Tweets Analysis

Related to this the user posting the tweets about the business. The user tweets are may contain positive or negative tweets.After the posting of user tweets,then tweets are loaded to user Twitter account. After entering the particular item for analysis. Then it shows the total number

of tweets belongs to that particular category. This shows that the total number of tweets and also positive or negative tweets with accuracy about the business. All these procedures shown in the following figures.

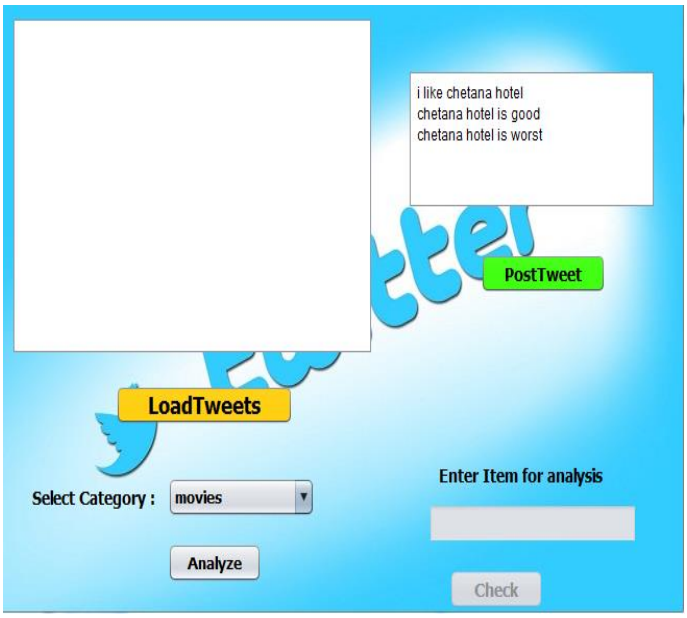


Fig-11: Posting of Tweets

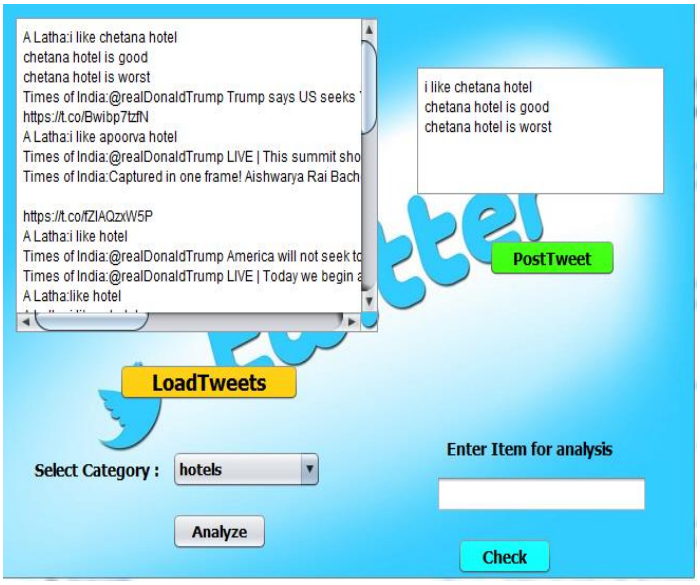


Fig-12: Loading of Tweets into user Twitter account

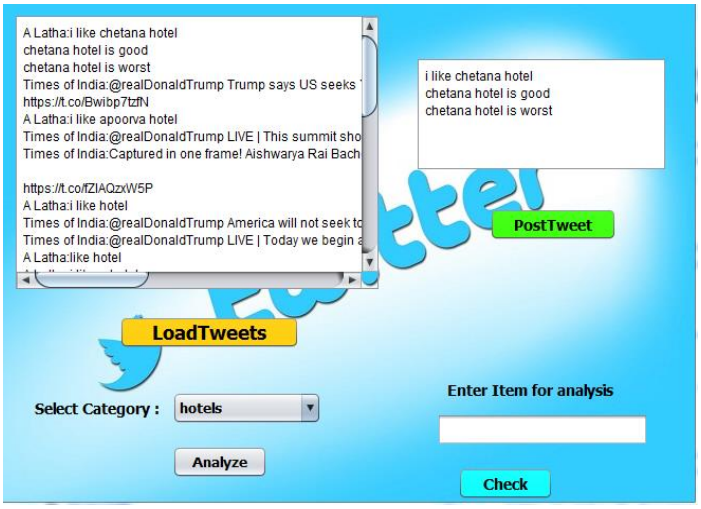


Fig-13: Selecting of particular Category

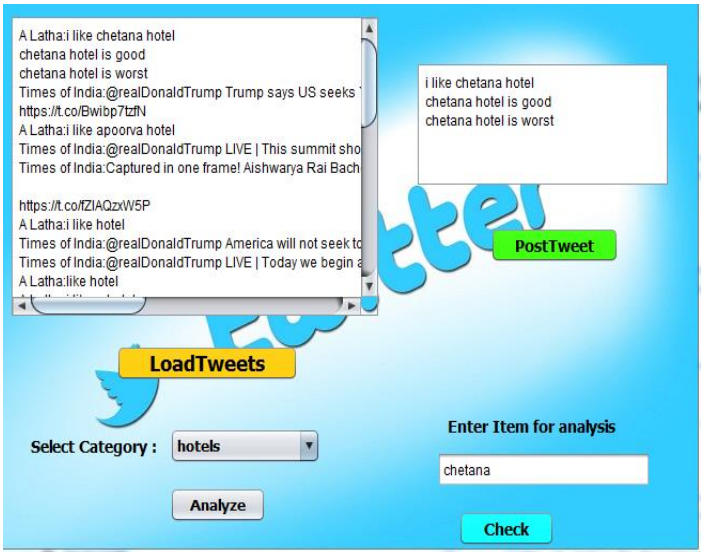


Fig-14: Analysis of Tweets of particular category

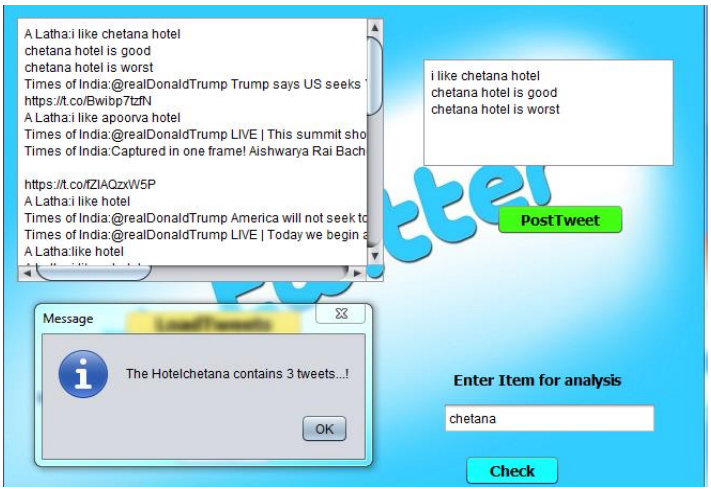


Fig-15: Analysis of posted Tweets

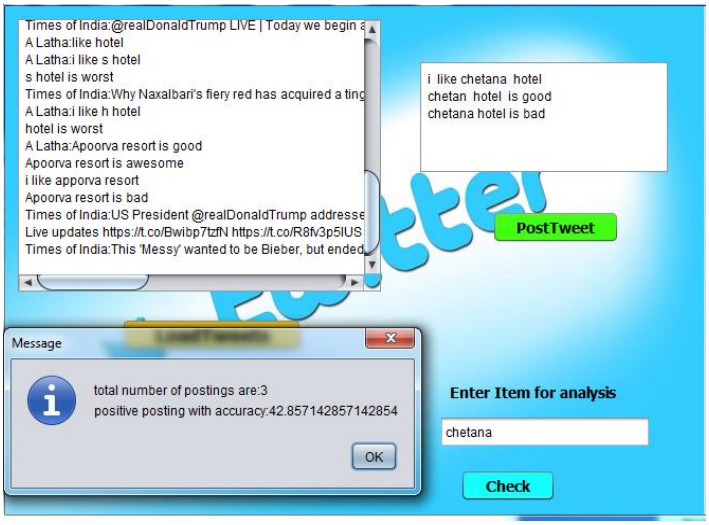


Fig-16: Analysis of posted Tweets with accuracy

5. CONCLUSION AND FUTURE WORK

The Knowledge base of small business is built by considering static and dynamic data. The static data includes business name, location, contact and website address of businesses. Web crawling applied on static data. And for the dynamic data, Twitter is considered. However, user comments on social media usually do not have ratings to indicate how much positive or negative their reactions are. So comments or reviews are analyzed. The user posted tweets into Twitter are analyzed to know how much positive or negative reactions about the businesses. Thus user opinions are useful to understand their preferences, and reputations of places or services they used.

For future work there is also the need to make sure the data are reliable. Since there always can be fake reviews from users who write reviews out of malicious purposes or are hired by businesses to write glowing reviews on social media. Sometimes information from users or websites can be out of date or not consistent with each other. These factors should be considered in the future as well to improve the reliability of knowledge discovery system.

5. REFERENCES

[1] Edd Dumbill, Forbes, Volume, Velocity, Variety: "What You Need to Know About Big Data", JAN 2012.

[2] Headd, Brian and Bruce Kirchhoff, "The growth, decline and survival of small businesses: Anexploratory study of life cycles," Journal of Small Business Management, pp. 531-550, October2009.

[3] Akerkar, R.A. and Sajja, P.S. 2009. Knowledge-based systems: Jones & Bartlett Publishers, Sudbury,MA, USA

[4] Ravindra Changala, D.Rajeswara Rao, T Janardhana Rao, P Kiran Kumar, Kareemunnisa ,2015 "Knowledge Discovery Process: The Next Step for Knowledge Search

[5] F.Morstatter, S. Kumar, H. Liu, and R. Maciejewski. Understanding Twitter Data with TweetXplorer. In Proceedings of the 2013 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013.

[6] Raheleh Makki,Axel j.Soto, Stephen Brooks 2016. Twitter message recommendation based on user interest profiles

[7] Elyasir, Ayoub Mohamed, and Kalaiarasi Sonai Muthu. "Focused Web Crawler." International Conference on Information and Knowledge Management (ICIKM 2012).

[8] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar."A Brief Overview on Data Mining Survey "2008 International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 1, Issue 3