Identifying Human Personalized Sentiment with Streaming Data

F. M. Tanvir Hossain
American International University-Bangladesh
Banani, Dhaka
Bangladesh

Maruf Ahmed
American International University-Bangladesh
Banani, Dhaka
Bangladesh

Anik Saha
American International University-Bangladesh
Banani, Dhaka
Bangladesh

Khandaker Tabin Hasan
American International University-Bangladesh
Banani, Dhaka
Bangladesh

ABSTRACT
Nowadays, social networks are becoming common platform of our emotion, sentiment, personality, and so on. A significant number of studies are also available about sentiment and emotion analysis from social network data. We observe that there are few studies are available those compute sentiment over real time data in Twitter and Foursquare. In this paper, we have conducted a research that can compute sentiment from real time data in a social network. We also use multiple techniques to compute sentiment such as sentiwordnet and textblob. We analyze the sentiments of a human from his/her twitter and from the location in foursquare of that person.

Keywords
Big Data, Sentiment Analysis, LBSN, Social Network ,Hadoop .

1. INTRODUCTION
Now-a-days online social networks have exploded in popularity and now rival the traditional Web in terms of usage and influencing our daily life in a huge amount. A few studies are available regarding sentiment analysis[1,9,11]. People spreading their ideas, daily routine, mental situation through social network in different ways. All these social networking mediums are interlinked. A number of studies have been conducted on values [17], personality [18], and preferences [19]. In this paper, we analyze the sentiments derived from the post/tweets that found in social networks.
Organizations use observations, opinion polls, and social media as a tackle to obtain feedback on their products and services. sentiment analysis is the computational study of opinions, sentiments, and emotions revealed in text. The use of sentiment analysis is becoming more widely leveraged because the information it yields can conclusion in the monetization of products and services. Numerous sites are dedicated to finding and maintaining contacts and to locating and sharing different types of information. For example, by obtaining consumer feedback on a marketing audit, an organization can measure the campaigns success or learn how to adjust it for greater success.

In current days, many systems recommend based on user's sentiment**.** Sentiment analysis focuses on what people think, and how they feel about something or somebody under certain circumstance. With the booming of social web, sentiment analysis brings us deeper understanding about online social network [6]. Micro-blogs such as Twitter provide huge amount of data which can be used to discover collective sentiment knowledge [7,8]. It refers to the branches of Computer Science like Natural Language Processing, Machine Learning, Text Mining and Information Theory and Coding to identify and extract subjective information in source materials. There are three levels in Sentiment analysis [9]. Document Level: Analyzing with whole document, we decided positive or negative sentiment. Aspect or Entity level: Fine-grained analysis are performed here. The aim of this level in Sentiment analysis is to find on entities or/and aspect those entities. For example my iPhone 6s has good picture quality but high price. So the sentiment on iPhone 6s picture quality is positive but price is negative. Sentence Level: this analysis is about find sentiment from a sentence which expressed a positive, negative or neutral sentiment. Basically this analysis related to subjectivity classification.

Product recompose is also helpful in building better products, which can have a direct impact on revenue, as well as comparing competitor offerings. First, we extracted data from tweeter and foursquare. Then, we analyzed the data to identify the human sentiment [1]. In this paper will describe the various types of sentiment classification, explore how to convert unstructured text into structured opinions, and address the current challenges in the field. So, we are taking the data from social network and analyzing the sentiment for each person.

In summary we have following contributions:

1) We analysis data in realtime from both Foursquare and Twitter.
2) We extract sentiment in different dimension i.e. season, gender, and weekdays.
3) We integrate hadoop in our realtime Sentiment analysis.
4) We recommend user dynamically based on their sentiment through Twitter.

2. RELATED WORKS
A plethora of works are available for recommendation Social recommendation system, Social network, Real time data analysis. In the following subsection, we described related works for each of the topics.

2.1 Social Recommendation System
A lots of research work has been done for building recommendation systems[2]. Which is mainly fall into three categories: memory-based approach, model-based approach and hybrid approach. Memory based approaches work with historical rating records to predict unknown rating without learning step , e.g. classical collaborating filtering methods. They focus on user-item rating matrix and attempt different

strategies to estimate missing ratings. Model-based approaches use the learned model from historical data to predict unknown ratings. They leverage statistics and machine learning techniques to learn models from data in order to predict the missing ratings. Hybrid approaches combine the two aforementioned approaches with certain fusion criterion.

2.2 Social Networks

Social Networks data is very effective for data analysis. There are many types of Social networking exists. Blogging, Online Journal, Photo Sharing, Virtual World, Video 3 Hosting, Micro-blogging, Learning Community, Music Sharing, Online Dating Sites, Online Gaming, Location Based Networking are the Social Networking systems widely used all over the world. Provided data for data analysis from Social Networks are not always reliable due to the privacy policy of the Networks. But still there are some social networks which allow data scientist or data analyzer, collect the desire data from the networks. There are many specialized social networks are being proposed for different purposes such as roles [20], location [8], and microblog [8].

2.2.1 Micro-blogging

Micro-blogging is the practice of posting small pieces of digital content which could be text, pictures, links, short videos, or other media on the Internet.

2.2.2 Location-Based service

Location-based service (LBS) is a software-level service that uses location data to control features. As such LBS is an information service and has a number of uses in social networking today as information, in entertainment or security, which is accessible with mobile device through the mobile network and which uses information on the geographical position of the mobile device. Micro-Blogging web site Twitter provides its streaming data through API. Twitter also provides its users data, only if the users set the access modifier as public. Location Based Social Networks like Foursquare also share its public data. These provided data is helpful for analysis many sectors in data science like Sentiment analysis, Business platform developing, Prediction of the future condition and possibilities etc.

2.3 Real Time Data Analysis

Real time data analysis is referred to the process of analyzing data at the moment it is produced or used. It is big data analysis is an iteractive process involving multiple tools and systems. Real time analysis is a form of big data analysis but rather focus on big data produced consumed, stored within a live environment. Such as analyzing mass amount of data as it is produced within stock exchanges, banks, branches and social networks throughout the globe. The scope of the analysis can be from multiple sources. It works by fetching importing big data stored within a system at run time and executes data big data analysis algorithms over it. Smith says that its helpful to divide the process into five phases: data distillation, model development, validation and deployment, real-time scoring, and model refresh. At each phase, the terms real time and big data are fluid in meaning. The definitions at each phase of the process are not carved into stone. Indeed, they are context dependent. Smiths five-phase process model is devised as a framework for predictive analysis. But it also works as a general framework for real-time big data analysis [3] [4]. There are some techniques where data diffuse automatically propagate information from one channel to another [21].

2.3.1 Data distillation

Like unrefined oil, data in the data layer is crude and messy. It lacks the structure required for building models or performing analysis. The data distillation phase includes extracting features for unstructured text, combining disparate data sources, filtering for populations of interest, selecting relevant features and outcomes for modeling, and exporting sets of distilled data to a local data mart.

2.3.2 Model development

Processes in this phase include feature selection, sampling and aggregation; variable transformation; model estimation; model refinement; and model benchmarking. The goal at this phase is creating a predictive model that is powerful, robust, comprehensible and implementable. The key requirements for data scientists at this phase are speed, flexibility, productivity, and reproducibility [5]. These requirements are critical in the context of big data: a data scientist will typically construct, refine and compare dozens of models in the search for a powerful and robust real-time algorithm.

2.3.3 Validation and deployment

The goal at this phase is testing the model to make sure that it works in the real world. The validation process involves re-extracting fresh data, running it against the model, and comparing results with outcomes run on data that's been withheld as a validation set. If the model works, it can be deployed into a production environment.

2.3.4 Real-time scoring

In real-time systems, scoring is triggered by actions at the decision layer (by consumers at a website or by an operational system through an API), and the actual communications are brokered by the integration layer. In the scoring phase, some real-time systems will use the same hardware that's used in the data layer, but they will not use the same data. At this phase of the process, the deployed scoring rules are divorced from the data in the data layer or data mart. Note also that at this phase, the limitations of Hadoop become apparent. Hadoop today is not particularly well suited for real-time scoring, although it can be used for near real-time applications such as populating large tables or pre-computing scores. Newer technologies such as Cloud eras Impala are designed to improve Hadoops real-time capabilities.

2.3.5 Model refresh

Data is always changing, so there needs to be a way to refresh the data and refresh the model built on the original data. The existing scripts or programs used to run the data and build the models can be re-used to refresh the models. Simple exploratory data analysis is also recommended, along with periodic (weekly, daily, or hourly) model refreshes. The refresh process, as well as validation and deployment, can be automated using web-based services such as RevoDeployR, a part of the Revolution R Enterprise solution.

3. DATA COLLECTION

We collect a total of data 205 Foursquare users' by observing the venues and these users are found in Twitter also. Then we collect a total of 5,74,000 tweets. We maximum, minimum and average tweets are 3220, 1375 and 2805 respectively. We also find that these users have in the range of 25 to 45 years. With different demo graphics and profession. We describe in details the process of data collection, tools and technologies in the methodology section.

4. METHODOLOGY

We build our methodology to identify human sentiments based on textual data. We consider twitter and foursquare as our data source. Regarding LSBN we collected data through specific users' id. Based on some parameter like time, gender, we separated the acquired data and tried to analyze human personalized sentiment.

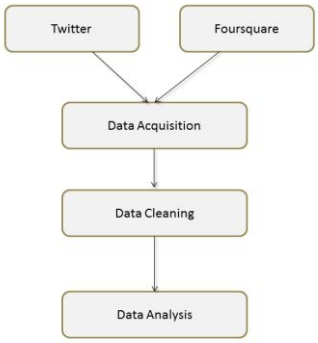

Fig. 1. Methodology of identifying human sentiment from streaming data

4.1 Data Acquisition

For Data Analysis, the first requirement is data. Data can be collected from various sources. But when this data analysis is related to the human behavior, the first data sources pop-up are Social Networks. Micro-blogging, virtual world and location base social networks are interacting with each other at the API level. Like Twitter, Facebook and Foursquare, each holding each other's data in their own way. Foursquare can be connected with Twitter and Facebook, therefore, user can tweet or post their data from foursquare. Here we extracted data from Twitter and Foursquare. We collected data of venues from foursquare, and found out the users listed into the venues. Then we found out the users twitter id from the users information of the foursquare.

4.1.1 Extracting data from Foursquare:

For using foursquare API, one has to have an App created in the Foursquare. Normal user can't create apps in foursquare. Foursquare provide this facility to a developer. After having the developer facility, one can go to the My App menu to create an app. For creating app in FourSquare, app name, redirect URL are required. By providing these, App can be created.

After creating the app, Client ID and Client Secret will be provided by the foursquare. Redirect URL is used in API request as a parameter. Foursquare sends its response to that redirected URL. For accessing the app one has to have access token provided by the foursquare. In here, we use Oauth2 authentication URL for getting the Access Token [13]. After sending the request for access token, Foursquare Oauth2 URL redirects the access token as a GET parameter to the redirected URL. Working with Foursquare API is relatively easy in Python. A Python library was built by Mike Lewis wrapping foursquare API, which made implementation much easier [14]. Using package manager pip, foursquare library can be installed from the gitHub repository foursquare.

```
## Construct the client object ##
client=foursquare . Foursquare (

        clientid ='YOUR CLIENT ID' ,

        clientsecret ='YOUR CLIENT SECRET' )
## Apply the returned access token to the client ##
client.set_access_token (access token )
## Access Token is for authentication ##
 client.set _access_token (access_token)
```

By importing foursquare library, *Foursquare()* function can be implement. This function takes Client ID and Client Secret as parameter. By implementing this function, Client instance can be created. This client instance deal with the responded data from the foursquare. Access Token have to set to the client instance for handling the access authentication. From this client instance another user or venues JSON object can be created.

```
## Self information of the user ##
 ## who created the app ##
user = client.users ( )
## user object is created here ##
 ## user id is given here as parameter ##
user = client.users ( userName )
## venue object is created here ##
 ## venue id is given here as parameter ##r
venues = client.venues ( venueID )
```

user() and *venues()* methods returns the JSON formatted data of user and venues. If the parameter of the user method is empty, then its returns the information of the owners, who create the app using Foursquare account. And if the user id is given in the *user()* method as a parameter then it returns the information of the users. Same goes for venues, if the venues id is given in the *venues()* method then it return information of the venues. 19 User and venues JOSN Object have created from the Client instance. User JSON contains the Contacts of the user only if the public property of the user account is set true. This contact instance contains the users Social Networks ID which are synchronized with the users Foursquare account. Venues JSON contains users ID, who checked in and liked the venues, left some tips, shared photos of the venues, tagged venues and listed by the venues. From the URL of Foursquare venues, Venues ID can be collected. We collected venues id form url like this (https://foursquare.com/v/venue-name/VENUEID) and put those id into the *venues()* method. Then we collect users id from venues JSON and put those users id into the *user()* method for getting the information about user and collecting Twitter id of the user (Those who connected Twitter account with Foursquare account).

4.1.2 Extracting data from Twitter:

Extracting data from Twitter: Same as Foursquare API, one has to create an app in Twitter for working with twitter API and extracting data from Twitter. From Twitter App site (https://apps.twitter.com) app can be created. For creating app, application details like application name, description, website fields must be provided and developer agreement must be checked by the user. After creating the app, twitter will provide four sets of KEYS. These keys are available in Keys and Access Tokens menu in app details of the twitter app site. These keys are Consumer Key (API Key), Consumer Secret (API Secret), Access Token, Access Token Secret. For working with Twitter API, a python library called *tweepy* is

available in gitHub. *Tweepy* was developed by Joshua Roesslein and Aaron Hill which is supported by twitter developer organization. This library was built using Twitter API wrapper in python [15]. Using package manager pip, *tweepy* library can be installed from the gitHub repository *tweepy*. 20 By importing Tweepy library, *tweepy* instance can be created. Using *tweepy* instance, *OAuthHandler* instance could be created. Into this *OAuthHandler* instance consumer token and consumer secrete have to be passed. If anyone has a web application and is using a callback URL, that needs to be supplied into the *OAuthHandler* instance.

```
## authorize twitter, initialize tweepy ##
auth=tweepy.OAuthHandler ( consumer token , consumer
secret )
## setting access tokens into the auth instance ##
auth.set_access token ( access_key , access _secret )
## creating API instance ##
api = tweepy. API ( auth )
```

Access Token and Access Token Secret have to set into the *OAuthHandler* instance by calling the *setaccesstoken()* method. This *setaccesstoken()* method takes access token and access token secret keys as parameters. After that, an API instance has to be created by supplying the *OAuthHandler* instance. This API instance has four Timeline methods, which are *hometimeline()*, *statuseslookup()*, *usertimeline()* and *retweetsofme()*. In here, we use *usertimeline()* method, which takes screen name, count and max id as parameters. Screen name is the user id of a twitter user, Count specifies the number of statuses to retrieve and Max ID returns only statuses with an ID greater than or equal to the specified ID.

```
## make initial request for most recent tweets ##
##  (200 is the maximum allowed count ) ##
Api.user timeline( screen_name= USER ID , count =200)

##  all subsequent requests use the ##
## maxidparam to prevent duplicates ##

 api.usertimeline( screenname = screenname , count
=200,maxid=oldest)
```

This user timeline() method returns the list of statuses posted from the authenticated user or the user specified by USER ID. Executing this process in a loop and passing the oldest post id, every tweets posted by the specified user can be retrieved. This retrieved data from the twitter can be stored in a file. We stored these tweets into a comma separated values (.csv) formatted file for farther analysis. Twitter user id which was collected from Foursquare is used in twitter data extracting portion. For getting the tweets of those users, who use twitter for micro-blogging and foursquare for Location Based Social networking we use these processes.

4.2 Data Cleaning

Foursquare contain detail information of its users. That's why it's possible to find out the gender of an users. We get venues and users data form foursquare using the Foursquare API. Now a day, API provides data in JSON format. But JOSN is Object formatted data. For using the data, we have to convert this JSON data into CSV or normal Text Format in a meaningful way. In here we convert these JSON data in CSV format. Where the required fields are arranged in column. After acquiring user data we separate the data by gender group and stored it in text format. We only save the twitter id information and gender information in these text files. After collecting the users twitter is and gender , we separate the whole data into two parts depending on gender. Using this twitter ids, we start collecting data from twitter. Twitter API also provides data in JSON format. We format these twitter data using Python library Pandas. Pandas library can manipulate data in desire form.

4.3 Data Analysis

For analyzing the data, there are many methodologies exists. In here we try to analyze the data based on sentiment of human personalized emotion. Take one user tweet and using text blob and find the sentiment polarities of the tweet. We see it as user sentiment level for the tweet creation time.

4.4 Algorithm

The processes we design for accomplish aforementioned tasks are described in this algorithm section.

a) Dynamic Sentiment Analyzer :

Input :- V_f

Output:- S_p , $Array_{ut\_sentiment\_po}$

BEGIN

Initialize $Array_{ut\_sentiment\_po}$ as Empty

Initialize Client Object from Foursquare Library.

For each venue ID V_f do the followings.
- Getting venue array.
- Find users U_f
- $Array_{uf} = U_f$

For each U_f do the followings
- For each $Array_{uf}$ => U_f
- Find User Twitter ID U_t

Getting All tweets of each U_t .
Initialize Api Object from Tweepy library.
- $Array_{ut\_tweets}$ = Api.UserTimeline(U_t)

For Sentiment Polarity Check
Initialize TextBlob Object from TextBlob library.

- For each $Array_{ut\_tweets}$ => Tweets
- S_p = TextBlob(Tweets)
- End For each $Array_{ut\_tweets}$

- $U_{t\_me\_setiment\_polarity}$ = Mean(S_p)
- Average $U_{t\_av\_sentiment\_polarity}$ =
- $U_{t\_me\_setiment\_polarity}$ / Tweets

- $Array_{ut\_sentiment\_po}$ = $U_{t\_av\_sentiment\_polarity}$
- End For each $Array_{uf}$

END

5. EXPERIMENT

After collecting and preprocessing our dataset, we conduct experiment in this section. First, we analysis users' sentiment based on season and then we also compute the sentiment tables for weekday basis over the same data for each of the aforementioned experiment, we determine result for male and female persons.

5.1 Season based sentiment analysis

Sentiment analysis level graph according to season and everyday in week are shown

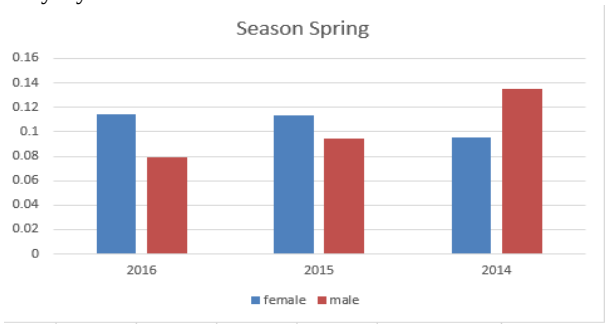

Fig. 2. Spring Season

The sentiment level in spring for female almost same and it's close to 0.1 to 0.12 but in the case of male it's decreasing, in the year 2014 sentiment level was 0.13 in 2015 0.09 and the year 2016 it's 0.08.

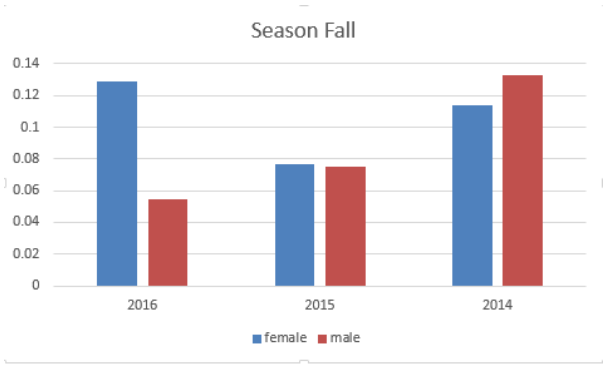

Fig. 3. Fall Season

In fall 2016 the sentiment level for female is too high & it's close to 0.13 where male sentiment level is .05 but in 2015 both level are same 0.08 . In the previous year (2014) male sentiment close to 0.13 & female 0.11 .

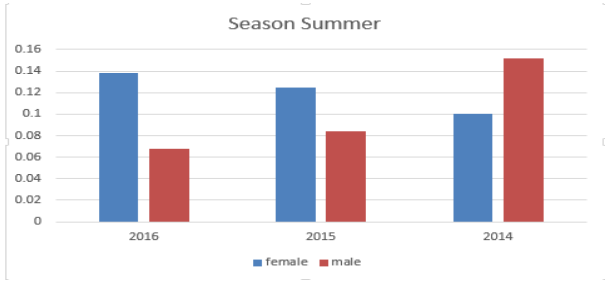

Fig. 4. Summer Season

Summer 2014 male sentiment still high and it's close to 0.15 but as like before it's decreasing in the year 2016 its close to 0.06. but female sentiment high in 2016 summer its 0.14 and better than 2014 when this sentiment level approximate 0.1 .

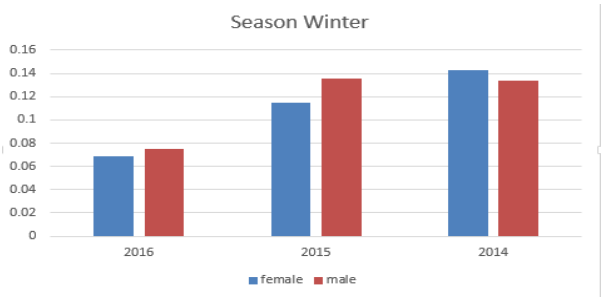

Fig. 5. Winter Season

Basically in winter sentiment level both male and female decreasing , but for male in 2014 is 0.13 2015 .13 and 2016 its close to 0.06 same as female where their sentiment level in 2016 is 0.07 but which was 0.12 in the year 2014.

5.2 Weekday based sentiment analysis

Here is a point to notify that sentiment level for male in any session still decreasing from 2015 to 2016. Cause of Unemployment problem which is average in 2015 but in 2016 it's cross the margin and it's directly affected male sentiment level.

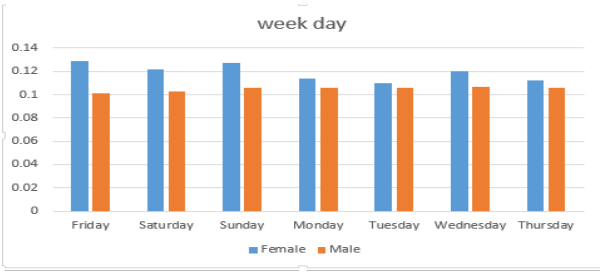

Fig. 6. Week Day

5.3 Sentiment Stability between male and female

Here is a point to notify that sentiment level for male in any session still decreasing from 2015 to 2016. Cause of Unemployment problem which is average in 2015 but in 2016 it's cross the margin and it's directly affected male sentiment level. From the Week Days data, it can be estimate that male sentiment is more stable than the female sentiment.

6. DISCUSSION

In this work, we use Hadoop to manage dynamic tweets for our selected users.

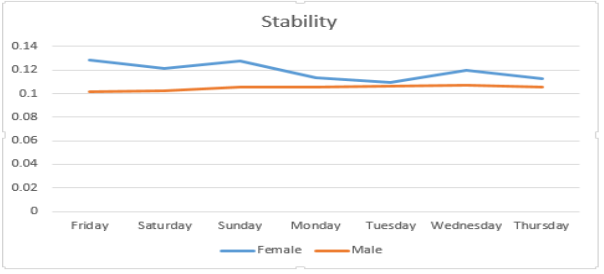

Fig. 7. Stability

Then we analyze users' sentiments based different parameters such as seasonal and weekday basis. We also show our results with sentiments in different scenario. Previously female users use social networks whereas in current times they use social network frequently. When we analyze users sentiment in

2016, we see that female users sentiments are more analyzable.

We also observe that female users share more personal and intimate information in social network, that is why we can analyze their sentiment based on weekdays also. However, we can develop a recommender system based on based on users' sentiment in real time.

7. CONCLUSION AND FUTURE WORKS

Sentiment analysis is a field of study that analyzes peoples sentiments, attitudes, or emotions in a certain entities. Sentiment analysis is an explode field with a variety of use applications. Although sentiment analysis tasks are challenging due to their natural language processing origins, a few progress has been made over the last few years due to the high demand for it. The growing need for product insights and the technical challenges currently facing the field will keep sentiment analysis and opinion mining relevant for the foreseeable future. Next-generation opinion mining systems need a deeper bind between complete knowledge bases with reasoning methods inspired by human thinking and psychology. This will lead to a better understanding of natural language opinions and will more efficiently bridge the gap between unstructured information in the form of human thoughts and structured data that can be analyzed and processed by a machine.

8. REFERENCES

[1] A. Katrekar, "An introduction to sentiment analysis," GlobalLogic Inc., June 2005.

[2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions.," IEEE Trans. on Knowl. and Data Eng., 17(6):734749,, June 2005.

[3] P. W. Zhigao Zheng, J. Liu3, and S. Sun, "real-time big data processing framework: Challenges and solutions, applied mathematics & information sciences an international journal," An International Journal, 2015.

[4] M. Barlow, "Real-time big data analytics: Emerging architecture," 1st ED. LONDON: OReilly Media., 2013.

[5] S. Loria, "Textblob python library or sentiment analysis," sloria/TextBlob on GitHub at commit eb08c12"Twitter via sms faq," April 13, 2012.

[6] Springer-Verlag, "A. bifet and e. frank. sentiment knowledge discovery in twitter streaming data.," In DS10, pages 115, Berlin, Heidelberg,, 2010.

[7] H. M. J. Bollen and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.," In ICWSM 11, 2011.

[8] B. Liu, "Sentiment analysis and opinion mining, morgan and claypool publishers," Morgan and Claypool Publishers, 2012.

[9] B. R. D. K. A. M. Michael Wiegand, Alexandra Balahur, "Asurvey on the role of negation in sentiment analysis.," Proceedings of the workshop on negation speculation in natural language processing 6068, Association for Computational Linguistics., 2010.

[10] T.-C. Peng and C.-C. Shih, "An unsupervised snippet-based sentiment classifi- cation method for chinese unknown phrases without using reference word pairs," IEEE/WIC/ACM International Conference on Web Intelligence and intelligent Agent Technology journal of computer, volume 2, issue 8, august 2010, issn 2151-9617 ., 2010.

[11] D. Hardt, "The oauth 2.0 authorization framework,," Internet Engineering Task Force (IETF), Request for Comments: 6749, Obsoletes: 5849 ,Category: Standards Track, ISSN: 2070-1721,, 2012.

[12] "Froursquare library for python, developed for manipulating and retrieving froursquare data," https://github.com/mLewisLogic/foursquare.

[13] "Tweepy library for python, developed for manipulating and retrieving twitter data,," http://www.tweepy.org/

[14] "Hadoop, powered by hadoop,"

[15] "What is hadoop," 2016.

[16] "Hadoop vs traditional database management system," 08-aug-2016.

[17] Mukta, Md Saddam Hossain, Mohammed Eunus Ali, and Jalal Mahmud. "User Generated vs. Supported Contents: Which One Can Better Predict Basic Human Values?." International Conference on Social Informatics. Springer International Publishing, 2016.

[18] Mukta, Md Saddam Hossain, Mohammed Eunus Ali, and Jalal Mahmud. "Identifying and validating personality traits-based homophilies for an egocentric network." Social Network Analysis and Mining 6.1 (2016): 74.

[19] Rahman, Md Mahabur, et al. "Can we predict eat-out preference of a person from tweets?." *Proceedings of the 8th ACM Conference on Web Science*. ACM, 2016.

[20] Giunchiglia, Fausto, et al. "Semantic enabled role based social network." International Journal of Intelligent Systems and Applications 4.12 (2012): 1.

[21] Hasan, Khandaker Tabin, et al. "Event-based content management by spontaneous metadata generation and diffusion." Computational Intelligence and Informatics (CINTI), 2012 IEEE 13th International Symposium on. IEEE, 2012.