Find like-minded user using Big Data Mining Technique: A Case Study on Twitter

Poonam Yadav[1]*, Kavita Khanna[2] & Jyoti Sahni[3]

[1]Research Scholar, CSE, The Northcap University, Gurugram, India

*Corresponding Author E-mail: *poonamyadav@ncuindia.edu*

[2]Associate Professor, CSE, The Northcap University, Gurugram, India

E-mail: *kavitakhanna@ncuindia.edu*

[3]Assistant Professor, CSE, The Northcap University, Gurugram, India

E-mail: *jyotisahni@ncuindia.edu*

Abstract: Twitter is a Social networking service, where users post and interact via tweet. Twitter becomes vehicles for businesses, organizations, and public figures to reach a broader audience. Twitter also allows people to connect with other person based on mutually shared interests. Connecting people is a very easy thing to do because there are so many platforms there to help e.g. Facebook, Instagram, twitter etc. But connecting like-minded people is not easy. Social network platforms cannot match you with people since they don't know much about yourself. For this first we find out the social interest of a person like politics, sports, news, movies, festivals, etc. by analyzing the profile of a person on twitter, which may belong to same social groups or may be random at all, on different factor like age group or the time period of a month the tweet may be focused on only one topic or may contain many topics, e.g tweet may be focused on the admission procedure in any school\ college\ university or may be consisting of various topics like recently released movie gossip, planning of a tour, any natural calamity, political talk, any ongoing tournament like Indian Premier League, Wimbledon Cup or may be the Olympics, doctor's appointment, which differs from person to person. This project categorized these people into the specific social group and then suggest or connect with like-minded people.

Keywords: Big data, Data Mining, Twitter, Sentiment analysis, Like-Minded User, Communities' Discovery, Text Mining

I. INTRODUCTION

Big data is a collection of the huge volume of structured and unstructured data sets that are so large and complex that it is difficult to process and analyse using traditional data processing applications. Nowadays, most of the task has been shifted to internet-based solutions rather being manual. Whether it is bill payments, food ordering, tax payment, online shopping, ticket booking or entertainment or social website. The data generated using these services provided on the internet is increasing exponentially. Hence, many times this data becomes difficult to manage.

Any data to be stated as Big Data has to satisfy already identified 7 V's, which are Variety (tweet contain plain text and media items), Velocity (500 million tweets per day), Value (big data, bigger value), Veracity, Volume (the volume of tweets is growing at around 30% per year), Variability (Constantly Changing), and newly proposed V, which is Volatility (liable to change unpredictably), all of which are present in this dataset. We understood what big data is and their dimensions, now look at the source of data and their usage.

1. Sensors produce data in the Internet of Things (IoT) to monitor activities of a smart device.

2. Movement of stars, satellites are analyzed in Astronomy to monitor the activities of asteroid bodies.

3. Video Surveillance or Recordings from CCTV cameras to analyze behavioral patterns for security and service improvement.

4. Medical records in Healthcare to aid in short-term health monitoring and long-term research programs.

5. Blog posts, tweets, social networking sites on Social Media to analyze the customer behavior pattern.

Many online activities like social networks, news reports, forums, online marketing, blogs publish content, which is used to understanding the opinions (positive, negative, neutral) of the general user. Multiple data analyses technique of web mining, data mining and text mining is used for data analysis. It plays the very important role in decision making in an organization. We can analyze social events like political movements and marketing campaigns.

This paper is organized in 4 Section We begin this paper with an Introduction to Big data and the importance of extracting useful information. Section 2 highlights related work on social media data mining on twitter. Section 3 outlines the methodology adopted and Algorithm formulation is also included in it. Finally, Conclusion and Future Work are presented in section 4.

II. RELATED WORK

Mining Social Media Data for Understanding Students' Learning Experiences [2]. To understand the issues and problems in educational experiences of engineering students, 35,000 tweets streamed at the location of Purdue University and use multi-label classification algorithm to classify tweets. They used the algorithm to know their educational experiences, feelings, current issues, opinions, interest, and concerns about the learning process.

Social media using data mining techniques: a survey of big data. They state that it's very difficult to modify and structure the data that the internet user generating daily on liking, poking, tweeting, chatting on social media via traditional databases. This paper deals with all these 5Vs of big data [3].

Improve user RTSE experience on the web through fast retrieval of social media content: They proposed the framework to enhance the system efficiency. Web 2.0 applications producing real-time content continuously i.e Twitter and Facebook, at a very rapid pace [4].

Topic modeling for social media content: In this paper, they explore an unsupervised topic modeling approach. That use LDA algorithm to find the topics in social media content [5].

A case study on mining social media data. They showcase how social media data can be utilized. They analyze social media comments to discover the inter-relationships among various factors and proposed a structured approach [6].

Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections: In this paper they provide a sentiment analysis of the Twitter discussion on the 2016 Austrian presidential elections. They analyzed the data-set consisting of 343645 Twitter messages related to the 2016 Austrian presidential elections [7].

Social media Mining -a brief introduction: social media has generated unmatched amounts of social data and provides the easily accessible platform for users to share information. To take out actionable patterns that can be beneficial for business, users, and consumers from that big social data novel challenges arise because social media data are noisy, unstructured, and dynamic in nature.

Twitter data analysis for studying communities of practice in the media industry: This article suggests a novel mixed-methods approach based on data to measure the role of Twitter for physical communities of practice [8].

Twitter based model for emotional state classification: In this paper, they proposed a model to classify or categorize an individual's current emotional state into eight predefined states. Then justify their prescribed approach, they also compare the results and accuracy of KNN, Decision Tree and Naive Bayes algorithm [9].

III. PROPOSED ALGORITHM

The process to take out a pattern from big data can be broken down into two main sub-processes: Data management and Data analytics as shown in Fig. 1. [1] these sub-processes further broken down into five stages:

We are presenting a methodology that shows how Informal social media data like twitter can relate two people by their common interest on the bases of words they used in the tweet. These are the various steps that are involved in the process of data mining from data collection to decision making.

Data management involves processes and supporting technologies to acquire and store data that prepare (clean) and retrieve it for analysis.

Data analytics techniques used to analyse and acquire intelligence from the selected data that helps to make the decision.

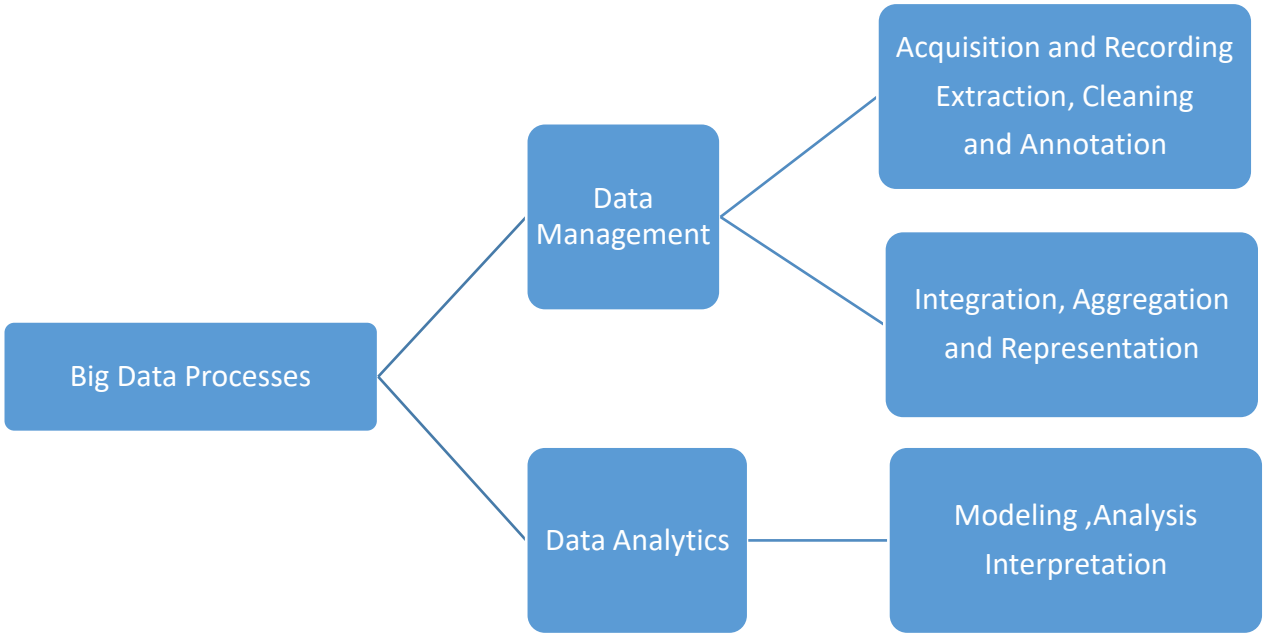

Fig. 1 Process of extracting Patterns from Big Data [1]

Data Integration: we collected and integrated the data from different sources like sensors, electronic device, news report, electronic media and images, blog posts, tweets, social networking sites, log details etc. In this project we stored twitter_data.txt is in JSON (JavaScript Object Notation) format.

Data Selection: all the data we have collected in the previous step is not required for analysis, so in data selection, we select only those data which we think useful for data mining project e.g tweet of different people for a defined date, time or different geographical areas.

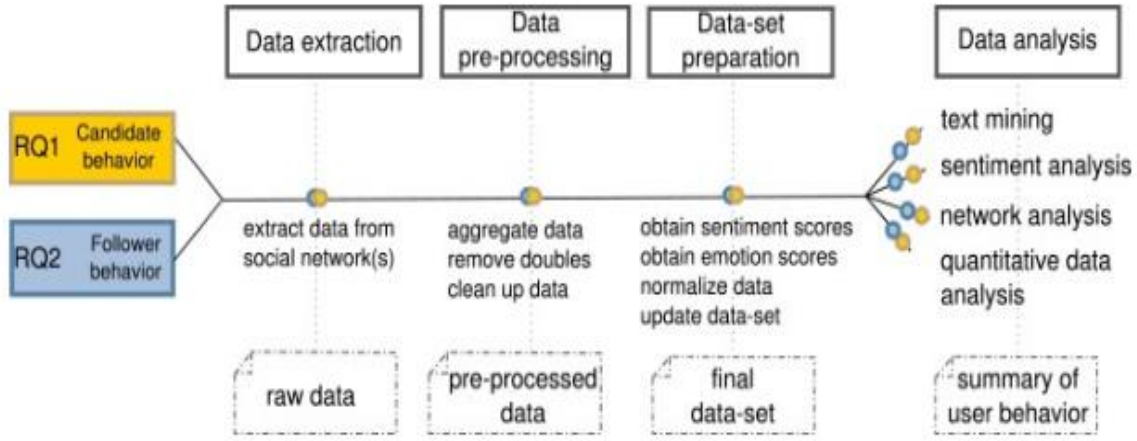

Fig-2. Approach overview: Sentiment analysis of the twitter data.

We select tweet of one week in the month of January.

Data Cleaning: The data we have collected is not clean and may contain errors, missing values. when we select one-week twitter data, we get the tweet of different language and inconsistent.So we need to apply different techniques to get rid of such anomalies. We are providing ways to clean data using Python.

a. The commonly occurring words (stop-words) should be removed. NLTK (Natural Language Toolkit) has a list of stop words stored in 16 different languages.

```
Import nltk
From nltk.corpus import stopwords
Set(stopwords.words('english'))
```

b. Punctuations like ".", ",","?" are important punctuations that should be retained while others need to be removed.

```
for char in my_str:
  if char not in punctuations:
     no_punct = no_punct + char
```

c. Removal of Expressions: human expressions like [laughing], [Crying], [Audience paused]. These expressions are non-relevant to content and hence need to be removed

d. Split Attached Words: words like WinterVacation, TheNorthCapUniversity etc. By using simple rules and regex method split into their normal forms.

e. Slangs lookup: like helo to hello. Apostrophe look up is used to convert slangs to standard words.

f. Removal of hyperlinks and URLs.

g. Spelling correction: algorithm like Levenshtein Distances, Dictionary Lookup etc is used for spelling correction.

Data Transformation: After cleaning the data we need to transform them into forms that are suitable for mining.

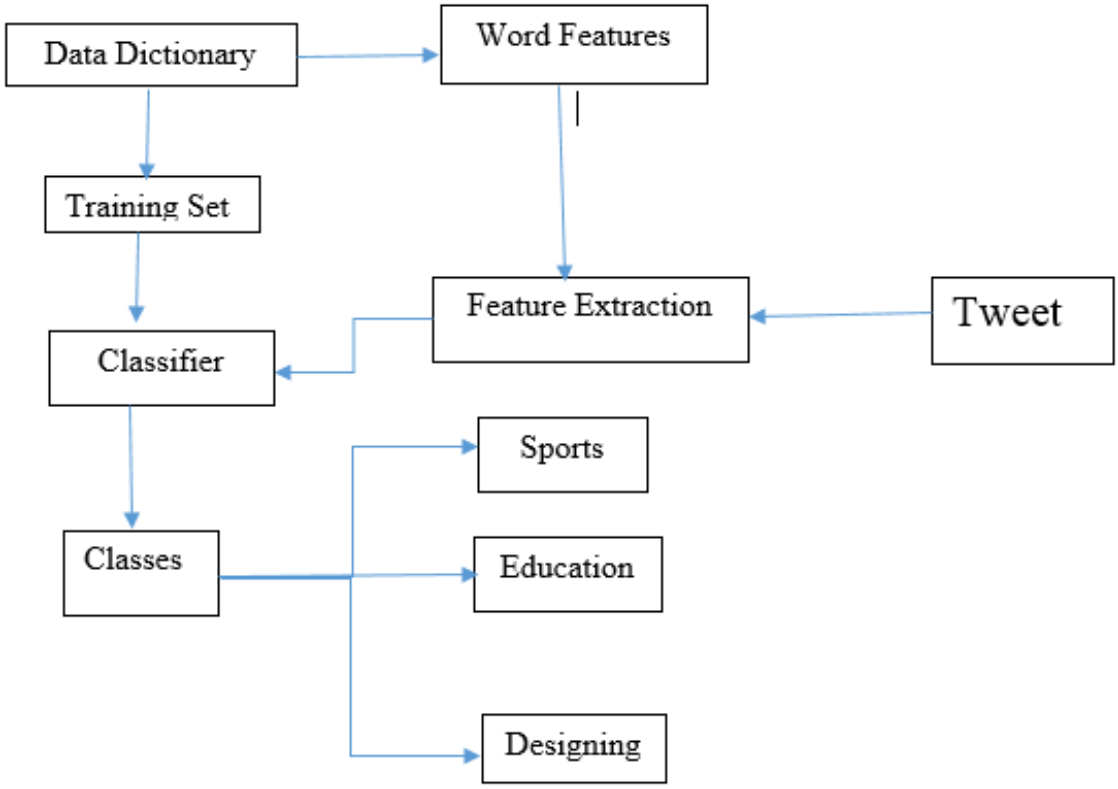

Fig. 3 Data Analysis Architecture

Data Mining: we are ready to apply data mining techniques directly on the data to discover the interesting patterns, techniques like clustering and association analysis is used in it.

Pattern Evaluation and Knowledge Presentation: This step involves visualization, a transformation of data and it also removing redundant unnecessary patterns.

The flow chart describing an overall system design of data analysis technique we applied on twitter data is shown in Fig 3. The user posts their tweet on twitter, these tweets are extracted from unstructured data. To extract the feature from unstructured dataset we convert it in the structured form. After selecting the feature of a word, the classification technique applied to them to group them into different classes. The classes are pre-defined like education, medical, sports etc, we can also divide these classes into sub classes' like sports into cricket, football, tennis.

After classification, we applied Sentiment analysis that is the important part of data analysis, a process that automates mining of attitudes, opinions, and views through Natural Language Processing (NLP). It converts text into the group like "positive" or "negative" or "neutral".

An example for Sentiment Analysis is as given below,

<SENTENCE> = XYZ school provide excellent teacher with great infrastructure.

<OPINION HOLDER> =<parent>

<OBJECT> = <school>

<FEATURE> = <teacher>< infrastructure>

<OPINION >= <excellent><great>

<POLARITY> = <positive>

In this project first, we get the interest area of twitter user, put that user in specific class and then connect the like-minded people based on their class.

Decisions / Use of Discovered Knowledge: we use the knowledge acquired to take better decisions. We can also use this knowledge to find the potential customer, social issues, current trends etc.

IV. CONCLUSION

Tweets convey opinions about the person, product and situation but tweets are unstructured. To obtain the overall understanding of these unstructured data can be very time-consuming. So, we divide these tweet in the different category and apply data analysis on the class.

The opinions on twitter are seen by the different users and thus creating an image about the person, products or services. To provide better picture we categorized the related tweets that generate fair judgment. The proposed system would be easy for a user to connect with the people who have the same opinion or political views as you. It is also used in the decision-making process in their daily life activity. Next step that we would be taking is to connect the likeminded people that help the person to disconnect from work and meet people with similar interests.

Conflict of interest: The authors declare that they have no conflict of interest.

Ethical statement: The authors declare that they have followed ethical responsibilities.

REFERENCES

[1] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.

[2] Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. "Mining social media data for understanding students' learning experiences." IEEE Transactions on Learning Technologies 7.3 (2014): 246-259

[3] Gole, Sheela, and Bharat Tidke. "A survey of big data in social media using data mining techniques." Advanced Computing and Communication Systems, 2015 International Conference on. IEEE, 2015.

[4] Quazilbash, Naveen Zehra, Syed Mazhar Hasan Qadri, and Shakeel Khoja. "Improved user RTSE experience on the web through fast retrieval of social media content." Multitopic Conference (INMIC), 2012 15th International. IEEE, 2012.

[5] Chan, Hing Kai, Ewelina Lacka, Rachel WY Yee, and Ming K. Lim. "A case study on mining social media data." In Industrial Engineering and Engineering Management (IEEM), 2014 IEEE International Conference on, pp. 593-596. IEEE, 2014.

[6] Gundecha, Pritam, and Huan Liu. "Mining social media: a brief introduction." New Directions in Informatics, Optimization, Logistics, and Production. Informs, 2012. 1-17.

[7] Kušen, Ema, and Mark Strembeck. "Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections." Onli. Soc. Networks and Media (2018): 37-50.

[8] Marlen Komorowski,Tien Do Huub, Nikos Deligiannisb," Twitter data analysis for studying communities of practice in the media industry"Telematics and informatics 35 (2018)195-212.

[9] Ahuja, Ravinder, Rohan Gupta, Saurabh Sharma, Ayush Govil, and Karthik Venkataraman. "Twitter based model for emotional state classification." In Signal Processing, Computing and Control (ISPCC), 2017 4th International Conference on, pp. 494-498. IEEE, 2017.

This volume is dedicated to Late Sh. Ram Singh Phanden, father of Dr. Rakesh Kumar Phanden.

