Privacy and Security in Online Social Media-a Big Challenge

Nripendra Narayan Das, Govind Singh Patel

Abstract ***Social media generate huge amount of unstructured data using multiple platforms. This study consists of various privacy and security concerns on social media, collect data from online social media analyze the data and visualize. The primary focus is establishes different aspects of security and privacy on social media. A novel framework has been developed for collecting data from online social network like Twitter, raising this content and visualizing this data in terms to understand whether the followers that genuine or fake. Extensive research has carried out to identify fake followers in social network by profile linking. A social network analysis system has been formed using text analysis approaches to extract meaningful inferences. Profile linking may be necessary for organizations. The experimental results showed that better profile linking is possible with past history of user handles.***

Keywords: ***Big Data, NLTK, Social media, Profile linking***

I. INTRODUCTION

Social media is of different types . different types of contents are getting generated on our social media . One popular type of social media networks is Facebook , twitter LinkedIn, So there are different ways in which social media content is getting generated . There is also social games . Virtual games and these are different types of content that are getting generated through these different types of social media services that are available . Some example of popular social networks of different categories is : YouTube , Facebook , twitter , google+ , four square linking Instagram , tumbler , Pinterest , tinder , whisper etc . Different categories mean different types of contents are getting generated in these networks , for example YouTube is one of the most popular video sharing service , four square is mostly the location based service , linked in which is professional based service . Facebook - It combination of many different types of content , Instagram which is for images , Twitter is for microblog of short content and also the combination of different types of content . So these sets of social networks that are available they are actually creating content with the networks of a particular category . based on the professional connections the we would like to have or we have . LinkedIn , Facebook , four square these are more traditional type of social networks . Whereas there has been other type of social networks that is getting more popular with Pinterest , vine , tumbler , whisper , snapchat these category of social networks can be categorised into things like social networks which are where the content is getting posted and it destroy by itself after some time and there is also social networks like whisper where the content that you post . It also non voice and who is posting the content is actually difficult to find . There are about 215 or 220 popular social networks services that are available now . Online social media is increasing or spreading at rapid rate , If you look at in 2013 , Facebook had only 2.5 million posts in 60 seconds whereas today it is bout 3.5 million posts every 60 second . Twitter in 2016-278000 posts but today they are getting 920000 posts . Due to increase in use of smart devices which are equipped with GPS (Global Positioning System) , location based services has / had become very prevalent , thus attracting the interest in research community [3] .

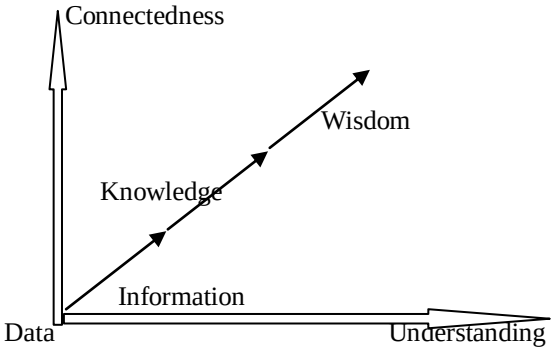

Fig-1. Wisdom Hierarchy of Information Knowledge

Revised Manuscript Received on April 18, 2019.

Dr. Nripendra Narayan Das, Associate Professor
Department Of Computer Science and Engineering
Manav Rachna International Institute of Research and Studies Faridabad, Haryana ,India
Dr. Govind Singh Patel, Professor SEEE Lovely rofessional University. Jalandhar, Punjab, India

When big data is analyzed with huge traditional enterprise companies, working on large volume of data will have better understanding of their business, which can have a significant impact on the business. In such situation , processing of information very quickly is a vital necessity for governance. Financial institutions which are involved in fetching external price information about stocks and other financial Institutions in real-time, in combination with some internal algorithms are used for purchasing or selling decision.

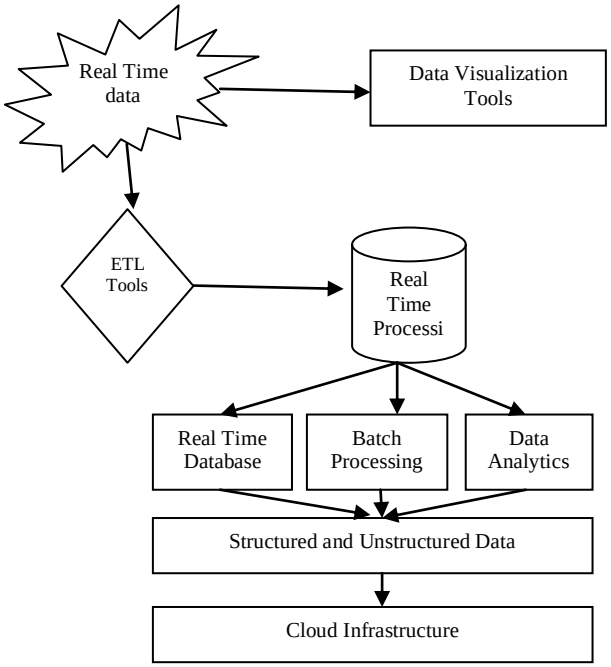

Fig-2. Big data with cloud computing.

Organizing and forming Big, Valuable Data is one of the costly process where many organizations are investing but to make a exact and high-quality volume of data huge resources are required. Currently, cloud- based solutions are used in more than 85% of Fortune 500 companies, and most of entrepreneur expects that 90% of new apps will be installed on cloud platforms[1]. During early days when Big Data are used by all the organization on the cloud computing platforms are installed in Hadoop clusters. provided by different Infrastructure-as-a-Service (IaaS) providers such as Flipcart Web Services and Rackspace for testing and development of the system , and analysis of existing datasets[9] . Data storage and data back-up are offered by these providers in a very low cost-effective manner. To store their structured and unstructured data such companies also provide a minimal cost and trustworthy environment that gives organizations the computing resources Software-as-a-Service (SaaS). The output in the form of analytics are given for the end users through a graphical interface . The queries formulation and incorporation to the data source on the cloud are prerequisites the organizations must need to undertake before any services can be delivered.

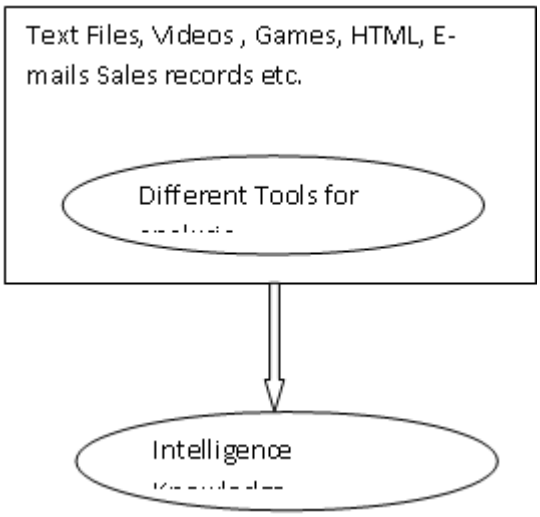

FIG-3. KNOWLEDGE DISCOVERY IN BIG DATA

Now a days Big data has changed the concept and one of the best ways for taking managerial decisions is provides . However , because big data has surpassed the capacity and capabilities of conventional storage and analytics systems, it is looking for new problem-solving approaches because of the convergence of social networking , mobility, wireless data, advanced database technologies and powerful computing.

II .4/5 v's OF ONLINE SOCIAL MEDIA

A.. Velocity

Velocity is the speed in which the data actually getting generated on these networks.

B. Variety

Variety of contents that are getting generated on social media.

C. Veracity

Veracity which is to see the confirmation, which is to find out whether that infomation which is posted on social media rgitimate and actually not very hard.

D. Volume

Volume is the size of the content that is getting posted.

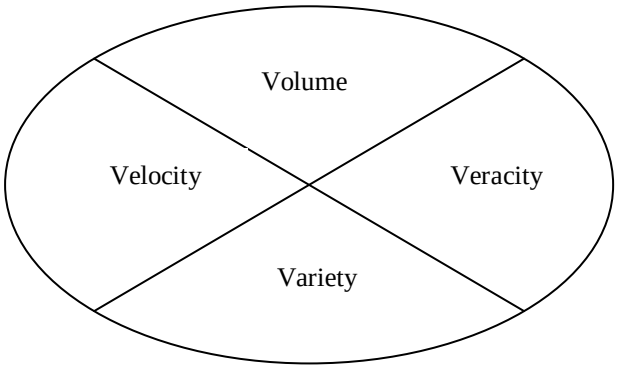

Fig-4. 4 v's OF ONLINE SOCIAL MEDIA

Our study consists of various privacy and security concerns on social media. , Collect data onto online social media analyze the data and visualize . The data onto in the context of privacy and security of online social media . Primary focuses is going to be different aspects of security and privacy on social media . Throughout the study we are also exposed to actually collect data onto online social network like Facebook , twitter , raising this content and visualizing this data onto terms understand whether the followers that i have on twitter are actually valid or fake . Then we look at social network analysis , text analysis that can be done using the content of social network and NTLK is one of the p/f that helps to analyze the contents from social networks ..

III. RELATED WORKS

A. Application program interface

Online social media API enables developers to interact with the online social media website programmatically . We use API's to extract data from twitter , Facebook etc. Each to interact with its own API and API rate limits[3]. In our study we would look for the Facebook and twitter API which would helps as to collect data from them. API returns data in the following two formates :-
JSON - Java script object notation.
XML - Extensible markup language.

B. FACEBOOK API

Access token : It is a key which opens door for Facebook API for you . Access token is an authentication string generated using the open authentication or the protocol which Facebook uses to verify the user or the authentication of the user[4]. Essentially any access tokenizes key corresponds to an application or a door which leads to the graphs API you cannot enter the graphs API without a key is the access token and you cannot generate a key of a door . E the application . Access token is used to extract data onto the API . In this query_field we have fields id . After clicking or submit API responds to the two fields that you asked for . The ' me ' part that we had in the query_tells the API that it needs to return the ID a name field for ' me ' which is the current authenticating user. Now these are other things which you can extract from API. Just click on the search for a field area and you can see what you want to search for . These access tokens have a very short life span for example : - If we open the access token then in access token debugger tool then you will notice that this token will expire in some minutes which mean if you make a request for the graph API using this particular token then after the specified minute the API will not return any data instead it will return an error saying that has token is invalid or expired[6] . If you intend data to collect program partially_for a prolog period if time you cannot keep on generating new token manually after one or every two hour and put in the code . In this case Facebook provides a way to extend the lifespan of the token but for that we need to create an application of our own . Graphs Representation and social network analysis :- . In this section we will learn how to represent social media data onto graph format consisting of nodes and edges and also learn the basics of social network analysis .

IV. GRAPH

Graph is a data structure which consists / consisted of a finite set of nodes and edges Nodes :- Entities(like users)
Edges :- Kind of relationships.
Various ways of representing a node edge graph
A - Adjacency matrix
B - Graph ML
C - CSV
Adjacency matrix :- It is a 2-D square matrix whost size is equals to number of nodes in the graph.
In this example since graph has 6 nodes so size of the matrix is 6*6. The cell at intersection of 1st row etc. The columns is 1 if an edge exist between node i and node j . Therewise 0. In this example there is an edge from node 1 to node 2 and 3. Therefore the cell at the intersection of 1st row and 2nd column sets a 1 similarly for the cell at 1st row and 3rd column sets 1. Adjacency matrix can easily constructed using an array data structure in any programming language . However if the input graph has high number of nodes and less edges then the resultant adjacency matrix can be very Spence and spacecrsuming . So this can be overcome by using another way i . E the graph ML format .
Graph ML : - It is an XML file format for graphs . It consists of an XML file containing a graph element within which an undated sequence of nodes and edge elements . Each node element should have a distinct ID attribute and Each edge element has source and forget attribute that identifies an end point of an edge of two nodes . We can use twecoll (a command lines / lined tool) to get twitter data onto graph ML format . using twecoll we can collect information about friends (followers on twitter , friend of friends) . Also we can use gephi tool to generate a network graph of your twitter data . Social network analysis matrics In a directed graph we have in degree , out degree and degree . Indegree : - It is a number of edges entering a node . Outdegree : - It is a number of edges leaving a node . Degree : - It is an integer and out degree . In this example edge are going away from node 2 to 4 and node 2 to 5 and node 2 also has a self loop therefore out degree of node 2 is 3 and similarly integer of node 2 is 2 . Centrality : - It find out which is the most central or important node . Indegree centrality :- . It finds the node with the highest integer it can signify the most instuendes node . example : - In case of twitter follower graph the user with highest number of followers . Outdegree centrality :- . It helps in locating the node whole out degree is highest . Other ways to measure centrality are :- . Betweeness centrality : - It is equal to number of the shortest path of all vertices to all others that pass through that node . Closeness centrality : - It helps to find the nodes with the lowest total distance from all other node . SNA other matrics : - . Community : A community is a graph of similar or strongly connected nodes. The measure to define the strength of community is modularity.
Modularity :- Fraction of the edges that fall within the given group.

V. PROFILE LINKING ON ONLINE SOCIAL MEDIA

Profile linking is a technique which gives the facility to connect profiles of a user on various social networks platform. Linked profiles can help companies like Dissey to make huge database for potential customers in a cost effective way[5] . Existing methods used to link profiles of user by taking high similarity between recent values of the attributes like name and username . However , many users change their attributes in a regular interval and choose different values in their profiles , these current values have low similarity . Knowing this can be useful for: - twitter , Facebook , blogs , digg , YouTube , linking etc . De-duplicating audience : - Facebook 437,632 likes , twitter 153K followers , linking 805,097 followers , So the total social audience is the sum of Facebook , twitter and linking or less so user have also same accounts in all of these so the challenges is little personal information descriptive opinions like twitter , YouTube then the heterogeneous online social networks gives quality and descriptive personal information in Facebook and professional information in linking . When the registration of same information on both online social networks then the attribute evolution gives the information evolved into one but not on other . Fake followers . Extensive research has provided methods to identify fake followers [1] , and spam accounts [2].

VI. PROFILE LINKING

(i) List common attributes

(ii) Compare attributes values using syntactic , semantic or graph based methods .

(iii) High similarity denotes profiles refer to a single user

(iv) Values considered here are the most recent values of the attributes .

Example : - How the people change the handle ? Registration : time

(v) @xyzgr in Facebook and @xyzgr in twitter . Observation : time

(vi) @xyzgrs in Facebook and @exp_xyzgr in twitter . Observation : time

(vii) @xyzgrs in Facebook and @imindian in twitter . This gives matching values of comparing with Facebook and twitter . - Then the problem is given two user profiles and the respective username sets , each composed of past and current usernames . It is only user names because It gives and provides a unique attribute of a user , universally and publicly available attribute , homogenous , character and length restricted and easier history collection methods for username as other attributes .

VII. GROUND TRUTH COLLECTION

(i) Self identification behavior .

(ii) Extrovert users . – The user has past usernames collection to understand in a given sample which is show - User ID : 856542874 - Past usernames on twitter : ["bigeasye_ " , "reez_ " , "epiceric4_ " , " soulanla " , "swampson_ " , " hebeth " , " swampkid_"] - Past Usernames on instagram : ["bigeasye_ " , "reez44_ " , "epiceric71_ " , " soulanla " , " swampson "] - There are many user which have a common name so if they want to create an own account . So it will be adding with number for an example a common name is rahul if he wants to create an own account in any social network sites then he uses user name such as rahul52 or rahul77 etc . Features Behavioral patterns of usernames sets : - - Username Creation Behavior - Username Reuse Behavior - Username Creation Behavior . In username creation behavior which is consisting static behaviour pattern and temporal behavioral patterns . The static behaviour pattern consists same length , same characters choice , same characters arrangement and the patterns behaviour. In username reuse behavior which consists occasional reuse patterns , frequent reuse patterns . The occasional reuse patterns consist common username , best similarity score , second best similarity score . The frequent reuse patterns consist common username set , temporal ordering and temporal symbol . Datasets The data is a collection of multiple social network of the past . and current user handles .

VIII. CONCLUSION

- Profile linking may be necessary for many organizations.
- Better profile linking is possible with past history of user handles.
- Linking profiles which have twitter - Instagram, twitter - tumblr and twitter - Facebook. Past usernames available for both profiles 21446 positive pairs, 21449 negative pairs. Past usernames available only on twitter but current username available on other profile 112,451 positive pairs, 112,451 negative pairs .The supervised are classified into independent supervised framework and cascaded supervised framework. The framework configuration is exact match, substring match, independent, cascaded (Native and SVM).

REFERENCES

1. D. M. Boyd and N. B. Ellison, “Social Network Sites: Definition, History, and Scholarship,” J. Comp.-Mediated Commun., vol. 13, no. 1, Oct. 2007, pp.210–30.
2. Wikipedia, “Social Network Service,” 2010; http://en.wikipedia.org/wiki/Social_network_service
S. B. Barnes, “A Privacy Paradox: Social Networking in the United States,” First Monday, vol. 11, no. 9, Sept. 2006.
3. [4] R. Gross and A. Acquisti, “Information Revelation and Privacy in Online
Social Networks,” Proc. WPES ‘05, Alexandria, VA, Nov. 2005.
B. Krishnamurthy and C. E. Wills, “Characterizing Privacy in Online Social Networks,” Proc. WOSN ‘08, Seattle, WA, Aug. 2008.
J. He and W. W. Chu, “Protecting Private Information in Online Social Networks,” in Intelligence and Security Informatics: Techniques and Applications, H. Chen and C. C. Yang, Eds., Springer, 2008
4. L. Banks and S. F. Wu, “All Friends are Not Created Equal: An Interaction Intensity Based Approach to Privacy in Online Social Networks,”Proc.WSPOSN ‘09, Vancouver, Canada, Aug. 2009.
H. Yu et al., “Sybilguard: Defending Against Sybil Attacks via Social Networks,” IEEE/ACM Trans. Net., vol. 16, no. 3, June 2008, pp. 576–89.
5. Jain, Vinay Kumar, and Shishir Kumar"Big Data Analytic Using Cloud Computing", 2015 Second International Conference on Advances in Computing and Communication Engineering, 2015.

6. Paridhi Jain, Ponnurangam Kumaraguru, Anupam Joshi. "Other times, other values: leveraging attribute history to link user profiles across online social networks", Social Network Analysis and Mining, 2016.
7. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, “Big Data: Issues and Challenges Moving Forward”, IEEE, 46th Hawaii International Conference on System Sciences, 2013.
8. Mervis, J. 2012. “Agencies Rally to Tackle Big Data”, Science, 36(4): 22, June 6, 2012
9. Bizer C, Boncz P, Brodie ML, Erling O. The meaningful use of big data: four perspectives–four challenges. ACM SIGMOD Record. 2012; 40(4):56–60
10. Ibrahim Abaker Targio Hashema, Ibrar Yaqooba, Nor Badrul Anuara, Salimah Mokhtara, Abdullah Gania, Samee Ullah Khanb. 2015. The rise of “big data” on cloud computing: Review and open research issues, Information Systems, Elsevier. 47: 98-115.
11. The Search for Analysts to Make Sense of Big Data. Yuki Noguchi. National Public Radio, Nov. 30, 2011
12. V. Mayer-Schonberger, K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt, 2013
13. Big data, http://en.wikipedia.org/wiki/Big_data, 2014
14. Labrinidis A, Jagadish HV (2012) Challenges and opportunities with big data. Proc VLDB Endowment 5(12):2032–2033
15. Big Data: 3 Open Source Tools to Know Firmex http://www.firmex.com/blog/big-data-3-open-source-tools-to-know.
16. Cisco Cloud Computing Data Center Strategy, Architecture, and Solutions [Online] http://www.cisco.com/web/strategy/education

AUTHORS PROFILE

Dr. Nripendra Narayan Das, did his Ph.D from Gautam Buddha University. He is currently as Associate Professor in Department of Computer Science and Engineering at MAnav Rachna International Institute of Research and Studies, Faridabad, Haryana, India. His research interests include Deep Web Search Engine, Big Data, Data Mining , Text Mining etc. He is having more than 22 years of experience in industry as well as in teaching. He has published his research papers in many reputed journals. He is also reviewer of many reputed Journals and Conferences.

Dr. Govind Singh Patel, received the Master degree in Instrumentation & Control Engineering from MD University, Rohtak, India. And he has done PhD in Electronics and Communication Engineering from Thapar University, Patiala, India. He is working as a Professor in Lovely Professional University, Jalandhar, PB. He has published more than 48 papers in National and International Journals. He is a reviewer of many international journals like Springer, JCTN etc.