DESIDOC Journal of Library & Information Technology, Vol. 43, No. 4, July 2023, pp. 258-265, DOI : 10.14429/djlit.43.4.19236

Text Mining of Twitter Data for Mapping the Digital Humanities Research Trends: A Case Study

Arti Sawale* and Paramjeet Kaur Walia

University of Delhi, Delhi- 110 007, India
**E- mail: artisawale@gmail.com*

ABSTRACT

Digital humanities have become a more relevant field of study due to the extraordinary growth in digitisation of the humanities data. Due to collaborative development of humanities and computing, many academics are convinced of the worth of digital humanities (DH) that actually provides the best insight into humanities studies. The panoramic view of the development of big data in humanities reflects its trendy directions and evoked new challenges in DH. It is complicated to analysis the objectives of digital humanities data with simple data analysis tools where as text mining can help to facilitate the qualitative findings in DH. In the humanities disciplines, data is often in the form of unstructured and text mining is a way of structuring and analysing digitised text-as-data. Twitter is a online social networking platform which offers an opportunity for quality information sharing, collaborative participation of digital humanities community. This paper is attempted to study the extensibility of digital humanities on twitter and also to interpret the evolution of twitter usage by analysing tweets posted related to DH via python data analysis.

Keywords: Digital humanities; DH; Twitter; Big data; Text mining; Python; Tweets

1. INTRODUCTION

Data is everywhere and very simple data can give a lot of really important patterns. With the rising interest in digital humanities and social science, big data analytics is no longer a domain solely for computer science. The extraordinary growth in the ability to store, process and analyse data has brought us to a focus on the big data around the globe. A massive amount of information has been generated in this modern digital era. On one side, we have a huge volume of unprocessed data like corpus, IoT, social media posts and comments, blogging etc. and on the other side, we are digitising a lot of things like google books, with the purpose of enhancing the understanding of individuals and collectives. Digital Humanities is a process of digitising the study related with humanities area and visualise DH results subject to new approaches in the humanities. It studies the digital media resources scenario in the human society disciplines and how these disciplines must approach the knowledge of computing. The advent of the world wide web (www) accelerated the transition in digital scholarship, like a massive growth in repository creation, to make a cultural legacy available globally for a wide range of purposes which leads to study the fundamental challenges that informs digital humanities linked towards big data.[1]

The objective behind using the Artificial Intelligence (AI) technology in humanities is to facilitate the digital humanities challenges[2] more precisely for current and future DH research. Text mining emphasises the broader analysis of unstructured data so large that no individual can read it in a reasonable period of time. With time, the majority of online data like articles, blogs, web pages, images, video audio, email attachments etc. are in the form of unstructured data. Twitter a microblogging[3] is considered as one of largest source of big data where millions of tweets get shared with unstructured data type in a day where big data characterised by volume (quantity), velocity (rapid growth) and variety (types and nature). Although text mining helps to make sense of these unstructured collections of data. Developing the values from the issues involved in the large-scale data analysis of cultural data by computational methods to explore the traditional humanities disciplines, also referred as humanities computing is an emerging area of research.[4] Humanities studies the aspects of human society and culture which leads to an interdisciplinary paradigm shift in big data for digital humanity.[5]

Received : 28 February 2023, Revised : 06 May 2023
Accepted : 28 July 2023, Online published : 04 September 2023

Social media and crowd sourcing uptakes the humanities community to digital humanities.[6] Twitter is the leading social media platform for researchers to connect online with other academics and professionals. Data analysis tools like python and RStudio rely on the API (application programming interfaces). Similarly, twitter API provides access to python for data collection. In this study, the methodology used to collect twitter data is without twitter API in python. Both ways are providing access to twitter data collection; the pros and cons about both methods are provided further in this study.

2. REVIEW OF LITERATURE

The literature has been reviewed to examine the digital humanities context in which the studies show how social networking informs and expands the findings of big data. Myers[7], et al. in his research findings studied the characteristics of twitter in both aspect of information network and a social network. Greenhow[8], et al. in their article 'Twitteracy' provided an overview of tweeting as a literacy practice for traditional and new literacies. To study the importance of web resources for understanding society, Grandjean[9], in his case study, broadly visualised the digital humanities community network on twitter to process the subjective nature of twitter user's analysis. Quan-Haase[10], et al. investigated how digital humanities community use twitter for informational gratifications. Ross[11], et al. studied twitter as a backchannel in academic conferences. Majdabadi[12], et al. provided a scoring algorithm to find the most relevant and coherent trends. Adapting the text mining method to humanities research gives rise to more refined and specific information from large quantity of unstructured datasets. Starbird & Palen[13], studied the microblogging information diffusion activity during the 2011 political events and analysed the nature of retweets on twitter. Manaskasemsak[14], et al. proposed an approach to ranking tweets, words and hashtags in each trend extraction using the graph clustering techniques with respect to their importance and relevance to the topics and analysing all tweets systematically to produce more reliable results. For mapping twitter population with twitter API, Bruns[15], et al. described the emerging uses of social media platform like twitter enabled the new research method to deal with big data. Kruchten[16] in his twitter case study conclude that digital humanities can be used to provide real world approaches to problems facing society.

3. THE SCOPE OF THE STUDY

The scope of the study is to measure the effectiveness of twitter for digital humanities scholars to share information, current research trends, publications and projects related to digital humanities in an interactive manner. Also, to know what kind of content resonates with digital humanities scholars based on tweet-specific user engagement data and which twitter influencers circulates references relevant to research interest and current trends in DH. This study highlights twitter usage most importantly, to present own views and initiatives, disseminate research projects, provide links to research works and information about conferences, seminars, workshops and scholarly communication related to digital humanities.

4. DATA COLLECTION

The main challenging task in the study is to collect the data from twitter platform. There are several processes to get twitter data, one of is twitter APIs (Application Programming Interfaces) which allow to access and download tweets where twitter's 'streaming API's i.e., for real time tweet provides collection up to 1 % of the total volume of tweets as per twitter data collection guidelines and another is twitter's REST (Representational State Transfer) API which allows to collect old tweets maximum up to 3,200 for a required topic.[17] This much limited tweets data is not sufficient to predict anything therefore different methodology is used to collect large number of tweets for this study.

In this study the data analysis tool python with scraper library used for tweets collection without twitter API. Twitter data includes unstructured data like URLs, emojis, user mentions, images, hashtags, links etc. Further for text mining processing a wide range of tools are provided by Natural Language Toolkit (nltk) which has been introduced. Python NLP (natural language processing) package to clean the tweets data for removing stop words, punctuation, tokenisation and etc. Python libraries and packages for NLP and allow easy scraping of tweets with a rather simple line of code. Python libraries facilitate to extract the tweets with the required twitter attributes like 'id', 'date', 'url', 'content', 'user', 'replyCount', 'likeCount', 'retweetCount', 'lang', 'source', 'sourceUrl', 'media', 'hashtags', 'cashtags', 'outlinks, 'qoutedTweet', 'mentionedUsers', 'place' and etc. https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet The best part is there is no limit in scraping with necessary attributes. The time factor of query running output cannot be avoided here as twitter continuous and explosive growth and quickly becomes overwhelming. Also, content gets rapidly pushed further down the scraping and miss content from even a few minutes ago in real time.

The advanced search facility of twitter uses hashtags or text phrases to search for the mentioned topic of interest. This is useful for tracking a specific twitter user account; however, it is very difficult to get the results at the wider corpus and analyse overall objectives of study. Text mining of digital humanities with the required twitter attributes (Figure 2) i.e., 'date', 'user', 'tweet', 'URL' and stored them in a data frame (df) to maintain the simplicity of this data set for this study. Twitter advanced search with keywords 'digital' AND 'humanities' year wise data has been collected and merged in the final dataset for further analysis and query execution output set limited due to the runtime.

```
# advanced search query to collect tweets for dataset year wise
query = "digital humanities until:2022-12-31 since:2022-01-01"
tweets = []
limit = 50000
```

Figure 1. Python code for year the 2022 DH tweets.

```
    # print(vars(tweet))
    if len(tweets) == limit:
        break
    else:
        tweets.append([tweet.date, tweet.username, tweet.content

df = pd.DataFrame(tweets, columns=['Date', 'User', 'Tweet', 'URL
print(df)
```

```
                          Date            User  \
0     2022-12-30 20:29:01+00:00         KrzyRuch
1     2022-12-30 19:41:12+00:00     paolomonella
2     2022-12-30 19:30:51+00:00         drdevroy
3     2022-12-30 19:09:36+00:00  GouldLibraryBot
4     2022-12-30 19:04:39+00:00      caroline32i
...                         ...              ...
43485 2021-01-01 07:23:25+00:00        ACNdigHum
43486 2021-01-01 02:59:15+00:00     infoliterati
43487 2021-01-01 00:19:39+00:00     lee_blog_bot
43488 2021-01-01 00:16:51+00:00        DHDefined
...                         ...              ...
43485 2021-01-01 07:23:25+00:00        ACNdigHum
43486 2021-01-01 02:59:15+00:00     infoliterati
43487 2021-01-01 00:19:39+00:00     lee_blog_bot
43488 2021-01-01 00:16:51+00:00        DHDefined
43489 2021-01-01 00:01:26+00:00  inklessEditions

                                                   Tweet  \
0      https://t.co/xb2QAsziIr - Specjalistyczny serw...
1      A job post is out for a post-doc in #DigitalHu...
2      Comparative literature and the digital humanit...
3       I am on first libe working on digital humanities
4      Digital Humanities and Material Religion: An I...
...                                                  ...
43485  Happy New Year from Animal Crossing: New Digit...
43486  This is a great special issue, well worth a br...
43487  On the only worked for *mes intimes* Digital H...
43488  Digital humanities is a field that uses contem...
43489  My family traditionally eats 12 grapes on the ...

                                                     URL
0      https://twitter.com/KrzyRuch/status/1608923129...
1      https://twitter.com/paolomonella/status/160891...
2      https://twitter.com/drdevroy/status/1608908493...
3      https://twitter.com/GouldLibraryBot/status/160...
4      https://twitter.com/caroline32i/status/1608901...
...                                                  ...
43485  https://twitter.com/ACNdigHum/status/134490706...
43486  https://twitter.com/infoliterati/status/134484...
43487  https://twitter.com/lee_blog_bot/status/134480...
43488  https://twitter.com/DHDefined/status/134479971...
43489  https://twitter.com/inklessEditions/status/134...

[43490 rows x 4 columns]
```

Figure 2. Python output with necessary twitter attributes for year the 2021-2022 DH tweets.

In Figure 2, the last row shows the total output of 43490 tweets for the query run (for the year 2021-2022). Similarly, the output from the year 2008 to 2022 is extracted year wise by python programming code as per Figure 1 and finally, the total of 2,92,878 tweets are collected from 2008 to 2022. The next stage of data cleaning and data analysis with reference to twitter hashtags, word frequency occurrence and twitter users is given in results and discussion.

5. RESULTS AND DISCUSSION

5.1 Twitter Hashtag Analysis

The hashtags are self-tagged topics which give valuable insights about the context of a discussion and a great way to reach targeted users. One of the main objectives of using hashtags on twitter is to get maximum visibility and encourage users' engagement. Tweets with '#DigitalHumanities' or '#DH' hashtag postings showing the interdisciplinary nature of tweets related to DH from 2008 to 2022. For example; tweets posted- "An amasing work on the 19th century literary history of Australia where newspapers are the main source of fiction #DigitalHumanities #literaryHistory #distantReading #textAnalysis #Australia", here the hashtags show the whole context of a tweet in just a few words.

Figure 3 shows that '#DigitalHumanities' has the largest word count, followed by the interdisciplinary topics associated with digital humanities conferences, new initiatives, workshops, seminars/webinars, dissection forums, research projects etc. at the global level. Exploring the hashtag data provides the best insights into the topics or trends analysis showing the topmost topics from the tweet's corpus of the last 15 years. Figure 4 is visualisation of the topic area using VOSviewer where a color scale shows the yearly involvement of the topics in DH tweets with the keyword 'India' and the density of circles showing the word occurrence of the topic which are occurred with digital humanities in India. It is also showing the gradually increased involvement of DH community in India as an emerging field of study and research over the period.

5.2 Word Frequency Analysis

The structure of twitter conversation for information sharing at conferences, workshops, seminars and webinars are analysed. Twitter activities during the DH conference and interaction help to summarize the events along with the online information sharing, trending DH topics with hashtags and links to additional materials such as presentation slides, publication, research articles etc. It has been observed that analysing the dataset just based on word frequency occurrence provided irrelevant results as the corpus has unstructured and noisy text. Several users have retweeted approximately 12,430 tweets that have been removed from the dataset. Also, after removing the stop words and noisy text from the dataset, figure 5 is the result of frequency of word occurrence related to DH conferences, workshops, seminars and webinars with the required URL (Uniform Resource Locator)links in original tweets by users across the world.URL links are important to shareas resources on social media which helps to recommend or locate the relevant and required web pages[18] of conference, webinar etc.

5.3 Digital Humanities Twitter Users Analysis

Analysing user engagement on precise topics is a challenging task in text mining. Social media analytics helps to design the appropriate user's engagement strategy for measuring the digital humanities research and activities on twitter. This study has attempted to analyze the topmost twitter users accounts, which are making continuous efforts to make available digital humanities research activities. However, the word frequency occurrence of 'digital' AND 'humanities' in tweets posted helps to

Figure 3. Word Cloud of topmost hashtags based on high frequency of occurrence in DH related tweets.

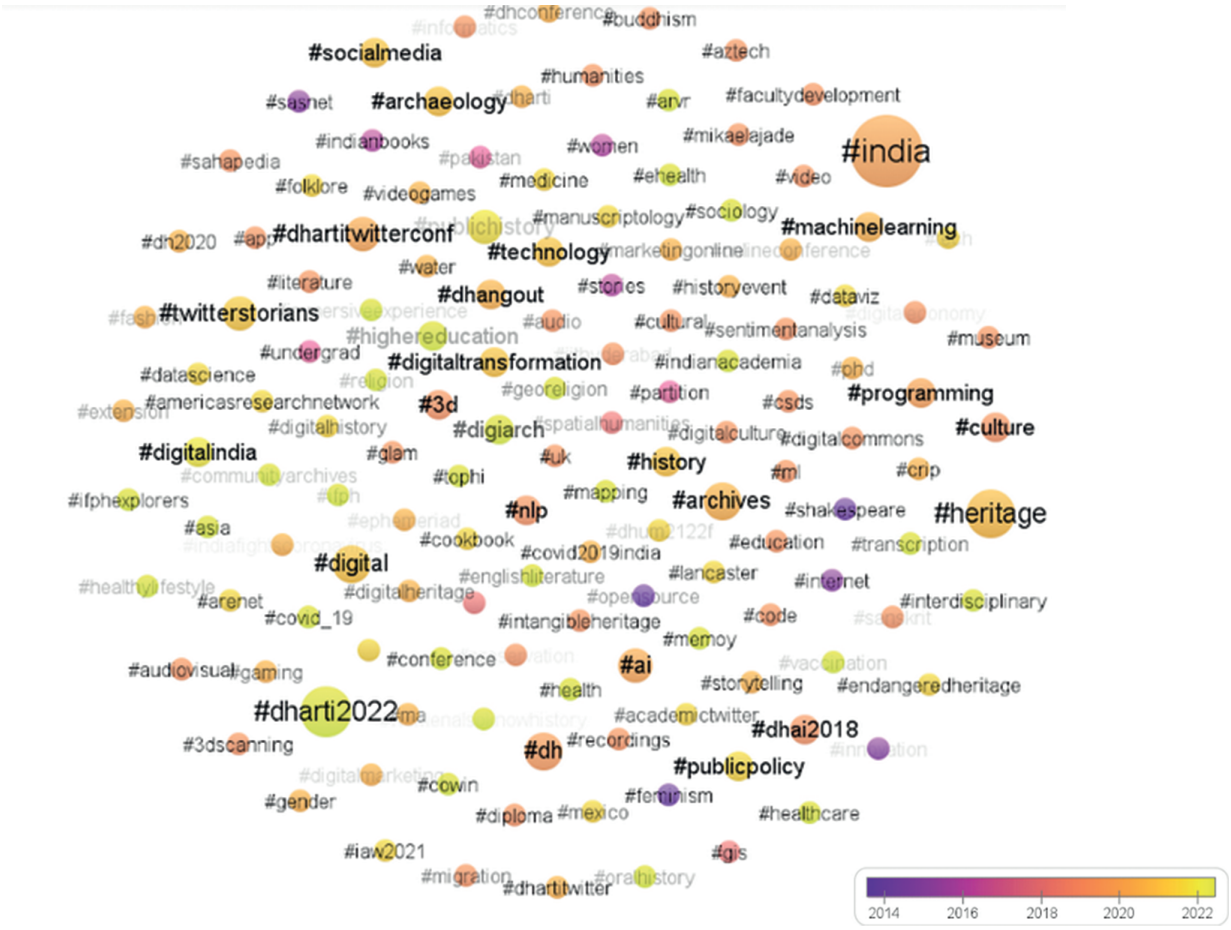

Figure 4. Tweets posted with hashtag '#digitalhumanities' and keyword 'India'.

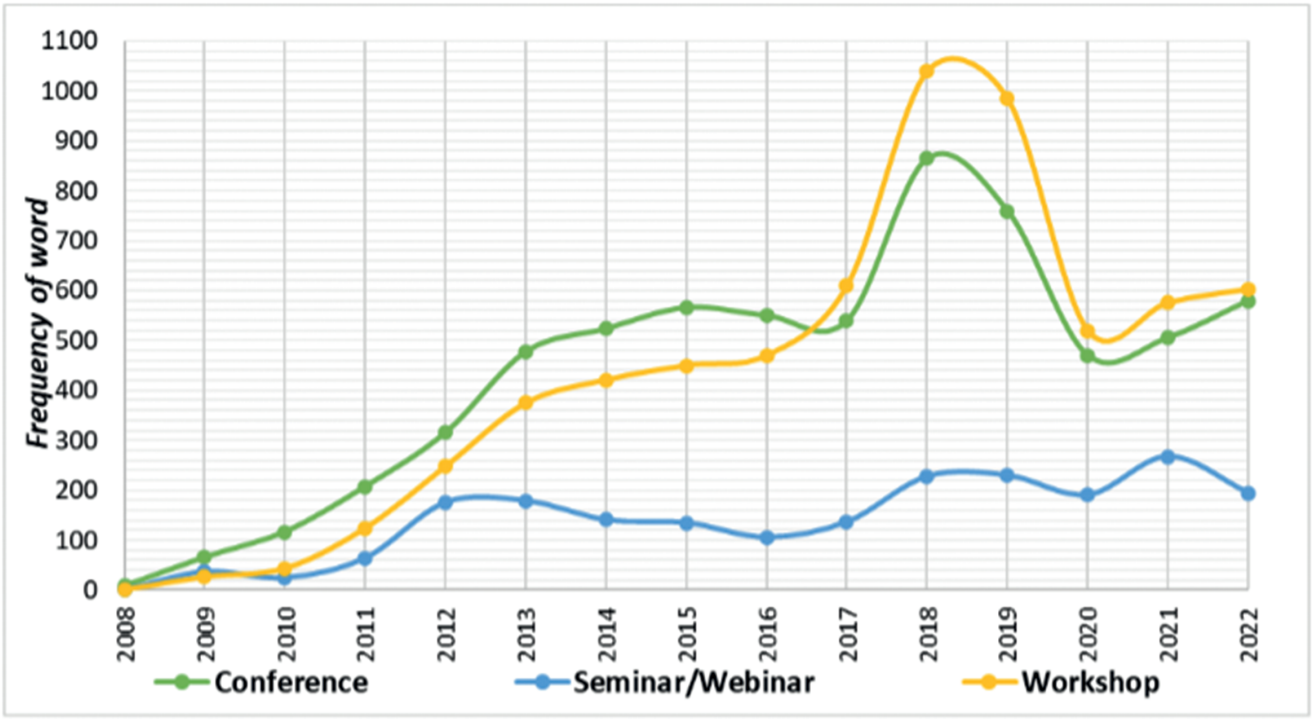

Figure 5. Year wise frequency of DH conferences, seminar/webinar and workshops with respective sources URLs (links).

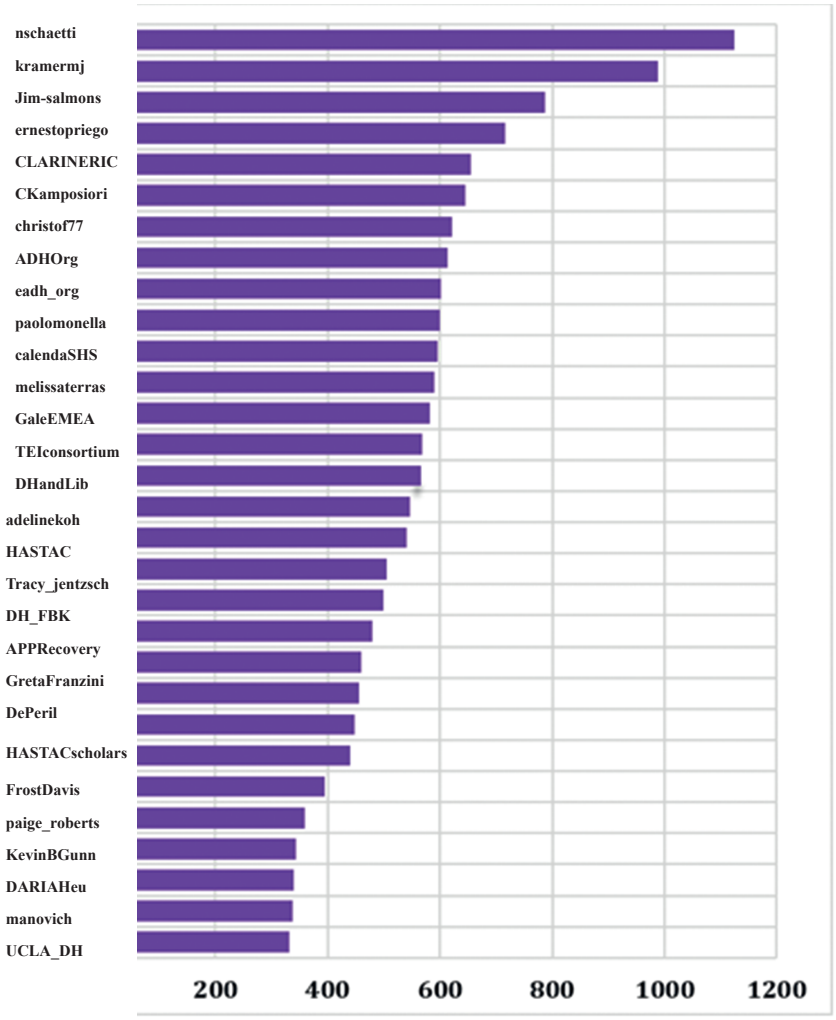

Figure 6. The topmost digital humanities twitter handles.

trace the most twitter handles in the past and present, but several types of indicators have to be considered to measure the success of a twitter in digital humanities, like 'retweets', 'replies' and 'likes'. Figure 6 is a list the topmost twitter handles/users from the collected dataset for this study. This list is based on the tweets posted about digital humanities articles, conferences, workshops, seminars/webinars, projects and other research activities with the respective URL links of the content source. This is one of the most important criteria in this study as mentioning the respective URL's links in the tweets helps users to reach the required source of digital humanities activities. It is observed that, not only the organisations but also the personal twitter accounts and a diverse group of historians, anthropologists, media studies scholars, curators, information technologists, digital humanists, archivists, librarians and library professionals are actively involved on twitter for scholarly communication in the field of digital humanities.

6. CONCLUSION

Text mining is an advanced technique for researchers to extract social networks from linguists and literature. Social media analytics is almost inherently supported in discovering large patterns of topic modeling and collective opinion formation. Tweets related with DH provides an insight into the digital humanities research[11] and academic professional usage of twitter. The interdisciplinary approach can be seen with the hashtags[19] used for trending topics in digital humanities. The gradual increase in twitter usage indicates that the researchers preferred digital format content on the web and twitter exposed broader participation, professional collaborations, sharing the digital humanities projects and other activities. Also, this paper is an attempt to identify the availability of the digital humanists' community on twitter, importantly considered as a great source for easy reach to digital humanities research. This study proposed text mining with the word frequency occurrence in tweets and limited to English language tweets. Further, it can be analysed for more precise research. This study is limited to data type i.e., text data in tweets, research can be extended by analysing more data types like geographical area, twitter images, analysis of URLs, sentiment analysis of user replies and retweets of DH tweets.

REFERENCES

1. Edmond, J. & Lehmann, J. Digital Humanities,

Knowledge Complexity, and the five Aporias of Digital Research. *Digital Scholarship in the Humanities*, 2021, **36**(2), 95-108.
Doi:10.1093/llc/fqab031.

2. Ye, J. The Application of Artificial Intelligence Technologies in Digital Humanities: Applying to Dunhuang Culture Inheritance, Development, and Innovation. *Journal of Computer Science and Technology Studies*, 2022, **4**(1), 31-38.
Doi: 10.32996/jcsts.2022.4.2.5.
3. Cheung, B.; Wong, CL.; Gardhouse, A.; Frank, C. & Budd, L. #Cgs2015: An Evaluation of Twitter Use at the Canadian Geriatrics Society Annual Scientific Meeting. *Canadian Geriatrics Journal*, 2018, **21**(2), 166–172.
Doi: 10.5770/cgj.21.302.
4. Luhmann, J. & Burghardt, M. Digital Humanities-A Discipline in Its Own Right? an Analysis of the Role and Position of Digital Humanities in the Academic Landscape. *Journal of the Association for Information Science and Technology*, 2021, **73**(2), 148-171.
Doi: 10.1002/asi.24533.
5. Xuyu, N. Art History research in the digital humanistic era: Paradigm shift and method transformation, *In* Proceedings of the 7th International Conference on Humanities and Social Science Research (ICHSSR), 2021, pp. 757-762.
Doi: 10.2991/assehr.k.210519.151.
6. Terras, M. Crowd sourcing in the digital humanities. A new companion to digital humanities. 2nded.In Schreibman, S. & Siemens, R. editors. Wiley-Blackwell. 2016, 420-439
Doi:10.1002/9781118680605
7. Myers, S.A.; Sharma, A.; Gupta, P. & Lin, J. Information network or social network? The structure of the twitter follow graph. *In* 23rd International World Wide Web Conference, 2014, pp. 493-498.
Doi: 10.1145/2567948.2576939
8. Greenhow, C. & Gleason, B. Twitteracy: Tweeting as a new literacy practice. *The Educational Forum*, 2012, **76**(4), 464-478.
Doi: 10.1080/00131725.2012.709032.
9. Grandjean, M. A social network analysis of Twitter: Mapping the digital humanities community, *Cogent Arts & Humanities*, 2016, **3**(1), 1171458.
Doi: 10.1080/23311983.2016.1171458.
10. Quan-Haase, A.; Martin, K. & McCayPeet, L. Networks of digital humanities scholars: The informational and social uses and gratifications of Twitter. *Big Data & Society*, 2015, **2**(1), 205395171558941.
Doi: 10.1177/2053951715589417.
11. Ross, C.; Terras, M.; Warwick, C. & Welsh, A. Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*, 2011, **67**(2), 214-237.
Doi: 10.1108/00220411111109449.
12. Majdabadi, Z. et al. Twitter trend extraction: A graph-based approach for tweet and hashtag ranking, utilising no-hashtag tweets. *In* 12th Conference of Language Resources and Evaluation Conference, 2020, pp. 6213-6219.
https://aclanthology.org/2020.lrec-1.762.pdf (Accessed on 24 January 2023)
13. Starbird, K. & Palen, L. (How) will the revolution be retweeted? Information Diffusion and the 2011 Egyptian Uprising. *In* Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, 2012.pp. 07-16.
Doi:10.1145/2145204.2145212.
14. Manaskasemsak, B.; Chinthanet, B. &Rungsawang, A. Graph Clustering-Based Emerging Event Detection from Twitter Data Stream. *In* Fifth International Conference on Network, Communication and Computing, 2016, pp. 37-41
Doi:10.1145/3033288.3033312.
15. Bruns, A.; Burgess, J. & Highfield, T. A Big Data Approach to Mapping the Australian Twitter sphere. In Arthur, P.L.; Bode, K. editors. Advancing Digital Humanities. Palgrave Macmillan, London,2014, pp.113-129.
Doi: 10.1057/97811373370168.
16. Kruchten, D. Keeping the human in digital humanities: A Twitter case study. Carolina Digital Repository. 2021. https://cdr.lib.unc.edu/concern/honors_theses/3r075346p
Doi: 10.17615/114q-e567.
17. Research guides: Resources for text and data mining: Twitter. https://guides.libraries.emory.edu/main/text-data-mining/twitter. (Accessed on 28 January 2023)
18. Nizam, N.; Watters, C. & Gruzd, A. Link sharing on twitter during popular events: Implications for social navigation on websites, *In* 47th Hawaii International Conference on System Sciences, 2014, pp. 1745-1754.
19. Tammaro, A.M. Evaluation of digital humanities: An interdisciplinary approach. *In* Catarci, T., Ferro, N., Poggi, A. editors. Bridging Between Cultural Heritage Institutions. Springer, Berlin, Heidelberg, 2014, pp.136-146.
Doi: 10.1007/978-3-642-54347-0_15.

CONTRIBUTORS

Ms Arti Sawale is a PhD Research Scholar at Department of Library & Information Science, University of Delhi and she received the Bachelor of Engineering in Electronics and Telecommunication from PRMIT & R, Amravati University. Currently, she is a Scientific & Technical Officer-I(LS) at INFLIBNET Centre, actively associated with SOUL library management software testing, data conversion, training, technical and operational assistance for the software.
For the present study, she conceived the idea of developing the proposal and presentation in the required format, software installation and programming, data analysis, evaluation, interpretation of study results, preparing and reviewing the manuscript.

Prof Paramjeet Kaur Walia has been teaching Library and Information Science for the last thirty-three years. Currently, she is teaching in the Department of Lib. & Info. Sci., University

of Delhi since 2003 and previously, at the Department of Lib. & Info. Sci., Panjab University, Chandigarh for thirteen years. Her area of specialisation are Library and Information Science Education, Government Information, Research in Technical Libraries and Information System and Programmes, Public Library and Information System also Management of Library and Information Centre. She has contributed several papers in of national and international repute journals. She has also presented papers in international and national conferences. She was one of the members of UGC Curriculum Development Committee and contributed in the curriculum development of Bachelor of Library and Information Science.

For the present study, she contributed in reviewing the manuscript and interpretation of study results more specifically.

Appendix I: List of top DH twitter handles and description provided in their respective twitter profiles

Twitterhandle/username	Twitter bio description
Dr. Nils Schaetti/@nschaetti	- Researcher in machine learning
Michael J. Kramer/@kramermj	- Professor of history
Jim Salmons/@Jim_Salmons	- Citizen scientist
Digital Humanities Now/@dhnow	- Web portal, digital humanities scholarship
Ernesto Priego/@ernestopriego	- Senior lecturer, editor at ComicsGrid, a peer-reviewed open access journal published by Open Library of Humanities
Clarin Eric/@CLARINERIC	- The Europeanresearchfor the Social Sciences & Humanities, supports research on cultural data, language resources and technology
Christina Kamposiori/@CKamposiori	- Executive program officer at Research Libraries UK
ChristofSchöch/@christof77	- Co-Director of the Trier Center for Digital Humanities, editor Journal of Computational Literary Studies (JCLS)
ADHO/@ADHOrg	- Organisation for digital research, publication, training to promote DH
EADH /@eadh_org	- Association for Digital Humanities in Europe
Paolo Monella/@paolomonella	- Researcher in Latin & DH at La Sapienza, Rome.
Calenda/@calendaSHS	- Web platform for DH digital resources
Melissa Terras/@melissaterras	- Prof of Digital Cultural Heritage at University of Edinburgh
GaleEMEA/@GaleEMEA	- Publish research in partner with global digital researchers
TEI/@TEIconsortium	- Consortium, training and develops software's for TEI tools
Dh+lib/@DHandLib	- Official twitter portal of Dh+lib, Digital Humanities and Libraries
Adeline Koh/@adelinekoh	- Former tenured professor postcolonial literature
HASTAC/@HASTAC	- An alliance and open community, academic DH network
Tracy H. Jentzsch/@Tracy_Jentzsch	- Digital humanist, supporter of Alt Ac Public Humanities.
DH Group at FBK/@DH_FBK	- DH Research Group at Fondazione Bruno Kessler
Recovering US Hispanic/@APPRecovery	- Recovery is an international project to locate, preserve and disseminate US Latino culture
Greta H. Franzini/@GretaFranzini	- Postdoc Researcher in Research Language Infrastructures & Digital Humanities
Aaron /@DrPeril	- Digital humanist, gamer, web designer
HASTAC Scholars/@HASTACscholars	- Scholar's program official account
Rebecca Frost Davis/@FrostDavis	- AVP for digital learning at St. Edward's University DH in undergrad curriculum & liberal arts in networked world
Paige Roberts/@paige_roberts	- Archivist, public historian
Kevin Gunn/@KevinBGunn	- A Coordinator of digital scholarship and Lecturer at the Catholic University of America
DARIAH ERIC/@DARIAHeu	- A network of expertise in DH
Manovich/@manovich	- Artist and writer - AI, digital art, media theory, digital humanities. Professor GC_CUNY cultural
UCLA DH/@UCLA_DH	- Research in the intersection of technology and humanities