

DISCOVERING THE OPEN ACCESS MOVEMENT ON TWITTER: AN EXPLORATORY STUDY

*Muhammad T. Sadiq
Akhilesh K. S. Yadav*

Mr. Muhammad Sadiq T.
Research Scholar,
Indian Institute of
Technology Madras,
Chennai – 600 036, TAMIL
NADU, INDIA
Email:
connectsadiqt@gmail.com

Dr. Akhilesh K. S. Yadav
Assistant Professor,
Tata Institute of Social
Sciences,
Mumbai – 400 088,
MAHARASHTRA, INDIA
Email:
akhilesh.yadav@tiss.edu
Corresponding Author

Open Access possesses unconfined reuse and free access of electronic resources. This research focuses on scholarly discussions on 'Open Access' in the most common microblogging platform Twitter. The main objectives of the study are to identify the locations, trends and applications used by scholars for frequent tweets; to apply text mining techniques to analyse unstructured text content on the Open Access; to find out the pattern, context with Open Access. Data collection process involved gathering tweets of one month using specific keyword 'Open Access'. During the research period, the highest number of tweets on Open Access was on 17th January 2018 and the least number of tweets was on 6th February 2018. The tweets posted on these days were on variety of topics, and most of the tweets were tweeted from United States. #OpenResearch, #OpenScience, #OpenScholarship and #OpenPR, #OpenData etc. were the most popular tweet hashtags used during the research.

Keywords: Open Access, Text Mining, Twitter, Twitter Trends; Open Science, Open Research

INTRODUCTION

The incredible development of computer hardware technology, happening since couple of years, has increased the credibility of data collection equipment, storage media, powerful and reasonable computers. Every single second, millions of data are generated around the globe from various sectors. Moreover, this development has significantly influenced the information industry. Looking at the timeline of information technology, one can certainly state that data and text mining is a revolutionary development in the field. Taking social media as one of the critical data sources, we could observe how it produces digital pictures, blogs, videos, virtual communities, and different kinds of social networks. In this data age, the main reason behind the acceptance of data and text mining in the information industry is due to the easy availability of a large amount of data and the market demand for converting such data to useful

knowledge and information (Gorunescu, 2011). The knowledge and information gained can be used for business management, market analysis, science surveys, to detect counterfeiting, to maintain customer relations etc. Data mining can be described as “an automatic process of extraction of non-trivial or implicit or previously unknown but potentially useful information or patterns from data in large databases, data warehouses, or in flat files (Han et al., 2012).”

World Wide Web (WWW) is also considered as a global information center for financial management, government, education, e-commerce, consumer information, media, advertisements, news etc. Thus, it comprehends a huge and rich collection of information about web page contents like hyperlink information, web documents, weblog information, hypertext structures, multimedia etc. Web mining is one of the applications of data mining methods to extract knowledge from the network and to discover structures and patterns (Prasad et al., 2012). A large quantity of information is gathered as text such as web pages, news articles, digital libraries, books, emails, technical papers, blogs etc. Around the globe, a huge amount of information is stored in an unstructured manner (Saini, 2014). Thus, arises a need for a useful technique that can extract needed information from unstructured text (Sarawagi, 2007). From raw textual data, text mining extracts new pieces of information using various techniques (Farzindar & Inkpen, 2015). One of the main goals of text mining is to extract valuable information from text (Talib et al., 2016). Text mining has many applications in several areas, for example, business intelligence, telecom

industry, customer chain management systems, commercial applications, sentiment analysis or opinion mining, etc. (Mukherjee & Bhattacharyya, 2013).

OPEN ACCESS

Open Access is defined as “Open Access literature is digital, online, free of charge, and free of most copyright and licensing restrictions (Suber, 2012)”. Open Access is a common term in the context of knowledge where information resources or publications are kept in public domains for public access without any subscription charges or with limited restrictions for the users. Open Access is promoted through online platforms; thus, Open Access developers and users should make sure that they have online connectivity for sharing knowledge. Free access to the work or publications benefits everyone, for authors, their work gets more citations, readers can access to more works and can update their area, and for funders, they get more audience for the work that they have funded.

It gives freedom to the author for publishing in cross platforms according to their choice. This helped the researchers of physics, mathematics, statistics, computer science etc. for self-archiving their scientific papers. Predominantly Open Access publications are available on two channels, Gold OA and Green OA. Recently, few more Open Access models are introduced by commercial publishers to get their authors’ content featured in Open Access domains. E-journal publishers adapt most of these models. Hybrid Open Access, Delayed Open Access, Short-term Open Access, Selected Open Access and Partial Open Access are popular Open Access models.

TWITTER

Twitter is a social networking system and microblogging platform founded in 2006 by Jack Dorsey, Noah Glass, Biz Stone, and Evan Williams. This is an online platform where users can share and interact with his/her followers. Soon it has attracted a huge number of users and became the most successful social interaction platform and popular broadcasting media of information. Users can share their views, ideas, the information in short messages instantly without any barriers. As of 2020, monthly active users of Twitter are more than 330 million. In a day, Twitter produces more than 500 million conversations referred as 'tweets'; tweets consist of 280 characters. There is an informal dictionary of words that are founded over these limited characters and are only used within social media platforms. Non-standard abbreviations, irony, typographical errors, sarcasm, and hashtags (#) were the trends of Twitter. Hashtags are referred as the trending topics, and it is one of the best contributions from Twitter to the social networking system. The highlight of Twitter is that it allows only 280 characters, and most of these tweets posted on different subject areas are considered as a meaningful source of information.

LITERATURE REVIEW

In any subject, before creating a new knowledge, there should be an understanding on background of the given subject. The literature review on this subject is conducted to establish a background on text mining approach for finding context. The reviews of the following literature deals with various aspects such as user's behavior, their preferences etc.

Borruto (2015) used one-year Twitter data for mining about the behavior and preference of Twitter users. Luo et al. (2015) have proposed a ranking model for Twitter opinion retrieval. This model combines information on social and opinion at edness for tweets' opinion retrieval. The findings of the study reveal that the opinion retrieval increases when links, mentions, author's information like status count or followers are taken into the user account. Huq et al. (2017) state that the character limit in Twitter forces the user to be precise and expressive at the same time. That is how Twitter becomes a massive source of belief mining and sentiment analysis. Park et al. (2016) conducted a Twitter analysis of four types of Asian restaurants' dinner perception (Chinese, Japanese, Korean, and Thai). The researcher had referred 86,015 tweets on these Asian restaurants. Text mining and sentiment analysis were used for finding the meaningful patterns, opinions, emotional states, and famous words that they have used. The finding was that, compared to the other three restaurants, Chinese restaurant's average sentiment scores were crucially low. Most of the positive tweets were about food quality, and negative tweets were about the quality of service in these restaurants.

Text mining techniques that follow the analytical method handled by patent analysis. Some of the text mining techniques used here are; text integration, summary abstraction, topic identification, term connection, cluster cohort, feature selection, and information plotting (Tseng et al. 2007). Similarly, Yan et al. (2015) had analyzed research papers on E-commerce. They recognized 68 phrases as primary keywords of

the E-commerce area using text mining techniques. Ajao et al. (2015) discuss location inference techniques on Twitter. Location inference can be used in many areas, and it is needed for consumer and marketing user profiling. It helps millions of videos on different topics to be shared on the world wide web. Pons-Porrata et al. (2007) analyzed indirect knowledge present in news streams using a topic discovery system. Scalfani (2017) conducted a text analysis to study chemistry thesis and dissertation titles using programmatic text analysis, particularly term frequency collections. Approximately 10,000 chemistry dissertation and thesis titles during the period 1911-2015 accumulated from nine major south-eastern (United States) research universities were taken into the analysis. This method was instrumental in identifying some of the most frequent terms of figuring in chemistry dissertation and thesis titles.

Palguna et al. (2015) come up with a theoretical formulation for sampling Twitter data. A large number of tweets from various applications needed more techniques that scale with the number of tweets. Anger and Kittl (2011) have attempted a grounded approximate for calculated individual Twitter accounts social networking potential. Boyd et al. (2010) mentioned that retweet had become a custom; user retweets for various purposes and in different styles. Here, the authors examined retweeting as a way by which users can be in information exchange. Using empirical data and series of case studies, the paper states that retweeting is a conversational exercise. Bild et al. (2015) focused on the quantitative description with total user behavior and their tweet graph. They argue

that the retweet chart is a balance free and tiny sphere as similar to the social graph.

In another research paper, researchers have applied a high dimensional visualization approach to testing the effectiveness of longitudinal analysis of Twitter message framing during the period. The finding of these mixed methods is, it allows enough reactivity to support and identify the analysis of both trending and non-trending subjects on Twitter (Uren & Dadzie, 2015). Soboleva et al. (2017) created an ideal mechanism for NGOs providing power to send tweets to all their network followers. The result was quite surprising that the most tweets were with limited mentions, and many of the tweets were without any mentions. Text mining is an emerging technological method to extract significant information from unstructured textual data. So, a series of research papers were collected, summarized, and analyzed for the literature review. There are many research papers available on text mining, especially on the application of text mining in sentiment analysis, topic detection, summarization, opinion mining etc. However, there is no research study conducted on 'Open Access' on Twitter data. Therefore, the researchers have attempted to find out the context and trends on Open Access by analyzing the tweets.

RESEARCH OBJECTIVES

1. To understand the scholarly discussions on Open Access in the most common microblogging platform Twitter.
2. To identify the locations, trends, and applications used by scholars for frequent tweets on Twitter.

3. To apply text mining techniques to analyze unstructured text content on the topic 'Open Access'.
4. To find out the pattern and context of Open Access.

RESEARCH METHODOLOGY

In this research, Twitter was used to collect data. The data collection process involved gathering tweets of one month's period between 7th January and 7th February 2018 using the specific keyword 'Open Access'. The data gathered is evaluated to identify the systems of Twitter action, which involves scholars and academic communities in social media discussions. Associations will be observed between: i) directed tweets from the academicians and mentions of the #OpenAccess by others on Twitter; and ii) comments from the academicians and comments from others on Twitter with #OpenAccess.

DATA ANALYSIS AND INTERPRETATION

Trend lines of tweets

On 7th January, Biblioteka PG (tweet ID) from Gdańsk, Polska (Poland) has initiated the

first tweet on OA and the day ended with 230 tweets. The highest number of tweets on Open Access was on 17th January. The tweets posted on this day were on variety of topics, and most of the tweets were from European countries. From 10th January to 14th January the tweets went down from 1,826 to 773. From 15th January to 18th January, we can see that there was a hike on OA tweets and 1604 to 1882 tweets were posted. #OpenResearch, #OpenScience, #OpenScholarship, #OpenPR, #OpenData etc. these were the common tweet topics on these days. Again, from 18th January to 31st January (890 to 769 tweets), there was a fall on number of tweets. The average tweets on OA on these periods were 688 tweets. When coming to 1st February, the new month was started with 813 tweets and in February as compared to month of January, there was no much discussion that happened on Open Access on Twitter. The least number of tweets on OA was on 6th February with 110 tweets and the highest number of tweets posted on Open Access was on 17th January with 1882 tweets (figure 1).

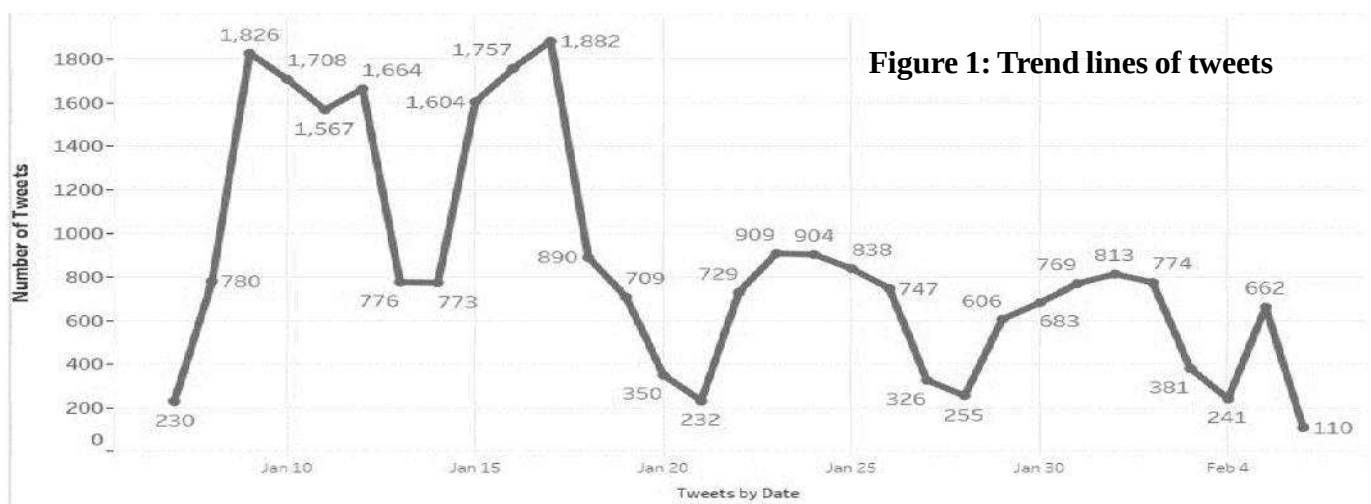


Figure 1: Trend lines of tweets

Applications used for posting on Twitter

The figure 2 shows the kind of application people have used for posting tweets on Open Access during the analysis period. They have used Twitter from different platforms. 9,799 (36%) users have used 'Twitter Web Client' application for posting tweets. The total 4,816 (18.18%) users have used Twitter iPhone applications and 4,345 (16.40%) users used Android applications. Thus, we can infer that most of the Twitter users prefer their phones for posting tweets. This may be because of users' comfort and access to phone. Schonfeld (2009) states that, there are around 21 third party Twitter clients for posting tweets. TwitPic, TwitterBerry and FriendsFeed are some of the popular Twitter clients. Out of 26,495 users, 5906 (22.29%) are from Apple's application e.g. iPhone, iPad, macOS and iOS Twitterbot. The least tweets were posted from Twitter Facebook application and the highest tweets on Open Access were posted from Twitter web client. TweetDeck, Hootsuite, Buffer, Twitter Lite, Sprout

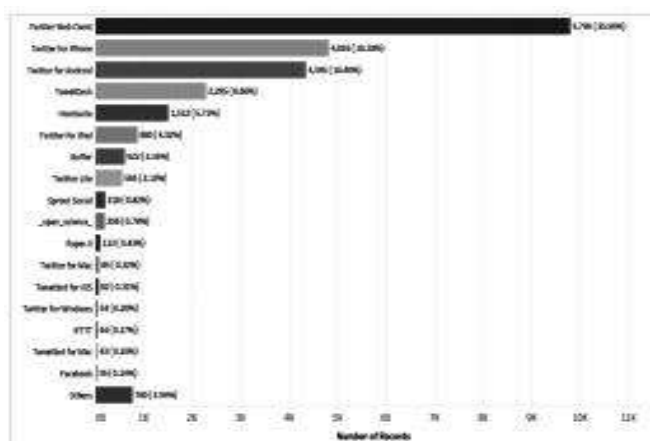


Figure 2: Applications used for tweets

Social, and IFTTT etc. are some of the other significant applications used for tweets. The very less numbers of tweets (30 users) were posted by using other applications.

Most of the tweets were tweeted from the United States of America. This is because compared to other

countries, Twitter is more prevalent in the USA, and during our analysis period, there was a discussion happening on Open Access and Open Science. For some of the tweets, their location information was missing because in Twitter profile, most of them haven't provided their locations, but somehow Twitter has managed to update location details and many tweets using 'current tweet locations'.

What are the languages used by users for posting tweets? So, all the tweets were analysed according to language that they were used for tweets. It was not a big surprise that from whole of 26,496 tweets, significantly 18,102 tweets (68.32%) were in US English language, followed by standard German language used for 2,252 tweets (8.50%), Spanish used for 1,595 tweets (6.02%), British English used for 1,297 tweets (4.90%) and French for 1,266 tweets (4.78%). As discussed before in the previous paragraph, most of the users were from USA and this may be the reason of English being the most preferred language for tweets. As we know, languages mentioned above are the most popular languages in the world.

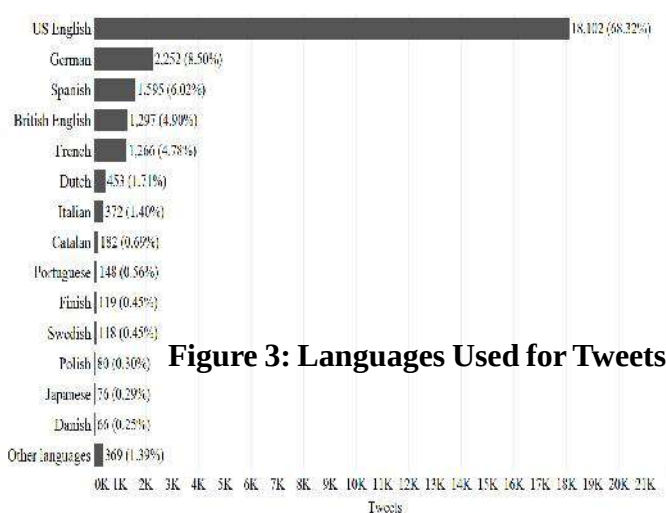


Figure 3: Languages Used for Tweets

It is found that ‘_open_science_’ was the most popular Twitter account which posted tweets on Open Access during our analysis period. ‘_Open_science_’ have 7882 followers on Twitter and so far, they have posted more than 63.5K tweets. Their tweets on Open Access was tweeted 208 times, however these tweets were not tagged. Some of the top tweets were from the following Twitter account: the MJA, UCLpress, Science_Open, ferli90, OpenResLeeds, SDawsonBerlin etc.

Timeline of Retweet and Favorite Count

The figure 4 basically describes the timeline of Retweet count and favorite count during the analysis period. 27,100 is the highest retweet count, this was on 15th January 2018. The results indicate that, on this day there was a discussion on “creative commons licenses” and around 40+ single user’s tweets were retweeted 150 times. In case of favorite count 3,211 is the highest count that happened on 10th January 2018. ‘GillC,

IamJamesWaatts, UTSePress, ossjay, Ellyll and panopen etc.’ were some of the Twitter users who have contributed for 27,100 retweets. When we compare retweets and favorite, there is a huge difference. Users prefer retweet because of many reasons, one of the main reasons is that it helps to increase followers. There are bots which helps for automatic retweet. Here the highest retweet counts are as following: 27,100 RT (15th January), 25,983 RT (17th January) and 25,883 RT (8th January). 3,211 FT (10th January), 3,134 FT (12th January) and 3,012 FT (9th January) are the highest favorite counts. 962 RT (4th February) is the least count in retweets and in favorite 77 FT (27th January) is the least count.

More the number of followers, more the retweets (Figure 5). Most of the grey circles (Retweet count) are near zero. It reveals that users with less number of followers get lesser retweets. The users who have less than 100K followers got 889 retweets. And the users with less than 200K

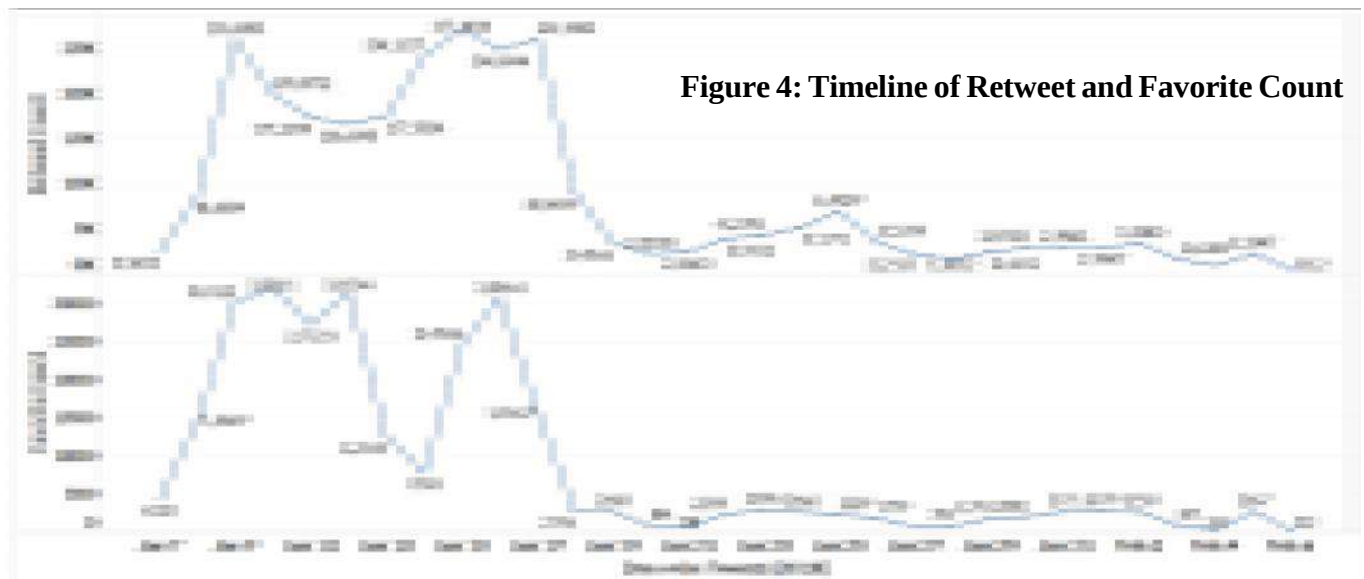


Figure 4: Timeline of Retweet and Favorite Count

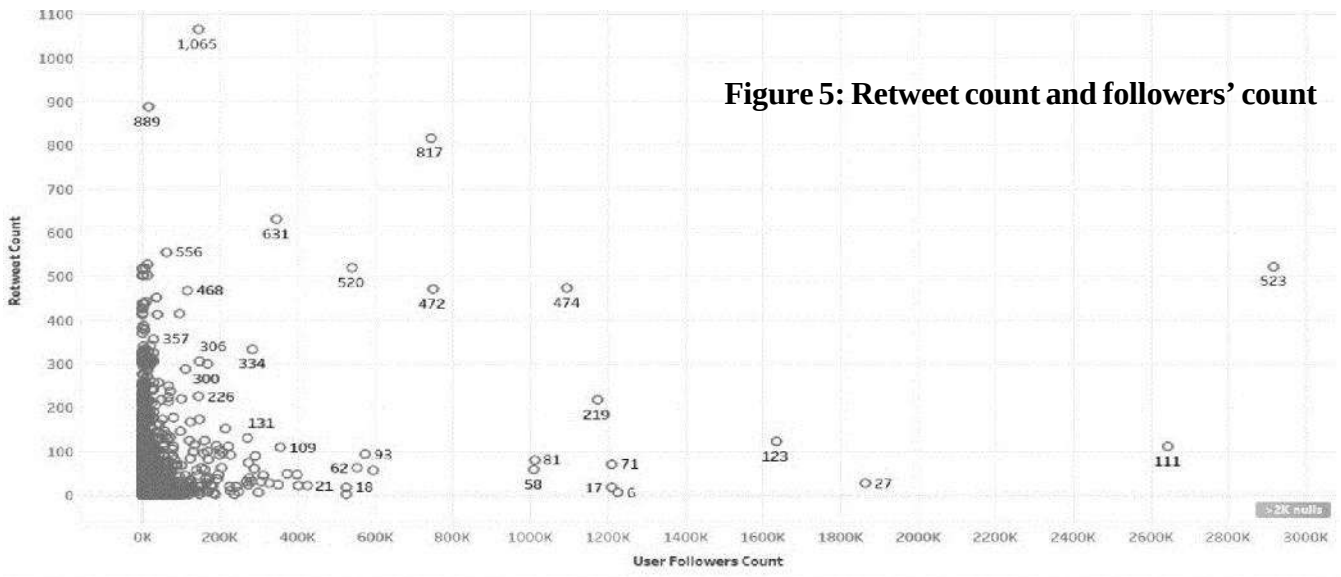


Figure 5: Retweet count and followers' count

got 1,065 retweets. On the other side, there is a user who has more than 2600K followers and he got only 111 retweets. Thus, with these results, we endorse to the statement that there will be more retweets if the number of followers are more. From the results, it is evident that users with more followers get more retweets, besides that users are considering the content of tweets. According to Zarrella (2009) users with more followers will get more retweets, some users are able to get lots of retweets without lots of followers; their content must be more contagious. Similarly, Boyd et al. (2010) also have the opinion that retweeting can be both a productive, communicative tool and a selfish act of attention seekers.

The results show that '@Jisc' got more retweets. '@Jisc' is an organization based in UK which has 30 years of experience providing digital solutions for education and research in the country. Since 2009, they are on Twitter. "RT @Jisc: Do you know how to use the Creative Commons licenses? #openaccess #JiscOA". This

tweet was posted on 13th January 2018 and it got 47,450 retweets during our analysis period. '@WikimediaUK' has also retweeted this as WikimediaUK is the supportive community of Wikipedia. This tweet was posted along with Creative Commons licenses informative chart. So, this can also be a reason for the huge number of tweets. The second top retweet was also from '@Jisc with 20,316 retweets for this tweet posted on 8th January. "The Creative Commons licenses are used widely, but do you know how to use them? #openaccess #JiscOA". Both the top tweets were from '@Jisc, posted as a continuation on the same topic 'Creative Commons licenses' and these tweets were posted with the chart. While analyzing the top trending retweets it was evident that most of the tweets were about new technologies. Interestingly, all these top or most trending retweets were from the account which has more followers. The same tweet which was posted by @Jisc listed in top retweet 'count' also got maximum number of 'favorite'. But the number of 'counts' is very less compared to number of

retweets. The maximum number of retweets were 47,450 and maximum favorites is 297. It's interesting that, only two of the tweets got more than 200 favorites. This shows the attitude of users towards favorites.

Frequent Words in the Tweets

Word cloud is a simple and instructional way for text analysis and to understand the textual data. The bold and biggest word has the highest term frequencies. The figure 6 is the word cloud of frequent terms of the entire tweets. The whole tweets were analysed by using Voyant Tools software. *Voyant Tools* is a web-based reading and analysis environment for digital texts. Here the term ‘openaccess’ is the most frequently used term among all the tweets analysed, having mentioned 26,724 times. The reason behind most frequent term is that our research is based on Open Access, and for the retrieval of tweets’ #openaccess’ was used. The corpus has 694,197 total words and 40,384 unique word forms. Here

the most frequent words from the corpus are as following: openaccess (26724), new (3248), research (2895), open(2386), openscience (2321), oa(2025), article(1957), journal(1906), read (1819), paper (1706), journals (1036), check (984), use (931), commons (796), great (673), work (654), latest (644), opendata (638), online (589), licenses (564), publish(562), creative (539), book (538), papers (538), health (537), academic (530), scholarly (514). All the stop words have been removed while extracting the frequent words.

SIGNIFICANCE OF THE STUDY

The research is an outcome of the observation and analysis of data collected for a short duration of time of only one month. The harvesting of the data on ‘Open Access’ from Twitter can be done in large-scale as high system configurations are required for harvesting large-scale data. There were many limitations with the harvesting of large-scale data but then also we have tried to get some insightful information out of harvested

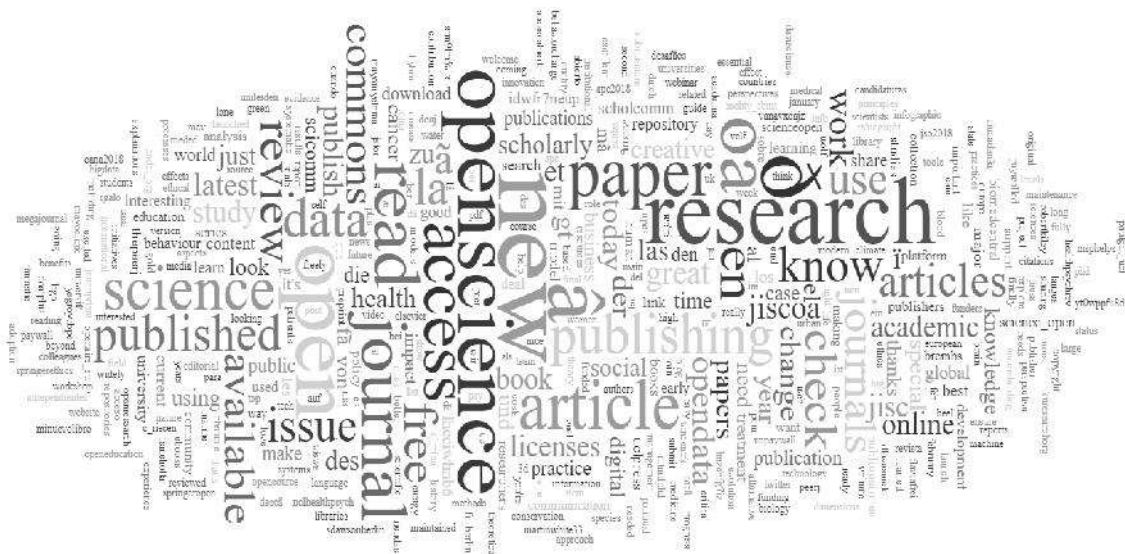


Figure 6: The Most Frequent Words Found in the Tweets

tweets of one month. The purpose was to understand the scholarly discussions happening on Open Access, as Open Access possesses unconfined reuse and free access. It could be inferred from the results that there are other related topics to 'open access' e.g. open science, digital scholarship, scholarly communications, institutional repositories, digital libraries, research ethics, creative common licenses, knowledge sharing, review practices etc.

CONCLUSION

Text mining methods are used to find new concepts and insightful information from textual data. This exploratory research aims to create a groundwork for future studies in text mining. Twitter is a platform where large amount of unstructured intellectual data is emerging. This was the rationale behind choosing Twitter as field of study. The purpose was to understand the scholarly discussions happening on Open Access, as Open Access possess unconfined reuse and free access. Data was harvested using the keyword '#openaccess' during the period 7th January to 7th February 2018. During this one month, we have harvested 26,497 tweets. On this day tweets were posted on a variety of topics, and most of the tweets were from United States. #OpenResearch, #OpenScience, #OpenScholarship and #OpenPR, #OpenData etc. were the most popular tweet hashtags used during the research. The highest number of users 9,799 (36%) have used 'Twitter Web Client' Application for posting tweets on the microblogging social media. 4,816 (18.18%) users have used 'Twitter iPhone' application and 4,345 (16.40%) users have used 'Twitter Android' application for posting tweets.

REFERENCES

1. Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science*, 41(6), 855-864.
2. Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, Article No.: 31, pp 1-4. <https://doi.org/10.1145/2024288.2024326>
3. Bild, D. R., Liu, Y., Dick, R. P., Mao, Z. M., & Wallach, D. S. (2015). Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. *ACM Transactions on Internet Technology*, 15(1), 1-24.
4. Borruto, G. (2015). Analysis of tweets in Twitter. *Webology*, 12(1). <http://www.webology.org/2015/v12n1/a131.pdf>
5. Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the Forty-Third Hawaii International Conference on System Sciences* (pp. 1-10). IEEE. <https://doi.org/10.1109/HICSS.2010.412>
6. Farzindar, A., & Inkpen, D. (2015). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies*, 8(2), 1-166.
7. Gorunescu, F. (2011). *Data mining: Concepts, models and techniques*. Berlin: Springer-Verlag.
8. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). San Francisco: Morgan Kaufmann.
9. Huq, M. R., Ali, A., & Rahman, A. (2017). Sentiment analysis on Twitter data using KNN and SVM. *International Journal of Advanced*

- Computer Science and Applications*, 8(6), 19-25.
10. Luo, Z., Osborne, M., & Wang, T. (2015). An effective approach to tweets opinion retrieval. *World Wide Web*, 18(3), 545-566.
 11. Mukherjee, S., & Bhattacharyya, P. (2013). Sentiment analysis: A literature survey. *arXiv:1304.4520*. <https://arxiv.org/pdf/1304.4520>
 12. Palguna, D. S., Joshi, V., Chakaravarthy, V., Kothari, R., & Subramaniam, L. V. (2015). Analysis of sampling algorithms for twitter. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 967-973). <https://dl.acm.org/doi/abs/10.5555/2832249.2832383>
 13. Park, S. B., Jang, J., & Ok, C. M. (2016). Analyzing Twitter to explore perceptions of Asian restaurants. *Journal of Hospitality and Tourism Technology*, 7(4), 405-422.
 14. Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques. *Information Processing and Management*, 43(3), 752-768.
 15. Prasad, M. R., Manjula, B., & Banu, A. (2012). Comparison of data mining and web mining. *IFRSA International Journal of Data Warehousing and Mining*, 2(1), 34-39.
 16. Saini, G. (2014). A survey report for data mining based on web research. *International Journal of Advance Research Ideas and Innovations in Technology*, 1(2), 1-7.
 17. Sarawagi, S. (2007). Information extraction. *Foundations and Trends in Databases*, 1(3), 261-377.
 18. Scalfani, V. F. (2017). Text analysis of chemistry thesis and dissertation titles. *Issues in Science and Technology Librarianship*, 86. <http://istl.org/17-spring/refereed3.html>
 19. Schonfeld, E. (2009, February 20). The top 21 Twitter clients (according to TwitStat). <https://techcrunch.com/2009/02/19/the-top-21-twitter-clients-according-to-twitstat/>
 20. Soboleva, A., Burton, S., Mallik, G., & Khan, A. (2017). 'Retweet for a chance to...': An analysis of what triggers consumers to engage in seeded eWOM on Twitter. *Journal of Marketing Management*, 33(13-14), 1120-1148.
 21. Suber, P. (2012). *Open Access*. Massachusetts: MIT Press.
 22. Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: Techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, 7(11), 414-418.
 23. Tseng, Y.H., Lin, C.J., & Lin, Y.I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, 43(5), 1216-1247.
 24. Uren, V., & Dadzie, A. S. (2015). Public science communication on Twitter: A visual analytic approach. *ASLIB Journal of Information Management*, 67(3), 337-355.
 25. Yan, B.N., Lee, T.S., & Lee, T.P. (2015). Analysis of research papers on E-commerce (2000-2013): Based on a text mining approach. *Scientometrics*, 105(1), 403-417.
 26. Zarrella, D. (2009). The science of retweets. <http://danzarrella.com/science-of-retweets.pdf>

