ASCILITE
SHAPING THE FUTURE OF TERTIARY EDUCATION
Show Me The Learning
Adelaide, Nov 27-30, 2016

Social Media #MOOC Mentions: Lessons for MOOC Research from Analysis of Twitter Data

Eamon Costello
National Institute for Digital Learning
Dublin City University

Binesh Nair
Business School
Dublin City University

Mark Brown
National Institute for Digital Learning
Dublin City University

Jingjing Zhang
Faculty of Education
Beijing Normal University

Mairéad Nic Giolla Mhichíl
National Institute for Digital Learning
Dublin City University

Enda Donlon
Institute of Education
Dublin City University

Theo Lynn
Business School
Dublin City University

There is a relative dearth of research into what is being said about MOOCs by users in social media, particularly through analysis of large datasets. In this paper we contribute to addressing this gap through an exploratory analysis of a Twitter dataset. We present an analysis of a dataset of tweets that contain the hashtag #MOOC. A three month sample of tweets from the global Twitter stream was obtained using the GNIP API. Using techniques for analysis of large microblogging datasets we conducted descriptive analysis and content analysis of the data. Our findings suggest that the set of tweets containing the hashtag #MOOC has some strong characteristics of an information network. Course providers and platforms are prominent in the data but teachers and learners are also evident. We draw lessons for further research based on our findings.

Keywords: HCI, MOOCs, Data Analytics, Twitter, Social Media, Big Data

Introduction

Although MOOCs learners are known to be educated, digitally literature (Jordan, 2104) and, socially networked (McAuley et al, 2010) little research has been undertaken into how MOOCs are portrayed in social media platforms such as Twitter. Studies to date are limited by relatively small datasets or from having taken samples of manually extracted tweets. One study of note in this area looked at users of the Sina Weibo platform, a popular Chinese microblogging website (Zhang et al, 2015). This study screen-scraped 95,015 postings with mentions of MOOC published by 62,074 users on Sina Weibo from a four year period and analyzed the volume of postings according to four time frames: year, month, day of the week, and the time of day. Their work outlined some trends and made an exploratory foray into this topic.

This paper contributes to research into MOOCs by a systematic extraction of a dataset from the global Twitter stream (utilizing the Twitter GNIP API) and interrogating this data via descriptive and content analyses. Our aim was to conduct exploratory analysis of the MOOC discourse on Twitter. We sought to determine, through big data analysis, what conversations are being conducted in the MOOC arena by the range of potential actors such as MOOC platform providers, traditional educational institutions providing MOOCs, MOOC teachers, MOOC leaners and MOOC researchers. Moreover, we sought to probe the use and meaning of the term MOOC itself as negotiated by users of the term on public social media via its hashtag.

Data and Methods

Twitter data for the MOOC dataset was extracted from GNIP API for the period September to December 2015 and augmented with additional data including Klout scores - a social network influencer measure as developed by Rao Spasojevic and Dsouza (2015). The GNIP API produces very large volumes of data and we used cloud computing, data extraction, storage and processing techniques to handle the data. The GNIP Stream API produces more reliable data than more manual techniques such as screen scraping of the public Twitter REST API (Driscoll & Walker, 2014), and also offers more data protection such as excluding data from deleted accounts. The hashtag '#MOOC' was used as a keyword to extract the required data. In this we followed the work of Zhang et al (2015).

The GNIP Stream API provides a file containing data for each 10 minute interval of a specified period. Complex analytics on the data were performed mainly in R (an open-source statistical tool). This study is in line with current state-of-the-art frameworks (Chae, 2015; Lynn et al. 2015) for descriptive analytics and content analytics on Twitter data. The methodology follows the approach for descriptive and content analytics outlined by Chae (2014) and extended by Lynn et al. (2015).

Findings

The MOOC dataset had 32,309 tweets of which 17,910 were original tweets and 14,399 were retweets. Replies constituted 8 percent (1,434) of the total number of the original tweets. The dataset had 4,980 unique hashtags. Obviously #MOOC features in most of the original tweets (17,263). Other popular co-occurring hashtags included #elearning (1,876), #edtech (1,134), #moocs (822), #highered (637), #coursera (631), and #education (594). The average number of hashtags in original tweets was 2.68.

There were 14,890 unique user screen names in the dataset. This indicates that each user on an average sends 1.2 tweets, 0.9 retweets and 0.1 replies. The most active and visible users were identified (See **Error! Reference source not found.**). Activity was calculated as per Chae (2015) i.e. the activity of a user was calculated as the sum of the number of tweets, retweets and replies which the user has contributed to the network. The visibility of a user was determined by the number of followers for each user as on 31st December 2015. **Error! Reference source not found.** shows a line graph to describe the relationship between active and visible users. It can clearly be observed from the figure that the most active users are not the most visible users and vice versa. For instance, @MOOCs (the most active user) is not the most visible user. Similarly, @edX, the most visible user, is not among the top 30 active users in this network.

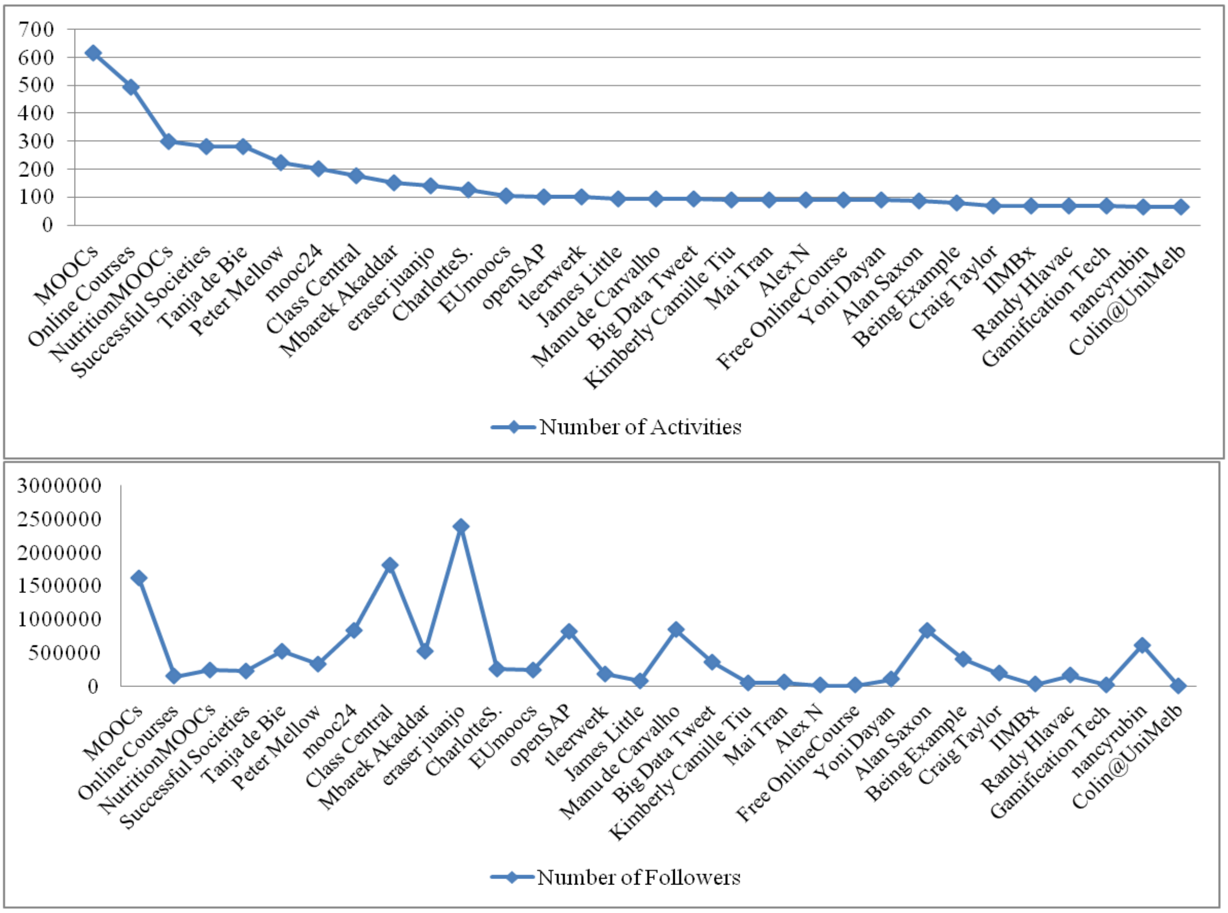

Figure 1: Active Users Vs Visible Users in #MOOC Dataset

Content analytics is primarily concerned with uncovering the patterns hidden inside content. Word analysis, hashtag analysis and sentiment analysis are the analyses which were performed in this category. For performing word analysis, the ‘tm’ library in R was used. Frequent words appearing in the tweets were discovered in order to identify the most popular words among the users in the network. The most popular words were unsurprising i.e. ‘MOOC’ (occurring 21,199 times), ‘course’ (2,733), ‘learn’ (2,686), ‘online’ (2,442), ‘elearn’ (2,079), ‘free’ (1,911), ‘coursera’ (1,410), ‘edxonline’ (1,332) and so on were some of the most popular words. The ‘ngram’ library in R was used to identify the most frequently co-occurring words in the dataset. The most popular co-occurring words included ‘mooc elearn’ (1,412 times), ‘online course’ (1,333), ‘edxonline mooc’ (869), ‘mooc course’ (496), ‘free mooc’ (478) and ‘mooc onlinecourse’ (467). The dataset had 4,980 unique hashtags. Some of the most popular hashtags were #mooc (17,263), #elearning (1,876), #edtech (1,134), #moocs (822), #highered (637), #coursera (631) and #education (594).

Peak detection algorithms were used to identify events of significance in the dataset. In line with Healy et al. (2015), the peak detection algorithms were those presented by Du et al. (2006), Palshikar (2009) and Lehmann et al. (2012). Due to the relatively small number of true peaks and low volume of tweets per peak, the topics were identified manually. Table 1 summarises the topics identified from the true peaks within the dataset. The table also mentions the originated tweet for the topic.

Table 1: Topics of True Peaks

Timestamp	Topic	Originating Tweet of the Topic
21st September 2015, 1900	Promotion of MOOC on “Cognitive Technology and its growing importance for business” by David Schatsky, course instructor and senior manager Deloitte LLP.	@dschatsky leads Deloitte’s #MOOC on #CognitiveTechnology and its growing importance for business: http://t.co/AxNIAePgDL
29th September 2015, 1400	Promotion of Deloitte’s MOOC on 3D printing.	Data from @DU_Press' #MOOC paints a picture of the future applications of #3DPrinting. #DeloitteReview http://t.co/PiCnozQ0ed
14th October 2015, 1300	Successful Societies (by Princeton University) promoting MOOC on ‘How can Governments Improve Citizen Services and Cabinet Office Coordination.’	@crownagents ISS #MOOC examines how govts improve citizen svcs, cabinet office coordination &amp; more. Starts 10/21/15 http://t.co/LJDCKKXL90
16th October 2015, 1400	Successful Societies promoting MOOC on ‘Making Government work in Hard Places’.	@USGLC These leaders made government work in hard places. Learn how. #MOOC: http://t.co/iZveyixK4B http://t.co/7aP0zRShE3
19th October 2015, 1500	Successful Societies promoting MOOC on ‘How Leaders Overcome Governance Challenges’.	@USGLC Still time to enroll! Princeton #MOOC on how leaders overcome #governance challenges. Starts 10/21/15. http://t.co/iZveyixK4B
23rd October 2015, 1600	Successful Societies promoting MOOC on ‘Writing Science of Delivery Case Studies’	@USGLC Enroll today! ISS #Princeton #MOOC on writing “Science of Delivery” case studies. Starts 10/28. https://t.co/AZuu2qFylP
27th October 2015, 1300	Successful Societies promoting MOOC on ‘Writing Science of Delivery Case Studies’	@EU_Commission Starts 10/28! Learn to write case studies on “Science of Delivery” in new free ISS #Princeton #MOOC. https://t.co/AZuu2qFylP
29th October 2015, 1400	Successful Societies promoting MOOC on ‘Writing Science of Delivery Case Studies’	@USGLC Just started 10/28! Learn to write case studies on “Science of Delivery” in new free ISS #Princeton #MOOC. https://t.co/AZuu2qFylP
2nd November 2015, 1500	NutritionMOOCs promoting MOOC on ‘Nutrition and Health: Micronutrients and Malnutrition’.	@APH008 Please RT: 9 November start #MOOC #NUTR102x “Nutrition and Health: Micronutrients and Malnutrition” https://t.co/q0NBluOac9
13th November 2015, 0900	NutritionMOOCs promoting 2nd part of MOOC on Nutrition and Health from Wageningen University.	@EatNutritious Please RT: Learn more about #nutrition and #health in 2nd part of our #MOOC @UniWageningen now: https://t.co/q0NBluOac9
8th December 2015, 1500	Promotion of Coursera’s MOOC on ‘Training TESOL Certificate Part 1: Teach English’.	RT NewsNeus More #Coursera #MOOC #Training TESOL Certificate, Part 1: Teach #English Now!
25th December 2015, 1800	Retweet of Quizalizeapp’s tweet ‘How to say Merry Christmas in 77 Languages’.	How to say Merry Christmas in 77 Languages. #Edtech #GBL #Langchat #MOOC #English https://t.co/5pBv47vjFP

Sentiment analysis is used to examine overall orientation (positive and negative) and intensity (strong or weak) of opinions in text (Pang & Lee, 2008). The ‘qdap’ library in R was used to perform sentiment analysis on this dataset. The average sentiment was found to be 0.095; suggesting that the tweets are highly neutral. The standard deviation of the sentiments across the tweets was found to be 0.202; indicating that the spread of the sentiments across the tweets was less. Further, a customized algorithm to analyse the distribution of tweets across different sentiment scores was implemented in R. If a tweet has more positive words, it will get a higher positive sentiment score. On the contrary, if a tweet has more words negative words then its sentiment score will be more negative. If a tweet has words which do not belong to either category then it qualifies as ‘neutral’. A tweet having a greater proportion of neutral words will have a neutral sentiment; that is a sentiment score of 0. Figure 2 provides a graphical representation for the sentiments distribution in the tweets.

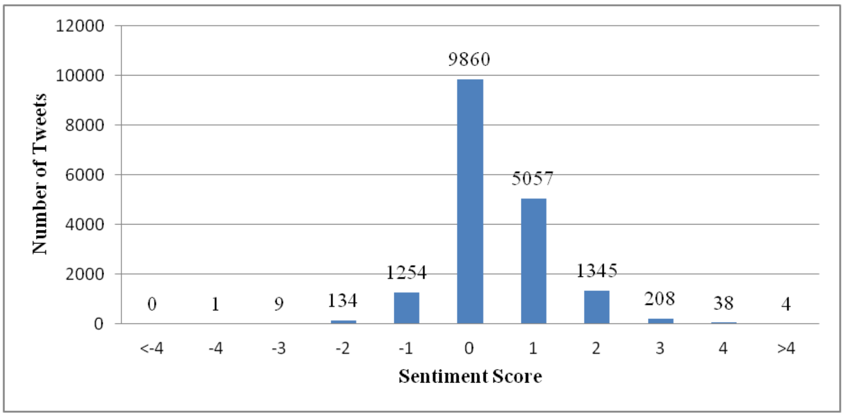

Figure 2: Sentiment Scores in the #MOOC Dataset

As can be easily observed from Figure 2, the MOOC dataset has a substantial amount (55 %) of neutral tweets. Positive tweets make up 38% of the total tweets and the remaining 7% percent constitutes negative tweets. Table 2 lists some exemplar tweets with strong sentiment.

Table 2: Tweets Showing Strong Sentiment

Exemplar Tweet	Sentiment
@thesiswhisperer I think the #MOOC is providing wonderful supportive pillow of trust &amp; honesty- glad I'm taking part- thank u #survivephd15	6
STUNNING #mathed animations from .@robertghrist in his calculus #MOOC. Beautiful and effective. Kudos. http://t.co/GgdN2KHFZf	4
Just discovered a great free #Social Innovation online course, on this cool learning platform - #iVersity #MOOC ~ http://t.co/kGPzmqONFq	4
#ememitalia Teixeira: focusing on dropout as a problem to criticize #MOOC education is a conceptual mistake	-4
CloudComputingApplications - definitely the worst @coursera #MOOC I've ever taken. Irrelevant videos &amp; useless tuts #unenrolled	-3

Finally URL analysis was performed in order to identify the popular URLs (most mentioned) in the network. It was found that URLs were widely used in the network with almost 60 percent of the tweets containing links. A subset of the top URLs are shown in Table 3.

Table 3: Top 15 URLs in the MOOC dataset

URL	Tweets
https://www.edx.org/course/nutrition-health-part-2-micronutrients-wageningenx-nutr102x	327
http://www.owensage.com/2/post/2015/04/how-i-lost-24-pounds-in-12-weeks-amidst-severe-personal-turmoilwithout-dieting-or-going-to-the-gym.html	204
https://www.futurelearn.com/courses/climate-from-space	168
http://www.europeanschoolnetacademy.eu/en/web/developing-digital-skills-in-your-classroom/course	161
http://www.startup365.fr/entrepreneur-courses/	159
https://www.canvas.net/browse/salto/courses/erasmus-funding-opportunities-2	156
https://www.edx.org/course/making-government-work-hard-places-princetonx-mgwx#!	143
https://www.edx.org/course/writing-case-studies-science-delivery-princetonx-casestudies101x	140
http://blog.coursera.org/post/132434298847/introducing-coursera-for-apple-tv-bringing-online	126
https://www.edx.org/xseries/data-science-analytics-context	120
http://www.moocsurvey.org	108
http://www.startup365.fr/the-1-small-business-guide-to-online-marketing/	103
http://Twitter.com/JimKim_WBG/status/661682878393266177/photo/1	96
https://www.youtube.com/watch?v=ahvuPvm-1YU	96
http://www.europeanschoolnetacademy.eu/web/introducing-computing-in-your-classroom	93
https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why	92

Discussion and Conclusion

Peak detection algorithms highlighted tweets of significance in the dataset which largely revolved around the promotion of several MOOCs. The course pages of several MOOCs from the peak detection are referred to in the top URLs. However, the URLs also indicate that the MOOC hashtag may be sometimes appropriated by, or be susceptible, to spam effects e.g. the prominence of weight loss slimming posts. URL 24 points to a book on amazon which contains negative reviews of people who claim to have been duped into following a Twitter link to the page.

The term MOOC may be a problematic one for use in defining networks of MOOC actors. The promotional nature of many tweets suggests this may be more of an informational than a social network (Myers et al., 2014). Beyond the scope of this paper are the findings of our Social Network Analysis (SNA) which confirmed these findings. Moreover, it may be that the term MOOC has particular currency only within particular communities such as the academic one. Some of the top tweets and URLs would appear to bear this out such as a link to a MOOC survey being conducted as part of an MSc. thesis – an item of as much interest to MOOC researchers as students. It is unknown how widely prevalent the term "MOOC" is in popular discourse and hence many MOOC students may go undetected. This may limit the value of using the term MOOC to make inferences about learners. Using other search constructs that would comprise course, platform, provider or some combinations of these might bring more learners into the dataset.

The sample of top tweets from the sentiment analysis does appear to show interesting data from MOOC learners however. All but one of these five tweets are from what we may infer to be a MOOC learner, or in one case prospective learner. The other tweet appears to be from a MOOC commentator/researcher. Of course researchers may also be MOOC students. Research has shown that MOOC learners have disproportionally high levels of educational attainment (Jordan, 2014). This is borne out here in that one of the sample tweets from the sentiment analysis is from well-known academic relating to a MOOC about "surviving" PhDs. Our findings suggest there may be a value in using sentiment analysis to filter a Twitter dataset before performing other types of analyses for researchers. For instance, it can be seen from a visual scan that peak tweets which are informational (and promotional) are relatively lacking in or have weak positive sentiment. This requires further analysis.

This paper has outlined the techniques we used and the theoretical basis by which we adopted these approaches in examining MOOCs in a Twitter dataset. We used descriptive and content analysis techniques to probe a sample of tweets using the hashtag #MOOC. Our results pose perhaps more questions that give definitive answers but we contribute by conducting exploratory analyses in an underexplored area namely research on MOOC actors using large Twitter datasets.

References

Abeywardena, I. S. (2014). Public opinion on OER and MOCC: A sentiment analysis of Twitter data. *Proceedings of the International Conference on Open and Flexible Education* (ICOFE 2014), Hong Kong SAR, China.

Chae, B. K. (2014). A complexity theory approach to IT-enabled services (IESs) and service innovation: Business analytics as an illustration of IES. Decision Support Systems, 57, 1–10.

Chae, B. K. (2015). Insights from hashtag# supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research. *International Journal of Production Economics*, *165*, 247-259.

Driscoll, K., & Walker, S. (2014). Big data, big questions| working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*, 20.

Du, P., Kibbe, W. A., & Lin, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics*, *22*(17), 2059-2065.

Healy, P., Hunt, G., Kilroy, S., Lynn, T., Morrison, J. P., & Venkatagiri, S. (2015, November). Evaluation of peak detection algorithms for social media event detection. In *Semantic and Social Media Adaptation and Personalization (SMAP), 2015 10th International Workshop on* (pp. 1-9). IEEE.

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(1).

Lehmann, J., Gonçalves, B., Ramasco, J. J., & Cattuto, C. (2012, April). Dynamical classes of collective attention in Twitter. *Proceedings of the 21st international conference on World Wide Web* (pp. 251-260). ACM.

Lynn, T., Healy, P., Kilroy, S., Hunt, G., van der Werff, L., Venkatagiri, S., & Morrison, J. (2015). Towards a general research framework for social media research using big data. *Proceedings of 2015 IEEE International Professional Communication Conference (IPCC)* (pp. 1-8). IEEE.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice. In *SSHRC Knowledge Synthesis Grant on the Digital Economy.* Retrieved from http://www.edukwest.com/wp-content/uploads/2011/07/MOOC_Final.pdf [viewed 08 July 2106]

Myers, S. A., Sharma, A., Gupta, P., & Lin, J. (2014). Information network or social network?: the structure of the Twitter follow graph. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 493-498). ACM.

Palshikar, G. (2009). Simple algorithms for peak detection in time-series. *Proceedings of 1st International Conference of Advanced Data Analysis, Business Analytics and Intelligence* (pp. 1-13).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Rao, A., Spasojevic, N., Li, Z., & Dsouza, T. (2015). Klout score: Measuring influence across multiple social networks. In Big Data (Big Data), 2015 IEEE International Conference on (pp. 2282-2289). IEEE.

Zhang, J., Perris, K., Zheng, Q., & Chen, L. (2015). Public Response to “the MOOC Movement” in China: Examining the Time Series of Microblogging. *The International Review of Research in Open and Distributed Learning*, *16*(5).

Please cite as: Costello, E., Binesh, N., Brown, M., Zhang, J., Giolla-Mhichíl, M.N., Donlon, E. & Lynn, T. (2016). Social Media #MOOC Mentions: Lessons for MOOC mentions from analysis of Twitter data. In S. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds.), *Show Me The Learning. Proceedings ASCILITE 2016 Adelaide* (pp. 157-162).

Note: All published papers are refereed, having undergone a double-blind peer-review process.