On the potential for detecting scientific issues and controversies on Twitter: a method for investigating conversations mentioning research

David Gunnarsson Lorentzen, Johan Eklund, Björn Ekström and Gustaf Nelhans

{david.gunnarsson_lorentzen, johan.eklund, björn ekström, gustaf.nelhans} @hb.se
University of Borås, Swedish School of Library and Information Science, S-501 90 Borås (Sweden)

Abstract

In this study, we demonstrate how to collect Twitter conversations emanating from or referring to scientific papers. We propose segmenting the conversational threads into smaller segments and then compare them using information retrieval techniques, in order to find differences and similarities between discussions and within discussions. While the method still can be improved, the study shows that it is possible to collect larger conversations about research on Twitter, and that these are suitable for various automated methods. We do however identify a need to analyse these with qualitative methods as well.

Introduction

The purpose of this paper is to propose a method for mapping issues within conversations on Twitter which in various ways refer to or mention scientific publications. The study builds on work done by Nelhans and Lorentzen (2016), who utilised the Twitter Streaming API to collect tweets including a reference to a digital object identifier (DOI) and the most active users in the collected dataset. By filtering the stream using the combination of search terms and users, they managed to mine conversational threads with references to DOIs. This stands apart from other means of identifying and extracting Twitter conversations that rely on hashtags for identifying tweets, which tend to miss large parts of the conversations. As a second step, this study also takes into consideration that conversations sometimes are divided into segments where new topics emerge. Through the identification of "bifurcations", i.e., parts of threads where Twitter conversations tend to take new directions, segments of a Twitter conversation can be partitioned off and treated as a coherent text to analyse. Such a treatment of parts of the conversation arguably makes it possible to find whether different issues are discussed in different parts of the thread or different perspectives on the same issue can be identified.

Building on the notions of issue mapping as an empiricist digital method for controversy analysis (Marres, 2015) we explore different network analysis-based techniques to identify, segment and measure user interactions conceived of as issues/controversies in Twitter conversation threads. In a paper presenting a checklist for the application of digital methods, Venturini et al. (2018) emphasised issues such as how the platform affords research, to what extent the study object plays out on the studied platform and if we are studying the object as it appears on the platform or if we use it as a proxy (e.g. for the public discourse). Given this, it is important to be aware of the affordances of Twitter at the time of the study. How is it possible to interact and how are conversations presented to the user? And how is it possible for a researcher to collect data? Moreover, we might view the interactions of Twitter as part of public discourse, but not a representation of it. By grounding the findings elsewhere (e.g. Rogers, 2013), we might get closer to a representation of the public discourse.

Through the analysis of Twitter conversations emanating from or including at least one reference to an academic paper, this study aims to further the understanding of the structure and content of Twitter conversations in the context of using them to identify the societal impact of research.

Literature review

The collection of Twitter data in the research literature has been mainly based on either hashtag-based or user-based methods. These methods only use tweets that contain a specific hashtag or keyword to identify the topic or limit the data collection to a set of users. How much of the conversations that are omitted by such methods has to our knowledge only been explored by D'heer et al. (2017) who saw their dataset of 1,719 tweets include 580 non-hashtagged replies and Lorentzen and Nolin (2017) who found an increase of 56 per cent new tweets through the inclusion of non-hashtagged tweets. Although the extent of the missing data will vary from topic to topic, using only hashtag or user-based data collection methods will inevitably render the data incomplete for a full understanding of the actual contents of the discourse.

While hashtags in a tweet can be compared to keywords in a scholarly article (Haunschild et al. 2018), at the same time, replies to, or retweets of other tweets, as well as mentions of a link (e.g. to a DOI) function as "internal" or "external" references, respectively (Haustein et al. 2014), thus corresponding to scholarly references. Mentions of another Twitter user (the @handle) does not have a clear corresponding function but serves both to signal an intended respondent as well as a means for highlighting interaction for this user, who would see an activity indicator in their Twitter interface. These different interactional aspects of the Twitter conversation are used in this study to grasp issues, sometimes in the form of controversies, highlighting both the interactive aspects of Twitter activity around tweets related to published research as their contents.

Collecting and analysing conversations in this sense is not common in Twitter research. Apart from the aforementioned works, Moon, Suzor and Matamoros-Fernandez (2016) found threads in a user-based set collected by Bruns, Burgess and Banks (2016) by following 2.8 million Australian Twitter accounts. Their study focused on conversations emanating from or including the word "uber". They argued that working with larger parts of texts would permit more comprehensive analysis of public opinion around a controversy and that analysis of these conversational threads contributes to a better understanding of social media communication. Another example of analysis of Twitter threads is provided by Zubiaga et al. (2016), who identified several internet rumours and then scraped the Twitter web interface for follow-on conversation attached to given tweets. The threads were then manually annotated as to whether the tweet was a rumour or an attempt to resolve the rumour as true or false.

Within altmetrics and similar areas, the focus has not been on the content and structure of conversations, however, but rather to what extent tweets can be used as a proxy for scientific impact. Even in an article using the term "conversation networks" in its title, Holmberg et al. (2014) explicitly state that it is not the full communication network, but rather the pairwise conversational connections that they study. Focusing on publications produced by Finnish researchers, Vainio and Holmberg (2017) found that those who referred to articles on Twitter "describe themselves more factually and by emphasizing their occupational expertise rather than personal interests". Didegah, Bowman and Holmberg (2018) studied factors behind altmetric scores compared to citations. Of special interest here is what makes a tweet about a research publication successful. Research funding was found to be most important, but journal impact factor and international collaboration also contributed to an increased number of tweets. Discipline-wise, research within medicine, natural sciences, and engineering and technology were more often tweeted than its counterpart within social sciences and humanities.

Nelhans and Lorentzen (2016) in a previous explorative study used a set of conversational threads that mentioned DOI references on Twitter to gain an understanding of the characteristics of the interaction and objects of discussion. Using both quantitative and qualitative methods, the authors characterised the objects of all mentioned DOI during a one-month period on Twitter. It was found that during the collection period articles and reviews from predominantly English-speaking countries at prestigious universities were mainly mentioned. 80 per cent of the mentioned literature was published in the same year as the data collection (which was done during the month of April). Content-wise, mentioned literature was heavily focussed on health, medicine, and the life sciences, as well as a broad range of social science topics, ranging from gender and learning, social media, and artificial intelligence, by way of social medicine and studies of the human condition, suggesting that broad social issues, was at the focus of interest of the tweeting users. In the study, a qualitative analysis of tweet practices was performed by categorising distinct conversational properties as "kinds" based on what was mentioned and how a DOI-referenced article was referenced. Specifically, different modes of discussing the contents, communicating about the context and different conversational practices were identified. In conclusion, it was found that digital object identifier URLs were mainly used for promoting a paper, as a conversation starter or as arguments in a discussion.

To a certain degree, contrasting findings regarding the promotional aspects of tweets were presented by Vainio and Holmberg (2017) who were only able to detect such use of tweets for marketing in the humanities and social sciences. Since that study did not focus on the conversational aspects of tweets, but on the user profiles mentioning highly tweeted articles, different aspects of the conversational practices were not studied. From the above, it is concluded that the study of the conversational properties of Twitter activity around scholarly publications are still in its infancy. The contribution by this study would be directed to social network analysis (SNA) aspects of Twitter conversations by expanding on the thread analysis of the structure and delineation of parts of conversation and identification of issues within the Twitter stream thread.

Method

Data collection was performed through filtering the Twitter stream using keywords and the most active users in the collected dataset, similarly to Nelhans and Lorentzen (2016), but instead of focusing only on DOIs, the present study tracked the keywords "dx doi org", dx.doi.org, arxiv.org, socarxiv.org, researchgate and academia.edu. This means that we did not attempt to collect data related to a particular topic, but rather any potential topic or discipline. Data collection started on August 23 2018 and ended two weeks later. When the data collection was finished, there were tweets in the database replying to tweets that had not been collected, most of them expected to be posted before the data collection started. The IDs of these missing tweets were put in a list and then used to query the statuses/lookup API endpoint, and if the tweet was a reply the ID of the new tweet was added to the list. This procedure added almost 10,000 new tweets, resulting in a total of 29,796 tweets. As tweets are identified by an ID and a reply is denoted by the ID of a tweet replied to, we can then string tweets together as a conversational thread. This procedure yielded a set of threads that varied in length and number of users in highly skewed distributions. The longest thread consisted of 1,458 tweets whereas the mean was 8.79 and the median 3. The largest number of participants was 59, with a mean of 2.7 and a median of 2. As noted in Nelhans and Lorentzen (2016), Twitter threads can take many different forms, including a chain-like, star-like or heavily bifurcated form, meaning that many new interactions could be identified where the discussion takes new directions. For further analysis, we chose two threads including 595 and 1024 tweets respectively, the first involving 30 participants and the second 28.

As few examples of Twitter research on conversational threads exist, a relevant aspect to explore is how to identify metrics for the activities. Such metrics could, for example, include the conversational impact of a tweet. One example of how a metric for statistical analysis of discussion threads was presented by Gómez, Kaltenbrunner and López (2008, p. 652-653): “the h-index h of a post is [...] the maximum nesting level i which has at least h > i comments, or in other words, h + 1 is the first nesting level i which has less than i comments.” However, this metric does not suit the conversational threads found on Twitter, where many threads involve bridges between tweets that spark a reaction from many users. Hence, a thread might start with one tweet replied to ten times (10 tweets at level 2), then one of these tweets is replied to once, and the reply is replied to once (1 tweet at levels 3 and 4), before this subsequent reply triggers a reaction with many replies. Similarly, we cannot rely on the nesting level only. A chain of 100 tweets without bifurcations would end up with a maximum nesting level of 100. In this exploratory work, we propose the identification of relevant threads to study based on the number of bifurcations resulting in at least two branches which include at least a total of 30 tweets.

To partition a conversational thread into segments of sufficient size, we initially identified all the bifurcations in the threads, that is, tweets replied to more than once. For each bifurcation, we then traversed the tree and counted the number of tweets emanating from the bifurcation. Following this step, we had tweet counts representing the number of posts from the point of a bifurcation, to the end of each of its branches. In the next step, we traversed the tree back to the root from the outermost bifurcations containing between 30 and 50 tweets, starting at the part of the thread which included the tweet with the latest timestamp. When reaching a previous bifurcation, we assigned the tweets belonging to that bifurcation a new segment ID if at least 30 tweets had been encountered. Finally, we joined the segments with one or two (depending on the size of the segment) adjacent segments to make the documents more suitable for automated text analysis. For the two examples included in this paper, this resulted in segments of varying sizes. While nine of 16 segments included between 100 and 125 tweets, the lengths of the segments varied from 65 to 157 tweets. The two threads were compared to each other using the cosine similarity measure, and we then focused on the longer thread to find out if it could be feasible to compare the segments within a thread with the rest of the thread. The texts were processed using the Porter stemmer and stop words were removed. In order to illustrate the topicality of the threads and the differences between the least similar segments, we created density maps of the documents using VOSviewer (van Eck and Waltman, 2014). Finally, for a topical analysis of the segments, we used Latent Dirichlet Allocation, LDA (e.g. Blei, Ng and Jordan, 2013). As training an LDA model with few documents, in this case, nine, renders instability, we tokenised the segments into sentences and used the sentences as training documents. From this, the most likely topic for each segment is induced as a list of ten terms ordered by a probability score. As each run results in a different set of terms for the topics, we trained the model in ten iterations and subsequently kept the ten terms most often coupled with each segment across the iterations.

Findings

Both threads are similar in that they both include large bifurcations, and they stretch over a few days, although thread 649 actually starts with a reply to a tweet posted more than three years before. Thread 1282 has a few more hubs which are tweets with many replies, but only a few replies result in larger branches. Topic-wise, the threads are different. The cosine similarity score for the thread comparison was 0.37. Judging by the density maps (Figure 1 and 2), thread 649 is about vaccination and related issues whereas thread 1282 is about learning, teaching and knowledge. The fact that the two threads differ much from each other comes as no surprise considering that the data collection was not restricted to one topic.

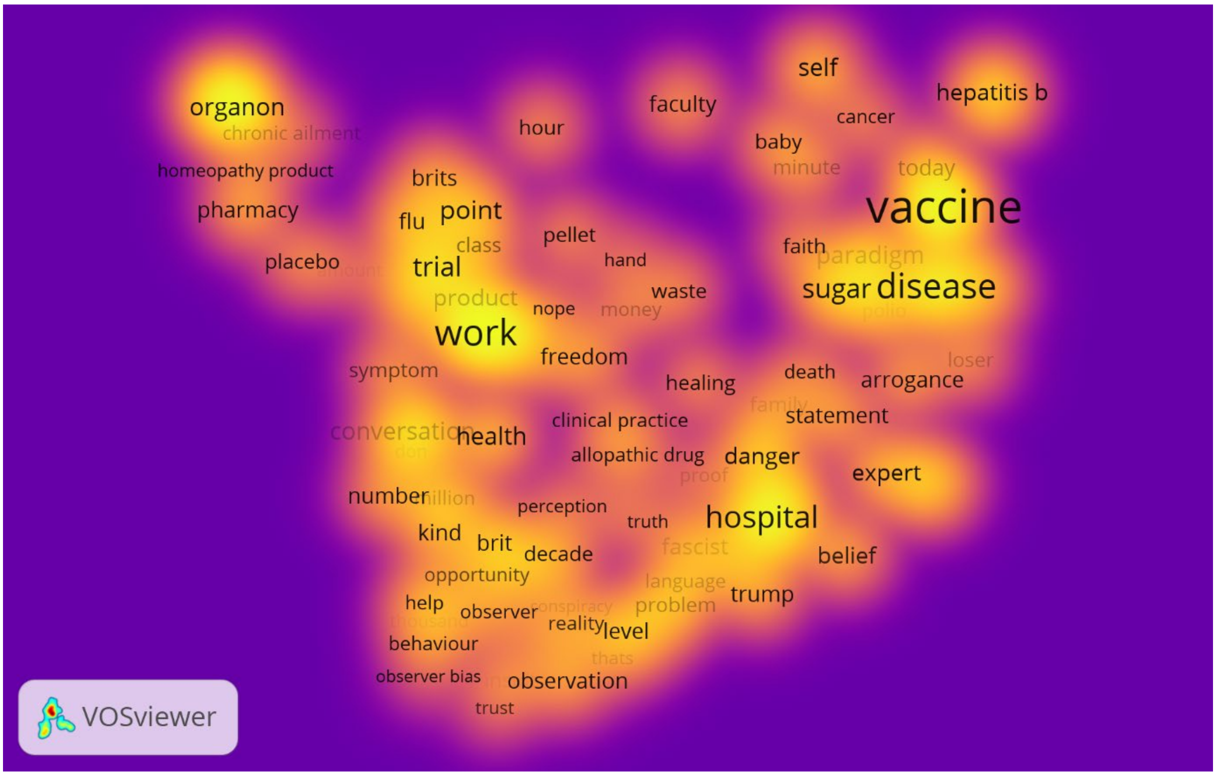

Figure 1. Density map of thread 649.

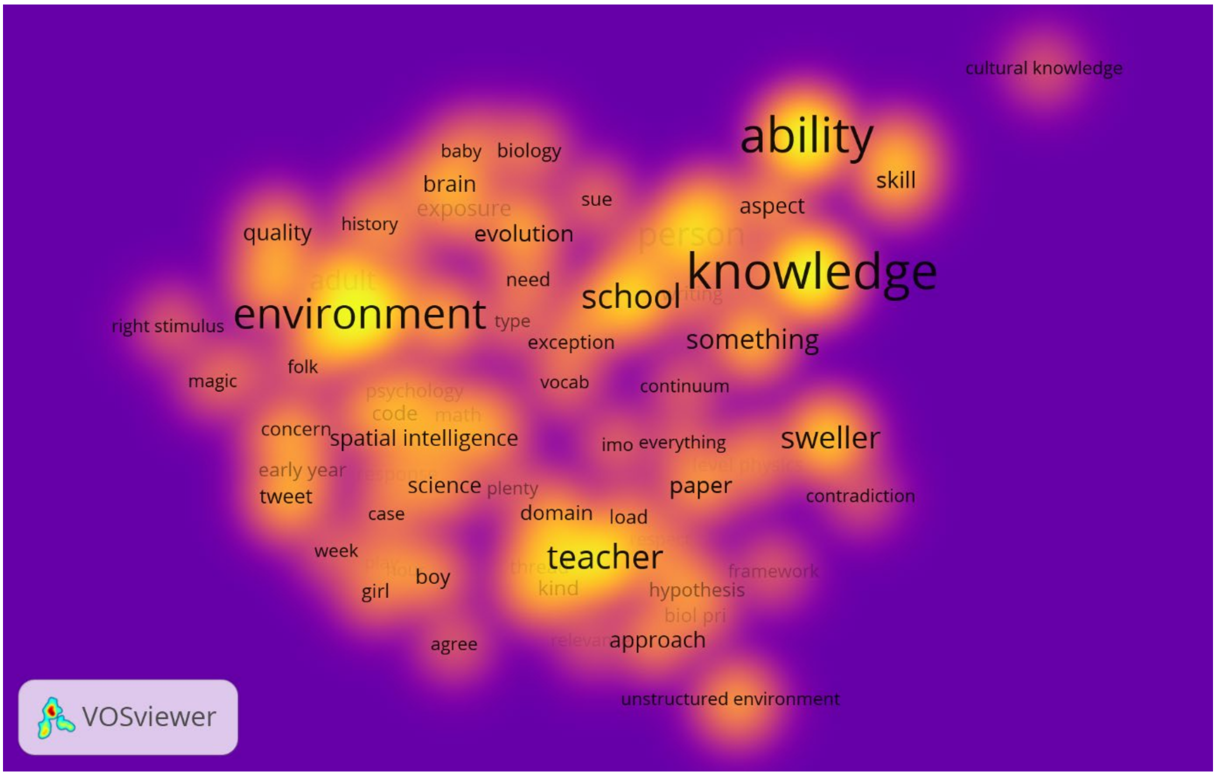

Figure 2. Density map of thread 1282.

We then focused on the longer of the threads for further analysis. The thread included many bifurcations which made it suitable for analysis of the topics within the discussion. Figure 3 shows its structure and the segments as identified by our algorithm. The arrow in the upper right corner points to the first tweet of the discussion.

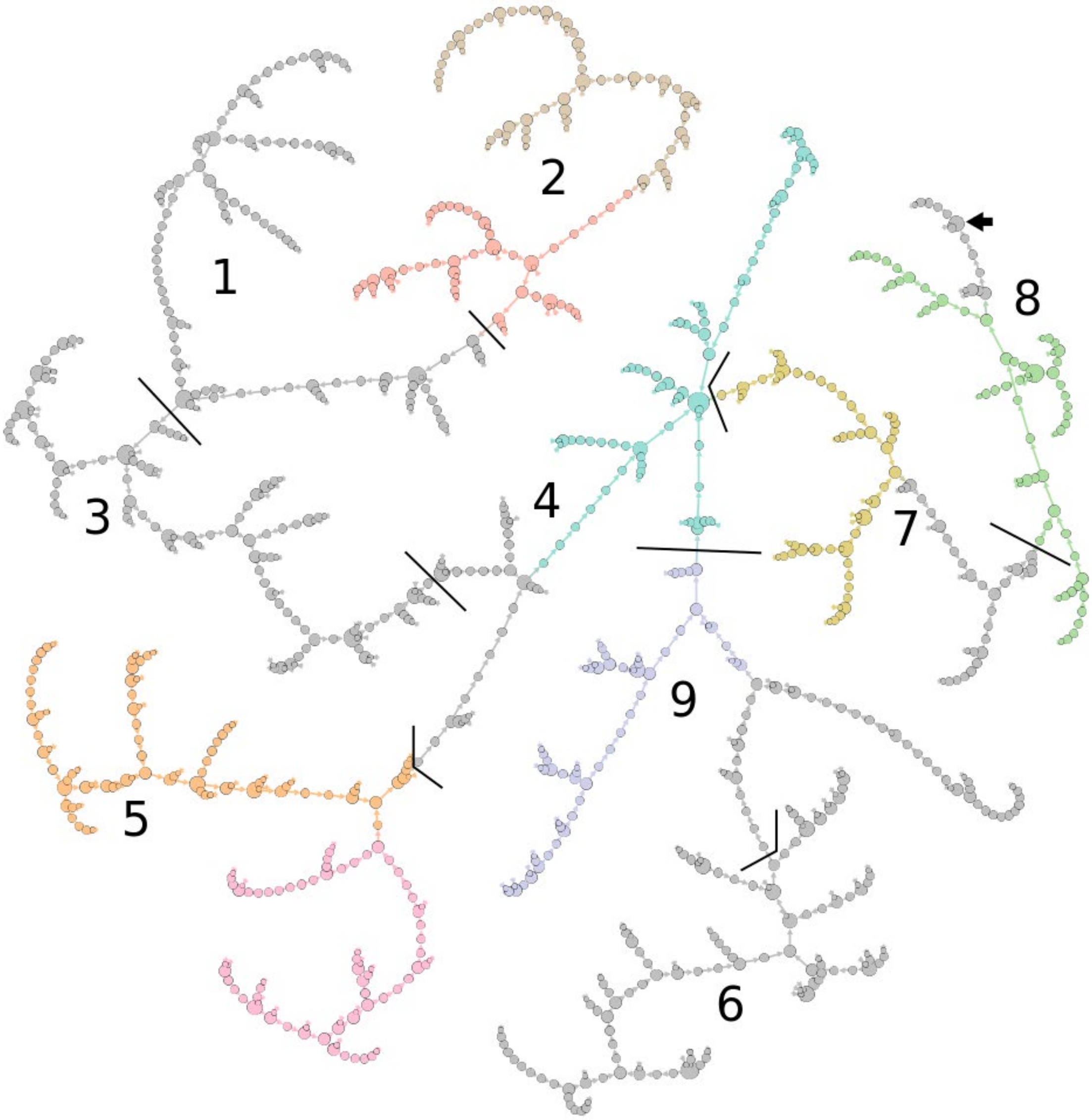

Figure 3. Thread 1282 with its segments. The lines denote where the thread is segmented. Nodes are sized according to the number of replies. The nodes are coloured according to clusters identified by the network analysis software Gephi.

The cosine similarity scores indicated that there were differences between the term frequency vectors representing the segments of the thread (Table 1). When comparing the different segments with each other, similarities were fairly low although they were still higher than the similarity between the threads. That is, the segments in thread 1282 differed less from each other than the two threads did. One hypothesis could be that smaller document sizes contributed to the lower scores as the segments deviated less from the thread when each segment was compared to the rest of the thread. With each segment treated as one document and the other eight segments as another document, the similarity scores were higher, ranging from 0.63 to 0.8.

When considering the top ten terms likely to be representative for each segment, we found numerous words in most of the segments, and it seems like the segments are quite similar to each other. Had the topic model resulted in lists of terms more distinct from each other, they could have been used as labels for the segments, however, in this case it seems as the segments do not differ much from each other according to the LDA model (Table 2).

Table 1. Cosine similarity scores between segments in thread 1282.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*Avg. sim.*	*Segment vs. thread*
1	1	0.40	0.41	0.36	0.47	0.44	0.38	0.39	0.35	0.47	0.63
2	0.40	1	0.47	0.37	0.49	0.43	0.43	0.43	0.40	0.49	0.69
3	0.41	0.47	1	0.41	0.50	0.48	0.44	0.41	0.47	0.51	0.73
4	0.36	0.37	0.41	1	0.48	0.47	0.62	0.40	0.42	0.50	0.67
5	0.47	0.49	0.50	0.48	1	0.56	0.50	0.47	0.47	0.55	0.80
6	0.44	0.43	0.48	0.47	0.56	1	0.44	0.37	0.47	0.52	0.72
7	0.38	0.43	0.44	0.62	0.50	0.44	1	0.46	0.46	0.52	0.72
8	0.39	0.43	0.41	0.40	0.47	0.37	0.46	1	0.36	0.48	0.65
9	0.35	0.40	0.47	0.42	0.47	0.47	0.46	0.36	1	0.49	0.70

Table 2. Ten most likely terms for each segment according to the LDA model.

Segment	*Ten most likely terms*
1	think, language, children, learn, teaching, learning, theory, instruction, words, geary
2	think, geary, learn, children, learning, teaching, instruction, child, language, taught
3	geary, think, children, instruction, learning, years, theory, language, reading, way
4	knowledge, ability, explicit, primary, taught, biologically, instruction, think, reading, children
5	think, learning, language, geary, words, see, explicit, instruction, speech, teaching
6	think, language, geary, teaching, child, environment, children, point, speech, spoken
7	geary, think, instruction, reading, explicit, primary, words, said, learn, speech
8	think, instruction, explicit, learning, knowledge, primary, speech, biologically, read, teaching
9	geary, think, learn, teaching, learning, language, children, speech, different, instruction

Words such as “think”, “learning”, “language”, “child”, “children” and the researcher David Geary are occurring at the top end of the terms in multiple segments. This is an interesting finding in itself, as one would expect that the participating user cannot overview the entire thread, but this one stays on the same topic. Incidentally, the segments most similar according to the cosine similarity score are the neighbouring segments 4 and 7. However, the segment that deviates most from the rest of the discussion (1), is also the one farthest away from the start, i.e. the bifurcation with the latest timestamp. The next segment with the lowest similarity with the rest of the thread is the thread start (8). These two segments are also deviating more from the other parts in our segment vs. segment analysis, and they have a fairly low similarity score too. Rather than focussing on the top terms it would be more relevant to focus on the unique words. For example, “biologically” is included in two segments and “environment” in one. These three segments might reveal a different topic than the rest of the thread, if analysed with manual methods, such as quantitative or qualitative content analysis. Seemingly, based on the results of the analysis of these two threads, the method was able to identify what issues were discussed, but as the segments were found to be quite similar, signs of controversies were not detected. What we do not see here, given these analysis methods, is if the thread bifurcates because of some kind of disagreement. Although a bifurcation does not seem to imply a topical shift, other methods might reveal disagreements within the segments, which could be a sign of controversy.

Discussion

We have presented a method for automatic analysis of conversations on Twitter, emanating from or referring to a research publication. We propose dividing a conversational thread into segments where the thread bifurcates. With information retrieval techniques such as the cosine similarity measure and LDA modelling, these segments can be compared with each other. While this measure did highlight differences, a further adaptation for the type of content produced on Twitter is highly recommended. Such adaptation includes the use of a specialised stop word list. Another issue is to train the topic model on a larger body of texts and not just the one thread containing the shorter segments as documents. We would also wish to stress the need for improving the algorithm for segmenting the thread into smaller parts, and an investigation into the optimal size of the segment for automatic text analysis. Considering the article on digital methods and controversy studies by Marres (2015), we conclude that the method presented here is useful for identifying possibly controversial issues as they are discussed on Twitter, but that they then need to be analysed qualitatively or with more sophisticated machine learning methods. For example, a similar approach as the one taken by Buntain and Golbeck (2017), who used a feature-based method for automatic detection of fake news, could be adapted and applied to these threads. Furthermore, standing alone, an analysis of Twitter conversations does not say much more than how people interact on this platform. Grounding the findings in other analyses of other types of conversations is recommended.

While limited, the analysis of Twitter conversations regarding research articles does provide an indication of what type of research a part of the public is interested in, how it is referred to and how it is used as arguments in the discussions. It has for example been found that academic papers also are referred to for promoting ideological views (e.g. Vainio & Holmberg, 2017). We recommend further analysis of this by taking a more comprehensive approach. If focussing on Twitter, we suggest to collect data so that next to complete conversations can be studied, implementing methods similar to those presented here to identify and map possible issues or controversies, and then take the process one step further with an analysis of how the interactions play out and how research is used in the public domain. Particularly of interest would be to investigate the level of disagreement within a branch as well as among the branches.

Finally, we must acknowledge a couple of limitations to this study that should be addressed in future endeavours of this kind. Firstly, the selection of keywords should include “doi.org” as the prefix dx is not needed. Secondly, it is important that the researcher is aware of the presence of bots for further analyses of the discussions, an issue that has been discussed previously (i.e. Haustein et al., 2016; Robinson-Garcia et al., 2017). While we stress that material from bots must be included so that threads are not broken, including bot detection algorithms to present the likelihood that a tweet is posted by a bot would be an important contribution. For example, in an experiment comparing different machine learning algorithms, Haidermota, Mitra and Pansare (2018) concluded that bots seem to be more predictable regarding the timing of the reply to a tweet, while other indicators could be helpful, for example follower counts and usage of URLs. Another interesting option is to make use of the application *Botometer*, previously known as *Botornot* (Davis et al., 2016). If we can identify bots prior to the conversation analysis, then we can also learn more about how bots participate in Twitter discussions, as well as how other users interact with them.

Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 770531.

References

Buntain, C., & Golbeck, J. (2017). Automatically Identifying Fake News in Popular Twitter Threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud),* New York, USA, Nov 3-5. IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research. 3 (4–5)* (pp. 993–1022).

Bruns, A., Burgess, J., & Banks, J. (2016). *TrISMA: Tracking Infrastructure for Social Media Analysis*. Retrieved May 27, 2019 from: https://trisma.org.

D’heer, E. et al. (2017). What are we missing? An empirical exploration in the structural biases of hashtag-based sampling on Twitter. *First Monday, 22*(2). Retrieved May 27, 2019 from: https://firstmonday.org/ojs/index.php/fm/article/view/6353/5758.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). Botornot: A system to evaluate social bots. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273-274).

Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for Finnish articles. *Journal of the Association for Information Science and Technology, 69*(6) (pp. 832-843).

Gómez, V., Kaltenbrunner, A., & López, V. (2008). Statistical analysis of the social network and discussion threads in Slashdot. Paper presented at the *WWW '08 Proceedings of the 17th international conference on World Wide Web.* Beijing, China, April 21-25 (pp. 645-654).

Haidermota, M., Mitra, i., & Pansare, A. (2018). Classifying Twitter user as a bot or not and comparing different classification algorithms. *International Journal of Advanced Research in Computer Science*, 9(3) (pp. 29-33).

Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2018). Does the public discuss other topics on climate change than researchers? A comparison of networks based on author keywords and hashtags. Retrieved May 27, 2019 from: https://arxiv.org/abs/1810.07456 (pp. 24).

Haustein, S., Bowman, T. D., Holmberg, K., Peters, I., & Larivière, V. (2014). Astrophysicists on Twitter: An in-depth analysis of tweeting and scientific publication behaviour. *Aslib Journal of Information Management*, 66(3) (pp. 279-296).

Haustein, S., Bowman, T., Holmberg, K., Tsou, A., Sugimoto, C., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated “bot” accounts on Twitter. *Journal of the Association for Information Science and Technology*, 67(1) (pp. 232–238).

Holmberg K, Bowman TD, Haustein S, Peters I (2014) Astrophysicists' Conversational Connections on Twitter. *PLoS ONE 9*(8): e106086.

Lorentzen, D. G., & Nolin, J. (2017). Approaching Completeness: Capturing a Hashtagged Twitter Conversation and its Follow-On Conversation. *Social Science Computer Review, 35*(2) (pp. 277-286).

Marres, N. (2015). Why Map Issues? On Controversy Analysis as a Digital Method. *Science, Technology, & Human Values, 40*(5) (pp. 655-686).

Moon, B., Suzor, N., & Matamoros-Fernandez, A. (2016). Beyond Hashtags: Collecting and Analysing Conversations on Twitter. Paper presented at *AoIR 2016: The 17th Annual Meeting of the Association of Internet Researchers.* Berlin, October 5-8. Germany: AoIR.

Nelhans, G. & Lorentzen, D. G. (2016). Twitter conversation patterns related to research papers. *Information Research, 21*(2), paper SM2. Retrieved May 27, 2019 from: http://www.informationr.net/ir/21-2/SM2.html.

Robinson-Garcia, N., Costas, C., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting - about journal articles. *PLoS ONE*, 12(8), e0183551.

Rogers, R. (2013). *Digital methods*. Cambridge, Massachusetts: The MIT Press.

Vainio, J. & Holmberg, K. (2017). Highly tweeted science articles: who tweets them? An analysis of Twitter user profile descriptions, *Scientometrics, 112*(1) (pp. 345.366).

van Eck, N.J. & Waltman, L. (2014). Visualizing bibliometric networks. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring scholarly impact: methods and practice*. Berlin: Springer (pp. 285-320).

Venturini, T., Bounegru, L., Gray, J., & Rogers, R. (2018). A reality check(list) for digital methods. *New Media & Society, 20*(11) (pp. 4195–4217).

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE 11*(3): e0150989.