Influential tweeters in relation to highly cited articles in altmetric big data

Saeed-Ul Hassan[1] · **Timothy D. Bowman**[2] · **Mudassir Shabbir**[1] · **Aqsa Akhtar**[1] · **Mubashir Imran**[1] · **Naif Radi Aljohani**[3]

Abstract

The relationship between influential tweeters and highly cited articles in the field of information sciences was analysed using Twitter data gathered by Altmetric.com from July 2011 through February 2017. The dataset consists of more than 10,000 tweets, and these mentions, retweets and followers were used to generate a connected, undirected graph. This graph reveals the most influential tweeters by identifying the largest drop in the eigenvalue of adjacency or affinity matrix of a graph when certain nodes are removed; those which, when deleted, cause the greatest drop in the eigenvalue of the graph are considered to be the most influential. The machine-learning model applied in this work utilizes a feature vector containing the accumulated sum of the rank scores of those influential users who tweet a given article, along with known altmetric features such as the user type and post counts for various social media. Finally, the supervised-learning model was trained using Random Forest and Support Vector Machine classifiers with 11 features, including the sum of the ranks of influential users who tweet a given article in our dataset. The results were analysed using Receiver Operating Characteristic (ROC) curves and Precision Recall (PR) curves, which give the commendable outcomes compared to the baseline model. We found that, for the classification of highly cited articles, Twitter users' score for influence is the most important feature. Finally, we show that our model—which was trained by taking the score for influence into consideration—outperforms the baseline, at 79% for ROC and 90% for PR with the Random Forest Model, effectively identifying the highly cited articles.

Keywords Altmetrics · Influential users · Twitter · Highly cited articles

✉ Saeed-Ul Hassan
saeed-ul-hassan@itu.edu.pk

Timothy D. Bowman
timothy.d.bowman@wayne.edu

Naif Radi Aljohani
nraljohani@kau.edu.sa

1 Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

2 School of Library and Information Science, Wayne State University, Detroit, MI, USA

3 Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

Introduction

It has been reported that, as of January 2017, approximately 2.8 billion people actively use social media worldwide (Kemp 2017). Social media sites such as Twitter, Facebook and YouTube serve as important platforms for people to communicate, interact, consume and disseminate information (Kalloubi et al. 2017).

Twitter is the most popular microblogging social media site of all the current online social media applications; it has 328 million active users per month (Escamilla et al. 2016). There are several features that the user can utilize and track. Twitter is a popular social media site for researchers to investigate, as the company allows developers to collect a portion of tweets freely through their API (application programming interface). Twitter data have been collected and analysed by various scientific studies, from predicting election results (Beauchamp 2017), surveying public health (Bates 2017) and identifying community sentiments corresponding to specific events (Jarwar et al. 2017) to analysing news reports during important events such as the Egyptian Arab Spring revolution (Lotan et al. 2011). In addition, Twitter has become a prominent platform used in altmetrics research to track the activities surrounding scientific documents. A recent study of over 1.1 million publications (with at least one citation count) shows that Twitter's coverage exceeds 91% of the total social media activities received by altmetrics (Hassan et al. 2017).

Given the vital role of Twitter's platform in generating altmetrics-related data, it is extremely important to learn more about its users who are sharing, liking, disseminating and communicating information related to scientific documents. Certain Twitter accounts have more influence in this context (Quercia et al. 2011), perhaps because the account holders are celebrities, politicians, social workers or companies and thus have more followers than others. Posts by these popular Twitter users may influence or impact on others within the network more than most (Hussain et al. 2012). Thus, in the context of altmetrics, it becomes extremely important to identify who, or what (e.g. journal, bot, lab, etc.), is tweeting about science.

The main objective of this paper is to detect the influential Twitter accounts among the tweeters in a certain discipline and to investigate their influence on the accumulation of scholarly citations. The research questions are as follows: (a) Who are the most influential tweeters in the field of information sciences during July 2011 to December 2015 (present in Altmertic.com version dataset-jun-4-2016.tar.gz)? (b) Are influential tweeters capable to discriminate highly cited articles from non-highly cited articles? and (c) More specifically, what is the role of influential tweeters in terms of discriminating highly cited articles?

The rest of the paper is organized as follows: the following section presents a literature review of altmetrics-related work. Next, we present our approach to the identification of influential tweeters, followed by the results and discussion section. Finally, we present our concluding remarks, along with future research directions.

Literature review

We review the literature in two parts: first, we make a brief review of altmetric studies, showcasing the importance of social media platforms, including Twitter, in disseminating research outputs. Second, we review the important studies on identifying influential tweeters or tweets on Twitter platform, in chronological order.

A brief review of altmetrics as an early predictor of citations

During the initial phase of altmetric research, the focus of many studies was finding out whether social media acts correlated to citations. Priem et al. (2011) studied the potential of altmetric data using linear regression to assess if altmetric counts contributed to the prediction of Web of Science citation counts. The authors found that altmetric events were a contributor to citation predictions. More recently, Costas et al. (2015) and Hassan et al. (2017) provided an extensive analysis of altmetric indicators and their relationship to citations across multiple disciplines; they found a weak correlation between citation and altmetric counts and that the number of altmetric acts for scientific publications were still catching up. Similar observations were made in further works, including by Thelwall et al. (2013), Sud and Thelwall (2014), Haustein et al. (2014) and Zahedi et al. (2014). Thus, altmetric researchers have turned their attention to the broader impact of research outside academia (Bornmann 2014, 2016; Bornmann and Haunschild 2016). To answer these questions, researchers must go beyond examining correlations to citations, and must start to examine who and why actors (both individuals and organizations) are sharing and discussing scientific publications on social media platforms.

In addition to these studies, others (see Sugimoto et al. 2017, for a thorough review of the literature) have examined various topics, including the impact on scholarly communication from sharing scientific documents in online environments (Shrivastava and Mahajan 2016), the identification of communities engaging with scientific documents online (Tsou et al. 2015), the variation in activity across online platforms (Haustein 2016) and the formulation of a theoretical lens with which to view altmetric activities (Haustein et al. 2015).

A brief review of detecting influential Twitter users

In this section, we present a brief review of studies pertaining to the identification of influential users or influential tweets on Twitter.

Alonso et al. (2010) made use of crowdsourcing, along with machine-learning algorithms, to find the *interesting* tweets in a batch of a thousand tweets. With the help of workers on Amazon Mechanical Turk (AMT), who took part in the evaluation procedure by labelling the tweets as '*interesting'* or *'not interesting'*, they analysed whether the presence of hyperlinks plays an important part in such classification. By training the data using a decision tree classifier, the model classified 85% tweets correctly; about 15% were misclassified. Anger and Kittl (2011), working on a dataset of Australian-based users, introduced a new measure of *social network potential* (SNP) to discover the most influential users. The authors analysed four features (*retweet, mention, follower* and *following*) as key factors to label users as influential or non-influential. They utilized the score from Klout, an online web-based service that tells users about their influential measures in social media sites, alongside three other major parameters that they defined as *Follower/Following Ration (rf)*, *Retweet and Mention Ratio (rRT)* and *Interactor Ratio (ri)*. They found that the mention and retweet ratios were of great importance when focusing on content-oriented interactions, while the interaction ratio was important in the case of conversation-oriented interactions. The SNP value was calculated using the sum of the *rRt* and *ri* ratios divided by two, where a result of 100% meant that all the tweets of a user were either mentioned or retweeted.

Yang et al. (2012) took a different approach to the identification of influential tweets on Twitter by focusing on the ranking of interesting tweets. In their analyses, users and tweets were considered nodes and the retweet relationship was considered an edge between these nodes; this approach differed from the standard approach of using HITS, as it was more influenced by the retweet behaviour of users. They found that HITSprop demonstrated even better results than standard HITSorig algorithm, and concluded that user authority is an important component of determining interesting tweets. More recently, Lee et al. (2017) used eigenvectors to find the influential users within a network. Working in the field of digital humanities, they formed a network of tweeters using the official AoIR (Association of Internet Researchers) conference hashtags from 2014, 2015 and 2016. Using degree centrality, PageRank and eigenvectors, they found the most influential users in the network. The authors then segmented tweets in the form of replies, mentions and tweets. They found that mentions represented the highest percentage of tweets, indicating that members of the AoIR used Twitter primarily to converse among themselves at conferences.

The above review summarizes scientific studies in two ways: first, we discussed the literature related to altmetrics as an early predictor of scholarly citations, then we presented studies pertaining to the identification of influential tweeters or tweets. To the best of our knowledge, no existing studies comprehensively discuss the role of influential tweeters in relation to scholarly citations. Thus, our work contributes significantly to a long debate dating back to the manifesto on determining the relationship between social mentions and scholarly citations that was initiated by Priem et al. (2010).

Data and methodology

Dataset

The dataset[1] used in this paper was obtained from Altmetric.com (version dataset-jun-4-2016.tar.gz). It contains over 4.5 million records, and the publications belong to various disciplines. Each record contains traces of the altmetric acts of a single scientific publication (article or dataset), as well as bibliometric information (such as DOI, number of authors, publication date, journal name, etc.). For this research, we selected publications from journals indexed in Scopus under the sub-discipline of Library and Information Sciences. Note that Scopus makes use of the All Science Journal Classification (ASJC) scheme to index journals. The ASJC classification maps journal and conferences across 27 broader disciplines, such as Agricultural and Biological Sciences, Chemistry, Computer Science and Social Sciences, along with more than 300 sub-disciplines including Artificial Intelligence, Human Computer Interaction, Safety Research, and Library and Information Sciences.

A total of 820 journal and conference publications (with at least one citation and associated tweet activity), indexed under Library and Information Science, were retrieved from the Scopus database during July 2011 to December 2015. Furthermore, all the tweets associated with these publications were procured. A combination of tweet ids (from Altmetric.com) and the Twitter API was then used to collect tweet

---

[1] The data and code used in this research can be downloaded at the following URL: https://github.com/slab-itu/influential_twitter_users/.

Table 1 Descriptive statistics of the Twitter dataset

Description	Value
Total number of articles	820
Retweets	4061
Mentions	8373
Total unique tweeters	5490
Total number of tweets	10,345
Publication time window	July 2011 to December 2015
Tweets and citation time window	July 2011 to February 2017

text, resulting in a dataset of 10,345 tweets made by 5490 unique users. These tweets contained 8373 mentions and 4061 retweets. All tweet information was then stored in a relational database. A summary of the retrieved Twitter dataset is shown in Table 1.

Identification of influential Twitter users

The first step was to identify the relationships between users by means of a social network graph. In our network, nodes represent Twitter users and edges represent three types of associations: *mentions*, *retweets* or *follow* relationships. The resulting undirected simple graph is represented by an adjacency matrix. A small subset of this graph is plotted in Fig. 1, and the mass of each node is indicative of its influence in the network. We found that institutional accounts, such as figshare, write4Research, SAGElibrarynew, and so on, lead in the chosen field of Library and Information Sciences.

It is well known that the network connectivity parameters and the spectral properties of the adjacency matrix of the corresponding graph are correlated. In particular, Chakrabarti et al. (2008) and Chung (1997) have reported that the maximum eigenvalue of a graph adjacency matrix is a good measure of its connectedness. This idea has been exploited by others to identify crucial nodes in a graph; if the removal of a node results in a significant difference in the largest eigenvalue, it may imply that the removed node was an influential entity in the original network. We used this observation and related algorithms in the literature to identify influential users in the Twitter social network who were tweeting about a particular set of publications.

For a given network $G=(V, E)$ with adjacency matrix A, the subset S of V with $\underset{S \in \binom{V}{k}}{\operatorname{argmax}} \left[\lambda_1(A) - \lambda_1(A[-S]) \right]$ where $\binom{V}{k}$ is the set of all k-subsets of vertices. Once rows and columns representing the nodes in the set S are removed, $A[-S]$ is the updated adjacency matrix and $\lambda_1(X)$ is a largest eigenvalue of a matrix X. Note that $\lambda_1(X)$ may not be unique, though this bears no relevance to the problem presented here.

The straightforward algorithm to solve this problem of finding k most influential nodes takes $O(n^k \cdot n^\sigma)$ time, where $O(n^\sigma)$ is the running time, to evaluate the spectrum of an n x n matrix. This running time is quite impractical for all useful values of k. Indeed, the problem was shown to be NP-Complete by a simple reduction from minimum vertex. To address this, an efficient approximation algorithm presented by Tariq et al. (2017) was used. For the sake of completeness, their algorithm/approximation technique is summarized below.
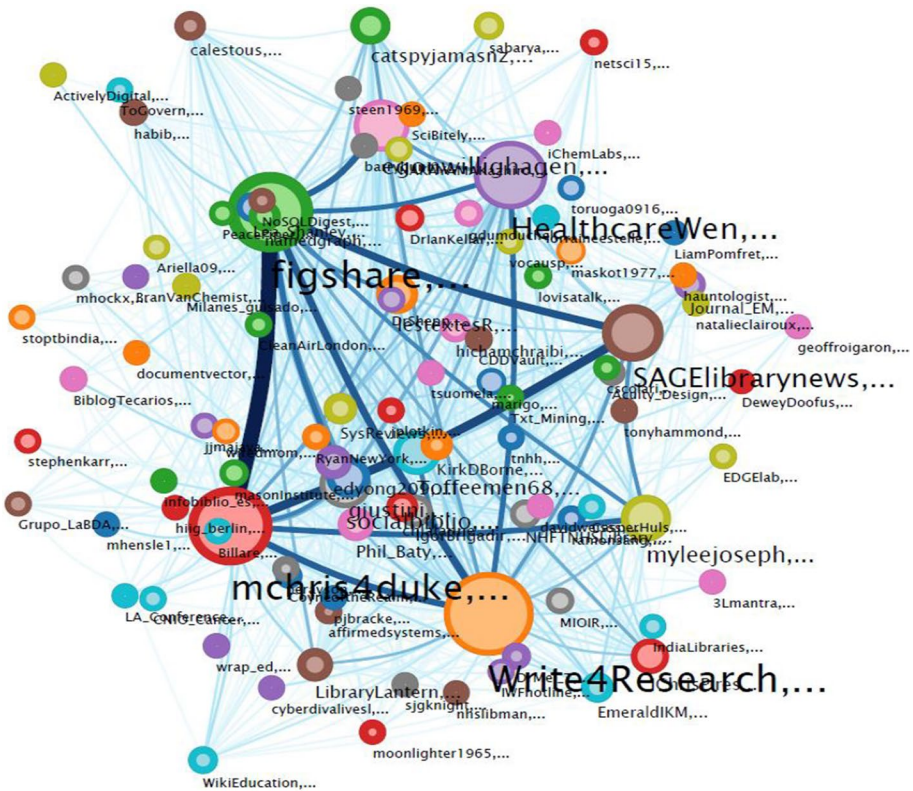
Fig. 1 Connection between tweeters in our dataset

Given an undirected graph $G=(V, E)$ with adjacency matrix A, associate a score $\psi = \mathrm{tr}(A^p) - \mathrm{tr}(A[-S]^p)$ with a subset S of vertices. Here p is a suitable constant (the larger the better, but larger is more time consuming) and tr is the classical trace function, which is defined in Eq. 1.

$$\mathrm{tr}(X) = \sum_{i=1}^{n} X[i, i] \tag{1}$$

It turns out that $\mathrm{tr}(A^p)$ is just the count of closed walks of length p in the given graph. Recall that a closed walk is a sequence $v_1, v_2, v_3 \dots v_l$ where $v_1 = v_l$ and each consecutive pair is an edge; that is, $v_i, v_{i+1} \in E$. Based on other results in graph theory, the following approximation is used, which can be computed for each vertex individually, as shown in Eq. 2.

$$\psi_G(v)' = 2d_G^2(v) + 4\left(\sum_{u \neq v} d_G(u, v)\right)^2 \tag{2}$$

Here, $d_{\mathrm{G}}(v)$ is the *degree* of the vertex v in the graph and $d_{\mathrm{G}}(u, v)$ is the *co-degree* of vertices u and v. The degrees $d_{\mathrm{G}}(v)$ are readily available and are traditionally stored along the diagonal entries of the adjacency matrix, while co-degrees $d_{\mathrm{G}}(u, v)$ can be computed by taking the intersection of characteristic vectors of corresponding rows in the matrix. Hence, the score $\psi_{\mathrm{G}}(v)'$ can be computed in linear time from the number of vertices and edges, which provides a measure of influence of a node in the graph.

Experiments and results

In this section, we use descriptive analysis and classification modelling techniques to highlight the importance of influential Twitter users in distinguishing highly cited articles. We divided the dataset of 820 articles into the top 50% articles with reference to their citation counts and labelled them as '1', while the rest of the articles were given the label '0'.

Descriptive analysis

The distribution of our data is revealed here through their mean and median. We next constructed a histogram of our data to analyse the relationship between the influential Twitter users and the number of citations in publications, both for the top 50% highly cited articles (HC) and the rest.

Table 2 shows that top 50% (HC) articles have a higher mean and median than the rest. This indicates that articles tweeted by influential users receive more citations. We found that the top 50% (HC) articles have, overall, greater mean and median scores than the total mean and median of the complete dataset. Furthermore, the histogram of user scores for both those publications in the top 50% (HC) and the rest (see Fig. 2) shows that for HC articles the sum of user scores is more spread out, indicating that articles tweeted by influential users achieve more citations than the rest, for which the sum of the user scores tends towards small values with a high frequency.

Supervised machine-learning models

Our goal was to design a supervised machine-learning model to distinguish the highly cited articles from the rest, using an array of known altmetric features along with the score for user influence. The features shown in Table 3 were chosen to train our classification model. The features relating to articles were extracted from the altmetric data: the number of authors (F2); the types of users in altmetrics (F3–F6); the most significant social media post count (F7–F11), also employed by Hassan et al. (2017) and Costas et al. (2015) and others; but not *User Influence* (F1), which is an accumulated value of the rank scores of all the Twitter users who tweeted a given publication. The Twitter users were ranked *1* through *n*, where *n* is the total number of users; the most influential tweeter was given the *n* rank score. Combining the list of top users with the *mentioned* and *retweeted* articles list obtained from Altmetric.com, we obtained a new dataset comprising all those articles that were retweeted or mentioned by the top tweeters.

For training and testing, the data were divided using a 10-fold cross-validation technique. Further, ROC and PR curves were used to evaluate the performance of the model. Important features were identified by using the Extra-Tree classifier, along with the PR

Table 2 Description of data used in study

	N	Mean	Median
Top 50% HC	410	67,758.54	20,299.50
Rest	410	41,029.35	15,811.00
Total	820	54,393.95	17,963.00
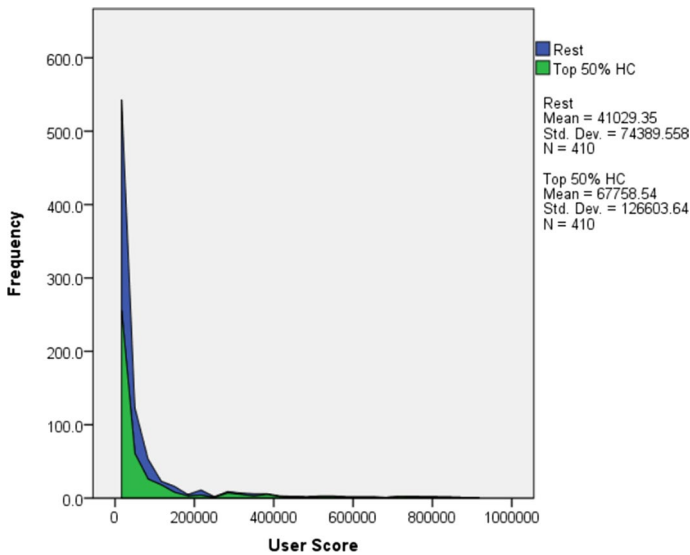
Fig. 2 User scores for articles in both the top 50% (HC) and the rest

area value of individual features. The classifiers applied to this study were Support Vector Machine (SVM) and Random Forest. Grid search algorithm was used to find the best parameter (Hesterman et al. 2010). The primary objective was to identify the relationship between the network of influential users and highly cited articles. For this purpose, a baseline was calculated, followed by a measure of the effectiveness of the results compared to the baseline.

Baseline model For the baseline experiment, we used all our features from F2 to F11, excluding F1; that is, *User Influence*. The performance of the baseline was evaluated by using both ROC and PR curves with that of the proposed model.

Figure 3 (left) indicates the ROC curve results using two different classifiers. The SVM displays the lowest area (0.66), while Random Forest achieved a score of 0.74. Figure 3 (right) also displays the results of the PR curve. Similar to the ROC results, Random Forest performed the best, with an PR area of 0.84, followed by SVM, which gave an area under the curve of 0.84.

Proposed model with user influence In this experiment, using all the features F1 through F11, a 10-fold cross-validation technique was used to train the classifiers. The technique divided the data into k subsets, resulting in 10 subsets. Each time one of the subsets was used as test data, the remainder of the $k-1$ subsets were used in training. This method

Table 3 Features for classification

| Features	Description
F1—user influence	User influence represents the network of Twitter users who were captured tweeting about a specific article. Each article in the dataset has multiple users who tweeted about it. User influence is the sum of the rank scores of those users
F2—number of authors	Number of co-authors of the article
F3—member of the public	As per the definition provided by Altmetric, these users do not link to scholarly literature
F4—researcher	As per the definition provided by Altmetric, these users are familiar with the scholarly literature
F5—practitioner	As per the definition provided by Altmetric, these users are clinicians or researchers who work in a clinic
F6—science communicator	As per the definition provided by Altmetric, these users link frequently to scientific articles from a variety of journals/publishers
F7—blog post count	Indicates how many times a publication has been mentioned or discussed in blogs
F8—Twitter post count	Indicates how many times the publication has been tweeted
F9—news post count	Indicates how many times a publication has been discussed or mentioned in news or magazines
F10—Facebook post count	Indicates how many times a publication appears on a public Facebook account's wall.
F11—Google+ post count	Indicates how many times users discussed a publication on Google+
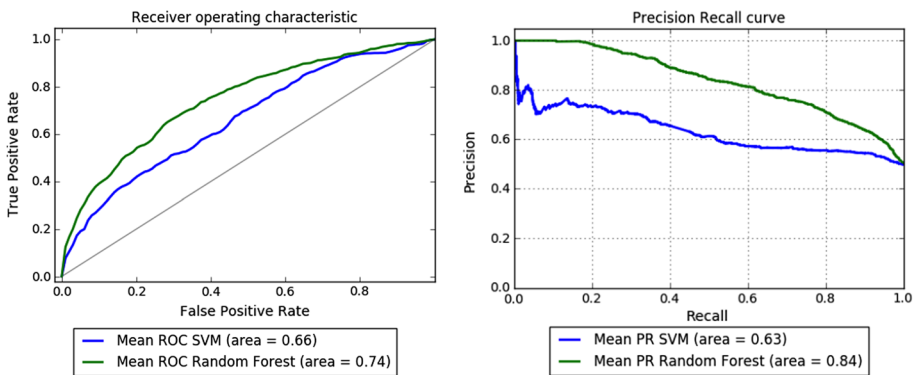
Fig. 3 Mean ROC and PR curves on baseline model

was repeated k times. It fit the model to 90% of the data, with 10% on test data. Two different classifiers, Random Forest and SVM, were implemented in a Python program using the *SciKit Learn* machine-learning library. SVM was useful as it ascertains a hyperplane, which was used in this case to separate the input class label in the case as highly cited or not; an RBF kernel was used for SVM in this instance. Multiple decision trees were used and ranked to receive an output using the Random Forest classifier; a maximum depth of 10 was set in Random Forest, to represent the number of questions to be asked before reaching an answer.
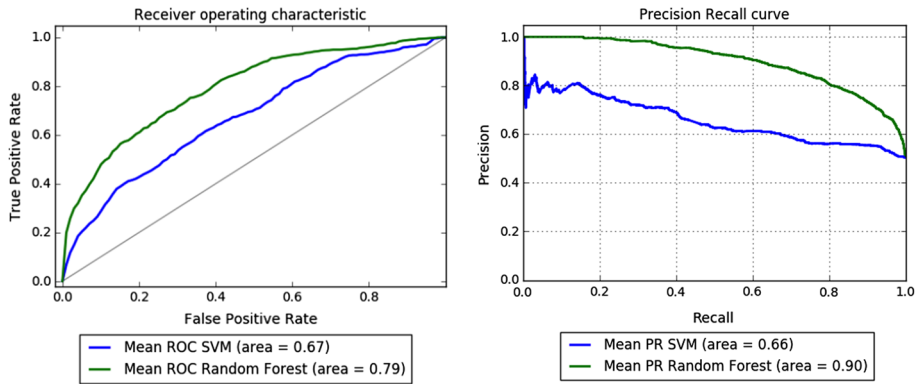
Fig. 4 Mean ROC and PR curves of proposed models

Again, the Grid search algorithm was used to obtain the optimal parameters to train the classifiers (Hesterman et al. 2010). Figure 4 (left) displays the mean ROC curves of the two different classifiers. It indicates the mean AUC using 10 folds, along with a mean AUROC value for each classifier. SVM showed an area of 0.67, whereas Random Forest gave an area of 0.79. The models performed better than the baseline, with Random Forest giving the best results of values 0.79.

To measure the effectiveness of the model at different recall levels, a PR curve was used. Figure 4 (right) displays the mean PR curves of the two classifiers. Similar to ROC, SVM gave an area with a curve value of 0.66, and Random Forest demonstrated good performance, exhibiting an area of 0.90. The models performed better than the baseline, with Random Forest providing the best results, with values of 0.90.

We compared the baseline results with the results obtained in the above section. The new model provided much better results for both PR analysis and ROC analysis. The best score of the baseline model in this experiment (0.84) was achieved by the Random Forest classifier, whereas the best score was (0.90) by Random Forest for PR. Similar results were obtained by ROC and, while Random Forest gave the best result (area = 0.74) in the case of the baseline, the worst results from the new model were also with Random Forest (area = 0.79).

Discussion on feature importance

Based on the results obtained in this work, Random Forest was chosen to obtain the most important features as it yielded better results than SVM. We examined the features' importance using the Extra-Tree Classifier. Along with PR curve analysis, it was employed to examine the effectiveness of individual features.

The Extra-Tree classifier was employed to rank the features by importance. Note that the Extra-Tree classification produces piece-wise multi-linear approximations (from a functional point of view), in contrast to the piece-wise constant ones of Random Forest. Our *SciKit-Learn*-based implementation assigned a score to each feature in such a way that the sum of all the features is always equal to 1. Table 4 shows the importance of individual features. The User influence (F1) feature proved to be the most important feature, with a score

Table 4 Importance score for each feature of the extra-tree classifier

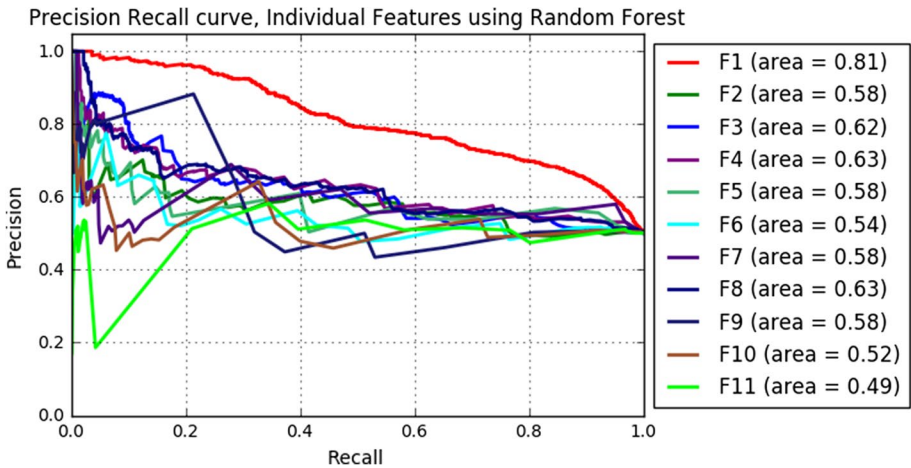| Type	Features	Score
F1	User influence	0.27
F2	Number of authors	0.15
F8	Twitter post count	0.12
F3	Member of public	0.11
F4	Researcher	0.10
F6	Science communicator	0.07
F7	Blog post count	0.05
F5	Practitioner	0.05
F10	Facebook post count	0.04
F11	Google+ post count	0.02
F9	News post count	0.02

Fig. 5 PR curves of individual features with Random Forest

of 0.27. Interestingly, the Twitter post count achieved the best result of all the social media post counts, topping the list at 0.12.

In addition to examining features' importance using the Extra-Tree Classifier, PR curve analysis was employed to examine the individual features using the Random Forest (see Fig. 5) classifier. Predictably, User influence gave the best PR area of 0.81, followed by Researcher and Twitter post count, each resulting in an area of 0.63. Four features achieved a similar performance (0.58): Number of authors; Practitioners; Blog post count; and News post count. Interestingly, both the analyses presented in Table 4 and Fig. 5 reveal User influence to be the most important feature in classification.

Concluding remarks

In this article, we investigated the relation between influential tweeters and highly cited articles. The tweets and user mentions, retweets and followers' links were modelled as an undirected graph, and this was used to find the most influential Twitter users in the dataset. We discovered that the features around influential users were highly competent in discriminating between highly cited and non-highly cited articles. We found that the score for influential users was the most important feature in the dataset, using the Extra-Tree Classifier and PR curves, when tested on individual features. In future, we seek to examine the influence of time to explore the duration for which a user remains influential within a subject area. The impact of time can also be used to examine differences across disciplines in order to investigate how influential users contribute to predicting citations across disciplines. In addition, in future studies we will consider the effect of bot Twitter accounts in identifying influential users in Twitter networks.

Last but not the least, instead of using social media as black box that generate social usage data, more research in needed to better study the underlying network structure of tweets and mentions that can directly or indirectly influence altmetric scores originated by Twitter platform.

References

Alonso, O., Carson, C., Gerster, D., Ji, X., & Nabar, U. S. (2010). Detecting uninteresting content in text streams. In *Proceedings of the SIGIR 2010 workshop on crowdsourcing for search evaluation* (*CSE 2010*) (pp. 39–42).

Anger, I., & Kittl, C. (2011). Measuring influence on Twitter. In *Proceedings of the 11th international conference on knowledge management and knowledge technologies*, *31*(*1*), pp. 1–31.

Bates, M. (2017). Tracking disease: Digital epidemiology offers new promise in predicting outbreaks. *IEEE Pulse.* https://doi.org/10.1109/mpul.2016.2627238.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science, 61,* 490–503. https://doi.org/10.1111/ajps.12274.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics, 8*(4), 895–903. https://doi.org/10.1016/j.joi.2014.09.005.

Bornmann, L. (2016). What do altmetrics counts mean? A plea for content analyses. *Journal of the Association for Information Science and Technology, 67*(4), 1016–1017.

Bornmann, L., & Haunschild, R. (2016). How to normalize Twitter counts? A first attempt based on journals in the Twitter Index. *Scientometrics, 107*(3), 1405–1422.

Chakrabarti, D., Wang, Y., Wang, C., Leskovec, J., & Faloutsos, C. (2008). Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC), 10*(4), 1.

Chung, F. R. (1997). *Spectral graph theory* (vol. 92, Regional Conference Series in Mathematics). Rhode Island: American Mathematical Society/Conference Board of the Mathematical Sciences. ISBN: 978-0-8218-0315-8.

Costas, R., Zahedi, Z., & Wouters, P. (2015). Do 'altmetrics' correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology, 66*(10), 2003–2019.

Escamilla, I., Torres-Ruiz, M., Moreno-Ibarra, M., Quintero, R., Guzmán, G., & Luna-Soto, V. (2016). Geocoding tweets approach based on conceptual representations in the context of the knowledge society. *International Journal on Semantic Web and Information Systems (IJSWIS), 12*(1), 44–61.

Hassan, S. U., Imran, M., Gillani, U., Aljohani, N. R., Bowman, T. D., & Didegah, F. (2017). Measuring social media activity of scientific literature: An exhaustive comparison of scopus and novel altmetrics big data. *Scientometrics, 113*(2), 1037–1057.

Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics, 108*(1), 412–423. https://doi.org/10.1007/s11192-016-1910-9.

Haustein, S., Bowman, T. D., & Costas, R. (2015). Interpreting 'altmetrics': Viewing acts on social media through the lens of citation and social theories. arXiv preprint arXiv:1502.05701.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology, 65*(4), 656–669.

Hesterman, J. Y., Caucci, L., Kupinski, M. A., Barrett, H. H., & Furenlid, L. R. (2010). Maximum-likelihood estimation with a contracting-grid search algorithm. *IEEE Transactions on Nuclear Science, 57*(3), 1077–1084.

Hussain, A. R., Hameed, M. A., & Sayeedunnissa, S. F. (2012). Measuring influence in social networks using a network amplification score-an analysis using cloud computing. In *2012 12th International conference on hybrid intelligent systems* (*HIS*).

Jarwar, M. A., Abbasi, R. A., Mushtaq, M., Maqbool, O., Aljohani, N. R., Daud, A., et al. (2017). CommuniMents: A framework for detecting community-based sentiments for events. *International Journal on Semantic Web and Information Systems (IJSWIS), 13*(2), 87–108.

Kalloubi, F., Nfaoui, E. H., & El Beqqali, O. (2017). Harnessing semantic features for large-scale content-based hashtag recommendations on microblogging platforms. *International Journal on Semantic Web and Information Systems (IJSWIS), 13*(1), 63–81.

Kemp, S. (2017). *Digital in 2017: Global overview*. Retrieved from 'We are social. https://wearesocial.com/blog/2017/01/digital-in-2017-global-overview. Accessed 10 June 2018.

Lee, M. K., Yoon, H. Y., Smith, M., Park, H. J., & Park, H. W. (2017). Mapping a Twitter scholarly communication network: A case of the association of internet researchers' conference. *Scientometrics, 112*(2), 767–797. https://doi.org/10.1007/s11192-017-2413-z.

Lotan, G., Ananny, M., Gaffney, D., & Pearce, I. (2011). The Arab Spring/The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication, 5,* 1375–1405.

Priem, J., Piwowar, H., & Hemminger, B. (2011). Altmetrics in the wild: An exploratory study of impact metrics based on social media. In *Metrics 2011: Symposium on informetric and scientometric research,* New Orleans, USA.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto.* Available online at http://altmetrics.org/manifesto/. Accessed 10 June 2018.

Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2011). In the mood for being influential on Twitter. In *Privacy, security, risk and trust (PASSAT) and 2011 IEEE 3rd international conference on social computing* (*SocialCom*) (pp. 307–314). IEEE.

Shrivastava, R., & Mahajan, P. (2016). Influence of social networking sites on scholarly communication: A study using literature in artificial intelligence. *Journal of Librarianship and Information Science.* https://doi.org/10.1177/0961000616678309.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics, 98*(2), 1131–1143.

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology, 68*(9), 2037–2062. https://doi.org/10.1002/asi.23833.

Tariq, J., Ahmad, M., Khan, I., & Shabbir, M. (2017). Scalable approximation algorithm for network immunization. arXiv preprint arXiv:1711.00784.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLoS ONE, 8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841.

Tsou, A., Bowman, T.D., Ghazinejad, A., & Sugimoto, C.R. (2015). Who tweets about science? In *Proceedings of the 2015 international society for scientometrics and informetrics* (pp. 95–100), Istanbul, Turkey.

Yang, M.-C., Lee, J.-T., Lee, S.-W., & Rim, A. H.-C. (2012). Finding interesting posts in Twitter based on retweet graph analysis. In *35th International ACM SIGIR conference on research and development in information retrieval* (pp. 1073–1074), August, Portland, OR.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics, 101*(2), 1491–1513.