Predicting Literature's Early Impact with Sentiment Analysis in Twitter

Saeed-Ul Hassan[a,], Naif R. Aljohani[b], Nimra Idrees[a], Raheem Sarwar[a], Raheel Nawaz[c], Eugenio Martínez-Cámara[d *], Sebastián Ventura[e,b], Francisco Herrera[d,b]

[a] *Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan*

[b] *Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia*.

[c] *Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Manchester, United Kingdom.*

[d] *Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, 18071 - Granada, Spain.*

[e] *Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Córdoba, 14071 - Córdoba, Spain.*

Abstract

Traditional bibliometric techniques gauge the impact of research through quantitative indices based on the citations data. However, due to the lag time involved in the citation-based indices, it may take years to comprehend the full impact of an article. This paper seeks to measure the early impact of research articles through the sentiments expressed in tweets about them. We claim that cited articles in either positive or neutral tweets have a more significant impact than those not cited at all or cited in negative tweets. We used the SentiStrength tool and improved it by incorporating new opinion-bearing words into its sentiment lexicon pertaining to scientific domains. Then, we classified the sentiment of 6,482,260 tweets linked to 1,083,535 publications covered by Altmetric.com. Using positive and negative tweets as an independent variable, and the citation count as the dependent variable, linear regression analysis showed a weak positive prediction of high citation counts across 16 broad disciplines in Scopus. Introducing an additional indicator to the regression model, i.e. 'number of unique Twitter users', improved the adjusted R-squared value of regression analysis in several disciplines. Overall, an encouraging positive correlation between tweet sentiments and citation counts showed that Twitter-based opinion may be exploited as a complementary predictor of literature's early impact.

Keywords: Altmetrics, Twitter, Sentiment Analysis, User Category, Predicting Citations

* Corresponding author

Email addresses: saeed-ul-hassan@itu.edu.pk (Saeed-Ul Hassan), nraljohani@kau.edu.sa (Naif R. Aljohani), nimraidrees@yahoo.com (Nimra Idrees), raheem.bwl@gmail.com (Raheem Sarwar), r.nawaz@mmu.ac.uk (Raheel Nawaz), emcamara@decsai.ugr.es (Eugenio Martínez-Cámara), sventura@uco.es (Sebastián Ventura), herrera@decsai.ugr.es (Francisco Herrera)

1. Introduction

Altmetrics is an umbrella term. Many social media platforms, such as Twitter[1], Facebook[2], CiteULike[3] and MendeleyReadership[4], can be used as article-level metrics to gauge the impact of research, and may be thus referred by this term (Haustein et al., 2015) too. With the growth in article-level indices data, there is a growing need to provide tools to allow researchers to employ these datasets. There are several altmetric data aggregators, including Altmetric.com, ImpactStory and Plum Analytics, available to capture article-level web activity and provide the data to researchers. More recently, scholars are increasingly using online platforms to read, bookmark, share, discuss and rate research, which results in a vast amount of online data. Mining these data may provide useful insights in an alternative way to traditional citation metrics (Priem et al., 2011). Although the popularity of altmetric techniques has been increasing (Nuzzolese et al., 2019), there is a paucity of both information and evidence on their effectiveness, and currently the major challenges are to ensure the use of standards and best practice (Haustein et al., 2015; Bornmann et al., 2019).

Traditional bibliometric methods gauge the impact of research through quantitative indices that are based on citations data (Waheed et al., 2018; Bonaccorsi et al., 2017a, b), such as the journal impact factor, h-index and source-normalized impact per article (Haddawy et al., 2017; Hassan et al., 2016, 2018). However, due to the lag time involved in the citations, which is a limitation associated with citation-based quantitative indices (Zhu et al., 2014; Hassan et al., 2012), it may take years before the full impact of an article can be comprehended. With increased usage of the web for scholarly communications, altmetric data are of enhanced interest as they capture real-time data from online platforms such as Twitter, Facebook and CiteULike (Priem et al., 2011). Therefore, altmetric techniques can be used to measure the early impact of scientific literature (Didegah et al., 2018). Twitter is a platform widely used by scholars to share opinion on research articles (Priem et al., 2011). There is a need to investigate how authentic is this way of measuring impact and whether a high tweet count for an article would lead to a high citation count in the future. Altmetric.com captures the tweet count for each research article and also other Twitter

[1] https://twitter.com
[2] https://www.facebook.com
[3] http://www.citeulike.org
[4] https://www.mendeley.com

demographics, such as the Twitter user category. Understanding the extent to which a tweet on a scholarly article conveys opinion about it might help us to understand the importance of the Twitter indicator as a measure of that article's impact.

In this paper we study the influence on the early impact of research literature by the opinions posted on Twitter and other Twitter data. We claim that the articles cited in positive and neutral tweets may have a greater impact than either those that are not cited or those that are cited in a negative tweet. In order to evaluate this claim, we measured the early impact of tweet sentiments associated with the research articles covered by Altmetric.com from July 2010 to June 2016 that were disseminated on the Twitter platform, using the text of over 6 million tweets. We explored the positive, negative and neutral sentiments in tweets, along with their unique Twitter user information. We performed multiple linear regression on our dataset to analyse the use of Twitter as a high or a low predictor of citation count. We first differentiated the counts of negative, positive and neutral tweets from the original altmetric tweet count then applied multiple linear regression, using positive and neutral sentiments as independent variables and citation count as the dependent variable. The inclusion of information on the unique Twitter user can help to normalise the effect of repeated dissemination of the same tweet by a given account (such as bot accounts) so, to overcome the effect of any inflated distribution, we introduced the unique Twitter user into the analysis as a third independent variable in predicting a high citation count.

For the tweet sentiment analysis, we used Twitter text messages consisting of a maximum of 140 characters[5], providing not just the text message but hashtags, usernames, pictures and URLs. We noted that Twitter texts often incorporate abbreviations, contractions and acronyms, and may contain shortened forms, truncated messages and slang. We found that the lexicon-based sentiment approach employed by SentiStrength (SentiStrength, 2017) was well suited to this context. Moreover, many studies have already shown its effectiveness compared to other tools to analyse the sentiments of tweet text. For instance, Friedrich et al. (2015) analysed two existing sentiment analysis tools, SentiStrength and Sentiment140, to detect the sentiment in tweets about academic articles. They concluded that the adaptation of the lexicon of SentiStrength to the scholarly domain allowed to improve the accuracy of the sentiment classification of SentiStrenth in the scholarly

[5] Note that the recent advancements on the Twitter platform now allow up to 280 characters.

domain. The SentiStrength uses an algorithm that simultaneously extracts both positive and negative sentiments from short, informal texts (Thelwall, Buckley & Paltoglou, 2012). In this study, we proposed improving SentiStrength[6] by incorporating new opinion-bearing words to update its sentiment lexicon and thus adapt it further to the scientific literature domain.

In this paper we attempt to answer the following research questions:

a) What is the influence of including the new opinion-bearing words in the research domains in SentiStrength for assessment of the impact of tweets on scientific literature?
b) What kind of opinions (positive, negative or neutral) do tweets convey about a linked research article?
c) What is the difference between disciplines regarding tweets containing positive, negative and neutral sentiments?
d) Which Twitter user categories share the most opinion when tweeting about research articles?
e) Does a high tweet count with positive sentiments about a research article lead to a high citation count for that article in the future?
f) Can a high tweet count with negative sentiments towards a research article lead to a low citation count for that article in the future?

The contributions of this study are as follows:

a) The prediction of the early impact of research articles by using the sentiment in tweets that mention those articles.
b) The adaptation of SentiStrength to the scientific literature domain by incorporating new opinion-bearing words into SentiStrength's sentiment lexicons.
c) To show the relationship between the citation count and the tweet sentiment associated with research articles by employing multiple linear regression.
d) To show that tweet sentiment can be used to indicate the early impact of scientific research.

The rest of the paper is organised as follows: Section 2 presents a review of related altmetric studies that seek to measure the impact of Twitter on the dissemination of scientific literature. Section 3 presents the dataset and the pre-processing approach to feed data to the SentiStrength model, along with the setup for the evaluation of classic SentiStrength against the adapted model. Next, Section

[6] SentiStrength is freely available for research purposes, and its lexicons can be adapted to the field of interest.

4 provides a detailed discussion of the results. Finally, Section 5 concludes the findings and highlights the directions for future research.

2. Literature Review

Many studies have shown that few tweets about research articles actually convey much positive or negative sentiment: most are neutral and solely for the purpose of information dissemination. Thelwall et al. (2013) performed a pilot study on 270 randomly collected tweets about research articles and analysed the kinds of opinions that the tweets conveyed about them and whether the ratio of negative tweets to the overall tweet count as a measure of research impact might be ignored. Their results showed that tweets about scholarly articles are mostly objective, consisting of either the article title or points from a brief summary. By contrast, our study conducted sentiment analyses on a dataset of over 6 million tweets linked to more than a million research articles, as captured by Altmetric.com.

Martin Fenner (2013) investigated article-level matrices of PLOS biology research that was published in 2010. The study showed that although some of the highly-cited articles had a great many online viewings, overall there was a low correlation between the number of citations and online views. Thelwall et al. (2013) analysed the degree of correlation between various altmetric sources and citation counts. The study compared 11 altmetric indicators with Web of Science citations for 208,739 *PubMed* articles. The results showed significant evidence of a correlation between a high altmetric score and a high citation count in Twitter, Facebook posts, blogs, research highlights, online media and forums, but very little or no correlation in Google+. There was insufficient data to support a correlation between the citation count and indicators such as LinkedIn, Pinterest, Q&A sites and Reddit.

Costa et al. (2015) undertook analysis of the various altmetric indicators provided by Altmetric.com and their correlation to citation counts. Their results showed that while there is a positive correlation, the value is very low, showing a weak correlation, and the authors conclude that altmetric indicators do not measure the same impact as do traditional methods, such as citation counts. Ravenscroft et al. (2017) investigated the correlation between altmetric scores and research

evaluation framework (REF) impact, and their results show that there is little significant correlation.

Houqiang Yu (2017) found that in all altmetric indicators there is a significant difference between the number of posts (NP) and the number of unique users (NUU). He identified a high to moderate Pearson correlation between NP and NUU for various altmetric indicators. He also analysed Twitter user count information for the various user categories of researcher, practitioner, science communicator and member of the public. Correlation analysis was conducted on the Twitter user count in each category and the citation count of the associated research article. The results revealed that the category of researcher yields the highest correlation value, yet the overall value remains low, similar to that in the findings of previous studies.

Several studies have aimed to establish the extent to which Twitter is an authentic measure of the research impact of an academic article and whether we can use it to predict the article's citation count (Priem et al., 2011; Thelwall et al., 2013; Costas et al., 2015; Haustein et al., 2015; Eysenbach, 2011; Holmberg & Thelwall, 2014). These studies used simple correlation techniques to find the relationship between the citation count and the raw tweet count. In this direction of research, Konkiel (2016) suggested that no single indicator could comprehensively measure the impact of research and that it would be beneficial for researchers to consider a combination of alternative metrics. Haustein et al. (2016) revealed that a large number of tweets about scholarly articles are from automated bot accounts, and the same account may tweet hundreds of times about the same article. This can affect Twitter's value as a measure of impact. In addition, Yu (2017) identified that there is a considerable disparity between the number of posts (NP) and the number of unique Twitter users. One possible reason might be multiple tweets about the same article by a single user for the sake of self-promotion, advertising or fraud. This makes altmetric indicators less valuable as a measure of research impact.

In this study, we explored whether the positive, negative and neutral sentiment conveyed in tweets, along with information on the unique Twitter user, may be used to predict the early impact of an article. We performed multiple linear regression analysis on our dataset to analyse the use of Twitter as a high or a low predictor of citation count. The results were noticeably improved in

disciplines such as earth and planetary sciences, health professions and nursing, mathematics, and medicine and medical sciences. The result of multiple linear regression analysis, which found that negative tweets predicted a low citation count, remained approximately zero across all disciplines – and in the discipline listed as 'general' it even exhibited a weak negative prediction – showing that in the multidisciplinary category negative sentiments in social media do not affect the achievement of a high citation rate.

3. Data and Methodology

This section presents the altmetric data and its pre-processing steps, along with the sentiment analysis approaches to infer the opinion mining of tweets that cite research articles. We also show the evaluation results of the classic SentiStrength model compared to the proposed model incorporating new opinion-bearing words.

3.1 Dataset

The corpus consisted of altmetric data captured by Altmetric.com[7] from July 2011 to June 2016. In total, 1,083,535 research articles with at least one tweet and one citation (to February 2017) were extracted from the altmetric dataset. Note that Altmetric.com provides a unique URL for each tweet pertaining to a given article. Using R code, the tweet text was scraped from Twitter.com by processing the URLs against all 6,482,260 tweets[8]. Altmetric.com indexes the Unique User Count (UUC) of those who tweet about each article, and this information was extracted from the dataset. In addition, user information was gathered, and they were categorised in: researchers, practitioners, science communicators or members of the public. These categories are assigned based on information in the user's profile, the types of journals that they are linked to and their 'friends' list (Altmetric LLP, 2017). Finally, the citation count of the research articles was collected from Scopus API and disciplinary information assigned to each as per the Scopus subject-category scheme employed by Haddawy et al. (2017).

[7] The data were received from Altmetrics.com in JSON file format. Under the agreement, the author cannot publicly disseminate any copy of this data. However, the same data may be freely obtained by the scientific community from the Altmetric.com for research purposes.

[8] The scraped tweet text and code can be downloaded from the following repository: https://github.com/slab-itu/kbs_tw_text/

3.2 Sentiment analysis approaches

We claim that papers cited in either positive or neutral tweets have a greater impact than those that are not cited or are cited in negative tweets. In general, sentiment analysis can be applied at three levels of detail: (1) document; (2) sentence; and (3) entity/aspect (Liu, 2012; Dragoni et al., 2019; Federici et al., 2016). Our evaluation is at the document/tweet level. The approach used in this study is a lexicon-based method that counts the words from a sentiment lexicon that appear in a given text (SentiStrength, 2017). Here, the sentiment is assigned by a lexicon-based sentiment classifier, a combination of the sentiment word scores and the query term–sentiment word proximity scores.

Following the work of Liu et al. (2017), we compiled a list of the terms most commonly used in tweets either to praise or to convey negative sentiment about a research article. Positive tweets about an article usually contain phrases such as 'compelling article', 'fundamental study', 'remarkable finding' or 'novel technique', while negative tweets contain words or phrases such as 'biased article', 'bad idea', 'fake', 'fallacy' or 'weak conclusion'. All such terms were searched for in the tweet dataset and the corresponding tweets were considered carefully. Further, terms that were found in many tweets, whether positive or negative, were added to the SentiStrength lexicons. Around 80 positive and negative terms were added in this way. In addition to these terms, we searched the tweets dataset for words from the SentiStrength lexicons, one by one, and those tweets containing them were analysed by two human annotators. Examining the lexicons, we noted that many of the terms used in tweets, such as 'death', 'war', 'accident', 'germs' and 'care', are science-specific terms and their use was not intended to convey the author's opinion of an article. However, these terms were causing false assignment of positive or negative sentiment, therefore we decided to remove them from the SentiStrength lexicon. In total, 148 such terms were removed.

We found that tweets sometimes contain research-specific terms taken from the article's title rather than used to express an opinion on the article. By comparing each word in a tweet text string to each word in the linked article's title string, then removing the word from the tweet's text, matched terms such as 'cancer', 'disaster' and 'harm' were excluded to prevent false assignment of sentiments, and this greatly improved SentiStrength's efficiency in detecting sentiment. Note that Friedrich et al. (2015) adapted this practice in their work on analysing tweet sentiments. Further,

URLs, # signs and user mentions (@username) were considered not to carry any opinion about the article and were duly removed from tweets' text to avoid false assignment of sentiment. Moreover, the specific language of each tweet's text was detected using the R programming language; tweets in any language other than English were filtered out, as the SentiStrength lexicons are composed of English words. Finally, using SentiStrength with our adapted lexicons, the sentiments of the remaining 5,341,800 tweets were detected. The data pertaining to the adapted lexicons can be found in Appendix A, Tables A-1 to A-3.

Finally, to identify the sentiments in tweet text we adapted our SentiStrength model's sentiment-strength value to range from -1 to -5 (for negative sentiments) and 1 to 5 (for positive sentiments). Further, tweets for which the sentiment strength was detected to be between 2 and 5 were regarded as positive tweets, and those for which it was between -2 and -5 were regarded as negative. The remainder, for a sentiment strength of between 1 and -1, were regarded as neutral. In this way, the counts of positive, negative and neutral tweets were established.

3.3 Evaluation of SentiStrength Models

In this section, we present our evaluation of the sentiment classification models. Accordingly, we annotated a subset of the tweets in the original dataset, containing 2,544 tweets in English about publications in various disciplines: biomedical and health sciences (20%); life and earth sciences (20%); mathematics and computer science (8%); physical sciences and engineering (32%); and social sciences and humanities (20%).

We manually annotated the tweets with the help of two independent annotators. Both are domain experts and well aware of the issues involved in the task of assigning tweet sentiment. Bearing in mind the context of the articles, the annotators marked the tweets as neutral, negative or neutral. The agreement of the annotators is 0.75 according to the Cohen's Kappa agreement coefficient (Cohen, 1960), which is a substantial agreement according to Landis and Koch (1977). Table 1 shows the percentage of tweets per label.

Table 1: Manually Annotated Tweets.

Tweet Labels	**Percentage of Tweets**
Positive	34.7%
Negative	35.7%
Neutral	29.6%

Table 2: Evaluation of classification models.

Models	**Precision**	**Recall**	**F1 Score**	**Accuracy**
New Lexicon	0.660	0.355	0.215	0.570
SentiStrength	0.569	0.489	0.476	0.659
SentiStrength + New Lexicon	0.642	0.581	0.576	0.721
SVM (- stop words, stemming, tf-df)	0.593	0.496	0.501	0.663
SVM (- stop words, stemming, tf-idf) + New Lexicon	0.670	0.593	0.603	0.728

Table 2 shows the evaluation results of the SentiStrength and New Lexicon models compared to our adapted SentiStrength + New Lexicon model. We found that our adapted model achieved great accuracy in predicting tweet sentiment, with an average accuracy of 72.1%, compared to the SentiStrength and New Lexicon models' 65.9% and 57.0% respectively. Our adapted model also achieved high F1 and recall scores compared to the SentiStrength and New Lexicon models. Interestingly, the New Lexicon model had the highest average precision, yet this was at the cost of a very low average recall.

In addition, we evaluated the performance of the SentiStrength model (unsupervised) against a standard supervised sentiment classifier, specifically the Support Vector Machines (SVM) algorithm. We formulated two SVM-based methods, and their performance is reported in Table 2. In the first method (i.e. SVM, TF-IDF), we pre-processed the tweets by removing stop words and applying the stemming process. We then used the bag-of-words (BoW) model to extract features from tweets, where TF-IDF (term frequency - inverse document frequency) scores are the feature values. After completing the feature extraction process, we applied the SVM model for tweet sentiment classification using 10-fold cross-validation. In the second method, we added the new lexicon as a feature in the same TF-IDF-based feature space that we used in the first method, and

applied the SVM model for tweet sentiment classification using same evaluation approach. The results show that incorporating the new lexicon in the feature space used by the first method (i.e. SVM, TF-IDF) improved the performance of the classification.

Note that the new lexicon words describe a scholar's attitude to a certain article and the properties upon which that opinion is about. However, creating a domain-specific lexicon is itself a complicated task because of its dependency on the subject domain. For instance, one word may express a positive opinion in one domain, for instance 'high-quality *material*', while in another context '*material* studies' conveys only neutral opinion. Hence, a better approach to constructing a list of opinion words is to develop for the desired domain a domain-specific lexicon instead of general-purpose lexicon. Another explanation is that some lexicon terms are actually generated by the user and do not appear in standard dictionaries. Therefore, a representative domain-specific lexicon facilitates the task of sentiment classification. Moreover, our experiments have shown that, while opinion lexicons are useful in sentiment classification, they are not sufficient in themselves and should be used only in conjunction with other features and tools.

Overall, the aim of this evaluation was to show the suitability of combining words from SentiStrength and New Lexicon. We demonstrated that the new words do not interfere with the SentiStrength terms. Instead, the combination of the two lexicons, SentiStrength and the new one, improves the recall at the cost of a drop in the score for precision, from 66% to 64.2%. This achieves a higher F1 score than by using the two lexicons independently. Notably, when the model uses the new lexicons the SVM-based evaluation shows increases in both the precision and the recall indices. Consequently, the new specific words in the domain of scientific literature provide valuable knowledge for improving the inferring of the opinion meaning. On the other hand, since the aim of this work is not the sentiment classification of tweets, and the performance difference among the unsupervised classification system and the supervised one is substantially short, we decided to use the unsupervised classification system, or in other words, the SentiStrenth algorithm enriched with the words of the scientific literature domain (New Lexicon) to conduct our study of the relation of the opinion meaning of tweets with the early prediction of the impact of an article.

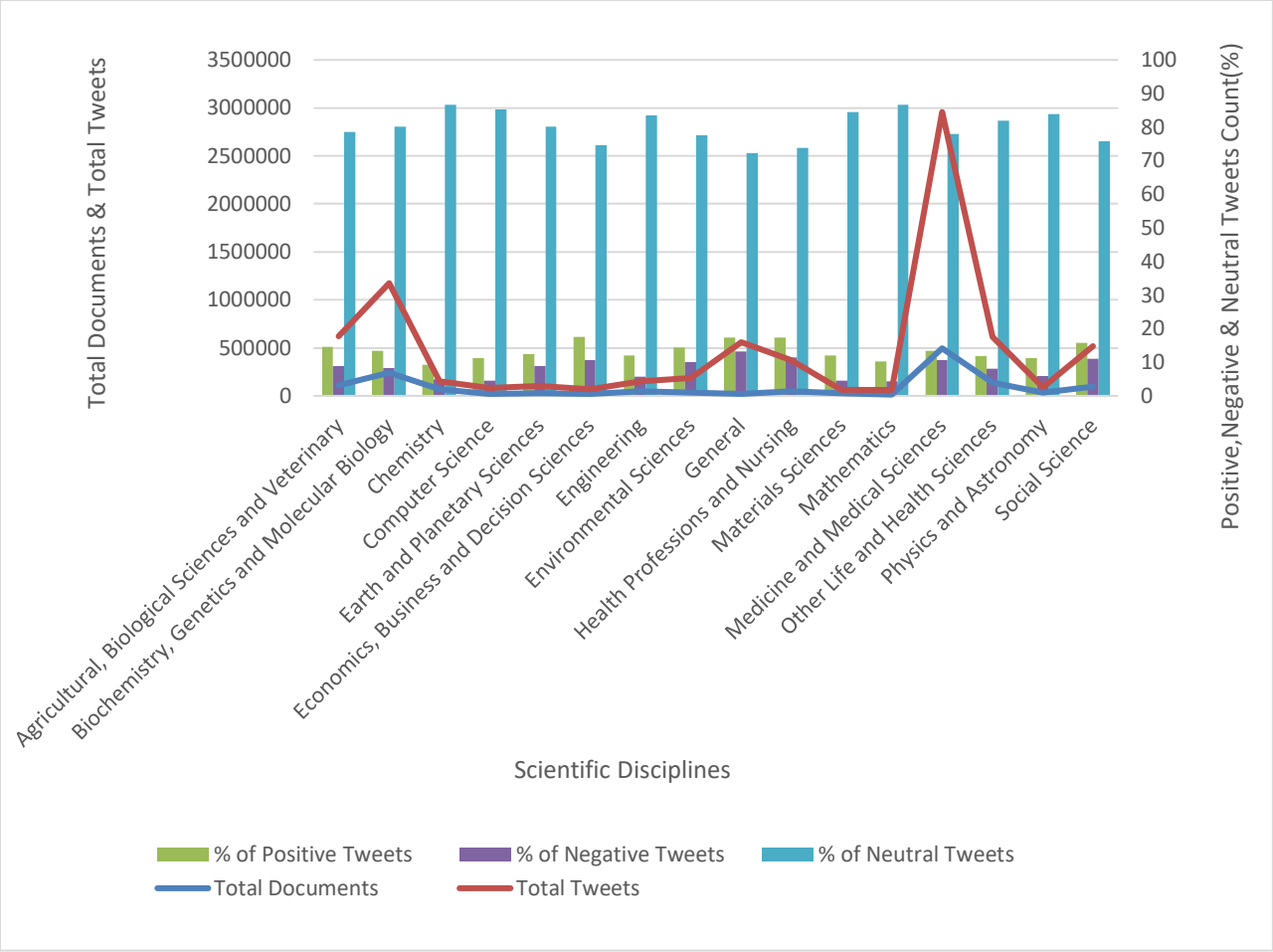

Figure 1. Distribution of tweets with positive, negative or neutral sentiments across scholarly disciplines.

4. Analyses and Results

This section presents the analysis and discussion of our results. Using the adapted SentiStrength with the new lexicon, we show the distribution of tweet sentiments across Twitter-user categories and broad disciplines, ranging from agriculture, biological sciences and veterinary studies, through the social sciences. Finally, we discuss the regression analysis conducted between tweet sentiments and citations.

4.1 Distribution of tweet sentiments across scholarly disciplines

Of the total 5,341,800 tweets, 75.7% are neutral, 14% positive and 10.3% negative. A cross-disciplinary analysis of the sentiments in these tweets showed that in all disciplines the majority are neutral, and that in most the percentage of positive sentiments is slightly higher than that of negative sentiments (see Fig. 1), confirming previous findings (Thelwall et al., 2013; Friedrich et

al., 2015). We found that the field of economics, business and decision sciences has a low tweet count of 67,946, yet it has the highest count of tweets that convey positive sentiments (17.5%). Moreover, in the disciplines of both 'general' and health professions and nursing, tweets linked to articles demonstrate a high percentage of both positive and negative sentiment; it can be concluded that they convey more sentiment than those in other disciplines.

4.2 Distribution of tweet sentiments among Twitter user categories

We counted the articles and the positive, negative and neutral tweets in each of the four user categories in which an article has at least one Twitter-user interaction. The data were analysed against these user categories, and the results showed that although the total tweet count is higher for the category 'member of the public', those in the other three categories (researcher, practitioner and science communicator) conveyed a greater number of positive and negative sentiments in their tweets (see Fig. 2). This comparatively high percentage in the latter three categories was expected, because these users interact and use research in their daily routine more than members of the public, thus they are more likely to convey sentiment and opinion in their tweets.

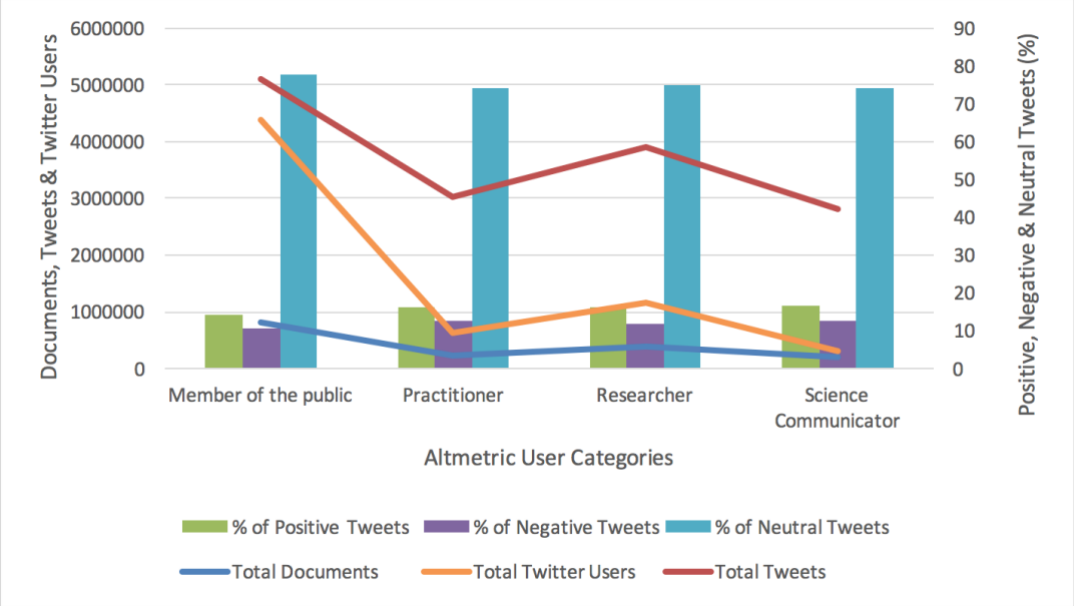

Figure 2. Distribution of tweets with positive, negative or neutral sentiments across user categories.

4.3 Multiple linear regression analysis on Twitter data

Previous studies have used raw tweet counts of research articles to analyse the correlation of tweet count to the citation count without considering whether the opinion on the article is positive or

negative (Priem et al., 2011; Thelwall et al., 2013; Costas et al., 2015; Haustein et al., 2015; Hassan et al., 2017a). In this study, we removed all tweets that conveyed negative sentiments about an article. We then performed multiple linear regression analysis on the tweet dataset to predict the citation count for that article. First, we applied linear regression to the remaining positive and neutral tweet counts as independent variables and to the citation count as a dependent variable. The R-squared and adjusted R-squared values of multiple linear regression analysis remained low, which indicates a weak prediction of the citation count. As a single Twitter user can send multiple tweets about any single article, to reduce the effect of inflated distribution we next introduced to the multiple linear regression model a third variable, UUC. In the variables that we used as independent variables in our regression model, as shown in Table 3, the P value is approximately zero. This makes it a significant variable.

Introducing the UUC variable to the multiple linear regression analysis improved the results slightly, and the adjusted R-squared value of regression analysis improved noticeably in disciplines such as earth and planetary sciences, health professions and nursing, mathematics, and medicine and medical sciences (see Fig. 3). We present the adjusted R-squared values throughout the analysis because R-squared increases every time that we add a new variable to the model, whereas the adjusted R-squared value increases only if the new variable improves the model.

Table 3: Coefficients of regression model with positive and neutral tweets.

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	12.761125	0.038188	334.16	<2e-16 ***
Positive tweets	0.254208	0.011579	21.95	<2e-16 ***
Neutral tweets	0.337588	0.004511	74.83	<2e-16 ***
UUC	0.035927	0.001179	30.48	<2e-16 ***

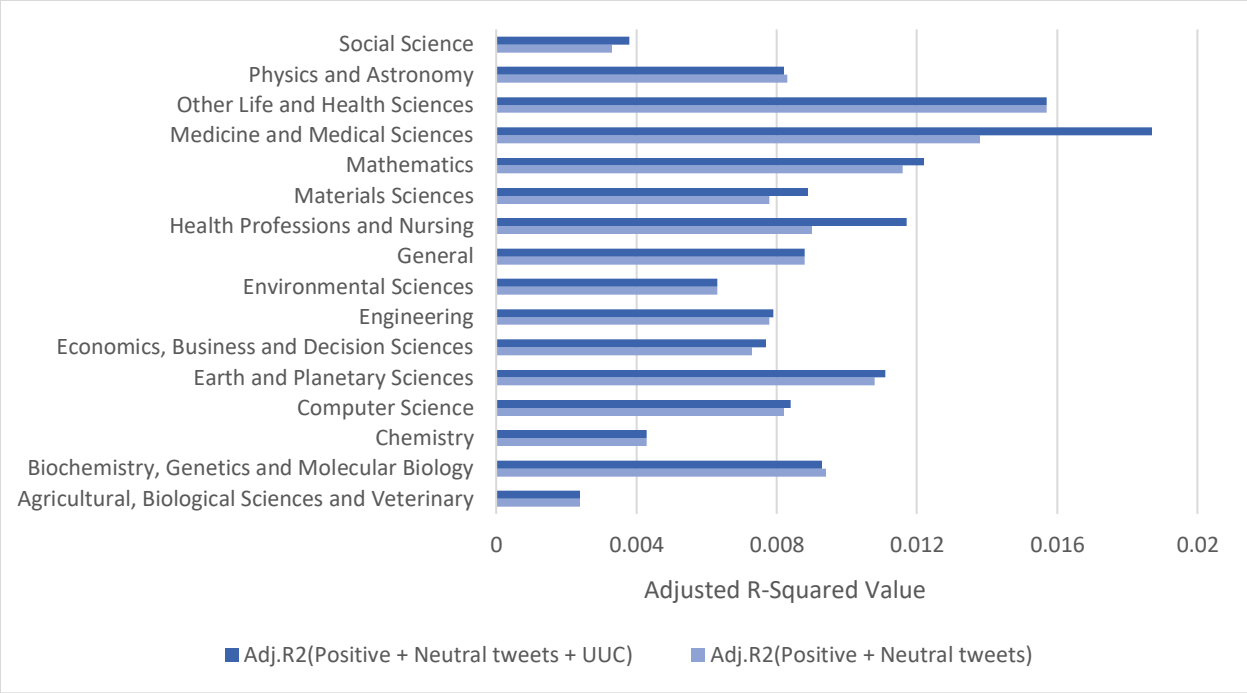

Figure 3. Adjusted R-squared value by using independent variables 'Positive + Neutral tweets' and 'Positive + Neutral tweets + UUC' across disciplines.

To analyse whether tweets with negative sentiments can be used as an early indicator of a low citation count in future, linear regression was applied to the negative tweet count as an independent variable and to the citation count as a dependent variable. The adjusted R-squared value of regression analysis remained low for all disciplines; indeed, the discipline of 'general' showed a weak negative prediction. Note that the 'general' discipline belongs to multidisciplinary publications indexed by prestigious journals such as *Science*, *Nature* or *PNAS*. This means that negative social media opinion does not affect the achievement of a high number of citations in multidisciplinary scientific research. This is a unique behaviour, contrasting with the other disciplines analysed in this study. Note that the variable of UUC was used as a second independent variable P value for the negative tweet count, and was approximately zero, which makes these variables in the regression model significant (see Table 4).

Table 4: Coefficients of regression model with negative tweets.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.867750	0.035818	387.18	<2e-16 ***
Negative tweets	0.455607	0.011987	38.01	<2e-16 ***
UUC	0.078219	0.001071	73.03	<2e-16 ***

Introducing UUC into the multiple linear regression on negative tweets increased the adjusted R-squared value in disciplines such as medicine and medical sciences, health professions and nursing, and material science (see Fig. 4). Nevertheless, the increase is simply the effect of UUC, supporting the theory that if there are large numbers of users performing altmetric activity on a scholarly article then the article must be popular and, most likely, will receive more citations in future. A good line for future work is the study of the influence of negative tweets on citations, in order to explore if negative tweets hinder future citation. Further, using both positive and negative tweets, we performed comparative analysis of the citation count and the adjusted R-squared value (see Fig. 5). As expected, the adjusted R-squared values were higher when positive tweet parameters were used than when negative ones were used. This is because when the count of positive tweets about an article is high then that article's citation count is also high.

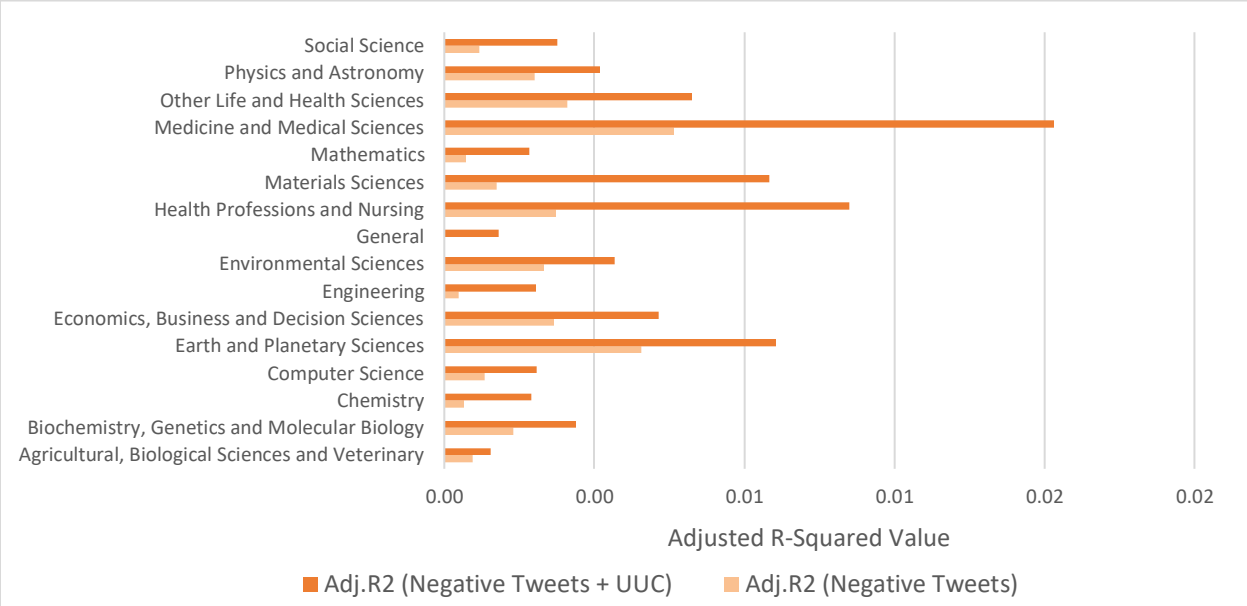

Figure 4. Adjusted R-squared value by using independent variables 'Negative tweets' and 'Negative tweets + UUC' across disciplines.

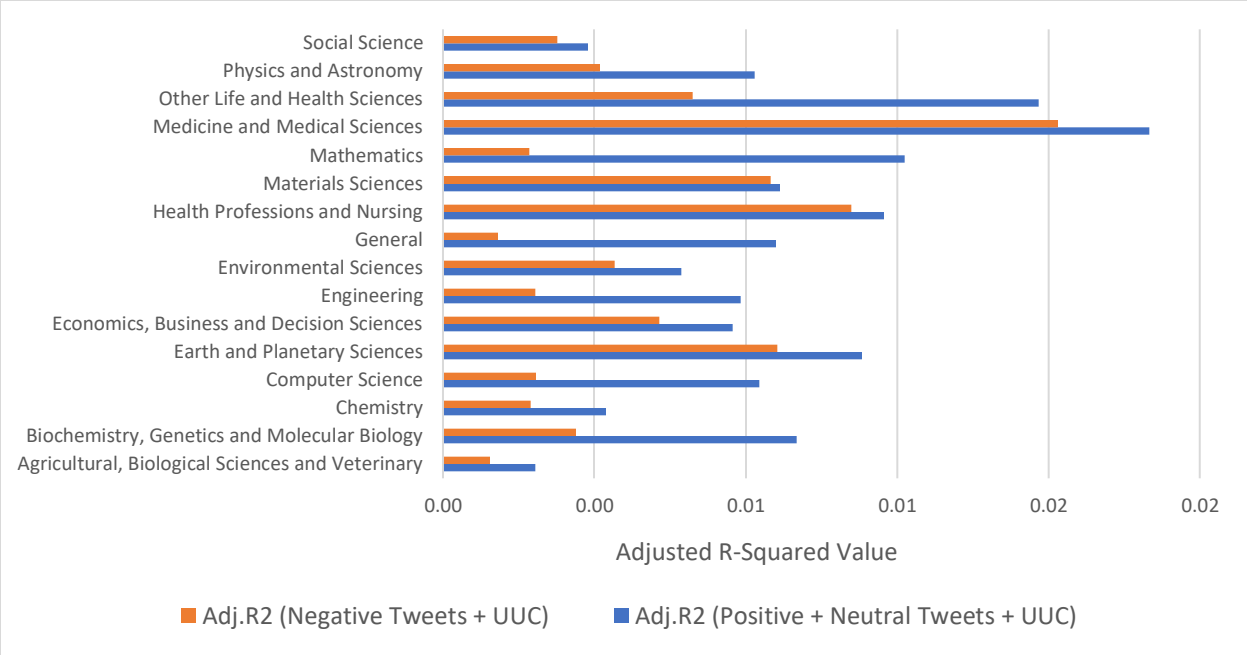

Figure 5. Positive and negative tweets for adjusted R-squared value distributed across disciplines.

5. Concluding Remarks

Traditional bibliometric techniques gauge the impact of research through citation-based quantitative indices, such as the journal impact factor and h-index. However, due to the citation lag time, which is a limitation associated with citation-based quantitative indices, it may take years before the full impact of an article can be comprehended. This study proposed measuring the early impact of tweet sentiments associated with research articles disseminated on Twitter. First, we improved SentiStrength, a sentiment analysis system, by incorporating new opinion-bearing words to update its lexicon to make it suitable for the assessment of the impact of tweets on scientific literature. We showed that the new opinion-bearing words in the research domain included in SentiStrength for the impact assessment of tweets on scientific literature improved the predictive power of the classic SentiStrength model. Thus, the techniques employed can be further exploited in the assessment of tweets pertaining to scientific literature. Further, to evaluate the use of Twitter as an altmetric means of gauging the early impact of a research article, sentiment analysis was performed using tweets about the scientific articles indexed in Altmetric.com from July 2011 to June 2016. We found that the papers cited in either positive or neutral tweets had a higher impact

than those not cited, or those cited in a negative tweet. Across the fields of economics, business and decision sciences, health professions and nursing, and 'general', tweets convey a comparatively high percentage of sentiments, while those in the field of chemistry convey the least. Furthermore, across the Twitter-user categories, tweet counts are lower among researchers, practitioners, and science communicators than among members of the public, yet those categories convey more sentiment in their tweets.

One of the limitations of this study is in the aggregation of sentiment counts across the user categories; a tweet_id can be assigned to multiple user categories in Altmetric.com data. Since Altmetric.com stores Twitter demographics at document rather than at a tweet level, there is no straightforward way to establish how many tweets have been sent in any specific category. Future studies could consider devising a means to differentiate this count to achieve superior analysis at the user-category level. Another limitation of our study is the use of user categories. Note that Altmetric.com assigns Twitter users to the categories of 'researcher', 'practitioner', and 'science communicator'. All other users are assigned as 'members of the public'. In other words, this is a catch-all category for Twitter users for whom Altmetric.com was unable to assign a proper category. Therefore, any results based on the category 'member of the public' are less useful than those for the other categories. Another issue is that less-good articles are sometimes used as a negative example in an article's literature review. Thus, future work could be undertaken on analysing the sentiment in a tweet in relation to a citation's opinion towards a scientific publication.

In future work, instead of assigning sentiment to one of three categories (positive, negative or neutral), we will seek to establish a tweet's strength of sentiment (Hassan et al., 2017b). Thus, tweets with greater positivity will be assigned a higher weight in evaluating an article's research impact, or other opinion distribution models can be exploited (Kim et al., 2018; Qiu et al., 2019) to improve citation prediction. It is possible that, while a recent article may receive much attention online because internet usage by scholars has recently increased, it may have a low citation count due to the short interval since its publication. Therefore, to improve the prediction of an article's citation count by regression analysis using tweet sentiments, we need to consider the length of time that has elapsed since its publication. Moreover, the meta-knowledge (Thompson et al., 2017;

Shardlow et al., 2018; Zhu et al., 2013) and discourse context (Hassan & Haddawy, 2015; Ananiadou et al., 2013) of a tweet's text can be exploited using state-of-the-art natural language processing (Batista-Navarro et al., 2013) and deep learning models (Jahangir et al., 2017) to understand the impact of tweets better. We recommend caution in making assertions that a high number of tweets about an article increases the likelihood of citation; if most of the tweets are negative, it is likely that the article will not be cited. We think that the study of the influence of negative tweets may be a good direction for future work. Future studies may include an experimental dataset formed of non-English tweets to secure better coverage of tweet sentiments in altmetric data.

We conclude that the correlation between Twitter-based sentiments and citations is an encouraging relationship between these indices and may be used as a complementary indicator to predict the early impact of literature; however, further investigation is desirable.

Acknowledgements

We should like to thank Altmetric.com for granting us access to their dataset for research purposes. This work was partially supported by the Spanish Ministry of Science and Technology under the projects TIN2017-89517-P and TIN2017-83445-P and a grant from the Fondo Europeo de Desarrollo Regional (FEDER). Saeed Ul Hassan was supported by NRPU Grant # 6857, received from Higher Education Commission of Pakistan. Eugenio Martínez Cámara was supported by the Juan de la Cierva Formación Programme (FJCI-2016-28353) of the Spanish government.

References

Altmetric LLP (2017), Fetching detailed article level metrics for an article, https://help.altmetric.com/support/solutions/articles/6000086844-sample-api-response, Retrieved on Dec 19, 2017.

Ananiadou, S., Thompson, P., & Nawaz, R. (2013). Enhancing search: Events and their discourse context. In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 318-334. Springer, Berlin, Heidelberg.

Batista-Navarro, R. T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., Korkontzelos, I. & Ananiadou, S. (2013). Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In International Conference on Intelligent Text Processing and Computational Linguistics, pp. 559-571. Springer, Berlin, Heidelberg.

Bonaccorsi, A., Cicero, T., Haddawy, P., & Hassan, S. U. (2017a). Explaining the transatlantic gap in research excellence. *Scientometrics*, 110(1), 217-241.

Bonaccorsi, A., Haddawy, P., Cicero, T., & Hassan, S. U. (2017b). The solitude of stars. An analysis of the distributed excellence model of European universities. *Journal of Informetrics*, 11(2), 435-454.

Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF). *Journal of Informetrics*, 13(1), 325-340.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

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.

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), 832-843.

Dragoni, M., Federici, M., & Rexha, A. (2019). ReUS: a real-time unsupervised system for monitoring opinion streams. *Cognitive Computation*, 11 (4), 469–488.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of medical Internet research*, 13(4), e123.

Fenner, M. (2013). What can article-level metrics do for you?. *PLoS biology*, 11(10), e1001687.

Friedrich, N., Bowman, T. D., Stock, W. G., & Haustein, S. (2015). Adapting sentiment analysis for tweets linking to scientific papers. *15th International Society of Scientometrics and Informetrics Conference, pp. 107-108.* Istanbul, Turkey.

Federici, M., & Dragoni, M. (2016). A knowledge-based approach for aspect-based opinion mining. In *Semantic Web Evaluation Challenge*, pp. 141-152. Springer, Cham.

Haddawy, P., Hassan, S. U., Abbey, C. W., & Lee, I. B. (2017). Uncovering fine-grained research excellence: The global research benchmarking system. *Journal of Informetrics*, 11(2), 389-406.

Hassan, S. U., Haddawy, P., Kuinkel, P., Degelsegger, A., & Blasy, C. (2012). A bibliometric study of research activity in ASEAN related to the EU in FP7 priority areas. *Scientometrics*, 91(3), 1035-1051.

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.

Hassan, S. U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, 116(2), 973-996.

Hassan, S. U., Sarwar, R., & Muazzam, A. (2016). Tapping into intra-and international collaborations of the Organization of Islamic Cooperation states across science and technology disciplines. *Science and Public Policy*, 43(5), 690-701.

Hassan, S. U., Akram, A., & Haddawy, P. (2017b). Identifying important citations using contextual information from full text. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1-8. IEEE.

Hassan, S. U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy. *Scientometrics*, 103(1), 33-46.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & 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), 232-238.

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. *PloS one*, 10(3), e0120495.

Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, 101(2), 1027-1042.

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. In 2017 Intelligent Systems Conference (IntelliSys), pp. 722-728. IEEE.

Kim, H. J., Lee, J., Chae, D. K., & Kim, S. W. (2018). Crowdsourced promotions in doubt: Analyzing effective crowdsourced promotions. Information Sciences, 432, 185-198.

Konkiel, S. (2016). Altmetrics: diversifying the understanding of influential scholarship. *Palgrave Communications*, 2, Available at http://dx.doi.org/10.1057/palcomms.2016.57

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics, 33(1):159–174.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1-167.

Liu, X. Z., & Fang, H. (2017). What we can learn from tweets linking to research papers. *Scientometrics*, 111(1), 349-369.

Nuzzolese, Andrea Giovanni, Paolo Ciancarini, Aldo Gangemi, Silvio Peroni, Francesco Poggi, and Valentina Presutti. (2019). Do altmetrics work for assessing research quality?." Scientometrics, 119 (2), 539–562

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.

Qiu, J., Lin, Z., & Shuai, Q. (2019). Investigating the opinions distribution in the controversy on social media. Information Sciences, 489, 274-288.

Ravenscroft, J., Liakata, M., Clare, A., & Duma, D. (2017). Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PloS one 12.3*, e0173152.

SentiStrength (2017), SentiStrength http://sentistrength.wlv.ac.uk/, Retrieve on March 10, 2017.

Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. BMC medical informatics and decision making, 18(1), 46. doi:10.1186/s12911-018-0639-1

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163-173.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, 17, 1-8. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=4533177

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.

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2017). Enriching news events with meta-knowledge information. Language Resources and Evaluation, 51(2), 409-438.

Waheed, H., Hassan, S. U., Aljohani, N. R., & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. Behaviour & Information Technology, 37(10-11), 941-957.

Yu, H. (2017). Context of altmetrics data matters: an investigation of count type and user category. *Scientometrics*, 111(1), 267-283.

Zhu, D., Wang, D., Hassan, S. U., & Haddawy, P. (2013). Small-world phenomenon of keywords network based on complex network. Scientometrics, 97(2), 435-442.

Zhu, J., Hassan, S. U., Mirza, H. T., & Xie, Q. (2014). Measuring recent research performance for Chinese universities using bibliometric methods. Scientometrics, 101(1), 429-443.

Appendix A

Table A1 lists all the terms that were added to the existing SentiStrength lexicon file (EmotionLookupTable.txt). Table A2 lists the adapted idioms in SentiStrength lexicons file (IdiomLookupTable.txt), Table A3 lists all the terms that were used as scientific terminologies and were causing false positive or false negative, and thus were removed from the SentiStrength lexicons file (EmotionLookupTable.txt). The original version of SentiStrength lexicons can be downloaded from: http://sentistrength.wlv.ac.uk/SentStrength_Data_Sept2011.zip.

Table A 1. Terms added to SentiStrength lexicon (EmotionLookupTable.txt)

Terms with a positive sentiment				
sober	soberness	fascinating	clearest (clear*[9])	fundamental
novel	novelties	fundamentalness	brac* (bracing, brace)	neat study
ground-breaking	novelness	fascinatingly	sound* (sound, sounding)	groundbreaking, ground breaking
worthy	fundament	fundamentally	astonish* (astonish, astonishing)	believe
big*	watershed	unprecedented	landmark	worthful
clever	astound* (astound, astounding)	comprehensive	serious (sentiment updated from negative to positive)	leap study
soberly	compelling	neat research	great systematic	stunning
novelly	remarkable	leap research	incredible	sobering
elegant	productive	extraordinary	fundamentalist	intriguing
Terms with a negative sentiment				
misreporting	fatuous	biased, bias	flaws	not right
joke (sentiment updated from positive to negative)		pilgrims, plagiarized		

[9] Some of the terms are ending with a wild card *, which means that it can be any word starting with that term, for example the term astonish* means both 'astonish' and 'astonishing'.

Table A 2. Adapted idioms in SentiStrength lexicon (IdiomLookupTable.txt)

Idiom Lookup			
wat up	new evidence	how are you	ground-breaking
new way	new research	what's good	less scientific
whats up	shock horror	game changer	new meta analysis
wuts good	breaking news	new analysis	thought provoking
new study	worth reading	felt compelled	thought-provoking
what's up	feel compelled	whats good	ground breaking
it hanging			

Table A 3. Terms removed from SentiStrength lexicon (EmotionLookupTable.txt)

Emotional Lookup			
bug	pains	confes*	prohibit*
war	rape*	corrupt	sufferer*
fat	shark	decease	suffering
foe	tears	default	hazardous
gay	fatty	disease	incurable
gun	fears	dispute	injurious
ill	fever	invade*	corruption
baby	flame	leakage	partition*
bomb	germs	leaking	slaughter*
burn	grave	molest*	unemployed
bury	abrupt	paining	catastrophe
care	absent	poison*	elimination
clog	addict	pollut*	emergencies
cold	afraid	poverty	incompatib*
dead	attack	prison*	compel (changed from negative to positive)
deny	babies	rapist*	dizzy
drag	bother	suffers	enemy
drop	brutal	suicide	jail*
duty	burden	terror*	kick*

envy	cancel	thirsty	adverse
leak	cancer	victim*	against
pain	charge	weapon*	capture
shy*	costly	fatigue	collide
evil	danger	illegal	Emotional

Table A 3. Terms removed from SentiStrength lexicon (EmotionLookupTable.txt) – continued

Emotional Lookup			
fear	injury	illness	outbreak*
feud	mourn*	jobless	paralysis
fist	pained	accident	paralyzed
germ	painf*	cannibal	hunter
grab	raping	casualty	arbitrary
hazy	severe	collapse	crime
homo	spill*	comfort*	death
alarm	suffer	contrary	decay
alien	terror	criminal	devil
argue	thirst	decrease	infect
avoid	feared	disorder	injure
blunt	fierce	pressur*	abandon
bribe	fought	suffered	absence
broke	gunmen	symptom*	collision
cared	hazard	terrori*	dizziness
chase	hunger	haziness	eliminate
choke	hungr*	abnormal*	emergency
cramp			