Exploiting tweet sentiments in altmetrics large-scale data

Saeed-Ul Hassan
Department of Computer Science, Information Technology University, Pakistan

Naif Radi Aljohani
Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia

Usman Iqbal Tarar
Department of Computer Science, Information Technology University, Pakistan

Iqra Safder
FAST School of Computing, FAST-NU Lahore, Pakistan

Raheem Sarwar
Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, United Kingdom

Salem Alelyani
Center for Artificial Intelligence (CAI), King Khalid University, Saudi Arabia; College of Computer Science, King Khalid University, Saudi Arabia

Raheel Nawaz
Staffordshire University, United Kingdom

Abstract
This article aims to exploit social exchanges on scientific literature, specifically tweets, to analyse social media users’ sentiments towards publications within a research field. First, we employ the SentiStrength tool, extended with newly created lexicon terms, to classify the sentiments of 6,482,260 tweets associated with 1,083,535 publications provided by Altmetric.com. Then, we propose harmonic means-based statistical measures to generate a specialised lexicon, using positive and negative sentiment scores and frequency metrics. Next, we adopt a novel article-level summarisation approach to domain-level sentiment analysis to gauge the opinion of social media users on Twitter about the scientific literature. Last, we propose and employ an aspect-based analytical approach to mine users’ expressions relating to various aspects of the article, such as tweets on its title, abstract, methodology, conclusion or results section. We show that research communities exhibit dissimilar sentiments towards their respective fields. The analysis of the field-wise distribution of article aspects shows that in Medicine, Economics, Business and Decision Sciences, tweet aspects are focused on the results section. In contrast, in Physics and Astronomy, Materials Sciences and Computer Science, these aspects are focused on the methodology section. Overall, the study helps us to understand the sentiments of online social exchanges of the scientific community on scientific literature. Specifically, such a fine-grained analysis may help research communities in improving their social media exchanges about the scientific articles to disseminate their scientific findings effectively and to further increase their societal impact.

Corresponding author:
Raheem Sarwar, Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, UK.
Email: R.Sarwar@mmu.ac.uk
1. Introduction

Traditionally, research impact has used citation as the main indicator of research’s standing; however, it takes years to see any measurable impact. On the contrary, researchers are increasingly going online to find and share information about science, as well as have been urged to consider how they can use social media platforms to engage with each other. With the increased usage of social media platforms for scholarly communications, altmetrics data are of enhanced interest as it captures real-time scholarly communication data from online platforms (e.g., Twitter and Facebook) and may be used as an early measure of the research impact. Scholars frequently use Twitter as a discussion platform to share their opinions on research. Perhaps, for this reason, digital libraries and journal websites are increasingly using tweet counts as a measure of the impact of research.

Altmetrics is the collective domain of social media platforms such as Twitter, Facebook, CiteULike, and MendeleyReadership in relation to research articles to provide metrics on their research impact [1–3]. Among several platforms, Twitter is widely used by scholars to share their opinions concerning research articles [4]. Recent studies show that tweet sentiments can help predict the early impact of the research articles. Specifically, the papers cited in positive and neutral tweets have a greater impact than those not cited or cited in a negative tweet. However, there is still a need to investigate tweeter data to analyse user sentiments relating to research articles in specific fields. Such a fine-grained investigation is required to fully utilise the findings of existing studies that may help research communities in improving their social media exchanges about the scientific articles to disseminate their scientific findings effectively and increase their research impact.

Specifically, we answer the following research questions in this article:

1. What are the differences among research communities of different domains regarding tweets containing positive, negative and neutral sentiments?
2. Are different research communities inclined towards different aspects of the articles such as methodology or conclusions?

As mentioned earlier, this article presents a quantitative study to exploit tweet data to analyse user sentiments relating to different aspects of research articles in specific fields. This study helps us to understand the sentiments of online social exchanges of the scientific community on scientific literature, specifically the sentiment of tweets, for better visibility and qualitative assessment of these interesting big data of altmetrics. We identify the sentiment of research communities with respect to their respective fields and to conduct an aspect-based analysis of user expressions related to their research articles. Such a fine-grained analysis may help research communities in improving their social media exchanges about the scientific articles to disseminate their scientific findings effectively and increase their impact.

The following are the three main contributions of the study:

- Lexicon generation: We design a harmonic means (HMs)–based statistical measure to generate a specialised lexicon to conduct this investigation, which helps improve the performance of the sentiment-analysis task. This is because general sentiment lexicons calculate the sentiment tendency of a word without considering domain knowledge. However, the sentiment contained in just a few words is inevitably domain-dependent. Therefore, generic sentiment lexicons used by SentiStrength report poor performance in various applications. For this reason, in this investigation, we design a new measure to generate a new lexicon for our altmetrics data to determine both domain-specific and expressive terms and then feed it to SentiStrength to identify the sentiments of the tweets. Specifically, we computed the rate and frequency metrics of each term or ‘token’. Next, we compiled statistical measures, such as the HM, using a cumulative distributive function for both positive and negative terms. The resulting descending-order list of lexicon terms shows the most meaningful and domain-dependent tokens in sentiment expressions and provides meaningful insights into the terms used in opinion mining in this altmetrics domain.

- Based on our newly generated lexicon, we designed a threshold-based mechanism to compute domain-wise article-level sentiment. We found that research communities exhibit dissimilar sentiment towards their respective fields.

- We design a method to perform an aspect-based analysis of user expressions related to the research article, such as its title, abstract, methodology, conclusion and results. We found that research communities focus on different
aspects of an article. Researchers in fields such as Medicine and Economics, Business and Decision Sciences show more interest in an article’s findings than its title, abstract or methodology. Interestingly, fields such as Engineering and Computer Sciences address more the techniques designed. Likewise, in Health Professions and Nursing, the scientific community primarily discusses articles on the basis of both their abstract and their findings.

The structure of the rest of this article is as follows: section 2 includes previous research work, associated concepts and a literature review. In section 3, we present the details of our dataset, followed by a discussion on our approach to lexicon creation and tweet sentiment analysis. Section 4 presents our data and insights on the results achieved. We end this study with concluding remarks.

2. Literature review

Altmetrics has a very broad scope, and many studies have been undertaken to define the extent of the term, the type of research measures that it may or may not provide, and whether there is enough data to indicate any impact [5]. Altmetrics can be regarded as an umbrella term for an article-level metrics of research impact that encompasses several social media platforms, such as Twitter, Facebook, MendeleyReaders, CiteULike and Google+ [6,7]. Altmetrics data are increasing all the time, and multiple organisations gather them, including Altmetric.com, Impact Story and Plum Analytics. These organisations collect all online activities concerning research articles and offer these data for research purposes. We have observed a promising increase in research into sentiment analysis and opinion mining of altmetrics dataset of researchers, publishers, universities and funders in the past few years, hence there is a growing demand for standards and new challenges to ensure best practice [7–9]. In the following subsections, we provide a brief overview of previous studies to highlight the quality and challenges of our altmetrics dataset and approaches that we used in sentiment analysis of Twitter altmetrics.

2.1. A brief review of altmetrics

Researchers and academics are increasingly using online research tools to access, download, bookmark, recommend, discuss, share and evaluate ongoing research. Through their online presence, they are creating huge volumes of online data that can be used in altmetrics. Traditionally, the relevance and actual impact of a research article have been gauged by its citation count, but this has the inherent problem of being sluggish. The use of this conventional citation metrics may be superseded by mining altmetrics data, as this can produce useful insights [10–15].

Of the altmetrics indicators such as Facebook, Google+, CiteULike, Mendeley, Wikipedia and other online blogs, Twitter is widely used by scholars and researchers, and many studies have investigated this use. Priem and Costello’s investigated 46,515 tweets from a sample of 28 scholars and examined their attitudes and practice to Twitter for scholarly discussion [4]. It explored how often they tweeted research articles, and the results revealed that, while they use it in this way, such a citation is different from the traditional citation. The study concluded that Twitter citations are much more rapid and that Twitter does indeed have an impact on scientific research. To find any common pattern of use among the disciplines or whether they are clearly different, a cross-disciplinary analysis was performed on how and why researchers use Twitter [16]. They analysed 10 diverse disciplines and categorised the tweets of selected scholars as Scholarly communication, Discipline-relevant, Not clear and Not about science. Their results show a clear difference in Twitter usage between scholars from these various disciplines. Priem and Costello discuss the quantity and quality of altmetrics data that are generated over the years [4]. As well as citation metrics, the authors correlated article-level metrics on various altmetrics platforms. They answered the main question, whether it can predict citation counts and is indeed an early measure of research impact, as their comparison of altmetrics and traditional citation revealed its significant contribution to the early prediction of citations. However, they concluded that altmetrics is different from citation count, as the impact that is captured is across a highly varied audience, which may suggest a much wider societal impact in multiple educational, cultural, environmental and economic fields. A recent study discussed both how social media signals are revealed in various scientific fields and that they differ by document type [17]. The results indicate that, in general, mentions of research articles on online platforms are somewhat low; however, Twitter has the best coverage of all social media platforms. The study also explored which altmetrics indicators have the most significant connection to citation count, and concluded that Twitter and online blogging have the best correlation with traditional metrics. Further analysis showed that shorter documents, such as editorials, news articles and letters, tend to receive more online coverage than longer, more complex documents [17].
2.2. Tweet sentiments of altmetrics

Sentiment-analysis algorithms either rely on machine learning or lexical methods. The machine learning methods partition text into words or word n-grams, learn which of these features are associated with sentiments based on human-coded text and use this information to predict the sentiment of the test sample. On the contrary, lexical methods use a list of sentiment words and their polarities with grammar structure knowledge such as a negation role to predict the sentiment of the text. Nevertheless, lexical methods report better accuracy for social media texts and are less probable to choose indirect indicators of sentiment that generate spurious sentiment patterns. For instance, machine learning methods may choose unpopular politicians’ names as negative features since they tend to occur in the negative text [18–20]. Typically, people use shortened forms of words and emoticons when writing on social web platforms, which increases the need to create tools to identify feelings in a short text [21,22]. Thelwall, Buckley and Paltoglou devised an algorithm, ‘SentiStrength’, that works in both supervised and unsupervised cases [23]. It adopts a lexical approach in which a list of terms is tagged with positive or negative sentiments on a scale of –5 to +5 and, on the basis of the occurrence of these terms, it predicts the sentiment of a text. The lexicon model may include additional information, such as emoticon lists and semantic rules for dealing with negation words. The SentiStrength algorithm shows good results when performing sentiment analysis on the datasets of web networks (Myspace, Twitter, Facebook, YouTube, BBC Forums). It works well with social web data for which no training dataset is available to detect sentiment, thus are recommended for applications in which direct, effective terms are exploited by performing sentiment analysis.

Scholars frequently use Twitter as a discussion platform to share their opinions on research. Perhaps, for this reason, digital libraries and journal websites are increasingly using tweet counts as a measure of the impact of research. To evaluate its use as an alternative measure of impact, several studies have raised the need to analyse the opinion expressed in tweets about articles. Researchers analysed the tweets of articles and reviews published in 2012 in WoS, as captured by Altmetric.com [24]. The dataset consisted of 487,610 tweets, mentioning 192,832 articles. The results showed that 11.0% contain positive sentiments and 7.3% negative, and 81.7% are neutral. Disciplinary analysis shows that fields such as Psychology, the Humanities and Social Sciences contain the most sentiment in their tweets, while fields such as Physics, Chemistry and Engineering express the least [25]. A recent study states that the Twitter-user influence score is a highly important feature in the classification of highly cited articles [26].

In addition, to ascertain scholarly impact through altmetrics events, there are challenges to be addressed. Studies have provided evidence that it is not actually the scientific merits or characteristics of an article that is captured by online or social media attention. A study reports that a curious or humorous article receives more tweets and that scientific journals may use social media as a platform for their promotional campaigns, creating an enhanced level of altmetrics events about certain research [27]. A study pointed out that usage of scholarly online and social platforms is almost devoid of sentiment and, in most cases, it offers no opinion [28]. However, citation presents the same issue: a study [29] revealed that the intentions behind creating a citation vary, and some actually relate to something other than the research itself. Researchers observed that long abstracts of medically related articles receive more citations, whereas longer titles in Psychology receive fewer [30]. Since many studies have explored techniques of sentiment analysis, certain aspects of citations using altmetrics data show a marked variation, aside from their scientific merit and approach. Most measurement of the sentiment and opinion of the people tweeting about research has been carried out quantitatively. Our study takes a more qualitative approach, exploiting tweet sentiment and opinion mining at a higher level, using document-level sentiment analysis and aspect-based sentiment analysis. This qualitative content analysis could introduce new viewpoints to altmetrics research.

3. Dataset and methodology

In this section, we discuss the proposed method. The proposed method consists of five parts: altmetrics data collection, tweet pre-processing, lexicon generation, combining article-level tweets and analysis (see Figure 1). Each part of the proposed method is explained in the following subsections.

3.1. Dataset

The corpus comprised altmetrics data collected by Altmetric.com from July 2011 to June 2016. Note that Altmetrics.com is the most important collector of social media content, offering these data for research purposes. The database consists of aggregated content from online platforms such as Twitter, Google+, Facebook and CiteULike. Twitter is the chief contributor. From the altmetrics data, we extracted 1,083,535 research articles that each had at least one citation and one tweet. While using the tweet URL, we fetched 6,482,260 tweets from Twitter, we retrieved the
articles’ citation count using Scopus API along with the disciplinary information provided by the Scopus subject-category scheme.

For cross-disciplinary analysis, the dataset was divided into scholarly disciplines by the All Science Subject Classification (ASJC) scheme. Inspired by a recent work [31], the top-level ASJC disciplines were merged into 16 disciplines by combining Agricultural, Biological Sciences and Veterinary; Biochemistry, Genetics and Molecular Biology; Chemistry; Computer Science; Earth Planetary Sciences; Engineering; Environmental Science; Economics, Business and Decision Sciences; General; Material Science; Health Professions and Nursing; Mathematics; Medicine and Medical Sciences; Physics and Astronomy; Social Sciences; and Other Life and Health Sciences.

3.2. Pre-processing

To demonstrate the need for pre-processing, Table 1 shows a few examples of the unprocessed tweet text. To obtain the clean text for lexicon creation and sentiment analysis, we performed the following pre-processing steps: (1) we detected and removed all non-English tweets; (2) since tweet text sometimes contains research-specific terms taken from the article’s title that are not actual opinion specific to the research article, we removed any such terms to avoid false allocation of sentiment [28]; (3) we used Beautiful Soup Python Library to decode HTML encoding, such as ‘&amp’, ‘&quot’ and so on, into general text; (4) we removed tags like ‘@mention’ from the tweet text using regular expressions (REs) ‘(r’ @[A-Za-z0-9]*)’; (5) we removed URLs using REs (r’ https://[A-Za-z0-9/.]+ and r’ www.[^] +’); (6) we found and removed any Unicode Transformation Format (UTF)-8 encoding patterns of characters ‘xef’xbf’xbd’ using UTF decoding; (7) we kept numbers as text, only removing the ‘#’ character using REs (^[^a-zA-Z]Z’; (8) we dropped any duplicate and null-text tweets; (9) we carefully handled negation words to avoid their destruction in pre-processing by preparing a
list of common negations (words with apostrophic combination), such as isn’t (is not), aren’t (are not), wasn’t (was not), weren’t (were not), haven’t (have not), hasn’t (has not), couldn’t (could not), shouldn’t (should not) and so on, converting them into two words; (10) last, we removed unnecessary blank spaces, performed tokenisation and lowercasing, and rejoined tokens to form proper sentences.

3.3. SentiStrength

Exploiting the tweet sentiment in altmetrics data requires a sentiment-analysis tool that performs well on social media text which is generally short and contains non-textual elements such as emoticons and is categorised as non-standard expressive text. In addition to this, it requires a sentiment-analysis tool that can determine the positive and negative sentiments simultaneously. This is because psychological research reports that humans can experience negative and positive emotions simultaneously. Furthermore, it requires a sentiment-analysis tool that works well with low or no training data. This is because, for some fields, there is less amount of Altmetrics data available for analysis. Unlike machine learning–based sentiment-analysis tools, the SentiStrength tools, which are a lexical method, have all these properties. SentiStrength uses a lexical approach to identify the sentiments of social media texts. Specifically, it simultaneously determines the strength of positive (on a scale of 1 − 5) and negative (on a scale of −1 to −5) emotions because of psychological research reports which state that humans can experience negative and positive emotions simultaneously.

3.4. Lexicon creation of tweets in altmetrics

By adapting to the SentiStrength tool for sentiment analysis of Altmetrics data, our whole data are tagged into positive, negative and neutral sentiments. Since SentiStrength is a generic lexicon-based tool, the use of words varies a lot from topic to topic. Therefore, generic sentiment lexicons used by SentiStrength report poor performance in various applications. Thus, we propose an improved scoring method for our Altmetrics corpus so as to see the most expressive terms of opinion for both positive and negative sentiments simultaneously. Specifically, we design a HMs-based statistical measure to generate a specialised lexicon to conduct this investigation which helps improve the performance of the sentiment-analysis task. Moreover specifically, we design a new measure to generate a new lexicon for our altmetrics data to determine both domain-specific and expressive terms and then feed it to SentiStrength to identify the sentiments of the tweets (see Algorithm 1).

We extracted 152,673 words/features from our dataset using the Python Count Vectorizer method. The Python Counter Vectorizer converts a collection of text documents to a matrix of token counts where the stopwords were ignored. In addition to this, Count Vectorizer uses an analyzer that does feature selection. Consequently, the resulting features are lower than the vocabulary size found by analysing the data. Our intuition is that, if a word appears more frequently in positive class as compared with a negative one, then it should be more characterised by a positive term.
Similarly, if a word appears more frequently in negative class as compared with the positive one, then it should be more characterised by negative terms. Thus, for each term in our dataset, we calculated Positive Rate (PR) and Negative Rate (NR). The PR of a term is calculated as the ratio of the relative frequency of the term in positively identified texts to the frequency of the term in all texts (see equation (1)), while its NR is the ratio of the relative frequency of the term in negatively identified texts to the total frequency of the term in all texts (see equation (2)).

\[
PR = \frac{\text{Positive frequency}}{\text{Total Frequency}} \tag{1}
\]

\[
NR = \frac{\text{Negative frequency}}{\text{Total Frequency}} \tag{2}
\]

We then sorted the terms by the rates and found no meaningful pattern in the top-scoring terms. Specifically, we found that words with the highest PR have zero frequency in negative tweets, but the overall frequency of these words is too low to consider it as a guideline. Next, we ascertained the rate of occurrence within a class by calculating Positive Frequency (PF) and Negative Frequency (NF) metrics, as shown in equations (3) and (4). This new metric resulted in almost the same ranking as the original term frequencies.

\[
PF = \frac{\text{Positive Frequency}}{\text{Sum of positive frequencies}} \tag{3}
\]

\[
NF = \frac{\text{Negative Frequency}}{\text{Sum of negative frequencies}} \tag{4}
\]

Since our intuition is to rank terms in order of their positive sentiment value, we generate the cumulative distribution function (CDF) values of PR and PF for the positive sentiment value, and CDF values of NR and NF for the negative sentiment values. CDF is a probability distribution function of \(X\) that is evaluated at \(x\), and it measures the probability that \(X\) will take a value less than or equal to \(x\), as shown in equation (5).

\[
F(x) = (X \leq x) \tag{5}
\]

The calculation of CDF value of PR or PF provides insight into their ranks in the distribution. Next, we combine CDF of PR and CDF of PF together to produce a metric that has a reflection of both PR and PF. That is, CDF helps find terms’ associations using their rate and frequency values. For instance, the term ‘Excellent’ scored 0.83 CDF of PR value and 1.00 CDF of PF. This means that roughly 83% of tokens will have a PR value of less than or equal to 0.99 and, for PF, all have a PF value of less than or equal to 0.001786. The CDF is used here to give the cumulative values of the distribution of PR and PF.

Next, we combine PR-CDF and PF-CDF together to produce a metric that has a reflection of both PR and the PF. Upon looking at the values, we found that the PR-CDF spans from 0 to 1 and the PF-CDF values are distributed in a smaller range, that is, 0–0.4. Consequently, taking the arithmetic average of these two numbers will dominate the PR over the PF value, thus instead we rely on the HM. Finally, we computed the HM of the CDF values for both the rate and frequency metrics. HM is the reciprocal of the arithmetic mean of that reciprocal. It is appropriate to use the HM when the metrics include outliers that could skew the results. Equations (6) and (7) show the HM for positive (HMP) and negative (HMN) terms, respectively, while \(n\) represents the number of metrics.

\[
\text{HMP(P) = Positive Score} = \frac{n}{\frac{1}{PR_{CDF}} + \frac{1}{PF_{CDF}}} = n(PR_{CDF}.PF_{CDF}) \frac{n(PR_{CDF}.PF_{CDF})}{PR_{CDF} + PF_{CDF}} \tag{6}
\]

\[
\text{HMN(N) = Negative Score} = \frac{n}{\frac{1}{NR_{CDF}} + \frac{1}{NF_{CDF}}} = n(PR_{CDF}.PF_{CDF}) \frac{n(PR_{CDF}.PF_{CDF})}{NR_{CDF} + NF_{CDF}} \tag{7}
\]

It is important to note that HM works the same as the F-score in terms of precision and recall metrics. Therefore, HM supports a cumulative score for all terms, providing a useful scoring mechanism for our tokens, as the descending-order list shows the most meaningful and domain-dependent tokens in sentiment expressions. Appendix 1 (Tables 6 and 7) lists the top-100 positive and negative words in our altmetrics dataset.
3.5. Article-level sentiments of tweets in altmetrics

To analyse the article-level sentiment of our altmetrics data on the basis of their Alt_ID, we combined the tweets about each article with at least 30 tweets. The objective of this level of analysis is to express a single sentiment for the whole article, and it assumes that all the sentences within a document refer to a single entity. We had a total of 61,233 distinct Alt_IDs for each research article with at least 30 tweets, and we computed sentiment scores for each using our newly created lexicon in SentiStrength. Once the scores were applied to the positive and negative terms, we achieved an average that ranged from 0.0 to 1.0. We referred to those values above 0.7 as positive and those below 0.3 as negative, and scores between the two as neutral.

3.6. Aspect analysis of tweets in altmetrics

An opinion can be defined as a quintuple: \((e_i, a_{ij}, h_k, t_l)\) [32]. Here, \(e_i\) and \(a_{ij}\) together represent the opinion target, where \(e_i\) is the entity as the main target of opinion, \(a_{ij}\) is an aspect of entity \(e_i\) for which opinion is being generated, \(h_k\) is the opinion holder and \(t_l\) is the time when the opinion is expressed by \(h_k\).

Using the above definitions, we performed domain-wise, article-level, aspect-based analysis of our altmetrics data. In this instance, the entity was a research article and the aspects were the title of the article, its abstract, methodology and the conclusion discussed at the end. The objective was to gauge community behaviour in tweeting about an article, by domain. First, on the basis of their Alt_IDS and domain code (QRR_IDS), we combined all tweets about each article with at least 30 tweets. Note that an article may fall into multiple domains, so the combined sum of articles (Alt_IDS) was 153,336, using the standard double counting method. We also identified the various aspects of an article that were expressed by researchers in their tweets, as typically stated in the keywords, as shown in Table 2. For every tweet in which the opinion referred to the entire article, that opinion was marked as a general aspect of the article.

4. Experimental results

In this section, we discuss the results obtained by our various analysis techniques, along with their significance to the different domains.

4.1. Distribution of tweet sentiments

Using SentiStrength with the domain- and emotion-specific terms, as prescribed in Hassan et al. [33], we classified as positive, negative and neutral a total of 6,482,260 tweets, relating to 1,083,535 altmetrics documents. We found that 22% were positive, around 14% negative and 64% neutral, as shown in Figure 2. Furthermore, we explored our altmetrics
tweets dataset to detect any significant change in behaviour in the usage of tweet sentiment. Figure 2 illustrates that there was no significant increase in tweet sentiment during the period 2012–2016.

4.2. Lexicon for altmetrics data for sentiment scores

SentiStrength is a generic lexicon-based tool with a generic text corpus. Since each text corpus is different in its nature and the use of words in subjects varies widely, we created a relative scoring technique based on our altmetrics corpus. We calculated the HM of CDF scores (normCDF_HM) for PR and PF, and NR and NF. The normCDF_HM provides a significant scoring pattern for the corpus unique terms. Table 3 gives the descending list of the most meaningful tokens in our corpus in terms of sentiment expression. Appendix 1 (Tables 6 and 7) contains the top-50 positive and negative tokens in the altmetrics dataset. Figure 3 illustrates the interesting pattern displayed by the normCDF_HM scores for both the rate and frequency metrics. The tokens shown at the top left are more positive, and the ones at the lower right are more negative. In this way, we created our own lexicon for the altmetrics corpus, and it will prove useful in the classification of tweets sentiment in future.

4.3. Article-level summarisation for altmetrics domains

To perform article-level summarisation, using SentiStrength, we combined all tweets about an article with at least 30 tweets into a single document and computed the document-level sentiment. To obtain reliable sentiment information for an article, we considered articles that have at least 30 tweets. Of the total of 61,233 unique articles, we found that around

### Table 3. Terms in descending order of positive harmonic mean.

| Token      | Total count | Positive HM score | Negative HM score | Token      | Total count | Positive HM score | Negative HM score |
|------------|-------------|-------------------|-------------------|------------|-------------|-------------------|-------------------|
| Excellent  | 6669        | 0.9077            | 0.2533            | Positive   | 4767        | 0.9008            | 0.2737            |
| Novel      | 3713        | 0.9064            | 0.2558            | Nice       | 14,630      | 0.9007            | 0.2886            |
| Amazing    | 2907        | 0.9054            | 0.2578            | Thanks     | 6169        | 0.9006            | 0.2768            |
| Congratulations | 1930 | 0.9051 | 0.2487 | Special  | 3458        | 0.8997            | 0.2739            |
| Awesome    | 3697        | 0.9040            | 0.2627            | Cool       | 9996        | 0.8991            | 0.2855            |
| Success    | 2319        | 0.9031            | 0.2606            | Interesting| 33,787      | 0.8987            | 0.3086            |
| Wow        | 3296        | 0.9030            | 0.2648            | Greater    | 4340        | 0.8963            | 0.2856            |
| Interested | 4181        | 0.9024            | 0.2678            | Hope       | 2277        | 0.8955            | 0.2779            |
| Exciting   | 2343        | 0.9023            | 0.2628            | Love       | 4026        | 0.8947            | 0.2892            |
| Great      | 25,625      | 0.9013            | 0.2952            | Pretty     | 2171        | 0.8937            | 0.2803            |

HM: harmonic mean.
82.55% contained neutral sentiments, followed by 17.35% with positive sentiments and only 0.1% of the articles were negative. The results suggest that, at article-level, the negative sentiments are quite insignificant.

In addition, we performed domain-level sentiment analysis using article-level summarisation in order to measure user behaviour across the domains. We aggregated article-level tweet documents on the basis of QRR_Field, and used SentiStrength, enriched by the new lexicons, to calculate the positive, negative and neutral sentiment scores. Table 4 gives a summary of the results, along with the normalised positive and negative sentiment scores from 0 to 1. Since the entity is not supposed to be single, we do not attempt to suggest that domain-level summarisation will give an opinion about the domain. Rather, it helps to show the intent and to indicate the behaviour of the users by their domain.

The results show that researchers expressed more positive opinion in domains such as Arts and Humanities, Computer Science and Chemistry, while the fields of Medicine, Health Professions and Nursing and Other Life and Health Sciences attracted more negative opinions from their respective scientific communities. Figure 4 presents a scatter plot to illustrate the community behaviours in each research domain. With reference to normalised sentiment score (HM Score), the domains expressing more positive opinions are at the top left, while those with a high value of negative sentiment are at lower right.

---

**Table 4.** Summary of domain-level sentiment-analysis results.

| Domain                                | No. of documents with at least 30 tweets | Avg. positive sentiment score | Avg. negative sentiment score |
|---------------------------------------|------------------------------------------|-------------------------------|-------------------------------|
| Arts and Humanities                   | 468                                      | 0.66                          | 0.46                          |
| Computer Science                      | 979                                      | 0.66                          | 0.45                          |
| Chemistry                             | 1414                                     | 0.66                          | 0.46                          |
| Mathematics                           | 950                                      | 0.65                          | 0.45                          |
| Physics and Astronomy                 | 1208                                     | 0.65                          | 0.46                          |
| Materials Sciences                    | 1094                                     | 0.65                          | 0.45                          |
| Economics, Business and Decision Sciences | 924                                      | 0.65                          | 0.47                          |
| Engineering                           | 2491                                     | 0.65                          | 0.46                          |
| Earth and Planetary Sciences          | 1438                                     | 0.64                          | 0.48                          |
| Environmental Sciences                | 2753                                     | 0.64                          | 0.49                          |
| Social Sciences                       | 6187                                     | 0.63                          | 0.49                          |
| Agricultural, Biological Science and Veterinary | 11,975                                   | 0.63                          | 0.48                          |
| General (Science, Nature, PNAS)      | 4520                                     | 0.63                          | 0.49                          |
| Biochemistry, Genetics and Molecular Biology | 17,047                                   | 0.63                          | 0.49                          |
| Other Life and Health Sciences        | 54,555                                   | 0.62                          | 0.50                          |
| Health Professions and Nursing        | 6068                                     | 0.61                          | 0.50                          |
| Medicine                              | 39,265                                   | 0.61                          | 0.51                          |

---

**Figure 4.** Scatter plot of research domains in terms of positive and negative sentiment.
Furthermore, we employed distribution analysis to see the difference from a normal distribution of Alt-Domains by fitting the tweet scores to a bell curve, as shown in Figure 5. The results indicate that Twitter usage in the domains of Arts and Humanities, Chemistry, Computer Sciences, Material Sciences and Mathematics are positive, while in Medicine, Health Professions and Nursing, and other Life and Health Sciences, it is towards the negative. It was found that domains such as Earth and Planetary Sciences are neutral, overall.

4.4. Article-level, aspect-based analysis

For domain-wise, aspect-based analysis, on the basis of their research domains, we compiled article-level tweet documents for the 61,233 unique articles in our altmetrics dataset that had at least 30 tweets each to obtain reliable sentiment information. Note that an article may fall into multiple domains, so the combined sum of articles is 153,336. Table 4 presents a summary of the results for article-level tweets that contain the users’ opinion of the title, the abstract, methodology, conclusion and results. Note that we created a separate category ‘Other’, for where a whole article is discussed in general. For articles in our dataset with at least 30 tweets, Table 5 shows the proportion that specifically discusses their various aspects in terms of their respective subject domain.

Regarding using the title in expressing an opinion, across the fields we found that General (Science, Nature, PNAS) was prominent, with 2.47% articles being debated in this way, followed by Arts and Humanities and Social Sciences, with 2.29% and 2.09% articles, respectively. In terms of abstract-based opinion, the domain of Health Professions and Nursing is significant, with 5.39% articles debated on this basis, followed by Arts and Humanities at 5.20% and Computer Sciences at 4.98%. We noted that in Material Sciences, 12.04% of article tweets concentrated on the article’s methodology, and in Physics and Astronomy and Chemistry, this was over 8%. Interestingly, it is important to note that researchers in fields such as Engineering and Computer Sciences address the techniques designed relatively more. This might be due to the fact that in these fields, the tasks at hand and the outputs are known a priori, and novel techniques are designed to achieve the desired output. For example, the authorship attribution task aims at identifying the original author of the anonymous text from a set of candidate authors. Researchers propose novel techniques to perform this task. Furthermore, in terms of debating an article on the basis of its results and conclusions aspects, we found that the domain of Economics and Business and Decision Sciences was the most notable of all domains, at 11.75%. Similarly, this domain (11.75%), Medicine (10.84%), Health Professions and Nursing (10.11%) and General (Science, Nature, PNAS).
Table 5. Summary of aspect-based analysis (all the numbers except the no. of documents are percentages).

| Questions-To-Reply-Rate-(QRR) field | No. of documents with at least 30 tweets | Title | Abstract | Methodology | Results and conclusions | Others removed |
|------------------------------------|-----------------------------------------|-------|----------|-------------|------------------------|----------------|
| Other Life and Health Sciences     | 54,555                                  | 1.43  | 3.56     | 3.73        | 9.53                   | 81.76          |
| Medicine                           | 39,265                                  | 1.38  | 3.32     | 3.01        | 10.84                  | 81.45          |
| Agricultural, Biological Sciences and Veterinary | 17,047                                  | 1.56  | 3.02     | 4.88        | 7.13                   | 83.40          |
| Social Sciences                    | 6187                                    | 2.01  | 4.80     | 3.66        | 8.64                   | 80.89          |
| Health Professions and Nursing     | 6068                                    | 1.49  | 5.39     | 2.70        | 10.11                  | 80.31          |
| General (Science, Nature, PNAS)    | 4520                                    | 2.74  | 4.19     | 5.90        | 10.03                  | 77.13          |
| Environmental Sciences             | 2753                                    | 0.90  | 4.05     | 2.97        | 7.74                   | 84.34          |
| Engineering                        | 2491                                    | 1.47  | 2.18     | 8.18        | 4.01                   | 84.15          |
| Earth and Planetary Sciences       | 1438                                    | 1.17  | 3.37     | 2.13        | 8.19                   | 85.13          |
| Chemistry                          | 1414                                    | 1.40  | 2.79     | 8.66        | 3.70                   | 83.45          |
| Physics and Astronomy              | 1208                                    | 1.47  | 4.17     | 8.74        | 4.00                   | 81.62          |
| Materials Sciences                 | 1094                                    | 1.09  | 1.90     | 12.04       | 1.90                   | 83.08          |
| Computer Science                   | 979                                     | 1.59  | 4.98     | 7.36        | 6.57                   | 79.50          |
| Mathematics                        | 950                                     | 1.14  | 2.79     | 7.64        | 5.68                   | 82.77          |
| Economics, Business and Decision Sciences | 924                                     | 1.71  | 3.10     | 4.06        | 11.75                  | 79.38          |
| Arts and Humanities                | 468                                     | 2.29  | 5.20     | 3.74        | 7.07                   | 81.70          |

(10.03%) appear to be most concerned to address aspects of articles’ results and conclusions. This might be due to the fact that the outcome of the economic policies, business decisions and new medicines are not known a priori.

The analysis of the field-wise distribution of an article’s aspects shows that in Medicine, Economics, and Business and Decision Sciences, researchers show more interest in the findings than the title, abstract or methodology. Likewise, the Health Professions and Nursing scientific community primarily discusses articles’ abstracts and findings. Those in General (Science, Nature, PNAS) are more focused on an article’s title and research results than are other fields. In the case of Material Sciences, 12% of all articles are debated on the basis of their methodology. This clearly indicates that this community is much concerned with the methods that are designed and presented by an article. Overall, the analysis suggests that researchers appear to be descriptive when exploring the various aspects of an article.

4.5. Discussion

With the increased usage of the social media platforms for scholarly communications, altmetrics data are of enhanced interest as it captures real-time scholarly communication data from online platforms and may be used as an early measure of the research impact. However, there is still a need to investigate tweeter data to analyse user sentiments relating to research articles in specific fields. Such a fine-grained investigation is required to fully utilise the findings of existing studies. We identify the sentiment of research communities with respect to their respective fields and to conduct an aspect-based sentiment analysis of user expressions related to their research articles. Such a fine-grained analysis may help research communities in improving their social media exchanges about the scientific articles to disseminate their scientific findings effectively and increase its impact.

We found that (1) Twitter usage in the domains of Mathematics, Engineering and Agriculture is inclined to the positive, while in Medicine and Environmental Sciences, it tends towards the negative. Fields of research such as Chemistry and the Social Sciences were found overall to be neutral. Thus, research communities exhibit dissimilar sentiment towards their respective fields; and (2) most tweets discuss research articles as a whole document; however, we saw a significant number where a specific aspect was discussed. Positive sentiment in tweets was observed to be more probable than negative. While the neutral sentiment is normally dominant in the whole-topic discussion, in aspect-based sentiment analysis, it is almost matched by other sentiment expressions. This shows that the Twitter user is inclined to be specific in his or her opinion when discussing the aspects of an article.

5. Concluding remarks

We design HMs-based statistical measures to generate a specialised lexicon to conduct this investigation which helps improve the performance of the sentiment-analysis task. Based on our newly generated lexicon, we designed a threshold-
Document-level sentiment analysis was used to establish any correlation between sentiment score and citation score. For this purpose, the documents were allocated to three bins on the basis of their score: above 0.85; above 0.8 and above 0.75. Using correlation analysis, we found that highly positive documents, those scoring over 0.85, showed a moderate correlation to citation score. This suggests that positive sentiment in a tweet about a research article does indeed predict the article’s popularity and has some relationship to it receiving somewhat more citations.

We also design a method to perform an aspect-based analysis of user expressions related to the research article, such as its title, abstract, methodology, conclusion and results. Various aspects of research articles were explored to examine which parts are commented upon by researchers in tweets. The results show that most tweets discuss research articles as a whole document; however, we saw a significant number where a specific aspect was discussed. Positive sentiment in tweets was observed to be more probable than negative. While the neutral sentiment is normally dominant in whole-topic discussion, in aspect-based sentiment analysis, it is almost matched by other sentiment expressions. This shows that the Twitter user is inclined to be specific in his or her opinion when discussing the aspects of an article.

5.1. Implications

Research impact has used citation as the main indicator of research’s standing; however, it takes years to see any measurable impact. On the contrary, researchers are increasingly going online to find and share information about science, as well as, they have been urged to consider how they can use social media platforms to engage with each other. With the increased usage of the social media platforms for scholarly communications, altmetrics data are of enhanced interest as it captures real-time scholarly communication data from online platforms (e.g., Twitter) and may be used as an early measure of the research impact. Specifically, the papers cited in positive and neutral tweets have a greater impact than those not cited or cited in a negative tweet. However, there is still a need to investigate tweeter data to analyse user sentiments relating to research articles in specific fields. Such a fine-grained investigation is required to fully utilise the findings of existing studies. As mentioned earlier, this article presents a quantitative study to exploit tweet data to analyse user sentiments relating to different aspects of research articles in specific fields. This study helps us to understand the sentiments of online social exchanges of the scientific community on scientific literature, specifically, the sentiment of tweets, for better visibility and qualitative assessment of these interesting big data of altmetrics. We identify the sentiment of research communities with respect to their respective fields and to conduct an aspect-based analysis of user expressions related to their research articles. Such a fine-grained analysis may help research communities in improving their social media exchanges about the scientific articles to disseminate their scientific findings effectively and increase their impact.

5.2. Limitations and future works

While there is a significant increase in Twitter usage in order to share research articles, the expression of opinion is still dominated by neutral sentiment, and the trends suggest no increase in sentiment expression (see Appendix 2). In further work that is undertaken over a longer duration, the Twitter mentions of a research article could be explored. Also, as we created a ranking of altmetrics articles on the basis of their Twitter popularity, we could follow up to see whether the topmost articles indeed attract a higher citation count, in time. There are on average 5.98 tweets associated with an article in our dataset and the percentage of retweets is less than 5% of the total number of tweets. Moreover, using SentiStrength on tweets as described in section 4.1 revealed that this percentage of retweets is roughly equally distributed among three sentiment classes in terms of the absolute number of positive, negative and neutral tweets. Consequently, avoiding retweets might have a negligible effect on the findings of our investigation. In future, a dataset containing a significant number of retweets can be developed for the detection of social media campaign style in tweeter about scholarly articles, as many retweets are evidently simple retweets. Moreover, research can be carried out to establish the significance of retweets in terms of any correlation with citation. In addition, in terms of scoring documents, we believe that the influence of a Twitter user is significant; that is, the sentiment score of a tweet from a particularly relevant user should be
heavily weighted. While aspect-based sentiment analysis was unable to capture a wide range of data, the aspects can be derived intellectually to increase the significance of these results. Moreover, 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.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship and/or publication of this article.

**ORCID iDs**

Saeed-Ul Hassan https://orcid.org/0000-0002-6509-9190
Raheem Sarwar https://orcid.org/0000-0002-0640-807X
Raheel Nawaz https://orcid.org/0000-0001-9588-0052

**Notes**

1. https://twitter.com
2. https://www.facebook.com
3. http://www.citeulike.org
4. https://www.mendeley.com
5. These data were obtained in JSON file format from Altmetrics.com. Under the closed agreement, no version of these data can be publicly disseminated by the publisher. Nevertheless, the same data can be freely accessed for research purposes by the Altmetric.com.
6. https://pypi.org/project/beautifulsoup4/
7. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

**References**

[1] Bornmann L, Haunschild R and Adams J. 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). *J Informetr* 2019; 13(1): 325–340.
[2] Drongstrup D, Malik S and Hassan SU. Altmetric study of economics. In: *ISSI 2019 conference: 17th international conference of the International Society for Scientometrics and Informetrics*, Rome, 2–5 September 2019, pp. 984–989. Leuven: International Society for Scientometrics and Informetrics.
[3] Bornmann L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *J Informetr* 2014; 8(4): 895–903.
[4] Priem J and Costello KL. How and why scholars cite on Twitter. *P Am Soc Inform Sci* 2010; 47(1): 1–4.
[5] Priem J, Taraborelli D, Groth P et al. Altmetrics: a manifesto, 2010, http://altmetrics.org/manifesto
[6] Haustein S, Bowman TD, Holmberg K et al. Tweets as impact indicators: examining the implications of automated ‘bot’ accounts on Twitter. *J Assoc Inf Sci Tech* 2016; 67(1): 232–238.
[7] El Rahman SA, AlOtaibi FA and AlShehri WA Sentiment analysis of Twitter data. In: *2019 international conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 3–4 April 2019, pp. 1–4. New York: IEEE.
[8] Vairetti C, Martínez-Cámara E, Maldonado S et al. Enhancing the classification of social media opinions by optimizing the structural information. *Future Gener Comp Sy* 2020; 102: 838–846.
[9] Wouters P, Zahedi Z and Costas R. Social media metrics for new research evaluation. In: W Glänzel, HF Moed, U Schmoch et al. (eds) *Springer handbook of science and technology indicators*. Cham: Springer, 2019, pp. 687–713.
[10] Diamantini C, Miccoli A, Potena D et al. Social information discovery enhanced by sentiment analysis techniques. *Future Gener Comp Sy* 2019; 95: 816–828.
[11] Gallo FR, Simari GI, Martínez MV et al. Predicting user reactions to Twitter feed content based on personality type and social cues. *Future Gener Comp Sy* 2019; 110: 918–930.
[12] Haustein S, Peters I, Bar-Ilan J et al. Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics* 2014; 101(2): 1145–1163.
[13] Hammarfelt B. Using altmetrics for assessing research impact in the humanities. *Scientometrics* 2014; 101(2): 1419–1430.
[14] Priem J, Groth P and Taraborelli D. The altmetrics collection. *PLoS ONE* 2012; 7(11): e48753.
Appendix I

Table 6. Top 50 positive lexicon terms, with their positive and negative scores.

| Token               | Positive score (pos cdf hmean) | Negative score (neg cdf hmean) | Token               | Positive score (pos cdf hmean) | Negative score (neg cdf hmean) |
|---------------------|--------------------------------|--------------------------------|---------------------|--------------------------------|--------------------------------|
| Excellent           | 0.907660308                    | 0.25328778                     | Improves            | 0.890779932                    | 0.304454795                    |
| Novel               | 0.906440324                    | 0.2557995                      | Promising           | 0.889829053                    | 0.28539616                     |
| Amazing             | 0.905406722                    | 0.257757634                    | Improved            | 0.888990591                    | 0.304827953                    |
| Congratulations     | 0.905110313                    | 0.248694131                    | Worth               | 0.888006505                    | 0.302092593                    |
| Awesome             | 0.903971472                    | 0.26275244                     | Best                | 0.886228491                    | 0.32246558                     |
| Success             | 0.903082892                    | 0.260565937                    | Useful              | 0.885983593                    | 0.317457826                    |
| Wow                 | 0.903005407                    | 0.264781511                    | Successful          | 0.88539703                    | 0.272322083                    |
| Interested          | 0.902443229                    | 0.267758917                    | Super               | 0.884806715                    | 0.27129548                     |
| Exciting            | 0.902263317                    | 0.262783203                    | Excited             | 0.883726536                    | 0.254014497                    |
| Great               | 0.901346915                    | 0.259209104                    | Neat                | 0.88322853                    | 0.269576177                     |

(continued)
Appendix 2

Evaluation of Models: To conduct this study, we annotated a subset of the tweets in the original dataset, containing 2544 tweets in English about publications in various disciplines considered in this article. Specifically, 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 [34], which is a substantial agreement according to Landis and Koch [35]. Table 9 shows the percentage of tweets per label.

Table 6. (continued)

| Token | Positive score (pos cdf hmean) | Negative score (neg cdf hmean) | Token | Positive score (pos cdf hmean) | Negative score (neg cdf hmean) |
|-------|---------------------------------|---------------------------------|-------|---------------------------------|---------------------------------|
| Positive | 0.900758366 | 0.273718577 | Fitness | 0.839039899 | 0.309667926 |
| Nice | 0.900716826 | 0.288556025 | Plays | 0.838340923 | 0.271447038 |
| Thanks | 0.900603947 | 0.276785536 | Save | 0.837633457 | 0.343615901 |
| Special | 0.899726703 | 0.273907076 | Stuff | 0.836890133 | 0.395107942 |
| Cool | 0.899106111 | 0.288502052 | Kind | 0.835788067 | 0.310951915 |
| Interesting | 0.898675388 | 0.308576729 | Very | 0.834852929 | 0.441671495 |
| Greater | 0.896300338 | 0.285598025 | Fun | 0.833788337 | 0.309385544 |
| Hope | 0.894977105 | 0.289159208 | Helpful | 0.830742627 | 0.357925343 |
| Love | 0.893660856 | 0.280259277 | Colleagues | 0.827570638 | 0.407013242 |
| Pretty | 0.89339334 | 0.285442743 | Wonderful | 0.829151154 | 0.254216552 |
| Improving | 0.89281883 | 0.279777332 | Overview | 0.828697218 | 0.325720107 |
| Huge | 0.892148183 | 0.277777332 | Improvements | 0.827931768 | 0.300823749 |
| Interest | 0.892070257 | 0.288772767 | Winner | 0.826416998 | 0.262400833 |
| Fascinating | 0.891114122 | 0.30054873 | Article | 0.826234942 | 0.45688716 |

Table 7. Top 50 negative lexicon terms, with their positive and negative scores.

| Token | Negative score (neg cdf hmean) | Positive score (pos cdf hmean) | Token | Negative score (neg cdf hmean) | Positive score (pos cdf hmean) |
|-------|---------------------------------|---------------------------------|-------|---------------------------------|---------------------------------|
| Depression | 0.956523417 | 0.14296229 | Sad | 0.933212243 | 0.146280421 |
| Failure | 0.955666946 | 0.145001895 | Decrease | 0.931970753 | 0.20996222 |
| Chronic | 0.954612919 | 0.148733036 | Threat | 0.931410764 | 0.154818844 |
| Anxiety | 0.95435904 | 0.14518852 | Sorry | 0.930641447 | 0.17683056 |
| Loss | 0.95376771 | 0.151883033 | Problem | 0.930362917 | 0.217497094 |
| Worse | 0.953155776 | 0.152592792 | Complications | 0.929943039 | 0.16253538 |
| Fight | 0.951942612 | 0.149667313 | Difficult | 0.929870572 | 0.182247957 |
| Poor | 0.948370281 | 0.167972753 | Fail | 0.92825984 | 0.149107358 |
| Obesity | 0.94758065 | 0.169641955 | Challenge | 0.927242281 | 0.195327438 |
| Abuse | 0.945487713 | 0.143557614 | Harm | 0.924565382 | 0.214261433 |
| Critical | 0.945486351 | 0.175523536 | Cross | 0.922892138 | 0.191351394 |
| Decline | 0.94469103 | 0.173960254 | Violence | 0.922127121 | 0.144070395 |
| Risks | 0.943912604 | 0.179202624 | Bad | 0.921307242 | 0.244515638 |
| Low | 0.943589836 | 0.187245872 | Aggressive | 0.919158237 | 0.146828486 |
| Lack | 0.94311165 | 0.181049788 | Weak | 0.867450812 | 0.183937168 |
| Source | 0.94305159 | 0.181707388 | Harms | 0.866356927 | 0.155131484 |
| Risk | 0.94302831 | 0.193847787 | Depressed | 0.865741126 | 0.151298593 |
| Stress | 0.94272419 | 0.182659655 | Factor | 0.864080846 | 0.359182693 |
| Missed | 0.94139016 | 0.183218571 | Regardless | 0.863985589 | 0.174197141 |
| Problems | 0.940341809 | 0.190235998 | Complicated | 0.863504401 | 0.195533743 |
| Wrong | 0.93989407 | 0.186005692 | Inequality | 0.863401924 | 0.18313997 |
| Dependent | 0.93967264 | 0.175491201 | Beware | 0.862710192 | 0.169002881 |
| Obese | 0.939746991 | 0.185052127 | Controversial | 0.862340075 | 0.183446669 |
| Challenges | 0.93442513 | 0.203175391 | Fighting | 0.862208247 | 0.144446439 |
| Crisis | 0.934260093 | 0.147841295 | Waste | 0.861882475 | 0.20016067 |
Table 8. Evaluations of classification models.

| Models                                      | Precision | Recall | F1 Score | Accuracy |
|---------------------------------------------|-----------|--------|----------|----------|
| SentiStregth                                | 0.569     | 0.489  | 0.476    | 0.659    |
| SVM (-Stop words, stemming, TF-IDF)         | 0.593     | 0.496  | 0.501    | 0.663    |
| SVM (-Stop words, stemming, TF-IDF) + New Lexicon | 0.650     | 0.573  | 0.583    | 0.708    |
| Out Method (SentiStregth + New Lexicon)     | 0.680     | 0.623  | 0.633    | 0.738    |

SVM: support vector machines.

Table 9. Manually annotated tweets.

| Tweet labels | Percentage of tweets |
|--------------|----------------------|
| Positive     | 34.7%                |
| Negative     | 35.7%                |
| Neutral      | 29.6%                |

Table 8 shows the evaluation results. We found that our method achieved great accuracy in predicting tweet sentiment, with an average accuracy of 73.8%, compared to the SentiStregth with 65.9% accuracy. Our method also achieved high F1 and recall scores compared to the SentiStregth. In addition, we evaluated the performance of the SentiStregth 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 8. 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. 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, as we proposed, 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.