Check for updates

Research Paper

JIS

Journal of Information Science
1–15
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0165551520930917
journals.sagepub.com/home/jis
SAGE

Sentiment analysis of tweets through Altmetrics: A machine learning approach

Saeed-Ul Hassan
Information Technology University, Pakistan

Aneela Saleem
Information Technology University, Pakistan

Saira Hanif Soroya
Department of Information Management, University of the Punjab, Pakistan

Iqra Safder
Information Technology University, Pakistan

Sehrish Iqbal
Information Technology University, Pakistan

Saqib Jamil
Department of Management Sciences, University of Okara, Pakistan

Faisal Bukhari
Punjab University College for Information Technology (PUCIT), University of the Punjab, Pakistan

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

Raheel Nawaz
School of Computer Science, Manchester Metropolitan University, UK

Abstract

The purpose of the study is to (a) contribute to annotating an Altmetrics dataset across five disciplines, (b) undertake sentiment analysis using various machine learning and natural language processing–based algorithms, (c) identify the best-performing model and (d) provide a Python library for sentiment analysis of an Altmetrics dataset. First, the researchers gave a set of guidelines to two human annotators familiar with the task of related tweet annotation of scientific literature. They duly labelled the sentiments, achieving an inter-annotator agreement (IAA) of 0.80 (Cohen's Kappa). Then, the same experiments were run on two versions of the dataset: one with tweets in English and the other with tweets in 23 languages, including English. Using 6388 tweets about 300 papers indexed in Web of Science, the effectiveness of employed machine learning and natural language processing models was measured by comparing with well-known sentiment analysis models, that is, SentiStrength and Sentiment140, as the baseline. It was proved that Support Vector Machine with uni-gram outperformed all the other classifiers and baseline methods employed, with an accuracy of over 85%, followed by Logistic Regression at 83% accuracy and Naïve Bayes at 80%. The precision, recall and F1 scores for Support Vector Machine, Logistic Regression and Naïve Bayes were (0.89, 0.86, 0.86), (0.86, 0.83, 0.80) and (0.85, 0.81, 0.76), respectively.

Keywords

Altmetrics; comparative analysis; machine learning; sentiment analysis; Twitter

Corresponding author:
Saira Hanif Soroya, Department of Information Management, University of the Punjab, Lahore 54590, Pakistan.
Email: sairasroya@gmail.com

1. Introduction

With the emergence of Web 2.0, the use of social networking websites has increased in all spheres of life [1]. As social media receives so much attention, plenty of public and private opinion on various subjects is spread from this platform. Due to its many advantages, it likewise attracts the scholarly community to spread scholarly work. The academics use social networking, social bookmarking, social data sharing, video, blogging, microblogging, wikis, media, data sharing and social recommending sites [2–4]. Scholars are increasingly reading, sharing, discussing and rating research articles publicly.

There are many events and traces (i.e. uses, downloads, bookmarks, etc.) that contribute useful metrics to the study of research impact apart from traditional citation metrics [5–7]. The data and indicators on these social networking websites make up the Altmetrics universe [8]. With the rise in the use of the web for scholarly communication, since 2010 Altmetrics has emerged as a new research area since social media is a source of metrics to evaluate scholarly impact [9] and serves as an early indicator of the impact and usefulness of an article [10]. Haustein et al. [11] suggest that a large part of Altmetrics can be termed Social Media Metrics, referring to data and indicators about social media use, reception and impact [8,12,13].

Of these social media sites, the microblogging service known as Twitter has emerged as a tool for communicating, building social relations and sharing interests among users [14,15]. Users do not restrict themselves to tweeting about their public activities: they also tweet about research topics. The academic community shares scientific advancements via web-based networking media, discussing the research informally [7,16]. According to Lee et al. [17], keyword searches on Google Scholar, based on the hit-count rate in July 2016, are as follows: Twitter (6.16 million), followed by Facebook (5.27 million), YouTube (3.98 million) and LinkedIn (1.12 million). Twitter is at the centre of social affairs and gathering information. It has become a notable vehicle for the fast creation, transmission and discourse of news stories [18–20]. Entirely due to the help from Twitter, Web 2.0 has evolved into a much bigger platform for marketing campaigns [21,22], research-work discussion and sentiment analysis [23–26]. Twitter is used by many research communities as a channel to increase visibility and reach broader audiences that are interested in discussing scientific literature [27,28], and the most Altmetrics Social Media Metrics research has either focused on or included Twitter [4].

There are several Altmetrics aggregators, such as Altmetric.com, Lagotto, PLoS ALM, Plum Analytics and Impact Story, that aggregate metrics for scholarly materials from several sources [29]. Of these, Altmetric.com has the most comprehensive coverage of tweets on research articles and uses this information as an indicator of the articles' impact [30,31]. Sentiment analysis and opinion mining are undertaken to gauge public opinion of social media posts [32,33], and the interest in using such strategies on news and blogs is growing substantially [34,35]. Since scholars are using social media platforms to share and discuss their research and Twitter is the focus of Altmetrics research, sentiment analysis of tweets about research articles can yield valuable insights into public opinion and the early impact of scientific literature.

In sentiment analysis, to analyse Altmetric.com datasets, specific tools are used, such as SentiStrength and Sentiment140. However, current tools have some limitations. Friedrich et al. [36] report that Sentiment140 and SentiStrength are unsuitable for understanding tweets on scientific subjects and that Sentiment140 is unable to detect negative tweets, while SentiStrength overestimates the results. Friedrich et al. [36] performed sentiment analysis on the Altmetrics dataset using SentiStrength and suggested that adapting the scientific lexicon leads to better accuracy. The objective of this research was to apply several machine learning–based models on Altmetrics data to analyse which performs the best. We built a set of guidelines to annotating the data from five disciplines, and two human annotators undertook the task. The study's contributions are as follows:

- The first and foremost contribution of our work is to present a compressive annotation guideline for manual coding of tweets.
- Second, we test which machine learning procedure works better for assigning a particular sentiment to tweet.
- Finally, we developed a library that can be imported into Python and used for sentiment analysis of the Altmetrics dataset. The library can be accessed at the following URL: https://github.com/slab-itu/tweet_sentiments_survey

Current sentiment analysis tools are not well suited to an Altmetrics dataset, because of most tweets linked to scientific articles neither praise nor criticise. This study explores the extent to which Twitter data reflect opinions on the research discussed on social media by exploiting various approaches, including lexicon-based and machine learning–based models, taking SentiStrength and Sentiment140 as the baseline. Friedrich et al. [36] conclude that, while adapting the SentiStrength tool's lexicons to scholarly terms and removing the title of the article from the text of scholarly tweets increase its efficiency in detecting their sentiment, since such tools cannot distinguish accurately if the tweet contains positive or negative words from the tweeted paper's title itself.

To overcome these limitations, for the current research, along with these sentiment analysis tools, we first process the data by adopting the method proposed by Friedrich et al. [36]. Furthermore, we used machine learning algorithms to determine which model is appropriate and effective for Altmetrics dataset. We first trained our models in three classes. The first was about positive, negative and neutral tweets and, in the other two, only about positive and negative tweets to see how the exclusion of neutral tweets enhanced performance. Thelwall et al. [16] found that tweets about academic articles are objective, containing either the article title or a brief note/summary. Thus, they are unlikely to contain any criticism; therefore, we could assume that all neutral tweets are positive and thus trained our models for a two-class dataset, which yielded improved results.

The rest of this article is organised as follows: the next section presents studies on Twitter as a popular microblogging platform to disseminate research on social media. Then we describe the dataset used in this study and the methodology adopted for sentiment analysis. This is followed by a discussion of the results, along with the conclusions and future research directions.

2. Literature review

A good number of available studies define Altmetrics, and its terms, potential benefits and challenges, the types of sentiment towards a scientific study that a tweet can convey and how detecting such sentiment can predict the article's future citation count. This section provides an insight into the work previously undertaken on the Altmetrics dataset.

The term 'Altmetrics' was introduced in 2010, meaning Alternative Metrics for performance assessment evaluation: this is an alternative to the traditional methods of bibliometrics and scientometrics [37]. These indicators are largely about social media metrics [29]. Social media has become highly popular with scholars to discuss their research. Erdt et al. [9] conducted a comprehensive overview of the Altmetrics landscape and defined it according to the Altmetrics literature. Besides the potential benefits and challenges of using Altmetrics to measure social impact, they discussed the various data sources and aggregators. From the existing studies, they concluded that Mendeley and Twitter were the most widely used data sources for Altmetrics. The researchers found a weak correlation between Altmetrics and citation count, ranging from 0.07 to 0.5, which showed that the Altmetrics result differs from the citation count.

Sentiment analysis of tweets linked to scientific articles can measure the interest in or the impact of research that is discussed on social media. However, in writing for social web platforms, people usually use short forms of terms and emoticons. This heightens the need to develop tools that can detect sentiment in brief texts. Thelwall et al. [38] presented an algorithm, SentiStrength, that works in both supervised and unsupervised cases. The algorithm adopts a lexicon approach whereby terms are coded as a positive or a negative sentiment of strength on a scale of –5 to +5. It predicts the sentiment according to the occurrence of these terms. With social web data, for which no training dataset is available, SentiStrength detects sentiment better and is thus recommended for applications in which effective direct terms are used to undertake sentiment analysis.

Thelwall et al. [16] conducted a pilot study on 270 tweets about academic articles, collected in 2012. The purpose was to analyse how scholarly articles are tweeted. For the annotations, Thelwall et al. [16] built coding schemes for use by three experienced human annotators, resulting in a moderate pairwise agreement using Cohen's Kappa statistic. The outcome demonstrated that tweets about academic articles are objective, for the most part: the content is either a summary of the article for a new audience, the article's title or a self-citation to increase the tweet count to boost the article's popularity. The authors concluded that tweets probably give little insight into researchers' responses to articles because few praise them and none criticise them.

Twitter is used as a data source by many aggregators, yet it is unclear to what extent tweets convey the positive or negative opinion of those in the various disciplines who click the link to the article in question. Haustein et al. [39] undertook sentiment analysis on 487,610 tweets that mentioned 192,832 articles. They performed the analysis on an Altmetrics dataset using SentiStrength, adapting its lexicon to the dataset used and found that few positive or negative sentiments were detected, as concluded by previous studies [16,40]: most tweets are neutral, only 20% conveying any sentiment. The disciplinary analysis showed that tweets from the psychology, the social sciences and humanities contain the highest level of sentiment, and physics, chemistry and engineering the least.

Friedrich et al. [41] used a dataset of 663,547 tweets from Altmetric.com and identified their polarity by using SentiStrength and Sentiment140, then compared the results through percentage overlap and Cohen's Kappa. The researchers argued that sentiment analysis of tweets linked to scientific literature is challenging because tweets seldom comment or express any sentiment on scientific articles. The study further suggests that the sentiment analysis tools currently available have limitations and cannot identify the sentiment in tweets about academic articles.

Different proposed classification and pure natural language processing (NLP)–based methods have different behaviours in predicting sentiment orientation [42]. The authors suggested that pure NLP-based methods have very low

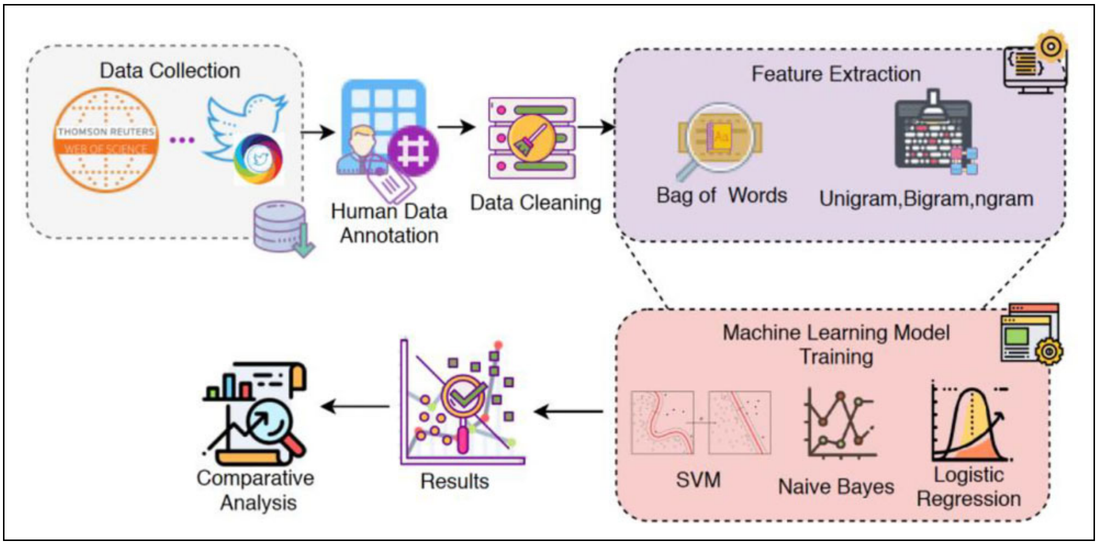

Figure 1. Overall approach of data input to analytics.

accuracy. However, when these methods are fused with machine learning–based methods (e.g. Support Vector Machine, tweets using Naïve Bayes and MaxEnt), they can achieve higher accuracy based on their distinct behaviours.

Another insight was presented by Robinson-Garcia et al. [43]. They conducted a sentiment analysis of 8206 tweets about 4358 dental articles between June 2011 and June 2016 from 2202 US-based accounts. They identified that bots were prevalent on Twitter, but that the frequency was only 2.5%. Interestingly, however, humans sometimes work as bots do; they duplicate tweets and behave mechanically. Much previous research has pointed to the scarcity of original content in tweets and the mostly neutral sentiments, advocating for an approach based on interactions [4,43–45]. In other words, ‘Twitter is less about what people tweet rather how they are connected’ [46].

Wakeling et al. [47] analysed 30,034 comments associated with 15,362 articles based on the data provided by PLOS. The researchers found that positive comments (praise) were found to be more prevalent than those, including some criticism, for almost all studied journals.

The review of the existing literature shows that existing state-of-the-art tools such as SentiStrength and Sentiment140 are unable to label scientific tweets correctly as positive or negative; tweets about scientific articles are mostly neutral or express little opinion; and that the strength of expressed sentiment varies between disciplines. In the current study, we regarded the existing state-of-the-art tools such as SentiStrength and Sentiment140 as the baseline against which to check the effectiveness of machine learning algorithms such as Naïve Bayes, Support Vector Machine and Logistic Regression on an Altmetrics dataset. We compared the models’ accuracy F1 scores. We used Receiver Operating Characteristic (ROC) curves to observe the behaviour of the classifiers.

3. Data and methodology

This section describes the dataset used in this study, the annotation guidelines for data preparation and the level of agreement between the two human annotators, who are familiar with the task of related tweet annotation of scientific literature, as evaluated by Cohen’s Kappa. It also explains the various techniques used to clean the text data to make it feasible to undertake machine learning, along with the feature extraction steps and overall methodology to perform sentiment analysis. Figure 1 shows the study’s overall data input approach to analytics.

3.1. Dataset

Several Altmetrics aggregators can collect and provide social media metrics: Altmetric.com, Lagotto, Plum Analytics and CrossRef Event Data. Of these, Altmetric.com has the best coverage and highest number of (re)tweets [29]. Therefore, for the current study, the Altmetric.com aggregator was used as the data source. Twitter data from Altmetric.com have been used in a previous study [48]. Our dataset was compiled in stages: first, we retrieved Web of Science (WoS) articles published in 2014; second, we searched each article for its DOI (Digital Object Identifier) through Altmetric.com and then, utilising the altmetrics.org database, the tweets obtained through links to each DOI. The DOIs were split into five

Table 1. Descriptive statistics of the Twitter dataset.

Measure	Value
Total number of tweets	6388
Total number of disciplines	5
Total number of articles	300
Total number of users	3099
Total number of followers of all users	15,046,672
Total number of languages	23

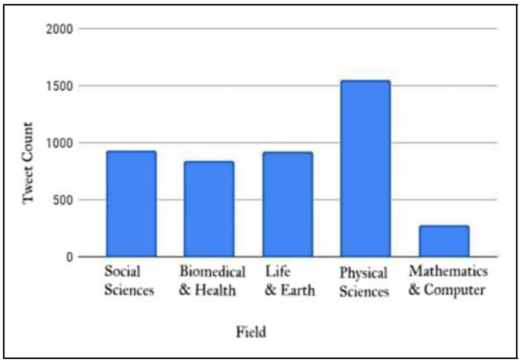

Figure 2. Tweet counts per field.

broad disciplines, based on the subject classification of journals, as developed at Thomson Reuters and later adapted by Centre for Science and Technology Studies (CWTS) at Leiden [49]. A sample of nearly 60 top-tweeted articles was extracted from each discipline, making 300 articles and 388 associated tweets in 23 languages (see Table 1).

The dataset contains 6388 tweets on publications in the disciplines of biomedical and health sciences, life and earth sciences, mathematics and computer science, physical sciences and engineering, and social sciences and humanities. Figure 2 gives an overview of the tweet counts in each discipline. As shown in Figure 3, 43.5% were labelled as neutral (class 0), 31.5% of tweets as negative (class –1) and the remaining 25% as positive (class 1). Of the 23 different languages found in our dataset, over 90% tweets were in the English language, 3% were in the Spanish language, 2% were in the Indonesian language, 1% tweets were in Portuguese and Japanese languages, whereas the following languages had less than 1% of the coverage: French, Russian, Arabic, Turkish, Italian, Dutch, Thai, Swedish, German, Korean, Malay, Filipino, Finnish, Danish, Polish, Chinese, Ukrainian and Greek. The tweet dataset was preprocessed to be consistent, then a survey of sentiment analysis was undertaken and comparative analysis carried out on existing techniques for opinion mining, such as machine-learning- and lexicon-based approaches, together with evaluation metrics. Later, our study took account of non-English tweets, and the results are to be reported.

3.2. Annotations

The annotation of the Altmetrics dataset was undertaken in three stages. In Stage 1, guidelines for manual annotation were set up (Table 2). Each tweet was compared with its related research article to detect its sentiment and whether the content was from the article's title, its abstract, its methodology or its conclusion. The manual annotation of tweets was to identify just three sentiments: neutral (0), negative (–1) and positive (1). See Table 2 for the annotation guidelines. The following protocols were observed:

- Tweets that contained only the article title were marked as neutral because they expressed no user sentiment (see Example 1) and lacked any evidence of engagement with the article [43].
- Tweets that contained self-citation, where authors cited their own article, were marked as neutral because they expressed no opinion on the article (see Example 2).
- Tweets in which users inquired about a topic were marked as neutral (see Example 3).
- Tweets that criticised an article were marked as negative, even if they used positive terms in sarcasm or expressed sentiments ironically.

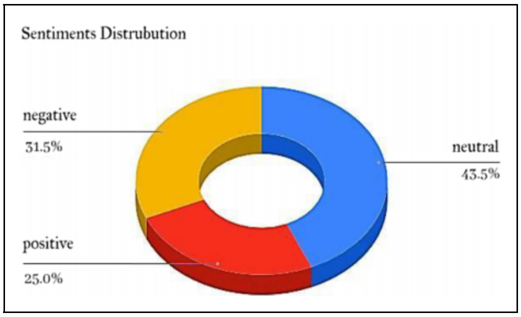

Figure 3. Distribution of tweet sentiment.

Table 2. Annotation guidelines and examples.

#	Property	Example
1	Paper title	*Paper title:* Kelp Gulls (*Larus dominicanus*) killed and injured by discarded monofilament lines at a marine recreational fishery in northern Patagonia *Tweet id:* 5258764 *Tweet content:* Kelp Gulls (*Larus dominicanus*) killed and injured by discarded monofilament lines at a marine recreational fishery in northern Patagonia
2	Self-citation	*Paper title:* The effect of phosphatidylserine administration on memory and symptoms of attention-deficit hyperactivity disorder: a randomised, double-blind, placebo-controlled clinical trial *Tweet id:* 5907295 *Tweet content:* Our latest clinical study has been published today in the *Journal of Human Nutrition and Dietetics*, the official . . . http://fb.me/FRNsFgJl
3	Questions	*Paper title:* Childhood Verbal Development and Drinking Behaviors from Adolescence to Young Adulthood: A Discordant Twin-Pair Analysis *Tweet id:* 8510316 *Tweet:* Thinking about how to give kids #abetterstart is not easy – better language development predicts earlier alcohol use?
4	Criticism	*Paper title:* Structural and electronic properties of chiral single-wall copper nanotubes (CuNTs) *Original tweet:* Oh dear. The difficulties of finding scientific abbreviations that work in all languages *Retweet id:* 5806620 *Retweet content:* @Danny_McMoomins You will love this from Prof Coxy
5	Highlights	*Paper title:* Cholesterol-lowering effects of oat b-glucan: a meta-analysis of randomized controlled trials *Tweet id:* 8671951 *Tweet content:* Oats or oat: food products reduces serum cholesterol in adults This tweet is taken from DISCUSSION part
6	Discussion	*Paper title:* A high dietary glycemic index increases total mortality in a Mediterranean population at high cardiovascular risk *Tweet id:* 2223908 *Tweet content:* Eating high GI food increase the risk dying early
7	Recommendations	*Paper title:* Why teach intelligence? *Tweet id:* 6564890 *Tweet content:* Mackintosh's nice contrib. 2 recent series of Intelligence articles on teaching intell. courses (Elsevier paywall)
8	Inappropriate terms	*Paper title:* Cognitive behavioral game design: a unified model for designing serious games *Tweet id:* 5907980 *Tweet content:* Social Cognitive Theory + Multiple Intelligence = Cognitive Behavioral Game Design aw shit yea This tweet is taken from ABSTRACT
9	Sarcastic salutations	*Paper title:* Structural and electronic properties of chiral single-wall copper nanotubes (CuNTs) *Tweet id:* 9220016 *Tweet content:* Oh dear. The difficulties of finding scientific abbreviations that work in all languages
10	Neutral terms	*Paper title:* The EBI RDF platform: linked open data for the life sciences *Tweet id:* 6240499 *Tweet content:* @LeNovereLab: new paper online. The EBI RDF platform http://t.co/QdsUvWhcUp?amp=1" #opendata #lod #bioinformatics

- The whole context of the tweet was considered (see Example 4). If the original tweet about the article contained criticism, then any retweet was also considered to contain criticism, based on this whole context.
- Tweets that referred to the abstract/methodology/discussion/conclusion of an article were marked as positive, as these sections are the highlights. We assumed that such parts provide a glimpse into the whole story and encourage people to read it, increasing the chances of the article receiving a further citation (see Example 5).
- Tweets that discussed an article to reveal a risk or alarming situation, even if using negative terms, were marked as positive (see Example 6). These tweets tend to highlight serious risks or messages attached to the article. Examples 5 and 6 are almost the same.
- Tweets where someone explicitly used positive adjectives about a study, that is, 'interesting', 'loved', 'nice' and so on, were marked as positive because they constituted a clear recommendation (see Example 7).
- Tweets that were taken from the abstract/methodology/discussion/conclusion, together with negative comments, were marked as negative.
- Tweets that contained sarcastic salutations, that is, 'Dear', were marked as negative on the basis of the context (see Example 9). Note that not all tweets with salutations were marked as negative; the annotation was strongly dependent on the context of the tweet.
- Tweets that contained opinions besides the article's title/summary could be positive, negative or neutral, depending on the text itself, and the hashtags had also to be considered (see Example 10, marked as positive).
- Furthermore, tweets containing the terms 'new', 'new article' and so on, were marked as neutral. Some negative terms, such as 'LOL', 'shit' and so on, were marked as negative because they were malign (see Example 8).

In Stage 2, using the above-defined guidelines, two coders manually annotated the tweets from the Altmetrics dataset. In Stage 3, both annotated all the tweets from scratch again to estimate the inter-annotator agreement (IAA) between them and to check the reliability of the dataset. Cohen's Kappa was used as the estimator for annotation.

3.2.1. IAA. Once the annotation task was completed, we performed a comparison to evaluate the quality of the annotation by calculating the IAA between individuals. Cohen's Kappa (for two raters) is one of the most commonly used statistics to calculate inter-rater agreement (IRR) for nominal variables [50]. The Kappa value calculated was 0.80, indicating a strong agreement between our annotators.

3.2.2. Annotations of non-English tweets. As described earlier, a total of 23 languages were seen in the Altmetrics dataset. Non-English tweets were first translated manually, using Google Translator, and then coded according to the annotation guidelines above. The sentiment analysis was undertaken on two types of Altmetrics datasets: one with only English tweets and the other with multiple languages. Finally, the results were compared.

3.3. Data cleaning

To perform text classification, we needed to extract useful data from the raw text. Twitter-specific affordances (i.e. URL, usernames, etc.) were removed. Of the total 6388 tweets, we were left with 4532 tweets after applying the following data cleaning on the data:

1. We removed the title of the article, as tweeters sometimes mention only this in their tweet yet, as a title expresses no opinion [50], it does not help to establish the sentiment of the tweet and could be misleading if the title should contain any positive or negative terms. So, we removed it to avoid false positives.
2. We removed special characters, including punctuation and symbols such as @, #, $, %, ^.
3. We removed the symbol # from the beginning of hashtags and kept the remaining characters as part of the normal text.
4. We removed all 'mentions' from the text because they are not useful in sentiment analysis.
5. We removed URLs from tweets because they are of no use to the sentiment analysis task.
6. When repeated characters were used as slang or just for fun, for example, 'happpyyyy', 'nooooooooo' and 'myyy', we adopted the usual spelling and used 'find and replace' to eliminate the specific repeated characters, as they make no sense.
7. Stop words were also removed.
8. Tokenisation, the process of splitting sentences into basic units such as words, was undertaken using Natural Language Toolkit (NLTK) Python tokenizer.[1] We then performed feature reduction, as mentioned above, such as stemming and stop-word removal using the NLTK Python library.[2]
9. We removed duplicates (retweets).
10. For feature reduction, we used stemming, specifically Porter stemmer.[3]

3.4. Feature extraction and classification models

Features were extracted from the processed dataset for later use in training the models for the classification task. Since we were undertaking text classification, we transformed our text data into bag-of-words (BOW) feature representations.

We experimented with the NLP model using n-gram with the value of $n = 1$ and $n = 2$ [51]. Friedrich et al. [52] performed sentiment classification on Internet Movie Database (IMDB) movie reviews and found that uni-gram outperform bi-gram on the short text. Our findings are in agreement because, in our case, uni-gram outperformed bi-gram when applied on short text, that is, tweets dataset.

After successfully cleaning and preprocessing the data, the next step was to analyse the dataset using machine learning models. Machine learning is a statistical approach to the prediction of output based on historical data, whereby algorithms make it possible for very simple document representations to train effective models [53]. As our problem is about classification, we could not use clustering, reinforcement learning or linear regression algorithms: we could use only classifiers. For this particular problem, sentiment analysis, our goal was to repeat the experiment using various algorithms and to report the precision, recall, F1 score and ROC, then choose the best ones, since it is a binary classification issue (i.e. positive or negative).

The best binary classifiers were found to be Support Vector Machine [21], Naïve Bayes [22] and Logistic Regression [54]. To measure the model performance on our dataset, we computed the accuracy, precision, recall, F1 score and area under the curve (AUC) scores. The following are brief details of the evaluation metrics employed.

3.4.1. Accuracy. Accuracy is a measure of the correctly predicted instances of the total number of instances. The following is the mathematical representation of accuracy (see equation (1))

$$Accuracy = \frac{Correct\ pred.\ positive\ samples\ +\ Correct\ pred.\ negative\ samples}{Total\ \#\ of\ samples} = \frac{TP + TN}{TP + TN + FN + FP} \tag{1}$$

where TP indicates the correctly predicted positive samples, and TN presents the correctly predicted negative samples. Similarly, FN and FP indicate, respectively, the misclassified negative instances and positive instances.

3.4.2. Precision. Precision is a measure of the positive tweets correctly predicted by the Artificial Intelligence (AI) model of the pool of total positive predicted tweets. The following is the mathematical fraction used for precision calculation (see equation (2))

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3.4.3. Recall. Recall is a measure of the positive tweets correctly classified by the AI model of the pool of total humanly annotated positive tweets. The following is the mathematical fraction used for Precision-Recall (PR) calculation (see equation (3))

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

3.4.4. F1 score. The weighted harmonic means of precision and recall. The following is the mathematical fraction used for F1-score calculation (see equation (4))

$$\text{F1 score} = 2 \times \frac{Precision\ \times\ recall}{Precision\ +\ recall} \tag{4}$$

3.4.5. ROC curve. The ROC measures the performance of the classification problem at various threshold settings. Mathematically, ROC is a probability curve, and the AUC measures the degree of separability that helps to find the model classification capability between tweet classes. The higher the AUC score, the better the tweet classification model is differentiating between positive, negative and neutral tweets. Furthermore, ROC curves were plotted with the true positive rate (TPR) and the false positive rate (FPR) on the x and y axes, respectively.

4. Results and discussion

This section describes the results of a series of experiments to improve performance. We divided our data into training and testing. We then scrutinised our results and reported which model worked best on the Altmetrics dataset. We ran the same

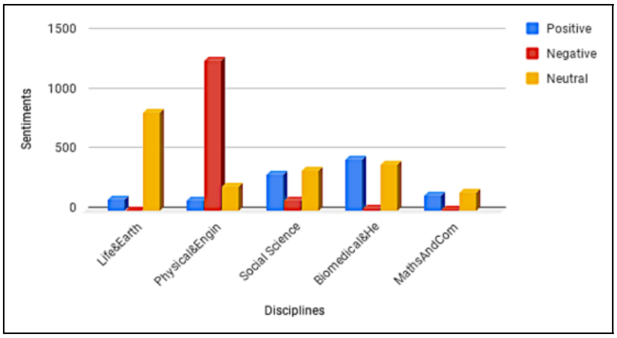

Figure 4. Cross-disciplinary distribution of sentiment types.

set of experiments on two versions of the dataset: one with tweets in English and the other with tweets in multiple languages. We used ROC with 10-fold cross-validation [55] to evaluate our models' performance. We used two language models of n-gram, that is, $n = 1$ and $n = 2$. Since most of the tweets were short, bi-gram did not perform well. Note that n-gram takes all the adjacent word combinations or n-length letters that we found in our source text [56]. For instance, a sentence 'this is a phrase' contains four uni-gram: 'this', 'is', 'a' and 'phrase'; three bi-gram: 'this is', 'is a' and 'a phrase'; and two 3-gram: 'this is a' and 'is a phrase'. Basically, an n-gram captures the syntax structure, such as what letter or word is likely to follow a given letter or word. The higher the value of n $(1, 2, 3, \ldots, n)$, the more context there is available to work. The optimum value of n depends heavily on the desired application and results.

4.1. Cross-disciplinary distribution of tweet sentiments

As seen in Figure 3, of 4532 tweets, 43.5% were coded as neutral, 25% as positive and 31.5% as negative. The cross-disciplinary distribution of the manually coded sentiments is shown in Figure 4. In all disciplines, most of the sentiments in tweets were neutral, and the percentage of positive sentiment was slightly higher than that of negative sentiment, confirming previous findings [10,57], apart from in physical sciences and engineering, in which negative tweets prevailed. The discipline of maths and computer science had a low total tweet count (290 tweets), yet it had the highest percentage of tweets with positive sentiments (45%); tweets from the disciplines of life and earth sciences and social sciences and humanities also expressed a high proportion of positive sentiment. The discipline of physical sciences and engineering had the highest total tweet count (1544 tweets) and also the highest proportion of negative sentiments (84.5%) of any discipline, which shows that its tweets about articles expressed more sentiment than those from other disciplines.

This interesting contrast among the fields leads to the additional research question, that is, Why tweets in specific fields of study are largely negative? Does it reflect the behaviour of the social media community that interacts with the scientific literature of the specific field? A recent work by Said et al. [58], on over 1.4 million tweets related to Altmetrics data, shows that Twitter users exhibit different social network structures across the fields. Thus, keeping in view the cross-field distribution of tweet sentiments presented in this study, this may also be implied that cross-field Twitter users exhibit not only different social network structures but also demonstrate different behaviour in giving their opinion on the scientific literature.

4.2. Experimental setup and parameter tuning

The emphasis of our experiment was to explore the extent to which Twitter data hold opinions about research that is discussed on social media by exploiting various approaches, including lexicon-based and machine-learning-based models. For this purpose, for the Altmetrics dataset, we used the F1 score and ROC to compare the existing state-of-the-art sentiment analysis tools and three classifiers based on Naïve Bayes, Support Vector Machine and Logistic Regression.

We used a grid-search approach, GridSearchCV from the scikit-learn library, to work out the best parameters for Support Vector Machine, Naïve Bayes and Logistic Regression. GridSearchCV is a meta-estimator that takes as its input an estimator whose parameters need to be optimised and a set of hyper-parameter settings to search all possible combinations (Cartesian product) to find the best combination of hyper-parameters [59]. In the Support Vector Machine, there were few parameters that could be fine-tuned apart from C, kernel and gamma (γ). Logistic Regression and Naïve Bayes gave the best results with the default parameters. Figure 5 shows the variations of C and gamma for kernel 'rbf'.

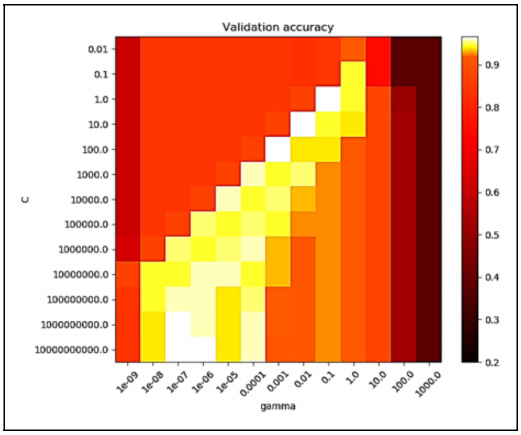

Figure 5. Variations of C and gamma for rbf kernel.

Table 3. Classification report for positive and negative for E-L and M-L tweets using NB, LR and SVM.

Precision		Recall		F1		Accuracy		
E-L	M-L	E-L	M-L	E-L	M-L	E-L	M-L	Classifier
0.85	0.84	0.81	0.79	0.76	0.73	0.8	0.79	NB
0.86	0.84	0.83	0.82	0.8	0.78	0.83	0.81	LR
0.89	0.89	0.86	0.87	0.86	0.87	0.85	0.86	SVM
0.78	0.8	0.72	0.7	0.75	0.75	0.65	0.64	SentiStrength
0.78	0.97	0.98	0.79	0.86	0.87	0.78	0.79	Sentiment140

E-L: English language; M-L: multi-language; NB: tweets using Naïve Bayes; LR: Logistic Regression; SVM: Support Vector Machine.

Support Vector Machine performed well for $C = 1000$ (a penalty parameter of error term) and gamma (γ) $= 0.001$ (a kernel coefficient). We set these parameters for kernel 'rbf' and saw satisfactory results. We also excluded neutral tweets from the dataset because these were not the focus of the research.

4.3. Comparative analysis of tweet classification

After selecting the combination of the best hyper-parameters and excluding neutral tweets, the overall classification results are in Table 3. For English-language tweets, the Support Vector Machine classifier performed better than the other models, with a precision of 89%, recall of 86%, F1 score of 86% and an accuracy of 85%. An observation about Sentiment140 is that its TPR was high compared with its true negative, false positive or false negative. Similarly, for multi-language tweets, Support Vector Machine outperformed the other models with an accuracy of 86%, the precision of 89%, recall of 87% and an F1 score of 87%. We show that this model performs better in terms of accuracy when 'all language' tweets are used. The machine learning models performed well compared with both the SentiStrength and Sentiment140 models.

4.4. Cross-disciplinary analysis of results

Table 4 shows the results of our best-performing classifier, namely, Support Vector Machine, in all five disciplines. Interestingly, we find that the classification results vary across the fields. We find that classification accuracy of physical sciences and engineering, along with mathematics and computer science, shows a high F1 score with overall classification accuracy, reaching up to 94% for physical sciences and engineering. In contrast with all the selected disciplines, the classification results of the tweets in the social sciences and humanities discipline remain below 80% across all the evaluation measures. This may reflect that the tweets related to social sciences and humanities is a challenging task and might require more training data to improve the classification accuracy. Nevertheless, we find that the disciplines, namely, biomedical and health science and life and earth sciences, show up to 90% predictive accuracy.

Table 4. Classification results of Support Vector Machine across scholarly disciplines.

	Precision	Recall	F1	Accuracy
Physical sciences and engineering	0.99	0.94	0.95	0.94
Social sciences and humanities	0.77	0.73	0.79	0.76
Biomedical and health sciences	0.90	0.93	0.88	0.91
Life and earth sciences	0.86	0.89	0.89	0.90
Mathematics and computer science	0.88	0.82	0.95	0.88

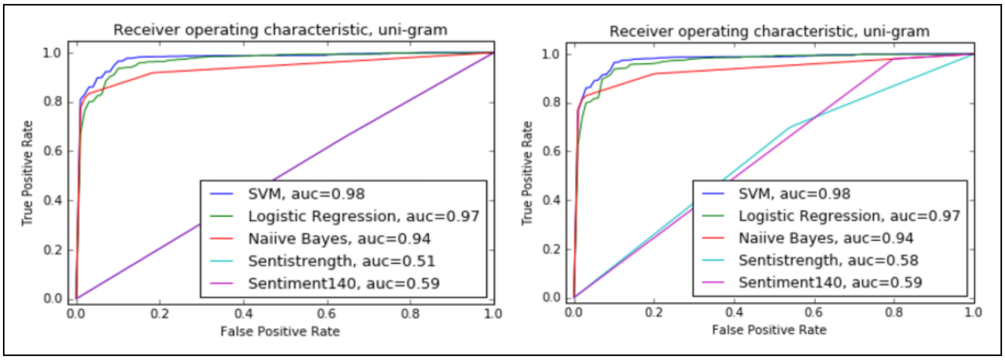

Figure 6. ROC of the classifiers on the Altmetrics dataset: English-language tweets (left) and multi-language tweets (right).

4.5. ROC curve analysis

ROC curves are recommended for use in classification problems when evaluating binary decision problems [60,61], yet they can present an overly optimistic view of an algorithm's performance if the class distribution is highly skewed. We evaluated our models through the AUC of ROC to compare the results of all classifiers using 10-fold cross-validation.

Figure 6 (left) shows the ROC of three classifiers, Support Vector Machine, Naïve Bayes and Logistic Regression, along with the ROC of our baseline models, SentiStrength and Sentiment140, for uni-gram. The AUC of our models is greater than that of the baseline models, SentiStrength and Sentiment140. The AUC is highest for Support Vector Machine, at 0.98, for the uni-gram. However, when tested with the bi-gram, AUC for Logistic Regression was 0.91. This agrees with our previous results in that our classifiers work better with the uni-gram than the bi-gram model. So, based on the AUC, the Support Vector Machine–based model outperforms the other models. Figure 6 (right) shows the curves drawn for the dataset using multiple languages for uni-gram. The AUC is greater for Support Vector Machine at 0.98. When tested, Logistic Regression has an AUC of 0.91 for the bi-gram. Our models perform very well compared with the baseline methods, so the results are somewhat similar to those for the English-language tweets in the previous section.

Overall, the baseline score in the ROC analysis obtained by SentiStrength and Sentiment140 was in the range of 0.58 to 0.59, whereas our models gave an AUC in the range of 0.94 to 0.98.

5. Concluding remarks

We have presented an annotated dataset for tweet sentiment classification using an Altmetrics dataset. Furthermore, we have tested the existing sentiment analysis tools and standard machine learning algorithms on an Altmetrics dataset. The tweets linked to WoS articles published in 2014 were annotated manually by two annotators, achieving IAA reliability of 0.80 using kappa's (κ), suggesting strong agreement, indicating that 6% to 81% data are reliable. As in the previous study by Friedrich et al. [36], using SentiStrength and Sentiment140, we considered these tools to be our baselines and compared machine learning algorithms such as Naïve Bayes, Support Vector Machine and Logistic Regression to see which performed best. We experimented with the NLP models of n-gram with the value of $n = 1$ and $n = 2$. We ran the same set of experiments on two versions of the dataset, one with tweets in English and the other with tweets in multiple languages. Support Vector Machine (uni-gram) outperformed the other models, with an accuracy of about 85%, although Logistic Regression was very close. For the Support Vector Machine–based model, the AUC of ROC was high for these experiments, at 0.98 and 0.99 for the Altmetrics English-only tweets and the Altmetrics multiple-language tweets, respectively. Support Vector Machine and Logistic Regression performed similarly and outperformed Naïve Bayes.

In the future, an interesting research direction could be sarcasm detection in Altmetrics tweets since many use positive words, although the context of the original tweet is negative, and vice versa. For instance, 'You will love this from Prof. Coxy' was coded as negative, although the positive words in the text increased the tweet's chance of being considered by the classifiers to be positive. So, further studies should focus on analysing the discourse of the original tweet text to improve the classifiers' automated sentiment analysis [62,63]. Our sentiment analysis algorithms did not take into consideration the use of emoticons, as these were removed while cleaning the data. In the future, we could analyse the effects of emoticons in tweets to find whether they play a role in establishing the polarity of sentiment by building meta-knowledge-based [64,65] and advanced deep-learning models [66].

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The author Saeed-Ul Hassan is grateful for the financial support received from the Higher Education Commission for this research for Crime Investigation and Prevention Lab (CIPL) affiliated with National Centre for Big Data and Cloud Computing (NCBC).

ORCID iD

Saira Hanif Soroya https://orcid.org/0000-0002-8153-1529
Raheel Nawaz https://orcid.org/0000-0001-9588-0052

Notes

1. See https://www.nltk.org/api/nltk.tokenize.html
2. See https://pythonspot.com/nltk-stop-words/
3. See https://www.nltk.org/howto/stem.html

References

[1] Imran M, Akhtar A, Said A, et al. Exploiting social networks of Twitter in Altmetrics big data. In: *Proceedings of the 23rd international conference on science and technology indicators (STI 2018)*, Leiden, 12–14 September 2018, pp. 1339–1344. Leiden: Centre for Science and Technology Studies (CWTS).

[2] Cheng LC and Lin MC. A hybrid recommender system for the mining of consumer preferences from their reviews. *J Inform Sci*. Epub ahead of print 10 July 2019. DOI: 10.1177/0165551519849510.

[3] Safder I and Hassan SU. Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications. *Scientometrics* 2019; 119(1): 257–277.

[4] Sugimoto CR, Work S, Larivière V, et al. Scholarly use of social media and Altmetrics: a review of the literature. *J Assoc Inform Sci Technol* 2017; 68(9): 2037–2062.

[5] Hassan SU, Imran M, Iftikhar T, et al. Deep stylometry and lexical & syntactic features based author attribution on PLoS digital repository. In: *Proceedings of the 19th international conference on Asian digital libraries*, Bangkok, Thailand, 13–15 November 2017; pp. 119–127. Cham, Switzerland: Springer.

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

[7] Priem J, Piwowar H and Hemminger B. Altmetrics in the wild: an exploratory study of impact metrics based on social media. In: *Proceedings of the Metrics 2011: Symposium on Informetric and Scientometric Research*. New Orleans, LA, 1 January 2011. San Francisco, CA: PLoS One.

[8] Wouters P, Zahedi Z and Costas R. Social media metrics for new research evaluation. In: Glänzel W, Moed HFSchmoch U, et al. (eds) *Springer handbook of science and technology indicators*. Cham, Switzerland: Springer, 2019, pp. 687–713.

[9] Erdt M, Nagarajan A, Sin SCJ, et al. Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics* 2016; 109(2): 1117–1166.

[10] Thelwall M, Tsou A, Weingart S, et al. Tweeting links to academic articles. *Cybermet Int J Scientomet Inform Bibliomet* 2013; 17: 1–8.

[11] Haustein S, Bowman TD and Costas R. Interpreting 'Altmetrics': viewing acts on social media through the lens of citation and social theories. In: Sugimoto CR (ed.) *Theories of informetrics and scholarly communication*. Berline: De Gruyter, 2016, pp. 372–405.

[12] Hassan SU, Safder I, Akram A, et al. A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics* 2018; 116(2): 973–996.

[13] Bonaccorsi A, Haddawy P, Cicero T, et al. The solitude of stars. An analysis of the distributed excellence model of European Universities. *J Informet* 2017; 11(2): 435–454.

[14] Bessagnet MN. A generic framework to perform comprehensive analysis of tweets. In: *Proceedings of the 7th Workshop on Bibliometric-enhanced information retrieval (BIR)*, Grenoble, 19 March 2018; pp. 80–85. Grenoble: CEUR-WS.org.

[15] Sarlan A, Nadam C and Basri S. Twitter sentiment analysis. In: *Proceedings of the 6th international conference, information technology and multimedia (ICIMU)*, Putrajaya, Malaysia, 18–20 November 2014, pp. 212–216. New York: IEEE.

[16] Thelwall M, Haustein S, Larivière V, et al. Do Altmetrics work? Twitter and ten other social web services. *PLoS One* 2013; 8(5): e64841.

[17] Lee MK, Yoon HY, Smith M, et al. Mapping a Twitter scholarly communication network: a case of the association of internet researchers' conference. *Scientometrics* 2017; 112(2): 767–797.

[18] Nawaz R, Thompson P and Ananiadou S. Identification of manner in bio-events. In: *Proceedings of the 8th international conference on language resources and evaluation LREC*, Istanbul, 21–27 May 2012, pp. 3505–3510. Paris: ELRA.

[19] Bruns A and Burgess J. Researching news discussion on Twitter: new methodologies. *Journal Stud* 2012; 13(5–6): 801–814.

[20] Shardlow M, Batista-Navarro R, Thompson P, et al. Identification of research hypotheses and new knowledge from scientific literature. *BMC Med Inform Decis Mak* 2018; 18(1): 46.

[21] Sampogna G, Henderson C, Thornicroft G, et al. Are social networks useful to challenge stigma attached to mental disorders? Findings from the time to change social marketing campaign 2009–2014. *Euro Psych* 2017; 41: S89.

[22] Houghton D, Hamdan ZA and Marder B. Structured abstract: political campaigning on Twitter – the use of language, message tone, and implications for political marketing communication from the UK general election 2015. In: Stieler M (ed.) *Creating marketing magic and innovative future marketing trends*. Cham, Switzerland: Springer, 2017, pp. 1413–1419.

[23] Jussila J, Vuori V, Okkonen J, et al. Reliability and perceived value of sentiment analysis for Twitter data. In: Kavoura A, Sakas DP and Tomaras P (eds) *Strategic innovative marketing*. Cham, Switzerland: Springer, 2017, pp. 43–48.

[24] Wang X, Rak R, Restificar A, et al. Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature. *BMC Bioinformat* 2011; 12(8): S11.

[25] Pak A and Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In: *Proceedings of the International Conference on Language Resources and Evaluation LREC*, Valletta, Malta, 17–23 May 2010, 1320–1326. https://www.researchgate.net/publication/220746311_Twitter_as_a_Corpus_for_Sentiment_Analysis_and_Opinion_Mining

[26] Bagheri A. Integrating word status for joint detection of sentiment and aspect in reviews. *J Inform Sci* 2019; 45(6): 736–755.

[27] 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.

[28] Letierce J, Passant A, Breslin JG, et al. Using Twitter during an academic conference: the #iswc2009 use-case. In: *Proceedings of the 4th international AAAI conference on weblogs and social media*, Washington, DC, 23–26 May 2010, https://www.researchgate.net/publication/221297840_Using_Twitter_During_an_Academic_Conference_The_iswc2009_Use-Case

[29] Wouters P, Zahedi Z and Costas R. Social media metrics for new research evaluation. In: Glänzel W, Moed HFSchmoch U, et al. (eds) *Springer Handbook of Science and Technology Indicators*. Cham, Switzerland: Springer, 2019, pp. 687–713.

[30] Bonaccorsi A, Cicero T, Haddawy P, et al. Explaining the transatlantic gap in research excellence. *Scientometrics* 2017; 110(1): 217–241.

[31] Zahedi Z and Costas R. General discussion of data quality challenges in social media metrics: extensive comparison of four major Altmetric data aggregators. *PLoS One* 2018; 13(5): e0197326.

[32] Safder I, Sarfraz J, Hassan SU, et al. Detecting target text related to algorithmic efficiency in scholarly big data using recurrent convolutional neural network model. In: *Proceedings of the 19th international conference on Asian digital libraries ICADL*, Bangkok, Thailand, 13–15 November 2017, pp. 30–40. Cham, Switzerland: Springer.

[33] Hassan SU, Imran M, Iftikhar T, et al. Deep stylometry and lexical & syntactic features based author attribution on PLoS digital repository. In: *Proceedings of the 19th international conference on Asian digital libraries ICADL*, Bangkok, Thailand, 13–15 November 2017, pp. 30–40. Cham, Switzerland: Springer.

[34] Pang B, Lee L and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on empirical methods in natural language processing*, Vol. 10, Philadelphia, July 2002, pp. 79–86. New York: ACM.

[35] Pang B and Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, Barcelona, Spain, July 2004, p. 271. https://www.cs.cornell.edu/home/llee/papers/cutsent.pdf

[36] Friedrich N, Bowman TD, Stock WG, et al. Adapting sentiment analysis for tweets linking to scientific papers. *Arxiv Preprint Arxiv*: 1507.01967, 2015.

[37] Priem J, Taraborelli D, Groth P, et al. *Altmetrics: a manifesto*, http://altmetrics.org/manifesto (accessed 26 October 2010).

[38] Thelwall M, Buckley K and Paltoglou G. Sentiment strength detection for the social web. *J Am Soci Inform Sci Technol* 2012; 63(1): 163–173.

[39] Haustein S, Costas R and Larivière V. Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *PLoS One* 2015; 10(3): e0120495.

[40] Friedrich N, Bowman TD and Haustein S. Do tweets to scientific articles contain positive or negative sentiments, http://altmetrics.org/altmetrics15/friedrich/ (accessed 9 October 2015).

[41] Friedrich N, Bowman TD and Haustein S. Do tweets to scientific articles contain positive or negative sentiments, http://altmetrics.org/altmetrics15/friedrich/ (accessed 9 October 2015).

[42] Emadi M and Rahgozar M. Twitter sentiment analysis using fuzzy integral classifier fusion. *J Inform Sci* 2020; 46(2): 226–242.

[43] Robinson-Garcia N, Costas R, Isett K, et al. The unbearable emptiness of tweeting – about journal articles. *PLoS One* 2017; 12(8): e0183551.

[44] Díaz-Faes AA, Bowman TD and Costas R. Towards a second generation of 'social media metrics': characterizing Twitter communities of attention around science. *PLoS One* 2019; 14(5): e0216408.

[45] Hellsten I and Leydesdorff L. Automated analysis of topic-actor networks on Twitter: new approach to the analysis of socio-semantic networks. *J Assoc Inf Sci Technol*. Epub ahead of print 18 March 2019. DOI: 10.1002/asi.24082.

[46] Haustein S. Scholarly Twitter metrics. In: Glänzel W, Moed HF, Schmoch U, et al. (eds) *Handbook of quantitative science and technology research*. Cham, Switzerland: Springer, 2019.

[47] Wakeling S, Willett P, Creaser C, et al. 'No comment'? A study of commenting on PLoS articles. *J Inform Sci*. Epub ahead of print 24 January 2019. DOI: 10.1177/0165551518819965.

[48] Didegah F, Mejlgaard N and Sørensen MP. Investigating the quality of interactions and public engagement around scientific papers on Twitter. *J Informet* 2018; 12(3): 960–971.

[49] Moed HF. CWTS crown indicator measures citation impact of a research group's publication oeuvre. Arxiv Preprint Arxiv: 1003.5884, 2010.

[50] Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol* 2012; 8(1): 23.

[51] Pang B, Lee L and Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on empirical methods in natural language processing*, Vol. 10, Philadelphia, July 2002, pp. 79–86. New York: ACM.

[52] Friedrich N, Bowman TD and Haustein S. Do tweets to scientific articles contain positive or negative sentiments. 2015. http://altmetrics.org/altmetrics15/friedrich

[53] Freitag D. Machine learning for information extraction in informal domains. *Mach Lear* 2000; 39(2–3): 169–202.

[54] Jussila J, Vuori V, Okkonen J, et al. Reliability and perceived value of sentiment analysis for Twitter data. In: Kavoura A, Sakas DP and Tomaras P (eds) *Strategic innovative marketing*. Cham, Switzerland: Springer, 2017, pp. 43–48.

[55] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int Joint Conf Artic Intel Aug* 1995; 14(2): 1137–1145.

[56] Liu B. *Web data mining: exploring hyperlinks, contents, and usage data*. 2nd ed. New York: Springer Science & Business Media, 2007.

[57] Holmberg K and Thelwall M. Disciplinary differences in Twitter scholarly communication. *Scientometrics* 2014; 101(2): 1027–1042.

[58] Said A, Bowman TD, Abbasi RA, et al. Mining network-level properties of Twitter Altmetrics data. *Scientometrics* 2019; 120(1): 217–235.

[59] Buitinck L, Louppe G, Blondel M, et al. API design for machine learning software: experiences from the scikit-learn project. *Arxiv Preprint Arxiv*: 1309.0238, 2013.

[60] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patt Recog* 1997; 30(7): 1145–1159.

[61] Provost F, Fawcett T and Kohavi R. The case against accuracy estimation for comparing induction algorithms. In: *Proceedings of the 15th international conference on machine learning*, 1998, pp. 445–453. https://www.researchgate.net/publication/2373067_The_Case_Against_Accuracy_Estimation_for_Comparing_Induction_Algorithms

[62] Ananiadou S, Thompson P and Nawaz R.Enhancing search: events and their discourse context. In: Gelbukh A (ed.) *Computational linguistics and intelligent text processing*. Berlin: Springer, 2013, pp. 318–334.

[63] Batista-Navarro RT, Kontonatsios G, Mihăilă C, et al. Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In: Gelbukh A (ed.) *Computational linguistics and intelligent text processing*. Berlin: Springer, 2013, pp. 318–334.

[64] Nawaz R, Thompson P and Ananiadou S. Negated bio-events: analysis and identification. *BMC Bioinform* 2013; 14(1): 14.

[65] Thompson P, Nawaz R, McNaught J, et al. Enriching news events with meta-knowledge information. *Lang Resour Eval* 2017; 51(2): 409–438.

[66] Jahangir M, Afzal H, Ahmed M, et al. An expert system for diabetes prediction using auto tuned multi-layer perceptron. In: *Proceedings of 2017 intelligent systems conference* (IntelliSys), London, 7–8 September 2017, pp. 722–728. New York: IEEE.