A Study of Altmetrics Using Sentiment Analysis

Luis Arredondo
Northern Illinois University
z1806788@students.niu.edu

Abstract

The increasing use of social media has found its way into the realm of professional and scholarly discussion. Whether it’s through negative or positive sentiment (or neither) people are discussing and bringing light to each other’s work in an informal manner through social media. This paper aims to analyze the social media data displayed in altmetrics and display any patterns or correlations between the sentiment data, the platforms, and the impact scores. In our findings we conclude that certain social media platforms hold more opinionated scholarly discussions and also that there is no correlation between social media sentiments and the social impact of a research publication.

Introduction

As time goes by, social media continues to gain traction as a professional communications outlet [1]. More and more discussions are taking place that can be used to gather the public’s opinion for sociological, commercial, and other purposes [2]. Of those now arising is the interest in analyzing the increasing social media attention on scholarly publications, as an alternative to the traditional peer review system that has been in place to this day. Thankfully, tools are being made to centralize and allow easy access and filters to these alternative metrics, or “altmetrics” for short, for further analysis [3].

Sentiment analysis, or opinion mining, aims to computationally determine if and how much opinion, sentiment, or subjectivity is present in a piece of text using natural language processing techniques [4]. Microblogging services such as Twitter, Facebook, etc. provide millions of opinion-rich messages every day about people’s thoughts and experiences [5]. Through altmetrics’ inherent focus on social media impact, this study focuses on analyzing the discussions taking place about scientific papers in social media and any pattern and correlations that can be identified with their sentiment data.

Data and Methods

The altmetric data used is a slightly modified data set that contains many pieces of information for each published paper in JSON format. From it, information such as the altmetric score, title, author(s), and individual Twitter posts can be obtained. In our case, the most important piece of data, the actual text in each social media post, was included as well. So there was no need for API fetches for Twitter, Facebook, etc.

The preprocessing done to the text in each individual posts included ignoring any reposts or retweets in the data set, as well as ignoring non-English posts (langdetect) and removing the research title from the text in the post. These steps were taken to ensure better sentiment analysis results [6]. Papers with altmetric scores of less than 10 were also taken out of the equation to make sure that enough relevant data is passed in for each paper. This resulted in a final count of around 4,250 papers (JSON files) with over 58,000 social media posts in total being analyzed.

NLTK for Python was used to apply sentiment analysis and store the results [7]. From it, SentimentIntensityAnalyzer passes in a piece of text to be analyzed and returns a compound score
were negative. 16.16% posts 58,529 analyzed 57.56% neutral, were the and 26.28% social studies, were ultimately set as the bounds for the three categories.

Related Work

With the proliferation of publications, researchers are utilizing academic social networks and reference management systems to find, store, and manage references [8]. Social and online reference management systems enable users to bookmark references to research content and to tag, review, and rate research content within their profiles. Scholarly tools such as these play an essential role in the organization of personal article collections and the generation of bibliographies across the research landscape today. Scholarly communities are sharing these digital reference libraries, and this open sharing encourages the formation of new research groups. Such online personal collections or repositories also accurately reflect researchers’ current and past reading, and indicate changes in their interests over time, making these datasets prime targets for recommendation analytics.

Several researchers have studied the relationship between citation-based and social-based metrics. Overall, research in this area has yielded weak to moderate correlations. For example, Haustein et al. [9] found that 9.4% of PubMed articles were mentioned in tweets, but there was only a low correlation between citations and tweets. Shuai et al. [10] reported a positive weak to moderate correlation between citations and Wikipedia mentions. Waltman and Costas [11] found a weak correlation between citations and F1000 recommendations. Bar-Ilan et al. [12] reported a moderate significant correlation between Scopus citations and Mendeley readership. Brody et al. [13] found a significant correlation between the citations and downloads of articles in physics, mathematics, astrophysics, and condensed matter. They used downloaded data from within six months after publication as a predictive feature. Costas et al. [14] reported weak correlations between citations and altmetrics, and they also found disciplinary differences in regard to the use of altmetrics.

Previous studies also explored the scholarly coverage of various social media platforms and reference managers [15]. Haustein et al. [16] studied the coverage of reference managers using data from the Web of Science (WoS) and Scopus. The DOIs of the articles were used to determine whether the articles were included in Mendeley and CiteULike. With 82% coverage, Mendeley significantly outperformed CiteULike for which the coverage was only 28%. Mendeley did lack in number of older articles but this is due to it being a newer platform. Similar results were reported by Zahedi et al. [17], who gathered publications from Impact Story and found that 62.6% of the articles had at least one bookmark in Mendeley. Twitter, Wikipedia, and Delicious each had less than 2% coverage of the publications. Hammarfelt [18] studied the publications of 30 Swedish universities and found that Mendeley had the highest coverage of articles and that 61.6% had a Google Scholar citation. Overall, according to these studies, Mendeley is the online reference manager with the most coverage of research articles.

Results

Of the 58,529 social media posts analyzed 26.28% were positive, 57.56% were neutral, and 16.16% were negative.
Figure 1: Histogram of sentiment scores

![Histogram of ALL Sentiment Scores](image)

Figure 2: Comparison of percentage of sentiment by platform

![Comparison of Platform Sentiments](image)

Table 1: Percentage numbers for each platform

<table>
<thead>
<tr>
<th></th>
<th>Positive(%)</th>
<th>Neutral(%)</th>
<th>Negative(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>19.29</td>
<td>64.77</td>
<td>15.95</td>
</tr>
<tr>
<td>Facebook</td>
<td>35.46</td>
<td>47.94</td>
<td>16.60</td>
</tr>
<tr>
<td>Google+</td>
<td>36.20</td>
<td>45.61</td>
<td>18.19</td>
</tr>
<tr>
<td>Blogs</td>
<td>37.60</td>
<td>46.13</td>
<td>16.27</td>
</tr>
</tbody>
</table>

Table 2: Actual numbers for each platform

<table>
<thead>
<tr>
<th></th>
<th>Positive(%)</th>
<th>Neutral(%)</th>
<th>Negative(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>6823</td>
<td>22911</td>
<td>5641</td>
</tr>
<tr>
<td>Facebook</td>
<td>2061</td>
<td>2787</td>
<td>965</td>
</tr>
<tr>
<td>Google+</td>
<td>577</td>
<td>727</td>
<td>290</td>
</tr>
<tr>
<td>Blogs</td>
<td>5919</td>
<td>7262</td>
<td>2561</td>
</tr>
</tbody>
</table>
What is interesting to note, however, is the difference in proportions of each platform’s sentiment. Twitter holds the least opinionated comments about a research publication, while blogs hold the most positive sentiment of the four platforms analyzed. It is not by much, but it holds true to each platforms typical intended use (Twitter a microblogging site, while blogging sites more typical for opinions and information). However, since the data set held contains more tweets than anything else, the totals throughout the paper get weighted more by twitter results (see Figure 1).

![Figure 3: Histogram of Twitter sentiment scores](image)

![Figure 4: Histogram of blog site sentiment scores](image)

When plotting the (average) sentiment score for each paper vs. the altmetric score it received, no discernable patterns appear to be visible. It is further confirmed when calculating the correlation between the two that no further similarities between the sentiment and social impact (altmetric score) exist.
Table 3: Correlation between sentiment score for each platform vs. altmetric score

<table>
<thead>
<tr>
<th>corr(x, y)</th>
<th>Altmetric Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>-0.01855</td>
</tr>
<tr>
<td>Facebook</td>
<td>-0.0082</td>
</tr>
<tr>
<td>Google+</td>
<td>-0.0166</td>
</tr>
<tr>
<td>Blogs</td>
<td>-0.00492</td>
</tr>
<tr>
<td>Overall Paper</td>
<td>-0.0355</td>
</tr>
</tbody>
</table>

The highest value of correlation is at -0.0355 for the overall sentiment of the paper vs. its altmetric score, but it is basically zero correlation.

Conclusions and Future Work

From the results gathered, the majority of posts on social media discussing research articles carry no opinion or subjectivity at all. However, by breaking up posts by platform, differences in overall sentiment can clearly be seen. Twitter is by far the least opinionated of the four platforms analyzed and the other three (Facebook, Google+, and blogging sites) hold more positive discussions. However, no correlation between a paper’s impact score and its social media sentiment exists. Subjective thoughts made by social media when discussing a paper hold no grounds to the paper’s success in outreach.

Future work could work on improving the range for sentiment scores to show more accurate results, as it might not have been conservative enough.

References


