

Using Machine Learning to Explore Social Media Engagement with Medical Publications

3

Authors: Adamos Spanashis, Nicola Lazzarini, Avgoustinos Filippoupolitis, Helen Stewart, Simon Francis, IQVIA, London, UK.

KEY TAKEAWAY

Machine learning models can predict, and provide insight into the drivers of, social media engagement with scientific publications. Journals with higher impact factors tend to receive more engagement on social media.

INTRODUCTION

IMPORTANCE OF SOCIAL MEDIA

Social media channels like Twitter (now X) are instrumental for Medical Affairs teams to communicate scientific research findings.¹

CHALLENGES

The effective use of social media as part of an evidence dissemination strategy is challenging due to the complex, fast-paced, and dynamic nature of user engagement on these platforms. This is further compounded by the need to communicate to diverse audiences and navigate complex regulations.

SOLUTION

We created an Impact Score to measure scientific publication engagement on social media. We then developed a state-of-the-art machine learning model to predict engagement and identify key factors driving interactions.

DATA SOURCES

Scientific evidence:

clinical study attributes extracted from PubMed abstracts and meta data using the IQVIA Linguamatics NLP engine.

Author information:

gathered from the IQVIA Onekey database, containing various metrics on healthcare providers, including their activity and influence within social media and the scientific community.

Publication journal information:

extracted from Scimago's journal ranking (SJR).

Twitter engagement:

collected using Twitter API and CrossRef events API.

METHODOLOGY

STEP 1

We defined an Impact Score to quantify and aggregate the engagement of a scientific study on Twitter:

$Impact\ Score = a * \sqrt{tweets} + b * \sqrt{likes} + c * \log(followers)$

We considered both the number of tweets and likes the study receives in the 12 months after publication. We also considered the number of followers of the user tweeting about the study.

STEP 2

We calculated the score for 7409 oncology papers published between 2017 and 2021. We used multiple data sources to characterise each publication and split it into training (80%) and test (20%) sets.

STEP 3

We used discussions with subject matter expert and data profiling to group the score into three categories of social media impact:

Of all papers: 2.7% 24.5% 72.8%

No Impact

Low Impact

High Impact

STEP 4

We developed a machine learning model² using a state-of-the-art algorithm, to predict the impact of each publication.

STEP 5

We used SHAP (SHapley Additive exPlanations)³ to interpret the model predictions and identify key driving factors for publication impact.

Study characteristics

- Use of randomisation in the study
- Type of study (e.g. clinical trial, observational)
- Type of arm (e.g. placebo, active comparator)
- Study design (e.g. blind/control)
- Support of real-world evidence
- Clinical outcome question (e.g. safety)

Source: Pubmed, Linguamatics (by IQVIA)

Scientific evidence

- Hazard ratio value for endpoint (e.g. overall survival)
- Hazard ratio p-value for endpoint
- Survival rate at X years (e.g. at 10 years)

Source: Pubmed, Linguamatics (by IQVIA)

Communication strategy

- Journal impact factor (SJR)
- Number of Twitter followers for the journal
- Author digital profile
- Author scientific profile
- Conference presence

Source: OneKey and Pharmaspectra (by IQVIA), SJR, Twitter

RESULTS

- Scimago's journal impact factor SJR is strongly associated with the Impact Score, while the publication journal being open access does not correlate much with the score.

Figure 2. A scatter plot of the papers' Impact Score against SJR (left) and the density of the Impact Score for open and non-open access journals (right)

- The machine learning model shows best test performance when predicting high-impact papers (15x better than baseline).
- It is harder to distinguish between low and no impact, as indicated by a small reduction in performance (2x and 1.2x better than baseline, respectively).

Figure 3. Confusion matrix comparing real and predicted labels (left). Overview of predictive performance (right). Precision indicates the % of predicted papers that are impactful (per class). Recall measures the % of impactful papers (per class) correctly predicted. F-score is the harmonic mean of the two

- The most influential predictor is the SJR, followed by the number of followers of the publishing journals' Twitter accounts. Author features (i.e. number of publications) do not contribute much to predicting the impact of the paper.

Figure 4. Top predictive features (using SHAP values). The length of the bar indicates the importance of each feature

CONCLUSIONS

Our model shows that machine learning can support the validation of publication decisions and guide journal selection.

SJR is the key predictor associated with higher engagement on Twitter, followed by the journal's follower count. In our model, author features, like publication count, have lower predictive value.

The proposed approach can be extended to other social media platforms (e.g. ResearchGate) to understand different types of impact.

Bibliography: 1. Fang, Z. et al. 2020. Scientometrics. 2020;124(3):2519–49. 2. Ke, G. et al. Advances in Neural Information Processing Systems 30 (NIPS 2017). 3. Lundberg, S.M. et al. Advances in Neural Information Processing Systems 30 (NIPS 2017).