Predicting the Popularity of Newly Published Research on Social Media

Daniel Shellstrom
Computer Science Department
Northern Illinois University
z1661193@students.niu.edu

David Gustafson
Computer Science Department
Northern Illinois University
z044140@students.niu.edu

Venkata Devesh Reddy Seethi
Computer Science Department
Northern Illinois University
z1839739@students.niu.edu

***Abstract*—The spread of published academic articles on social media can vary depending on many factors. This research focuses on determining whether a newly published work will be popular within social media by classifying information that is present before an academic article is published. We extract existing features from a large data set consisting of over 200,000 academic research articles and develop newly created features with a scoring system that allows us to train better models. This research uses a decision based models to predict whether a newly published article will be popular based on it's altmetric score.**

I. Introduction

The nascent of Web 2.0 brought user-generated content and with it, a shift in how society functions. A groundbreaking phenomenon, users are now able to create their own content instead of being bystanders in the digital age. Social media has become ubiquitous with everyday life allowing users to interact, collaborate, and share ideas in real time. It has changed the way we do business, understand politics, and formulate research. It has been shown that social media can reshape opinions and exert influence over a range of topics. Communication between users in real-time offer a new and unique manner on how thoughts and ideas are formed. Evidence suggests that social media is impacting how we make decisions and thereby influencing our choices and opinions [14].

Scholarly articles serve as a channel to present research and share knowledge as well as to discuss, criticize, and formulate ideas. The growth of social media in higher education, has created an alternative forum for the discussion of research articles. Through social media, content can reach an audience larger than conventional means [18]. Social media also allows for the immediate assessment of research after publication [21]. Altmetrics, an alternative form of metrics is used to assess the social impact of research, focusing on measuring social media platforms. It has been shown that altmetrics offer several benefits over traditional bibliometrics [24]. Altmetrics can measure the impact of an article outside of academia, as well as measure other forms of scholarly production than just papers. It can measure the impact upon social media in real time and is freely available. Using altmetrics to measure social media is drawing more support and is being used to determine the usefulness of an article.

We have identified that social media has a substantial grasp on modern society and that altmetrics have unique qualities to measure the perception of scholarly articles in social media. While altmetrics usage and research is becoming more popular, questions remain about possible functions and scenarios involving the impact altmetrics [13]. This paper contributes to this discussion by attempting to measure the scholarly impact of an article at publication. Predicting the popularity of content is often difficult for many reasons. External factors such as cultural and geopolitical forces often influence users. However, we attempt to generalize this prediction by foregoing external factors and focusing on the altmetric score and citation characteristics of the article. The altmetric score is a weighted ranking indicating the amount of recognition an article has received across social media. This paper attempts to build a model that can predict, whether a scholarly article will be popular (have a high altmetric score), at the moment of publication. Understanding the societal impact, can allow researchers to identify areas of interest that are prevalent today and focus on subjects that can make an immediate impact.

II. Related Work

Initial Interest cannot be interpreted as a determining factor for sustainable popularity [17] By using the altmetric score, we can predict the non-scholarly impact of an article [22]. Using Natural Language Processing, we can categorize peoples views as Altmetric scores have strong relation to traditional metrics [5].Academic impact can be positively correlated to the author and article availability on Wikipedia [16].Categorizing based on peoples views and diversity is a valuable metric [19]. Multiple social media platforms through time, determine associations with popularity [21]. In some areas, altmetric scores cannot be used as a determining factor. Varying attributes like Journal citation scores in blogs and altmetric scores gives better predictions [2].

Measuring article popularity can be based on scholarly and demographic impacts [23]. However, a different metric scale should be used. Some of the most popular articles can be heavily read and saved by scholars, but rarely cited [15].This increases ambiguity and makes it harder to measure impact [11]. Twitter data is biased on top journals , this would skew the popularity [8].

Prior research shows that reader counts and citation counts over a period time is a useful indicator to determine impact [20]. Initial user counts indicate early momentum [11] and

popularity can be based on a general rule that older articles are still popular due to better research [7]. Different platforms like Twitter and Mendeley determine popularity in different social, scientific groups [6]. Data can be consumed quite quickly through news aggregators. While this system offers prompt recognition, the interest gained can fade just as rapidly. Sites that host videos where views and likes are accrued, will take more time for the user to consume data and for the content to gain popularity.[18].

Eysenbach G. [3], performed categorical classification and a linear multivariate regression model to predict the popularity of scholarly articles through Twitter. Alternatively, Peoples B. K. [12] tested the amount of citations an article received through the use of a generalized linear mixed regression model using Twitter activity, the journal impact factor, and time. Maclaughlin A. [9] focused on ranking scholarly articles with the amount of news coverage. Finch T. [4] used the AAS (Altmetric Attention Score) across four different platforms Twitter, Facebook, news sites, and blogs using negative binomial generalized linear models. Ahmed M. [1] goes a step further and measures the content of social media over a predetermined time to predict popularity rise, with hits over time, clustering those points, and classifying them. The results are measured using linear regression against a K-Means baseline for 3 different data sets.

III. Data Set Preparation

A. Original Data

The data set used is from https://www.altmetric.com. The original data set comprised of over 100 gigabits, and was randomly sampled to obtain approximately 1.3 million published research articles. Among the random sample, articles were chosen that contained non-null values for the selected features. By cleaning the articles that contained null values, the data was reduced to around 180,000 articles (train/test set).

B. List Set

The new features that were created involved using existing data to develop measures for the predictions. To avoid data leakage and bias, a separate data set was sampled and cleaned in order to build the model. This set contained approximately 180,000 articles (list set) each containing values for the selected features.

C. Selecting Altmetric Score

The ability to predict whether a published research article will be popular in social media is determined by the ability to use the altmtric score's average. To establish a baseline average, the mean was taken from the original 1.3 million articles was $\mu = 3.88$. To remove outliers and sparse data from the training set, a distribution was taken and showed that close to 90% of the data's altmetric scores fell between the ranges of 0 and 10. Because of large and sparse varying ranges of scores, the train set was reduced to only contain published research articles with altmetric scores ranging from 0 to 10. The average altmetric score for this range was $\mu = 3.0$, which implies that the sample is close to representative of the original data. In order to balance the train set, oversampling was preformed to contain an even amount of articles above and below the mean.

IV. Feature Extraction

When examining the altmetric dataset, a limited number of attributes qualified as features in the target function while the rest were deemed irrelevant. We identified 2 types of attributes that would exist at the time of publication. The first type consisted of the categorical attributes author, journal, publisher, and subject. The second type consisted of attributes made up of text, including title and abstract.

A. List Features

- Author Relevant
  - The first author of an article was selected.
  - Author Productivity: How productive an author is. How many articles an author has published.
  - Author Prestige: The total altmetric score of all articles the author has contributed to, divided by the number of articles.
- Journal Relevant
  - Journal Distribution: The number of articles that appear in any given journal.
  - Journal Prestige: The total altmetric score of all articles that appeared in a journal, divided by the number of articles.
- Publisher Relevant
  - Publisher Distribution: The number of articles that have been published by any given publisher.
  - Publisher Prestige: The total altmetric score of all articles published by a publisher, divided by the number of articles.
- Subject Relevant
  - The first subject of an article was selected.
  - Subject Distribution: The number of articles that lists any given subject
  - Subject Prestige: The total altmetric score of all articles that lists a given subject, divided by the number of articles.

B. Text Features

Scopus Subjects: The frequency in which a scopus subject repeat in a corpus of research papers show a significant a trend. And the textual data can be used for the behavioural analysis. However, in this dataset 26,000 papers plotting the mean occurences of the subjects showed an inclination towards health and biology subjects. This has reduced the diversity and the behavioural analysis couldnt be performed.

Named Entity Recognition: A readers engagement depends on how closely the title and abstracts are related. High similarity equates to lesser deviation from problem statement, A low similarity may reflect a deviation from the topic or rigorous development on the problem statement. We used Named Entity

Recognition in our analysis on Title and Abstract, to split them into categories based on their grammar. This experiment gave interesting results as a negative correlation between the title and abstract was observed. Due to a sparse dataset this feature was not included in our analysis. In future, with a bigger dataset with full research paper we can use sentimental analysis in textual data to predict the popularity of the paper.

V. Classification Models

Three classification models were used to make predictions on new articles and label them as either popular or unpopular. If the altmetric score was less than 3, then it was considered unpopular with a label of "no", otherwise it was considered popular and was labeled as "yes". A mean value of $\mu = 3.0$ was used as the determining factor of popularity. The classification models used for the predictions were decision trees, naive Bayes, and random forest. The decision trees were built with a max depth of 3 to avoid over-fitting and the leafs had to contain at least 5 samples to be considered a node. A test set of 30% was taken from the original data to test the models. Since the author information was sparse, when the null values were removed from author features there were around 15,000 articles left to train and test. The idea to fill the null values with the mean score in order to preserve data was plausible, but would introduce higher scores for authors which would effect the prediction. In order to build the models with a larger amount of data, the decision tree classifier was ran two times, once with author features included and the other without author features. The naive Bayes and random forest models excluded author features when trained.

VI. Results

A. Decision Tree

Testing the classifier on all of the features except authors gave a accuracy score of 63.73% The root node split used the feature 'Journal Prestige' which had a gini index value of 0.5%. Figure 1 describes the accuracy measures.

	precision	recall	f1-score	support
no	0.67	0.54	0.60	26958
yes	0.62	0.74	0.67	27152

Fig. 1. Results Excluding Author Features - Decision Tree

A second test was ran on all of the features. This test used around 15,000 articles since a large majority of author feature values were null. The accuracy of the second test was 62.55% The root node split also used the feature 'Journal Prestige' which had a gini index value of 0.496%. Figure 2 describes the accuracy measures.

B. Naive Bayes

The naive Bayes classifier presented similar results compared to the decision tree, with the total performance reaching 56.79%. Out of a total of 54110 points from the test set, the classifier mislabeled close to half at 23383. The label of 0 represents unpopular, while the label of 1 represents popular. Figure 3 describes the results.

	precision	recall	f1-score	support
no	0.65	0.64	0.64	2498
yes	0.60	0.61	0.61	2215

Fig. 2. Results Including Author Features - Decision Tree

	precision	recall	f1-score	support
0	0.50	0.63	0.56	27029
1	0.50	0.38	0.43	27081

Fig. 3. Results Excluding Author Features - Naive Bayes

C. Random Forest

The random forest classifier presented the best prediction results. The accuracy was 71.22%. A few random forest classifiers were ran with a different number of estimators to choose the best preforming model. An estimator with 50 gave the best results. The label of 0 represents unpopular, while the label of 1 represents popular. Figure 4 describes the results.

	precision	recall	f1-score	support
0	0.72	0.72	0.72	27394
1	0.71	0.71	0.71	26716

Fig. 4. Results Excluding Author Features - Random Forest

VII. Discussion

While we did not have extraordinary results, we believe that we can greatly increase the accuracy of the model through several measures.

First, we would like to increase the sample of the list data set. We initially used a sample of 180,370 to measure the productivity, distribution, and prestige metrics. Our model relies on a bank of authors, journals, publishers, and subjects to collect pre-existing data. With a limited amount of data, we were unable to achieve the greaterresults that we had hoped for. Increasing the sample set applies to the number of names in the list, as well as having an accurate portrayal of the metrics used.

Second, we would like to increase the percentage of usable authors extracted from the list sample set as well as the train/test set. From the list data set, we extracted 122,758 authors out of 180,370 or 68.06%. However, when assigning author relevant features to the train/test data we had an 8.81% match rate for a total of 15,887 authors to fit and test a model. This severely limited what we had hypnotized would be the greatest feature in determining popularity.

Third, we would like to modify the subject content from scopus subject, to the publisher subject or specific subject

of the article. We choose scopus subject due to the inherent binning to make classification easier. However, binning seems to greatly reduce the prediction as it becomes generalized and loses specificity.

Fourth, for more accurate altmetric prediction with NLP, we would like to extend our analysis to other parts of paper like the work ,references , etc. A bigger dataset would enable us to pick some fine details about paper including patterns of grammer, content-similarities that determine its future scope.

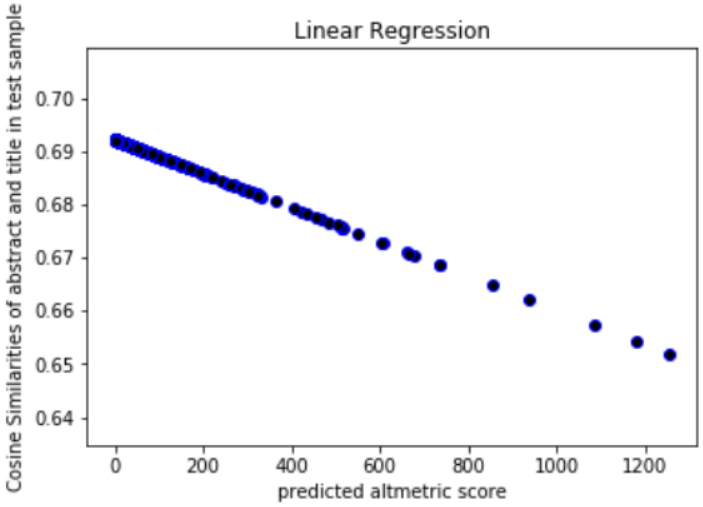

Fig. 5. cosine similarity

VIII. Conclusion and Future Work

In this article we have presented our method for predicting the popularity of scholarly articles across social media. We developed a model based on the citation attributes of the article, thereby developing a generalized prediction free from external factors. We were successful in developing a classifier using random forest that determined whether a scholarly article would be popular 71.22% of the time.

Going forward, we would like to extend the current model so that it is able to predict popularity of specific social media sites. The altmetric score is the result of an algorithm that gives a weighted count to a number of sites including Twitter, Reddit, Wikipedia, etc. We would also like to include time and geography in the model, so that we can identify when and where a specific article will make an impact.

References

[1] Ahmed, M., Spagna, S., Huici, F., Niccolini, S. (2013). A peek into the future: Predicting the evolution of popularity in user generated content. Proceedings of the sixth ACM international conference on Web search and data mining, pages 607-616

[2] Costas, Z. Zahedi, and P. Wouters,(2014) "Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective," arXiv:1401.4321, Jan 2014.

[3] Eysenbach G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research, 13(4), e123. doi:10.2196/jmir.2012

[4] Finch T., OHanlon N., Dudley S. P. (2017). Tweeting birds: online mentions predict future citations in ornithology. R. Soc. Open sci. 4: 171371.

[5] Galligan, F., Dyas-Correia, S. (2013) "Altmetrics: Rethinking the Way We Measure. Serials Review, 39:1, 56-61, DOI: 10.1080/00987913.2013.10765486

[6] Haustein S, Lariviere V., Thelwall M., Amyot D., Peters I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ?. Journal of the Association for Information Sciences and Technology. it - Information Technology, 56(5), pages 207-215

[7] Kudlow, Paul et al. (2107). Online distribution channel increases article usage on Mendeley: a randomized controlled trial. Scientometrics.

[8] Kwak, H, Lee, J. G.(2014). Has Much Potential but Biased: Exploring the Scholarly Landscape in Twitter. Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion, 7 Apr. 2014, pp. 563564., doi:10.1145/2567948.2576956

[9] Maclaughlin A., Wihbey J., Smith, D. A. (2018). Predicting news coverage of scientific articles. Proceedings of the Twelfth International AAAI Conference on Web and Social Media. Pages 192-200.

[10] Maflahi, N., Thelwall, M. (2018). How quickly do publications get read? The evolution of mendeley reader counts for new articles. JASIST 69 : 158-167.

[11] Maflahi, N., Thelwall, M. (2016). When are readership counts as useful as citation counts? Scopus versus Mendeley for LIS journals. JASIST 67 (2016): 191-199.

[12] Peoples, B.K., Midway, S.R., Sackett D., Lynch, A., Cooney, P.B. (2016). Twitter Predicts Citation Rates of Ecological Research. PLoS ONE 11(11): e0166570

[13] Peters, I., Beutelspacher, L., Maghferat, P., Terliesner, J. (2012). "Scientific Bloggers under the altmetric microscope. Proceedings Of the American Society". For Information Science And Technology , 49 (1), 14, doi:10.1002/meet.14504901305.

[14] Power, D. J., Phillips-Wren, G. (2011). "Impact of Social Media and Web 2.0 on Decision-Making". Journal of Decision Systems, 20(3), 249261. doi:10.3166/jds.20.249-261

[15] Priem, J., Piwowar, H. A., Hemminger, B. M. (2012) Altmetrics in the wild: Using social media to explore scholarly impact. ArXiv e-prints, Mar. 2012

[16] Shuai, X., Jiang, Z., Liu, X., Bollen, J. (2013). A comparative study of academic and Wikipedia ranking. Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, July 22-26, 2013, Indianapolis, Indiana, USA [doi¿10.1145/2467696.2467746]

[17] Starbuck E., Purtee S., (2017) "Altmetric scores: short-term popularity or long-term scientific importance", Digital Library Perspectives, Vol. 33 Issue: 4, pp.314-323

[18] Szabo, G., Huberman, B. A., (2008) "Predicting the Popularity of Online Content".dx.doi.org/10.2139/ssrn.1295610

[19] Thelwall M. (2016). Interpreting correlations between citation counts and other indicators, Scientometrics. 10.1007/s11192-016-1973-7, 108, 1, (337-347).

[20] Thelwall, M. (2018). Early Mendeley readers correlate with later citation counts. Scientometrics : 1-10.

[21] Thelwall. M., Haustein, S., Larivire, V., Sugimoto, C.R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLoS ONE 8(5): e64841

[22] Thelwall, M., Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts?. Journal of Informetrics, 10.1016/j.joi.2018.01.008, 12, 1, (237-248).

[23] Weller K. (2015). Social Media and Altmetrics: An Overview of Current Alternative Approaches to Measuring Scholarly Impact. In: Welpe I., Wollersheim J., Ringelhan S., Osterloh M. (eds) Incentives and Performance. Springer, Cham

[24] Wouters, P., Costas, R. (2012). "Users, Narcissism and control - Tracking the impact of scholarly publications in the 21st Century". Utrecht, The Netherlands: SURFfoundation