Using Social Media and Scholarly Text to Predict Public Understanding of Science

Harish Varma Siravuri
Northern Illinois University
DeKalb, Illinois
hsiravuri@niu.edu

Akhil Pandey Akella
Northern Illinois University
DeKalb, Illinois
aakella@niu.edu

Christian Bailey
Northern Illinois University
DeKalb, Illinois
cbailey10@niu.edu

Hamed Alhoori
Northern Illinois University
DeKalb, Illinois
alhoori@niu.edu

ABSTRACT

People often struggle to understand scientific texts, which leads to miscommunication and often to inaccurate and even sensationalistic reports of research. Identifying and achieving a better understanding of the factors that affect comprehension would be helpful to analyze what improves public understanding of science. In this study, we generate features from scientific text that represent some common text structures and use them to predict the semantic similarity between the scientific text and the textual content posted by the general public about the same scientific text online. In this endeavor, we built regression models to achieve this purpose and evaluated them based on their R-squared values and mean squared errors. R-squared values as high as 0.73 were observed, indicating a high chance of a relationship between certain textual features and the public's understanding of science.

KEYWORDS

Public Understanding of Science, Text Comprehension, Altmetrics

ACM Reference Format:
Harish Varma Siravuri, Akhil Pandey Akella, Christian Bailey, and Hamed Alhoori. 2018. Using Social Media and Scholarly Text to Predict Public Understanding of Science. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, Article 4, 2 pages. https://doi.org/10.1145/3197026.3203890

1 BACKGROUND

1.1 Introduction

Promoting public understanding of science has always been an important consideration associated with the broader societal impact of research. Scientific literacy has been promoted as an important aspect of citizenship [6]. According to McGinn and Roth [5], scientific literacy is an important quality in promoting "good citizenship practices" such as participation in scientific laboratories, activist movements, the judicial system, and other communities. Further, scientific literacy is a significant driver of economic growth, and for this reason virtually every modern society has shown a commitment to promoting scientific study and determining the public's understanding of scientific discoveries and advances.

"The Public Understanding of Science" [1], a report published by the Royal Society, is widely considered to have given rise to the current interest in understanding and promoting scientific literacy. Interest in this area is fueled by the widely held belief that science will be the ultimate beneficiary of any gains in scientific literacy among the public [8]. In this regard, some studies investigate the relationship between textual complexity and reading comprehension [2, 9]. The existence of a similar relationship between scientific texts and public understanding of science would make it possible to identify research that is likely to be well understood by general readers.

In Conant's [3] view, there is a need to adopt scientific methodologies in order to promote critical and strategic thinking among the general public. A survey of students and teachers [10] suggests that there is a misapprehension about the purpose of scientific research among the general public. The previous study identifies the public's lack of understanding in regard to the purpose of science and offers the conclusion that the public would be well served if the purpose and place of science in society were clarified. However, merely promoting the purpose of science will not suffice: people need to read and understand scientific texts in order to understand the rationale underlying them and the relative validity of the results and their implications. Unfortunately, the vocabulary used in scientific texts is challenging to the general reader. Evans and Durant [4] proposed a two-dimensional structure for measuring civic scientific literacy, and Miller [7] suggested a three-dimensional structure to fulfill the same purpose. It is worth noting that in both of these studies, the concept for measuring civic scientific literacy includes the dimension of vocabulary as a basic scientific construct. Most previous studies focused on measuring or improving scientific comprehension rely on surveys using a sample of science students. In this study, we focused on examining readers' comprehension of the scientific text and predicting it based on various features derived from the text.

1.2 Data Collection

The data used in this study was obtained from Altmetric.com. The data includes information regarding online activity relating to 5.2

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203890

million scholarly research outputs. Initial analysis showed that of the 5.2 million articles, over 1.7 million articles had been shared and talked about on blogs. We further randomly sampled 1% of the dataset and extracted text from the abstract section and the blog posts to build a smaller dataset consisting of 17,736 data points for further analysis. We used regular expressions to filter out texts that contained only hyperlinks to the scholarly text. Also, we removed textual content that exactly matched the title or sentences from the abstract to avoid any bias caused by social media content in which only the scholarly output is noted without an accompanying discussion of it.

2 METHODS

2.1 Feature Generation

We generated a set of five features - a target variable and four predictors - which we later used to build the regression models.

2.1.1 Target Variable. A 'comprehension score' that reflects the extent to which readers understood the scientific text. The target variable is representative of how semantically similar the text from the blogs is to the abstract. We used cosine similarity to compute the semantic similarity because the magnitude of the vectorized forms of words from either documents is significant. Using Euclidean distance instead of cosine similarity does not take this into consideration. Also, we used L2 normalization form instead of L1 normalization form for normalizing the vectors. The reason for this is we needed a single analytical solution while normalizing the vectors which could be achieved only by using L2 norm.

2.1.2 Predictors. We also generated the following four features using the scientific text to use as predictors in the regression models.

(1) Lexical diversity of the abstract - The ratio of unique word stems to the total words computed. It is an effective measure of the richness of vocabulary or verbal creativity of a text. We used Yules I measure [11] instead of a simple frequency-based measure, since it yields an unbiased result irrespective of the length of the text.
(2) Average word length - The mean of the number of characters in each word in the abstract.
(3) Average sentence length - The mean of the number of words in each sentence in the abstract.
(4) Frequency of words longer than the average word length - A measure of the number of long words that have more characters than the average word in the abstract.

2.2 Regression

Using the processed data, we built five regression models: Decision Tree Regressor, Random Forest Regressor with 100 estimators, Support Vector Regressor, KNN Regressor, and a Gradient Boost Regressor. The models used the predictors generated in 2.1.2 to to predict the the comprehension score calculated in 2.1.1. We evaluated the models based on their coefficient of determination (R squared values) and mean squared errors, as shown in Table 1.

The Decision Tree Regressor and the Random Forest Regressor were observed to perform best compared to the other models. We calculated the Gini importance of each feature for both models to determine the relative significance of each feature to the public understanding. The results are shown in Table 2.

Table 1: Coefficient of determination and mean squared errors for different models

	R Squared	Mean Squared Error
Decision Tree Regression	0.7356	0.01148
Random Forest Regression	0.6486	0.0074
Support Vector Regression	0.1202	0.0089
KNN Regression	0.0824	0.0074
Gradient Boost Regression	0.0609	0.0071

Table 2: Gini importance of each feature in respect to the Decision Tree and Random Forest Regressors

	Decision Tree	Random Forest
Lexical Diversity	0.2588	0.2657
Average Word Length	0.3122	0.2799
Average Sentence Length	0.2476	0.2622
Frequency of words longer than average word length	0.1815	0.1922

3 CONCLUSIONS AND FUTURE WORK

The results indicate the existence of a relationship between the scientific text and the public understanding of the text. The models establish a way to understand and predict how well readers are likely to comprehend a body of scientific text based solely on features derived from the text itself. In the future, we plan to augment the models using additional features and full texts instead of text drawn only from the abstract sections.

REFERENCES

[1] 1985. The Public Understanding of Science - Royal Society. (1985).
[2] Rebekah Georger Benjamin and Paula J Schwanenflugel. 2010. Text complexity and oral reading prosody in young readers. *Reading Research Quarterly* 45, 4 (2010), 388–404.
[3] James B. Conant. 1952. Science and Common Sense. *The Journal of Philosophy* 49, 1 (1952), 11–20.
[4] Geoffrey Evans and John Durant. 1995. The relationship between knowledge and attitudes in the public understanding of science in Britain. *Public Understanding of Science* 4, 1 (1995), 57–74.
[5] Michelle K. McGinn and Wolff-Michael Roth. 1999. Preparing Students for Competent Scientific Practice: Implications of Recent Research in Science and Technology Studies. *Educational Researcher* 28, 3 (1999), 14–24.
[6] Jon D. Miller. 1989. Scientific Literacy: A Conceptual and Empirical Review. *Daedalus* 112, 2 (1989), 29–48.
[7] Jon D Miller. 1998. The measurement of civic scientific literacy. *Public understanding of science* 7, 3 (1998), 203–223.
[8] Steve Miller. 2001. Public understanding of science at the crossroads. *Public Understanding of Science* 10, 1 (2001), 115–120.
[9] Sanja Stajner, Richard Evans, Constantin Orasan, and Mitkov Ruslan. 2012. What can readability measures really tell us about text complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility.*
[10] Leland L. Wilson. 1954. A study of opinions related to the nature of science and its purpose in society. *Science Education* 38, 2 (1954), 159–164.
[11] G. Udny Yule. 1912. On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society* 75, 6 (1912), 579–652.