Investigating Scientific Misinformation Using Different Modes of Learning

Autumn Toney-Wails*, Kornraphop Kawintiranon, Lisa Singh, Paul-Emmanuel Courtines, Alexander Lin and Haofei Zhang

Georgetown University, Washington, D.C. USA

Abstract

This paper presents an initial analysis of scientific misinformation from three areas of research: Computer Science, Environmental Science, and Medicine. We investigate keywords in publication titles and abstracts from retracted scientific publications, which we view as a proxy for misinformation publications. Using the Altmetric Attention Score as a signal of publication popularity, we group articles into low-popularity and high-popularity subsets. We apply three modes of learning (unsupervised, semi-supervised, and supervised), to identify main themes from scientific research publications and compare the results between publication popularity sets. We find that while there is overlap among the terms identified by different methods, they are not the same. However, general topic coverage using different words is similar, highlighting the difficulty in identifying keyword "markers" for popular, poor-quality scientific information.

Keywords

scientific documents, misinformation, altmetrics

1. Introduction

From vaccines to climate change, there are known controversial scientific research areas that have discrepancies surrounding their scientific validity, particularly in politically-charged environments [1, 2, 3]. Recent studies have shown a rise in public skepticism of scientists and scientific research, with 35% of Americans believing that the scientific method may be used to produce "any result a researcher wants" and less than 20% of Americans believing that scientists are transparent in their work and hold themselves accountable for mistakes in their publications [4, 3, 5]. This scientific distrust and controversy is a leading factor in research focusing on *scientific misinformation*, as it undermines the public's ability to consume and trust scientific information [2].

While there is no universal set of steps that leads to scientific discovery, there are particular characteristics of research across all disciplines of science that distinguish it from general inquiry, which make it rigorous and reliable. Generally, the scientific method involves 1) developing a theory or hypothesis, 2) conducting qualitative and/or quantitative experiments to measure observations and collect results, and 3) deriving conclusions from experimentation [6, 7]. Thus, scientific research is considered to be principled, as it relies on reproducible experiments and evidence-based conclusions. However, with the increase of scientists, publication venues, and online platforms for information sharing partnered with the "publish or perish" reality, the challenge of preserving the rigour and reliability of scientific research is magnified [8, 5].

Scientific misinformation is difficult to characterize, and as a result, difficult to identify [9, 10, 3]. We adopt the following scientific misinformation definition from Southwell et al. [3]: "publicly available information that is misleading or deceptive relative to the best available scientific evidence and that runs contrary to statements by actors or institutions who adhere to scientific principles." The majority of research on misinformation focuses on news articles and social media in the context of fake news and propaganda campaigns and analyzes how these stories disseminate through social networks. We found that a critical limitation of this avenue of work is that scientific misinformation is not yet well-researched and there are no available ground-truth datasets.

In this paper, we link scientific misinformation content to popularity. We are interested in understanding if it is possible to tease out themes of those pieces of scientific thought that are poor quality and popular from those that are not. Here, we use retracted publications as a proxy for identifying publications with a high potential for misinformation and the Altmetric[1] Attention Score as a proxy for publication popularity. For this exploratory analysis, we compare text analysis techniques that employ different modes of learning: unsupervised, semi-supervised, and supervised. Each text analysis technique is performed on retracted scientific publications with

The AAAI-23 Workshop on Scientific Document Understanding, February 14, 2023, Washington, D.C.

*Corresponding author.
autumn.toney@georgetown.edu (A. Toney-Wails); kk1155@georgetown.edu (K. Kawintiranon); lisa.singh@georgetown.edu (L. Singh); pyc25@georgetown.edu (P. Courtines); al1439@georgetown.edu (A. Lin); hz251h@georgetown.edu (H. Zhang)

[1] www.altmetric.com

low popularity and high popularity in major research domains: Computer Science, Environmental Science, and Medicine. We find that all three methods produce complimentary, non-overlapping, but not contradictory results, highlighting the complexity of identifying "markers" for popular, poor-quality scientific information.

To summarize, the main contributions of this paper are as follows: 1) analyzing scientific misinformation across different domains of research, 2) measuring the prevalence of scientific misinformation, and 3) comparing learning text analysis techniques applied to scientific research publications.

2. Experimental Design

We apply three modes of learning for text analysis on our data. First, we use unsupervised learning methods for traditional keyword extraction. Next, we employ a semi-supervised, generative topic model that uses expert identified seed terms to guide the topic discovery process.Lastly, we run an interpretable, supervised machine learning model that predicts popularity and identify keyword features that are used to separate the classes. Figure 1 shows the overview process. Each method uses text from the titles and abstracts of scientific publications. We normalize the text by setting all tokens to lowercase, removing urls, digits, symbols, and the word *retracted*. This normalized text is the input to all of our models.

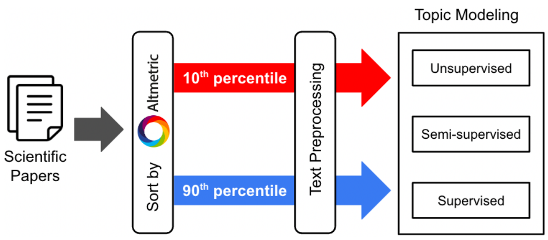

Figure 1: The overview process of our experimentation.

Keyword Extraction Methods (unsupervised): We use the three keyword extraction methods as shown in Table 1: 1) term frequency-inverse document frequence (TF-IDF), 2) YAKE [11], and 3) KeyBERT [12]. Each method provides a different approach to keyword extraction (term frequency, unsupervised feature extraction, and contextualized word embeddings), enabling us to compare results across extraction methods. The last two columns of the table show the Python package used and the non-default parameters in cases where the default parameters were not used.

Generative Modeling (semi-supervised): Because we have some domain knowledge, we test a semi-supervised topic model, Guided Topic-Noise Model (GTM) [14]. In addition to text, GTM takes a set of seed words for topics as input and implements the Generalized Polya Urn (GPU) sampling method to help keep seed words within a single topic together during the generation process. We selected GTM because we have short, noisy text, and GTM generates both a topic and noise distribution, removing words that are domain-specific but appear across a large number of topics. It also identifies other topics that domain experts may have missed. In our implementation, we use the default parameters for GTM.

Predictive Modeling (supervised): We train a Decision Tree on our datasets to test if we can identify important n-gram features (key terms) in predicting if a research publication is in the top or bottom 10% of Altmetric Attention Scores. We use sklearn's tree implementation and its default parameters.

3. Datasets

For our analysis, we use retracted publications as a proxy for scientific research that *could* be scientific misinformation. By using these scientific publications in our study we are not definitively labeling them as scientific misinformation. An example of a peer-reviewed, retracted (due to misinformation) publication is *Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis* [15]. This publication is in the top 5% of all research outputs, from any year, scored by Altmetric [16]. Figure 2 displays the overview of attention found on Altmeteric for this publication, which received an Altmetric Attenion Score of 22,503. We are interested in the comprehensive Altmetric Attention Score (displayed in the colorful circle), which represents a combination of all the attention a publication receives (displayed in the category counts on the far left).

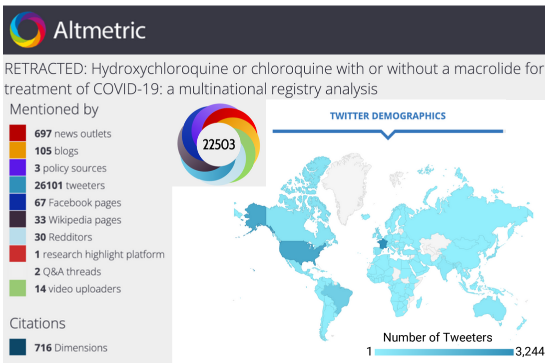

Figure 2: Altmertic.com overview of attention.

Method	Description	Tool	Parameters
TFIDF	Calculates the frequency of a word in a document and multiplies the logarithm of the number of documents divided by the number of documents containing the word	TfidfVectorizer [13]	[stop_words='english']
YAKE	Uses statistical text features from single documents without using term frequency	yake [11]	[language = "en", deduplication_thresold = 0.9, deduplication_algo = 'seqm', numOfKeywords = 20]
KeyBERT	Computes the cosine similarity between the subphrases in a document that are the most similar to the document itself using BERT-embeddings	KeyBERT [12]	*default*

Table 1
Keyword extraction descriptions.

Retraction Watch Database: We used the publicly available, manually curated Retraction Watch Database [17]. Retraction Watch contains 22,614 articles with a DOI, enabling us to link the articles to Dimensions, a large scientific literature database, and obtain their titles and abstracts for analysis. Because Retraction Watch is manually curated, each retracted paper is labeled with at least one reason for retraction; there are 105 unique reasons, such as Investigation by Journal/Publisher, Concerns/Issues About Data, and Unreliable Results. Table 2 provides the top five retraction reasons by number of publications for the research areas that we analyze. There is minimal overlap in the top five reasons across research areas, but at least three of the five reasons are concerned with scientific integrity related to data and methods.

	Computer Science	Environmental Science	Medicine
1	Fake Peer Review	Investigation by Journal/Publisher	Concerns/Issues About Data
2	Plagiarism of Article	Upgrade/Update of Prior Notice	Notice - Limited or No Information
3	Date of Retraction/Other Unknown	Fake Peer Review	Investigation by Journal/Publisher
4	Euphemisms for Plagiarism	Rogue Editor	Unreliable Results
5	Duplication of Article	Randomly Generated Content	Withdrawal

Table 2
Top five retraction reason by research area.

Dimensions: Our dataset of paper titles and abstracts is sourced from Dimensions, an inter-linked research information system provided by Digital Science [18]. We have three sets of scientific research articles that we select from Dimensions: Computer Science, Environmental Science, and Medicine. Each publication in Dimensions is labeled with a broad area of research, which we use to create our subsets of publications. Using the DOIs from these three publication sets, we query the Altmetric API to identify publications with Altmetric attention scores [16]. The Altmetric Attention Score is a weighted count of the online attention a research publication receives from various groups, such as scientists, policy-makers, news sources, and the general public. The Altmetric Attention Score is not an indicator of scientific impact.

Popularity	Comp. Science	Environ. Science	Medicine
Low	97	43	355
High	34	18	210

Table 3
Number of retracted publications in each popularity set across the three broad areas of research.

For each of the three subsets of research publications (Computer Science, Environmental Science, and Medicine) with Altmetric scores, we generate two subcategories, **low-popularity** and **high-popularity**. We select the publications with a bottom 10% Altmetric Attention Score as low-popularity and the publications with a top 10% Altmetric Attention Score as high-popularity. Table 3 displays the number of retracted publications in each of the six categories we analyze. Medicine has significantly more publications with Altmertic data compared to Computer and Environmental Science.

4. Empirical Evaluation

We perform our text analysis on the low-popularity and high-popularity sets of scientific research publications from our three domains. For all methods, except GTM, the only input required is the input text; GTM also requires a seed set of words organized by topics. We implemented noiseless Latent Dirichlet Allocation on all sets

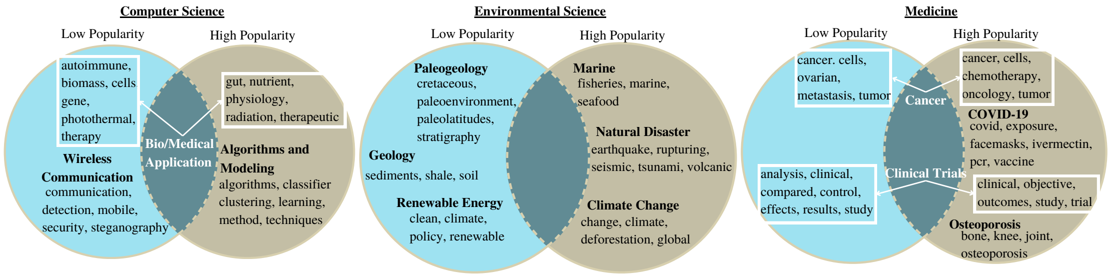

Figure 3: Scientific publication themes and key terms by research area and publication popularity. The white box outline highlights topics that appear in both the low and high popularity publication sets.

of publications to find candidate seed words that could be organized into coherent topics and then manually select the final list of seed words. Table 4 displays the seed words selected for the GTM experiments.

We first compared our results across all five methods for each subset of research and publication popularity and found that no terms appeared in all five methods for any subset of publications that we analyzed. However, we did find that *different* words related to the same theme appeared across all five methods; for example the high popularity, Medicine results has facemasks (TF-IDF), adult exposure (KeyBERT), ivermectin (YAKE), pcr (decision tree), and covid (GTM).

While the keyword results across all five methods varied, we find general themes for each research area and popularity (see Figure 3). Under each theme we provide a sample of keywords that appeared from at least one of the methods. Computer Science and Medicine have overlapping themes between the low popularity and high popularity publications, whereas Environmental Science does not. Additionally, Computer Science has a theme relating to biology and medicine applications in both low and high popularity subsets, which resulted in words that are not directly related to computer science, such as biomass and radiation.

We find that the Medicine subset of research publications produced the most coherent results, perhaps indicating that these methods perform best on larger sets of documents.

Corpus	**Seed Words**
Computer Science: *Low Popularity*	1. network, detection, realtime 2. data, largescale, information 3. mobile, wireless, application
Computer Science: *High Popularity*	1. algorithm, method, system 2. data, learning, study 3. weight, cells, theraputic
Environmental Science: *Low Popularity*	1. soil, area, sea 2. china policy
Environmental Science: *High Popularity*	1. climate, change, control 2. water, resources
Medicine: *Low Popularity*	1. cancer, cells, ovarian, breast 2. disease, ci, survival, liver 3. mir, metastasis, tumor
Medicine: *High Popularity*	1. cancer, cells, tumor, brain 2. coronavirus, sarscov, covid 3. diabetes, insulin, weight 4. children, vaccine, age

Table 4
GTM seed words.

5. Conclusions

In this work, we investigate scientific research misinformation. As an initial analysis, we select publications from three broad areas of research (Computer Science, Environmental Science, and Medicine) and attempt to identify keyword differences between low popularity and high popularity scientific misinformation using unsupervised, semi-supervised, and supervised modes of learning on scientific research publication text. We find that across all experimental results, we are able to identify themes of research topics in each research area using different learning approaches, but some themes overlap in popularity levels, highlighting the complexity of using keywords as indicators for this task. Future work will consider using network metrics to identify popular poor quality scientific information.

Acknowledgments

This work was supported in part by the Massive Data Institute (MDI), the Fritz Family Fellows Program, and the Center for Security and Emerging Technology (CSET) at Georgetown University.

References

[1] D. A. Scheufele, N. M. Krause, Science audiences, misinformation, and fake news, Proceedings of the National Academy of Sciences 116 (2019) 7662–7669.

[2] J. Farrell, K. McConnell, R. Brulle, Evidence-based strategies to combat scientific misinformation, Nature climate change 9 (2019) 191–195.

[3] B. G. Southwell, J. S. B. Brennen, R. Paquin, V. Boudewyns, J. Zeng, Defining and measuring scientific misinformation, American Academy of Political and Social Science 700 (2022) 98–111.

[4] G. C. Kabat, Taking distrust of science seriously: To overcome public distrust in science, scientists need to stop pretending that there is a scientific consensus on controversial issues when there is not, EMBO reports 18 (2017) 1052–1055.

[5] C. Funk, Key findings about americans' confidence in science and their views on scientists' role in society, 2020.

[6] H. G. Gauch, Scientific method in practice, Cambridge University Press, 2003.

[7] National Academies of Sciences, Engineering, and Medicine, Reproducibility and Replicability in Science, Technical Report, 2019.

[8] D. Sarewitz, The pressure to publish pushes down quality, Nature 533 (2016) 147–147.

[9] E. K. Vraga, L. Bode, Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation, Political Communication 37 (2020) 136–144.

[10] J. N. Druckman, Threats to science: Politicization, misinformation, and inequalities, The ANNALS of the American Academy of Political and Social Science 700 (2022) 8–24.

[11] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences (2020).

[12] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265, https://doi.org/10.5281/zenodo.4461265.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research (2011).

[14] R. Churchill, L. Singh, A guided topic-noise model for short texts, in: International World Wide Web Conference, 2022.

[15] M. R. Mehra, S. S. Desai, F. Ruschitzka, A. N. Patel, Retracted: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of covid-19: a multinational registry analysis, Lancet (London, England) (2020) S0140–6736.

[16] Altmetric, Altmetric.com, www.almetric.com/, 2012. Accessed: 2022-01-25.

[17] The Center For Scientific Integrity, The Retraction Watch Database, http://retractiondatabase.org/, 2018. ISSN: 2692-465X. Accessed: 2022-01-25.

[18] D. W. Hook, S. J. Porter, C. Herzog, Dimensions: building context for search and evaluation, Frontiers in Research Metrics and Analytics 3 (2018) 23.