Gathering New Insights using Altmetrics

Hamed Alhoori

Northern Illinois University

Outline

- ❑ **Understanding Scholarly Information Behavior and Altmetrics**
- ❑ Recommending Scholarly Venues
- ❑ Predicting Scholarly and Societal Impact

Understanding Scholarly Information Behavior

More than **$2,000,000,000,000** were spent internationally on research and development in 2017 (R&D Magazine, 2018)

- H. Alhoori, R. Furuta, M. Samaka, and E. Fox, "Anatomy of Scholarly Information Behavior Patterns in the Wake of Academic Social Media Platforms," International Journal on Digital Libraries 2018.
- H. Alhoori, C. Thompson, R. Furuta, J. Impagliazzo, E. Fox, M. Samaka, and S. Al-Maadeed, "The Evolution of Scholarly Digital Library Needs in an International Environment: Social Reference Management Systems and Qatar," ICADL, 2013.

Some Findings

- ❑ Academic social networks
- ❑ Inefficient search
- ❑ Holistic solution
- ❑ Scholarly recommendations
- ❑ Social media reluctance
- ❑ Awareness and misconceptions

Publication Overload

- Affects **78%** of researchers
- At least **2.5 million** articles published yearly (Harnad et al., 2008)
- At least **114 million articles** available on the web (Khabsa and Giles, 2014)
- Inadequate literature review

Where is our Research Community?

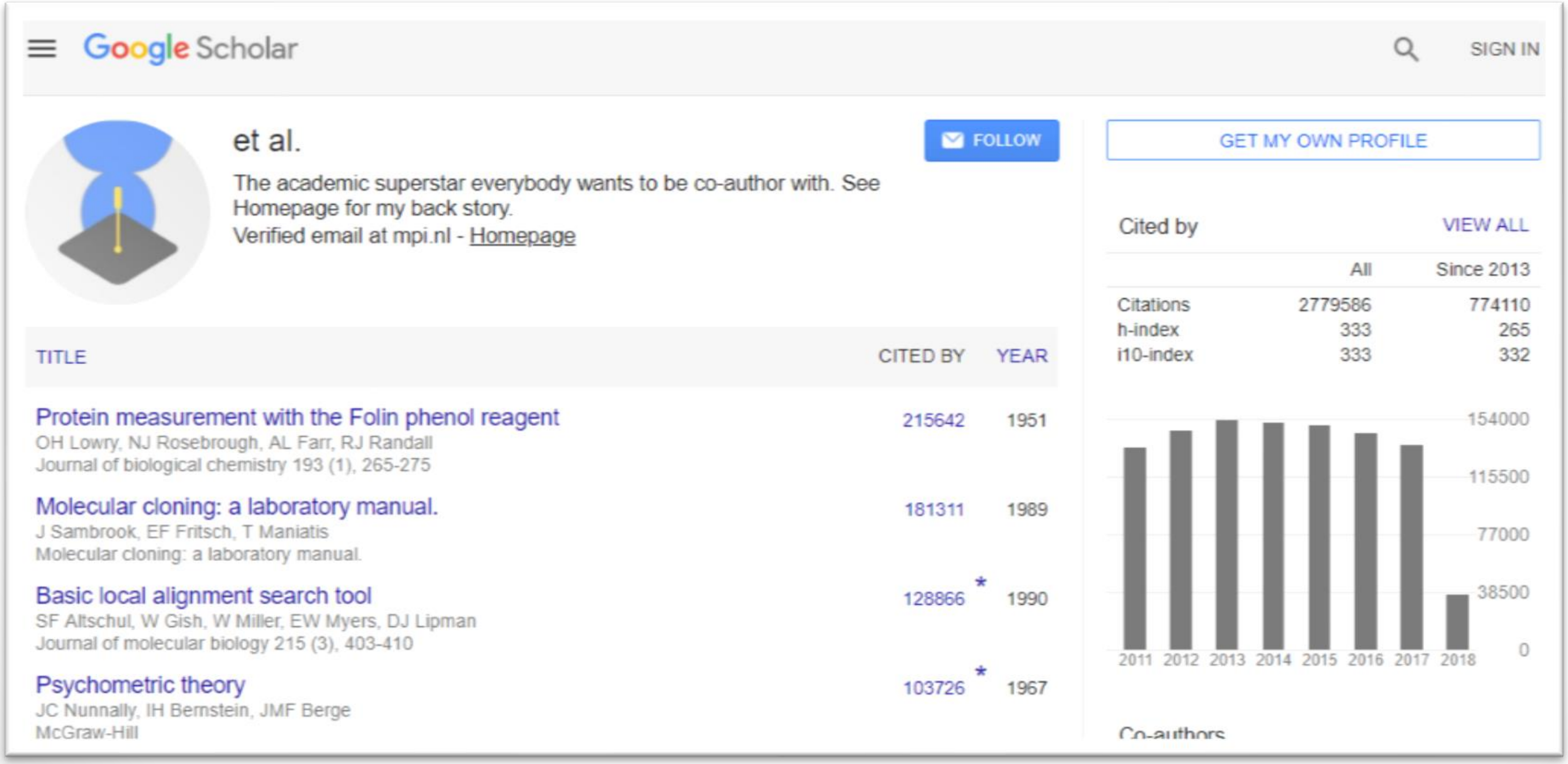
- ❑ Academic administration
- ❑ Library
- ❑ Funding agencies

H. Alhoori "How to Identify Specialized Research Communities Related to a Researcher's Changing Interests," in Proceedings of 2016 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

How to Measure Research Impact?

Drawbacks of Citation Count

- ❑ Time, one measure, disciplinary differences, and page limit.
- ❑ Goodhart's law: “Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes”



et al.

The academic superstar everybody wants to be co-author with. See Homepage for my back story.
Verified email at mpi.nl - Homepage

TITLE	CITED BY	YEAR
Protein measurement with the Folin phenol reagent OH Lowry, NJ Rosebrough, AL Farr, RJ Randall Journal of biological chemistry 193 (1), 265-275	215642	1951
Molecular cloning: a laboratory manual. J Sambrook, EF Fritsch, T Maniatis Molecular cloning: a laboratory manual.	181311	1989
Basic local alignment search tool SF Altschul, W Gish, W Miller, EW Myers, DJ Lipman Journal of molecular biology 215 (3), 403-410	128866 *	1990
Psychometric theory JC Nunnally, IH Bernstein, JMF Berge McGraw-Hill	103726 *	1967

Cited by

	All	Since 2013
Citations	2779586	774110
h-index	333	265
i10-index	333	332

Measuring Research Impact

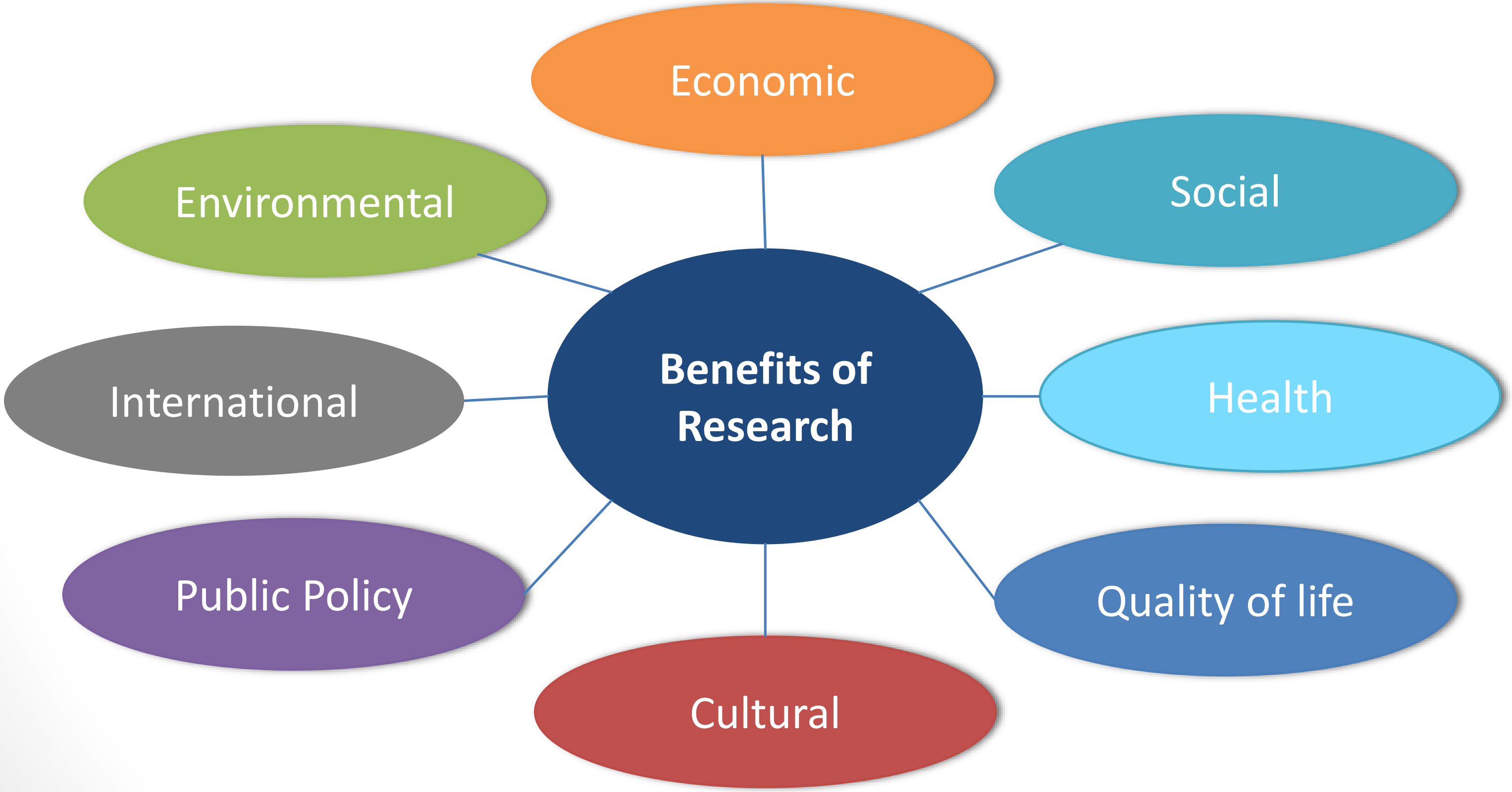
- ❑ Views, downloads, bookmarks, and comments
  - ❑ Each has its benefits and limitations (Neylon, C, 2009)
- ❑ "The notion of scientific impact is a **multi-dimensional** construct that cannot be adequately measured by any single indicator" (Bollen et al., 2009)

metrics?

- ❑ Who moved my ~~cheese~~?

Broader Impacts

- ❑ From publication-based to **product-based assessments** (Heather, 2013)
- ❑ UK, the higher education funding bodies “the impact element will include all kinds of social, economic and cultural benefits and impacts **beyond academia**”

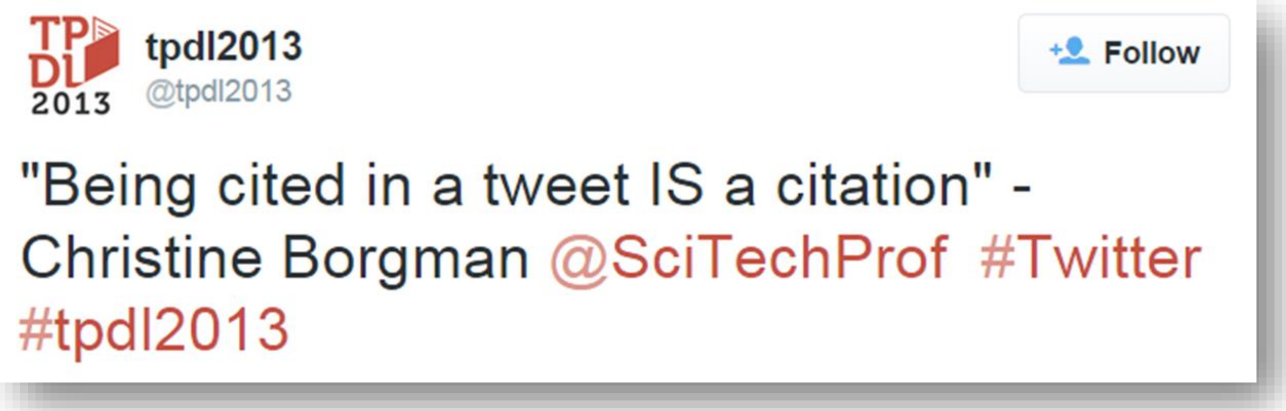


Beyond Citations

twitter
f
MENDELEY
zotero
Research Blogging
crossref
citeulike
ScienceSeeker
Science news from science newsmakers
ResearchGate
Scientific Network
figshare
credit for all your research
slideshare
Present Yourself
DRYAD
academia.edu
follow research
Wikipedia
Google scholar

Altmetrics

Altmetrics



tpdl2013
@tpdl2013

"Being cited in a tweet IS a citation" - Christine Borgman @SciTechProf #Twitter #tpdl2013

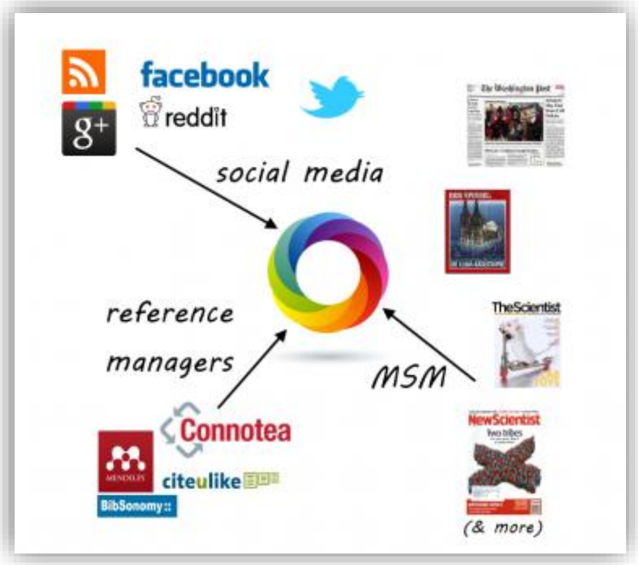
- ❑ Increase in research articles shared on social media **5–10%** a month (Euan and Roe,2013).

Understanding Altmetrics

1. How do **social media platforms** differ in the coverage, usage, and distribution of scholarly works?
2. Is the online attention received by research articles related to **scholarly impact** or due to other factors?
3. Do **open access** articles receive more altmetrics than non-open access articles?

Data

- ❑ Used various **data sources** such as: Twitter, Facebook, CiteULike, Mendeley, F1000, blogs, mainstream news outlets, Google Plus, Pinterest, Reddit, Sina Weibo, the peer review sites PubPeer and Publons, policy documents, and sites running Stack Exchange (Q&A).
- ❑ From **5** to **19 million** scholarly articles.



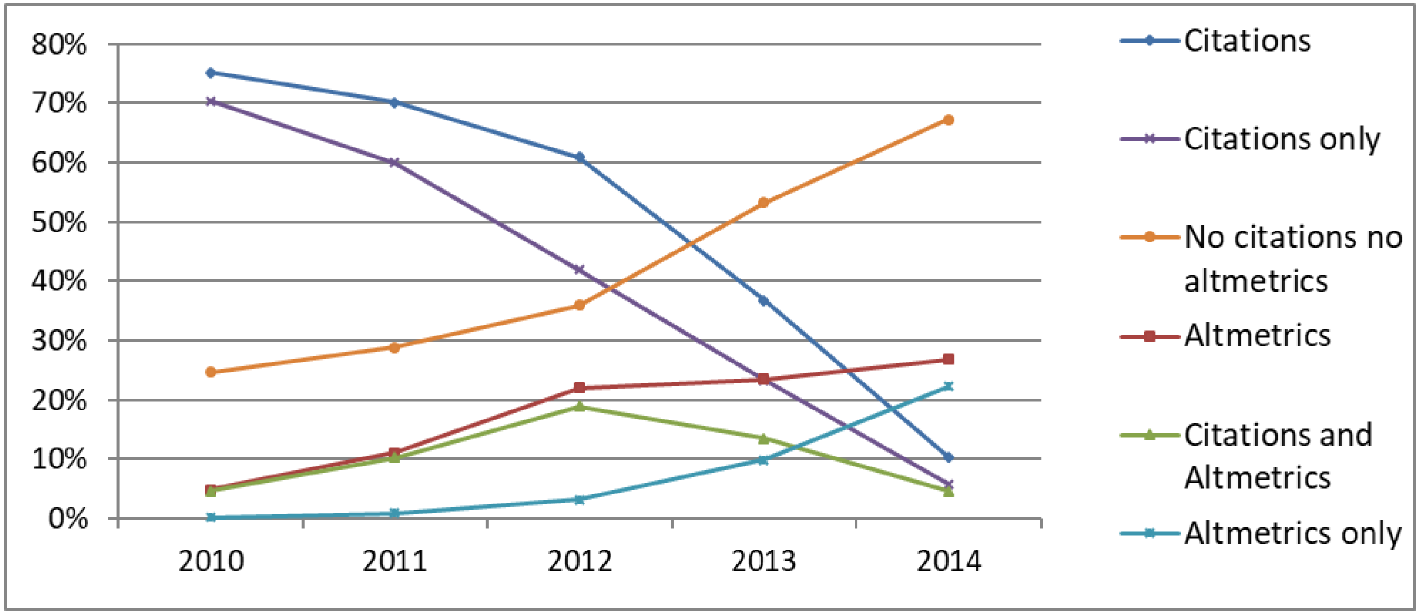
Article-level Altmetrics

- Weak correlations
  - citation-based metrics
  - among themselves
- **Academic social networks** have the highest correlations with citation-based metrics

MENDELEY citeulike

H. Alhoori and R. Furuta, “Do Altmetrics Follow the Crowd or Does the Crowd Follow Altmetrics?,” in Proceedings of 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)

Altmetrics vs. Citations



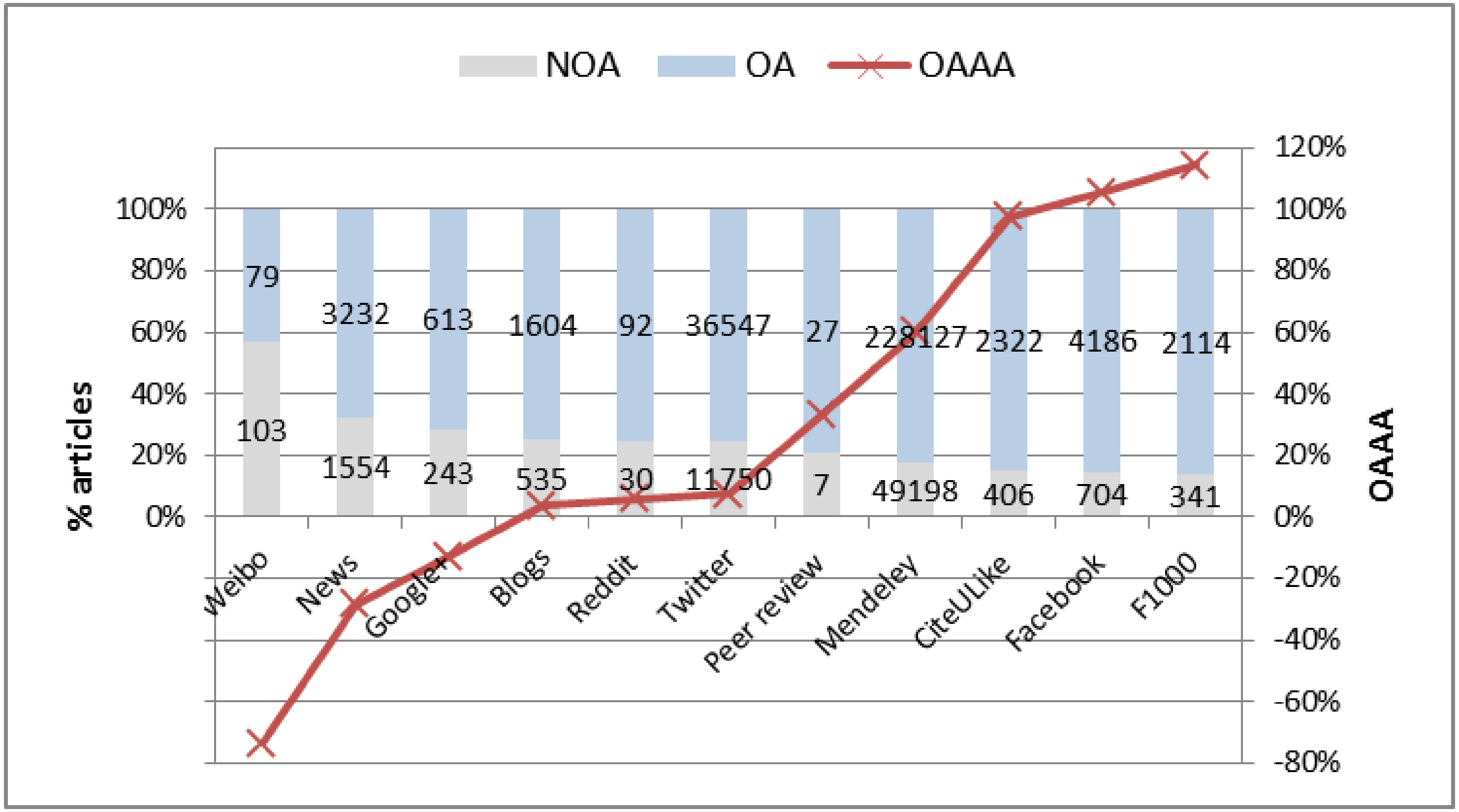
Access-Level Altmetrics

- 23 NOA and OA journals and 42,582 articles
- A parser for Google Scholar
  - Article title search
  - Returns either a web link (OA) or no link (NOA)
- 27,011 articles after filtering (DOIs only, remove duplicates, years outside 2010-2014)

$$OA\ Altmetric\ Advantage(\boldsymbol{OAAA}) = \frac{\overline{OA} - \overline{NOA}}{\overline{NOA}}$$

H. Alhoori, S. Choudhury, T. Kanan, R. Furuta, E. Fox, and C.L. Giles "On the Relationship between Open Access and Altmetrics," iConference 2015.

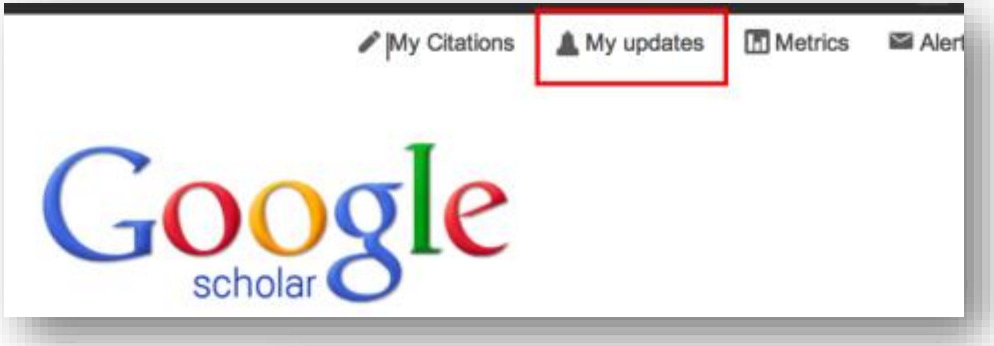
Altmetric-based OAAA

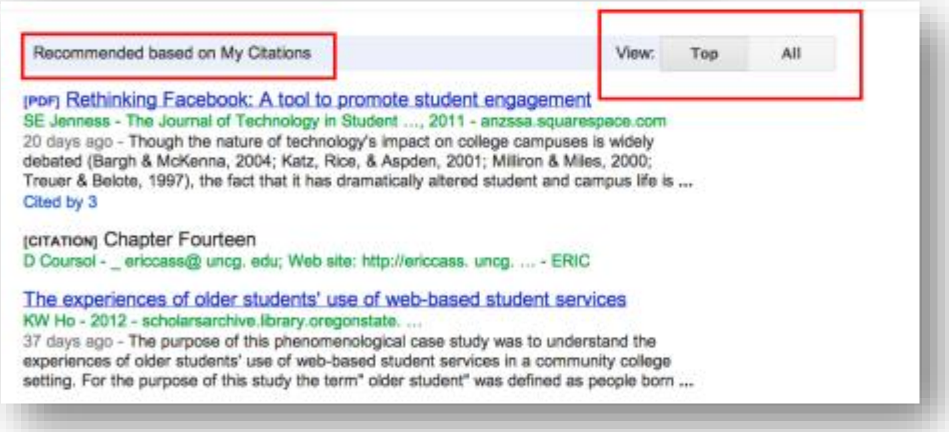


Outline

- ✓ Understanding Scholarly Information Behavior and Altmetrics
- ❑ **Recommending Scholarly Venues**
- ❑ Predicting Scholarly and Societal Impact

Scholarly Recommendation System





Recommended based on My Citations

[PDF] Rethinking Facebook: A tool to promote student engagement
SE Jenness - The Journal of Technology in Student ..., 2011 - anzssa.squarespace.com
20 days ago - Though the nature of technology's impact on college campuses is widely debated (Bargh & McKenna, 2004; Katz, Rice, & Aspden, 2001; Milliron & Miles, 2000; Treuer & Belote, 1997), the fact that it has dramatically altered student and campus life is ...
Cited by 3

[CITATION] Chapter Fourteen
D Coursol - _ericcass@ uncg. edu; Web site: http://ericcass. uncg. ... - ERIC

The experiences of older students' use of web-based student services
KW Ho - 2012 - scholarsarchive.library.oregonstate. ...
37 days ago - The purpose of this phenomenological case study was to understand the experiences of older students' use of web-based student services in a community college setting. For the purpose of this study the term" older student" was defined as people born ...

- ❑ Recommendations based on publications
- ❑ Recommendations based on current interests?

Personal Venue Rating (PVR)

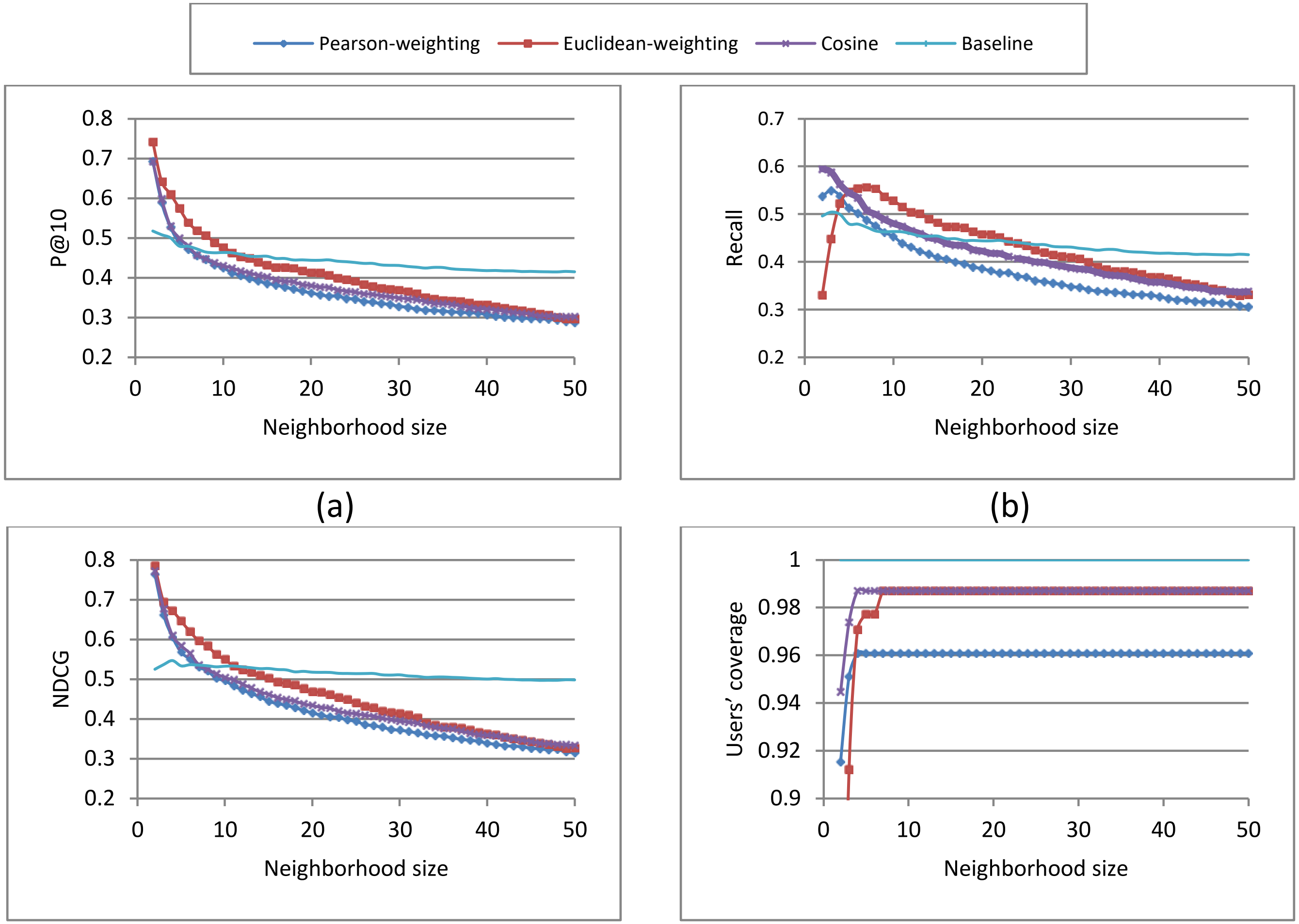
$$r_{u,v} = \sum_{i=y}^{1} w \log(v_{u,i} + 1)$$

$$sim_{x,u} = \frac{\sum_{v=1}^{n}\left(r_{x,v} - \overline{r_x}\right)\left(r_{u,v} - \overline{r_u}\right)}{\sqrt{\sum_{v=1}^{n}\left(r_{x,v} - \overline{r_x}\right)^2}\ \sqrt{\sum_{v=1}^{n}\left(r_{u,v} - \overline{r_u}\right)^2}}$$

$$p_{x,v} = \overline{r_x} + \frac{\sum_{u \in U_v(x)}\left(r_{u,v} - \overline{r_u}\right) sim_{x,u}}{\sum_{u \in U_v(x)} |sim_{x,u}|}$$

H. Alhoori, R. Furuta, “Recommendation of Scholarly Venues Based on Dynamic User Interests,” Journal of Informetrics 2017.

Comparing the user-based CF algorithm with the baseline at different neighborhood sizes



(a)

(b)

(c)

(d)

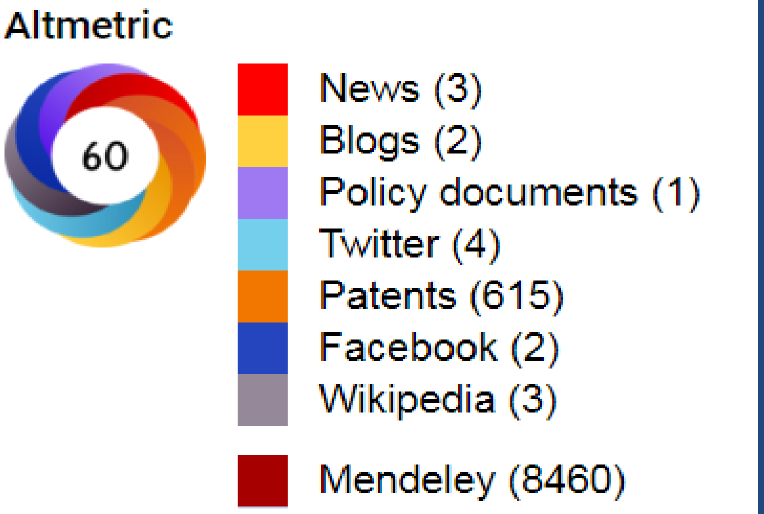
Outline

- ✓ Understanding Scholarly Information Behavior and Altmetrics
- ✓ Recommending Scholarly Venues
- ❑ **Predicting Scholarly and Societal Impact**

A Machine Learning Approach

❑ We built machine learning models that used altmetrics and other features to predict research and societal impact of a research article:

1. Public policy citations
2. News mentions
3. Patent citations
4. Scholarly citations
5. Public understanding of science



Altmetric

60

🟥 News (3)
🟨 Blogs (2)
🟪 Policy documents (1)
🟦 Twitter (4)
🟧 Patents (615)
🟦 Facebook (2)
⬜ Wikipedia (3)

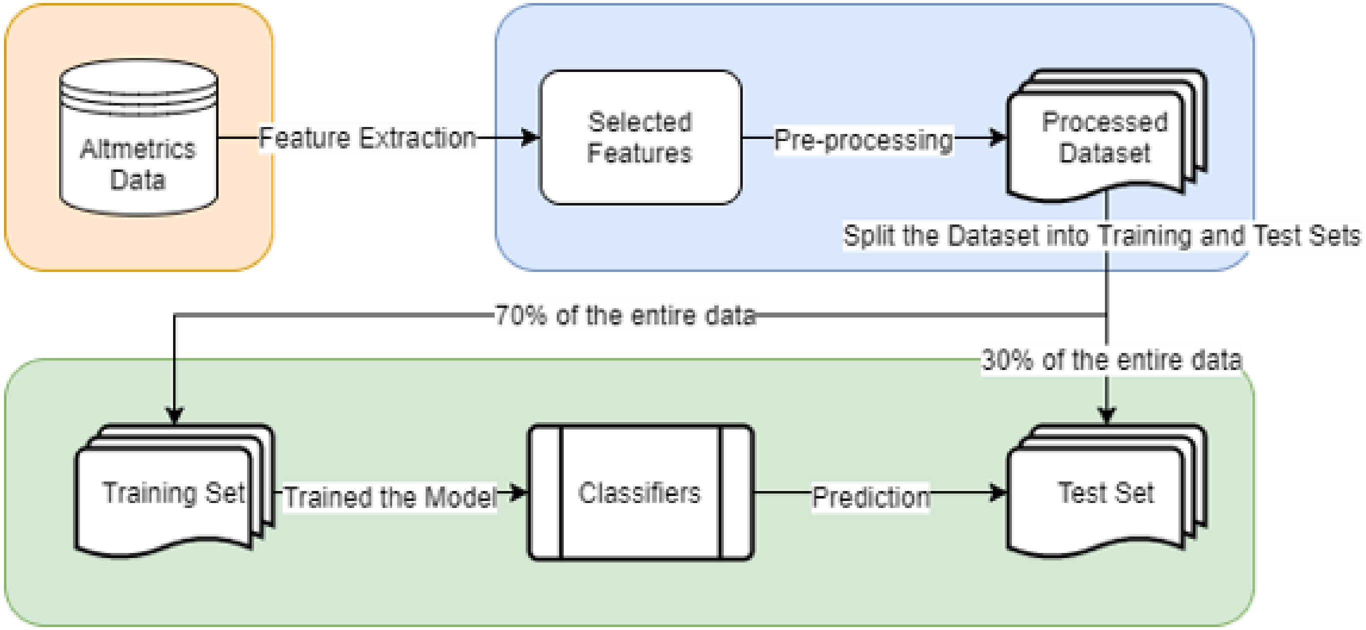
🟥 Mendeley (8460)

Policy Documents

- Evidence based policy making is being encouraged in all areas of public service.
- Number of citations in public policy documents.
- Approximately 180,000 research papers were used.
- **Features** used include: Mendeley, Google+, Wikipedia, Reddit, Blogs, YouTube, Facebook, Twitter, Peer Reviews, and Weibo.

B. Kale, H. V. Siravuri, H. Alhoori, and M. E. Papka, “Predicting research that will be cited in policy documents”, in Proceedings of the 2017 ACM on Web Science Conference, ACM, 2017.

The Classification Process



Classification Results

Model	Accuracy	Precision	Recall	F1 Score
Multinomial Naive Bayes	0.842	0.802	0.905	0.850
Random Forest	0.870	0.826	0.870	0.844
Support Vector Machine	0.868	0.820	0.868	0.824

News

- Attention from news outlets
- Stories are vetted by journalists to decide if they are newsworthy.

H. V. Siravuri and H. Alhoori, “What makes a research article newsworthy?”, Proceedings of the Association for Information Science and Technology, vol. 54, no. 1, 2017.

Results

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.924	0.796	0.658	0.720
Support Vector Machine	0.888	0.806	0.326	0.465
Multinomial Naive Bayes	0.782	0.302	0.365	0.331

Economic Impact of Research

- ❑ A crucial goal of funding R&D has always been to advance economic development.
- ❑ Predicting amount of patent citations can be helpful in measuring the economic impact of research and in understanding how knowledge is commercialized.
- ❑ We found a moderate positive correlation between scholarly citations and patent citations.

A. Shaikh and H. Alhoori, "Predicting Patent Citations to measure Economic Impact of Scholarly Research," in Proceedings of 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL).

Results

	Accuracy	F1-score	Precision	Recall
Logistic Regression	89.7	90.3	90.2	90.4
Decision Tree	92.6	93	92.6	93.4
Naïve Bayes	90.5	90.4	90.7	90.1
Random Forest	93.9	94.5	94.2	94.8

Predicting Scholarly Impact using Altmetrics

Akhil Pandey, Hamed Alhoori, Pavan Kondamudi, Cole Freeman, and Haiming Zhou. "Predicting Scholarly Impact with Altmetrics" (*under review*).

Feature Engineering

- Total of 22 features were used:
  - Tweets
  - Retweets
  - Profession on Twitter
  - Mentions in Tweets (@)
  - Max. Followers on Twitter
  - Hashtags (#)
  - Facebook posts
  - Mendeley Readership
  - Academic Status (Mendeley)
  - CiteULike Readership
  - Total Platforms
  - Platform with Max Mentions
  - Post Length
  - Peer Review sites
  - News Mentions
  - Author Count
  - Publication Age
  - Countries
  - Reddit
  - Blog Mentions
  - Wikipedia citations
  - GooglePlus Mentions

Expt. 1: Articles with non-zero citations

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.862	0.863	1.0	0.927
Decision Tree	0.863	0.863	1.0	0.927
Gradient Boosting	0.863	0.863	1.0	0.927
AdaBoost	**0.866**	0.87	0.993	**0.928**
Bernoulli Naive Bayes	0.835	0.877	0.942	0.908
KNN	0.851	0.883	0.953	0.917
Neural Network	0.860	0.860	1.0	0.925
SVM	0.862	0.863	0.998	0.926

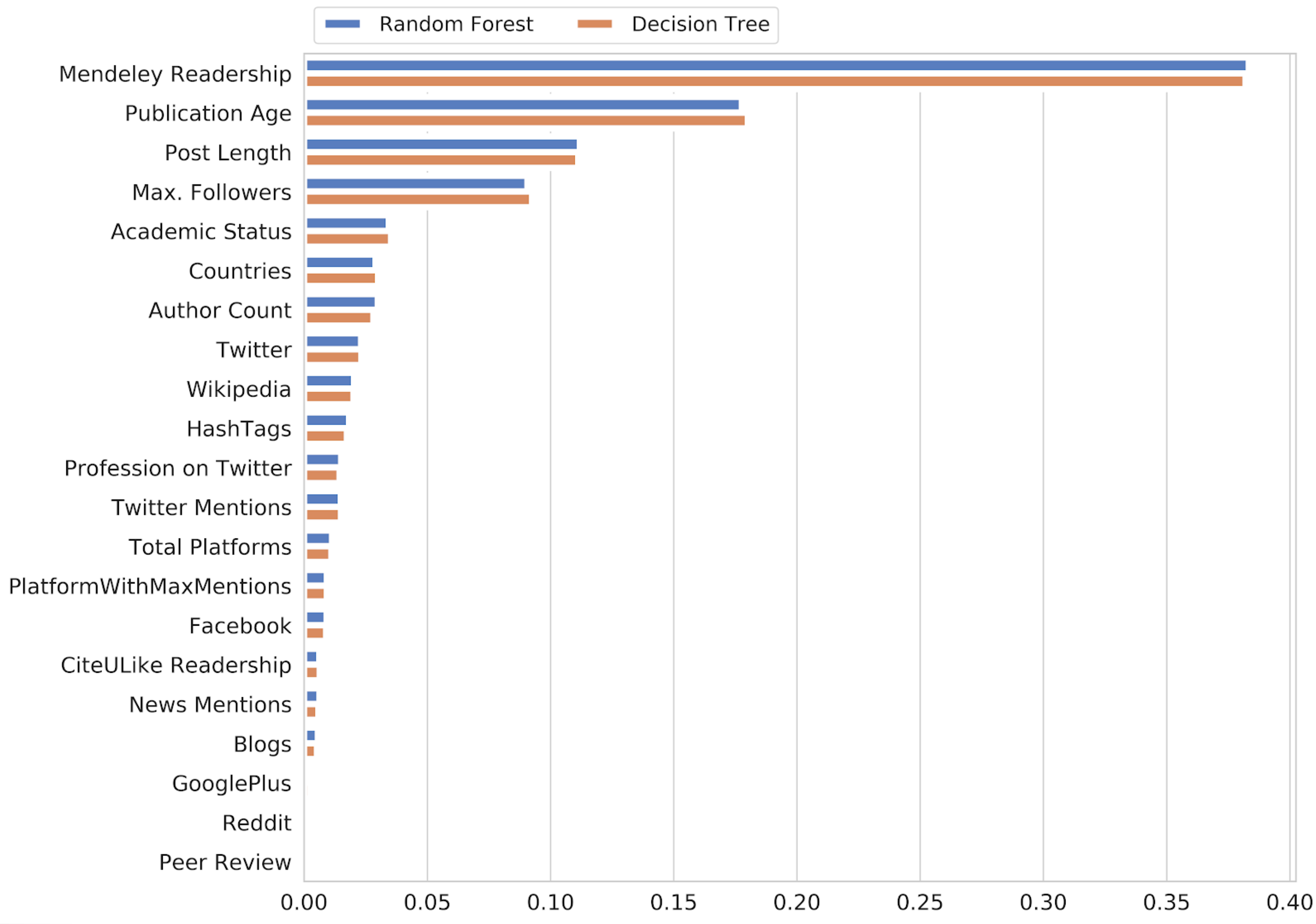
Expt. 2: Articles with least median citations

Model	Accuracy	Precision	Recall	F-1
Random Forest	0.771	0.732	**0.825**	0.776
Decision Tree	0.787	0.783	0.77	0.776
Gradient Boosting	0.793	0.808	0.747	0.776
AdaBoost	**0.797**	0.806	0.760	**0.782**
Bernoulli Naive Bayes	0.675	0.74	0.501	0.597
KNN	0.753	0.777	0.680	0.726
Neural Network	0.791	**0.815**	0.730	0.77
SVM	0.519	0.464	0.006	0.012

Expt. 3: Predict citations count

Model	MSE	R-squared
Random Forest	1.334	0.510
Decision Tree	1.659	0.390
Linear Model	1.759	0.354
Neural Network	**1.288**	**0.522**

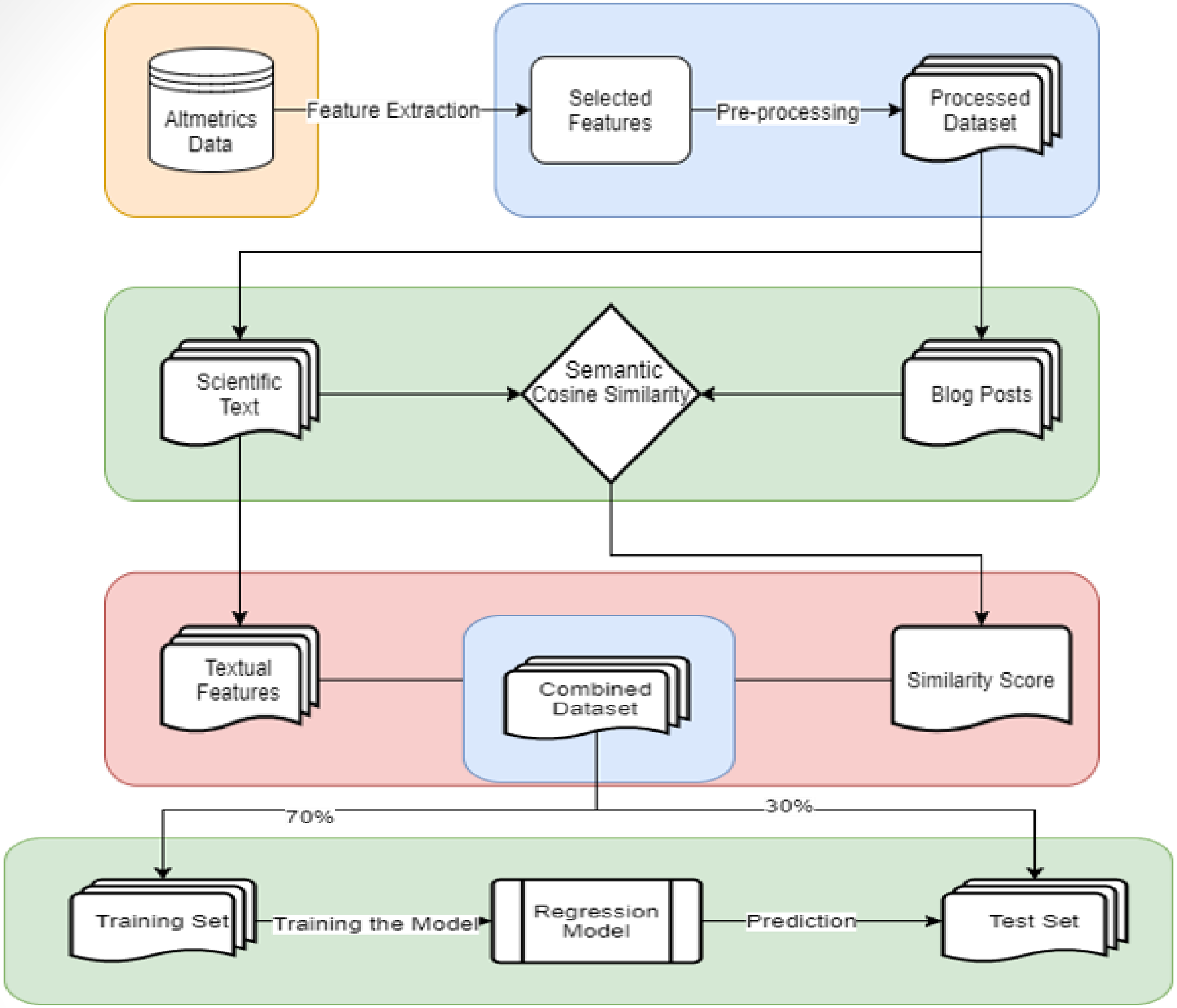
Feature Importance



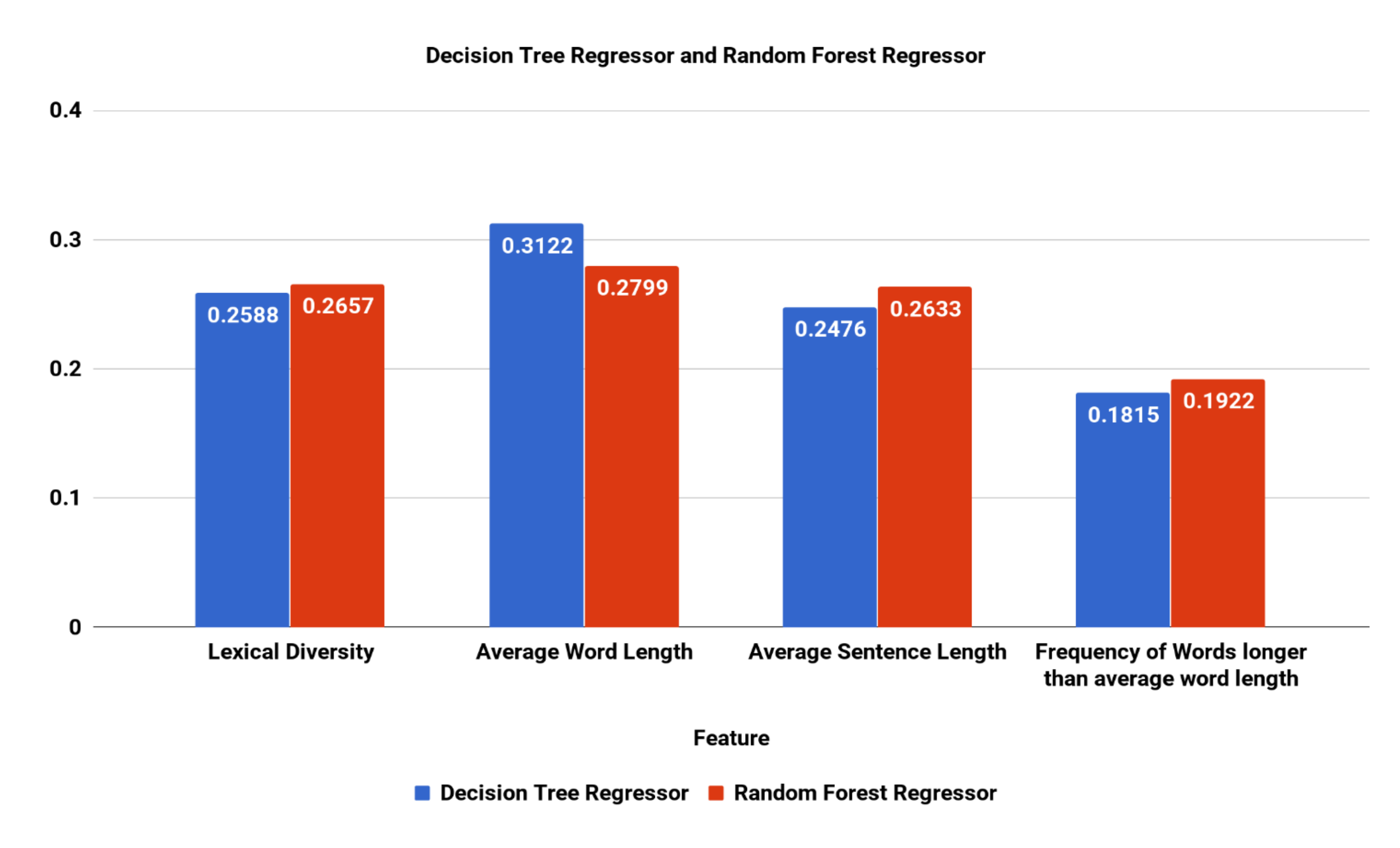
Public Understanding of Science

- Well understood research is more likely to shape public opinion.
- Public opinion influences important decisions.

H. V. Siravuri, A. P. Akella, C. Bailey, and H. Alhoori, “Using Social Media and Scholarly Text to Predict Public Understanding of Science”, in Proceedings of the 2018 ACM/IEEE JCDL.


Altmetrics Data
Feature Extraction
Selected Features
Pre-processing
Processed Dataset
Scientific Text
Semantic Cosine Similarity
Blog Posts
Textual Features
Combined Dataset
Similarity Score
70%
30%
Training Set
Training the Model
Regression Model
Prediction
Test Set

Feature Importance


Decision Tree Regressor and Random Forest Regressor
Lexical Diversity
Average Word Length
Average Sentence Length
Frequency of Words longer than average word length
Feature
Decision Tree Regressor
Random Forest Regressor