Crucial factors determining the popularity of scientific articles

ROBERT JANKOWSKI
JULIAN SIENKIEWICZ

Faculty of Physics, Warsaw University of Technology

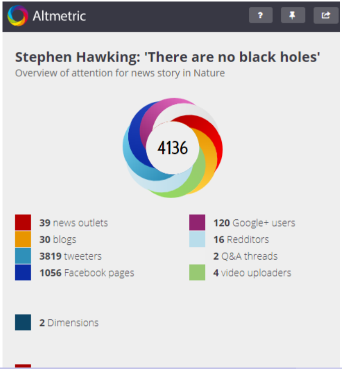
Świerk, 03-07-19

Goals

- Find critical factors which determine the popularity of scientific articles
- Calculate the popularity threshold of the articles

DIFFERENT POPULARITY METRICS

1. Google Scholar
2. PLoS ONE
3. Scopus
4. Web Of Science

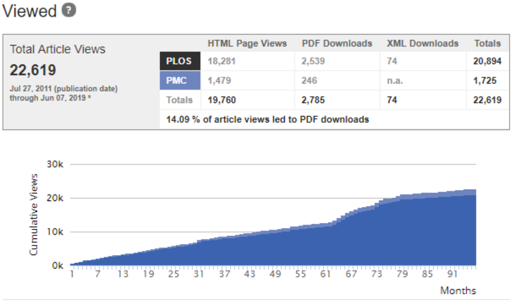


Stephen Hawking: 'There are no black holes'

Overview of attention for news story in Nature

🟥 **39** news outlets	🟪 **120** Google+ users
🟧 **30** blogs	🟦 **16** Redditors
🟦 **3819** tweeters	**2** Q&A threads
🟦 **1056** Facebook pages	🟩 **4** video uploaders

🟦 **2** Dimensions



Viewed ?

Total Article Views

22,619

Jul 27, 2011 (publication date)
through Jun 07, 2019 *

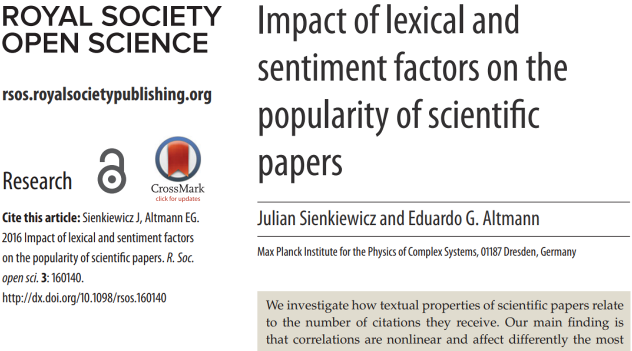
	HTML Page Views	PDF Downloads	XML Downloads	Totals
PLOS	18,281	2,539	74	20,894
PMC	1,479	246	n.a.	1,725
Totals	19,760	2,785	74	22,619

14.09 % of article views led to PDF downloads

Impact of lexical and sentiment factors on the popularity of scientific papers

Introduction

- over **4.3 million** papers, over **1500** different journals
- text length, text complexity, sentiment



ROYAL SOCIETY
OPEN SCIENCE

rsos.royalsocietypublishing.org

Research

CrossMark
click for updates

Cite this article: Sienkiewicz J, Altmann EG. 2016 Impact of lexical and sentiment factors on the popularity of scientific papers. *R. Soc. open sci.* **3**: 160140. http://dx.doi.org/10.1098/rsos.160140

Impact of lexical and sentiment factors on the popularity of scientific papers

Julian Sienkiewicz and Eduardo G. Altmann

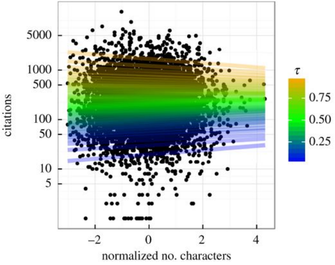
Max Planck Institute for the Physics of Complex Systems, 01187 Dresden, Germany

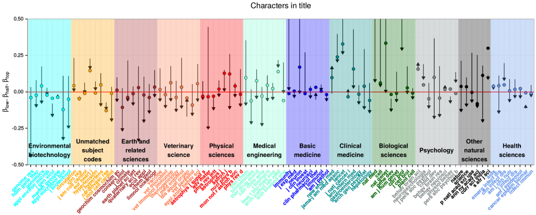
We investigate how textual properties of scientific papers relate to the number of citations they receive. Our main finding is that correlations are nonlinear and affect differently the most

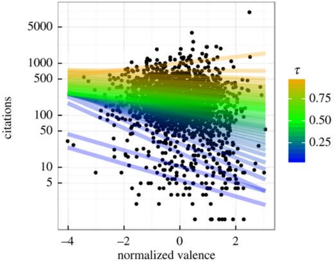
Results

Summary

- Correlations are **non-linear** and affect differently most-cited and typical papers
- In most journals short titles correlate positively with citations only for the most cited papers, for typical papers the correlation is in most cases negative
- Large variability across journals







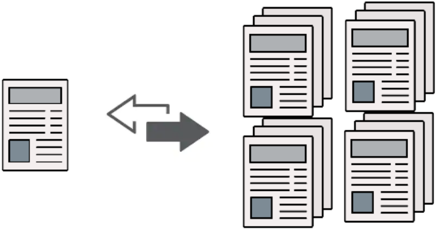
ANALYSIS FROM ANOTHER PERSPECTIVE

METHODS

- Classification vs Statistical analysis

DATA

- One journal vs 1500 different journals
- Number of views vs Number of citations



DATA

PLoS ONE service

FILTERING

- one part over 140.000
- second part over 80.000

OUTCOME

- over 70 000 papers from **2003** to **2014**
- information about the title, authors, full abstract contents and number of views per month
- mean of the **total** number of views

METRICS

LENGTH

- number of characters
- number of words
- number of sentences

Metrics

Length

- number of characters
- number of words
- number of sentences

Complexity

- Fog index: $F = \left(\frac{\#words}{\#sentences} + 100 \frac{\#complex\ words}{\#words} \right)$
- Herdan's C: $C = \frac{\log N}{\log M}$, M - text length, N - vocabulary size

Metrics

Length

- number of characters
- number of words
- number of sentences

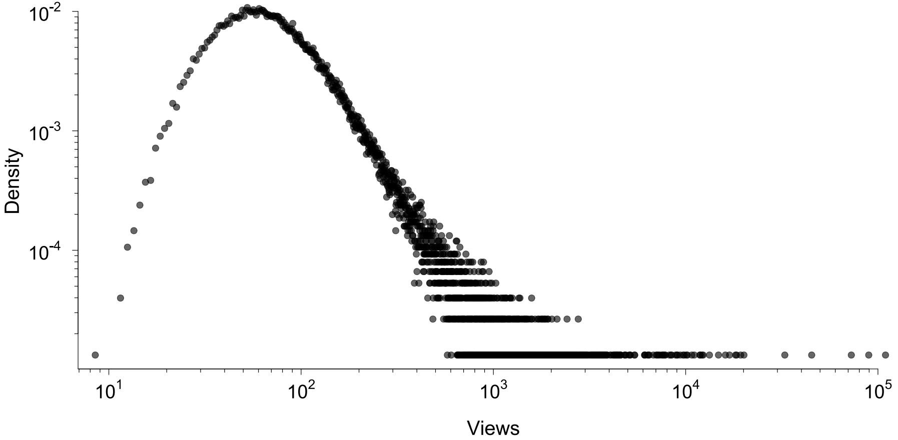
Complexity

- Fog index: $F = \left(\frac{\#words}{\#sentences} + 100\frac{\#complex\ words}{\#words}\right)$
- Herdan's C: $C = \frac{\log N}{\log M}$, M - text length, N - vocabulary size

Sentiment

- valence - emotional sign of the text
- arousal - level of emotional activation

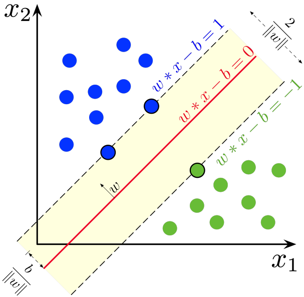
VIEWS DISTRIBUTION



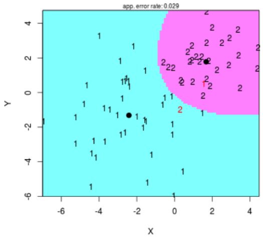
MODEL

CLASSIFICATION MODELS

- LDA, QDA
- SVM (Support-vector machine)
- Random forest



(a) SVM

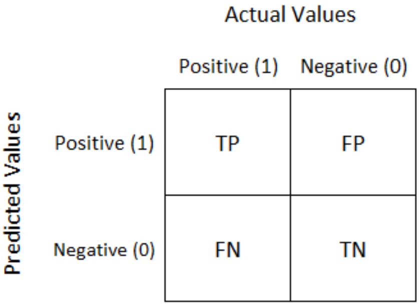


(b) QDA

Measure

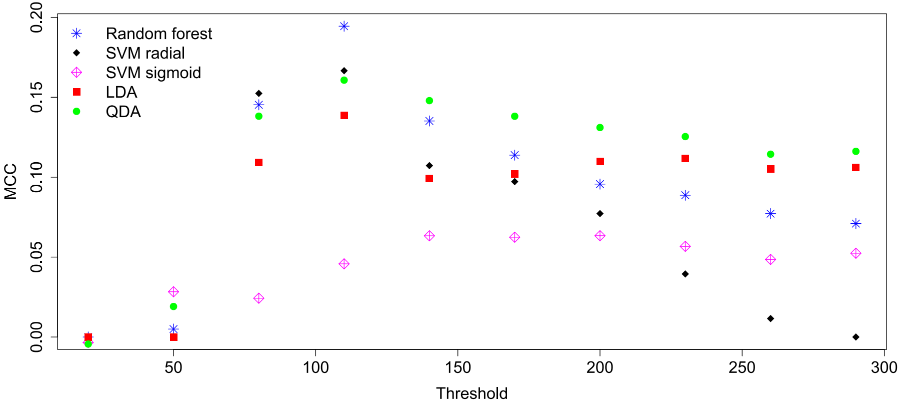
Metrics

1. F1 score ($F1 \in [0, 1]$)
2. Matthews correlation coefficient ($MCC \in [-1, 1]$)

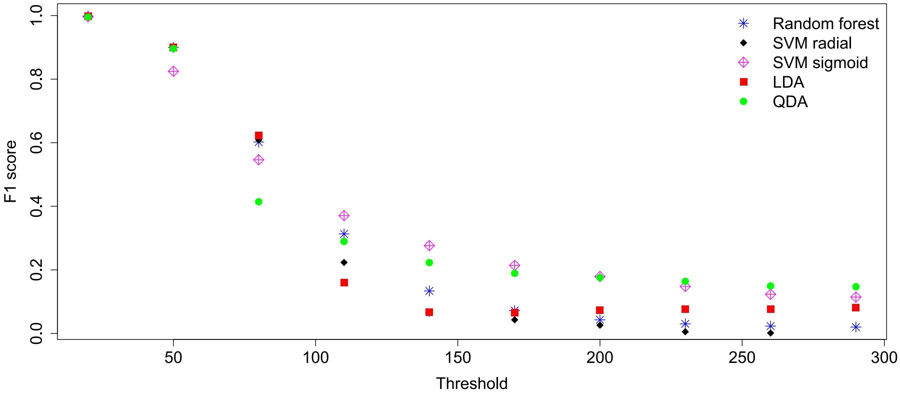


		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

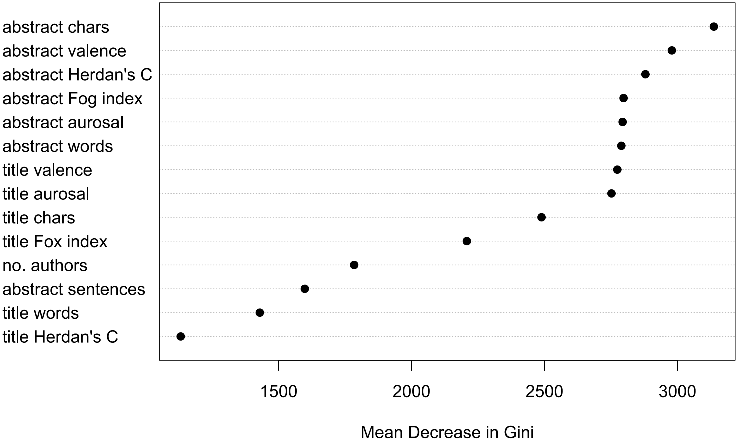
IMPLEMENTATION



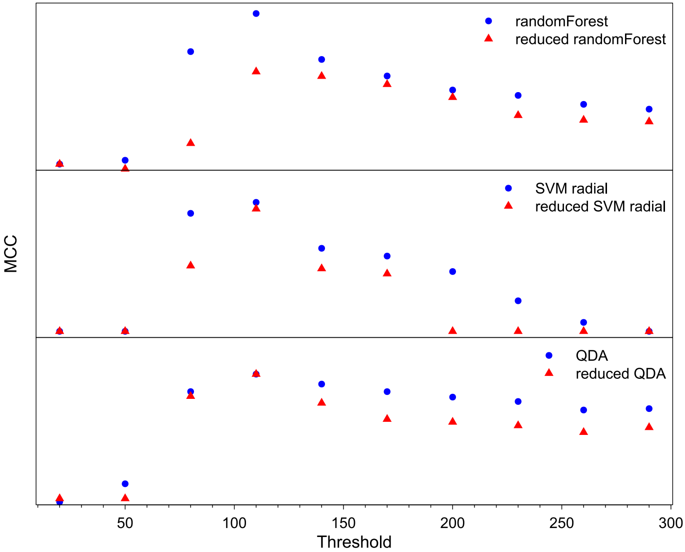
IMPLEMENTATION



REDUCED MODELS



REDUCED MODELS



RESULTS

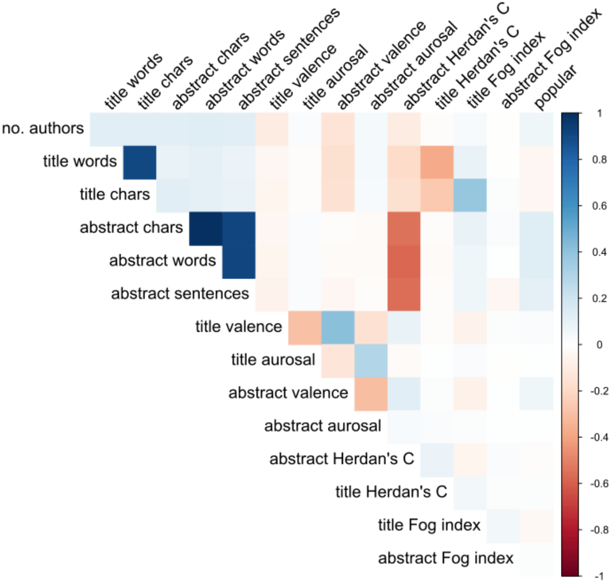
RESULTS

- the best popularity threshold for classification - **80-140** views
- number of **characters** and **valence** in abstract - critical factors
- inferior classification for the reduced number of features

FURTHER WORK

- sentiment **in each part** of the article full text (e.g introduction, discussion) and **around citation**
- a **yearly** number of views from date of publication

CORRELATION MATRIX


title words
title chars
abstract chars
abstract words
abstract sentences
title valence
title aurosal
abstract valence
abstract aurosal
abstract Herdan's C
title Herdan's C
title Fog index
abstract Fog index
popular
no. authors
title words
title chars
abstract chars
abstract words
abstract sentences
title valence
title aurosal
abstract valence
abstract aurosal
abstract Herdan's C
title Herdan's C
title Fog index
abstract Fog index