# Measuring the Novelty of Scientific Papers

Pavel Savov (pavel.savov@pja.edu.pl)

Supervisors:
Dr hab. Jerzy Paweł Nowacki, prof. PJAIT
Dr Radosław Nielek, prof. PJAIT

POLISH-JAPANESE ACADEMY OF INFORMATION TECHNOLOGY

# Pavel Savov

- M.Sc. in Informatics: University of Warsaw, 2007
- Ph.D. Candidate: Polish-Japanese Academy of Information Technology
  - Since 2021: Research and Teaching Assistant
- Software Engineer: 2004 - present
  - Since 2019: Big Data Engineer at Allegro
- Interests: NLP, Machine Learning, Knowledge Discovery

# Publications


- Savov, Pavel, and Radoslaw Nielek. "Ridiculously Expensive Watches and Surprisingly Many Reviewers: A Study of Irony." 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI). IEEE, 2016.
- Savov, Pavel, Adam Jatowt, and Radoslaw Nielek. "Towards understanding the evolution of the WWW conference." Proceedings of the 26th international conference on world wide web companion. 2017.
- Savov, Pavel, Adam Jatowt, and Radoslaw Nielek. "Identifying breakthrough scientific papers." Information Processing & Management 57.2 (2020): 102168.
- Savov, Pavel, Adam Jatowt, and Radoslaw Nielek. "Innovativeness analysis of scholarly publications by age prediction using ordinal regression." International Conference on Computational Science. Springer, Cham, 2020.
- Savov, Pavel, Adam Jatowt, and Radoslaw Nielek. "Predicting the Age of Scientific Papers." International Conference on Computational Science. Springer, Cham, 2021.

# Introduction

- As the number of papers published each year keeps growing, it is becoming increasingly difficult to follow all research, even in one's own area.
- Researchers and funding bodies rely on citation-based metrics for identifying promising and potentially breakthrough research
- Citation-based metrics are used for evaluating the output of researchers
- Can a supplementary approach to traditional scientometrics be offered for assessing the merit of publications without requiring expert knowledge and/or manual labor?

# Related Work: Identifying Breakthroughs

- Citation-based approaches:
  - J W Schneider and R Costas. Identifying potential “breakthrough” publications using refined citation analyses: Three related explorative approaches. Journal of the Association for Information Science and Technology, 68(3):709–723,2017
  - I V Ponomarev, D E Williams, C J Hackett, J D Schnell, and L L Haak. Predicting highly cited papers: A method for early detection of candidate breakthroughs. Technological Forecasting and Social Change, 81:49–55, 2014
  - JJ Winnink and R JW Tijssen. Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. Scientometrics, 102(1):113–134, 2015
- Classification approach:
  - H N Wolcott, M J Fouch, E R Hsu, L G DiJoseph, C A Bernaciak, J G Corrigan, and D E Williams. Modeling time-dependent and-independent indicators to facilitate identification of breakthrough research papers. Scientometrics, 107(2):807–817, 2016
- Analogy mining:
  - T Hope, J Chan, A Kittur, and D Shahaf. Accelerating innovation through analogy mining. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 235–243, 2017

# Related Work: Document Dating

- Temporal Language Models:
  - A Dalli and Y Wilks. Automatic dating of documents and temporal text classification. In Proceedings of the Workshop on Annotating and Reasoning about Time and Events, pages 17–22, 2006
  - N Kanhabua and K Nørvag. Using temporal language models for document dating. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 738–741. Springer, 2009
  - A Jatowt and R Campos.Interactive system for reasoning about document age. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, pages 2471–2474, New York, NY, USA, 2017
- Classification-based approaches:
  - H Salaberri, I Salaberri, O Arregi, and B Zapirain. Ixagroupehudiac: A multiple approach system towards the diachronic evaluation of texts. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 840–845, Denver, Colorado, 2015. Association for Computational Linguistics
  - V Niculae, M Zampieri, L P Dinu, and A M Ciobanu. Temporal text ranking and automatic dating of texts. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pages 17–21, 2014
  - O Popescu and C Strapparava. Semeval 2015, task 7: Diachronic text evaluation. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 870–878, 2015

# Related Work: Development of Research Fields

- Topic modelling:
  - D Hall, D Jurafsky, and C D Manning. Studying the History of Ideas Using Topic Models. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP ’08, pages 363–371, Stroudsburg, PA,USA, 2008. Association for Computational Linguistics
  - B Chen, S Tsutsui, Y Ding, and F Ma. Understanding the topic evolution in a scientific domain: An exploratory study for the field of information retrieval. Journal of Informetrics, 11(4):1175–1189, 2017
  - L Sun and Y Yin. Discovering themes and trends in transportation research using topic modeling. Transportation Research Part C: Emerging Technologies, 77:49–66, 2017
- Word embeddings:
  - K Hu, Q Luo, K Qi, S Yang, J Mao, X Fu, J Zheng, H Wu, Y Guo, and Q Zhu. Understanding the topic evolution of scientific literatures like an evolving city: Using google word2vec model and spatial autocorrelation analysis. Information Processing & Management, 56(4):1185–1203,2019

# Problems with Citation Analysis

- Citing prominent publications, following the crowd
- “The rich get richer and the poor get poorer”
- Google Scholar Effect
- Attention stealing
- Ignoring the purpose of citations (support vs criticism)
- Slowness: It may take several years to acquire the first citations

# Main Contributions

- Method for dating scientific papers in a given domain using topic models and ordinal regression, based solely on textual content
- Paper Innovation Score – a real-number measure of scientific paper novelty based on age prediction errors
- Improved method of dating scientific papers using state-of-the-art word embeddings and ordinal regression

# Method Outline

- Given a diachronic corpus of scientific papers in a specific domain:
  - Train topic model
  - Using topics as features train ordinal regression model for publication year prediction
- For each analyzed paper:
  - Predict publication year
  - Based on the prediction error calculate Paper Innovation Score
  - The more positive the error is, the more the paper resembles those published in the future, and the greater its innovation score
  - Adjust the Paper Innovation Score for the publication year, since minimum and maximum prediction errors decrease as the publication year increases

# Topic Model Training and Selection

- Select a range [$k_{min}$,$k_{max}$] dependent on size of corpus and domain
- Train *k*-topic Correlated Topic Models (CTM) for each *k* in [$k_{min}$,$k_{max}$]
- Calculate $C_V$ Topic Coherence for each model
- Select the model with the highest $C_V$ value

# Topic Coherence

- Parameters: W - sliding window size, *n* - number of top words
- "Context vectors" $v_i$ for each word $w_i$ in the top *n* words for each topic:

$$v_{ij} = \frac{log\frac{P(w_i,w_j)+\epsilon}{P(w_i)\cdot P(w_j)}}{-log(P(w_i,w_j)+\epsilon)}$$

where $P(w_i)$ - observed occurrence probability of $w_i$, $P(w_i, w_j)$ - observed probability that $w_i$ and $w_j$ co-occur within a sliding window of size W

- Calculate cosine similarity for each pair of context vectors *u* and *v*:

$$s_{cos}(\vec{u},\vec{v}) = \frac{\sum_{i=1}^{|W|} u_i \cdot v_i}{||\vec{u}||_2 \cdot ||\vec{v}||_2}$$

- Calculate topic coherence:

$$C_V = \mu(\{s_{cos}(\vec{u},\vec{w}) \mid \forall(\vec{u},\vec{w})\})$$

# Why Topic Coherence?

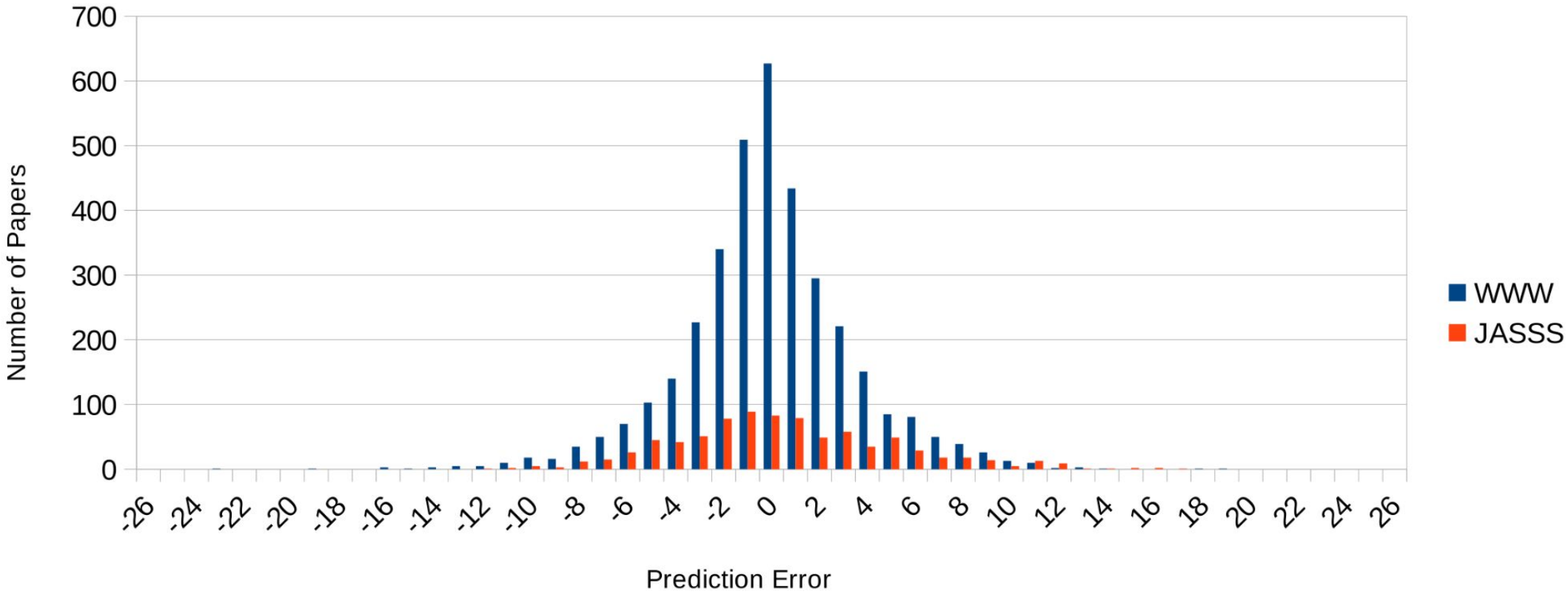
- $C_V$ has been shown to reflect topic "interpretability" by humans
- Traditional likelihood or perplexity-based approaches have been shown to result in topics which are more difficult to understand
- Paper Innovation Scores calculated using different topic models have been shown to be strongly correlated.
  Mean pairwise Spearman's $\rho$ on test corpora:
  - 0.75 (std. deviation: 0.04)
  - 0.65 (std. deviation: 0.05)

# Publication Year Prediction

- Latent topic probabilities as features
- Ordinal Regression model for predicting publication years:
  - For each pair of consecutive years in the corpus train a *before-after* binary classifier
  - Given predicted class membership probabilities calculate overall model confidence that paper *p* was published in year *Y*:

$$conf(p, Y) = \prod_{y=Y_{min}}^{Y} P(Y_p \leq y) \cdot \prod_{y=Y+1}^{Y_{max}} (1 - P(Y_p \leq y))$$

# Prediction Error Distribution



# Paper Innovation Score

$$S_P(p) = \frac{\sum_y conf(p,y) \cdot (y - Y_p)}{\sum_y conf(p,y)}$$

$Y_p$ - publication year of paper *p*

- Problem: Minimum and maximum prediction errors decrease as the publication year increases and so does the mean unadjusted score ($S_P$)

# Adjusting Innovation Scores for Publication Year

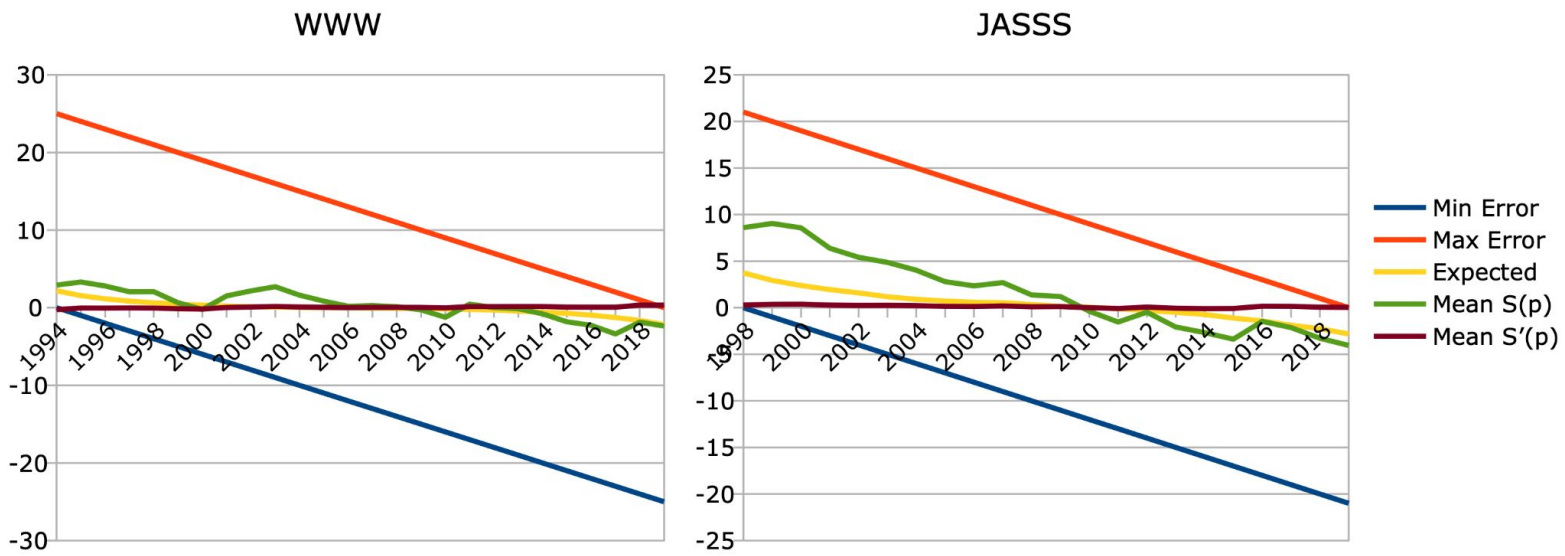
- Let us suppose the prediction error for papers published in year *Y* is a discrete random variable $Err_Y$.
- Based on the actual prediction error distributions, let us define the expected prediction error for papers from year *Y*:

$$E(Err_Y) = \sum_{n=Y_{min}-Y}^{Y_{max}-Y} n \cdot Pr(Err_Y = n)$$

- Probabilities $Pr(Err_Y)$ are calculated using prediction error distribution truncated to the range $<Y_{min} - Y, Y_{max} - Y>$ (minimum and maximum prediction errors for year *Y*)

# Innovation Score Adjusted for Publication Year

$$S'_P(p) = \begin{cases} \dfrac{S_P(p) - E(Err_{Y_p})}{E(Err_{Y_p}) - (Y_{min} - Y_p)} & \text{if } S_P(p) < E(Err_{Y_p}) \\ \dfrac{S_P(p) - E(Err_{Y_p})}{Y_{max} - Y_p - E(Err_{Y_p})} & \text{if } S_P(p) \geq E(Err_{Y_p}) \end{cases}$$



# Datasets

- WWW -- The Web Conference
  - 3,577 papers published between 1994 and 2019
- JASSS -- Journal of Artificial Societies and Social Simulation
  - 835 articles published between 1998 and 2019
- Yearly time slices

# Publication Year Prediction Results

- Mean Absolute Error:
  - WWW: 2.56 years
  - JASSS: 3.56 years

# Correlation Between Innovation Scores and Citation Counts

- Spearman's $\rho$:
  - WWW: 0.28, p-value: $1.21 \cdot 10^{-41}$
  - JASSS: 0.32, p-value: $1.91 \cdot 10^{-6}$
- For papers at least 5 years old:
  - WWW: 0.3
  - JASSS: 0.37

# Top 3 Papers: WWW

| Year | Author(s) and Title | Score | Citations |
|---|---|---|---|
| 2011 | C. Budak, D. Agrawal, A. El Abbadi, *Limiting the Spread of Misinformation in Social Networks* | 0.971 | 607 |
| 2010 | A. Sala, L. Cao, Ch. Wilson, R. Zablit, H. Zheng, B. Y. Zhao, *Measurement-calibrated Graph Models for Social Network Experiments* | 0.963 | 189 |
| 2018 | H. Wu, Ch. Wang, J. Yin, K. Lu, L. Zhu, *Sharing Deep Neural Network Models with Interpretation* | 0.955 | 7 |

# Top 3 Papers: JASSS

| Year | Author(s) and Title | Score | Citations |
|---|---|---|---|
| 2001 | K. Auer, T. Norris, "*ArrierosAlife*" *a Multi-Agent Approach Simulating the Evolution of a Social System: Modeling the Emergence of Social Networks with "Ascape"* | 0.868 | 13 |
| 2000 | B. G. Lawson, S. Park, *Asynchronous Time Evolution in an Artificial Society Model* | 0.841 | 13 |
| 2008 | R. Bhavnani, D. Miodownik, J. Nart, *REsCape: an Agent-Based Framework for Modeling Resources, Ethnicity, and Conflict* | 0.788 | 51 |

# Predicting Publication Years Using BERT

- Predict the age of each sentence
  - Ordinal regression model at sentence level
  - SciBERT models fine-tuned for sequence classification
  - BERT models trained on scientific publications
- (Optional) Remove irrelevant sentences
  - Containing citations
  - Entire *Related Work* section
- Aggregate results for sentences to determine document age

# Datasets

- WWW -- The Web Conference
  - 3,896 papers published between 1994 and 2020
  - 1M sentences
- JASSS -- Journal of Artificial Societies and Social Simulation
  - 884 articles published between 1998 and 2020
  - 321k sentences
- Yearly time slices

# Result Aggregation

- Newest Sentence
- Arithmetic Mean
- Weighted Mean w/Sentence Offset
- Weighted Mean w/Sentence Importance (TextRank)

# Results: WWW

| | Mean Absolute Error (Years) | |
|---|---|---|
| Document-level | 2.56 | |
| Sentence-level | All Sentences | No Citations |
| Newest Sentence | 8.959 | 8.946 |
| Arithmetic Mean | 0.833 | 0.816 |
| Weighted Mean w/Sentence Offset | 0.709 | **0.684** |
| Weighted Mean w/TextRank | 0.741 | 0.725 |

# Results: JASSS

| | Mean Absolute Error (Years) | |
|---|---|---|
| Document-level | 3.56 | |
| Sentence-level | All Sentences | No Citations |
| Newest Sentence | 8.267 | 8.33 |
| Arithmetic Mean | 0.743 | 0.67 |
| Weighted Mean w/Sentence Offset | 0.738 | 0.645 |
| Weighted Mean w/TextRank | 0.67 | **0.636** |

# Conclusions

- Novel method of analyzing corpora of publications from multiple year periods has been proposed
- None of its steps require expert knowledge or manual intervention
- The version with topic models is explainable - publication year prediction error depends on topic popularity over time

# Weaknesses

- Only a snapshot in time is captured
- As new papers in the studied domain are published and new time slices are added, the topic model and prediction model need to be retrained
- The latent topics discovered in the updated corpus may change
  - Not a problem if BERT is used
- Sensitivity to shifts in the scope of the analyzed publication venues
  - Train the models on papers from multiple venues in a given domain
  - Previously researched topics occurring in a new context may indicate innovation

# Weaknesses

- Innovation Score adjusted for publication year by linear function
- The observed prediction error distribution is non-uniform, small deviations from the expected value are more likely than large ones, and therefore - less significant
- Computationally expensive

# Questions?