# Predicting the Age of Scientific Papers

Pavel Savov[1], Adam Jatowt[2], and Radoslaw Nielek[1]

[1] Polish-Japanese Academy of Information Technology, ul. Koszykowa 86,
02-008 Warszawa, Poland
{pavel.savov,nielek}@pja.edu.pl
[2] University of Innsbruck, Innrain 52, 6020 Innsbruck, Austria
adam.jatowt@uibk.ac.at

**Abstract.** In this paper we show how the age of scientific papers can be predicted given a diachronic corpus of papers from a particular domain published over a certain time period. We first train ordinal regression models for the task of predicting the age of individual sentences by fine-tuning series of BERT models for binary classification. We then aggregate the prediction results on individual sentences into a final result for entire papers. Using two corpora of publications from the International World Wide Web Conference and the Journal of Artificial Societies and Social Simulation, we compare various result aggregation methods, and show that the sentence-based approach produces better results than the direct document-level method.

**Keywords:** Scientometrics · Text Age Prediction · Embeddings · Ordinal Regression

## 1 Introduction

Document dating or timestamping is the process of inferring the age of a document, if it is either unknown or unreliable, based on its textual content. In the scientific domain, publication dates of documents are usually known, but the results of document timestamping may be used to complement traditional scientometric methods in assessing the innovativeness of research papers [20] or identifying novelty. At a basic level, the larger the difference between the actual timestamp and the predicted timestamp of a target scientific document, the higher is its potential innovativeness or novelty of the target paper. This may be useful to non-expert readers of technical documents, such as potential investors or decision makers at funding bodies, who wish to know how new or innovative the ideas or methods covered by these documents were at the time of their creation. Furthermore, in practical scenarios, the timestamping models specialized for scientific corpora can also be applied to other types of documents that may discuss scientific technology and domain-focused research, or quote content from scientific papers. Such documents may not have explicit timestamps (e.g., web pages) and the determination of their age (as well as the related concept of timeliness) can be useful in many cases. Thus, in general, scientific document age prediction can be used for discovering the content parts in a scientific

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_58

publication that are novel or innovative, or perhaps obsolete/outdated when considering the document publication date [20] as well as for determining the age of science-related content in non-scholarly documents that lack timestamps.

In this paper we focus on improving the accuracy of scientific paper age prediction by using state-of-the-art word embedding models trained on two corpora of papers from related but distinct domains, published at leading publication venues in their respective fields. Typical approaches to automatic document dating are based on modeling language change over time and shifts in word usage. Examples of temporal language models, i.e. time series of statistical language models include [5, 9]. Jatowt and Campos [8] have implemented an online visual and interactive system based on $n$-gram frequency analysis. Garcia-Fernandez et al. [7] used SVM classifiers on feature vectors of word and n-gram frequencies. Ordinal regression models were used for document dating by Niculae et al. [16], or Popescu and Strapparava [18]. Another approach to temporal language modeling are neural language models based on word embeddings such as Word2Vec [15]. Kim et al. [10] studied the shift in word semantics over time by training a model for each time interval and then plotting the words' cosine similarities to their reference points. Soni et al. [21] used diachronic word embeddings to show that scientific papers using words in their newer meanings tend to receive more citations. Vashishth et al. [23] proposed a deep learning approach to document dating, exploiting syntactic and temporal document graph structures. Unlike the above-mentioned methods, which work mainly on news articles or generic documents, we focus on a particular genre of scholarly publications. We also approach the document dating task at a sentence-level, and we test several sentence aggregation approaches.

## 2 Datasets

We study the following two corpora: (1) *WWW*: 3,896 papers published at the International World Wide Web Conference between 1994 and 2020, containing 1,037,051 sentences, (2) *JASSS*: 884 articles published in the Journal of Artificial Societies and Social Simulation[3] between 1998 and 2020, containing 321,589 sentences. Both corpora contain entire papers. However, we have removed page headers and footers, *References*, *Bibliography* and *Acknowledgments* sections as "noise" irrelevant to the papers' contents. All papers published in the JASSS journal are available in HTML at the journal's website[3]. Papers from the proceedings of the WWW conference are available at `https://thewebconf.org/` in different formats for different years. Most are available in PDF, some in HTML and a small number of older papers in PostScript. We used the *pdftotext* tool[4] to extract plain text from PDF documents. We divided the documents into sentences using the Punkt sentence tokenizer for the English language implemented in the Natural Language Toolkit (NLTK) Python library [4]. Conversion to lower case and tokenization were performed by the BERT tokenizer.

[3] http://jasss.soc.surrey.ac.uk/

[4] https://www.xpdfreader.com/pdftotext-man.html

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_58

## 3 Method

We propose to approach the problem of scientific document's age prediction by first predicting the age of its sentences. Thanks to focusing on sentences instead of entire documents we can use more labelled data instances for training, which is quite important for relatively narrow scientific domains with constrained datasets (e.g., proceedings of conferences dedicated to a particular research subfield). Thus, our approach is composed of two steps: (1) predicting the age of sentences and (2) aggregating sentence age to determine the document age. We describe these two steps below.

### 3.1 Predicting Sentence Age

As time units are clearly ordinal values, we predict the age of individual sentences by means of Ordinal Regression, a.k.a. Ordinal Classification, based on the framework proposed by Li and Lin [11]. Ordinal Regression was also used by Martin et al. [13] for photograph dating. An $N$-class ordinal regression model consists of $N-1$ *before-after* binary classifiers, i.e. for each pair of consecutive years a classifier is trained, which assigns sentences to one of two classes: "year $y$ or before" and "year $y+1$ or after". Given the class membership probabilities predicted by these classifiers, the overall classifier confidence that sentence $s$ was written in the year $Y$ is then determined, as in [13], by Eqs. 1 and 2:

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

where $Y_{min}$ and $Y_{max}$ are the first and last year in the corpus, and $Y_s$ is the publication year of the paper that $s$ comes from.

Thus, the predicted year for the sentence $s$ is:

$$\hat{Y_s} = \underset{y \in [Y_{min}, Y_{max}]}{\operatorname{argmax}} \; conf(s,y) \tag{2}$$

Unlike the approaches of [11] and [13], we used the Huggingface Transformers[5] [24] Python library to fine-tune SciBERT models [3] for sequence classification in binary *before-after* classification. SciBERT is a BERT [6] model trained on 1.14M scientific papers from the `semanticscholar.org` corpus. The maximum sequence length supported out-of-the-box is 512, however over 95% of the sentences in our corpora contain up to 64 tokens (see Fig. 1). We have, therefore, decided to cap the maximum sequence length at 64. We have not observed any significant differences in the predictive performance of the models, expressed as Mean Absolute Error, for maximum sequence lengths of 64, 128, and 512 tokens. We trained each model for two epochs, the batch size was 32, and the learning rate: 2e-5. The BERT authors recommend fine-tuning the models for 2 to 4 epochs, but we have found our models to overfit the training data when fine-tuned for more than 2 epochs. In most cases the differences in average loss and accuracy on the validation set for models trained for two epochs vs. one were minimal.

[5] `https://huggingface.co/transformers/`

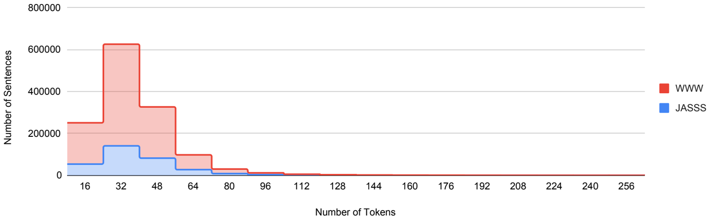

Fig. 1: Number of Tokens per Sentence

We have made an 80/20 split on the document level so as to make sure of the clean separation of training and testing sentences. Although our approach yielded poor prediction results on the sentence-level (4.49 years for JASSS and 3.56 years for WWW, see Fig. 2), as we will show later, the final prediction of document age produces quite good results.

### 3.2 Predicting Document Age

As stated above, we predict the age of entire papers by aggregating the results of individual sentence age prediction using various aggregation functions. We have experimented with rejecting sentences for which the model's confidence was below a certain threshold in the range from 0 to 0.5. For values greater than 0.5 in some documents no sentences exceeded that threshold.

**Newest Sentence** As a baseline approach we assume the age of the paper $p$ equals the age of its newest sentence. Since most papers contain at least one sentence the most probable age of which is predicted as 0 years, we only take into account the sentence predicted as the newest among those, for which the model's confidence exceeds 0.5. This value was chosen, as it gave the best results.

**Topic distribution based classifier** As another baseline approach, which works purely on the document-level, we used a method based on SVM classifier on vectors of latent topic distributions derived from document collections [20].

**Arithmetic Mean** In this approach we calculated the predicted age of paper $p$ as the mean predicted age of all its sentences.

**Weighted Mean w/Sentence Offset** We assumed that the sooner a sentence appears in the paper, the more important it is. We, therefore, defined the predicted age of paper $p$ as the weighted mean predicted age of its sentences, where the weight of each sentence was its ordinal number within the paper $p$ divided by the number of sentences in $p$:

$$\hat{Y}_p = \frac{\sum_{s \in p} \hat{Y}_s \cdot \frac{n_s}{|\{s \in p\}|}}{\sum_{s \in p} \frac{n_s}{|\{s \in p\}|}}$$

where $n_s$ is the ordinal number of the sentence $s$ within $p$.

This concept is a simplified approach to weighted zoning [12], where each sentence is assigned a weight, depending on which section of the paper it appears in, e.g. Abstract: 1, Introduction: 0.8, Related Work: 0.3, everything else: 0.5.

**Weighted Mean w/TextRank** TextRank by Mihalcea and Tarau [14] is an unsupervised graph-based algorithm for keyword extraction and text summarization, based on PageRank [17]. Its variant for text summarization finds the most important sentences by running a variation of PageRank on a graph, whose vertices represent the document's sentences. Each edge has a weight corresponding to the similarity of the sentences represented by the vertices connected by that edge. In contrast to PageRank, the graph constructed by TextRank is undirected, since the similarity between sentences is symmetric. Various sentence similarity measures may be used, but Barrios et al. [2] showed that a variation of the Okapi-BM25 [19] ranking function, which is itself a variation of the TF-IDF model using a probabilistic model, yields the best results. We used the implementation of TextRank with the BM25 ranking function from the *gensim*[6] Python library to find importance scores for all sentences in each document. We then used these scores as weights to calculate the predicted publication year of each paper $p$ defined as the weighted mean of the years of its sentences:

$$\hat{Y}_p = \frac{\sum_{s \in p} Imp_p^s \cdot \hat{Y}_s}{\sum_{s \in p} Imp_p^s}$$

where $Imp_p^s$ is the TextRank importance score of $s$ within the paper $p$.

**Citation Removal** In this approach we make the assumption that any sentences citing other papers are unimportant for the content of the paper being analyzed or introduce concepts and ideas from older papers (hence potentially negatively impacting the age detection process). Thus, we remove all sentences containing citations and proceed to calculate the predicted publication year using any of the approaches described above. As shown in Section 4, in most cases citation removal improves the prediction results in terms of Mean Absolute Error. Another possible extension could be removing entire *Related Work* sections.

## 4 Results

As stated before, the mean absolute age prediction error (MAE) for individual sentences is 4.49 years for the JASSS corpus and 3.56 for WWW. The prediction error distribution is shown in Fig. 2. Although these results are not satisfactory, we obtain much better results for entire documents. As shown in Tab. 1, the sentence-based approach aggregating individual predictions of many sentences gives much better results in predicting paper publication dates. Except for the naive *newest sentence* baseline, the MAE is always less than 1 year. Also the document level approach proposed in [20] performs much worse.

Weighting the sentence age predictions by sentence offsets performed better on the WWW corpus, while TextRank weights gave better results for JASSS. In all cases, however, removing sentences containing citations improved the document age predictions significantly. This supports our assumption that sentences citing other articles could introduce noise.

[6] `https://radimrehurek.com/gensim_3.8.3/index.html`

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_58

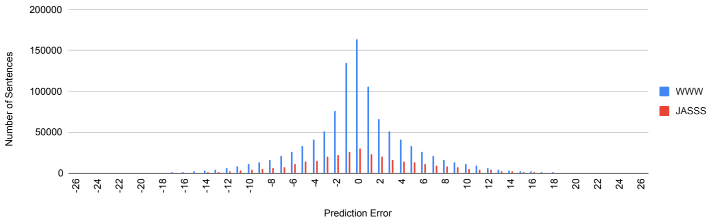

Fig. 2: Sentence Prediction Error Distributions

Table 1: Results of prediction methods (Mean Absolute Error: #years).

| | WWW | | JASSS | |
|---|---|---|---|---|
| Document-level [20] | 2.56 | | 3.56 | |
| Sentence-level | All Sentences | Citations Removed | All Sentences | Citations Removed |
| Newest Sentence | 8.959 | 8.946 | 8.267 | 8.33 |
| Arithmetic Mean | 0.833 | 0.816 | 0.743 | 0.67 |
| Weighted Mean w/Sentence Offset | 0.709 | **0.684** | 0.738 | 0.645 |
| Weighted Mean w/TextRank | 0.741 | 0.725 | 0.67 | **0.636** |

## 5 Conclusions and Future Work

In this paper we have shown how the accuracy of scientific paper age prediction can be improved by using state-of-the-art word embedding models at the sentence level, and then aggregating the results. Interestingly, for all aggregation methods except for the most basic baseline approach, i.e. *newest sentence*, increasing the value of the confidence threshold led to worse results. This suggests that unless sentences are rejected based on domain-specific knowledge, e.g. rejecting sentences containing citations, the more predictions are aggregated into the final result the better, similarly to the "wisdom of the crowds" effect, where the aggregated predictions of multiple agents are far closer to the actual value than most of the individual predictions [22]. Finally, we note that as our approach works on the sentence-level, it could also be used to assess the age of text excerpts (e.g., in web pages) about specialized scientific topics, and, therefore, potentially help readers better understand their actual novelty and age.

Having achieved a mean prediction error of less than a year, we plan on experimenting with datasets having narrower time slices, e.g. the Covid-19 dataset from Kaggle[7]. We will also try weighting sentences containing scientific claims [1].

## References

1. Achakulvisut, T., Bhagavatula, C., Acuna, D., Kording, K.: Claim extraction in biomedical publications using deep discourse model and transfer learning. arXiv preprint arXiv:1907.00962 (2019)
2. Barrios, F., López, F., Argerich, L., Wachenchauzer, R.: Variations of the similarity function of textrank for automated summarization. arXiv preprint arXiv:1602.03606 (2016)

[7] `https://www.kaggle.com/imdevskp/corona-virus-report`

3. Beltagy, I., Lo, K., Cohan, A.: Scibert: Pretrained language model for scientific text. In: EMNLP (2019)
4. Bird, S.: Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. pp. 69–72 (2006)
5. De Jong, F., Rode, H., Hiemstra, D.: Temporal language models for the disclosure of historical text. In: International Conference of the Association for History and Computing (AHC 2005). pp. 161–168. Koninklijke Nederlandse Academie van Wetenschappen, Amsterdam, the Netherlands (2005)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
7. Garcia-Fernandez, A., Ligozat, A.L., Dinarelli, M., Bernhard, D.: When was it written? automatically determining publication dates. In: SPIRE. pp. 221–236 (2011)
8. Jatowt, A., Campos, R.: Interactive system for reasoning about document age. In: CIKM'17. pp. 2471–2474. ACM
9. Kanhabua, N., Nørvåg, K.: Using temporal language models for document dating. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 738–741. Springer (2009)
10. Kim, Y., Chiu, Y.I., Hanaki, K., Hegde, D., Petrov, S.: Temporal analysis of language through neural language models. arXiv preprint arXiv:1405.3515 (2014)
11. Li, L., Lin, H.T.: Ordinal regression by extended binary classification. In: Advances in neural information processing systems. pp. 865–872 (2007)
12. Manning, C.D., Raghavan, P., Schütze, H.: Scoring, term weighting and the vector space model. Introduction to information retrieval **100**, 2–4 (2008)
13. Martin, P., Doucet, A., Jurie, F.: Dating color images with ordinal classification. In: ICMR. pp. 447–450 (2014)
14. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: EMNLP. pp. 404–411 (2004)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Niculae, V., Zampieri, M., Dinu, L.P., Ciobanu, A.M.: Temporal text ranking and automatic dating of texts. In: EACL. pp. 17–21 (2014)
17. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
18. Popescu, O., Strapparava, C.: Semeval 2015, task 7: Diachronic text evaluation. In: Proceedings of SemEval 2015. pp. 870–878 (2015)
19. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M., et al.: Okapi at trec-3. Nist Special Publication Sp **109**, 109 (1995)
20. Savov, P., Jatowt, A., Nielek, R.: Innovativeness analysis of scholarly publications by age prediction using ordinal regression. In: ICCS. pp. 646–660. Springer (2020)
21. Soni, S., Lerman, K., Eisenstein, J.: Follow the leader: Documents on the leading edge of semantic change get more citations. JASIST (2020)
22. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
23. Vashishth, S., Dasgupta, S.S., Ray, S.N., Talukdar, P.: Dating documents using graph convolution networks. In: Proceedings of ACL. pp. 1605–1615 (2018)
24. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: EMNLP: System Demonstrations. pp. 38–45 (2020)