Building a corpus of patents-articles siblings

Comparing scientific publication and patent abstracts using contextual text embeddings

Jean-Marc Deltorn[1,2], Domnique Guellec[3], Chenyin Wu[1], Jianying Liu[3]

(1) Center for International Intellectual Property Studies (CEIPI) - University of Strasbourg, France
(2) Bureau of Theoretical and Applied Economics (BETA) - University of Strasbourg, France
(3) Science and Technology Observatory (OST) (OST), Paris, France

Innovation Information Initiative (I3) Technical Working Group Meeting, Cambridge, MA, 2-3 Dec. 2022

Scientific objectives

- **How does basic research links to inventions and technological trends ?**
  - Basic research can produce information that is directly useful to technology
  - Basic research gives a more general framework for understanding certain technical effects, a framework that can be used to identify promising directions of technological research

→ Research question : "***What observable characteristics of scientific discoveries make them amenable or not to subsequent technological developments?***"

Prerequisite: **a mapping between basic science and technological development**...

How to relate basic science and technological development ?

Hypothesis:
- Scientific publications as proxies to Basic research
- Patent applications as proxies to technological development

Relating patents & scientific publications (1): citations

Citations as often used as a proxy of relatedness between scientific publications and patents

Benefits: A direct link from patents to scientific publications (much) more rarely from scientific publications to patents. Patents can cite "non-patent literature" (NPL) references (e.g. scientific publications).

- front page citations — USPTO: prior art against which the patent itself was defined as novel and non-obvious. (≠ @EPO)
- in-text citations (may include further citations from the front page)

Limitations of patent → NPL citations:

- "*front-page citations may be overgenerous as applicants attempt to impress examiners with a long list of prior art against which the present invention is (supposedly) distinct*" (Marx & Fuegi, 2020)
- citations depend on the strategies of the patent's applicant (may omit essential references or include others that do not reflect directly the sources of a technical contribution)
- citations do not always reflect a close technical proximity (e.g.: may relate to a general context, "*didactic & illustrative*", cf. Meyer, 2000: only "*a third of all patent citations have a close proximity to the citing patent*")

see: Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. Scientometrics, 49(1), 93–123; Marx, M. and Fuegi, A., 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. Strategic Management Journal, 41(9), pp.1572-1594.

→ Relying solely on a citation analysis to draw a picture of the relation between research developments and technological applications therefore leads to a partial — and largely biased — picture.

Relating patents & scientific publications (2): text

What about text-to-text relatedness?

Benefits: rely solely no the description by author(s) and applicant.

- Independent claims should recite the essential features of the invention (the abstract is (in practice) in close proximity to the original main independent claim)
- Scientific publications abstract should provide a technical description of the underlying content

Limitations

textual description may be incomplete
language/form
functions
public/audience
drafters
and
relating the semantics of two words/texts is an open problem: text is "brittle"...

Idea → Patent

- abstract = original claim (not final "invention")
- no legal constraint (possible obfuscation)
- motivation = "make it easy for the patent attorney!"
- drafted by attorney (not scientist/inventor)

Idea → Scientific publication

- abstract = "make it interesting so that people/editors/peers keep reading!"
- focus on (broad) context, key technical features and main results
- drafted by scientist(s)

Relating patents & scientific publications (2): text

What about text-to-text relatedness?

Benefits: rely solely no the description by author(s) and applicant.

- Independent claims should recite the essential features of the invention (the abstract is (in practice) in close proximity to the original main independent claim)
- Scientific publications abstract should provide a technical description of the underlying content

Limitations

textual description may be incomplete
language/form
functions
public/audience
drafters
and
relating the semantics of two words/texts is an open problem: text is "brittle"...

How to test the hypothesis that text representations can capture the underlying proximity between two documents ?

Questions:

- Can these two modes of expression be meaningfully compared?
- Do they express similarly the underlying technical content?
- Can we quantify the "dissimilarity" between the abstracts from patent and in scientific publications?

Patent paper pairs / "siblings"

→ **A possibility**: to identify a set of documents that express (essentially) the same technical content in the form of patent applications and scientific publications : "**patent publication pairs" (PPP) or patent publications "siblings"**

The majority of works have relied on **manual examination or basic text-mining techniques to collect PPPs** (cf. Coward and Franklin, 1989; Ducor, 2000; Lissoni and Montobbio, 2008; Lissoni et al., 2013; Murray and Stern, 2007).

Main uses:

- **Studying the relationship between authorship and inventorship**:

  - Noyons et al. (1994) analyzed papers and patents related to medical applications of x-rays produced by public research institutions in Europe, showing an increasing trend in the number of papers coauthored by researchers in industry and research institutions.
  - Ducor (2000) compared the numbers of authors in articles and inventors in patents in a set of 40 PPPs in the field of molecular biology.
  - Lissoni and Montobbio (2008) and Lissoni et al. (2013) showed that the number of coauthors was higher than that of coinventors, and that authors who are also inventors tend to be the first or last authors (focusing on four technical fields in Italy between 1975 and 2002)

PPPs elsewhere

- **Studying the flow o f information and the influence between scientific publications and patents**

  - Murray (2002) Innovation as co-evolution of scientific and technological networks: Exploring tissue engineering
  - Murray and Stern (2004) compared the influence of papers between PPPs and non-PPPs in biotechnology.
  - Boyack, K. W., & Klavans, R. (2008). Measuring science–technology interaction using rare inventor–author names. Journal of Informetrics, 2, 173–182.
  - Chang Y-W. et al. (2017) studied the interaction between science and technology in the field of fuel cells based on patent paper analysis (based on 247 articles matched to 155 patents).
  - Martinelli, A. and Romito, E. (2019) : When authors become inventors: An empirical analysis on patent-paper pairs in medical research, LEM Working Paper Series, No. 2019/32, Scuola Superiore Sant'Anna (379 PPPs)
  - La, H.L. and Bekkers, R. (2021). Science and Technology Relatedness: The Case of DNA Nanoscience and DNA Nanotechnology.

However, these works do not attempt to quantify the proximity between scientific publications and patent applications

Relating patents & scientific publications: PPP & text

With a few exceptions... using text from patent and scientific paper siblings to relate science and technology

Using keywords to link texts:

- Magerman, T., Van Looy, B. and Debackere, K. (2015), rely on the number of common terms between documents (hence facing"brittelness" shortcomings).

Using Latent Semantic Analysis:

- Tom Magerman, B. Van Looy, B. Baesens and Koenraad Debackere (2011) matched patent and publication documents based on content similarity scores using Latent Semantic Analysis (LSA). Patent-publication combinations having a high content similarity are regarded to originate from the same inventive event. However LSA failed to discriminate between true PPP and non-PPP documents: " LSA- based measures tend to overestimate similarity and not grasp the real topic similarity of patent and publication documents. Expert validation of 250 cases confirmed the poor performance of LSA based measures."

Should we give up on text-based similarity metrics?

Lately: developments in text embeddings may offer new avenues to determine the proximity between patents and scientific publications

Example (1) :

Patent (US 20090299929 A1- Methods of improved learning in simultaneous recurrent neural network- Inventor Robert Kozma, Paul J. Werbos, Applicant: University of Memphis Research Foundation) **abstract**

Methods, computer-readable media, and systems are provided for machine learning in a **simultaneous recurrent neural network**. One embodiment of the invention provides a method including initializing one or more weight in the network, initializing parameters of an **extended Kalman filter**, setting a Jacobian matrix to an empty matrix, augmenting the Jacobian matrix for each of a plurality of training patterns, adjusting the one or more weights using the extended Kalman filter formulas, and calculating a network output for one or more testing patterns.

Article (Ilin, R., R. Thijs Kozma and P.J. Werbos. "Efficient Learning in Cellular Simultaneous Recurrent Neural Networks - The Case of Maze Navigation Problem." 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning) **abstract**

Cellular **simultaneous recurrent neural networks** (SRN) show great promise in solving complex function approximation problems. In particular, approximate dynamic programming is an important application area where SRNs have significant potential advantages compared to other approximation methods. Learning in SRNs, however, proved to be a notoriously difficult problem, which prevented their broader use. This paper introduces an **extended Kalman filter** approach to train SRNs. Using the two-dimensional maze navigation problem as a testbed, we illustrate the operation of the method and demonstrate its benefits in generalization and testing performance

Example (2) :

Patent (US11321604B2, "Systems and devices for compressing neural network parameters"
Inventors: Jiecao YU Andrew Lukefahr David Palframan Ganesh Dasika Reetuparnda Das Scott Mahlke
ARM Ltd University of Michigan
Priority: 2017-06-21**) abstract**

"Subject matter disclosed herein may relate to storage and/or processing of signals and/or states representative of neural network parameters in a computing device, and may relate more particularly to **compressing signals and/or states representative of neural network nodes in a computing device.**"

Article (Jiecao Yu, A. Lukefahr, D. Palframan G. Dasika Reetuparnda Das, S. Mahlke "Scalpel: Customizing DNN pruning"2017) **abstract**

"As the size of Deep Neural Networks (DNNs) continues to grow to increase accuracy and solve more complex problems, their energy footprint also scales. **Weight pruning reduces DNN model size and the computation by removing redundant weights.** However, we implemented weight pruning for several popular networks on a variety of hardware platforms and observed surprising results. For many networks, the network sparsity caused by weight pruning will actually hurt the overall performance despite large reductions in the model size and required multiply-accumulate operations. Also, encoding the sparse format of pruned networks incurs additional storage space overhead. To overcome these challenges, we propose Scalpel that customizes DNN pruning to the underlying hardware by matching the pruned network structure to the data-parallel hardware organization. Scalpel consists of two techniques: SIMD-aware weight pruning and node pruning. For low-parallelism hardware (e.g., microcontroller), SIMD-aware weight pruning maintains weights in aligned fixed-size groups to fully utilize the SIMD units. For high-parallelism hardware (e.g., GPU), node pruning removes redundant nodes, not redundant weights, thereby reducing computation without sacrificing the dense matrix format. For hardware with moderate parallelism (e.g., desktop CPU), SIMD-aware weight pruning and node pruning are synergistically applied together. Across the microcontroller, CPU and GPU, Scalpel achieves mean speedups of 3.54x, 2.61x, and 1.25x while reducing the model sizes by 88%, 82%, and 53%. In comparison, traditional weight pruning achieves mean speedups of 1.90x, 1.06x, 0.41x across the three platforms."

Transformers and contextual embeddings

The growing dominance of vector embeddings to represent words/documents

- TF-IDF (Hans Peter Luhn (1957), and Karen Spärck Jones frequency (1972))
- word2vec/doc2vec (Mikolov 2013, Le & Mikolov 2015)
- "attention is all you need" (Vaswani et al., 2017) → Transformers
- BERT (Devlin et al., 2018)

Specialised models trained on technical documents (patents or scientific publications)

- PatentBert (2019)
- SciBert (2019)
- SPECTER (2020) → **document-level embedding** of scientific documents based on pretraining a Transformer language model
- PatentSBERTa (2021)
- Bert for patents (2022)

However, while showing great strength at specific applications, some of these models are optimised for classification (PatentBert) and do not focus on relating documents.
Furthermore, these models are either based on patent sources or on a (limited) set of technical/scientific domains
→ the impact on the language model of both the language (form) and of the choice of technical domain remains open.

→ Can the latest generations of language models based on contextual word embeddings provide a sufficiently precise representation of the underlying technical content of patent and article abstracts?

Sub-questions

Can contextual text-embeddings capture — at least in part— a measure of technical similarity between a patent and a scientific publication?

i.e. if A is technically closer to B than C then we would expect: sim(A,B) < sim(A,C)

→ PPP offer a "ground truth" corpus in selected technical domains (where A ≃ B ⇒ sim(A,B) is minimized) to put contextual embeddings to the test:

- **Can contextual embeddings select/identify/discriminate between PPP and non-PPPs?**
- **Since no 2 embeddings are equal: necessity to test different embeddings**

Should contextual embedding be able to capture such technical proximity, this may open perspectives to automate the process of identifying PPP and to fine-tune existing models (with a focus on improving the technical similarity measures).

Out of scope:

- The goal is not to create a complete sample of PPPs (we opt to favor a high specificity rather than a high recall).
- The goal is not to build a tool to identify the closest prior art

Methodology

Manual "ground truth" corpus

- Collect reference corpus of true “patent-article” pairs corresponding to identical technical innovations in the field of AI
- Build suite of patents DB → Semantic Scholar automation tools
- Focus on core sub-fields (maximize precision, i.e. true positives, not recall)

Compute embeddings

- Different types of embeddings: PatentBert, Specter, PatentSBERTa, bert-for-patents, Doc2vec
- Computation of the embeddings : CLS vs mean pooling / titles, abstracts / claims (original or granted)
- Computation of the (cosine) similarity between documents

Analysis

Measurement of the effectiveness of each embedding at:
(i) identifying closest documents (rank statistics, distances)
(ii) at discriminating true patent-article pairs from spurious ones (F1 test)

Application:

Build a classifier to automate the end-to-end detection of true patent-article pairs

Methodology

- **Manual selection of PPPs in subdomains of machine learning.**
- Evaluation of different embeddings
- Development of a classifier

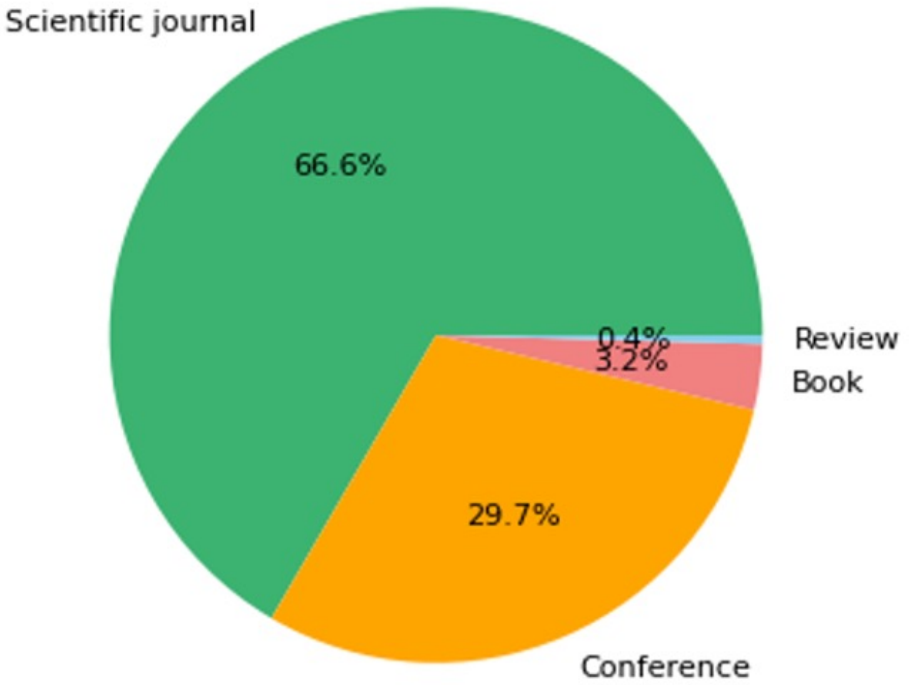
- **Selection of "functional AI/ML" sub-domains**
  - Mapped to CPC (in functional ML fields) and specific keywords (GANs, Transformers, RL)
  - Extraction of patents (granted and pending) from USPTO

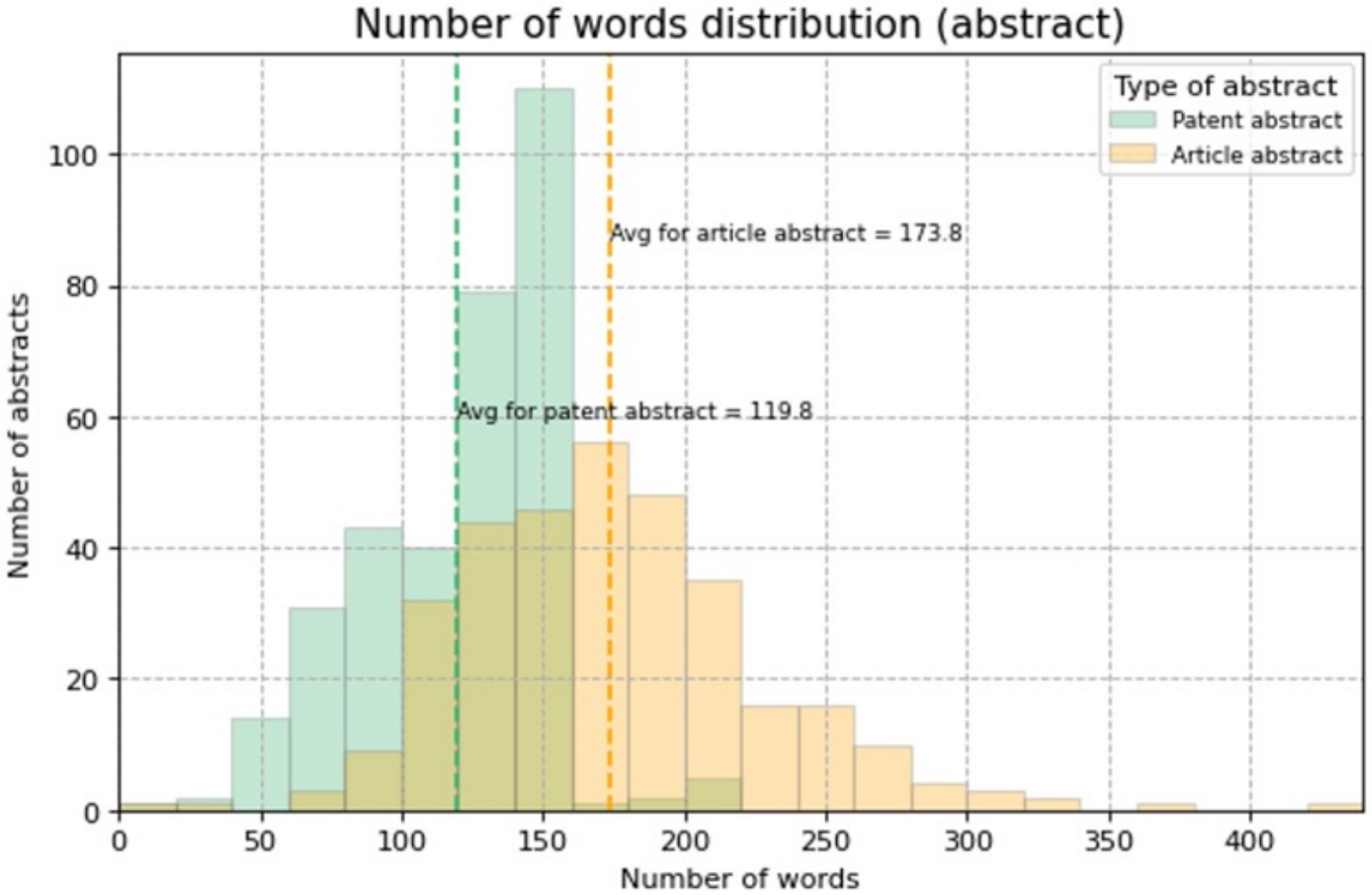
- **Selection of corresponding scientific articles**
  - based on inventors' (last) names, publication date vs priority date
  - query "article corpus" (semantic scholar's API: elastic search where: {authors}={inventors})
  - manual inspection of top N (up to 20) results
  - selection of "sibling" (using abstract and description/figures in both patent and article)

→ 462 PPPs (329 unique pairs) in ML

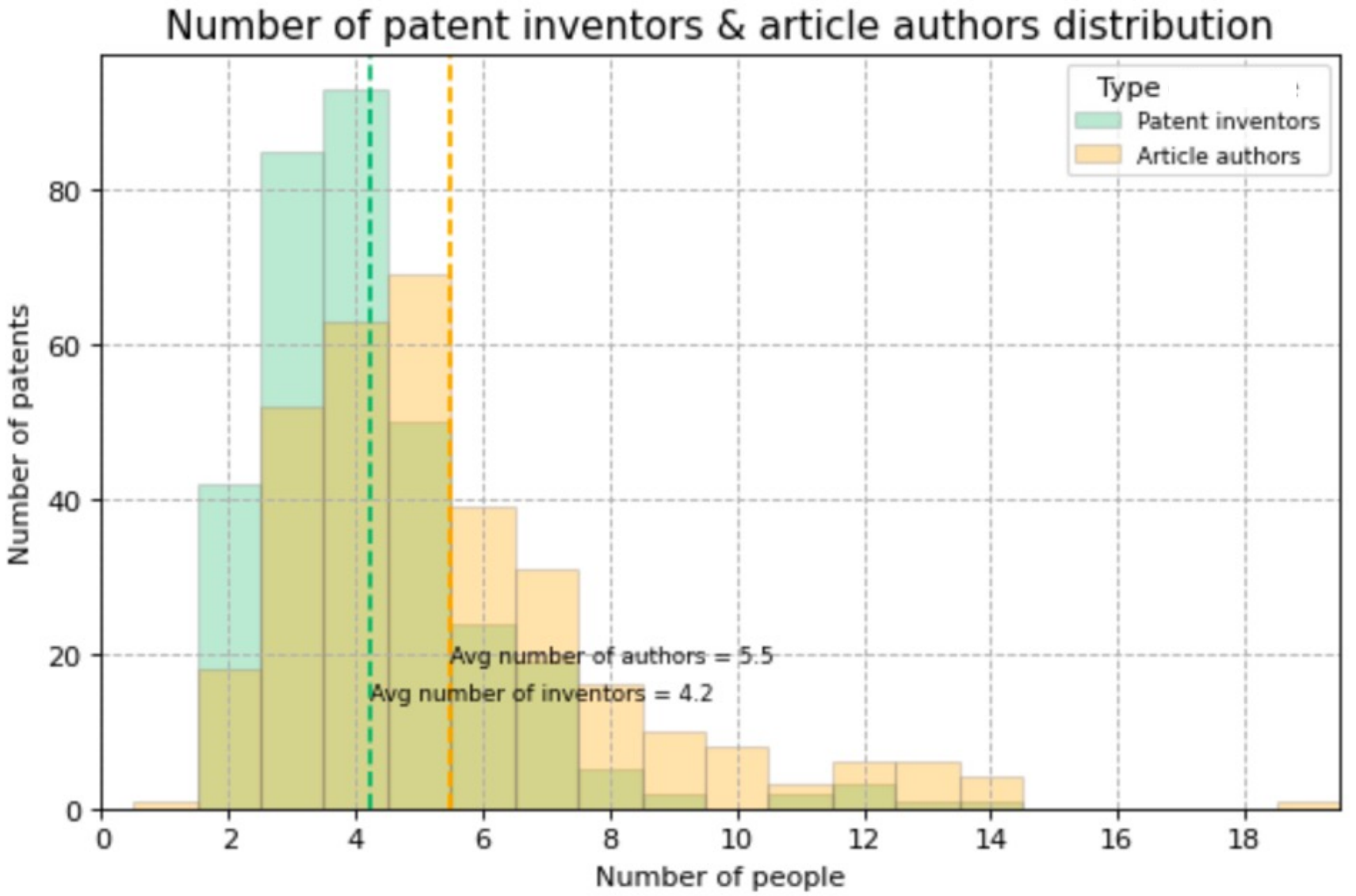
20% of all selected AI/ML patents lead to a matched article in semantic scholar top 10 results.

Publication types and abstracts lengths

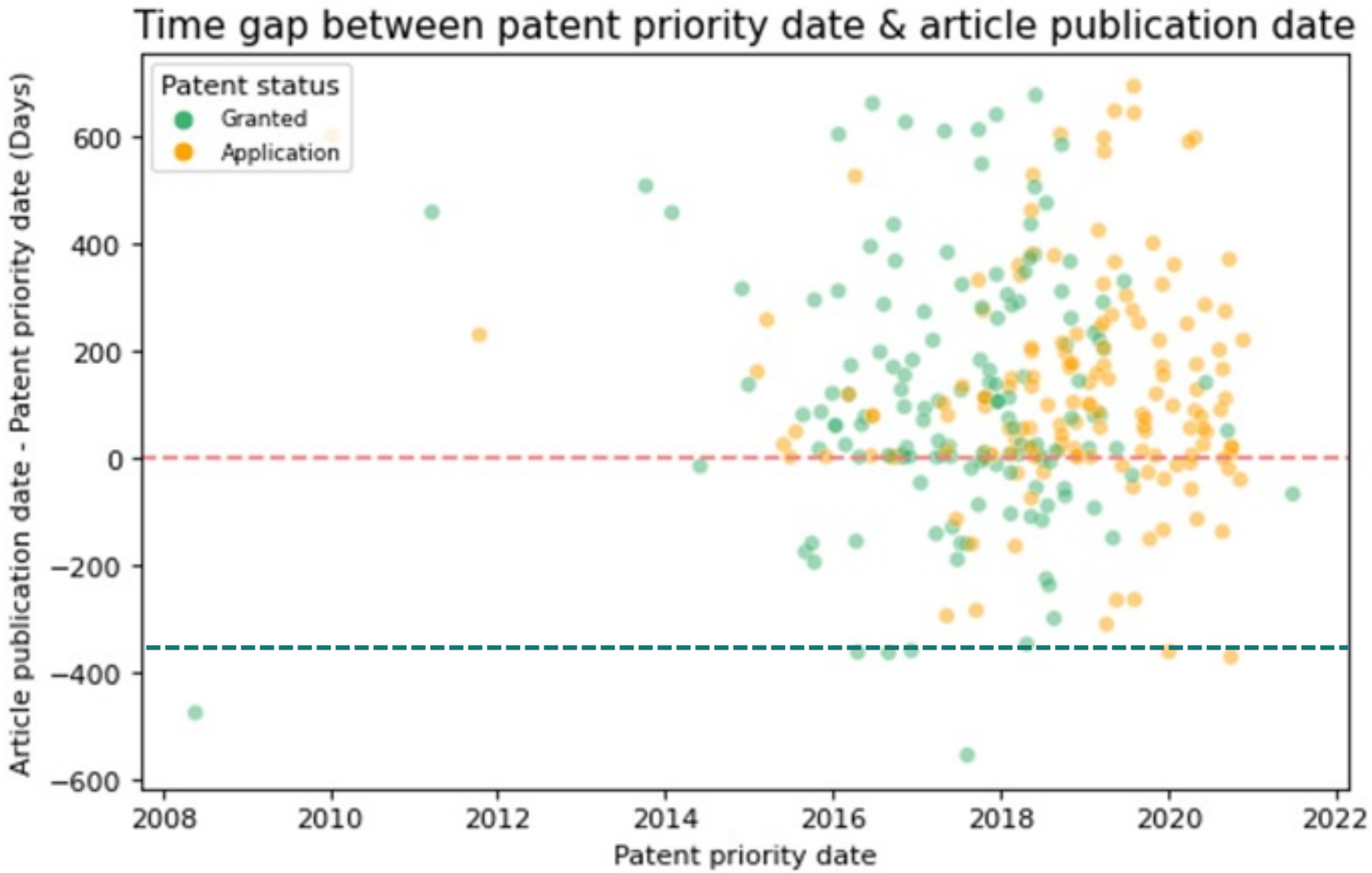




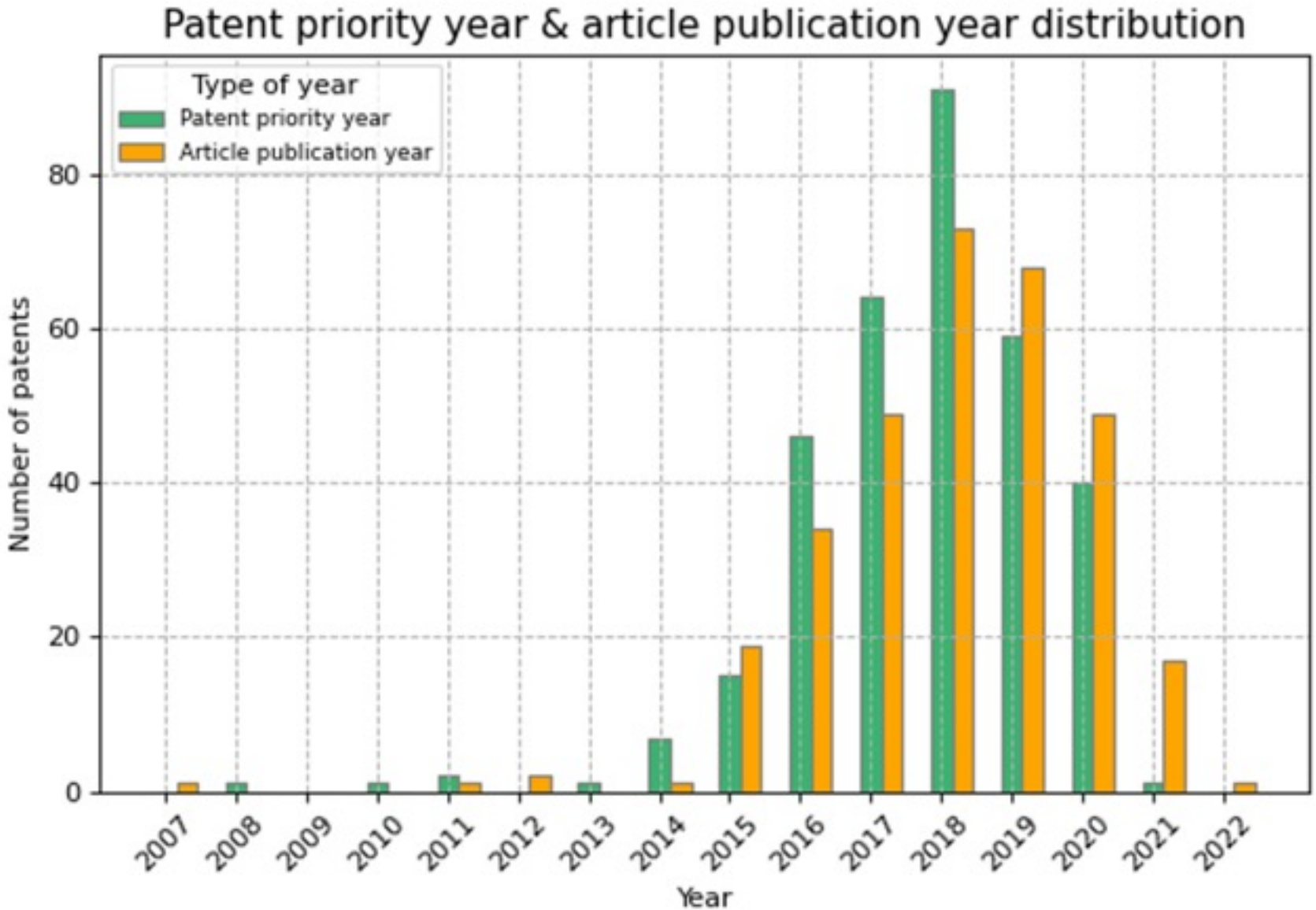
Authors & inventors



Time gap between patent and scientific publications



Publications dates



From texts to vector representations (embeddings)

Model	Training Data	Original Mission
Doc2Vec (benchmark)	Trained on the English Wikipedia data	Not domain-specific Not mission-specific
Specter	Trained based on transformer model architecture & document-level relatedness: the citation graph	For scientific documents Not mission-specific
PatentBert	Fine-tuned with over 2 million patents, based on a pre-trained BERT model	For patents For CPC classification
Bert-for-Patents	Retrained based on "BERT Large" model with more than 100M patent documents (abs, claims, desc.)	For patents Not mission-specific
PatentSBERTa	Retrained based on Sentence-BERT & full patent claims	For patents For patent-to-patent similarity

Generation of vector representations — 2 options:

- **[CLS]:** special token **used for classification task (not specifically document embedding, although often used as proxy)**
- **"Mean pooling" (MP)** : average the second to last hidden layer of each token → produces a single vector as the document representation.

Patent-article cosine similarities: reference sets

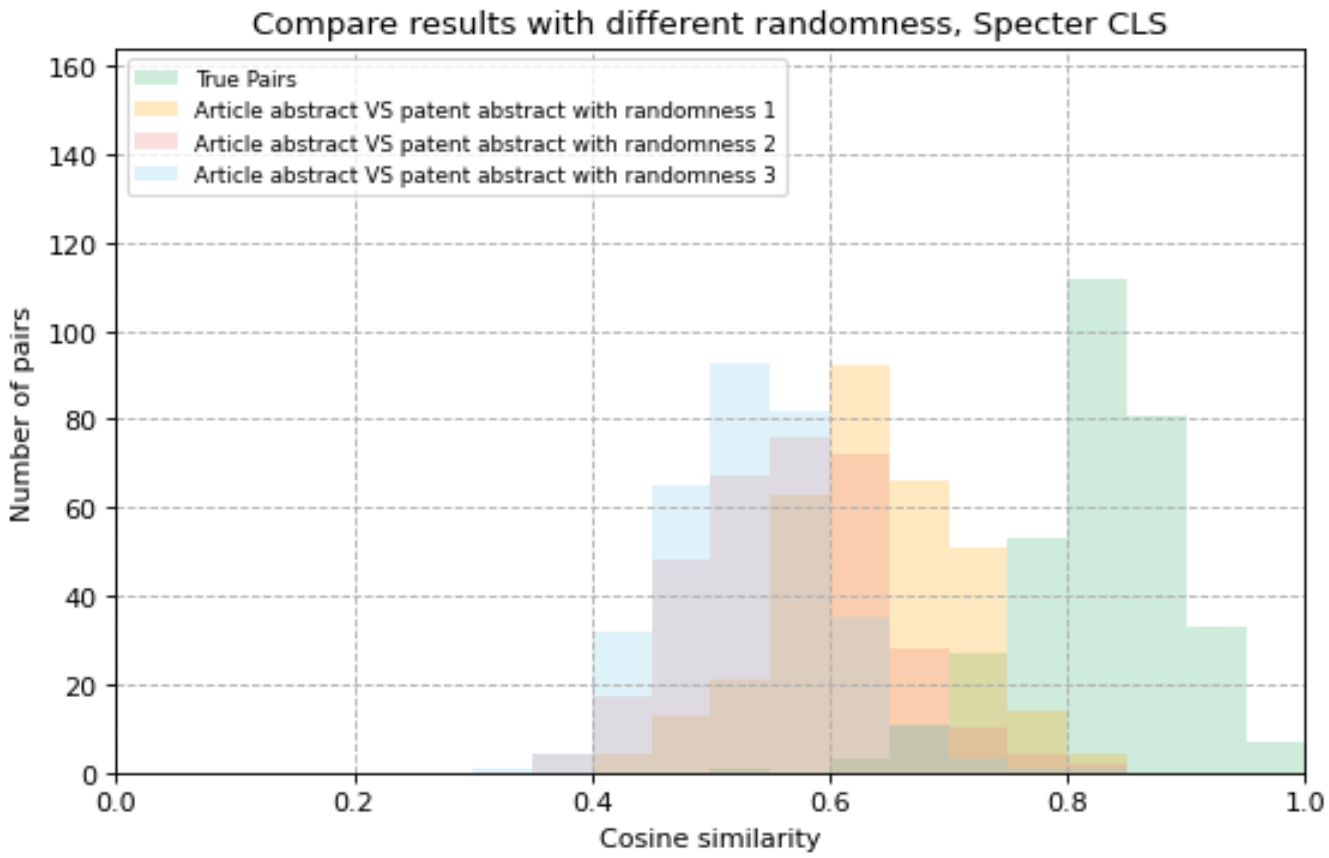
Comparison of $\boldsymbol{\sigma}_{cos}$= **cosine similarity (** $\varepsilon_{i,P}^{M}$ **,** $\varepsilon_{i,A}^{M}$ **)** between patent and articles

1 reference (true) and 3 randomized reference sets:

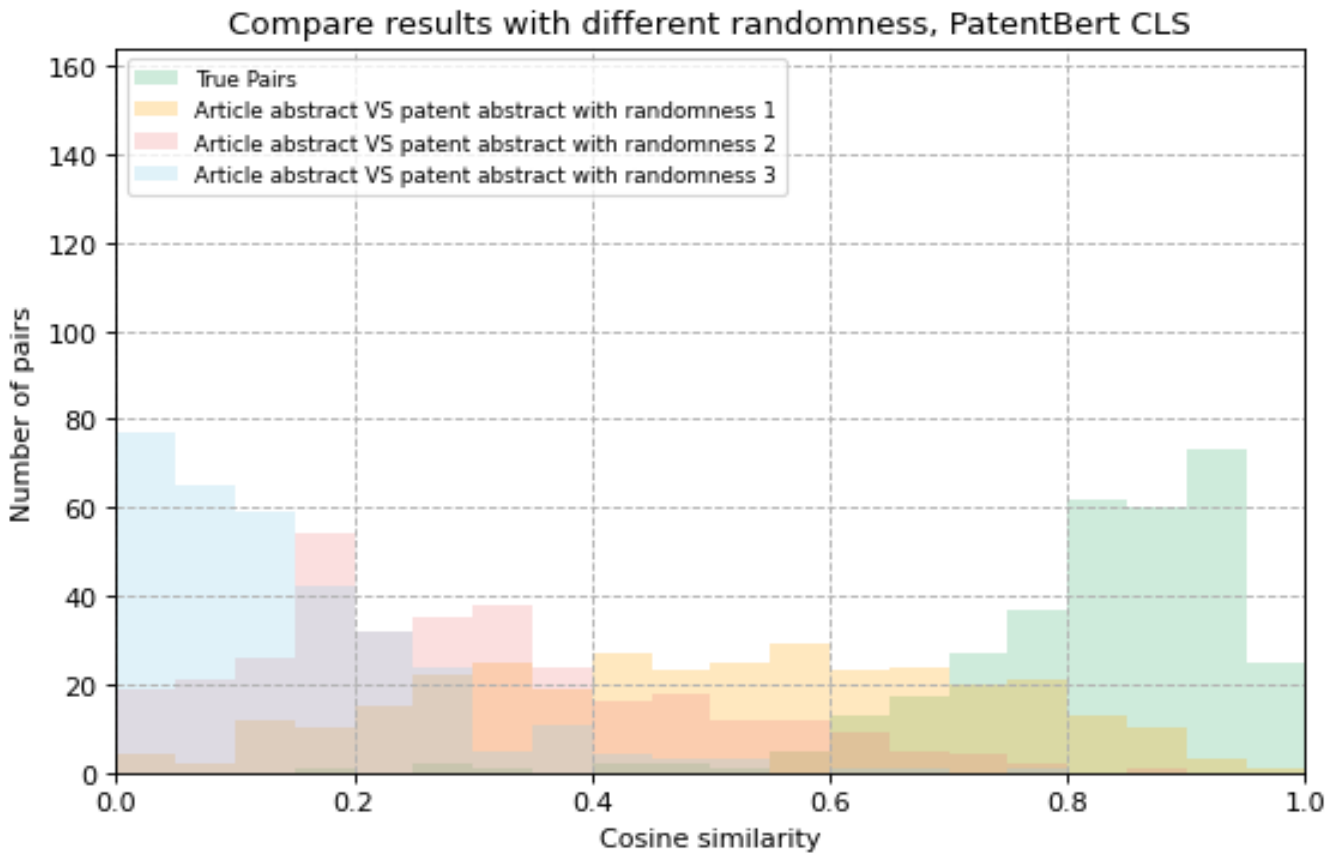
- S_{true} : similarity between manually selected ("true") pairs
- $S_{rnd,true}$: similarity of article to random patents in the manually selected "true" set
- S_{AI} : similarity of article to random AI patent (based on CPC, i.e. at least one CPC in G06N3-5,10, G06K9, G10L11-17, G06F40, etc.)
- $S_{non\ AI}$: similarity of article to random non-AI patent similarity (no CPC in G06N3-5,10, G06K9, G10L11-17, G06F40, etc.)

→ **We expect:** $\overline{S}_{true} > \overline{S}_{rnd,\ true} > \overline{S}_{AI} > \overline{S}_{non\ AI}$

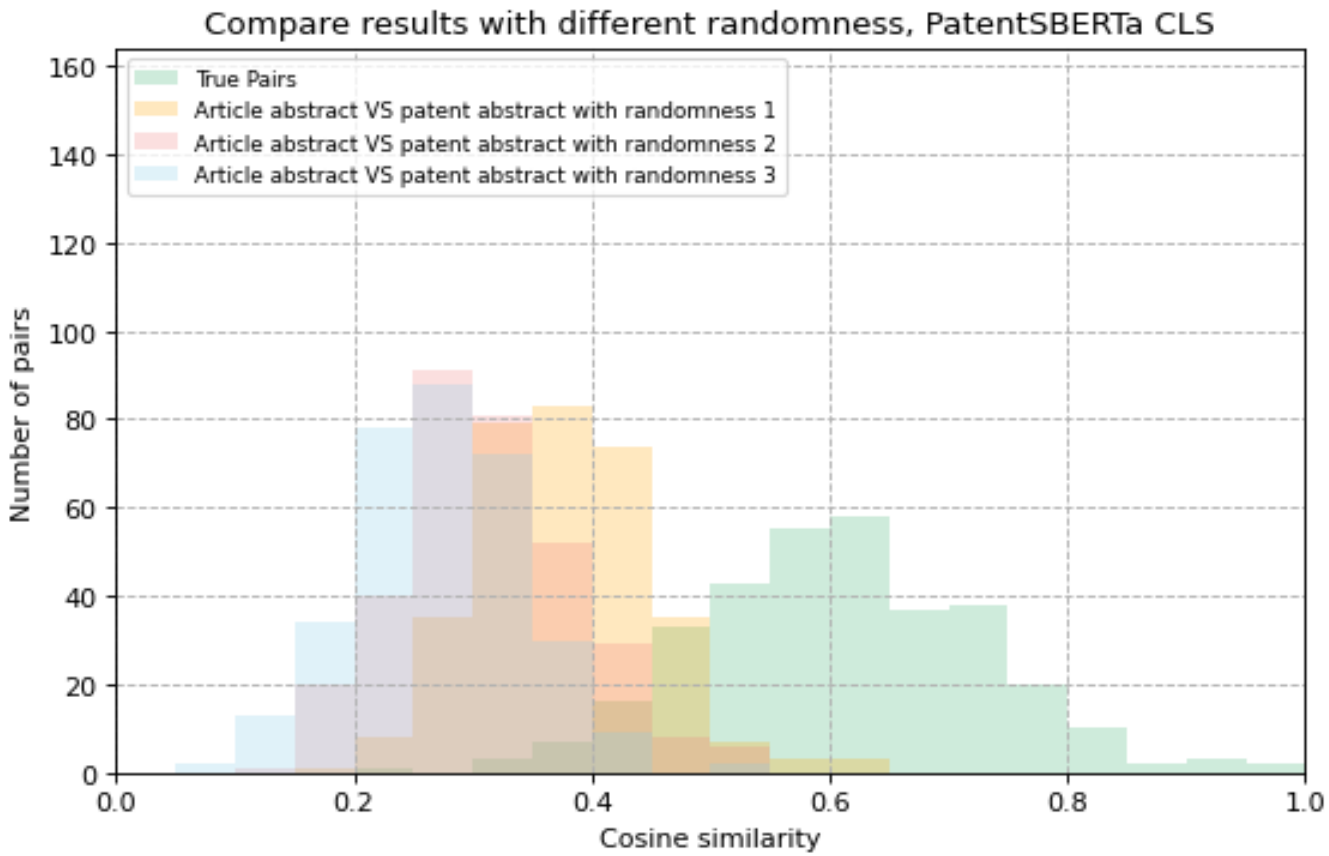
Patent-article cosine similarities: SPECTER [CLS]



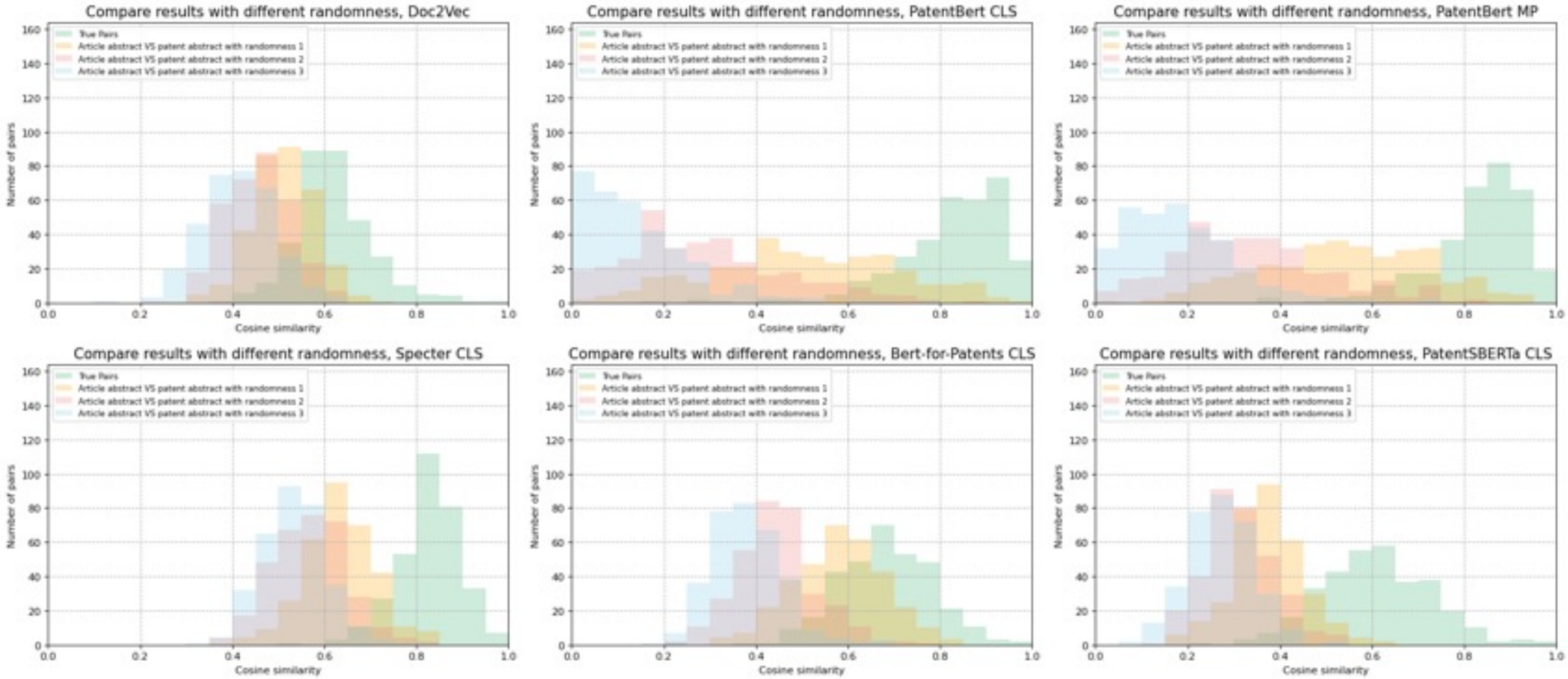
Patent-article cosine similarities: PatentBert (MP)



Patent-article cosine similarities: PatentSBERTa [CLS]



Patent-article cosine similarities



- contextual embeddings improve over Doc2vec
- behaviour of embeddings varies significantly (S_{AI}: F1=95.3% @ $\boldsymbol{\sigma}_{cos}$=0.7 for SPECTER while
- SPECTER and PatentSBERTa are capable of discriminating true PPP from non PPP

Rank of true sibblings

What is the rank of the true match compared to all other documents in the S_{AI} sample?

Quantile (top Q%) : fraction of patents for which the similarity to the matched article is in the top Q%

	Doc2Vec	PatentBert [CLS]	PatentBert [MP]	SPECTER [CLS]	Bert for Patent [CLS]	PatentSBERTa [CLS]
		PatentBert [CLS]				
top 1%	61.28%	73.78%	77.44%	87.81%	43.60%	86.28%
top 3%	76.22%	88.42%	88.11%	94.51%	76.83%	94.51%
top 5%	81.10%	93.90%	93.29%	97.56%	86.59%	95.43%

- Performance of contextual embeddings does vary
- SPECTER and PatentSBERTa are best in class

→ ($\varepsilon_{i,P}^{SPECTER}$, $\varepsilon_{i,A}^{SPECTER}$) to build a classifier

Classifier pipeline

Objective: build a classifier to predict PPPs (i.e. whether an article is a match to a patent)

- **Features:**
  - authors / inventors:
    - fraction of authors listed as inventors
    - fraction of inventors listed as authors
    - Jaccard index (inventors, authors): J(Inv, Auth) = $\frac{|\mathbf{Inv} \cap \mathbf{Auth}|}{|\mathbf{Inv} \cup \mathbf{Auth}\,|}$
  - dates → $\Delta T = T_{priority} - T_{publication}$
  - text embeddings (abstracts) → $\sigma_{cos}^{SPECTER}$ = **cosine similarity (** $\varepsilon_{i,P}^{SPECTER}$ **,** $\varepsilon_{i,A}^{SPECTER}$)

Train a **random forest** classifier (80% train - 20% test)

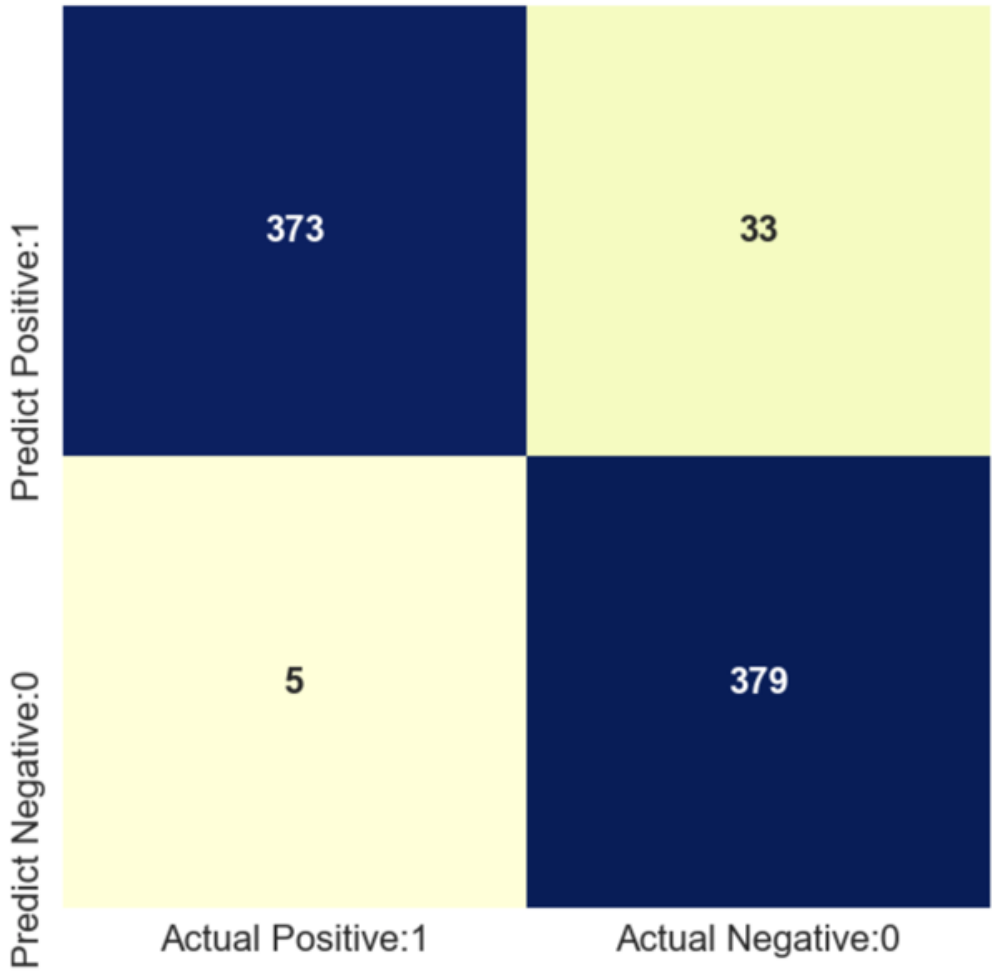
Towards an "AI PPP" classifier

- Features: (authors/inventors, dates, abstracts) from patents and articles
- **Random Forest**

```
Model accuracy score: 0.9519
Training set score: 0.9997
Test set score: 0.9519
True Positives(TP) =  373
True Negatives(TN) =  379
False Positives(FP) =  33
False Negatives(FN) =  5
ROC AUC score: 0.9910457717569786
              precision    recall  f1-score   support

           0       0.99      0.92      0.95       406
           1       0.92      0.99      0.95       384

    accuracy                           0.95       790
   macro avg       0.95      0.95      0.95       790
weighted avg       0.95      0.95      0.95       790
```



Next steps

- **Ongoing work:**
  - **Build on current results to automate the detection of PPP in a variety of technical fields**
    - AI/ML
    - Cryptography
    - Quantum computing
    - ARNm

  → Build a publicly available corpus of field dependent PPPs

- **Next:**
  - Extend to other souces: preprints/other media
  - fine tune language models to improve similarity of PPPs

Remaining issues / limitations:

- patent and article do not necessarily cover the same subject matter tested on claims: same as article (could similarity of article content / patent description help?)
- reliance on authors/inventors names (pb of normalization/errors)
- abstracts missing (<0.1%)