Mapping Inventions in the Space of Ideas, 1836–2022: Representation, Measurement, and Validation

Ina Ganguli[a], Jeffrey Lin[b], Vitaly Meursault[b], Nicholas Reynolds[c]

[a] *University of Massachusetts Amherst and NBER*
[b] *Federal Reserve Bank of Philadelphia*
[c] *University of Essex*

April 2, 2024

Abstract

How well can different methods meaningfully represent inventions in the "space of ideas"? We evaluate four leading natural language processing (NLP) models, each of which produces a different numerical representation of patent text. We design three novel, domain-specific validation tasks to select between these representations. Sentence-BERT (S-BERT) significantly outperforms other widely used NLP models, creating metrics better aligned with both expert and non-expert human judgment about patent similarity. The choice of representation matters significantly for economic measurement. According to S-BERT, contemporaneous patents have declined in similarity over more than a century, as inventions have "spread out" on an expanding knowledge frontier. Other representations report ambiguous or diverging patterns. We reproduce the S-BERT result using newly digitized records of historical interferences, which show secular declines in the rate of multiple invention. Our results highlight the importance of validation and model selection as an essential step in constructing and using measures derived from patent text.

We are extending our analysis to include the latest generation of "ChatGPT-era" embedding models. OpenAI's latest embeddings significantly outperform S-BERT in our main validation task. We are in the process of fully integrating these new results into our paper.

JEL classification: O31, C81, L19

Keywords:
Innovation Economics, Natural Language Processing, Deep Learning, Patent Interferences

Author information: Ina Ganguli, iganguli@umass.edu; Jeffrey Lin, jeff.lin@phil.frb.org; Vitaly Meursault, vitaly.meursault@phil.frb.org; Nicholas Reynolds, nicholas.reynolds@essex.ac.uk

Acknowledgements: We gratefully acknowledge support from an NBER Innovation Policy Grant. We also received excellent RA support from Aaron Rosenbaum, Joseph Huang, Cameron Fen, Annette Gailliot, Jake Moore, and Isaac Rand. Finally, we received useful feedback from Matt Clancy, Gaétan de Rassenfosse, Luise Eisfeld, Semyon Malamud, Roxana Mihet, Darya Davydova, the participants of the seminar at EPFL, the participants at the NBER Innovation Information Initiative Technical Working Group Meeting, and TADA 2023. First version: December 21, 2023.

1. Introduction

We need to measure invention similarity. When evaluating patent applications, an examiner must assess the similarity of its claims against prior patents and applications. For inventors racing for priority, the overlap between their own claims and those of their competitors is a crucial consideration. Policy makers may want to measure idea similarity to encourage or discourage the duplication of independent effort on specific inventions.

Each of these settings requires a reliable measurement of invention similarity. Furthermore, invention similarity has long been a core concern for researchers of innovation. Similarity may be a factor in the value of a patent, the strength of knowledge spillovers, the direction of technological change, or the optimality of R&D investments (Griliches, 1979; Jaffe, 1986). Measuring similarity is also a first step in calculating many other measures of economic interest. For example, Kelly et al. (2021) measure "breakthrough" patents as those that are dissimilar to prior patents but very similar to future patents.

Many methods have been used to measure patent similarity. How should researchers choose among them? Our main contribution is to develop and implement a pipeline for the construction, validation, and selection of measures of economic interest derived from patent text. This pipeline is a step-by-step guide for researchers constructing other measures. We emphasize domain-specific validation and model selection as an essential step. Our pipeline produces a validated representation of every US patent, 1836–2022, in a "space of ideas."

The construction of patent similarity and other measures can be usefully separated into three distinct steps: (1) representation, (2) measurement, and (3) validation. The first step maps each patent to a location in idea space; i.e., it *represents* each patent as a vector in R^n. This mapping could be based on, e.g., patent office classifications, traditional Natural Language Processing (NLP) methods that count words, or modern NLP methods that produce distributed embeddings. These vector representations define a space of ideas. The second step *measures* a concept of economic interest using representations produced by each of several candidate models. Patent similarity is a classic quantity of interest; other concepts

might be motivated from theory, derived from a structural model, or based on intuition.

Different representations lead to different measurements of the same concept. Therefore, the third step *validates* these representations using purpose-built, domain-specific tasks to select the "best" mapping. This step is rare in economics and is a central focus of this paper. Our approach accords with the common view in the NLP literature that there is no single best method for all tasks, and therefore, methods should be assessed on their usefulness for each particular task (Ash and Hansen, 2023). For example, two of the six "key principles for text analysis" in the textbook by Grimmer et al. (2022) are "the best method depends on the task" and "validations are essential and depend on the theory and the task."

We follow these guidelines to construct validated measures of invention similarity for US utility patents, 1836–2022. We design three novel, domain-specific validation tasks that compare the performance of four leading and widely-used NLP models: (1) Term Frequency-Inverse Document Frequency (**TF-IDF**; Sparck Jones, 1972), (2) **doc2vec** (Le and Mikolov, 2014; Mikolov et al., 2013), (3) Universal Sentence Encoder (**USE**) (Cer et al., 2018), and (4) Sentence-BERT (**S-BERT**) (Reimers and Gurevych, 2019; Devlin et al., 2019). Each task uses a sample of patent pairs with human judgments of similarity. We assess how well different representations agree with human judgment. Our validation tasks vary across three important dimensions. One, they span both expert and non-expert human judgment. Two, they span text from both modern and historical patents. Three, they span a range from near-identical similarity to coarser judgments of relatedness. We find that S-BERT significantly outperforms other widely-used NLP models in all three of our tasks.

Our first validation task uses patent interference cases, combining modern patent application text and expert judgment of near-identical similarity. Interference cases were triggered when the US Patent and Trademark Office (USPTO) received simultaneous, identical claims of invention from multiple independent parties (Ganguli et al., 2020).[1] Therefore, interfer-

[1]Interferences cases were used to determine priority of invention under the "first to invent" rule that prevailed in the US until 2014.

ences are indicators of "multiple invention" (Merton, 1957, 1973). Interferences were declared by examiners that specialized in particular technology classes, and thus indicate expert human judgment of near-identical *legal* similarity. We use applications in interferences decided 1998–2014. S-BERT significantly outperforms other models in classifying interfering pairs, with a precision-recall area under curve (PR AUC) of 48%, followed by TF-IDF at 43%. This difference is economically meaningful: a law firm switching from TF-IDF to S-BERT-based measures to screen potential interference pairs could reduce false positives (i.e., paralegal time) by 21%–61%. USE and doc2vec perform significantly worse, producing around 5 and 70 times more false positives than S-BERT, respectively, for a given number of true positives. In an ongoing analysis of recent (post-2023) embedding models, we show that OpenAI embeddings offer further dramatic improvements, reducing the number of false positives in one of the tests by 63% relative to S-BERT.

Our second validation task uses patent classifications, 1836–2022. The USPTO uses classifications to group patents with a common subject matter. Patent classifications indicate expert human judgment that there is broad (but not exact) technological similarity between different patents. Therefore, they measure similarity at a coarser scale compared with interferences. S-BERT again significantly outperforms TF-IDF and the other models.

Our third validation task uses a general, non-expert sense of similarity for historical patents, 1880–1920. A challenge is that humans have difficulty precisely quantifying the similarity of a pair of patents. Our solution is to sample patent *triples*: a reference patent A and two comparison patents B and C. The triples are somewhat similar according to both S-BERT and TF-IDF, but the models disagree about whether B or C is more similar to A. We ask humans to resolve the disagreement. Humans prefer S-BERT 71% of the time.

Taken together, S-BERT representations produce measures of patent similarity which more closely match human judgment. Our publicly-available S-BERT representations can be used to calculate the similarity of any patent pair. Our results suggest these representations should be the current standard. Of course, there is rapid development in new models; our

pipeline provides guidance on how to evaluate them. Newer models are not always better: TF-IDF (from the 1970s) significantly outperforms two recent models, USE and doc2vec.

The choice of model matters for the measurement of long-run trends in patent similarity. We analyze the full text of US patents, 1836–2022, split into four corpora: titles, abstracts, descriptions, and claims. Figure 1 shows average pairwise similarity by model (S-BERT vs. TF-IDF), corpora, and year (see details in Section 4.1). According to S-BERT, across corpora, patent similarity exhibits a secular decline. In contrast, measures based on TF-IDF show ambiguous or diverging patterns. Using TF-IDF, abstracts and claims are becoming more similar over time, while descriptions and titles have unclear trends. The results of our prior validation step are therefore *essential* for interpreting this evidence. We also find that the size of the S-BERT idea space has steadily increased over time, suggesting that inventors are "spreading out" on an expanding knowledge frontier. Finally, according to S-BERT, patent similarity has declined both within and between classes, suggesting that traditional approaches using patent classifications to measure similarity may miss important margins.

We also re-examine the evidence on trends in "breakthrough" inventions following Kelly et al. (2021). We are able to replicate some qualitative features of their study, but with fewer discretionary researcher choices. Based on these results and our comparative measurements of invention similarity, we conclude that S-BERT is more robust compared with TF-IDF.

We reproduce the S-BERT finding on declining invention similarity by estimating the rate of interference over nearly 150 years. In theory, the rate of interference declines when inventors choose projects that are less similar. We combine newly-digitized 1864–1901 *Registers of Interferences* with summary statistics of interferences for 1950–1962 and 1981–1993 and our 1998–2014 database of interference decisions. Consistent with S-BERT-based estimates, the rate of interference also exhibits a secular decline. And, since interferences before 1998 were not used to validate the representations produced by S-BERT, this result represents independent, out-of-sample confirmation of the long-run decline in invention similarity.

In a companion paper (Ganguli et al., 2024), we develop a theory where the decline

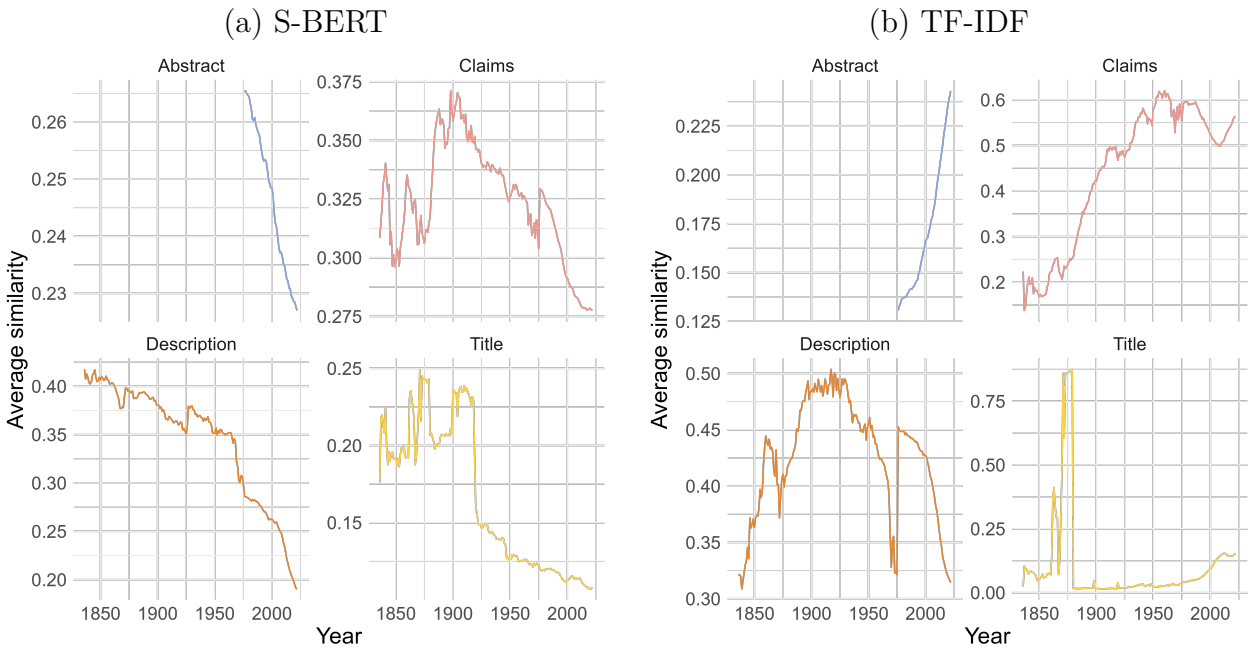


Figure 1: Average pairwise patent similarity by model, corpora, and year

in contemporaneous invention similarity is related to recent findings on long-run invention dynamics, including the increasing "burden of knowledge" (Jones, 2009), increasing R&D spending (Hirschey et al., 2012), declining R&D productivity (Bloom et al., 2020), and constant R&D spillovers (Lucking et al., 2019). The increasing burden of knowledge raises the fixed costs of inventing over time. This restricts entry into invention as the space of inventions grows, leading inventors to "spread out" over an expanding knowledge frontier. Ideas get "harder to find" because there are weaker positive knowledge spillovers from "neighbors" that are now more distant in idea space. Inventors increase their own R&D inputs in response to weaker spillovers, thus reducing own-R&D productivity. (On net, total spillovers may be roughly constant as increasing idea distance is offset by increases in own-R&D investment.)

Our paper builds on work measuring the similarity of inventions and ideas. Some have used features such as overlapping patent classifications (Fleming, 2001; Clancy, 2018; Akcigit et al., 2017), keywords (Azoulay et al., 2019), or citations (Wang et al., 2017; Berkes and Gaetani, 2020). Others use the workhorse model TF-IDF (Kelly et al., 2021) or newer NLP models including doc2vec (Feng, 2020) and S-BERT (Lee and Hsiang, 2019). Some

recent work evaluates the performance of NLP models (Arts et al., 2021, 2018; Cheng et al., 2022). A typical approach is to validate a single representation using expert judgment (e.g., classifications) or choice behavior (e.g., citations). Compared with this work, our analysis contributes a comparative design that evaluates several leading NLP models against a common battery of validation tasks. We also provide general guidelines for innovation researchers using NLP methods to measure economic quantities of interest, design novel validation tasks, and document new facts about invention similarity over time.

Our analysis focuses on pairwise, contemporaneous invention similarity. This measure is distinct from the related concepts of "novel" (Akcigit et al., 2017), "disruptive" (Park et al., 2023), "breakthrough" (Kelly et al., 2021), or "unconventional" (Berkes and Gaetani, 2020) innovations analyzed previously. In some cases, similarity is a direct input into a derived measure. Our contribution is to highlight the importance of validation and model selection for constructing measures based on patent text.

Finally, our results are useful to many applications in innovation economics. For example, similarity measures may be used for constructing matched controls in studies of localized knowledge spillovers (Jaffe et al., 1993; Thompson and Fox-Kean, 2005; Murata et al., 2014; Ganguli et al., 2020). Similarity measures seem especially useful for empirical study of theories of idea space (Dasgupta and Maskin, 1987; Akcigit et al., 2017; Clancy, 2018).

The rest of the paper is structured as follows. Section 2 outlines a pipeline for the construction and validation of measures (including similarity) from patent text. Section 3 compares the performance of different representations in our validation tasks. Section 4 shows that representation affects measurement. Section 5 presents additional results and potential explanations for S-BERT's superior performance compared with TF-IDF.

2. Framework and Pipeline

A researcher wants to measure a concept or quantity of economic interest based on patent text. This concept might be based on researcher intuition, might be motivated informally

from theory, or explicitly derived from a structural economic model. Many such concepts start with representations of patents as vectors in R^n, and therefore they define a space of ideas. Then, measurements of the economic concept or quantity are functions of the representations. Importantly, the formulation of a concept or quantity is distinct from a particular representation or mapping of patents to idea space.

This section outlines a general pipeline for the construction and validation of similarity measures and other measures of economic interest from patent data. Figure 2 provides a schematic view. Given an economic concept or quantity, the construction of many measures used in the literature can be usefully separated into several distinct steps. Step 1 maps each patent to a location in idea space, i.e., *represents* each patent as a vector. Step 2 *measures* the concept or quantity of interest, e.g., pairwise cosine similarities.

The contribution of this paper is to emphasize Step 3, *validation-based selection* between alternative representations of patent text. Validation uses external measures which have clear theoretical connections with the concept of interest but are available only in a subsample of the data. We use the performance of different representations on multiple common validation tasks to select the model which most closely accords with human judgment (or "ground truth"). This goes beyond some prior usage of "validation" to mean any correlation of a measure based on a single representation to external judgment. Taken together, these validation tasks allow us to select the representation with the best concept validity (Step 4).

Work in progress. Recent advancements in generative AI have introduced a suite of novel embedding models. This development necessitates a rigorous evaluation of these models using domain-specific validation tasks. We show that on the interference task OpenAI embeddings significantly outperform S-BERT. We are in the process of incorporating the novel embeddings throughout the paper.

2.1. Data

We use the full text of all US utility patents issued 1836–2022. For historical patents issued 1836–1975, we rely on the Patents Core database by ProQuest. These are digitized

Text of patent 1
Text of patent 2
Text of patent 3

Step 1: Numerical representations in three idea spaces $(C_i(m_k), k \in \{A, B, C\})$

Repr. A

$$\begin{bmatrix} 0.44 & 0.03 & 0.55 & 0.44 \\ 0.42 & 0.33 & 0.2 & 0.62 \\ 0.3 & 0.27 & 0.62 & 0.53 \end{bmatrix}$$

Repr. B

$$\begin{bmatrix} 0.13 & 0.51 & 0.18 \\ 0.85 & 0.49 & 0.85 \\ 0.51 & 0.07 & 0.43 \end{bmatrix}$$

Repr. C

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Step 2: Measurements of pairwise similarities $(Sim^{m_k}(p_i, p_j))$

Repr. A

Pat 1	Pat 2	Cos. Sim.
1	2	0.82
1	3	0.94
2	3	0.87

Repr. B

Pat 1	Pat 2	Cos. Sim.
1	2	0.71
1	3	0.48
2	3	0.96

Repr. C

Pat 1	Pat 2	Cos. Sim.
1	2	1
1	3	0
2	3	0

Step 3: Validation-based selection $\left(V^l(m_k), l \in \{(i), (ii), (iii)\}\right)$

Task (i)

Repr.	Perf.	Rank
Repr. A	0.91	1
Repr. B	0.87	2
Repr. C	0.84	3
Baseline	0.51	4

Task (ii)

Repr.	Perf.	Rank
Repr. A	0.46	1
Repr. B	0.23	2
Repr. C	0.18	3
Baseline	0.03	4

Task (iii)

Repr.	Perf.	Rank
Repr. A	0.85	2
Repr. B	0.93	1
Repr. C	0.73	3
Baseline	0.05	4

Step 4: Compute downstream measure based on the best representation:

E.g., "Breakthrough" patents $(q^m(p_i))$ (Kelly et al., 2021) or average patent pair similarity $(q^m(p_i, p_j))$ by year (this paper).

Figure 2: Overview of the NLP pipeline

patent text using ProQuest's proprietary methods. For modern patents issued 1976–2022, we use full-text patent data from PatentsView (U.S. Patent and Trademark Office, 2023). From the same source, we also use patent metadata, including patent classifications. We also use the text of modern patent applications, historical and modern data on patent interferences, and human-annotated data. These data are described as they are used in later sections.

2.2. Representation: Mapping patents to idea space

We denote a representation of patent text p_i to a location in idea space as:

$$m(p_i) = C_i^m \tag{1}$$

where m refers to the particular method or model used to map the patent to a location in idea space and C_i^m refers to the coordinate vector for patent i based on method m.

Many different methods have been used to map patents to idea space. A traditional approach uses patent classifications (e.g., Jaffe, 1986; Jaffe et al., 1993). Classifications are assigned by patent examiners who are specialized subject-matter experts. They are primarily administrative tools designed to facilitate searches for relevant prior art. Currently, the USPTO uses the Cooperative Patent Classification (CPC), which is divided into nine top-level sections (e.g., "human necessities," "textiles," or "electricity").[2]

In our framework, one could define a class-based mapping of patents to idea space in which each patent is represented by a vector with 1s in the position of its assigned class(es) and 0s in all other positions. Representation C in Figure 2 has the broad features of a class-based representation. A limit of such class-based mappings is their coarse granularity. Implicitly, with this representation, all patents in the same class are equally close to each other and all patents which are not in the same class are equally far apart.

[2]These are further subdivided into 129 classes (e.g., "inorganic chemistry," "manufacture of fertilizers," "semiconductor devices") and 250,000 classifications overall. The USPTO periodically re-classifies all patents to the current CPC. We use the current CPC classification as of February 2023.

More recent NLP models use patent texts for their mappings to idea space.[3] The first and second representations in Figure 2 have the broad features of a text-based representations.

For example, Kelly et al. (2021) use a variant of the traditional NLP method TF-IDF. TF-IDF is based on word counting. The TF-IDF coordinate vector for patent i will be a vector of length K, where the entry for each unique word k is:

$$c_{i,k}^{TFIDF} \equiv TF_{i,k} \cdot IDF_{i,k} \tag{2}$$

The first term, "Term Frequency," is $TF_{i,k} \equiv n_{i,k}/\sum_j n_{i,j}$, or the number of times $n_{i,k}$ word k appears in patent i divided by the total number of words in patent i. The second term, "Inverse Document Frequency," is $IDF_{i,k} \equiv log(\frac{\text{\# of patents in corpus}}{\text{\# of patents in corpus with word k}})$. Intuitively, two patents are similar in TF-IDF idea space if they share many of the same words, especially if those shared words are otherwise rare.

Despite its early origins in the 1970s, TF-IDF continues to be popular due to its transparency, ease of use, and satisfactory performance in many cases. However, recently there has been explosive growth in the use of neural networks and deep learning trained on large corpora of text (e.g., Wikipedia). These models produce vector representations (called "word embeddings" or "document embeddings"). They differ from earlier approaches in that they (i) capture the meaning of individual words within context and (ii) distribute each word's meaning across an entire word or document vector (hence "distributed representations").[4] These models "learn" that different words are similar when they are frequently used in the same context; for example, word2vec uses information from the surrounding five to 10 words.

A priori, one might expect that newer models will outperform word-counting methods due to their ability to capture similarity in meaning in a broader, contextual sense. Ultimately, this is an empirical question best answered through validation-based selection.

[3]See Gentzkow et al. (2019); Grimmer et al. (2022); Bochkay et al. (2023) for reviews of the use of NLP methods in economics and neighboring disciplines.

[4]See Smith (2020) for an introduction to the evolution of word representations in NLP.

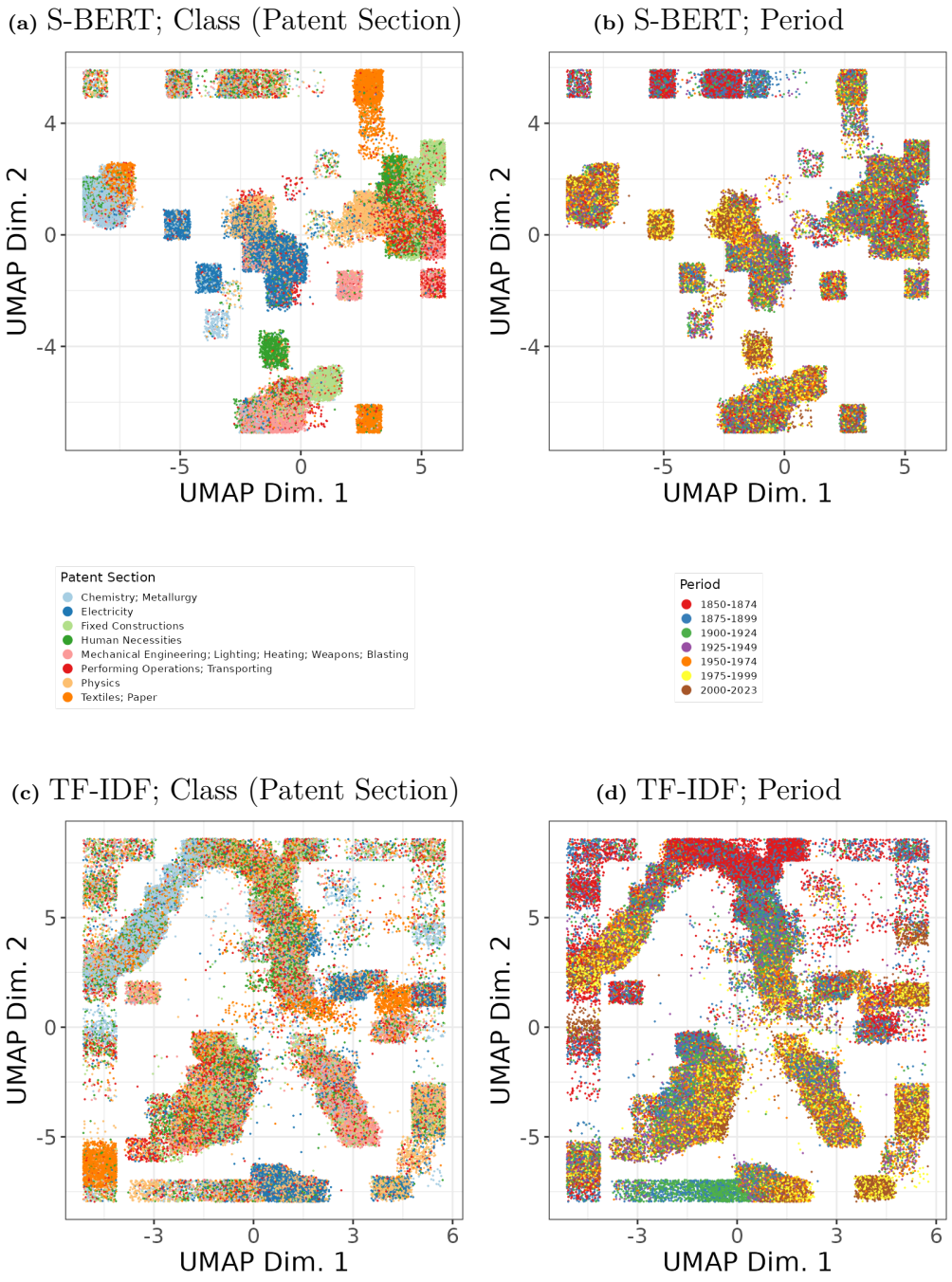


Figure 3: Uniform Manifold Approximation and Projection (UMAP) plots for S-BERT and TF-IDF representations

Notes: The plot is based on a sample of 111,251 patents stratified by patent class (USPTO Section) and quarter-century period. To constrain extreme values, the data were winsorized at the 5% and 95% levels along both axes.

To develop intuition, Figure 3 displays two-dimensional projections of high-dimensional idea spaces based on (panels a, b) S-BERT and (c, d) TF-IDF representations (see details in Appendix A.) Each point represents a patent and colors denote (a, c) different CPC top-level sections or (b, d) quarter-century time periods. Visually, the different methods appear to represent idea space quite differently. We expand on this comparison below.

2.3. Measuring concepts from patent text

Having obtained a text representation, a researcher can proceed to define a measure of interest. A core measure of interest is pairwise similarity. Following common NLP practice, define the similarity between patent texts p_i and p_j as the cosine similarity between their vector representations (based on representations m) in idea space.[5]

$$Sim^m(p_i, p_j) \equiv \frac{C_i^m \cdot C_j^m}{||C_i^m||||C_j^m||}. \tag{3}$$

Clearly, different choices of mappings m may result in different measures of pairwise similarity (see Figure 2, Step 2). A mapping based on patent classifications yields a measure of similarity that reflects the number of shared classes. A mapping based on TF-IDF yields a measure of similarity that reflects shared word usage, especially if those words are otherwise rare. Mappings based on distributed representations, such as S-BERT, yield measures of similarity that reflect shared semantic meaning in the sense of the specific model.

A number of other measures from prior literature can also be written as vector functions of patent representations. For example, Kelly et al. (2021) develop a clever measure of the "importance" of patents, and study patterns in the number of *breakthrough* patents—those that are in the upper tail of importance. In our notation, their measure of importance is:

$$q^m(p_i) \equiv \frac{\sum_{j \in \mathcal{F}} Sim^m(p_i, p_j)/|\mathcal{F}|}{\sum_{k \in \mathcal{B}} Sim^m(p_i, p_j)/|\mathcal{B}|} \tag{4}$$

[5]A natural alternative, Euclidean distance, has the undesirable property of depending on document length when used with word-counting-based methods such as TF-IDF (Grimmer et al., 2022).

where $\mathcal{F}$ denotes the set of patents published in the 5 years prior to patent i, $\mathcal{B}$ denotes the set of patents published in the 5 years after patent j, and $|\mathcal{B}|$ is the number of patents in set $\mathcal{B}$. Thus, importance is the ratio of "forward similarity," the similarity of patent i to subsequent patents, to "backwards similarity," the similarity of patent i to preceding patents.

Our framework makes clear that the choice of a conceptual measure of interest is distinct from the choice of a mapping of patents to idea space. Kelly et al. (2021) use a variant of TF-IDF as their mapping m. Different choices for mappings m may result in different measures of importance and may select different patents as breakthroughs.

2.4. Validation-based selection

Given a concept, how should a researcher choose between alternative representations? Our third step is validation-based selection. Validation requires external measures which have a clear theoretical connection with the concept of interest ("ground truth"), perhaps for a sub-sample of the data. For us, these are external assessments of patent similarity.

Formally, given a concept of interest c, a representation-specific measurement f^{m_i}, and a score function S that quantifies the correspondence between concept of interest and measurement, validation evaluates each representation m_i to select the best mapping.

$$V(m_i) = S\left(\{f^{m_i}(\mathbf{p}), c(\mathbf{p})\}\right) \tag{5}$$

Ideally, if the concept of interest were available for all observations $\mathbf{p}$, then there would be no need for validation; we could simply use observed $c(\mathbf{p})$. We resort to NLP when direct measures of c are unavailable. Often, though, we have access to various ground truths—external measures that align with the concept of interest for certain observations, perhaps imperfectly.[6] If no external measures exist, annotation is often a viable alternative.

[6] "Ground truth" sometimes suggests uniqueness and absolute precision and accuracy. Instead, we use "ground truth" to signify an imperfectly-accurate representation that may be imprecise or non-unique.

Validation-based selection implements an empirically-feasible version of equation 5:

$$V^j(m_i) = S^j\left(\{f^{m_i}(\mathbf{p^j}), g^j(\mathbf{p^j})\} \mid \mathbf{p^j}\right), \tag{6}$$

where $g^j(\mathbf{p^j})$ is a ground truth, $f^{m_i}(\mathbf{p^j})$ is a measurement based on representation m_i, and the score S^j quantifies the correspondence level between measures derived from the representation m_i and ground truth. Measurement and ground truth share the same domain $\mathbf{p^j}$, which is typically small and may not be randomly selected from the population $\mathbf{p}$. Ground truth must be theoretically aligned with the concept of interest; multiple ground truths j correspond to different validation tasks. S^j could be correlation or mean squared error. The functional form of S^j may vary across different validation tasks j.

$V^j(m_i)$ is then a score or ranking for each model m_i of the set $\{m_1, m_2, \ldots, m_n\}$. This identifies the best representation according to validation criterion j. If different validations suggest different representations as optimal, a final decision should use human judgment regarding the relative importance of each validation task.

For concreteness, we can apply this framework to our analysis. In the interference validation task, $\mathbf{p^j}$ represents the set of patent applications in interference cases. The ground truth function $g(\mathbf{p^j})$ creates pairwise combinations of these applications and produces a Boolean vector whose entries are 1 if the corresponding pair was in an interference case. This approach is theoretically grounded on the premise that subject-matter experts (i.e., examiners) can identify application pairs with a high degree of similarity. The function $f^{m_i}(\mathbf{p^j})$ computes pairwise similarities based on the representation m_i. The score function S^j is the Receiver Operating Characteristic Area Under the Curve (ROC AUC) or the Precision-Recall Area Under Curve (PR AUC) score using $f^{m_i}(\mathbf{p^j})$ as the signal and $g(\mathbf{p^j})$ as the label.

A second ground truth is patent classifications. To continue developing intuition, return to Figure 3, panels (a) and (c). S-BERT representations clearly produce a mapping in which patents in the same CPC top-level class cluster together, as indicated by color. In contrast,

TF-IDF representations present less clear clustering by class. This is visual evidence that S-BERT provides a better representation of idea space than TF-IDF, at least according to this particular validation task. Section 3.2 quantifies this result more precisely.

Intriguingly, other features of this visualization besides patent classifications accord with external human judgment. In Figure 3a, the blue square near (-5, 0) contains many semiconductor patents (in the Electricity top-level class). This cluster is positioned between a light-blue square of materials science patents (Chemistry and Metallurgy) to its left and a broader Electricity cluster to its right. This relative positioning does not score S-BERT any points on the patent class validation task, but it accords with external knowledge that semiconductor innovations combine materials science and electricity. This result also highlights that a single ground truth is unlikely to perfectly capture patent similarity. Therefore, multiple validation tasks, as we employ in Section 3, are desirable.

Note that we use the term "validation" differently from some prior literature. For us, validation is integrated with model selection. This domain-specific validation-based selection is common in many fields but less prevalent in economics (Grimmer et al., 2022).[7] In contrast, a common usage is to "validate" an NLP measure by correlating with an external ground truth. For example, our interference validation could evaluate only a single representation chosen *ex ante* (e.g., TF-IDF). We could then compare TF-IDF's score to a random-guess benchmark. This is a low bar for concept validity and model selection: TF-IDF easily passes. In our framework, this is implicitly a form of validation-based selection that simply reaffirms the initial choice of representation. However, this approach fails to recognize that *all* of our candidate models easily pass, even though there are very large performance differences.

[7] Ash and Hansen (2023) provide examples outside of innovation economics where different representations lead to different conclusions. In economics, a similar concept to validation-based selection is econometric model selection. There, various models might be evaluated using common criteria such as the Akaike Information Criterion (AIC). Validation-based selection also parallels model selection based on out-of-sample testing in Machine Learning. There, m_i is one model from a set of candidates, $\mathbf{p^j}$ is the sample not used for model training, $g(\mathbf{p^j})$ are the realized values of the dependent variable, and $f^{m_i}(\mathbf{p^j})$ are the predicted values. The score function could be, e.g., root mean squared error.

3. Validation Task Results

3.1. Interferences

Our first validation task uses patent interferences. Patent interferences were USPTO administrative proceedings that decided priority of invention when two or more independent parties claimed to have invented the same thing at the same time. An interference was suggested by a specialized patent examiner when, during their search for relevant prior art, they encountered at least two US patent applications contained the "same patentable invention" (37 CFR §1.601).[8] Thus, patent interferences represent *expert judgment* that two independent patent applications contain identical legal claims.

3.1.1. Data

We select patent applications from a database of 215 interference cases decided 1998–2014.[9] These decisions were publicly available through the USPTO's "e-FOIA Reading Room" and encoded by Ganguli et al. (2020). Each interference case involves two or more independent *parties* with competing, simultaneous claims to the same patentable invention. Each party has one or more patent *applications* corresponding to the content of the interference. In our database of 215 cases, we identify 440 distinct patent applications. Using these interference cases and applications, we construct 96,580 ($= \frac{1}{2}(440^2 - 440)$) *application pairs*. Of these application pairs, we identify 322 *interfering pairs*—meaning two patent applications from independent (opposing) parties that make overlapping claims of invention.

We represent the text of each application using patent class, TF-IDF, doc2vec, USE, and S-BERT. For every pair of application representations, we compute their cosine similarity.

[8]Patent interferences owe their existence to the unique "first to invent" rule that prevailed in the US until 2013. Under "first to invent," the inventor who conceived and reduced to practice first was awarded patent protection. This contrasts with the "first to file" system that prevails in the US today and in much of the rest of the world, where the patent is awarded to the first inventor to file a patent application.

[9]This is a subset of the 1,329 interference cases used in (Ganguli et al., 2020). Our present sample is limited to interference cases where (i) we observe the number of distinct claims in interference and (ii) we observe applications for both parties. Interference claims data are coded from decisions; some decisions omit this information. Patent application data are available only after March 15, 2001.

Table 1: Example rows from the patent pair dataset used for interference validation

		Cosine similarity based on:					
ID App. 1	ID App. 2	Class	TF-IDF	doc2vec	USE	S-BERT	Int.
12714708	13775784	0.00	0.09	0.75	0.46	0.52	0
10054638	10739610	0.25	0.01	0.91	0.41	0.58	0
10388111	10461256	0.00	0.00	0.38	0.06	0.11	0
10278437	10923413	0.00	0.02	0.76	0.37	0.40	0
12714205	12714708	1.00	0.48	0.83	0.65	0.94	1

Notes: Columns show patent IDs, similarity scores from different patent representations, and a binary label indicating whether this pair was part of an interference case.

Table 1 shows an excerpt of the resulting application-pair database. Each row is a unique pair. Columns are application identifiers, similarity scores, and a true interference indicator.

3.1.2. Evaluation

Next, we evaluate the performance of alternative representations. The task is to classify application pairs in interference. Evaluating such a binary classification problem is surprisingly complex. We highlight these complexities by considering a hypothetical scenario where a patent examiner wants to identify application pairs that are likely to be in interference.

First, the examiner cares about (i) how often a classifier correctly classifies true interfering pairs (true positives), (ii) how often a classifier incorrectly classifies non-interfering pairs as interfering pairs (false positives), and (iii) the relative weights on these two criteria.[10] Here, true positives represent, pre-2014, the statutory obligation of the USPTO to determine priority of invention, while false positives incur investigation costs through examiner and paralegal time. Thus, there may be a trade-off between classifiers which detect many true positives but also many false positives, and those which detect fewer true positives but also fewer false positives. This can be formalized as the tradeoff between *recall* and *precision*. The *recall* of a classifier is the share of total interferences it correctly identifies. The *precision*

[10]Formally, this is referred to as cost-sensitive classification (Elkan, 2001)).

Table 2: Rankings based on threshold-based metrics

(a) Separate F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	S-BERT	154	97	0.55
2	TF-IDF	117	70	0.47
3	USE	90	67	0.38
4	doc2vec	52	79	0.23
5	Class	111	899	0.17

(b) Separate F10-max. thresh.

Rank	Repr.	TP	FP	F10
1	S-BERT	285	3,231	0.83
2	TF-IDF	275	4,369	0.77
3	USE	266	5,447	0.73
4	Class	232	6,942	0.61
5	doc2vec	243	24,498	0.44

(c) TP fixed at TF-IDF's F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	S-BERT	117	55	0.48
2	TF-IDF	117	70	0.47
3	USE	117	191	0.38
4	Class	117	972	0.17
5	doc2vec	117	3,399	0.06

(d) TP fixed at S-BERT's F1-max. thresh.

Rank	Repr.	TP	FP	F1
1	S-BERT	154	97	0.55
2	TF-IDF	154	252	0.43
3	USE	154	615	0.28
4	Class	154	1,636	0.15
5	doc2vec	154	7,992	0.04

Notes: F1/F10 scores and underlying true positives and false positives with a different thresholding strategy in each panel. The total number of patents is 440; the total number of patent pairs is 96,580; the total number of interference cases is 312.

of a classifier is the share of pairs correctly classified as interfering.

Second, note that many different classifiers can be built from a given similarity measure. The natural way to use the similarity measures to build a classifier is to pick a threshold level of the similarity score, and classify those pairs above the threshold as interfering pairs and those below as not. Different threshold levels will lead to classifiers with different performance in terms of selecting true positives and false positives.

As a starting point, consider the case where the examiner values identifying promising cases (recall) and not overburdening staff (precision) equally. We consider how the different similarity measures perform when the threshold for each classifier is chosen to maximize the so-called F1 score, which weights precision and recall equally. The results in Table 2a show that S-BERT has the highest F1 score, followed by TF-IDF, at each measure's F1-maximizing threshold. However, while S-BERT results in more true positives, it also results in more false positives than TF-IDF.

Next, consider the case where the examiner is more concerned with identifying potential interferences (true positives) than wasted effort (false positives)—i.e., staff time is relatively cheap. The F10 score weights recall ten times more than precision. Panel (b) shows that at each measure's F10-maximizing threshold, S-BERT retrieves around 3% more true positives than TF-IDF, while reducing false positives by a remarkable 24%. The higher true-positive rate surfaces more high-likelihood interferences, while the smaller false-positive rate reduces unnecessary investigations by almost a third.

Another way to compare classifier performance is to fix the number of true positives at one model's F1-maximizing level, and compare the number of false positives. To do so we first select a threshold for each measure which yields the same number of true positives as at the TF-IDF F1-optimal level. Panel (c) shows that S-BERT can achieve this number of true positives while selecting 21% fewer false positives. Next, we select a threshold for each measure which yields the same number of true positives as at the S-BERT F1-optimal level. Panel (d) shows that S-BERT can achieve this higher number of true positives while selecting 62% fewer false positives than TF-IDF.

Notably, on these threshold-based evaluations S-BERT and TF-IDF significantly outperform the classifiers based on the other two NLP methods, USE and doc2vec. We also report the performance of a classifier based on the number of shared CPC classes between application pairs, which consistently lags behind S-BERT and TF-IDF.

We next report results based on two metrics which summarize classifier performance across all possible thresholds. Receiver Operating Characteristic Area Under the Curve (ROC AUC) evaluates the trade-off between true positive and false positive rates across all possible thresholds. Precision-Recall Area Under Curve (PR AUC) measures the trade-off between precision and recall across all possible thresholds.[11]

Across both ROC AUC and PR AUC, we find that S-BERT best predicts interference cases, followed by TF-IDF, and then the other models (Table 3). The PR AUC differences

[11]See Davis and Goadrich 2006 for a comparison of the two measures.

Table 3: Rankings based on non-threshold-based metrics

(a) ROC AUC

Rank	Repr.	ROC AUC
1	S-BERT	0.99
2	TF-IDF	0.98
3	USE	0.97
4	Class	0.85
5	doc2vec	0.84

(b) PR AUC

Rank	Repr.	PR AUC
1	S-BERT	0.51
2	TF-IDF	0.43
3	USE	0.34
4	Class	0.20
5	doc2vec	0.15

Notes: ROC and PR AUC scores for different patent text representations on predicting interference cases.

are more pronounced, as expected for an imbalanced binary prediction problem.

Across threshold- and non-threshold-based comparisons, classifiers based on S-BERT consistently uncover a higher number of true positives at substantially lower false positive costs than the other NLP methods. The performance gain is economically significant—a patent examiner upgrading to a S-BERT-based classifier from an alternative methods could substantially reduce costs. TF-IDF is a clear second-best, ahead of USE and doc2vec.

3.1.3. OpenAI embeddings (`text-embeddings-large3`) and other recent (as of February 2024) embeddings

Table 4 presents preliminary results for more recent embedding models. We selected embedding models based on their performance on the generalist Massive Text Embedding Benchmark (MTEB) leaderboard as of February 2024[12]. Among the best-performing models on MTEB, we select OpenAI's `text-embedding-3-large`, Voyage AI's `voyage-code-2` embeddings, and `UAE-Large-V1` embeddings introduced by Li and Li (2023).

The `OpenAI` model performs exceptionally well. At the F-10 maximizing threshold, the `OpenAI` model reduces false positives compared with both `S-BERT` and `TF-IDF` (1,118 versus 3,001 and 5,306, respectively, with a similar number of true positives). This suggests that the `OpenAI` embeddings outperform `S-BERT` to a similar extent as `S-BERT` surpasses `TF-IDF`.

[12]Created by Muennighoff et al. (2023), available at https://huggingface.co/spaces/mteb/leaderboard.

Table 4: Rankings of recent embeddings models

(a) PR AUC

Rank	Repr.	PR AUC
1	OpenAI	0.62
2	Voyage	0.59
3	S-BERT	0.52
4	Angle	0.51
5	TF-IDF	0.44
6	USE	0.36
7	Class	0.21
8	doc2vec	0.16

(b) Separate F10-max. thresh.

Rank	Repr.	TP	FP	F10
1	OpenAI	255	1,118	0.89
2	Voyage AI	254	1,665	0.87
3	Angle	244	1,701	0.83
4	S-BERT	250	3,001	0.82
5	TF-IDF	253	5,306	0.77
6	USE	235	4,984	0.72
7	Class	209	6,255	0.62
8	doc2vec	198	17,944	0.44

Notes: PR AUC and F10 scores for different patent text representations on predicting interference cases.

We are currently focused on fully incorporating these new embeddings into our paper.

3.2. Patent Class and Time

We classify patents into 56 section-by-time-period groups according to (i) eight top-level CPC technology sections and (ii) date of issue in seven quarter-century periods from 1850 to 2022. We draw random samples of 200 patents from each group, yielding 11,200 total patents and 62,714,400 unique pairs of patents. For each pair, we create indicators for common section or common time period. As in the prior task, we then evaluate the performance of similarity scores based on TF-IDF and S-BERT representations at classifying patent pairs as belonging to the same class or the same period.

Figure 4 shows that S-BERT representations better predict shared CPC sections among patent pairs, while TF-IDF representations better predict shared date of issue. (This more precisely characterizes the results in Figure 3.) Following its superior performance in the interference task, we view this as another validation that S-BERT better represents similarity in idea space, even if patents were issued in different eras. As we discuss in Section 5, TF-IDF's basis in counting words makes it sensitive to changing word usage. Compared with S-BERT, patent pairs issued in different periods are farther away in TF-IDF's idea space.

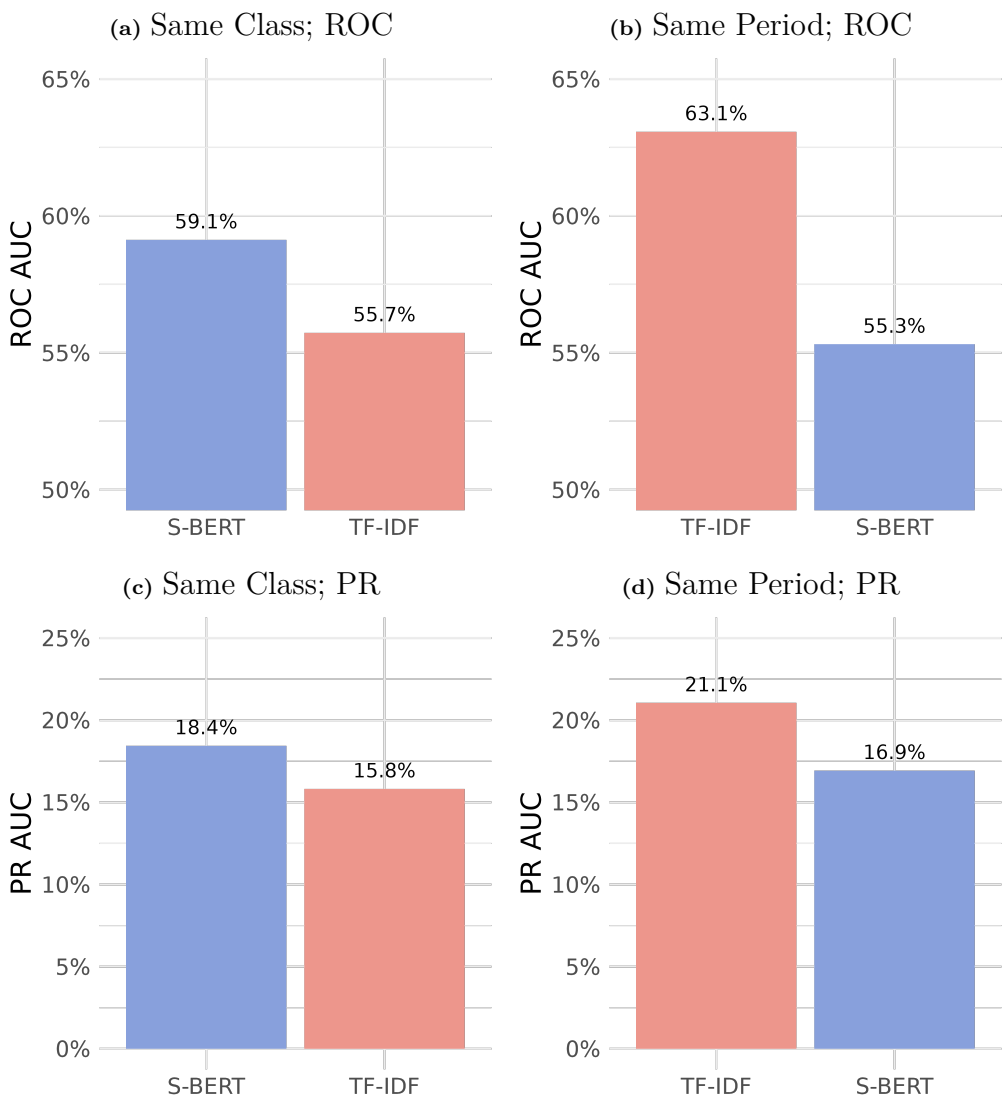


Figure 4: Representation performance on same class and same period classification task

3.3. Nonexpert human judgment

We design and implement a non-expert human validation task. This task complements the interference validation task, as it draws patents from a different time period (1880-1920) and focuses on patents which are of a moderate level of similarity (versus nearly identical).

A main challenge is that humans without special training struggle to place objects on absolute scales (Carlson and Montgomery, 2017). Therefore, we asked five research assistants (RAs) to make *relative* judgments of similarity for the same 100 patent triples (see Table 5

for an example).[13] We presented them with a Focal Patent and asked them to select either Patent 1 or Patent 2 as more similar to the Focal Patent.[14] RAs were told to use any criteria they wished. They were not required to understand the technical details of inventions, but they were allowed to do web searches for technical terms if they wished.

In initial pilots, humans had trouble assessing the relative similarity of patents that were extremely dissimilar. Through iterative exploration on patent triples not used for final annotation, we found that the 75th percentile of similarity across all patent pairs was similar enough for human perception. We required both comparison patents to have at least that level of similarity to the Focal Patent, for both S-BERT and TF-IDF representations.

To increase power compared with random sampling, we only sampled triples for which S-BERT and TF-IDF *disagreed* on the relative similarity between the Focal Patent and Patent 1 or Patent 2. Therefore, both models deemed Patents 1 and 2 as somewhat similar to the Focal Patent, but they disagreed on which is more similar. Annotators were tasked with resolving this disagreement. Additionally, we chose to present annotators with a concise extract from the summary paragraph[15] and the first 120 characters of the claims section.

3.3.1. Results

Table 6 shows results. We estimate a regression with a dependent variable indicator that the S-BERT (vs. TF-IDF) representation of Patent 1 is more similar to the Focal Patent. The main explanatory variable is an indicator that the annotator chose Patent 1 as more similar to the Focal Patent, i.e., they agreed with S-BERT. The sum of the estimated intercept and coefficient is equal to the share of patent triples for which the annotator agreed with S-BERT. Overall, the five annotators agreed with S-BERT 71% of the time and TF-

[13]These RAs were not involved in the project in any other capacity. This task was inspired by Carlson and Montgomery (2017)

[14]We also allowed them to select "Can't choose." The observations annotated as "Can't choose" were excluded from the analysis.

[15]The text between "improvement(s) in" and the next "." are present in the majority of patents.

Table 5: Example annotation task for the human validation

Focal Text	Text 1	Text 2
IMPR: improvements in music-scales CLAIMS: 1. A scale device for a music-perforation spacer, comprising strips, each having imprinted thereon a signature, and a plurali	IMPR: Improvements in the Mounting of Piano-Keys CLAIMS: The combination of a piano-key having a solid top, with a balance-pin firmly affixed in the under side of said key and projec	IMPR: Improvement in Pianofortes CLAIMS: , and desire to secure by Let-5 ters Patent of the United States, is- In combination in a piano, the frame, the double suppor

Notes: Annotators were presented with 100 triples of texts and were asked whether Text 1 or Text 2 is more similar to the Focal Text. Triples were selected so that S-BERT and TF-IDF representations disagree about this relative similarity. The texts preserve errors in optical character recognition and parsing that were presented to human annotators.

IDF 29% of the time.[16] Each of the five annotators tended to select the patents preferred by S-BERT with the rate ranging from 67% to 75%. The regression estimates suggest we can reject the null hypothesis that S-BERT is no better compared with TF-IDF at aligning with human judgment. These findings therefore suggest that S-BERT also better captures a non-expert human sense of similarity for historical patents.

4. Validation Matters for Downstream Economic Measurement

We find that the average similarity of patents has been declining since at least the early 20th century. This robust finding is evident only when idea space is represented by S-BERT, the model that performed best across all three validation tasks described above. In contrast, representations based on TF-IDF suggest that similarity has been increasing over time, and are much less robust across different corpora. The results from the above validation tests are thus essential for interpreting these results.

In a companion paper (Ganguli et al., 2024), we develop a theory where the decline in contemporaneous invention similarity is related to recent findings on long-run invention

[16]There was moderate consensus across annotators: 68% for label 1, 58% for label 2, and 26% for "Can't choose." A moderate level of agreement is expected given the open-ended nature of the similarity assessment criteria.

Table 6: Annotator agreement with S-BERT vs. TF-IDF in the human validation task.

	Dep. Var.: S-BERT=1					
	Pooled	Ann.1	Ann.2	Ann.3	Ann.4	Ann.5
(Intercept)	0.19***	0.26***	0.14	0.19**	0.09	0.21*
	(0.03)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
Choice=1	0.52***	0.45***	0.52***	0.55***	0.64***	0.50***
	(0.04)	(0.09)	(0.09)	(0.09)	(0.11)	(0.11)
R^2	0.27	0.20	0.26	0.30	0.40	0.24
Adj. R^2	0.27	0.19	0.25	0.30	0.39	0.23
Num. obs.	402	91	92	93	56	70

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Notes: Annotators were presented with 100 triplets of texts and were asked whether the Focal Text is more similar to Text 1 or Text 2. Each of the texts is preferred by a similarity metric based on either TF-IDF or S-BERT. In the regression, the dependent variable, S-BERT=1, is an indicator equal to 1 when S-BERT indicates that Text 1 is more similar to the focal text than Text 2 (which is preferred by TF-IDF). The independent variable, Choice=1, is an indicator variable equal to one when the annotator chooses Text 1 as more similar to the Focal Text. The first column is based on the pooled data of the 5 annotators and the rest of the columns are based on the data of individual annotators. All annotators were given the same 100 triples. Sample sizes differ because the annotators had an option not to choose when they were unable to do so and those annotations were excluded.

dynamics, including the increasing "burden of knowledge" (Jones, 2009), increasing R&D spending (Hirschey et al., 2012), declining R&D productivity (Bloom et al., 2020), and constant R&D spillovers (Lucking et al., 2019). As the increasing burden of knowledge raises the fixed costs of inventing, inventors "spread out" over an expanding knowledge frontier. In this model, ideas get "harder to find" (Bloom et al., 2020) because inventors respond to the expansion of idea space by spreading out to avoid competition. In turn, this makes invention harder because there are weaker positive knowledge spillovers from "neighbors" that are now more distant in idea space. Inventors increase their own R&D inputs in response to weaker spillovers, reducing own-R&D productivity. (On net, total spillovers may be roughly constant as increasing idea distance is offset by increases in own-R&D investment.)

4.1. Declining invention similarity using S-BERT

We measure invention similarity over time using S-BERT. We use S-BERT representations for every issued patent, 1836–2022, split into four components: abstracts, claims, descriptions, and titles. The abstract is a short synopsis of the patented invention, usually

about a paragraph.[17] Claims are a precise set of numbered statements that define the scope of invention and determine its legal boundaries. The description is the remaining text that often includes background information and a description of prior art, and includes abstracts prior to 1976. A patent title is typically around 3–5 words long.

Figure 1a shows average pairwise similarity based on S-BERT. Each corpora exhibits declining similarity. The decline in description similarity spans the entire range. The decline in the similarity of claims and titles dates to at least the early 20th century. Abstracts have declined steadily in similarity since 1976, when they are first parsed in our data.

Some minor features of Figure 1 deserve comment. The number of issued patents per year increases over time, which accounts for reduced volatility over time. Post-1976 patent text does not rely on OCR, which could account for the modest structural breaks seen in series for claims and descriptions. Finally, there are a number of indicators that suggest undocumented discrete changes in how ProQuest processed the OCR outputs for titles of patents issued after 1919. This could account for the drop in title similarity after that year.

A related measure of interest is the size of the knowledge space. How "wide" is the frontier of knowledge? One measure of this might be the convex hull of the vector representations of all patents issued in one year. Unfortunately, the large number of patents in a given year and the high dimensionality of S-BERT vectors makes this exercise computationally infeasible.

To make progress, we first use principal component analysis to reduce the dimensionality of S-BERT vectors to seven principal components. Then, we use the quickhull algorithm to compute the volume of the convex hull containing the principal components from that year.

Figure 5 shows the results. The lower-dimension space spanned by the seven principal components has generally and steadily increased in size over time. By using an alternative measure, this result provides a check on the invention similarity results presented in Figure 1. Moreover, it adds a geometric intuition to the similarity results: the size of the "knowledge frontier" has expanded over time. Taken together with the results on invention similarity,

[17] Abstracts are separately parsed in our data only in 1976 and later.

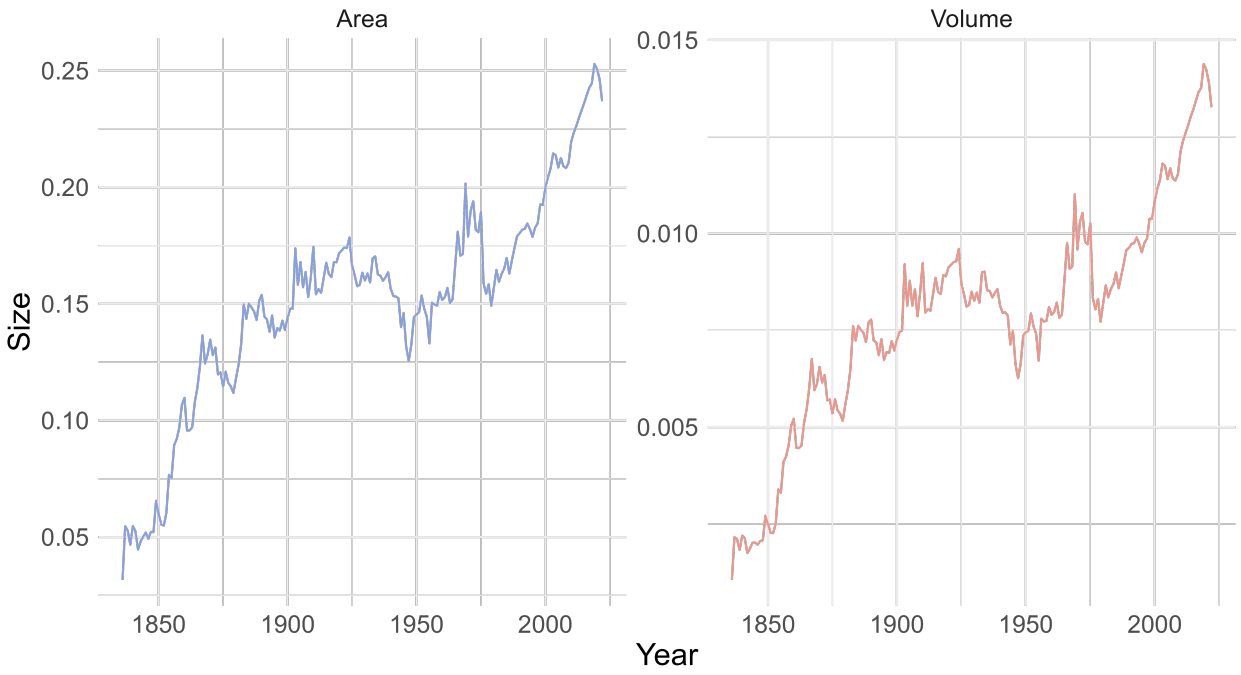


Figure 5: Size of convex hull of 7 principal components of S-BERT vectors by year

inventors appear to be "spreading out" on an expanding knowledge frontier.

4.2. Measurement depends on representation

S-BERT and TF-IDF produce significantly different pictures of the evolution of average similarity over time. Figure 1b shows average pairwise similarity based on TF-IDF. TF-IDF delivers results which are (i) often opposite S-BERT and (ii) inconsistent across corpora. Abstract and claim similarity has been increasing over much of observed history. In contrast, descriptions increase in similarity in the 19th century and decline in the 20th century.

These results emphasize the importance of validation. Without validation, researchers would have little guidance on which "stylized facts" to trust. The validation results increase confidence in our conclusion that average patent similarity has declined over the long run.

To further demonstrate that measurement depends on representation, we revisited the analysis of "breakthrough" patents by Kelly et al. (2021). (See details in Appendix B.) Overall, our analysis confirms the Kelly et al. (2021) finding that the rate of breakthrough inventions is higher today compared with prior decades. That said, the choice of representa-

tion still matters. Compared with the TF-BIDF representations used by Kelly et al. (2021), S-BERT-based measures suggest that the recent increase in breakthrough inventions is less unusual compared with historical patterns. Moreover, S-BERT-based measures appear to be more robust and less sensitive to decisions about how to process and residualize the data.

4.3. Declines within and across patent technology classifications

Figure 6 shows that both within-class and between-class similarity have been generally declining over time. We group patents into each of the 129 CPC classifications. We then calculate average pairwise similarity for patents in the same class and for those in different classes, returning to our validated S-BERT based representations and using claims as our corpus. Over most of the period both within-class and between-class similarity have been declining (with the exception being a sharp increase in the late 19th century in within-class similarity). Both of these margins of decline would be missed by more traditional approaches using only classifications to measure similarity. Those measures implicitly treat all patents in a class as equally similar and all patents in different classes as equally dissimilar.

4.4. Declines in Interferences

In this section, we validate the S-BERT results on declining similarity by constructing a time series of interference rates over 150 years. This is an out-of-sample reproduction of the finding of declining patent similarity over time. While interferences were used to validate S-BERT-based measures of similarity, only applications in interference cases post-1998 were used, while this section documents trends in interference rates over 150 years.

We estimate the annual rate of interferences per issued patent. This is approximately the probability that an issued patent was involved in an interference. We combine four different sources of data: (i) a database of interferences from newly-digitized *Registers of Interferences* from the National Archives 1864–1901; (ii) summary statistics on patent interferences from 1950–1962 (Di Simone et al., 1963) and (iii) 1980–1993 (Calvert and Sofocleous, 1982, 1986, 1989, 1992, 1995) and (iv) a database of patent interference decisions from 1998–2014.

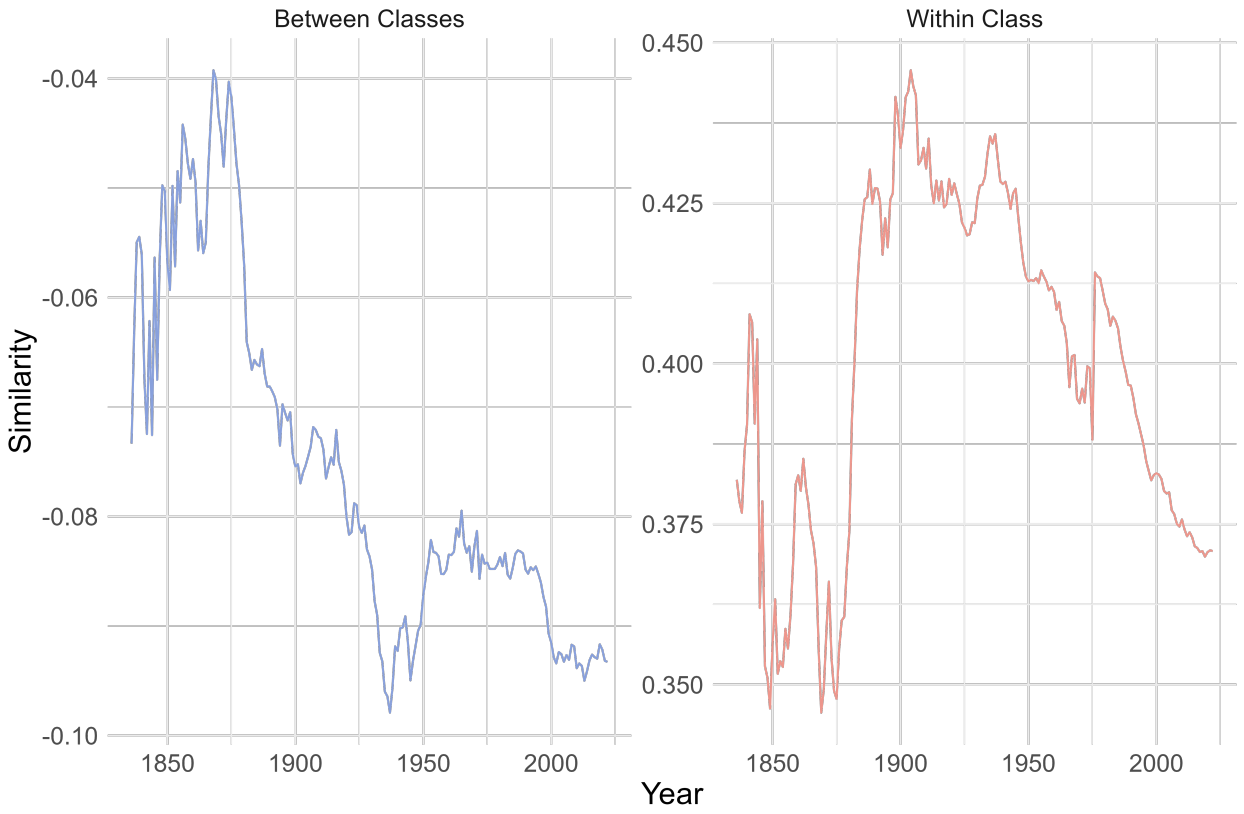


Figure 6: Between versus within class similarity over time

Notes: Similarity within and between technology classes, using S-BERT representations and 129 CPC classes. Annual averages 1836–2022 shown.

First, we used purpose-digitized data from the *Registers of Interferences* in the USPTO Records of the National Archives. We scanned and digitized 19,388 interference cases in 21 volumes of the Registers that spanned 1860–1908. (Appendix D.1 provides an example.) The registers are organized chronologically by hearing date. We recorded the decision or termination dates for each case, and total the number of cases terminated in each year. Based on the ranges of hearing dates, the early (1860–1863) and late (1902–1908) volumes appear to represent only partial years, so we use only years 1864–1901. On average, there were 504 interferences terminated in each year during this period 1864–1901.

Second, we used summary statistics reported by Di Simone et al. (1963), Table 1.2, on interferences terminated by year, 1950–1962. Only summary statistics—no case-level data—were reported. Technically, these statistics were reported for each fiscal year versus calendar

year. In these data, on average, 640 interferences were terminated in each year 1950–1962.

Third, we used summary statistics reported by Calvert and Sofocleous (1982, 1986, 1989, 1992, 1995) of totals interferences terminated in 3-year periods. (Again, these totals are for fiscal years.) On average, 237 interferences were terminated yearly 1980–1994.

Finally, we used interference case decisions 1998–2014 issued by the Board of Patent Interferences and encoded from the USPTO eFOIA site by Ganguli et al. (2020). Based on these data, on average, 76 interferences were terminated annually during this period.[18]

Next, we estimate the rate of interference, that is, the probability that an issued patent interfered with another application.[19] For each year, we divide the number of total interferences by the number of total patents issued. Figure 7 shows a steady decline in the rate of interference in the 150 years between 1864–2014. Based on the Registers data, the average rate of interference over 1864–1901 was 2.71%. Based on Di Simone et al. (1963), the average rate of interference over 1950–1962 was 1.43%. Based on Calvert and Sofocleous (1982) et al., the average rate of interference over 1980–1994 was 0.30% Finally, based on the eFOIA decisions, the average rate of interference over 1998–2014 was 0.05%. Thus, the decline in interference rate is consistent with measured declines in similarity according to S-BERT-based representations.

5. Why is S-BERT Better?

In this section, we further explore the performance differences between S-BERT and TF-IDF. First, we compare a 21st-century bicycle patent and a 19th-century velocipede patent to illustrate S-BERT's ability to identify semantic similarities. Second, we examine unigram frequencies in the Google Books Ngram database. Unigrams characteristic of patent pairs

[18]It is likely that this slightly undercounts the actual number of interferences, since some interferences were terminated before they reached the Board of Patent Interferences. We can determine this based on interference numbers, which are assigned sequentially. Among the case decisions issued between 1998–2014, we can thus infer that there were 2,403 cases declared between 1991 and 2014, or about 104 interferences declared per year.

[19]A limitation of this exercise is that dates of interference termination and patent issuance likely lag behind the dates of actual invention and application. Typical lags might be up to 2 years.

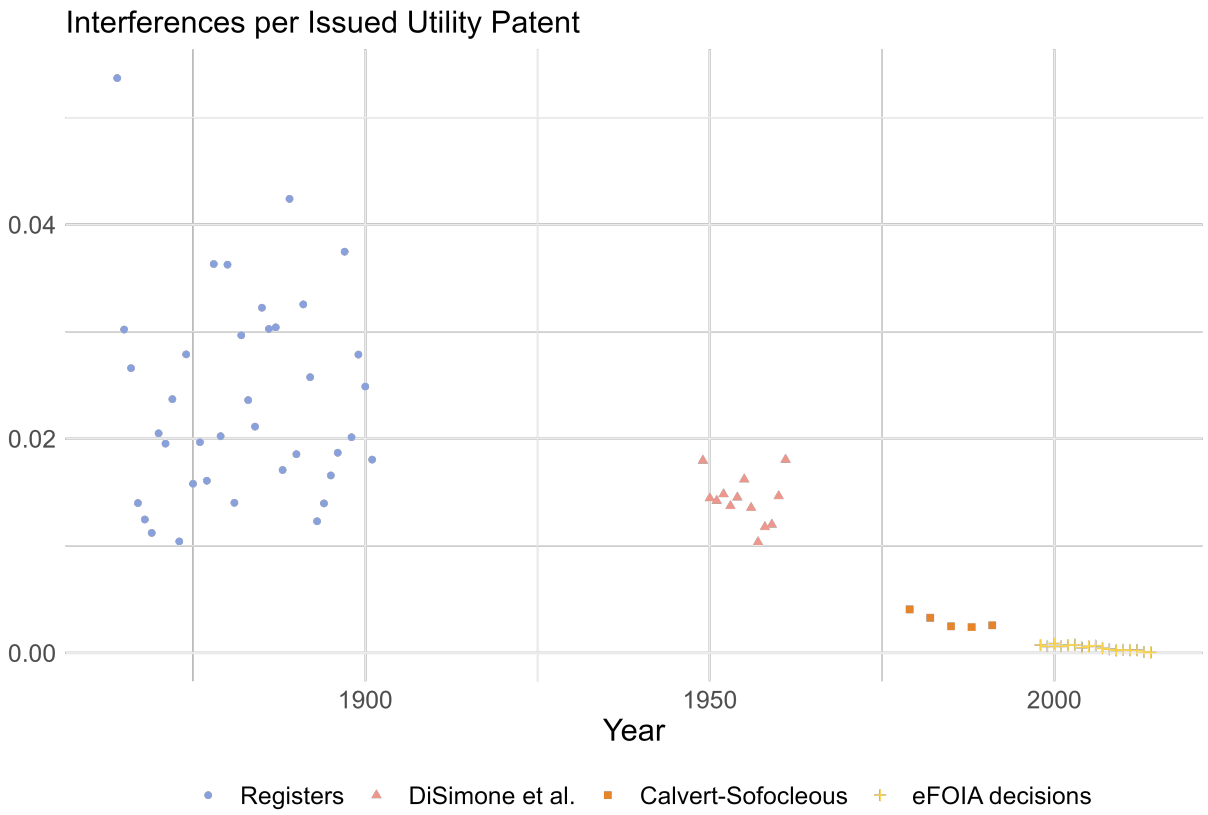


Figure 7: Interferences per issued patent

with high TF-IDF similarity overweight period-specific language, thus explaining its effectiveness in period- but not class-based tasks. Appendix C presents details of the characteristic unigram methodology, an additional Google Books Ngram analysis, and a synonym-based analysis that further highlights S-BERT's ability to capture semantic similarity.

5.1. Example: Bicycle versus velocipede

Figure 8 shows a bicycle patent from the 21st century and a velocipede patent from the 19th century. Despite these patents originating from different time periods and employing distinct terminologies, S-BERT successfully identifies them as similar, positioning them in the 87th percentile of similarity. At the same time, the similarity according to TF-IDF is 0. This example illustrates the S-BERT's ability to capture semantic nuances and contextual similarities despite changes in language.

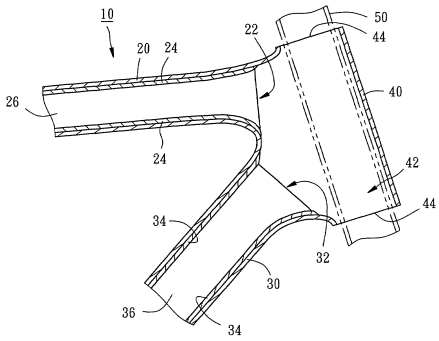
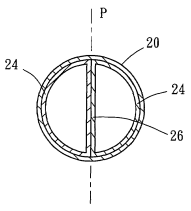
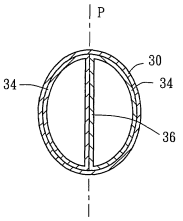
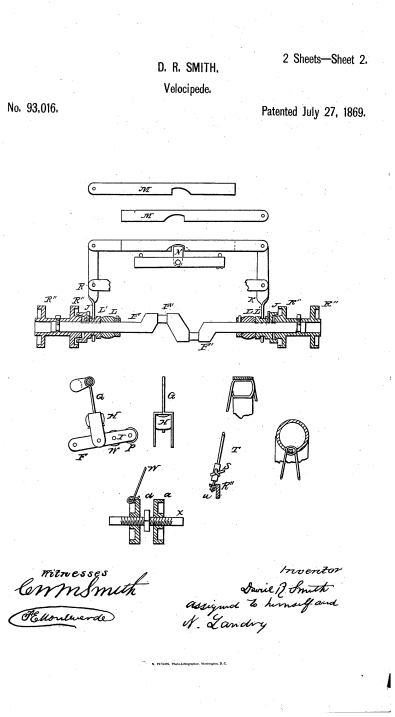
Patent 1: US7562890B2 (2009)	**Patent 2: US93016A (1869)**
Front frame for a bicycle. 1. A front frame for a bicycle, comprising: two first inner tubes abutted together; two second inner tubes abutted together; an upper tube of cured multiple layers of fiber reinforced rein material wound around the two first inner tubes so that there is no crack between the upper tube and ...	IMPROVED VELOCIPEDE. In the velocipede as constructed, and in combination therewith, the friction-clutch, spurs, arms, cross-bar, cam, guide-wheel, with hollow rim and axle, arranged and operated substantially as described. In witness whereof, I have hereunto set my hand and seal.
 FIG. 4  FIG. 5  FIG. 6	

Figure 8: A conceptually similar pair of patents from different time periods

Notes: Velocipede is a type of bicycle. The text is truncated to the title and the beginning of the claims section of the patents. Typos due to OCR were fixed for this illustrative example. According to S-BERT, these patents are in the 87th percentile of similarity, whereas according to TF-IDF, the similarity is 0.

Both patents introduce improvements in the design or function of two-wheeled vehicles. A velocipede is an archaic term for a type of bicycle. Although Patent 1 focuses on the "front frame for a bicycle" while Patent 2 is more broadly about an "improved velocipede," they both involve common mechanical features such as tubes, frames, and axles. However, the patents do not share many common terms. Patent 1 talks about "front frame," "inner tubes," "upper tube," while Patent 2 mentions "friction-clutch," "spurs," "arms," etc.

S-BERT takes into account not just specific words, but also the context in which these words appear. Words with similar meaning that frequently appear in similar contexts will be assigned similar S-BERT vectors. Thus, S-BERT representations reflect that both patents are about two-wheeled vehicles, even if they use different terms. S-BERT is trained on a diverse dataset, which includes technical language. It can therefore encode terms like "frame," "tubes," and "axle" as related in general, even if they appear in different contexts.

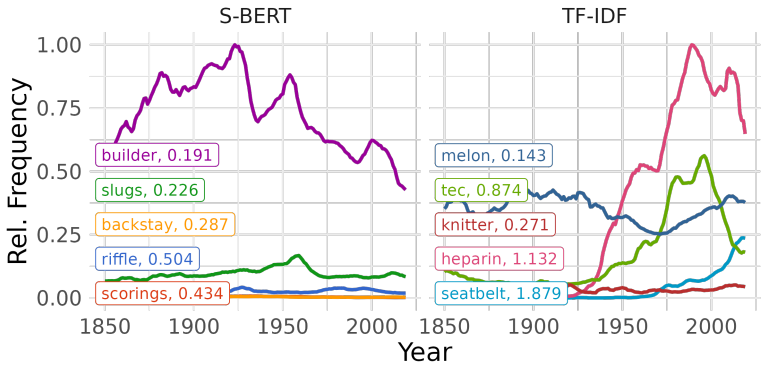
TF-IDF is a simpler bag-of-words model that does not capture meaning in the same way (see Smith, 2020). It considers only the frequency of individual words in each document and in the corpus as a whole. TF-IDF treats distinct terms such as "bicycle" and "velocipede" as unrelated concepts. In sum, S-BERT is able to better capture the semantic and contextual similarities between these two patents that describe similar inventions but do not share a common vocabulary.

5.2. TF-IDF overweights period-specific words versus universal synonyms

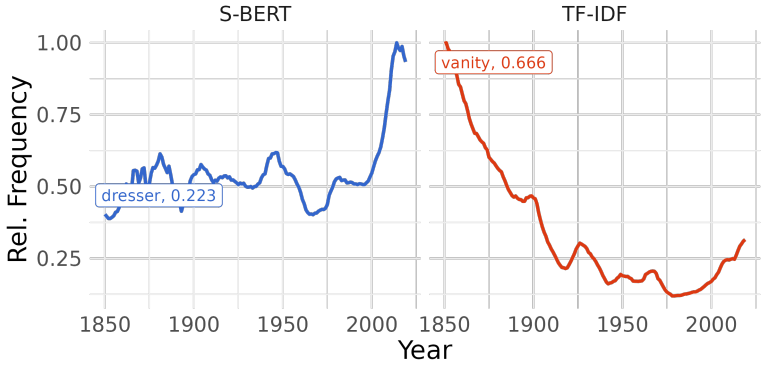
The bicycle/velocipede example suggests that TF-IDF overweights period-specific terms like velocipede, leading it to assign low similarity to pairs that might describe the same idea with different terms. Here we extend that analysis. We hypothesize that terms used in patent pairs assigned high similarity by TF-IDF should have a higher variance of usage over time. These period-specific terms might be archaic or modern, or they may have irregular fluctuations in usage.

We use the Google Books Ngram database. We identify characteristic tokens that differentiate patent pairs based on their similarity scores. Our analysis categorizes patent pairs

(a) Top-5 characteristic unigrams for each representation



(b) Hand-picked example 1



(c) Hand-picked example 2

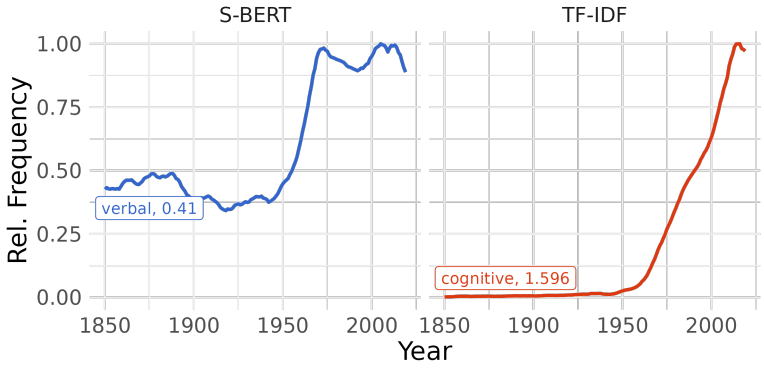


Figure 9: Frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

Notes: The plot is based on the Google Ngram Corpus (1850–2019). Frequency is normalized to the largest frequency on each plot. The number after the unigram label is the coefficient of variation, defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe et al. (2017) algorithm.

into three groups: (i) those identified as similar by both S-BERT and TF-IDF, (ii) those recognized as similar only by S-BERT, and (iii) those recognized as similar only by TF-IDF. We exclude pairs with mutual agreement between models and determine characteristic unigrams for the latter two categories. Appendix C details the methodology and presents additional analyzes.

Figure 9 presents some illustrative examples of unigram frequencies over time. Among the top-five most characteristic unigrams, TF-IDF unigrams are more volatile, which indicates more time-specific word usage.

We further hand-picked examples of conceptually-similar words in panel (b). "Dresser," characteristic of S-BERT similar pairs, exhibits moderate use with little variation until the 2000s. In contrast, "vanity," characteristic of TF-IDF similar pairs, exhibits more volatility, steadily dropping in usage throughout the period between 1850 and 1970, followed by a small rise. Another example is shown in panel (c). "Verbal" and "cognitive" both increase after 1950. But the increase is more dramatic for "cognitive," and therefore this term characteristic of TF-IDF similar pairs has a larger coefficient of variation.

6. Conclusion

In this paper, we developed a pipeline for the construction, validation, and selection of measures of economic interest derived from patent text. Innovation economists should pay attention to the choice of text representation, since different choices can significantly affect conceptual validity and the results of subsequent economic analyzes.

The construction of similarity and other measures based on patent text can be separated into three distinct steps: representation, measurement, and validation. The first step maps each patent to a location in idea space; the second step measures a concept of economic interest using representations produced by each of several candidate models; and the third step validates these representations using purpose-built, domain-specific tasks to select the best mapping.

We designed three novel, domain-specific validation tasks that compare the performance of four leading and widely-used NLP models. Each task uses a sample of patent pairs with human judgments of similarity. We then assessed how well different representations agree with human judgment. Our validation results suggest that S-BERT produces measures of patent similarity that more closely match human judgment compared with other leading NLP models.

Finally, we constructed validated measures of invention similarity for US utility patents issued 1836—2022. S-BERT-based estimates of patent similarity show a secular decline in invention similarity. In contrast, measures based on TF-IDF show ambiguous or diverging patterns. The rate of patent interference also exhibits a secular decline, reproducing the S-BERT result.

Our results are useful to many applications in the economics of science and innovation. Our publicly-available S-BERT vector representations of the text of every US issued patent 1836—2022 can be used by researchers to compute patent similarity or other downstream measures. These can be used in many applications, such as for constructing matched controls in studies of localized knowledge spillovers, or for empirical implementations of theories that involve the space of ideas. Our approach can also be used to develop new measures of the space of ideas using other types of text, such as scientific papers.

References

Akcigit, U., W. R. Kerr, and T. Nicholas (2017): "The mechanics of endogenous innovation and growth: Evidence from historical US patents," https://economics.harvard.edu/files/economics/files/kerr-william_mechanics_of_endogenous_innovation_patents_sbbi-2-3-17_0.pdf.

Arts, S., B. Cassiman, and J. C. Gomez (2018): "Text matching to measure patent similarity," *Strategic Management Journal*, 39, 62–84.

Arts, S., J. Hou, and J. C. Gomez (2021): "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, 50, 104144.

Ash, E. and S. Hansen (2023): "Text algorithms in economics," *Annual Review of Economics*, 15, 659–688.

Azoulay, P., C. Fons-Rosen, and J. S. Graff Zivin (2019): "Does science advance one funeral at a time?" *American Economic Review*, 109, 2889–2920.

Berkes, E. and R. Gaetani (2020): "The geography of unconventional innovation," *Economic Journal*, 131, 1466–1514.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020): "Are ideas getting harder to find?" *American Economic Review*, 110, 1104–1144.

Bochkay, K., S. V. Brown, A. J. Leone, and J. W. Tucker (2023): "Textual analysis in accounting: What's next?" *Contemporary accounting research*, 40, 765–805.

Calvert, I. A. and M. Sofocleous (1982): "Three years of interference statistics," *Journal of the Patent Office Society*, 64, 699.

——— (1986): "Interference statistics for fiscal years 1983 to 1985," *Journal of the Patent & Trademark Office Society*, 68, 385.

——— (1989): "Interference statistics for fiscal years 1986 to 1988," *Journal of the Patent & Trademark Office Society*, 71, 399.

——— (1992): "Interference statistics for fiscal years 1989 to 1991," *Journal of the Patent & Trademark Office Society*, 74, 822.

——— (1995): "Interference statistics for fiscal years 1992 to 1994," *Journal of the Patent & Trademark Office Society*, 77, 417.

Carlson, David. and J. M. Montgomery (2017): "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts," *American Political Science Review*, 111, 835–843.

Carmody, S. (2023): *Ngramr: Retrieve and Plot Google n-Gram Data.*

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil (2018): "Universal sentence encoder for English," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium: Association for Computational Linguistics, 169–174.

Cheng, Z., D. Lee, and P. Tambe (2022): "InnoVAE: Generative AI for understanding patents and innovation," http://dx.doi.org/10.2139/ssrn.3868599.

Clancy, M. S. (2018): "Inventing by combining pre-existing technologies: Patent evidence on learning and fishing out," *Research Policy*, 47, 252–265.

Dasgupta, P. and E. Maskin (1987): "The simple economics of research portfolios," *Economic Journal*, 97, 581–595.

Davis, J. and M. Goadrich (2006): "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, New York, NY, USA: Association for Computing Machinery, ICML '06, 233–240.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019): "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 4171–4186.

Di Simone, D. V., J. B. Gambell, and C. F. Gareau (1963): "Characteristics of interference practice," *Journal of the Patent Office Society*, 45, 503–591.

Elkan, C. (2001): "The foundations of cost-sensitive learning," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence—Volume 2*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., IJCAI'01, 973–978.

Feng, S. (2020): "The proximity of ideas: An analysis of patent text using machine learning," *PLOS ONE*, 15, 1–19.

Fleming, L. (2001): "Recombinant uncertainty in technological search," *Management Science*, 47, 117–132.

Ganguli, I., J. Lin, V. Meursault, and N. Reynolds (2024): "Declining invention similarity: Theory, implications, and evidence," working paper.

Ganguli, I., J. Lin, and N. Reynolds (2020): "The paper trail of knowledge spillovers: Evidence from patent interferences," *American Economic Journal: Applied Economics*, 12, 278–302.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as data," *Journal of Economic Literature*, 57, 535–574.

Griliches, Z. (1979): "Issues in assessing the contribution of research and development to productivity growth," *Bell Journal of Economics*, 10, 92–116.

Grimmer, J., M. Roberts, and B. Stewart (2022): *Text as Data: A New Framework for Machine Learning and the Social Sciences*, Princeton University Press.

Hirschey, M., H. Skiba, and M. B. Wintoki (2012): "The size, concentration and evolution of corporate R&D spending in US firms from 1976 to 2010: Evidence and implications," *Journal of Corporate Finance*, 18, 496–518.

Jaffe, A. B. (1986): "Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits, and market value," *American Economic Review*, 76, 984–1001.

Jaffe, A. B., M. Trajtenberg, and R. Henderson (1993): "Geographic localization of knowledge spillovers as evidenced by patent citations," *Quarterly Journal of Economics*, 108, 577–598.

Jones, B. F. (2009): "The burden of knowledge and the "death of the renaissance man": Is innovation getting harder?" *Review of Economic Studies*, 76, 283–317.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021): "Measuring technological innovation over the long run," *American Economic Review: Insights*, 3, 303–20.

Le, Q. and T. Mikolov (2014): "Distributed representations of sentences and documents," in *Proceedings of the 31st International Conference on Machine Learning*, ed. by E. P. Xing and T. Jebara, Beijing, China: PMLR, vol. 32 of *Proceedings of Machine Learning Research*, 1188–1196.

Lee, J. and J. Hsiang (2019): "PatentBERT: Patent classification with fine-tuning a pre-trained BERT Model," http://arxiv.org/abs/1906.02124.

Li, X. and J. Li (2023): "AnglE-optimized text embeddings," https://doi.org/10.48550/arXiv.2309.12871.

Lucking, B., N. Bloom, and J. Van Reenen (2019): "Have R&D spillovers declined in the 21st century?" *Fiscal Studies*, 40, 561–590.

Merton, R. K. (1957): "Priorities in scientific discovery: A chapter in the sociology of science," *American Sociological Review*, 22, 635–659.

——— (1973): *The Sociology of Science: Theoretical and Empirical Investigations*, University of Chicago press.

Mikolov, T., K. Chen, G. S. Corrado, and J. Dean (2013): "Efficient estimation of word representations in vector space," in *International Conference on Learning Representations*.

Miller, G. A. (1992): "WordNet: A lexical database for English," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Monroe, B. L., M. P. Colaresi, and K. M. Quinn (2017): "Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict," *Political Analysis*, 16, 372–403.

Muennighoff, N., N. Tazi, L. Magne, and N. Reimers (2023): "MTEB: Massive text embedding benchmark," https://doi.org/10.48550/arXiv.2210.07316.

Murata, Y., R. Nakajima, R. Okamoto, and R. Tamura (2014): "Localized knowledge spillovers and patent citations: A distance-based approach," *Review of Economics and Statistics*, 96, 967–985.

Park, M., E. Leahey, and R. J. Funk (2023): "Papers and patents are becoming less disruptive over time," *Nature*, 613, 138–144.

Reimers, N. and I. Gurevych (2019): "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Conference on Empirical Methods in Natural Language Processing*.

Schnoebelen, T., J. Silge, and A. Hayes (2022): *Tidylo: Weighted Tidy Log Odds Ratio*.

Smith, N. A. (2020): "Contextual word representations: Putting words into computers," *Communications of the ACM*, 63, 66–74.

Sparck Jones, K. (1972): "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, 28, 11–21.

Thompson, P. and M. Fox-Kean (2005): "Patent citations and the geography of knowledge spillovers: A reassessment," *American Economic Review*, 95, 450–460.

U.S. Patent and Trademark Office (2023): "Data download tables," PatentsView, accessed February 9, 2023.

Wang, J., R. Veugelers, and P. Stephan (2017): "Bias against novelty in science: A cautionary tale for users of bibliometric indicators," *Research Policy*, 46, 1416–1436.

Appendix A. Visualization of Embedding Spaces

This section describes the process we followed to generate the visualizations discussed in Section 2.

The raw data are obtained using the same sampling strategy outlined in the class and period validation section (3.2). This strategy involves sampling patents from specified classes as categorized by the USPTO, across distinct 25-year periods ranging from 1850 to 2023.

We then plot 2-dimensional projections of the embedding spaces, where individual patents are marked with color according to their respective class or period. This visualization technique provides a geometrically intuitive perspective of the innovation space. It also lays a visual foundation for comparing the efficacy of different embedding techniques like S-BERT and TF-IDF.

Appendix A.1. Methodology

The primary method we employ for visualization is dimensionality reduction through the Uniform Manifold Approximation and Projection (UMAP) technique. UMAP is noted for its ability to preserve both global and local structures during reduction, making it, roughly speaking, a non-linear variant of Principal Component Analysis (PCA).

To speed up the computation, we conduct the initial dimension reduction using PCA, which reduces the dimensionality of the S-BERT and TF-IDF representations to 50. Subsequently, UMAP is applied to these reduced representations. This two-step process harnesses the computational efficiency of PCA while benefiting from the geometric qualities of UMAP.

We manually tuned UMAP hyperparameters to achieve a more clustered representation that looked more like an "archipelago" than a singular "continent." This tuning aids in better visual separation among clusters within the innovation space.

Appendix A.2. Plotting

One of the challenges encountered during visualization was the overlapping of data points, especially in dense clusters. To mitigate this, a jittering technique was employed which

disperses each point slightly within its local neighborhood to reduce overlap, hence enhancing the visibility of individual clusters. The jittering results in a boxier scatter plot, which is a compromise for better clarity.

The plots (refer to Figure 3) primarily serve as illustrative tools, providing a more tangible notion of the idea space. We use color coding to denote different patent classes and 25-year periods in both S-BERT and TF-IDF projections. Despite the inherent distortions, some observations could hint at underlying structural differences between the representations.

At first glance, it's clear how the representations reflect the class and period structure. S-BERT representations show clearer class boundaries compared to TF-IDF representations, suggesting that patent clustering is closer to the class structure. On the other hand, TF-IDF periods seem less mixed compared to S-BERT periods, although this difference is more subtle. These visual patterns match the results we discussed in Section 3.2, where we evaluated how well the representations classify patent pairs into the same class and same period categories. This consistency between visual observations and analytical findings is encouraging.

It is harder to draw conclusions from the general layout because of the distortions inherent in the projection project. However, some observations stand out. For example, TF-IDF has more "dust" compared to S-BERT, which has more of an "empty space." Also, the extended x and y tails in TF-IDF, hidden due to winsorizing, hint at a possible trend where variability in expressing similar ideas with different words pushes these representations farther from the core.

Lastly, we explored the clusters qualitatively using an interactive tool. While we don't expect every aspect of patent positions to be interpretable, some interesting observations came to light. For instance, in Panel A of Figure 3, a blue square around (-5, 0), representing the electricity class, contains many semiconductor patents. This square sits between the light blue square on its left representing materials science patents (Chemistry and Metallurgy class) and a more general blue electricity patent cluster on its right. Although such observations are anecdotal, they help build trust in the model, especially when supported by

more rigorous analyzes. Such qualitative insights, alongside quantitative evaluations, enrich our understanding of the embedding spaces and their ability to capture the complex nature of innovation.

The visualizations provide insight into how different representations can result in meaningfully different similarity measures, highlighting the importance of making grounded choices in representations when studying innovation.

Appendix B. Robustness of S-BERT representations

In this section, we provide additional evidence that economic measurement depends on representation by performing a robustness analysis of Kelly et al. (2021).

Kelly et al. (2021) use a backward-looking variant of the traditional NLP method TF-IDF, which they call TF-BIDF ("B" for backward). They measure "breakthrough" patents as those that are dissimilar to past patents but very similar to future patents. Intuitively, using only the backward-looking corpus to measure (inverse) document frequency avoids penalizing especially influential patents that introduce terms that are widely-used in subsequent patents.

Here, we briefly summarize some key steps in the Kelly et al. (2021) pipeline. First, create representations of patent texts using TF-BIDF. This is the "breakthrough" or "importance" measure. Second, residualize this breakthrough measure on year fixed effects, so that importance is measured relative to the average issued patent in each year. Third, identify the all-time top 10% of patents in the residualized breakthrough measure. Finally, plot the rate of breakthrough patents normalized by total US population in each year. The result is their Figure 4, Panel A, reproduced in Figure B.10 below.

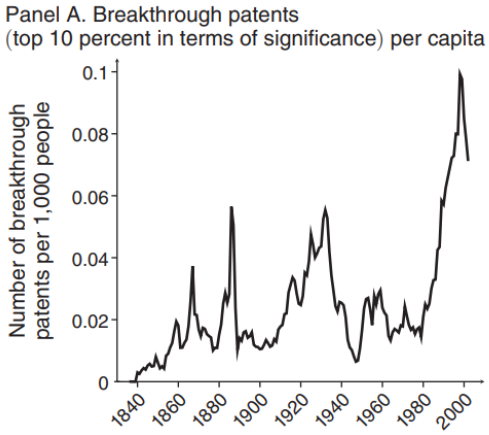


Figure B.10: Reproduction of Kelly et al. (2021), Figure 4, Panel A

We explore the sensitivity of their results along several dimensions. The most important and relevant for the current study is (i) using S-BERT versus TF-BIDF for the representations. We also explore robustness to two other key decisions in the Kelly et al. (2021) pipeline: (ii) residualizing the breakthrough measure on year fixed effects, (iii) normalizing the rate of breakthrough patents by total US population.

Figure B.11, Panel A shows our replication of the Kelly et al. (2021) result. There are two primary differences in our replication. One, we are using a different source corpus—our source is ProQuest database of claims versus Kelly et al. (2021)'s Google Patents digitized text. Two, for computational reasons we simplify the computation of the backward-looking IDFs to the prior five calendar years. Kelly et al. (2021) instead compute a backward IDF for each patent up to five years prior to the date of issue. Thus, there are slight differences in our replication methodology.

Overall, comparing Figure B.11, Panel A with Figure B.10, we are able to closely replicate the Kelly et al. (2021) result. The qualitative dynamics are very similar, with fluctuations in the rate (per US population) of breakthrough patents, followed by a sharp increase starting around 1980. The overall correlation coefficient between the two series is 0.729.

Figure B.11, Panel B shows the number of breakthrough patents by year, without normalizing by US population. There is a secular increase in the number of breakthrough patents

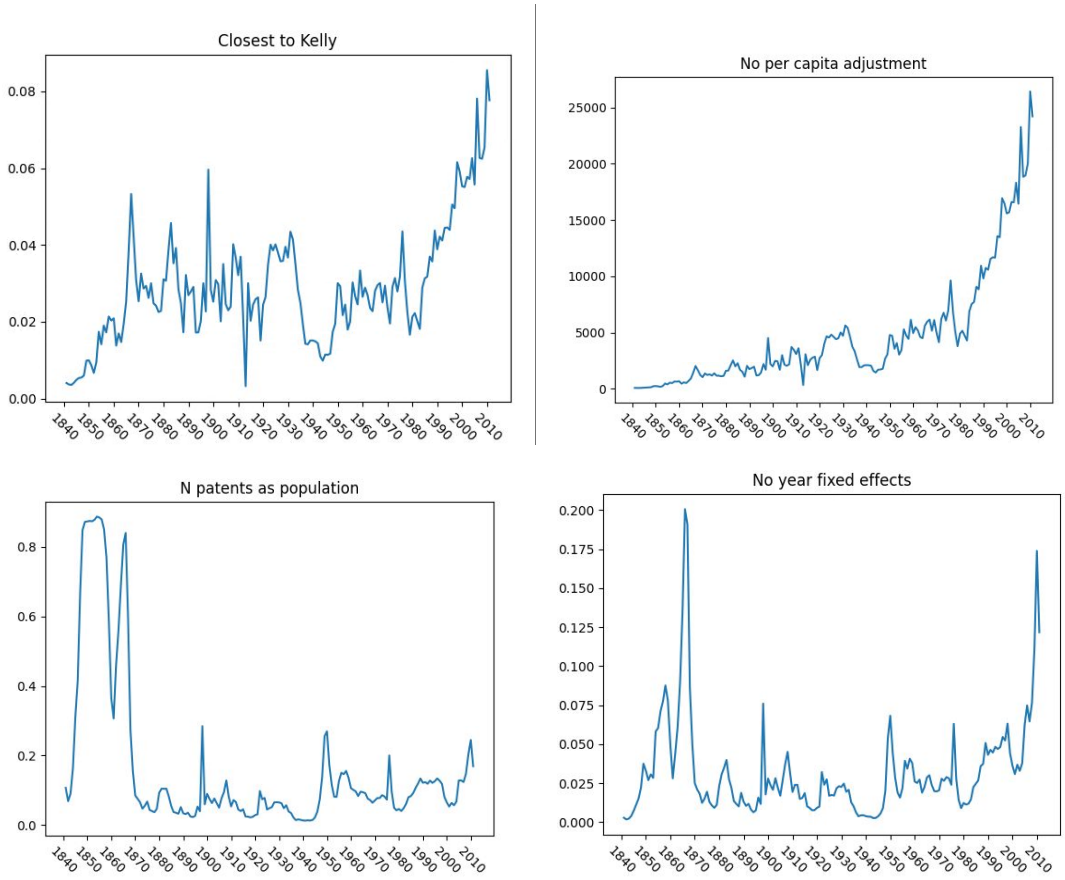


Figure B.11: Replication and robustness of (Kelly et al., 2021) using TF-BIDF representations

over time.

Figure B.11, Panel C shows an alternative normalization, using total issued patents by year instead of total US population by year. Unlike the first two panels, the choice of normalization is consequential: while there were more breakthrough patents in recent years, as a share of total issued patents, the peak rate of breakthrough patents was before 1870.

Finally, Figure B.11, Panel D shows the effect of residualizing the breakthrough measure on year fixed effects. While there is still an increase in the rate of breakthrough patents since 1980, there was also a similarly high rate of breakthrough patents in the 1860s.

Figure B.12, Panel A shows robustness of the baseline Kelly et al. (2021) result to using S-BERT representations versus TF-BIDF. Qualitatively, there are some similar dynamic features. The rate of breakthrough patents appeared to increase significantly after 1980.

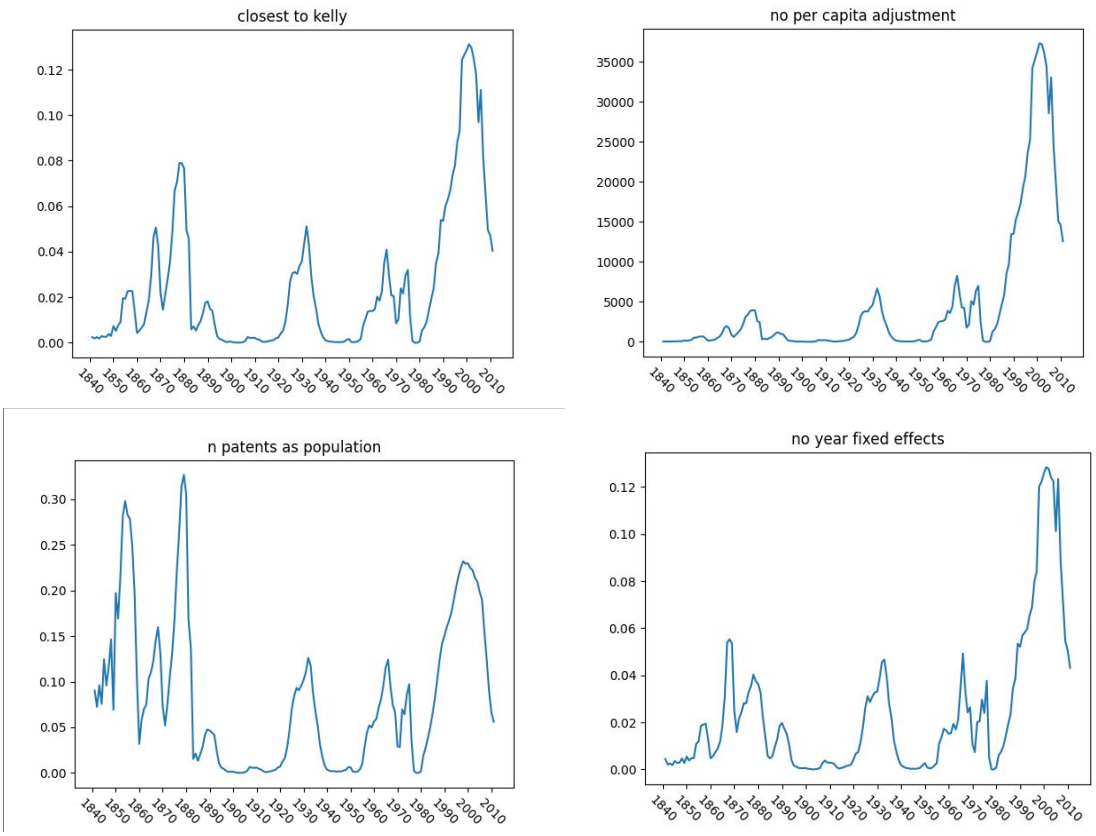


Figure B.12: Robustness of (Kelly et al., 2021) using S-BERT representations

However, there were similar, although more modest in magnitude, booms in the rate of breakthrough patents in the 1870s, 1930s, and 1960s. Thus, the recent increase in the rate of breakthrough patents appears less unusual compared with historical episodes versus the TF-BIDF results. Overall, the correlation with the TF-BIDF-based measure is 0.577.

Panel B shows similar dynamics compared with TF-BIDF-based measures in the total number of breakthrough patents. Patent C shows similar sensitivity to the choice of normalization.

Finally, Figure B.12, Panel D shows that the choice to residualize on year fixed effects is less consequential using S-BERT representations versus TF-BIDF representations. The comparison of Panels A and D in this figure implies similar trends in breakthrough patents. In contrast, the comparison of Panels A and D in Figure B.11 implies different trends in breakthrough patents.

Overall, our analysis confirms the Kelly et al. (2021) finding that the rate of breakthrough inventions is higher today compared with prior decades. That said, the choice of representation matters for measurement. Compared with TF-BIDF, S-BERT-based measures suggest that the recent increase in breakthrough inventions is less unusual compared with historical patterns. Moreover, S-BERT-based measures appear to be more robust and less sensitive to decisions about how to process and residualize the data.

Appendix C. Why is S-BERT better?

In this section, we aim to further elucidate the performance differences between S-BERT and TF-IDF.

Appendix C.1. Google Ngrams Analysis

To gain insights into the time-specific nature of the words that TF-IDF focuses on, we turn to examining the tokens characteristic of patent pairs located closely in the TF-IDF space through the lens of Google Ngrams data. This analysis demonstrates that the unigrams characteristic of patent pairs with high TF-IDF similarity tend be more heavily used in specific time periods compared to the S-BERT unigrams, which can explain the outperformance of TF-IDF in the period classification task.

The Google Books Ngrams dataset is a collection of word frequencies derived from the Google Books corpus,[20] which contains a vast array of books published over several centuries. This dataset enables the analysis of the usage patterns of words and phrases over time, providing a valuable resource for studying the evolution of language.

In NLP, characteristic tokens or words are specific lexical features that are highly indicative of a particular category, topic, or sentiment. These tokens serve as markers that can help in classifying or differentiating texts based on the target concept of interest, such as the party alignment of a political speech, or, in our case, whether a patent pair is deemed

[20]Specifically, we use the "English 2019" corpus accessed using *ngramr* library in R programming language (Carmody, 2023).

similar by S-BERT or TF-IDF. We use the Monroe et al. (2017) method implemented in the Schnoebelen et al. (2022) R library to systematically identify characteristic words. The method employs Bayesian shrinkage and regularization techniques to select and evaluate the relative importance of words that capture the target semantic concept.

Finding characteristic words requires a corpus of text split according to a categorical variable, which we obtain the following way. From the corpus of 11,200 patents used in the class and period validation task, we selected pairs that were in the top quartile of similarity scores according to S-BERT, TF-IDF, or both. We then categorized these pairs into three classes:

1. The representations agree
2. S-BERT identifies as similar, but TF-IDF does not *S-BERT Yes* category
3. TF-IDF identifies as similar, but S-BERT does not *TF-IDF Yes* category

We discard the pairs where both representations agreed and use the rest of the pairs as the input to Monroe et al. (2017) algorithm to find unigrams most characteristic of S-BERT and TF-IDF similarity. The output of the algorithm is the list of characteristic words for the categories *S-BERT Yes* and *TF-IDF Yes* along with the weighted log-odds that quantify the extent to which a unigram is more likely to appear in one category of patent pairs compared to the other.

Once the characteristic unigrams are obtained, we analyze their frequency from 1850 to the present using the Google Books Ngram corpus. For each unigram, we calculate the mean and standard deviation of its frequency over time. To obtain a measure of variation that is comparable between different unigrams we compute the coefficient of variation, defined as the standard deviation divided by the mean.

Figure C.13 demonstrates the average coefficient of variation for *S-BERT Yes* and *TF-IDF Yes* characteristic unigrams. The difference is large, especially for the unigrams with the highest weighted log-odds. For the top 100 unigrams, the S-BERT coefficient of variation is 0.7 compared to 1.2 for TF-IDF (which means that the average standard deviation is 70%

and 120% of the mean, respectively). As we increase the number of unigrams we include in the computation, the difference becomes smaller, but is always large: for all unigrams, the S-BERT coefficient of variation is 0.74 compared to 0.95 for TF-IDF.

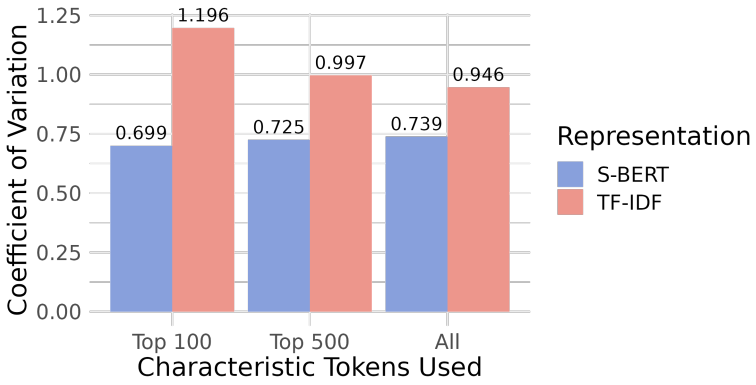


Figure C.13: Average over-time coefficient of variation of the frequency of characteristic unigrams of the pairs of patents classified as similar by S-BERT and TF-IDF

Notes: The unigram frequency information is from the Google Ngram Corpus (1850–2019). The coefficient of variation is defined as the standard deviation divided by the mean. The characteristic unigrams are computed using the Monroe et al. (2017) algorithm.

The higher coefficient of variation of unigrams in the *TF-IDF Yes* category suggests that TF-IDF is sensitive to the linguistic peculiarities of specific time periods. This provides strong evidence for why TF-IDF is more effective at categorizing patents based on their temporal context.

Appendix C.2. Synonyms Analysis

The objective of this analysis is to delve deeper into the contrasting types of similarity captured by S-BERT and TF-IDF, particularly focusing on why S-BERT excels in class validation while TF-IDF shines in the period task. Our hypothesis posits that S-BERT, unlike TF-IDF, assigns a relatively lower weight to exactly overlapping words when determining similarity between patent pairs, and leans more towards semantic similarity and other forms of word "interchangeability." This distinction becomes apparent when analyzing patents within the same period that tend to exhibit period-specific overlapping language, even if they belong to different classes. Conversely, patents from the same class but different peri-

ods are more likely to exhibit similarity at a conceptual or idea level, which is the main type of similarity we aim to capture.

In preparing the data for analysis, we further stratified patent pairs from the Class/Period validation sample into two strata: `tfidf_yes`, `S-BERT_yes`, and `agree` (using the 75th percentile similarity cutoff for yes). For instance, `S-BERT_yes` implies that according to S-BERT this pair is similar, but according to TF-IDF, it is not. We further categorized them as `same_class`, `same_period`, `both_same`, and `neither_same`. To focus on informative cases, pairs in `agree`, `both_same`, and `neither_same` categories were excluded. A sample of 200 pairs from each of the 4 strata (800 pairs in total) was selected.

To enrich our analysis, we employed WordNet, a lexical database of English (Miller, 1992). In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct word sense. These synsets are interlinked by means of semantic relations. The relations include hypernyms (more abstract terms), hyponyms (more specific terms). For each word in each patent, we listed all word senses. For each word sense, we found the set of synonyms, hypernyms, and hyponyms. These, along with the original word, were concatenated. For instance, for the word "air," we obtained a set of related terms encompassing synonyms like "breeze," hypernyms like "gas," and hyponyms like "zephyr."

Each patent was then represented as the set of unique tokens in it (each counted once) and separately as the set of unique tokens plus their synonyms, hypernyms, and hyponyms. For each document pair, we calculated the exact word overlap and the word plus synonym plus hypernym plus hyponym overlap (Word+ overlap).

We then conducted a pair of analyzes with the aim of investigating whether the same text characteristics drive both S-BERT similarity and belonging to the `same_class` category, as well as TF-IDF similarity and belonging to the `same_period` category. In the first analysis of the pair, we ran regressions with S-BERT and TF-IDF on the LHS and the text characteristics (exact word overlap and Word+ overlap) on the RHS. This analysis aimed

to explore the relationship between the similarity scores generated by S-BERT and TF-IDF and the text characteristics.

In the second analysis of the pair, we conducted a PR AUC analysis with `same_class` and `same_period` categories as the dependent variables and the text characteristics as predictors. This analysis aimed to explore how well the text characteristics predict the categorization of patents into `same_class` and `same_period` categories.

The findings from both analyzes exhibited similar patterns: S-BERT similarity and `same_class` categorization were both driven by Word+ overlap, while TF-IDF similarity and `same_period` categorization were both driven by direct word overlap. These patterns led us to conclude that S-BERT's superior performance in `same_class` categorization can be attributed to its ability to capture the semantic similarity of words present in the patents, whereas TF-IDF's superior performance in `same_period` categorization can be attributed to its ability to capture direct word overlap.

The findings are shown in Table C.7 and Figure C.14, exhibiting expected patterns. Table C.7 quantitatively shows how WordNet-derived measures relate to S-BERT and TF-IDF similarity scores. The regression coefficients indicate that S-BERT's similarity scores are negatively associated with direct word overlap but positively associated with Word+ overlap, suggesting a stronger emphasis on semantic similarity (the negative coefficient on direct word overlap is not surprising, given our sampling strategy's focus on patent pairs where the two models disagree). Conversely, TF-IDF's similarity scores are positively associated with direct word overlap, indicating a preference for exact lexical matching.

Table C.7: Regression results for similirty scores and Wordnet-based measures on the `S-BERT_yes` and `tfidf_yes` patent sample

	TF-IDF	S-BERT
(Intercept)	0.31^{***}	0.58^{***}
	(0.02)	(0.02)
Word Overlap	0.39^{***}	-0.29^{***}
	(0.04)	(0.04)
Word+ Overlap	-0.01	0.13^{**}
	(0.04)	(0.04)
R^2	0.15	0.06
Adj. R^2	0.15	0.06
Num. obs.	800	800

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Notes: The table presents the coefficients from a regression analysis where the dependent variables are the similarity scores generated by TF-IDF and S-BERT. The independent variables are Word Overlap, representing the exact word overlap between patent pairs, and Word+ Overlap, representing the overlap including synonyms, hypernyms, and hyponyms. The negative coefficients for S-BERT on Word Overlap and for TF-IDF on Word+ Overlap are observed due to the sampling strategy focusing on patents where the two models disagree.

Following the tabular analysis, Figure C.14 visually represents the Precision-Recall Area Under Curve (PR AUC) values for Word and Word+ overlap measures across `same_class` and `same_period` categorizations. In the `same_class` categorization, it is discernible from the figure that Word+ overlap (`sim_combined`) yields a higher PR AUC value of 0.49 compared to the Word overlap (`sim_1_2`) value of 0.43, underscoring the importance of capturing semantic relationships in addition to exact word overlap for classifying patents within the same class. Conversely, in the `same_period` categorization, Word overlap outperforms Word+ overlap with a PR AUC value of 0.588 against 0.512, indicating that direct word overlap is more pertinent for capturing period-specific similarities. The Figure also shows that, S-BERT performs best on `same_class` task and TF-IDF performs `same_period` task on the sub-sample used in this analysis, conforming with the full sample results discussed in Section 3.2.

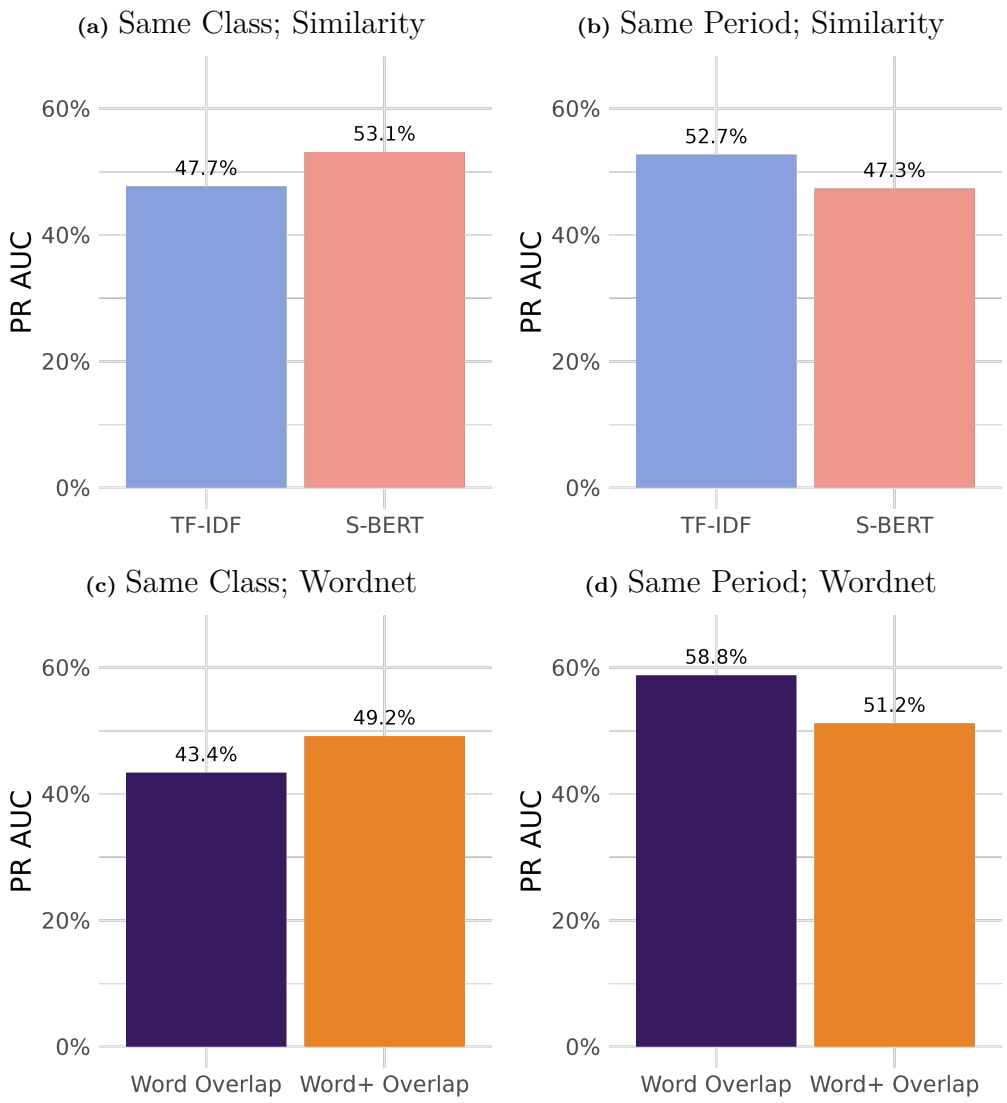


Figure C.14: Similarity scores based on the S-BERT and TF-IDF representations and Wordnet-based measures for categorizing patent pairs as belonging to the same class and period

Notes: The sample includes patent pairs in the `S-BERT_yes` and `tfidf_yes` categories. We evaluate how well patent pairs can be classified as belonging to the same class or the same quarter-century period using two sets of similarity scores, based on S-BERT and TF-IDF representations, and two sets of Wordnet-based measures, Word Overlap and Word+ Overlap. "Word" represents exact word overlap and "Word+" encompasses word overlap along with their synonyms, hypernyms, and hyponyms as derived from Wordbet, a lexical database grouping English words into sets of synonyms and recording their semantic relationships.

In conclusion, one of the mechanisms through which S-BERT better captures idea similarity is through its ability to assign similar vectors to words located closely in the semantic graph (synonyms, hypernyms, hyponyms). This is consistent with the properties theoreti-

cally expected from S-BERT based on its architecture and training procedure. Our results show that these properties are useful in innovation economics by allowing S-BERT to capture the similarity of ideas in a way that transcends period-specific language.

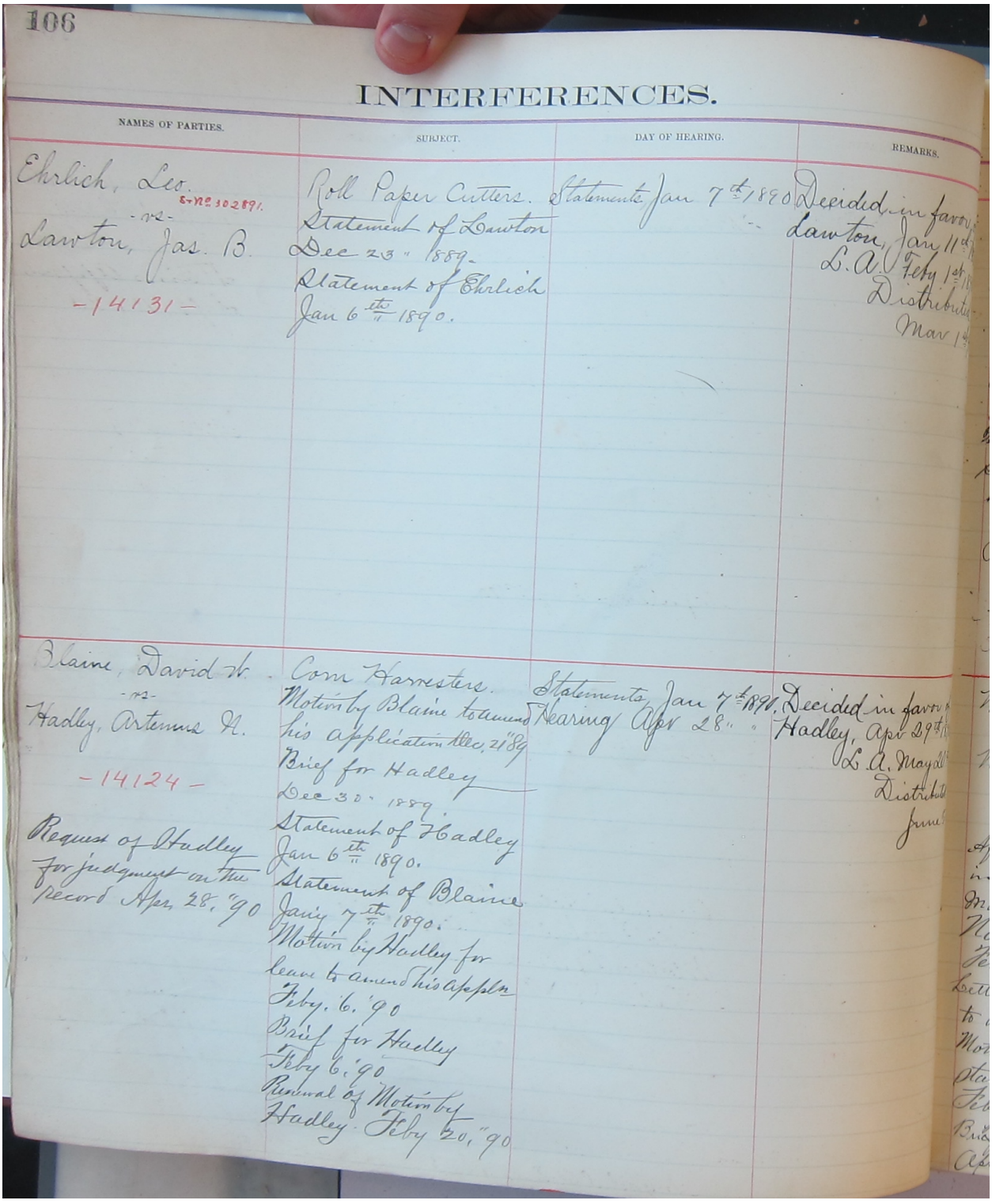
Appendix C.3. Why is S-BERT better? Conclusion

The Google Ngrams analysis and the patent pair example collectively offer robust evidence to support our initial observations. TF-IDF's strength lies in identifying patents from the same time period, primarily due to its sensitivity to words that are popular within specific temporal contexts. Conversely, S-BERT proves superior at classifying patents into the same technical class, given its ability to understand and capture the semantic essence of the text, highlighted by its association with synonym, hypernym, and hyponym overlap as opposed to the exact word overlap. These insights are important for choosing the more appropriate model for specific downstream tasks.

Appendix D. Miscellanea

Appendix D.1. Photograph of the register of interferences

Figure D.15 shows an example page from one of the Register volumes. It displays two cases. Both cases record hearing dates of January 7, 1890. The subject of the first case was roll paper cutters and the competing inventors were named Ehrlich and Lawton. The case was decided in favor of Lawton on January 11. The subject of the second case, Blaine v. Hadley, was corn harvesters; the case was decided in favor of Hadley on April 29th.



106

INTERFERENCES.

NAMES OF PARTIES.	SUBJECT.	DAY OF HEARING.	REMARKS.
Ehrlich, Leo. E No. 302891. -vs- Lawton, Jas. B. -14131-	Roll Paper Cutters. Statement of Lawton Dec 23 " 1889. Statement of Ehrlich Jan 6th " 1890.	Statements Jan 7th 1890	Decided in favor Lawton, Jan 11th L. A. Feby 1st Distributed Mar 1st
Blaine, David W. -vs- Hadley, Artemus H. -14124- Request of Hadley for judgment on the record Apr. 28. '90	Corn Harvesters. Motion by Blaine to amend his application Dec. 21 '89 Brief for Hadley Dec 30. 1889. Statement of Hadley Jan 6th " 1890. Statement of Blaine Jany 7th " 1890. Motion by Hadley for leave to amend his applcn Feby. 6. '90 Brief for Hadley Feby 6. '90 Renewal of Motion by Hadley Feby 20. '90	Statements Jan 7th 1890. Hearing Apr 28. "	Decided in favor Hadley, Apr 29th L. A. Mayall Distributed June

Figure D.15: Example page from Register of Interferences