

DELab

DIGITAL RESEARCH STUDIES

WORKING PAPER

AUGUST 23, 2023

Validation of automated (dis)similarity in legal texts: the case of regional trade agreements

Authors:

Łukasz Nawaro

Magdalena Słok-Wódkowska

CITATION

Nawaro, Ł., Słok-Wódkowska, M. (2023). Validation of automated (dis)similarity in legal texts: the case of regional trade agreements. University of Warsaw. DOI: 10.5281/zenodo.14070401



ABSTRACT

Regulatory patterns within the ever-growing set of regional trade agreements (RTAs) emerge from a continuous cycle of inspiration, paraphrasing or copy-pasting of provisions from previous treaties. The sheer number and length of RTAs makes it difficult to trace the origin of a given provision, and this is where automated text mining comes to the rescue. In this article we validate the assumptions and then develop automatic tools which make it possible to determine who creates the rules and who copies them. To this end, we split the articles into various units, from short sequences of characters to long sequences of words, and check the similarity of articles using multiple methods of assessing similarity, from simple Jaccard similarity to document embeddings. The results of automatic similarity detection are then compared to a number of expert-created challenges, to discover which method matches the ground truth best. Even though dimensionality reduction outperforms simpler bag-of-words methods in assessing pairwise similarity between articles, it is often mistaken in finding the ultimate source for the article. Jaccard similarity with a low threshold is sufficient to solve the tricky cases. Sequences of words – which increase the importance of minor changes and modifications – fail to improve the results: analysis of single words or sequences of characters is more effective. Even under conservative assumptions, about half of articles in RTAs are copied from another RTA. Our research lends empirical corroboration to quantitative assessment of the rule-making power of actors within the international economic law.

1. INTRODUCTION

International economic law consists of hundreds or thousands of treaties containing millions of words. Regional trade agreements contain over 7 million words (excluding annexes); chapters on trade in services alone consist of about 400 thousand words, which is about twice the length of Herman Melville's "Moby Dick". Many of the provisions within the treaties are considerably similar. For example, the provision on conditional most-favored nation, originally used in 1778 in a treaty between the United States and France, has since proliferated in multiple bilateral trade agreements (Houde, 2006). Similarity is more common if the treaty sides are geographically close (Latrille, 2016). In order to determine the origin and usage of a provision, researchers used to painstakingly scour tens or hundreds of treaties. Can this tedious and time-consuming work be automated?

The pioneering work in the field of text mining in international economic law was done by Allee and Lugg (2016) with the focus on Trans-Pacific Partnership and simultaneously by Alschner and Skougarevskiy (2016), who investigated international investment agreements. A year later, together with Seiermann (Alschner et al., 2017), they extended their work to regional trade agreements (RTAs), which they published on GitHub as Texts of Trade Agreements (ToTA). While their work was a milestone in the field and opened up international law to automated analysis, neither approach is justified qualitatively or quantitatively: Allee and Lugg (2016) simply decided to use the same approach as Corley et al. (2011), while Alschner and Skougarevskiy (2016) split strings just as Spirling (2012). The methodological limitations are, among others, related to the choice of the unit of analysis.

Let us consider the example used by Alschner and Skougarevskiy (2016) and compare two sentences: 'He is guilty and must not be acquitted' and 'He is not guilty and must be acquitted'. These sentences use the same words, so similarity defined as the number of unique words present in both texts divided by the number of words present in either or both is 1, while the similarity using 5-character components instead of words is 0.62. However, using 4-word sequences (he_is_guilty_and, is_guilty_and_must, . . .), the similarity is 0.0, which may be more appropriate for sentences of precisely the opposite meaning. Allee and Lugg (2016) and Allee and Elsig (2019) use even longer word sequences, up to 10-word. To prove that this approach may likewise not work perfectly, consider the first articles in the chapter "Movement of workers" in the European Union's Association Agreements (and RTAs) with Croatia and Serbia.

1. Subject to the conditions and modalities applicable in each Member State: — treatment accorded to workers who are Croatian nationals and who are legally employed in the territory of a Member State shall be free of any discrimination based on nationality, as regards working conditions, remuneration or dismissal, compared to its own nationals; (...) (EU-Croatia)
2. Subject to the conditions and modalities applicable in each Member State: (a) treatment accorded to workers who are nationals of Serbia and who are legally employed in the territory of a Member State shall be free of any discrimination based on nationality, as regards working conditions, remuneration or dismissal, compared to nationals of that Member State; (...) (EU-Serbia)

The second provision was clearly inspired by the first one. Apart from minor stylistic deviations which can be alleviated by removal of punctuation and short words, two major differences arise. "Croatian nationals" was changed to "nationals of Serbia", and "its own nationals" to "nationals of that Member State". The meaning is identical with regards to the corresponding country. Using word sequences, the similarity exceeds 80%, while for 6-word sequences it is about 50%. The incidental dissimilarity is caused by propagation of small changes in n-grams: if just one word is changed, up to n unique word sequences emerge in the article and up to n disappear from it. From the opposite perspective, a common boilerplate style for a particular side's treaties (e.g. Member States instead of countries for European Union's RTAs) does not imply meaningful similarity. Removal of country names and adjectives solves only one issue, the issue with reference to nationals would require domain-specific changes if we were to maintain long word sequences as units of analysis.

The temporal dimension remains not studied thoroughly: all treaties are simply compared to all other treaties, while the original creator of the provision - the rule-maker - remains unknown. The first-move advantage means that character sequences and too short word sequences will exhibit irrelevant similarity, and consequently rule-making ability cannot be ascribed with certainty - at best it means that a country was the first to notice an area to regulate, or to use particular words while doing so, but not to choose a particular regulatory approach. On the other hand, imperfections in optical character recognition, the possibility of paraphrasing

without changing the meaning, swapping a country name and so on call for caution when using long word sequences.

In this paper, we propose how to test whether a methodology works as intended. In particular, our work concerns articles as a unit of analysis and their (dis)similarity across treaties to check which methods accurately measure similarity, especially in tricky cases. We make as few assumptions as possible and discuss the advantages and disadvantages of particular methods. Apart from methodologies inspired by Allee and Lugg (2016) and Alschner and Skougarevskiy (2016), we study dimensionality reduction and document embeddings as alternative methods. The main aim of the study is to compare various methods and discover which one of them best suits the purpose of comparing various provisions. Our findings also shed light on another intriguing question: how many original articles are there in RTAs, and how many are copied?

2. METHODOLOGY

2.1. Preprocessing

A majority of RTAs were available in "ToTA: Texts of Trade Agreements" on GitHub (Alschner et al., 2017). The database was augmented with 27 newer agreements, including USMCA, European Union-Japan, and United Kingdom-Japan. As most of the United Kingdom's agreements were rolled over from previous European Union agreements, with some not even having a consolidated text available, they were skipped with three exceptions (European Union, Kenya – without a comparable agreement with the EU, and Japan, which differs from the EU deal). Three purely digital economy agreements outside the WTO notification scope were added: US-Japan Digital Trade Agreement; Digital Economy Partnership Agreement between Chile, New Zealand, and Singapore; and Singapore-Australia Digital Economy Agreement. The source of metadata is WTO's Regional Trade Agreements database. Additional texts of English-language agreements were transformed from PDF by pdftotext Linux tool.

Texts were transformed to lowercase, punctuation was removed, and spelling rules were adjusted to British English to be consistent with Alschner et al. (2017). Tokenization – splitting the text into words – was performed using WordPunctTokenizer from the NLTK package.

In legal texts, precise wording, including definite articles and conjunctions, may matter. Consequently, following Stavrianou et al. (2007) who argued for domain-specific transformations and earlier work in text mining in the area of RTAs, we skipped removal of the most commonly used words in the language (stopwords). Similarly, we did not use stemming or lemmatization – transformation of the word into its base form.

2.2. Similarity

In order to achieve the goal of finding who is the rule-maker in international economic law, we need to ascribe a single, most probable source for the article. As minor differences in articles are common (see EU's RTAs with Croatia and Serbia shown in the Introduction), approximate similarity detection is required. Four techniques that we study are Jaccard, cosine, and Levenshtein similarities of units in articles, and cosine similarity of document embeddings, where articles are

treated as documents. An article could be assigned as a copy of another article if similarity exceeds a user-defined threshold.

2.2.1. Threshold-based statistics -- assignment

We use the following assignment algorithm:

1. Create a mapping dictionary between article and its ID, initially empty
2. Iterate over all articles in a treaty
 - (a) If no article in the mapping dictionary exceeds a similarity threshold, add this article as a key and assign a unique number
 - (b) If there are such articles, assign the number of the earliest article in the mapping dictionary to the article in the treaty
3. Repeat step 2 for all treaties

One issue with such an algorithm is that it does not guarantee transitivity. Suppose we have three articles, A, B, and C, created in this order. If an article A is similar to article B above threshold and article B is similar to article C above threshold, article A may not be similar to article C above threshold. Using this algorithm, articles A and B have the same number assigned to them and are considered similar, while article C uses a different one despite similar or even higher possibility of being based on article B than possibility of article B being based on article A. However, as article B is similar above threshold to article A, its creator country should not be considered a rulemaker, and as article C is similar below threshold to article A, the country which created article A would not be considered a rule-maker under this assumption. Checking for transitive similarity is also computationally more expensive and likely to quickly collapse – the worst-case scenario with a 50% threshold after just seven steps is less than 1% similarity to the original article.

2.2.2. Jaccard similarity

Alschner and Skougarevskiy (2016) use Jaccard similarity to measure similarity at the treaty and chapter levels. It divides the number of units which occur in both compared entities (documents) by the number of units which occur in at least one document. This is a set-based statistic, ignoring unit order and the number of unit occurrences. The units may be in particular words, which we use as units of analysis in Examples 2.1 and 2.2.

Example 2.1.

Article a) humans shall be equal in dignity and rights

Article b) human dignity shall be inviolable

Common words 3, words only in a) 5, words only in b) 2

Jaccard similarity = $3 / (3 + 5 + 2) = 3 / 10 = 0.3$

Example 2.2.

Articles emerged in documents in the following order:

Article a) humans shall be equal in dignity and rights

Article b) human dignity shall be inviolable

Article c) humans shall be equal in dignity

Article d) dignity of humans shall be equal

Table 1. Jaccard similarity values in Example 2.2

similarity	a)	b)	c)	d)
a)	1.0	0.3	0.75	0.56
b)	0.3	1.0	0.375	0.375
c)	0.75	0.375	1.0	0.71
d)	0.56	0.375	0.71	1.0

At the threshold of 70%, a) is similar to c), and c) is similar to d). However, a) is not sufficiently similar to d), consequently while c) is being treated as a copy of a), d) is treated as an original article.

2.2.3. Cosine similarity

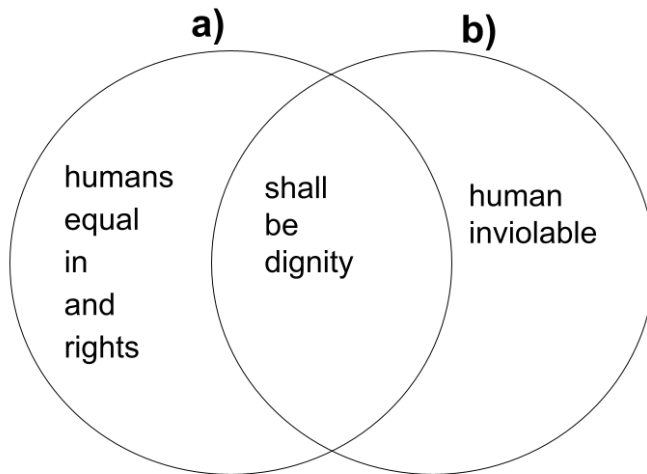
Yet another commonly used statistic is cosine similarity. It allows us to use further transformations, like assigning weight to words depending on their frequency. Documents are

transformed into a vector. A dot product of two vectors divided by the arithmetic products of the first vector's norm and the second vector's norm. The basic method to transform text into a vector is bag-of-words. First, a corpus of all words is created and numbers from 0 to $M - 1$ (where M is the number of all words in the corpus) are assigned to words. In the vector assigned for the text, the number at position m is determined by the number of occurrences of the corresponding word. Its disadvantage is ignoring word order and importance – rarely occurring words tend to be more meaningful.

We transform the bag-of-words vectors to tf-idf (term frequency-inverse document frequency) vectors. Term frequency is the number of occurrences of the word in the document. Inverse document frequency is defined as the logarithm of the number of documents divided by the number of documents which contain the word in the documentation of scikit-learn (Pedregosa et al., 2011). The Euclidean length of the resulting vector is normalized to 1 (L2 normalization), i.e. sum of squares of vector elements is 1. Cosine similarity is relatively quick to compute thanks to sparsity – the vast majority of elements is equal to 0.

Words can be transformed into topics using one of the topic modeling techniques. One of the simplest ones is latent semantic analysis (LSA), which uses truncated singular value decomposition (SVD) to produce a matrix of topics in documents. The term-document matrix X is decomposed to $U\Sigma V^T$, where U and V are orthogonal ($U^T U = U U^T = V^T V = I$), and Σ is diagonal. The smallest $N-K$ elements in the Σ matrix can be discarded to get an approximation of the original matrix. By multiplying the Σ_k matrix

Figure 1. Venn diagram of words in Example 2.1



by the document-vector, we get a k-dimensional representation of the document. Eckart and Young (1936) showed that SVD is the best least-square approximation of the larger matrix, allowing us to find latent factors. Here, the speed of the cosine computation lies in the reduced dimensionality – the number of arithmetic operations is lowered.

Example 2.3.

Article a) humans shall be equal in dignity and rights

Article b) human dignity shall be inviolable

Article c) humans shall be equal in dignity

Table 2. Cosine similarity values in Example 2.3

bag-of-words	humans	shall	be	equal	in	dignity	and	rights	human	inviolable
a)	1	1	1	1	1	1	1	1	0	0
b)	0	1	1	0	0	1	0	0	1	1
c)	1	1	1	1	1	1	0	0	0	0

$$\text{Cosine similarity (a, b)} = (1 + 1 + 1) / (\sqrt{8} * \sqrt{5}) = 3 / (\sqrt{8} * \sqrt{5}) = 0.47$$

Cosine similarity (a, c) = $(1 + 1 + 1 + 1 + 1 + 1) / (\sqrt{8} * \sqrt{6}) = 6 / (\sqrt{8} * \sqrt{6}) = 0.87$

Cosine similarity (b, c) = $(1 + 1 + 1) / (\sqrt{5} * \sqrt{6}) = 3 / (\sqrt{5} * \sqrt{6}) = 0.56$

Table 3. normalized tf-idf cosine similarity values in Example 2.3

tf-idf normalized	humans	shall	be	equal	in	dignity	and	rights	human	inviolable
a)	0.35	0.27	0.27	0.35	0.35	0.27	0.46	0.46	0	0
b)	0	0.34	0.34	0	0	0.34	0	0	0.57	0.57
c)	0.46	0.35	0.35	0.46	0.46	0.35	0	0	0	0

“shall”, “be”, and “dignity” occur three times in the set of documents, “humans”, “equal”, and “in” – twice, while all other words occur just once.

Cosine similarity tf-idf norm(a, b) = $0.27 * 0.34 * 3 = 0.28$

Cosine similarity tf-idf norm(a, c) = $0.35 * 0.46 * 3 + 0.27 * 0.35 * 3 = 0.77$

Cosine similarity tf-idf norm(b, c) = $0.34 * 0.35 * 3 = 0.36$

2.2.4. Levenshtein similarity

Levenshtein distance (Levenshtein et al., 1966) is the minimum number of single-unit edits required to transform one string into another. It can be naturally extended to all types of ordered data, including words and lists of words or characters. The measure takes into account word order, although substitution ignores distance between words. In the legal context, Palmirani et al. (2021) utilized Levenshtein distance to examine corrigenda in EU legislation. We used the Levenshtein Python library with score_cutoff parameter set to the threshold. The library treats substitution as two changes – one insertion and one deletion. Ratio is the result of division of the number of required unit insertions and deletions by the sum of the ordered list lengths measured in units subtracted from 1, so it is normalized to the [0, 1] range.

Example 2.4.

Article a) humans shall be equal in dignity and rights

Article b) human dignity shall be inviolable

Article c) humans shall be equal in dignity

a → b: remove "humans", "equal", "in", "dignity", "and", "rights"; add "human", "dignity", "inviolable"

Levenshtein similarity(a, b) = $1 - 9/(8 + 5) = 1 - 9/13 = 0.31$

a → c: remove "and", "rights"

Levenshtein similarity(a, c) = $1 - 2/(8 + 6) = 1 - 2/14 = 0.86$

b → c: remove "human", "dignity", "inviolable"; add "humans", "equal", "in", "dignity"

Levenshtein similarity(b, c) = $1 - 7/(6 + 5) = 1 - 7/11 = 0.36$

2.3. Document vectors

One worthwhile consideration for text analysis is the possibility of performing arithmetic operations on words and documents (for example, king + woman - man = queen). Cosine similarity on a bag-of-words matrix disregards word order and synonyms; similarly, operations on words are limited to co-occurrences. First, Mikolov et al. (2013) introduced word2vec, which assigns a vector to all the words, enabling arithmetic operations. Further development was done by Le and Mikolov (2014), who introduced document embeddings with doc2vec: vectors are assigned not only to words, but also to lists of words such as paragraphs and documents.

The most-often used model in the legal area is Law2Vec (Chalkidis and Kampas, 2019), which is a word2vec model. Law2Vec is trained on over 123 thousand documents and results in a 100- or 200-dimensional vectors assigned to each word which occurs at least 10 times in the corpus. The most straightforward way to create document embeddings out of word embeddings, vector averaging, loses word order and performs poorly relative to doc2vec (Campr and Ježek (2015), Lee and Yoon (2018)), although it may be an improvement over simpler methods such as SVD.

Document embeddings trained on the particular dataset in a legal setting have been used by Zhang and Zhou (2019) for Chinese aviation law and Ranera et al. (2019), as well as Yang et al.

(2020), in similar court case recommendations. Our study represents a pioneering effort in employing doc2vec to check the similarity of articles in trade agreements.

In our work, we use gensim's implementation of doc2vec with default settings. Both Distributed Memory (DM: classifying a word based on document id and nearby words) and Distributed Bag of Words (DBOW: classifying multiple words based on document id, ignoring word order) (Le and Mikolov, 2014) are tested. DBOW is limited to document vectors; DM produces word vectors as well.

2.4. Words, character n-grams, and word n-grams

All similarity detection techniques can be performed on single words and ngrams – items joined together into groups of n – which preserve local order. The items can be either characters – as in Alschner and Skougarevskiy (2016) – or words, which is the approach used by Allee and Lugg (2016).

In our study, characters are joined together into groups of characters or groups of words (including a "group" of 1 word). These are the basic units of analysis. For example, if the Preamble to the Japan-Thailand RTA begins with words "Japan and the Kingdom of Thailand", the corresponding wordbased 4-grams are: 'preamble_japan_and_the', 'japan_and_the_kingdom', 'and_the_kingdom_of', 'the_kingdom_of_thailand', while character 5-grams are 'preamb', 'reamb', 'eambl', 'amble', 'mble_', 'ble_j' and so on.

2.5. Tested methods

We define "method" as a combination of the similarity detection technique with its settings. The threshold-based statistics use thresholds of 30%, 40%, 50%, 55%, 60%, 65%, 70%, 75%,

80%, 85%, 90%, and 95%. The Jaccard similarity, cosine similarity on a tf-idf matrix, and cosine similarity on this matrix transformed by SVD into 50 and 300-dimensional vectors.¹

Vector sizes tested are 10, 50, 100, 200 and 300. A vector is assigned to each article in the database. In DM, word vectors are created as well. Learning rate, window size, min_count and all other parameters are not changed. All methods are tested with four possibilities with regards to creating treaty and chapter vectors in addition to article vectors: both, only one, or none.

For the sake of brevity, "cosine article" means computing cosine similarity on tf-idf-transformed bag-of-words matrix, "cosine SVD" – on SVD matrix, "vector" – on article embedding.

Five units of analysis are used in methods' settings: 5-character n-grams as in Alschner and Skougarevskiy (2016), 10-character n-grams, words, 4- word n-grams, 6-word n-grams as in Allee and Lugg (2016).

¹ We could use the word "tokens" instead of "words", but as we remove punctuation, all the tokens are words despite using WordPunctTokenizer.

3. EVALUATING ARTICLE SIMILARITY

3.1. Groups of articles

In order to verify the similarity identified by various methods the examples are chosen to compare articles which are close in wording but different in meaning or the opposite – the kind of task relatively easy for human lawyers, but difficult for algorithms. The following list is sorted from the most similar groups of articles, comparing Jaccard similarity by words, to the most dissimilar. If there are two similar groups, they are presented together, ordered using the higher Jaccard similarity – see Table 4 for values sorted by maximum 1-word similarity. The studied groups, joined together if they concern the same general topic, are:

- Open government data:** These kind of articles appears in fairly few RTAs, as it is a new regulatory invention. Japan and Singapore both signed two treaties containing a provision on open government data – Japan with the United States and United Kingdom, Singapore with Australia and Chile/New Zealand. The Japanese articles are very similar in wording, the Singaporean share some similarities, but also have minor, insignificant differences. We compare them with the articles in EU-United Kingdom and USMCA. The pairs which we expect to be sufficiently similar are United Kingdom-Japan & United States-Japan, as well as Digital Economy Partnership Agreement & Australia-Singapore Digital Economy Agreement. The other 13 pairs are different, as shown by Mazur and Słok-Wódkowska (2022).
- Affirmation of rights & obligations:** these are general and common provisions. The difficulty with them is that they may refer to various chapters, with usage of similar wording. The RTAs of the United States and Thailand with Chile affirm (or reaffirm) existing obligations in general. The article in US-Australia RTA is limited to sanitary and phytosanitary requirements.
- Internal taxation:** these are very common provisions, that appear in almost every RTA, modelled after GATT's Article III. the TurkeyAlbania RTA differs by changing "contracting parties" to "parties", "products originating in the territory of the other contracting party" to "products originating in the other party" and "one of the parties" to "the parties". All changes are minor and insignificant for meaning.

- **Security exceptions:** this kind of provision is used in multiple RTAs, although with no clear model and various wording. The Latvia-Sweden RTA contains both minor (“party” instead of “state party”) and somewhat significant, but not crucial to the point of the article (the limitation regarding “traffic in arms, ammunition and implements of war” regarding conditions of competition), differences to the corresponding article in EFTA-Turkey RTA.
- **MFN investment:** this is one of the most traditional clauses in international economic law, used not only toward investment but also i.a. trade in goods and services. There are usually multiple MFN clauses in every agreement. Two EU articles with Balkan countries are almost identical, they only differ in the country name (Croatia and Serbia) and a few minor details. The third agreement, US-Colombia, contains an explicit most-favored-nation article regarding investment with the same meaning, but different wording.
- **GATT XI:** article XI of GATT is about limitation of quantitative restrictions. Some treaties include it *mutatis mutandis* (EFTA-Colombia), other copy it (EU-Ghana with a limitation regarding safeguards, EUChile). The difference between EU-Chile and EFTA-Colombia is even smaller than between the two EU treaties, despite the difference in wording.
- **Enquiries:** provision related to conducting enquiries can refer to various aspects of trade. Three articles in RTAs of Mexico, Malaysia, and Chile with Japan contain requirements about designating enquiry points regarding sanitary and phytosanitary (SPS) measures. One of them, with Mexico, is outside the SPS chapter, but concerns SPS explicitly. The Pakistan-Malaysia RTA, on the other hand, is different – the enquiry point is not related to SPS, but customs matters in general, so it is materially different to all other articles.
- **Data flow:** this is a fairly new topic, hence there is only limited number of RTAs that contain such provisions. Three articles which were assigned category 1 in 1.28.1 of TAPED database by Burri and Polanco (2020) are “soft” – the parties “affirm the importance of working” or “shall endeavor” to facilitate cross-border flows of information. The article from Korea-US RTA concerns only information flows, while Nicaragua-Taiwan and Canada-Korea “affirm the importance” of multiple areas of cooperation in electronic commerce. Code 3 means “hard” requirements: “shall allow” or “shall [not] prohibit or restrict”, and treaties with articles being compared are TPP, USMCA (just like Korea-US, regulating less areas), and Indonesia-Australia. The implications of these two groups of articles are meaningfully different, despite similar wording about cross-border transfer of information.
- **Electronic transmissions:** this provision directly implements the WTO’s decision, which is however temporary. In both Australian RTAs the regulation forbidding customs on

electronic transmissions is the same. The additional sentence in the RTA with Singapore only mentions that it is also the “current practice”, but it has no relevance for the future in which the treaty has legal force.

- **Goods/services:** we compare two articles in US RTAs with Chile and Singapore about national treatment for goods and services, respectively. The two differences between them is the area regulated, and a clause with a limitation in the Chilean goods article.
- **Universal service:** in all analyzed articles (EU-Chile, US-Singapore, EU-CARIFORUM States EPA), the obligation must be:
 - transparent
 - non-discriminatory
 - competitively neutral
 - not more burdensome than necessary

All articles should be considered the same despite paraphrases.

- **Arrangement for particular goods:** EU’s analyzed arrangements are about textiles in its RTA with Croatia, and agricultural goods in RTA with Norway.
- **Application:** EFTA-Turkey RTA concerns territorial application of the treaty, while Japan-Mongolia – application of sanitary and phytosanitary requirements. This word, despite its relative rarity which makes it impactful in tf-idf, is too general.
- **T&Cs:** although both articles contain the phrase “terms and conditions”, the Russia-Azerbaijan RTA concerns T&Cs of accession to the treaty, while Trans-Pacific Partnership – of cooperation.
- **Workers’ treatment:** EU’s agreement with Hungary is binding regarding discrimination, while with Colombia and Peru sides only “recognise the importance of promoting equality (...) with a view to eliminating any discrimination”

Table 4 presents Jaccard similarity within groups and the desired outcome.

Table 4. Jaccard similarity within groups (in percentages), desired outcome and challenge present in the group.

Group	5-character (%)	1-word (%)	4-word (%)	Desired outcome
-------	-----------------	------------	------------	-----------------

Open government data	93.6, 48.9, 48.1, 49.4, 42.2, 41.0, 49.3, 34.6, 30.9, 36.2, 29.4, 25.1,	93.7, 52.2, 55.2, 54.0, 49.1, 49.1, 37.6, 37.6, 34.5, 31.7, 36.3, 30.3,	70.4, 27.0, 28.8, 22.3, 23.8, 24.8, 15.7, 17.8, 12.7, 20.5, 12.8, 11.1,	Japan same Singapore same EU-UK different USMCA different
Affirmation of rights & obligations	94.2, 52.4, 49.2	91.3, 56.5, 50.0	74.2, 32.4, 22.5	Chile same US-Australia different
Internal taxation	86.7	91.1	41.8	same
Security exceptions	79.2	83.5	69.4	same
MFN investment	73.6, 10.8, 9.3	78.5, 14.8, 13.7	49.8, 0.0, 0.0	same
GATT XI	66.6, 21.2, 19.9	76.0, 27.3, 25.9	41.0, 4.7, 5.9	same
Enquiries	56.8, 64.5, 74.8, 11.9, 15.6, 13.4	70.6, 68.6, 68.6, 22.2, 22.2, 20.0	12.7, 18.9, 31.1, 1.0, 2.1, 0.0	Japan same Pakistan-Malaysia different
Data flow	65.5, 63.8, 58.8, 57.9, 11.9, 9.7, 10.7, 11.2, 12.2, 10.5, 11.5, 9.3, 10.2, 7.4, 8.8	68.0, 63.8, 68.3, 62.5, 10.9, 15.6, 15.5, 15.4, 15.2, 14.0, 11.5, 11.1	53.0, 35.6, 33.3, 28.0, 0.0, 0.0, 14.0, 15.2, 12.4, 11.6, 11.5, 11.1	soft same hard same
Electronic transmissions	45.8	45.5	20.0	same
Goods/services	30.5	41.1	16.0	different
Universal service	42.6, 34.0, 31.6	41.4, 39.6, 36.1	14.9, 4.9, 9.0	same
Arrangements for particular goods	28.0	38.1	0.0	different
Application	19.3	33.3	0.0	different
T&Cs	15.1	31.4	2.0	different
Workers' treatment	11.7	17.8	0.0	different

3. 2. Measuring the quality of methods

First, we found the sources for the articles in a pair. The source is the treaty in which the article with the article's ID assigned by an algorithm described in Section 2.2.1 first appeared. We checked whether the sources are the same or different and whether the result matches the desired outcome column. Desired outcome matching is not sufficient; the source for the article

should also be reasonable and not random. For that reason, as a second check we examined whether both suspected sources for the articles can be accepted by the expert. If both the first (pair of sources matching with desired outcome) and the second (sources accepted by expert analysis) checks pass, we consider that pair correctly evaluated.

There are multiple available metrics to aggregate the results. We present the following ones:

1. **number of pairs of articles for which the result is correct:** this is the most straightforward metric, which simply checks whether the method returns a correct result and sorts the methods by the number of such article pairs. A potential issue with it is that it places excessive weight on articles in larger groups -- one article is compared up to 5 times, and one group contains up to 15 article pairs.
2. **number of pairs of articles for which the result is correct with each group's impact equalized:** if there are n pairs of articles in a group, each pair has $\frac{1}{n}$ impact. Methods which score high in this metric perform best in a diverse set of tasks, but may fail more often in specific cases.
3. **number of pairs of articles with the same desired outcome for which the result is correct:** it is trivial to achieve 100% correctness (in the first check; ignoring the expert analysis of sources) with all pairs belonging to either "same" or "different" group: all pairs of articles are classified to the respective group. However, the naïve method achieves 0% correctness for the other desired outcome. We compute the harmonic mean of the share of correct pairs, which penalizes methods performing very poorly for either desired outcome, to see which methods perform well in both cases. This metric is meant to alleviate the imbalance between the number of pairs which are supposed to be the same and different.

4. RESULTS

In the full evaluation of both the outcome (automatic matching) and the sources (expert analysis), Jaccard similarity on 5-character n-grams and threshold of 30% achieves the superior result of 45 correctly evaluated pairs of articles (Table 5). Cosine similarity on 10-character n-grams and the same threshold could reach 45 such pairs as well, but one was determined by an expert to be wrong. Both methods find about ten thousand unique source articles in the database: 10229 and 9650, respectively. Several methods get 43 pairs right – they include Levenshtein similarity as well and find a larger number of source articles, from 15.5 thousand to 20.5 thousand. Just below we find the embedding-based methods, which have in common the threshold of 90%, 100 dimensions, and using CBOW. They differ in the unit of analysis and created vectors (5-gram character-based with only article vectors, and 1-gram word-based with both article and chapter vectors).

The top score with regards to matching the desired outcome only (Appendix Table 1)) is achieved by cosine SVD with 50 dimensions on 1-word n-grams and 85% threshold: 49 pairs matching the desired outcome with 2836 source articles. The tied-second best score can be achieved with as little as 38 source articles by document embeddings, and this is clearly too little. It shows that expert evaluation is indispensable. Similarly, document embeddings are superior with the metric averaging groups, also with four available settings not to be worse than any other method, and cosine SVD is back as the top performer in the metric utilizing the harmonic mean. Such methods more often ascribe random, irrelevant articles as the sources, but later do absolutely fine when comparing pairs.

Document embeddings do better when the groups' impact is normalized (Table 6), with 2 out of top 3 methods being document embeddings, one with 10-character n-grams, 100 dimensions and article vectors only, another with 5-character n-grams, 50 dimensions, and article, chapter, and treaty vectors – but they still fail to outperform the aforementioned Jaccard similarity with the appropriate settings. Jaccard similarity on 5-character n-grams and threshold of 30% outperforms all other methods in the metric based on harmonic mean of the share of correctly evaluated same/different pairs (Table 7) as well.

The differences between the top six methods sorted by the number-of pairs metric with the particular focus on the two top-performing ones are the following:

- **Open government data:** the top methods overstate the importance of DEPA and USMCA, which helps them in the comparison between DEPA and DEA (stemming from the same source), but causes them to make mistakes for Australia-Singapore and the UK treaties. Cosine similarity on 5-character n-grams and 65% threshold performs even worse.
- **Internal taxation:** the top cosine-based method finds two separate sources for the articles. The top Jaccard-based method finds Yaoundé I as the source of both articles, while EU-Switzerland-Liechtenstein is considered the source by the other methods.
- **GATT XI:** EU's RTAs are considered stemming from the same EFTA source by four methods – the two top-performing ones and the two cosine-based out of four worse-performing ones. The other two look for the source of the article in EU-Ghana in other EU treaties.
- **Enquiries:** cosine similarity on 1-word n-grams with 75% threshold fails in three cases due to considering Japan-Malaysia and Chile-Japan articles as original.
- **Electronic transmissions:** Singapore-Australia is correctly chosen as the source by the two top-performing methods, other methods treat the articles as original.
- **Goods/services:** the two top-performing methods find a difference due to one article emerging from CUSFTA and the other – from NAFTA, while three of four others claim that both articles stem from NAFTA.
- **Universal service:** the two top-performing methods find all articles stemming from the EU-Chile RTA. Cosine similarity correctly finds US-Singapore to be a copy of EU-Chile, but falsely assigns the article from EU-CARIFORUM States EPA as original.

Arrangements for particular goods: in this case the two topperforming methods are worse – Jaccard-based claims that both articles stem from the same source, while cosine-based is rejected by the expert due to finding the article in EU-Croatia as a copy of an article in EC-Faroe Islands.

Table 5. Top-performing methods (number of correctly evaluated pairs of articles).

Method	Threshold (%)	n-grams	Dimensions	Result
Jaccard	30	5-gram character		45
cosine	30	10-gram character		44
Levenshtein	60	10-gram character		43

cosine	65	5-gram character		43
cosine	75	1-gram word		43
Jaccard	60	1-gram word		43
Levenshtein	55	10-gram character		42
embeddings	90	1-gram word	100	42
embeddings	90	5-gram character	100	42
cosine	80	5-gram character		42
cosine	70	5-gram character		42
cosine	55	10-gram character		42
cosine SVD	90	1-gram word	300	42

Table 6. Top-performing methods (number of correctly evaluated pairs of articles with each group's impact equalized).

Method	Threshold (%)	n-grams	Dimensions	Result
Jaccard	30	5-gram character		12.266667
embeddings	85	10-gram character	100	12.133333
embeddings	75	5-gram character	50	12.066667
cosine	50	1-gram word		11.866667
Levenshtein	30	4-gram word		11.600000
embeddings	95	1-gram word	300	11.533333
cosine SVD	65	1-gram word	300	11.466667
Levenshtein	40	1-gram word		11.400000
Levenshtein	30	5-gram character		11.400000

Levenshtein	40	10-gram character		11.333333
cosine SVD	90	1-gram word	300	11.333333

Table 7. Top-performing methods (harmonic mean of the share of correctly evaluated same and different pairs of articles).

Method	Threshold (%)	n-grams	Dimensions	Result
Jaccard	30	5-gram character		0.782895
cosine	30	10-gram character		0.770134
embeddings	75	5-gram character	300	0.719424
Levenshtein	30	5-gram character		0.713504
cosine	65	5-gram character		0.709790
embeddings	75	5-gram character	200	0.705882
embeddings	90	1-gram word	100	0.700000
embeddings	70	5-gram character	200	0.691729
embeddings	85	10-gram character	100	0.691729
Levenshtein	30	4-gram word		0.689781
Levenshtein	60	5-gram character		0.689781
cosine	40	10-gram character		0.689781

5. LIMITATIONS AND CONCLUSIONS

The metrics show that short sequences of words and sequences of characters outperform more complex units. At the same time, dimensionality reduction is recommended when solely pairwise article comparison is required.

The number of methods tested is already vast -- in the thousands. With such a number, we may run into overfitting of the chosen method. The top methods could be accidentally effective, but would not generalize well. We argue that the stability of the number of unique articles among the top methods shows that many methods found themselves near the global optimum, and overfitting is prevented by using unsupervised methods. We do not finetune the settings to search for local overfitted optima. A large number of methods using document embeddings reached a conclusion that the number of articles is close to 10 thousand, but not a single one has outperformed the simpler methods.

After testing multiple methods of finding similar articles, some being an extension of previous methods and some being novel, we found that Jaccard similarity on 5-character n-grams introduced by Alschner and Skougarevskiy (2016) achieves the best results when combined with expert evaluation. At the same time, pairwise comparison points to the fact that dimensionality reduction is better in evaluating whether articles are sufficiently similar – the conclusions regarding similarity of articles reached by cosine or Jaccard similarity are wrong more often. However, an article created a long time ago may be wrongly considered as a potential source. For example, SVD clearly overstates the importance of words "measures" and "paragraph", so an unrelated article in Yaoundé I becomes a source to an article in USMCA about Open government data. Possibly, manual cleaning of domain-specific stopwords would help.

The results are robust to the choice of the unit of analysis, as long as the unit length is not excessive. Poor results of methods using long sequences of words may be due to data quality issues. Optical character recognition is not perfect and minor mistakes "infect" more units, when the units are defined more broadly as multiple-word components. Publishing law in an interoperable, unified standard would be perfect; until this admittedly unlikely feat is achieved, issues with data will occur often. We focused on the methods, while leaving the data cleaning and transforming just as before for comparability with previous studies. Future work can focus on data quality and possible modifications.

The number of unique source articles in the two top-performing methods is very close to 10 thousand in both the number-of-pairs and the harmonic mean metrics, with slightly worse methods finding up to 23 thousand source articles. The numbers are smaller in the metric with each group's impact equalized, but it is generally over 5 thousand. Even with rather strict assumptions about similarity, we can say that at least about half of all RTA articles (42277 in our database) are copied from another article.

Regional trade agreements do not meet most definitions of big data. Although numerous and complex, they are far from 825 GiB datasets used to train language models such as GPT-NeoX-20B Black et al. (2022). Consequently, taking into account unit frequency does not create major improvement. Integration of knowledge stemming from other data sources could make machine learning methods find the actual "meaning" of words; on the other hand, the particular area of international law is small and specific, distinct to the common law system prevalent in English texts. Here, document embeddings, just like other methods, were only given access to the RTA dataset, so the more "traditional" methods perform just as well. Both legal (Chalkidis and Kamps, 2019) and general (Choi et al., 2023) datasets can be used to create an AI lawyer, who can then be supported by simple, interpretable metrics.

The source-article method has an inherent limitation, as only articles in regional trade agreements are potential sources. We did not include the General Agreement on Tariffs and Trade as a source, as it is not a regional trade agreement, but some provisions are clearly inspired by it. A treaty side introducing a GATT-style wording may still provide some information, one could suspect it to be a supporter of multilateralism. For a discussion about this topic using Jaccard similarity on 1-word n-grams, see Alschner et al. (2022).

FUNDING

This research was funded by the National Science Centre, Poland (project number 2019/35/B/HS5/02107).

DATA AVAILABILITY STATEMENT

Data associated with the research is available on https://osf.io/bmgw9/?view_only=05d5697f9c3a430b8fe69a4f09bbe8c2. New regional trade agreements (transform_xml) will be made available on a public repository upon acceptance.

REFERENCES

- Allee, T., Elsig, M., 2019. Are the contents of international treaties copied and pasted? Evidence from preferential trade agreements. *International Studies Quarterly* 63, 603–613.
- Allee, T., Lugg, A., 2016. Who wrote the rules for the Trans-Pacific Partnership? *Research & Politics* 3, 2053168016658919.
- Alschner, W., Elsig, M., Wüthrich, S., 2022. Main act or side show? model agreements by international institutions and their reuse in investment treaty texts. *Journal of International Economic Law* 25, 592–610.
- Alschner, W., Seiermann, J., Skougarevskiy, D., 2017. Text-as-data analysis of preferential trade agreements: mapping the PTA landscape. UN.
- Alschner, W., Skougarevskiy, D., 2016. Mapping the universe of international investment agreements. *Journal of International Economic Law* 19, 561– 588.
- Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonnell, K., Phang, J., et al., 2022. GPTNeoX-20B: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745 .
- Burri, M., Polanco, R., 2020. Digital trade provisions in preferential trade agreements: introducing a new dataset. *Journal of International Economic Law* 23, 187–220.
- Campr, M., Ježek, K., 2015. Comparing semantic models for evaluating automatic document summarization, in: *Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen, Czech Republic, September 14-17, 2015, Proceedings 18*, Springer. pp. 252–260.
- Chalkidis, I., Kampas, D., 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* 27, 171–198.
- Choi, J.H., Hickman, K.E., Monahan, A., Schwarcz, D., 2023. Chatgpt goes to law school. Available at SSRN .

Corley, P.C., Collins Jr, P.M., Calvin, B., 2011. Lower court influence on US Supreme Court opinion content. *The Journal of Politics* 73, 31–44.

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1, 211–218.

Houde, M., 2006. Most-favoured-nation treatment in international investment law. *Bulletin of Comparative Labour Relations* 60, 69.

Latrille, P., 2016. Services rules in regional trade agreements: How diverse or creative are they compared to the multilateral rules. *Regional Trade Agreements and the Multilateral Trading System*, 421–493.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents, in: *International conference on machine learning*, PMLR. pp. 1188– 1196.

Lee, H., Yoon, Y., 2018. Engineering doc2vec for automatic classification of product descriptions on o2o applications. *Electronic Commerce Research* 18, 433–456.

Levenshtein, V.I., et al., 1966. Binary codes capable of correcting deletions, insertions, and reversals, in: *Soviet physics doklady, Soviet Union*. pp. 707–710.

Mazur, J., Słok-Wódkowska, M., 2022. Access to information and data in international law: How to find a path forward from human rights-oriented and market-oriented approach? *Nordic Journal of International Law* 91, 310–338.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Palmirani, M., Sovrano, F., Liga, D., Sapienza, S., Vitali, F., 2021. Hybrid ai framework for legal analysis of the eu legislation corrigenda, in: *Legal knowledge and information systems*. IOS Press, pp. 68–75.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,

Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830. 2

Ranera, L.T.B., Solano, G.A., Oco, N., 2019. Retrieval of semantically similar philippine supreme court case decisions using doc2vec, in: 2019 International Symposium on Multimedia and Communication Technology (ISMTC), IEEE. pp. 1–6.

Spirling, A., 2012. US treaty making with American Indians: institutional change and relative power, 1784–1911. *American Journal of Political Science* 56, 84–97.

Stavrianou, A., Andritsos, P., Nicoloyannis, N., 2007. Overview and semantic issues of text mining. *ACM Sigmod Record* 36, 23–34.

Yang, F., Chen, J., Huang, Y., Li, C., 2020. Court similar case recommendation model based on word embedding and word frequency, in: 2020 12th International Conference on Advanced Computational Intelligence (ICACI), IEEE. pp. 165–170.

Zhang, H., Zhou, L., 2019. Similarity judgment of civil aviation regulations based on doc2vec deep learning algorithm, in: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), IEEE. pp. 1–8.

A TABLES

For clarity, only 10 best results (and those tied with 10th) are presented below.

Table 1. Top-performing methods (number of cases in which the outcome matches with the desired outcome).

Method	Threshold (%)	n-grams	Dimensions	Result
cosine SVD	85	1-gram word	50	49
embeddings	30	10-gram character	100	46
cosine SVD	65	1-gram word	300	46
embeddings	55	1-gram word	300	45
embeddings	40	1-gram word	300	45
Jaccard	30	5-gram character	45	45
cosine	30	10-gram character	45	45
embeddings	40	1-gram word	200	44
embeddings	85	1-gram word	100	44
embeddings	80	10-gram character	50	44
cosine SVD	85	5-gram character	50	44
cosine SVD	80	5-gram character	50	44
cosine SVD	60	10-gram character	300	44
cosine SVD	60	1-gram word	300	44

Table 2. Top-performing methods (number of cases in which the outcome matches with the desired outcome with each group's impact equalized).

Method	Threshold (%)	n-grams	Dimensions	Result
embeddings	80	1-gram word	100	12.800000
embeddings	75	5-gram character	10	12.700000
embeddings	80	10-gram character	300	12.600000
embeddings	80	10-gram character	200	12.433333
Levenshtein	30	1-gram word		12.333333

embeddings	75	5-gram character	50	12.333333
embeddings	85	10-gram character	100	12.333333
cosine SVD	65	1-gram word	300	12.333333
embeddings	50	5-gram character	50	12.333333
cosine	50	1-gram word		12.333333

Table 3. Top-performing methods (harmonic mean of the share of the percentage of cases in which the outcome matches with the desired outcome for same and different pairs of articles).

Method	Threshold (%)	n-grams	Dimensions	Result
cosine SVD	85	1-gram word	50	0.845455
embeddings	30	10-gram character	100	0.825472
embeddings	40	1-gram word	300	0.805732
cosine SVD	65	1-gram word	300	0.795161
Jaccard	30	5-gram character		0.782895
cosine	30	10-gram character		0.782895
embeddings	40	1-gram word	200	0.770134
embeddings	40	1-gram word	100	0.765306
cosine	50	1-gram word		0.753448
embeddings	75	5-gram character	10	0.750000