


Automatic Labeling of Articles in International Investment Agreements

Machine Learning and Word Embeddings

Conference Paper

Author(s):

[Rao, Susie Xi](#) 

Publication date:

2017-12

Permanent link:

<https://doi.org/10.3929/ethz-b-000228391>

Rights / license:

[In Copyright - Non-Commercial Use Permitted](#)

Automatic Labeling of Articles in International Investment Agreements

Machine Learning and Word Embeddings

Susie Xi RAO^{a,b,1}

^a*Institute of Computational Linguistics, University of Zurich, Switzerland*

^b*KOF Swiss Economic Institute, ETH Zurich, Switzerland*

Abstract. International investment agreements (IIAs) are international commitments amongst contracting parties to protect and promote investment. Although each treaty has a distinctive structure regarding placement and organization of information, IIAs as instruments of international law share underlying textual and legal structures. Treaty articles are important components in IIAs: Some articles have been assigned with titles, while the other articles remained untitled. In this paper, the author investigates various machine learning methods using features such as bag-of-words (BoW) and word embeddings. Comparing the performances of supervised and unsupervised systems leads to the conclusion that word embeddings can effectively expand the semantic features for words and documents, which enables the accurate categorization of texts from closely related sub-fields of one research area. To the best of the author's knowledge, this work is the first endeavor to categorize treaty articles in the existing IIAs across languages. It is believed that treaty article categorization can assist mapping treaty texts to their inherent structures. The resulting simplified structure of a treaty is represented by IIA topics, which is beneficial to organizing treaties in information retrieval systems or databases.

Keywords. document embeddings, international investment agreements (IIAs), partially supervised clustering, supervised learning, semi-supervised learning, text categorization, word embeddings

1. Introduction and Background

International investment agreements (IIAs) are “essentially instruments of international law” [1]. A fundamental purpose of investment treaties is to protect and promote investment. It has been commonly agreed on that in the literature for treaty content and structure, although there has been no uniform treaty structure and the degree of agreement varies across treaties, essentially all investment treaties address the same issues and follow similar legal and textual structures (cf. [1, 2]). As a result, despite the variations in the language usage from treaty to treaty, we argue that because

¹Susie Xi Rao is a recent graduate from the master programs in Multilingual Text Analysis and Computational Linguistics with the Institute of Computational Linguistics, the University of Zurich. She currently works at the KOF Swiss Economic Institute, ETH Zurich (e-mail: rao@kof.ethz.ch, xi.rao@uzh.ch). The author appreciates the valuable comments and support from Prof. Dr. Martin Volk and Dr. Kyoko Sugisaki from the Institute of Computational Linguistics at the University of Zurich and Prof. Dr. Peter Egger from KOF Swiss Economic Institute, ETH Zurich, who has been in charge of manually assigning titles to the treaty articles as gold standards in the evaluation. The author also thanks for the support of the SNIS project *Diffusion of International Law* and numerous feedback from the UZH CL community. This draft version was presented at the 2017 Legal Data Analysis workshop (LDA2017) and will be published in 2018.

December 2017

ARTICLE 1
Definitions
For the purposes of this Agreement:
1. "Investment" means every kind of asset that is invested by investors of either Party in the territory of the other Party, in accordance with laws and provisions of this Party.
In particular, though not exclusively, the term "Investment" means:

Article 1
The Parties shall, subject to the laws and regulations and investment policies in force in their respective countries, take all appropriate measures to facilitate, promote, strengthen and diversify trade, economic relations and investment, with the aim of achieving a mutually beneficial expansion of trade, economic relations and investment.

Figure 1. Example of a titled article

Figure 2. Example of an untitled article

of the strong commonality among them, more than 3,300 individual investment treaties negotiated over the last six decades constitute a single, integrated global regime for investment.

Generally speaking, a treaty is composed of preface, preamble (e.g. title page and table of contents), text body (i.e. articles and paragraphs), conclusion (e.g. signatures) and sometimes annex (cf. [3]), with articles as thematic units. Figures 1 and 2 show examples of a titled article (Article 1 entitled "Definitions") and an untitled article, respectively.

In order to understand the negotiation behaviors of contracting parties, we can use the content and structure of IIAs as a posteriori proxies and reflection of the negotiation processes. Therefore, analyzing treaty structure and content of IIAs as a body of law instruments has established itself as a research area that continuously gains more interest from various disciplines, such as law, economics, political science. One goal of the IIA studies is to build up a comprehensive database with categorized information, where a full-text query of specific types of provisions is allowed.

A joint project on IIAs, *Diffusion of International Law: A Textual Analysis of International Investment Agreements*², was launched from various disciplines (law, economics, political science, computational linguistics (CL)) under the Swiss Network for International Studies (SNIS) network, with the goals to understand the design, evolution, and effects of the IIAs currently in practice. Under this project, an English corpus of IIAs (the SNIS corpus hereafter) was created by converting "a broad variety of formats" [3] (e.g. PDF, HTML, *Microsoft Word*, etc.) and into XML documents and automatically translating non-English treaties into English. In total, the corpus has 2,823 English treaties.

Techniques of textual analysis (e.g. clustering, text similarity measures, cf. [2, 4, 5]) have been receiving much attention due to their efficacy in transforming textual data into meaningful and operationalisable representations. Current research on text classification in the legal domain has mainly focused on a document as a whole. Treaty article as the unit of analysis has not yet been the focus of research in the legal domain, although the exploration of smaller analysis unit (e.g. sentential, cf. [6–9]) has begun. Alschner and Skougarevskiy [10] have implemented multinomial inverse regression (trained by a human-generated IIA taxonomy) to classify 22,500 IIA articles that were originally in English; however, they did not report the performance of the classifier. When we look at the IIA treaties at the article level, some articles come with titles; others do not. As certain treaty articles are marked with titles which summarize the content described succinctly, we can utilize article titles as an assistance to grasp the structure and content in treaties without reading through treaty texts. Hence, if we aim at representing treaty structure with articles, we will first need to assign titles to the untitled text blocks by learning the knowledge encoded in the titled articles.

In this paper, we discuss and experiment various methods of text categorization to assign titles to untitled treaty articles automatically. The remainder of this work is organized as follows: Section 2 describes the steps to extract and preprocess the titled and untitled articles, followed by Section 3 on a naive experiment of article classification. Sections 4, 5, 6 and 7 are devoted to the applicability of machine learning methods (e.g. supervised and semi-supervised) to assign titles to untitled articles, where the efficacy of different methods is evaluated. We conclude the paper with Section 8 about important findings, implications and future work.

²http://www.snis.ch/project_diffusion-international-law-textual-analysis-international-investment-agreements (accessed 26 Jan 2017).

Table 1. Lowercased token and type counts for titled and untitled articles **Table 2.** Unique normalized titles after preprocessing

	article	token	type
<i>untitled</i>	10,074	1,809,743	22,304
<i>titled</i>	34,524	7,505,258	38,953
total	44,598	9,315,001	52,171

unique normalized titles	frequency
"admission", "investment"	28
"contracting", "dispute", "settlement", "state"	38
"compensation", "dispossession"	72

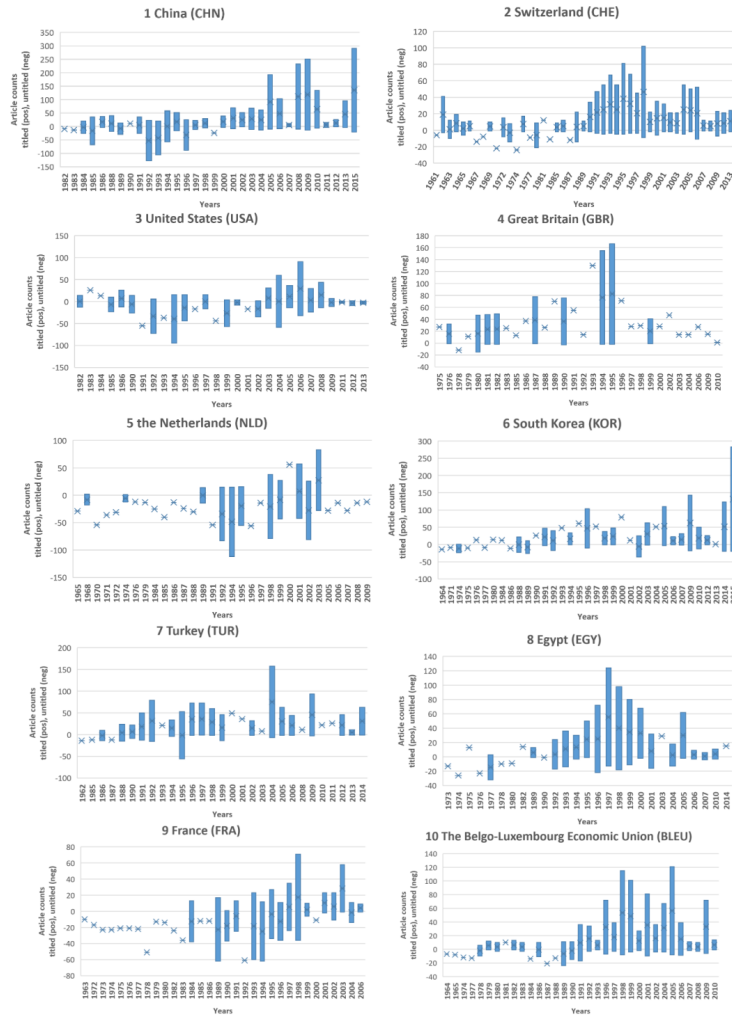


Figure 3. Boxplots of the numbers of the titled and untitled articles in Top 10 negotiators, titled: positive, untitled: negative

2. Extraction of Treaty Articles

In order to perform article categorization, we first need to extract from the titled articles their titles and the corresponding texts as training material as well as the untitled text blocks as test data. Despite the variety of article structure (be it nested or not) in XML, we managed to develop strategies of extracting titles and texts as much as possible.

The tokens and types for *titled* and *untitled* articles are shown in Table 1. We observe that while the number of the titled articles is three times that of the untitled articles, titled and untitled

December 2017

articles do not share a large portion of the vocabularies. In total, we have extracted 10,074 untitled and 34,524 titled articles with more than nine million tokens. Interestingly, we found out that untitled and titled articles are mixed in 453 treaties (16% of the SNIS corpus).

According to expert consultation, it is solely due to formality reasons that some articles from certain signatories are titled, while others are untitled-prone. It is also hard to conclude the pattern of titled/untitled distribution for each negotiator across years. In Figure 3 we summarize the distribution of titled and untitled articles signed by the top 10 negotiators. The negative/positive numbers represent the counts of untitled/titled articles. We observe no clear pattern across countries and years, except for the fact that France and the Netherlands have negotiated more untitled titles than the titled ones.

The average length of titled articles is 217.39 tokens, the untitled 160.66 tokens. It can be seen that the length discrepancy between titled and untitled parts is not as large as expected since one might assume that untitled articles are much shorter than titled articles. The average length of article titles is 3.84 tokens. Besides, the translated texts in the untitled and titled parts are 23% and 17%, respectively.

We observe from 34,524 article titles that they differ from each other in word forms (singular or plural), in format (upper or lower case), in stop words (with or without) and in word order, etc. In order to gain a more condensed representation of distinctive titles, we lemmatized and lowercased titles, removed stop words and sorted words in titles in the alphabetical orders. This renders the list of 34,524 retrieved titles in 5,101 unique normalized forms. Table 2 lists three examples of unique normalized titles and their frequencies. The frequency distribution of normalized titles is extremely uneven: We have 65.3% of titles which appear only once; 33.2% of the titles appear more than once yet less than 50 times; only eleven normalized titles have been used very frequently, i.e. more than 500 times. As a result, it would be challenging to categorize articles from such an uneven frequency distribution. In the next sections, we discuss the choice of methods suitable to our data structure thoroughly.

3. Supervised Article Classification of the *Frequent* Articles

We firstly ran two classification tests with the articles from the most frequent twenty (*top20*) and fifty (*top50*) normalized titles. The sample of *top20* titles consists of 11,846 titled articles and 1,308 untitled articles; the *top50* sample 14,413 titled and 1,587 untitled articles. Both of the tests were implemented with the stochastic gradient descent classifier `SGDClassifier` in the Python `sklearn` library³ in a five-fold cross-validation setting. The tests have rendered the test accuracy of 96.04% and 95.02%, with the 500 and 1,000 TF/IDF⁴-weighted lowercased content words as features, respectively. The best set of hyperparameters is the learning rate of 0.00001, the loss function of hinge and the regularization of L1.

Although this method produces a high accuracy score for the articles with high frequencies, it cannot be used to classify titles that are not in the training set. In addition, this naive method only covers as training material less than a half of the titled articles.

4. Word and Article Embeddings

As discussed above, it is not feasible to use 5,101 classes in a categorization task; hence, we first need to compress the titles into more condensed, meaningful categories and then use those in our

³See <http://scikit-learn.org/stable/modules/sgd.html>, http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html (accessed 10 May 2017).

⁴Term Frequency/Inverse Document Frequency.

Table 3. Settings in partially supervised clustering

no. cluster	word embedding	average	initial centroids	vector composition	passes
9	pretrained	no weights	random	addition	100
10	retrained	TF/IDF weights	topic definitions	weighted addition	200
11					

text categorization task⁵.

As we have a relatively small corpus, expansion of corpus-customized word embeddings based on pretrained *Google News*⁶ embeddings is the key to success in clustering and classification tasks. For the titled corpus, we first generate the title embeddings and text embeddings separately and then construct document embeddings for articles through vector composition. We consider four criteria in computing article embeddings: (1) the type of word embeddings (pretrained or retrained, cf. [12]); (2) the choice of algorithm (*doc2vec* or *word2vec*⁷ (cf. [12, 13])); (3) the weighting scheme of word embeddings (the average of word vectors or the TF/IDF-weighted average of word vectors (cf. [11, 14])); (4) how to assign weights to titles and texts in additive and multiplicative functions (i.e. finding optimal weights α and β in $v_{article} = \alpha v_{title} + \beta v_{text}$, cf. [15]).

We selected the algorithm *word2vec* to create document embeddings because it has the capacity of differentiating similar title pairs from dissimilar ones. The other three criteria remain to be examined together with the results of article clustering (through the Silhouette coefficient, topic modeling and visualization) in Section 5.

For articles that share the same normalized title after preprocessing, the variability of texts is high. For a condense representation of texts sharing the identical normalized title, we took the average article embeddings for those articles. In the end, we obtained 5,101 article embeddings which we could use as features in k-means clustering.

5. Semi-Supervised Article Clustering of the Titled Articles

In our experiment, it is infeasible to use 5,101 classes when labeling the untitled articles because we cannot generalize our learned knowledge with a large number of classes. Consequently, we need to make use of document clustering to “compress” this list of normalized titles. We follow the method of weak supervision in document clustering as Aggarwal et al. [16] in 2004 explained in their experiment with unlabeled data. They termed the approach of using a priori knowledge as “centroids” *partially supervised clustering*, one type of semi-supervised learning. Ten main topics of IIAs are commonly agreed upon to form an exhaustive list of topics an investment treaty can cover, as summarized by Salacuse [1] in 2015. We calculate the document embeddings of the definitions from [1]⁸.

Table 3 gives an overview of possible settings we could test in the k-means clustering: the number of clusters, the type of word embeddings, weights for average, vector composition, initial centroids and passes.

Evaluation of k-means clustering is to decide the best partitioning method of the dataset. In relation to this, we discover the number of clusters with the help of our prior knowledge of IIA topics. In this experiment, we combine three methods to evaluate the results of clustering: (1) the Silhouette coefficient⁹ which measures the distance between the mean of instances in the same

⁵Throughout this work, we make a clear distinction among the three terms, *categorization*, *classification* and *clustering*. *Text classification* is defined as “trying to organize text documents into various categories based on inherent properties or attributes of each text document” [11] with supervised learning techniques. *Text clustering* is also known as *document clustering*, where documents are clustered into groups “purely based on their features, similarity and attributes, without training any model on previously labeled data” [11]. The term *text categorization* is referred here as the hypernym of the previous two terms; therefore it is used in thesis as the broadest term to address labeling texts with certain taxonomy.

⁶<https://code.google.com/archive/p/word2vec/> (accessed 24 April, 2017).

⁷<https://github.com/jhlau/doc2vec> (accessed 01 May 2017).

⁸Manger and Peinhardt [17] and Alschner and Skougarevskiy [2] have introduced the other existing taxonomies of categorizing the IIA content such as [18, 19] that can be used in the future work as the robustness checks of our chosen categories (i.e. “centroids” in clustering).

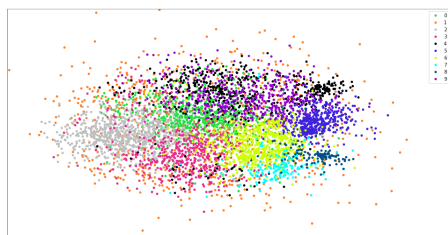
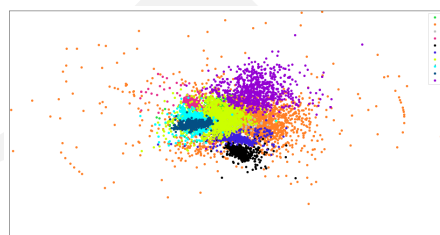
⁹http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (accessed 10 May 2017).

Table 4. Final list of topics in clustering

cluster	topics
0	Conditions for the entry of foreign investment and investors
1	Definitions and scope of application
2	Others (other political, economical,cultural, technological and scientific cooperation)
3	Operational and other conditions
4	Treaty entry, exceptions, modifications and terminations
5	Dispute settlement
6	General standards of treatment of foreign investments and investors
7	Monetary transfers
8	Compensation (expropriation and dispossession/losses from armed conflict or internal disorder)
9	International governance and regime in IIAs

Table 5. Keywords of ten clustered topics

cluster	count	keywords
0	761	"technical", "regulation", "transparency", "assessment", "procurement", "conformity"
1	641	"general", "application", "scope", "definition"
2	794	"trade", "economic", "environment", "social", "technology", "technical", "drug"
3	606	"operation", "material", "person", "intellectual", "access", "border", "cross",
4	554	"force", "entry", "duration", "termination", "amendment", "elimination", "restriction", "annexe"
5	537	"dispute", "settlement", "arbitration", "state", "court", "tribunal", "claim", "procedure", "resolution"
6	667	"national", "protection", "promotion", "nation", "obligation", "right",
7	162	"transfer", "capital", "repatriation", "payment", "profit", "income", "free", "revenue", "asset", "currency", "fund"
8	127	"compensation", "expropriation", "loss", "nationalization", "indemnification", "war", "nationalize", "deprivation", "dispossession"
9	446	"committee", "consultation", "joint", "council", "procedure", "commission", "panel", "meeting", "cooperation"

**Figure 4.** Best clustering settings for the *titled* articles**Figure 5.** Best clustering settings for the *untitled* articles

cluster (i.e. intra-cluster) and the mean of instances from the nearest cluster; (2) visualization with multidimensional scaling (MDS); (3) topic keywords generated by topic modeling using Latent Dirichlet Allocation (LDA cf. Blei et al. [20]).

MDS is applied to project a high-dimensional representation of data into a low-dimensional space and to analyze similarity or dissimilarity of data as "distances in a geometric space"¹⁰. Topic modeling is applied to each cluster of documents to generate 20 representative key words for that cluster.

We obtained the gold labels for 100 titled articles created by Prof. Dr. Peter Egger, an expert in IIAs and international trade. The expert was provided with the article titles and their corresponding texts, for which he chose one label from the given set of labels, i.e. the topics of the resulting clusters.

The clusters of the best settings using k-means (ten clusters, the average retrained word embeddings, the additive vector composition, the initialized centroids with topic definitions) are visualized in Figure 4. We also tested the best settings with the iteration passes of 100 and 200 and found out that compared with those of 100 passes, the clustering results of 200 passes have not changed substantially regarding the Silhouette coefficient, the MDS cluster visualization, and the topic keywords. Hence, we report only the results of 100 passes. Table 5 lists the keywords in each

¹⁰<http://scikit-learn.org/stable/modules/manifold.html#multidimensional-scaling> (accessed 10 May 2017).

Table 6. Accuracy for partially supervised clustering in each cluster for 100 titled instances

cluster	0	1	2	3	4	5	6	7	8	9
gold	5	31	9	2	3	7	8	20	10	5
accurate	0	1	4	0	1	7	4	20	10	3
accuracy	0.00%	3.23%	44.44%	0.00%	33.33%	100.00%	50.00%	100.00%	100.00%	60.00%

cluster. The cluster numbers and their corresponding topics can be found in Table 4. We adopted mostly the topics of IIAs from [1] and added two topics “others” and “international governance and regime”.

As we can see from Figure 4, the distribution of data points in the two-dimensional space exemplifies the characteristics of each topic and their relations with one another. We start from the left bottom, clusters 2, 0, 3, 6, 4, 8, 5 (cluster IDs cf. Table 4) have very condensed intra-cluster distributions. Data points belong to the seven clusters mentioned above group closer to their group members than to the data points from other topic clusters. The more fluid clusters are 9 and 4 (cluster IDs cf. Table 4) that partially overlap. This can be explained by their textual and legal interconnectivity with each other. For instance, “international governance and regime” (cluster 9) covers the principles, norms, rules, decision-making processes of IIAs; the provisions of “entry, exceptions, terminations, modifications” (cluster 4) can intertwine with cluster 9 because both clusters deal with the institutional provisions. The remaining clusters cover mainly the negotiated terms and conditions on the concrete issues and matters of international investment and investors. Cluster 1 is one very scattered one that spreads surrounding the other clusters. It indicates the topic “definitions and scope of application” where the definitions used throughout the treaties are explained, and the applicability of terms is specified. It is expected that definitions are composed of various terms which are then further specified in other articles of the same treaty. Therefore, cluster 1 can be rather scattered in the semantic space.

The overall accuracy of clustering is 50% across all clusters. From Table 6 we can see that accuracies in clusters 5, 7, 8 have reached 100%, followed by cluster 9 with an accuracy of 60%, cluster 6 with 50%, cluster 2 with 44.4% and cluster 4 with 33.33%. Only one instance out of cluster 1 fell into the correct cluster. Instances in clusters 0 and 3 have all been clustered wrongly. The clusters with the high accuracies have condensed clusters as shown in Figure 4.

Hence, it is not particularly surprising that cluster 1 has a lower accuracy, given the fact that it has a fluid intra-cluster structure. Each instance within that cluster can be linked to other clusters by addressing the basic definitions and the scope of conditions for other pertaining articles. Therefore, it is crucial to define our goal in the evaluation of clustering: If we are only interested in certain clusters, we shall pay attention the change of precision and recall in that cluster. Moreover, we should bear in mind that the clustering performance can directly influence the performance in the text classification tasks described in Section 6.

6. Supervised Article Classification of the *Untitled* Articles

The titled articles together with their assigned cluster membership were then be used as training and tuning material (34,524 titled articles in ten classes) for text classification. As we do not have access to the true label for each titled article, we simply assume that the assigned cluster membership can act as a proxy for the true label.

We applied six different classifiers, K-nearest neighbor (KNN), linear support vector machine (SVM), non-linear SVM, multi-layer perceptron (MLP), stochastic gradient descent (SGD) and convolutional neural network (CNN), to our training and tuning sets generated by k-means clustering in Section 5 (cf. [21–23] for more technical details). The first five classifiers were trained using

December 2017

Table 7. Overall accuracy for 100 untitled instances, / nonavailable, * Till writing the paper, run 2 epoches, acc: ~20%, randEMB: random embeddings

features	ML techniques						
	KNN	linear SVM	non-linear SVM	MLP	SGD	CNN	semi-k-means
BoW	4%	10%	15%	2%	11%	/	/
randEMB	/	/	/	/	/	46%	/
pretrainEMB	57%	53%	58%	52%	51%	/*	/
retrainEMB	18%	17%	18%	19%	17%	/*	60%

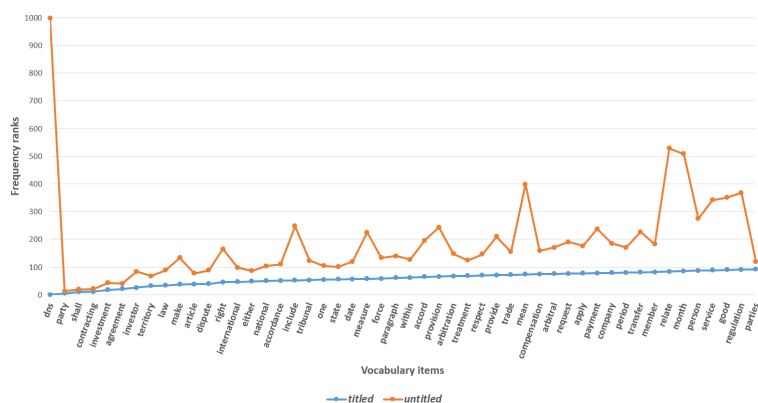


Figure 6. Top 50 BoW features in scikit-learn, benchmark: the titled corpus

scikit-learn, the last classifier, CNN¹¹, was trained with the TensorFlow¹² library. We used 5-fold cross-validation in the grid search¹³ to optimize the hyperparameters. We tested the classifiers on our untitled part of the corpus (10,074 articles, lemmatized, lowercased). We reported the results with two types of features: BoW (the most frequent 500 content words in the titled part) and word embeddings (pretrained and retrained), with the former as the baseline.

In order to evaluate the classifier performance on the untitled corpus, we randomly selected 100 untitled articles and obtained their human annotated labels from the expert annotation. The overall results of supervised learning are not ideal, with the highest overall accuracy of 58% (by non-linear SVM with the pretrained embeddings), followed by CNN with the random embeddings (46%), and by Gaussian kernel SVM 15% with the BoW features (cf. Table 7). We analyzed the classification precision/recall/f-score of each class in the best performing systems as shown in Table 8. It is obvious that systems with word embeddings outperformed those with the simple BoW features. The classes 5, 7, 8 have 100% recall in systems with word embeddings, which is in line with the training performance of partially supervised clustering in those classes (cf. Table 6). To further understand the failure of the BoW baseline, Figure 6 lists the most common 50 word features used in the baseline BoW models and the frequency distribution of those items in the titled and untitled parts. We can assume that the distribution discrepancy has led to the weaker prediction power of BoW on the untitled part, compared with the embedding features that take into account the context of words used in the same articles.

As a comparison with the text classification, we report the efficacy of using the retrained word embeddings from the titled corpus to cluster the untitled articles in the next section.

¹¹The best set of hyperparameters: randomized embeddings dimensions: 50, filter size: (3,4,5), 20 filters, dropout rate: 0.1, batch size: 64, 10 epoches.

¹²https://www.tensorflow.org/api_docs/python/ (accessed 20 April 2017).

¹³http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed 10 May 2017).

Table 8. Comparison of the precision, recall and f-score across the best systems, **P**: precision, **R**: recall, **F1**: f-score

class	non-linear SVM (BoW)			non-linear SVM (pretrainEMB)			CNN (randEMB)			semi-k-means (retrainEMB)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
0	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
1	13%	39%	20%	0%	0%	0%	0%	0%	0%	88%	39%	54%
2	14%	10%	12%	58%	70%	64%	100%	10%	18%	50%	70%	58%
3	0%	0%	0%	10%	11%	11%	0%	0%	0%	21%	33%	26%
4	0%	0%	0%	40%	67%	50%	67%	67%	67%	83%	83%	83%
5	100%	13%	22%	73%	100%	84%	73%	100%	84%	100%	88%	93%
6	50%	22%	30%	68%	83%	75%	44%	87%	59%	100%	43%	61%
7	0%	0%	0%	71%	100%	83%	83%	100%	91%	100%	60%	75%
8	0%	0%	0%	78%	100%	88%	78%	100%	88%	78%	100%	88%
9	25%	8%	12%	100%	54%	70%	100%	8%	14%	55%	85%	67%
avg	20%	9%	10%	50%	58%	52%	54%	47%	42%	68%	60%	60%

7. Semi-Supervised Article Clustering of the *Untitled* Articles

We used the untitled part of the corpus (lowercased, lemmatized) in a setting of *partially supervised k-means clustering*. The goal is to cluster the untitled articles based on their document embeddings computed by averaging word embeddings retrained on the titled part of the corpus, and to compare the accuracy in categorizing 100 untitled instances with that of the best supervised learning systems.

The identical setting from Section 5 that achieved the best cluster partition was applied here again: ten clusters, the average retrained word embeddings, the additive vector composition, the initialized centroids with topic definitions. The resulting clusters are visualized in Figure 5.

Similar to the results in the titled part of the corpus, articles from cluster 1 on “definitions and scope of application” spread across the space and encompass the other clusters. Clusters 0 (“entry of foreign investment and investors”) and 2 (“others”) are hidden behind cluster 6 (“general standards of treatment for foreign investment and investors”). The explanation for this might be that terms about foreign investment and investors can be covered both in cluster 0 or 6. For cluster 2, it can be due to the fact that there are not many instances from that cluster in the untitled part. The rest of the clusters have condensed intra-cluster distributions and are allocated nicely in the space.

We evaluated the clustering of the untitled articles with 100 annotated untitled instances. The overall accuracy of prediction has reached 60%, with an increase of 2% compared with 58% accuracy achieved by the non-linear SVM with pretrained embeddings.

Compared with the best supervised classifier (non-linear SVM (pretrainEMB)), the precision of each cluster trained by k-means clustering in Table 8 has increased at the cost of recall. In the setting of semi-supervised k-means clustering with retrained word embeddings, the average precision, recall and f-score are the highest amongst all systems.

Moreover, we also computed the keywords for each cluster which show the similar mapping patterns between the keywords and the ten topics as our mappings summarized in Table 5. Finally, the k-means clustering does not have a tendency of predicting a particular cluster label because it predicts the membership for clusters based on the intra-cluster similarity and inter-cluster dissimilarity.

For a better understanding of the difference in feature selection, we computed the average of the Jaccard distance¹⁴ and the normalized Levenshtein distance for each cluster in the evaluation set with 100 annotated instances for the untitled articles. This comparison discloses that even the articles belong to the same cluster have high lexical dissimilarity (higher than 70%). It is well known that word embeddings can capture certain linear semantic and syntactic regularities (cf. [24]); as a result, the k-means clustering which makes use of document embeddings has the

¹⁴ Alschner and Skougarevskiy [2] performed Jaccard distance on the character-level ngrams to measure the similarity and dissimilarity between the articles.

source format	source lang	text	gold label	BoW					pretrainEMB					randEMB	retrainEMB
				KNN	linear SVM	non-linear SVM	MLP	SGD	KNN	linear SVM	non-linear SVM	MLP	SGD	CNN	semi-k-means
msWord	bg	Each Contracting Party hereby consents to submit any legal dispute arising between that Contracting Party and a national of the other Contracting Party concerning an investment of that national ... In case a Contracting Party is not a Contracting State to the said Convention,...	5 Dispute settlement	5	5	1	1	5	5	5	5	5	5	5	3
html	en	1. Any dispute between an investor of one Contracting Party and the other Contracting Party in connection with an investments shall ... 2. If a dispute between one Contracting party and an investor of the other Contracting Party cannot amicably be settled within six (6) months, ... 3. An international arbitral tribunal shall be constituted ... 4. Each party to the dispute shall bear the cost ...		0	4	1	3	3	5	5	5	5	5	5	5
html	en	1. Any investment dispute shall be notified by the investor of one Contracting Party ... 2. In the absence of an agreement between the parties to the dispute within six months of the date of its notification, the dispute shall, ... 3. A Contracting Party which is a party to a dispute shall not, ...		0	4	0	4	4	5	5	5	5	5	5	5

Figure 7. Error analysis of three samples from class 5

advantage over the context-counting classifiers that learn from the BoW model. Because of this, the k-means algorithm outperformed the BoW non-linear SVM in cluster 1 with an increase of precision by 75%.

We also find out that for certain clusters (5, 7, 8) where the technical jargons are of particular use to the topics (e.g. jargons on “dispute settlement”, on “monetary transfer”) and rarely intertwine with other topics, both the non-linear SVM with pretrained embeddings and CNN with randomized embeddings can reach high recall. However, semi-k-means with retrained embeddings has not outperformed the best supervised systems in those clusters (cf. Figure 7). Furthermore, the supervised method has performed even better than the k-means clustering, because the lexical items in those clusters do not vary from article to article largely. Last but not least, we confirm that due to ease in configuration, retrained word embeddings customized to our SNIS corpus are better than randomly initialized embeddings, especially when our corpus is relatively small.

The comparison between the supervised and semi-supervised learners shows that the systems (e.g. non-linear SVM, CNN and k-means) that utilize word embeddings have the strongest predicting power as opposed to those classifiers which use only surface lexical and some distributional features.

8. Conclusions and Future Work

This paper is an endeavor devoted to an interdisciplinary research topic: how to better understand the inherent structures of IIAs. As the first step to explore the structure of IIAs with ten topics, this work has enhanced our understanding of the applicability of text categorization, be it classification or clustering, to capture the inherent content structure.

We have devised a pipeline which extracted and preprocessed the titled and untitled articles (34,524 and 10,047 snippets, respectively) from 2,823 treaties in the SNIS corpus. In order to expand the word semantics in our domain-specific corpus, we retrained the word embeddings with the pretrained embeddings from the *Google News* corpus. We then performed partially supervised clustering where we compressed the document semantics of 5,101 unique formal titles and their corresponding texts and then generated the article labels (out of the ten topics) for the titled part of the corpus.

We then compared the efficacy of the supervised and semi-supervised techniques with BoW and embeddings as features in the same learning problem. This led us to the conclusion that the k-means clustering with the retrained word embeddings customized to the SNIS corpus has brought about an increase of 30% in accuracy compared to a simple CNN classifier which has outperformed the other five supervised learners with BoW features. K-means clustering also outperformed the best classifier with pretrained word embeddings (+ 2%).

In a nutshell, systems with word embeddings as features outperform those with BoW features. Semi-supervised system outperforms the supervised systems and is easier to train: the best

December 2017

supervised system (non-linear SVM) with pretrained word embeddings could not outperform the semi-supervised k-means clustering with retrained word embeddings tailored to the IIA corpus.

This work has highlighted the importance of expanding the semantic features of documents in text categorization. The present findings have important implications for improving the supervised classifier: We could use the retrained word embeddings as features in a supervised setting; we could initialize the word embeddings in a CNN classifier with our retrained representations.

We hope that this work will be beneficial to the construction of IIA database in the future because it has tested different techniques to decipher the structure of IIAs by categorizing text snippets into the interlinking topics from the same domain.

It should also be examined, whether we can perform partially supervised clustering with the whole corpus (the titled and untitled parts included) and assign the labels for the untitled articles based on their cluster membership, as we have access to the article titles of the titled part. It would also be interesting to compare the output of affinity propagation (AP) (where no number of clusters should be specified) and that of the k-means clustering. We have also found out that the CNN classifier and the k-means clustering perform differently in various topics. It remains to be tested if topic-specific learning techniques should be devised to tackle the variability of semantics and syntax in each topic. It would be worth testing whether using the annotated texts (e.g. part-of-speech tagged, syntactically parsed) can improve text categorization. These topics are reserved for our future work.

References

- [1] Jeswald W Salacuse. *The Law of Investment Treaties*. OUP Oxford, 2015.
- [2] Wolfgang Alschner and Dmitriy Skougarevskiy. Mapping the universe of international investment agreements. *Journal of International Economic Law*, pages 561–588, 2016.
- [3] Kyoko Sugisaki, Martin Volk, Rodrigo Polanco, Wolfgang Alschner, and Dmitriy Skougarevskiy. Building a corpus of multi-lingual and multi-format international investment agreements. In *Legal Knowledge and Information Systems - JURIX 2016: The 29th Annual Conference*, pages 203–206, 2016.
- [4] Wolfgang Alschner and Dmitriy Skougarevskiy. Treaty texts as data-developing new tools for negotiators and litigators to compare bilateral investment treaties. In *Legal Knowledge and Information Systems - JURIX 2015: The 28th Annual Conference*, pages 141–144, 2015.
- [5] Wolfgang Alschner and Dmitriy Skougarevskiy. Rule-takers or rule-makers? a new look at african bilateral investment treaty practice. Technical Report 7, World Trade Institute (University of Berne), Swiss National Centre of Competence in Research, 6 2016.
- [6] Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Claudia Soria. *Automatic Classification and Analysis of Provisions in Italian Legal Texts: A Case Study*, pages 593–604. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
- [7] Emile de Maat and Radboud Winkels. Automatic classification of sentences in Dutch laws. In E. Francesconi et al., editor, *Legal Knowledge and Information Systems*, volume 189, pages 207–216. IOS Press, 2008.
- [8] Emile de Maat and Radboud Winkels. A next step towards automated modelling of sources of law. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 31–39, New York, NY, USA, 2009. ACM.
- [9] Emile de Maat and Radboud Winkels. *Automated Classification of Norms in Sources of Law*, pages 170–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
- [10] Wolfgang Alschner and Dmitriy Skougarevskiy. Convergence and divergence in the investment treaty universe-scoping the potential for multilateral consolidation. *Trade L. & Dev.*, 8:189, 2016.
- [11] Dipanjan Sarkar. *Text Analytics with Python: A Practical Real-world Approach to Gaining Actionable Insights from Your Data*. Apress, 2016.
- [12] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany, August 2016. ACL.
- [13] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, volume 14, pages 1188–1196, 2014.
- [14] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Dls@cu: Sentence similarity from word alignment and semantic vector composition. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 148–153, 2015.
- [15] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Proceedings of the 8th ACL: Human Language Technology Conference (HLT)*, pages 236–244, Columbus, Ohio, USA, June 2008.
- [16] Charu C Aggarwal, Stephen C Gates, and Philip S Yu. On using partial supervision for text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):245–255, 2004.
- [17] Mark S. Manger and Clint Peinhardt. Learning and the precision of international investment agreements. *International Interactions*, pages 1–21, 2017.
- [18] BILATERAL INVESTMENT TREATIES UNCTAD. Treaties 1995-2006: Trends in investment rule making, un doc. Technical report, UNCTAD/ITE/IIT/2006/5, UN Sales No. E. 06. II. D. 16, 2007.
- [19] Chester Brown. *Commentaries on Selected Model Investment Treaties*. OUP Oxford, 2013.
- [20] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [21] Sebastian Raschka. *Python Machine Learning*. Packt Publishing, Birmingham, UK, 2015. ISBN 1783555130.
- [22] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 17461751, Doha, Qata, 2014. ACL.
- [23] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. 2015. URL <https://arxiv.org/abs/1510.03820>.
- [24] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, volume 13, pages 746–751, 2013.