Textual Entailment for Cybersecurity: An Applicative Case

Giovanni Siragusa
University of Turin, Turin, Italy
siragusa@di.unito.it

Livio Robaldo
Legal Innovation Lab Wales, Swansea University, United Kingdom
Nomotika SRL, Turin, Italy
livio.robaldo@gmail.com

Luigi Di Caro
University of Turin, Turin, Italy
Nomotika SRL, Turin, Italy
dicaro@di.unito.it

Andrea Violato
Nomotika SRL, Turin, Italy
andrea.violato@nomotika.it

Abstract

Recognizing Textual Entailment (RTE) is the task of recognizing the relation between two sentences, in order to measure whether and to what extent one of the two is inferred from the other. It is used in many Natural Language Processing (NLP) tasks. In the last decades, with the digitization of many legal documents, NLP applied to the legal domain has became prominent, due to the need of knowing which norms are complied with in case other norms are. In this context, from a set of obligations that are known to be complied with, RTE may be used to infer which other norms are complied with as well. We propose a dataset, regarding cybersecurity controls, for RTE on the legal domain. The dataset has been constructed using information available online, provided by domain experts from NIST (https://www.nist.gov).

Vol. 8 No. 4 2021
Journal of Applied Logics — IfCoLog Journal of Logics and their Applications

1 Introduction

It is well-known that laws can be pragmatically interpreted in multiple, and often incompatible, ways, even in the same context. Handling multiple interpretations of legal norms is perhaps the best known problem in Legal Informatics.

On the one hand, since it is impossible to predict a priori every possible context where the norms will be deployed, legislators tend to use vague terms that are flexible enough to be adapted to a multitude of contexts and, within certain limits, to the technological advancements of the society.

On the other hand, what makes legal texts so much dependent on subjective human interpretation is that they are used in disputes that represent different interests, so that the interpretation of the norms tends to be stretched depending on the interest involved.

It is eventually up to judges and other appointed authorities to decide the interpretation of norms in context. According to the seminal work in [13], legal authorities expand or restrict the core of determinate meaning of norms by filling legal gaps to connect *legal requirements* (formal compliance) and *operational requirements* (substantive compliance), i.e., how and to what extent the legal requirements from legislation are met in real-world scenarios.

More generally, the connection from legal to operational requirements recalls the notion of "concept holism" (see [7, 11, 27, 3], among others): one cannot say to have the complete meaning of a legal requirement without knowing the whole system of constitutive rules and the web of concepts with which the meaning of that requirement is intertwined.

In order to take decisions about the interpretation of norms, judges often consult the relevant literature in the area. For this reason, other legal authorities, standardization bodies, or associations representing categories of involved entities produce additional documents that contain recommendations, guidelines, standards, etc. specifying how to be compliant with the legislation in specific situations. In many cases, this is even explicitly required by the legislation itself, as in the case of the General Data Protection Regulation (GDPR), which requires controllers to define their own data protection policies (cf. GDPR, Artt. 13, 14, and 24(2)), invites associations and other bodies representing categories of controllers or processors to prepare codes of conducts (see, e.g., GDPR, Art. 40), the European Data Protection Board has the duty to release guidelines and recommendations (see, e.g., GDPR, Art. 70(1)(d)), etc. See discussion in [22] and [25].

Recommendations, guidelines, standards, etc. are not typically part of legislation; therefore, their adoption do not automatically provide compliance with the regulations. However, by certifying the adoption of a standard, an organisation can

argue in favour of its proactive attitude and best efforts to be compliant with the regulations. In other words, such certifications provide strong arguments of compliance to be possibly used in auditing procedures or even in court.

On the other hand, since operational requirements are usually scattered around several documents in different format released, at different times, by different associations and other bodies, with different authoritative power and reputation in a certain domain, finding correlations between legal and operational requirements requires to build, maintain, and analyze an up-to-date archive of all these documents, which may be rather time-consuming, burdensome, and, therefore, unmanageable.

In light of this, Natural Language Processing (NLP) applications, in particular Textual Entailment (TE) applications [16], can provide valid help in creating and maintaining such an holistic network of legal/operational requirements, specifically in identifying when a requirement semantically entails another one.

The main problem of TE regarding legal documents is the availability of dataset used to train machine learning algorithms to recognize the relation expressed. The few existing ones are generally based on case laws, as the one proposed in Competition on Legal Information Extraction/Entailment (COLIEE) Workshop[1]. There is no dataset regarding standard procedures that a company has to implement to protect their data, where the adoption of TE techniques are crucial to verify if they have been defined.

Such procedures are generally defined by ISO[2] (the International Organization for Standardization). The standard includes 114 *controls*, that a company needs to check in order to consider itself as "secure" enough from cybersecurity attacks. In order to assess compliance with the standard, a company hires specialized auditors, who, after an inspection, decides whether the company is compliant with the standard or has to revise some of its internal business processes.

However, the ISO/IEC 27001:2013 controls, expressed in Natural Language, are quite vague and leave plenty of room for subjective interpretations.

For this reason, several public institutions, e.g., NIST[3] (National Institute of Standards and Technology), release more context-specific standards that refine the ISO/IEC 27001:2013. In this paper, we focus in particular on the NIST 800-53 rev.4[4], which implements 256 controls while specifying how they relate to the 114 controls of ISO/IEC 27001:2013 and viceversa. Specifically, it contains annexes that explain which controls of one of the two standards are satisfied by the controls of the other (and viceversa), in the sense that if a company implements one of the two,

[1] https://sites.ualberta.ca/~rabelo/COLIEE2019/

[2] https://www.iso.org/home.html

[3] https://www.nist.gov

[4] https://nvd.nist.gov/800-53

then it is assumed that the company also implements the associated controls in the other standard.

The present paper starts from the assumption that the ISO/IEC 27001:2013 and NIST 800-53 rev.4, and, in particular, the annexes included in the latter, which specify correspondences between the two standards, are precious raw sources for building a dataset for RTE. The latter has been recently identified in [4] as a challenging research topic.

Note that ISO/IEC 27001:2013 and NIST 800-53 rev.4 are just the two running examples that we will use in this paper. Many other cybersecurity standards are available on the Web, as well as corresponding tables inter-linking their controls. In other words, this paper has to be considered as the first step of a bigger research project to create a dataset made of a *network* of inter-connected technical documents in the cybersecurity domain. The advocated dataset, and the RTE classifiers trained, tuned, and evaluated on it, would be a precious resource for cybersecurity auditors and companies collaborating with them, e.g., Nomotika SRL.

In this paper, we propose:

- A dataset for RTE regarding cybersecurity. We constructed the dataset using the correspondences between controls that we found in the ISO/IEC 27001:2013 and NIST 800-53 rev.4.
- An evaluation of several RTE classifiers on the dataset, where we conducted a three-step evaluation. In the first step, we evaluated the dataset using cross-fold validation to see if it could be used to train the RTE classifiers. In the second step, we trained the classifiers using the COLIEE dataset[5] for Legal Textual Entailment; the idea is to check whether it is possible to transfer the knowledge acquired from a domain-oriented dataset to our one. Finally, in the third step, we checked if it is possible to transfer the knowledge acquired on our dataset to other legal ones.

The remainder of the paper is structured as follows: Section 2 describes the construction of the dataset, reporting the number of pairs it contains, the average length of sentences and the vocabulary size; Section 3 describes the used models and their performance on the two datasets. Section 4 describes some related works on legal domain and RTE. Section 5 concludes the paper.

[5] `https://sites.ualberta.ca/~rabelo/COLIEE2019/`

2 A dataset for Recognizing Textual Entailment (RTE)

We defined a dataset for RTE in the cybersecurity domain, called *cybersecurity entailment*. We constructed the dataset using ISO controls covered by the NIST ones[6], and the controls of the same NIST document[7] covered by the ISO/IEC 27001. For NIST and ISO documents, each <NIST control, ISO control> pair is constructed using the table[8] reported in the NIST document, and it could be seen as an entailment pair. We then extended the entailment pairs with neutral ones, applying a cartesian product between NIST and ISO controls and removing duplicated ones. An example of positive <NIST control, ISO control> pair follows:

NIST: "*The organization employs the principle of least privilege, allowing only authorized accesses for users (or processes acting on behalf of users) which are necessary to accomplish assigned tasks in accordance with organizational missions and business functions.*"

ISO: "*The allocation and use of privileged access rights shall be restricted and controlled.*"

The following one is an example of neutral <NIST control, ISO control> pair:

NIST: "*The information system enforces approved authorizations for logical access to information and system resources in accordance with applicable access control policies.*"

ISO: "*Users shall ensure that unattended equipment has appropriate protection.*"

We repeated the process of constructing the pairs for the controls between the NIST document and the ISO/IEC 27001. In this case, each ISO/IEC control has a *related_to* tag that expresses connections with other NIST ones, reporting their IDs. The IDs in each tag are assigned by a domain expert. We used the tag to construct the entailment pairs. We then extended the set with neutral pairs as for the previous set. A <NIST control, ISO/IEC control> pair that expresses an entailment relation follows:

sent.: *The organization implements a tamper protection program for the information system, system component, or information system service.*

[6] `https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf`, appendix H

[7] `https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf`, appendix F

[8] The table is created by a domain expert when the document is redacted.

related_to: *The organization protects against supply chain threats to the information system, system component, or information system service by employing as part of a comprehensive, defense-in-breadth information security strategy.*

The following pair is an example of neutral <NIST control, ISO/IEC control> pair:

sent.: *The information system maintains a separate execution domain for each executing process.*

related_to: *The information system separates user functionality (including user interface services) from information system management functionality.*

Finally, we merged the two sets of pairs to create the *cybersecurity entailment* dataset. We balanced the resulting dataset in order to have the same number of entailment and neutral pairs. An interesting fact is that the constructed dataset does not contain contradiction pairs because a control cannot be in contraposition with an another one.

Table 1 reports the number of pairs, the average sentence length, and the size of unique terms. The table highlights that the vocabulary of neutral pairs is contained in the entailment one. We also report the frequency of the Part-Of-Speech (POS) tags in Figure 1. We used OpenNLP[9] to assign the POS tags. From the image, it is possible to see that the majority of words are nouns, followed by adjectives.

The proposed dataset is availabled at `https://drive.google.com/drive/folders/1swYciO8yOtaM1pCTS9ySEpNZ-Ac8A569?usp=sharing`

	# pairs	avg. sentence length	vocabulary size
all dataset	2898	110.0	1912
entailment	1449	115.37	1912
neutral	1449	104.59	1905

Table 1: The table reports the number of pairs, the average sentence length, and the size of unique terms of the *cybersecurity entailment* dataset.

2.1 XML representation of the dataset

We stored the dataset into an XML file to simplify sharing and interoperability. We encapsulated each <premise, hypothesis> pair inside the *pair* tag. In each tag, the first element of the pair is contained in the *t* tag (the premise) and the second element in the *h* tag (the hypothesis). Furthermore, each *pair* tag has the attribute *entailment* which expresses the

[9] `https://opennlp.apache.org`

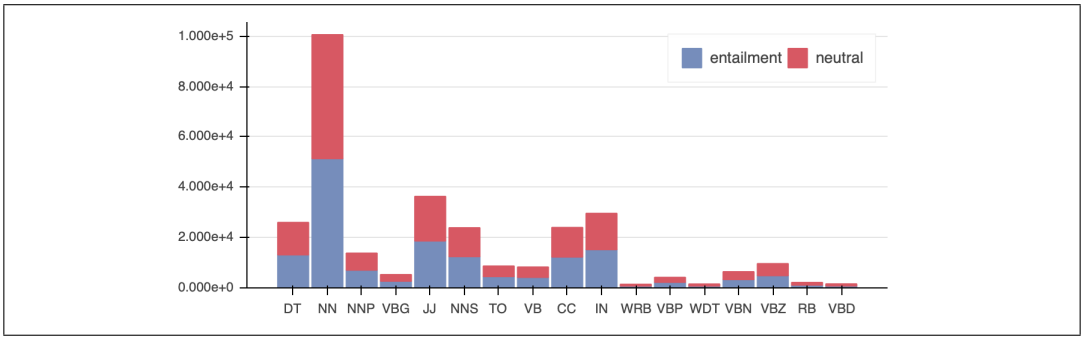

Figure 1: The POS frequency of the *cybersecurity entailment* dataset.

relation: entailment or neutral. It also has two other attributes: *id* which is an identifier of the pair, and *task* which is required by Excitement Open Platform (EOP) framework [17, 21]. All those entries are contained under the tag *entailment-corpus*. Figure 2 depicts an excerpt of the xml.

The main advantage of such XML structure is that it can be passed as an input to the EOP framework (or another one for TE) in order to train a classifier.

3 Evaluation Study of the Cybersecurity Dataset

In this section, we will perform three different analysis on our dataset:

- we will evaluate whether it is possible to train a classifier for RTE in order to recognize the entailment relation expressed in the <NIST, ISO> pair. For this evaluation, we will perform a cross-fold validation on our dataset;
- we will evaluate whether it is possible to transfer the knowledge acquired from another dataset to our own. In this evaluation, we would like to check the complexity of our dataset, i.e. if the relation expressed in the pairs can be easily recognized;
- we will evaluate whether it is possible to use our dataset to recognize the TE relations present in another dataset for legal domain, i.e. if a classifier trained on our dataset can generalize on unseen data. For this evaluation, we will train the classifiers on the *cybersecurity entailment* dataset and we will test them on the COLIEE testset.

.

We will follow RTE evaluation using *accuracy* measure: the ratio between the number of test instances correctly predicted and the number of test instances.

3.1 Preprocessing of the Sentences

In order to train the classifiers, we have to process the dataset sentences to extract the relevant features. The preprocessing consists of the following steps: tokenization, stopwords removal, stemming and feature extraction.

```
<?xml version="1.0" ?>
<entailment-corpus lang="EN">
  <pair entailment="ENTAILMENT" id="0" task="IR">
    <h>The organization retains audit records for [Assignment: organization-defined time period consistent with records retention policy] to provide support for after-the-fact investigations of
      security incidents and to meet regulatory and organizational information retention requirements.</h>
    <t>The information system provides the capability for authorized users to select a user session to capture/record or view/hear.</t>
  </pair>
  <pair entailment="ENTAILMENT" id="1" task="IR">
    <h>The organization: Develops and implements anti-counterfeit policy and procedures that include the means to detect and prevent counterfeit components from entering the information system; and
      Reports counterfeit information system components to [Selection (one or more): source of counterfeit component; [Assignment: organization-defined external reporting organizations];
      [Assignment: organization-defined personnel or roles]].</h>
    <t>The organization protects against supply chain threats to the information system, system component, or information system service by employing [Assignment: organization-defined security
      safeguards] as part of a comprehensive, defense-in-breadth information security strategy.</t>
  </pair>
  <pair entailment="ENTAILMENT" id="2" task="IR">
    <h>The organization: Develops, approves, and maintains a list of individuals with authorized access to the facility where the information system resides; Issues authorization credentials for
      facility access; Reviews the access list detailing authorized facility access by individuals [Assignment: organization-defined frequency]; and Removes individuals from the facility access
      list when access is no longer required.</h>
    <t>The organization implements an insider threat program that includes a cross-discipline insider threat incident handling team.</t>
  </pair>
  <pair entailment="ENTAILMENT" id="3" task="IR">
    <h>The organization: Authorizes connections from the information system to other information systems through the use of Interconnection Security Agreements; Documents, for each interconnection,
      the interface characteristics, security requirements, and the nature of the information communicated; and Reviews and updates Interconnection Security Agreements [Assignment:
      organization-defined frequency].</h>
    <t>The organization: Requires that providers of external information system services comply with organizational information security requirements and employ [Assignment: organization-defined
      security controls] in accordance with applicable federal laws, Executive Orders, directives, policies, regulations, standards, and guidance; Defines and documents government oversight and
      user roles and responsibilities with regard  to external information system services; and Employs [Assignment: organization-defined processes, methods, and techniques] to monitor security
      control compliance by external service providers on an ongoing basis.</t>
  </pair>
  <pair entailment="ENTAILMENT" id="4" task="IR">
    <h>A set of policies for information security shall be defined, approved by management, published and communicated to employees and relevant external parties.</h>
    <t>The organization: Develops, documents, and disseminates to [Assignment: organization-defined personnel or roles]: Reviews and updates the current: An incident response policy that addresses
      purpose, scope, roles, responsibilities, management commitment, coordination among organizational entities, and compliance; and Procedures to facilitate the implementation of the incident
      response policy and associated incident response controls; and Incident response policy [Assignment: organization-defined frequency]; and Incident response procedures [Assignment:
      organization-defined frequency].</t>
  </pair>
  <pair entailment="ENTAILMENT" id="5" task="IR">
    <h>The organization: Configures the information system to provide only essential capabilities; and Prohibits or restricts the use of the following functions, ports, protocols, and/or services:
      [Assignment: organization-defined prohibited or restricted functions, ports, protocols, and/or services].</h>
    <t>The information system enforces approved authorizations for controlling the flow of information within the system and between interconnected systems based on [Assignment: organization-define
      information flow control policies].</t>
  </pair>
```

Figure 2: The image shows a small section of the xml.

We start by computing the Part-Of-Speech (POS) tags of the words. Those are necessary to extract the features in the last step. We used OpenNLP[10] to obtain the POS tags of each word. Then, we tokenized the sentences using the NLTK[11] module. We filtered the stopwords out using the list provided by this latter framework. We also used a regular expression to remove all non-alphanumerical tokens because they are not relevant to recognize the TE relation. Finally, we stemmed and lowercased the remaining tokens to obtain a less diversified vocabulary.

Once the list of salient tokens is obtained through the above mentioned steps, we proceeded to extract the features for the classifiers. We first selected the POS tags corresponding

[10] `https://opennlp.apache.org`

[11] `https://www.nltk.org`

to the remaining words; then, we computed words n-grams and POS n-grams with n comprises in the range $[1, 5]$. We used a Term Frequency - Inverse Document Frequency schema to weight the extracted n-grams and to obtain the features.

3.2 Ablation Study

In the previous section, we said that the classifiers will use both word n-grams and POS n-grams. In this section, we will analyze the impact of these features on the performances. For this evaluation, we will compare the *Support Vector Machine* (SVM) classifier with the *Maximum Entropy* (ME) of EOP framework since both generally perform well in RTE tasks. We will use a *Random* classifier, which assigns the entailment relation with a probability of 0.5, as a baseline one.

We will train and test the classifiers on the *cybersecurity entailment* dataset to check the impact of the features. Since we do not have a testset, we will perform a cross-fold evaluation, with the fold number sets to 10. Each fold contains about 175 TE pairs. In detail, we will train the classifier on 9 folds and test on the remaining one. We will repeat this process leaving out a different fold for the test. Each classifier will be trained using the following features:

unigram: The classifier uses only unigrams as features. We decided to use such features as a baseline;

n-grams: The classifier uses n-grams as features, with n comprises in the range $[1, 5]$;

n-grams + POS: The classifier uses both word n-grams and POS tag n-grams, with n comprises in the range $[1, 5]$.

Table 2 reports the result of this evaluation. We can see that the n-grams slightly increased the accuracy of both classifiers. The accuracy is further increased with the adoption of the POS n-grams. It is interesting to notice that the n-grams had a major impact on the ME classifier than on the SVM one.

Model		Accuracy
SVM	+ unigrams	82.12%
	+ n-grams	82.58%
	+ n-grams + POS	83.06%
ME	+ unigrams	82.04%
	+ n-grams	82.75%
	+ n-grams + POS	83.10%

Table 2: The table reports the accuracy of ME and SVM classifiers with the different features.

3.3 Evaluation

In this section, we will evaluate several classifiers to check which one performs better in recognizing the expressed TE relation. The proposed classifiers are:

Random: it assigns a label randomly to each pair in the *cybersecurity entailment* dataset, which has a fixed accuracy of 50%. We used this classifier as baseline;

SVM: a Support Vector Machine that uses the extracted features (word and POS n-grams) to classify the pairs;

NB: a Naive Bayes classifier that uses the same features of SVM;

RF: since each pair of premise and hypothesis could contain specific words or POS tags that bring the entailment or neutral relation out, we decided to use a Random Forest classifier to capture them. The Random Forest creates a decision tree in which each branch contains word and POS n-grams features useful to distinguish the relation;

ME: a Maximum Entropy classifier with word and POS n-grams features. We used the implementation provided by the EOP framework since it contains state-of-the-art methods and classifiers for the TE task;

ME+WN+VO: it extends the features of the previous Maximum Entropy classifier with Wordnet [18] synsets (WN) and Verb Ocean [10] (VO) classes, i.e. a semantic network for verbs. VO reports for each verb: (1) the semantic relation with other verbs (e.g., to make and to create have a similarity relation), (2) if the verb is transitive and (3) if it is symmetric. Those features are used to handle possible periphrases and synonyms in the pairs. As for *ME*, we used the implementation provided by the EOP framework.

For the SVM, Naive Bayes and Random Forest classifiers, we used their implementation provided by the scikit-learn framework[12].

For the first evaluation, we decided to check if it is possible to train a classifier on the *cybersecurity entailment* dataset and generalize on similar data. Since we do not have a testset, we used the cross-fold validation. We divided the dataset in 10-fold, training the classifiers on nine folds and testing on the remaining one. Table 3 reports the results of this evaluation.

From the table, we can see that the SVM, the ME and the ME+WN+VO classifiers are able to recognize the relations expressed in those pairs, obtaining outstanding results. We can also notice that both the ME classifiers obtained an accuracy higher than the SVM, about 0.04 percentage points; since such difference is not significant, it is possible to use either the SVM or the ME. Both ME and ME+WN+VO classifiers have the same accuracy, meaning that the addition of WordNet synsets and Verb Ocean classes to the features is not relevant to recognize the TE relation.

We analyzed the errors made by SVM and ME classifiers. We found that they tend to mistake a neutral relation for an entailment one when both the premise and the hypothesis regard different topics of the same argument (e.g., auditing records storage vs. auditing events). An example of missclassified pair follows:

[12] `https://scikit-learn.org/stable/`

Classifier	Avg. Accuracy
Random	50%
SVM	83.06%
NB	82.90%
RF	79.42%
ME	83.10%
ME+WN+VO	83.10%

Table 3: The table reports the average accuracy for the cross-fold evaluation.

NIST: "*The organization: Schedules, performs, documents, and reviews records of maintenance and repairs on information system components in accordance with manufacturer or vendor specifications and/or organizational requirements; Approves and monitors all maintenance activities, whether performed on site or remotely and whether the equipment is serviced on site or removed to another location; Requires that [Assignment: organization-defined personnel or roles] explicitly approve the removal of the information system or system components from organizational facilities for off-site maintenance or repairs; Sanitizes equipment to remove all information from associated media prior to removal from organizational facilities for off-site maintenance or repairs; Checks all potentially impacted security controls to verify that the controls are still functioning properly following maintenance or repair actions; and Includes [Assignment: organization-defined maintenance-related information] in organizational maintenance records.* "

ISO: "*The organization: Documents and monitors individual information system security training activities including basic security awareness training and specific information system security training; and Retains individual training records for [Assignment: organization-defined time period].* "

In the proposed example, both the controls regard the *information system*, but they are not related to the same topic. The NIST one describes that the organization should maintain documents regarding maintenance activities and changes to the information systems, while the ISO one regards the training activities on the information systems.

We conducted a second evaluation to see whether it is possible to train the classifiers on a dataset, and use such acquired knowledge to recognize the relation expressed in our own one. In other words, we would like to verify if it is possible to generalize on our TE pairs. For this evaluation, we trained the classifiers using the legal textual entailment dataset proposed in COLIEE 2019[13] task 2. The dataset is composed of 362 pairs, divided into 182 pairs that express an entailment relation and 182 that express a contradiction one. Since this dataset does not present any neutral relation, we treated the neutral pairs of our dataset as negative ones to perform a proper evaluation. We applied the same preprocessing phase

[13] `https://sites.ualberta.ca/~rabelo/COLIEE2019/`

to the COLIEE trainset. Table 4 reports the results of this second evaluation. From the table, we can notice that only the SVM classifier slightly surpassed the Random one. Those results highlight the fact that our cybersecurity dataset contains complex pairs, making hard to generalize from a legal TE dataset to our one.

Classifier	Accuracy
Random	50%
SVM	50.48%
NB	49.44%
RF	48.86%
ME	50%
ME+WN+VO	50%

Table 4: The table reports the results obtained training the classifier on COLIEE dataset and testing on our one.

Finally, we conducted a third evaluation to see whether it is possible to transfer the knowledge that the classifier acquired from our dataset to other legal-based ones. For this experiment, we decided to evaluate the classifiers on the COLIEE testset. For a completed evaluation, we also reported the accuracy of the classifiers when they are trained and tested on only the COLIEE one. We expect that the accuracy will be high in this latter case, surpassing certainly the random classifier, while being lower for the generalization from the *cybersecurity entailment* dataset to the COLIEE one. Table 5 reports the results of this last evaluation.

Trainset	Classifier	Accuracy
-	Random	50%
COLIEE	SVM	71.11%
	NB	64.44%
	RF	67.80%
	ME	71.20%
	ME+WN+VO	71.06%
Cybersecurity	SVM	45.55%
	NB	46.70%
	RF	47.00%
	ME	47.00%
	ME+WN+VO	47.00%

Table 5: The table reports the results obtained training the classifier on COLIEE dataset and testing on our one.

As we expected, the accuracy of the classifiers trained and tested on COLIEE dataset surpassed the Random one. However, if we train them on our dataset, we obtain very poor performances; in this latter case, the classifiers have a lower accuracy, meaning that they found difficult to generalize on unseen data. Those results confirm again that our dataset contains more complex and semantic distant pairs than the COLIEE ones.

This could be verified computing the cosine distance between premise (or hypothesis) sentences of our dataset with the ones of COLIEE. More in detail, we calculated the average cosine distance between the premise (or hypothesis) sentences of the *cybersecurity entailment* dataset and the COLIEE ones. Table 6 reports the cosine distance for both the premise and hypothesis. From the table, it is possible to see that the two datasets do not have neither a jargon nor a syntactic structure in common.

Pairs	Cosine Distance
Premise	0.97
Hypothesis	0.95

Table 6: This table shows the cosine distance between the Cybersecurity dataset and the COLIEE one.

We report a distant pair for both the Premise and the Hypothesis. Table 7 shows a Premise pair and its score, while Table 8 shows an Hypothesis pair.

4 Related Works

Nowadays, Recognizing Textual Entailment (RTE) is an interesting task since it predicts the relation of two sentences. For instance, in legal domain could be used to see whether a law has a relation (i.e., entails) another one, or in case of European Union, we can see whether a law of a member state implements a directive of the EU (cf. [20]).

In general, research on generic RTE is conducted with the use of Neural Networks, where one important research works is [8]. In the article, the authors proposed both a dataset constructed through crowd-sourcing, and a Multi-Layer Perceptron (MLP) to classify the relation of a <premise, hypothesis> pair. After this article, researcher started to experiment with deep-learning models, also re-adapting idea coming from different NLP fields, such as Machine Translation. [26] proposed an encoder with attention for textual entailment. First, the authors sequentially read the premise and the hypothesis tokens with an LSTM, producing a list of encoded representation for the words. Then, they applied an attention mechanism to understand the correlation between the premise and the hypothesis words. They found that the attention mechanism is able to capture small semantic difference (e.g., the colour) in similar sentences. Finally, [29, 12, 24] found that, for some datasets, the hypothesis is all you need. According to them, the hypothesis contains very salient information that can be used by a Neural Network to unravel the relation. [23] used such models to evaluate Natural Language Inference Problems, defining several evaluation frameworks. Other works related to the legal domain include [6, 19, 2, 1].

Cybersecurity	The organization: Establishes and makes readily available to individuals requiring access to the information system, the rules that describe their responsibilities and expected behavior with regard to information and information system usage; Receives a signed acknowledgment from such individuals, indicating that they have read, understand, and agree to abide by the rules of behavior, before authorizing access to information and the information system; Reviews and updates the rules of behavior [Assignment: organization-defined frequency]; and Requires individuals who have signed a previous version of the rules of behavior to read and re-sign when the rules of behavior are revised/updated.
COLIEE	The proceeding at issue was a Motion for Summary Judgment under the previous Rules of the Court in regard to summary judgments. The Rules have been amended prior to the hearing of the motion to provide for summary trials but those amendments had no material effect on this matter.
Distance score	0.75

Table 7: The table reports a Premise sentence pair with its cosine distance score.

Cybersecurity	Information security shall be addressed in project management, regardless of the type of the project.
COLIEE	Further, this Court and the Federal Court of Appeal have both observed that a statement of claim should only be struck in the clearest and most obvious of cases.
Distance score	0.79

Table 8: The table reports an Hypothesis sentence pair with its cosine distance score.

Other researchers, instead, tried to apply the Recognizing Textual Entailment task on different domains, also starting competitions to see which models could perform well. In the field of Legal Informatics, we can find COLIEE (Competition on Legal Information

Extraction/Entailment)[14]. COLIEE started in 2014, and defined a competition every year up to now. Each competition is composed of four tasks: two regarding information retrieval on legal text, one regarding question answering and one regarding RTE on legal text. The task datasets are free to access upon request. In this paper, we decided to use their dataset for RTE in order to train our classifiers, since our dataset, to the best of our knowledge, is the first one regarding the cybersecurity domain. In this competition, [28] proposed a Multi-Layer Perceptron (MLP), composed of two hidden layers, with a decomposable attention model to find relations between words pairs. In detail, they started collecting articles from a civil code. Then, they ranked those articles according to a given query. Finally, they paired the best article (after the ranking) with the query to construct the training dataset for the MLP. [9] proposed a method similar to the one in [28], where they used n-grams, extracted using lexical and morphological characteristics, to retrieve articles from the civil code. Another one close to the work of Son et al. is [15]. In this work, the authors tried a convolutional neural network to see whether two legal articles are related to each other or not. Finally, [14] proposed a complex model to solve both legal information retrieval and textual entailment for COLIEE 2016. For the former one, they proposed an ensemble similarity method using least mean square and linear discriminant analysis. For the latter, they applied a majority vote schema of three classifiers: a decision tree, an SVM, and a convolutional neural network. As features for the classifiers, they used word overlap, cosine similarity, WordNet [18] similarity score, and substring similarity.

5 Conclusion

We presented a dataset for Recognizing Textual Entailment on the legal domain. All pairs of the dataset regard cybersecurity controls extracted from NIST, ISO and ISO/IEC 27001 documents. To the best of our knowledge, this is the first dataset for cybersecurity RTE.

We conducted three evaluations on our dataset using several classifiers. We first checked whether it is possible to train a classifier to recognize the relations expressed in our dataset. Since there is no testset, we used a cross-fold evaluation. We obtained an average accuracy of about 83% for the Support Vector Machine and the Maximum Entropy classifiers. We also reported that only word and POS tag n-grams are relevant as features to predict the relation. This is also confirmed by the ablation study. However, the classifiers tend to predict the wrong label when both the premise and hypothesis regard different aspects of the same topic (e.g., information system training vs information system maintenance). To solve this problem, we think that the classifiers require features that are able to capture the the topics of the NIST and ISO controls. For such reason, we will adopt the Topic Model proposed by [5].

We then performed a second evaluation, checking whether it is possible to transfer the knowledge acquired from a legal dataset for RTE to our one. Thus, we trained the classifiers using Task 2 dataset of COLIEE competition. We obtained an accuracy of about 50.48%, slightly surpassing the *Random* classifier. Such analysis showed that our dataset contains complex pairs, for both language and content, that do not allow classifiers trained

[14] `https://sites.ualberta.ca/~rabelo/COLIEE2019/`

on other datasets to generalize well. This has been confirmed by the third evaluation, where we evaluated if it is possible to transfer the knowledge acquired on our dataset to other legal ones for RTE. We decided to test the classifiers on the COLIEE testset, obtaining an accuracy of about 47%.

Finally, we discovered that all the proposed models have an accuracy on entailment pairs close to 100%. They however find difficult to recognize neutral pairs, obtaining an accuracy on these pairs at most of 10%. We think that a further classification of the pairs into the three classes *entailment*, *neutral* and *contradiction* will be useful to boost the performance of the machine learning models.

In our future works, we aim at integrating and inter-linking more cybersecurity standards. Specifically, we want to create a *unified* inter-connected corpus of technical documents in Natural Language for the cybersecurity domain, on which training and evaluating RTE classifiers to be later used by auditors as well as by companies collaborating with them, such as Nomotika SRL.

Acknowledgments

This research was supported by the EU's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts" (`http://www.mirelproject.eu`). We would like to thank prof. Cleo Condoravdi for fruitful discussions and feedback during the secondment of Giovanni Siragusa and Livio Robaldo at Stanford University in the context of the MIREL project.

References

[1] Kolawole Adebayo, Luigi Di Caro, and Guido Boella. Siamese network with soft attention for semantic text understanding. In *Proceedings of the 13th International Conference on Semantic Systems*, pages 160–167, 2017.

[2] Kolawole John Adebayo, Luigi Di Caro, Livio Robaldo, and Guido Boella. Textual inference with tree-structured lstm. In *BNCAI*, pages 17–31, 2016.

[3] G. Ajani, G. Boella, L. DI Caro, L. Robaldo, L. Humphreys, S. Praduroux, P. Rossi, and A. Violato. The European Taxonomy Syllabus: A multi-lingual, multi-level ontology framework to untangle the web of European legal terminology. *Applied Ontology*, 11(4), 2016.

[4] Luisa Bentivogli, Ido Dagan, and Bernardo Magnini. *The Recognizing Textual Entailment Challenges: Datasets and Methodologies*, pages 1119–1147. Springer Netherlands, Dordrecht, 2017.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[6] G. Boella, L. di Caro, L. Humphreys, L. Robaldo, and L. van der Torre. NLP Challenges for Eunomos, a Tool to Build and Manage Legal Knowledge. Proceedings of the International Conference on Language Resources and Evaluation, 2012.

[7] G. Boella, G. Governatori, A. Rotolo, and L. van der Torre. A logical understanding of legal interpretation. In *Int. Conference on the Principles of Knowledge Representation and Reasoning*, 2010.

[8] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[9] Danilo S Carvalho, Minh-Tien Nguyen, Chien-Xuan Tran, and Minh-Le Nguyen. Lexical-morphological modeling for legal text analysis. In *JSAI International Symposium on Artificial Intelligence*, pages 295–311. Springer, 2015.

[10] Timothy Chklovski and Patrick Pantel. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004.

[11] D. Grossi., C. Meyer, and F. Dignum. The many faces of counts-as: A formal analysis of constitutive-rules. *Journal of Applied Logic*, 6(2), 2008.

[12] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 107–112, 2018.

[13] H. Hart. *The Concept of Law*. Oxford: Clarendon Press, 1994.

[14] Kiyoun Kim, Seongwan Heo, Sungchul Jung, Kihyun Hong, and Young-Yik Rhim. An ensemble based legal information retrieval and entailment system. In *Tenth International Workshop on Juris-informatics (JURISIN)*, 2016.

[15] Mi-Young Kim, Randy Goebel, and S Ken. Coliee-2015: evaluation of legal question answering. In *Ninth International Workshop on Juris-informatics (JURISIN 2015)*, 2015.

[16] Daniel Z. Korman, Eric Mack, Jacob Jett, and Allen H. Renear. Defining textual entailment. *Journal of the Association for Information Science and Technology*, 69(6), 2018.

[17] Bernardo Magnini, Roberto Zanoli, Ido Dagan, Kathrin Eichler, Günter Neumann, Tae-Gil Noh, Sebastian Pado, Asher Stern, and Omer Levy. The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48, 2014.

[18] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[19] Rohan Nanda, Kolawole John Adebayo, Luigi Di Caro, Guido Boella, and Livio Robaldo. Legal information retrieval using topic clustering and neural networks. In *COLIEE@ ICAIL*, pages 68–78, 2017.

[20] Rohan Nanda, Luigi Di Caro, Guido Boella, Hristo Konstantinov, Tenyo Tyankov, Daniel Traykov, Hristo Hristov, Francesco Costamagna, Llio Humphreys, Livio Robaldo, and Michele Romano. A unifying similarity measure for automated identification of national implementations of european union directives. In *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, ICAIL '17, pages 149–158, New York, NY, USA, 2017. ACM.

[21] Sebastian Padó, Tae-Gil Noh, Asher Stern, Rui Wang, and Roberto Zanoli. Design and realization of a modular architecture for textual entailment. *Natural Language Engineering*, 21(2):167–200, 2015.

[22] Monica Palmirani, Michele Martoni, Arianna Rossi, Cesare Bartolini, and Livio Robaldo. Legal ontology for modelling GDPR concepts and norms. In Monica Palmirani, editor, *Legal Knowledge and Information Systems - JURIX 2018: The Thirty-first Annual Conference, Groningen, The Netherlands.*, pages 91–100, 2018.

[23] Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81, 2018.

[24] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, 2018.

[25] L. Robaldo, C. Bartolini, M. Palmirani, A. Rossi, M. Martoni, and G. Lenzini. Formalizing GDPR Provisions in Reified I/O Logic: The DAPRECO Knowledge Base. *Journal of Logic, Language and Information*, to appear, 2019.

[26] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. Reasoning about entailment with neural attention. In *Proceedings of the 2015 International Conference on Learning Representations*, 2015.

[27] A. Rotolo and C. Roversi. Constitutive rules and coherence in legal argumentation: The case of extensive and restrictive interpretation. *Legal Argumentation Theory*, 2012.

[28] Nguyen Truong Son, Viet-Anh Phan, and Nguyen Le Minh. Recognizing entailments in legal texts using sentence encoding-based and decomposable attention models. In *COLIEE@ ICAIL*, pages 31–42, 2017.

[29] Masatoshi Tsuchiya. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

 Received 1 February 2021