A Legal Perspective on Training Models for Natural Language Processing

Richard Eckart de Castilho[♣], Giulia Dore[♠], Thomas Margoni[♠], Penny Labropoulou[♡], Iryna Gurevych[♣]

[♣]Technische Universität Darmstadt, Department of Computer Science, UKP Lab

[♠]University of Glasgow, School of Law - CREATe Centre

[♡]Athena RC, Institute for Language and Speech Processing

Abstract

A significant concern in processing natural language data is the often unclear legal status of the input and output data/resources. In this paper, we investigate this problem by discussing a typical activity in Natural Language Processing: the training of a machine learning model from an annotated corpus. We examine which legal rules apply at relevant steps and how they affect the legal status of the results, especially in terms of copyright and copyright-related rights.

Keywords: Copyright, Licensing, Machine Learning, Annotated Corpora

1. Introduction

The state-of-the-art in many areas of Natural Language Processing (NLP) and Text Mining (TM) is based on Machine Learning (ML). Algorithms learn abstract probabilistic models from texts annotated with labels (e.g. named entities, part-of-speech tags, sentiment tags, etc.) in order to predict such labels on unseen text. NLP tasks usually require the deployment of multiple components each using specialised models. As training models can be tedious and computationally intensive, pre-trained models are a valuable resource.

However, the legal status of these models is often dubious, as in many cases it is unclear (a) whether a model can be trained from a corpus in absence of specific authorisation, (b) which licence (if any) can or must be assigned to them, and (c) if and in which cases the licence(s) of the original corpus and annotations affect the licensing of a model. This legal uncertainty often constitutes a hurdle, if not a real barrier for the development of research infrastructures and repositories, such as CLARIN[1] or OpenMinTeD,[2] where models are shared and used.

In this paper, we explore the process of training a model both from an NLP and a legal perspective. We discuss under which circumstances annotated corpora may be used for training a model and if their legal status may restrict the choice of licences that can be applied to the trained model. We use EU copyright law as the main reference, although the analysis may find application beyond the EU (and may need some degree of adjustment in different EU Member States).

2. Background

This section starts introducing the typical actions and resources involved in ML models construction and deployment and successively discusses the relevant legal concepts.

2.1. NLP perspective

Models are constructed through a training process involving a *learning algorithm* and *training data* to learn from. The model captures abstract probabilistic characteristics from the training data, which can then be used to predict the learned labels on unseen data. For illustration purposes, we focus on Named Entity Recognition (NER), an example of a sequence classification task.

In general, constructing a model consists of the following steps: (1) corpus compilation, (2) corpus pre-processing, (3) corpus annotation, and (4) training of the model. Depending on the availability of an annotated corpus, one or more of these steps can be skipped. We briefly describe all of these steps here and elaborate on the training step in the main parts of our investigation.

Corpus compilation. Each corpus is compiled to capture specific aspects of real world language. For best results, the ML algorithm must be trained on a corpus (i.e. set of texts) that is similar to the corpus to which it is later applied; i.e., it must be of the same language and domain or text type and annotated with the appropriate labels, e.g., "*English*", "*Social Sciences*", "*scholarly publications*" and "*named entities*" (NE), respectively. The corpus texts are selected and obtained from one or more sources (e.g. publishers, journals, web sites, etc.).

Pre-processing. This involves all kinds of (usually automatic) processes required to convert the textual content into a format that can be further processed by the NLP tools, such as conversion of PDF or HTML files into plain text, removal of images, tables, etc.

Annotation. This is the task of manual or automatic enrichment of texts with labels relevant to the target task, possibly further corrected by experts. Annotations are often arranged in layers, e.g. grammatical categories, morphological or syntactic features, etc. In all cases, the human annotator or the tool "*reads*" the text which is segmented into units (e.g. words, phrases) and assigns to some or all of them the appropriate labels. The inventory of labels is defined in *annotation resources*, such as tagsets, ontologies, thesauri, etc. The assignment usually follows *instructions* (e.g. guidelines, grammar rules, statistical data) that define when to assign labels and how to disambiguate if multiple candidate labels exist.

Training. The *training tool* is a software programme that implements an ML algorithm which is applied to the anno-

[1]https://www.clarin.eu

[2]https://www.openminted.eu

tated corpus, analyses its features and extracts from it the appropriate probabilistic and statistical characteristics. The model can thus be regarded as an abstraction of the annotated corpus based on statistical observations which can then be used with a second software tool (*tagger*) to predict the learned labels (e.g. NEs) on unseen text.
For the present analysis, we assume that ML tools (trainers and taggers) are governed by licenses that do not impose restrictions on the models they create. A similar assumption is made in relation to the use of predictions ML tools make. Accordingly, the analysis focuses on the licensing terms of the annotated corpora and the actions performed on them while training a model.

2.2. Legal perspective

Before proceeding to the discussion of the three scenarios, this section clarifies some basic copyright law concepts.

Texts and literary works. Most corpora employed in NLP consist of web pages, publications, articles, newspaper texts, blog posts or even tweets, annotated or not. All these resources possess the potential to be protected by *copyright law*. To be eligible for copyright protection a work must be original. The originality standard has been harmonised by EU law at the level of the *author's own intellectual creation*. Current legislation[3] and relevant case law[4]. indicate that this harmonised level of originality, despite the evocative formula employed, is placed at a rather low level (Margoni, 2016) and the Court of Justice of the EU (CJEU) has held that 11 consecutive words can in certain circumstances be considered the author's own intellectual creation, thus protected by copyright. This must be verified on a case-by-case basis and it is achieved when an author is able to put their personal stamp onto the work through free and creative choices. Therefore it cannot be excluded that even single sentences, if original, can be object of copyright protection. In conclusion, it can be assumed that most corpora used for TM/NLP, especially those of a literary and scientific character, such as scholarly articles, are protected copyright.

Databases. Under EU law, as well as under the law of many other countries,[5] databases are defined as collections of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means.[6] Copyright exists if originality is found in the *selection or arrangement of the content*, i.e., the "*intellectual creation*" has to be found in the database structure. Consequently, copyright in databases protects only the structure and does not extend to the content. The content, in turn, can be autonomously protected by copyright (a database of scholarly articles), related rights (a database of sound recordings), or be in the public domain (a database of unprotected facts or of medieval texts).

In addition, EU law, unlike the law of most other countries in the world, has introduced a new right protecting non-original databases when a substantial investment has been put in the obtaining, verification or presentation of the data – but importantly not in the creation of the data. In this case, the database maker (usually the person or entity who bears the financial risk) enjoys a *sui generis database right* (SGDR), which protects the *content* of the database from substantial extractions. In other words, even databases of unprotected facts could become object of a proprietary right that extends to the database content in the light of the aforementioned substantial investment (Hugenholtz, 2016; Guibault and Wiebe, 2013). Therefore, certain collections of corpora (e.g. the database of Institute X that over the years has collected public domain corpora investing substantial time and work resources in the process) could be protected by the SGDR.
Copyright and database rights are probably the two most relevant rights potentially covering the annotated and unannotated corpora forming the basis for any training activity (Stamatoudi and Torremans, 2000; Firdhous, 2012; Payne and Landry, 2012; Borghi and Karapapa, 2013; Tsiavos et al., 2014; Truyens and Van Eecke, 2014; Triaille et al., 2014; Handke et al., 2015). In the following sections, we examine this aspect more closely in the context of three basic scenarios. Specific attention will be paid to: (a) the *right of reproduction* (making copies) of the resources in question and whether the distinction between temporary and permanent copies matters; (b) the right of *adaptation and/or translation*, i.e. the creation of works based on those resources, what is often –but imprecisely from a EU law point of view– called a *derivative work*; and (c) the specific *licence types* under which annotated corpora are distributed. Nevertheless, no specific attention will be dedicated to the annotation process as many corpora are already available in an annotated form.

3. Scenario I: Liberally Licensed Corpora

We start our investigation with a straightforward scenario: training a model on an annotated corpus with a liberal TM/NLP-friendly licence carrying only an attribution clause. Here, we choose the popular Creative Commons Public Licence with the Attribution clause in the latest version available (CC BY 4.0). This licence is particularly "*friendly*" for TM/NLP activities because it authorises licensees to perform all the aforementioned rights (reproduction, redistribution, communication to the public, adaptation, etc.) under the only main condition that attribution be maintained.

3.1. Scenario description

Corpus. Despite version 4.0 of the CC licences being available since 2013, many of the existing CC BY licensed corpora are still distributed under older versions such as CC BY 2.5 (e.g. the Wikinews texts of the GUM corpus (Zeldes, 2017)) or CC BY 3.0 (e.g. the IULA Spanish LSP Treebank (Marimon et al., 2014) or the CRAFT corpus (Verspoor et al., 2012)).[7] We only consider the latest 4.0 version in the present analysis, but it should be noted that

[3] E.g. Directive 2009/24/EC, OJ L 111, 5.5.2009, 16–22, Article 1(3).

[4] E.g. Judgment of 16 July 2009, *Infopaq International v. Danske Dagblades Forening*, C-5/08, ECLI:EU:C:2009:465

[5] General Agreement on Trade-Related Aspects of Intellectual Property (TRIPS), 1869 U.N.T.S. 299, 33 I.L.M. 1197; WIPO Copyright Treaty (WCT), 105-17 (1997), 36 ILM 65(1997)

[6] Directive 96/9/EC, OJ L 77, 27.3.1996, Article 1

[7] It is notably difficult to locate corpora under specific CC licence versions. E.g. at the time of writing, META-SHARE (http://www.meta-share.org) lists over 200 corpora under CC BY,

different licence versions could lead to different assessments especially in relation to the SGDR right. Examples for corpora under this licence version are the recent GermEVAL 2014 dataset for NER (Benikova et al., 2014) or the Coptic Treebank (Schroeder and Zeldes, 2016).[8]

Training. In our scenarios, we train a NER model with the Stanford NER tool (Manning et al., 2014). To determine the relation of the trained model to the original data, we examine which information goes into the model. We describe the process in the present and following scenarios at increasing levels of detail, as required by the respective legal analyses. The training process requires the creation of a usually temporary copy (i.e. a reproduction) of the original data and usually its transformation into the training data format. The training data format used by the Stanford NER tool is very simple: a two-column format in which the first column contains a *token* (word or punctuation mark) and the second column contains a *label*. Sentences are separated by a blank line. If the word is a named entity, the label indicates the entity type (e.g. person, organisation, etc.), otherwise it contains a special "*no category*" label. To handle cases where a NE consists of multiple tokens, the type is either prefixed with *B-* to indicate the first token of the NE, *I-* for the other tokens of the NE and *O* for single-token NEs. This so-called *BIO*-encoding is a technical convention allowing the NER tool to learn how to correctly detect multi-token NEs.
Before considering the training process in more detail in Scenario II, we investigate whether the process up to this point is permitted by the licence.

3.2. Scenario analysis

Is reproduction permitted? According to the terms of the CC BY 4.0 licence, the act of making reproductions is expressly permitted, as per Section 2 of the licence text.[9]
In the present case, thanks to the liberal conditions established by the licence, it is not necessary to investigate whether the *results* of the training activity constitute a reproduction of the original corpora. The applicable licence permits any type of reproduction, being it the transient reproduction necessary for the conversion of the corpora into a machine processable format, or the final results of the training process Therefore, annotated corpora under these licences may be reproduced as part of the model training process on the basis of the licence (in those cases when this act is not covered by applicable exceptions and limitations, a situation that would not trigger the terms of the CC licence, see Scenario III below).

Is the result an adaptation (derivative work)? As stated above, the scenario is predicated on the assumption that all input resources (raw texts and annotations) are covered by a CC BY 4.0 license.

but does not carry information about the licence version. The LINDAT/CLARIN repository (http://lindat.mff.cuni.cz/) includes the licence version metadata, but the search interface does not allow filtering resources by it.

[8] We should note here that we have not investigated whether the assignment of these licences to the respective corpora is indeed valid.

[9] https://creativecommons.org/licenses/by/4.0/legalcode, s. 2 Section 2a of this licence expressly permits the creation of adaptations and defines it contractually, although ultimately referring to the applicable copyright law. Since the creation of an adaptation is explicitly permitted under the terms of the licences, if the act of training a model constitutes a derivative work, which is indeed not trivial to determine (see Scenario II below), in the present scenario this activity is permitted, under a mere attribution condition. It is worth pointing out that the attribution requirement in the present case would require: a) retaining attribution, copyright and licence notices, and providing a URI or hyperlink to the licensed material to the extent reasonably practicable; b) indicating modifications to the licensed material and retain an indication of any previous modifications; c) indicating the licensed material is licensed under this public license, including the text of, or the URI or hyperlink to the licence. No other limitations apply to the model. In principle, the model could also be re-licensed under any arbitrary licence if it qualifies for copyright protection on its own right, an aspect that will be discussed below in Scenario II.
In conclusion, training models on the basis of liberally licensed corpora does not present major legal obstacles, although proper attribution should be given.

4. Scenario II: Corpora with a Reciprocal Licence

We continue our investigation with a slightly more complicated case: training a model on an annotated corpus with a TM/NLP-friendly licence which includes a share-alike clause, such as the Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0). This means that works adapted (derived) from the original work must carry the same licence as the original work. Since this is a reciprocal condition, it is important to determine whether a model is an adaptation (derivative work) of the corpora under the conditions of the licence.

4.1. Scenario description

Corpus. We consider the corpus to be licensed under the CC BY-SA 4.0 license, such as the latest version of the SETimes.HR dataset (Agić and Ljubešić, 2014).[10]

Training. The basic scenario is the same NER training process we have already started to describe in Scenario I. Assuming the corpus at hand has already been transformed into the two-column format described in Scenario I, the process of training a model is rather straightforward (although it may require significant computational resources).
As a first step, a configuration file for the Stanford NER tool needs to be created. This file contains the names of the files that comprise the training corpus and a name to be used for the output file (i.e. the model to be created), as well as a set of parameters controlling which features are extracted from the training data and used for training the classifier. An example parameter file (Figure 1) can be found in the documentation of the Stanford NER tool.[11]

[10] https://github.com/ffnlp/sethr (Agić and Ljubešić, 2014) is based on texts from the Croatian translation of the SETimes portal, which were freely shared with attribution to the source.

[11] https://nlp.stanford.edu/software/crf-faq.html – A more exten-

```
trainFile = training-data.col
serializeTo = ner-model.ser.gz
map = word=0,answer=1

useClassFeature=true
useWord=true
useNGrams=true
noMidNGrams=true
maxNGramLeng=6
usePrev=true
useNext=true
useSequences=true
usePrevSequences=true
maxLeft=1
useTypeSeqs=true
useTypeSeqs2=true
useTypeySequences=true
wordShape=chris2useLC
useDisjunctive=true
```

Figure 1: Example parameter file *params.prop*.

As a second step, the Stanford NER tool is started in training mode using the command *java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop params.prop*. From this point on, the process runs fully automatically without further interaction from the person training the model.

4.2. Scenario analysis

Is reproduction permitted? Once again, the act of making copies (reproductions) is expressly allowed by the CC BY-SA 4.0 in the same terms analysed in Scenario I.

Is the result an adaptation (derivative work)? The creation of adapted materials is also expressly permitted, as it was with CC BY 4.0 in the previous scenario. However, the SA clause that applies in the present scenario requires that distribution of the adapted material be made under the terms of the same licence or a later version with the same terms.
Therefore, it is important to determine whether the trained model is an adaptation of the original annotated corpora. If it is, the SA clause requires that the same licence be applied to the trained model.
What constitutes adapted material is defined in Section 1(a) of the licence[12] as "*... material subject to Copyright and Similar Rights that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission under the Copyright and Similar Rights held by the Licensor.*" To establish when this happens is a determination that can be done only against a specific domestic legal framework even within the EU. In fact, differently from other rights, the right of adaptation is not harmonised by EU law.[13]
However, not all modifications lead to the creation of an adapted work. EU law seems to suggest that such a new (adapted) work is created only when the process of modification involves an original contribution (in the sense of the author's own intellectual creation). This appears to be also the view of the licences developers. [14] Absent enough originality in the modification, the unoriginally modified material does not constitute a new copyright protected work, but rather a mere reproduction (even if partial or "*in any form*"), which, unless authorised by law or by contract, infringes the copyright (the right of reproduction) in the original work.
Given this definition of adapted material, we need to consider whether the act of training a model using the outlined procedure meets the licence requirements.
The simplicity of the training process outlined above and the limited choices in the parametrisation of a largely automated process suggest that there is no space for the free and creative choices that allow the author to express their personality into the work. In particular, it seems that even when certain choices in parametrisation are available these are dictated mostly by technical considerations and by the "*rules of the game*" of model training in a way that any equally skilled technician would achieve a similar or identical result. Under this interpretation, the model is not a creative adaptation of the underlying annotated text corpora and thus does not qualify as adapted material under the SA clause of the CC license.
This means that the trained model, not being an adaptation of the underlying corpora, does not trigger the SA clause. Training a model, as seen above, requires other types of copyright relevant acts, namely reproduction, which must be authorised or excused –statutorily or contractually– to avoid infringement. In the present case, it means that if the trained model, which does not qualify as adapted material, is nonetheless a "*reproduction in part*" of the original corpora, that part –and only that part– remains under the conditions of the original licence, in the present case a CC BY-SA 4.0.
Similar conclusions would be reached if the resources employed in the training process were licensed under a CC BY-ND 4.0. The licence in question, in fact, although allowing the creation and reproduction of adapted works, does not allow for their distribution (alias sharing), as specified in its Section 2(a).[15] If the model is not an adapted work in the meaning outlined above, then the NoDerivatives (ND) clause will not be triggered.
Finally, does this mean that the trained model can be arbitrarily licensed by its developer? In the present case, the trained model lacks sufficient originality to qualify as a derivative as well as an independent work. Therefore, it is not a work of authorship for copyright law purposes. In theory a license could still be applied but this would only have contractual effects and not be based on an underlying property right (a very relevant difference that cannot be explained here, suffice to say that most copyright licences are based on a valid underlying property right: if this is not present the effects of the licence are limited. In the case of CC licenses, as well as most "*open*" licences, if the licence is applied to something that is not protected by copyright or related rights the licence is not triggered).

sive list of parameters can be found in the documentation to the Stanford NER Java class *NERFeatureFactory*.

[12] https://creativecommons.org/licenses/by/4.0/legalcode, s. 1

[13] Judgement of 22 Jan 2015, *Art and Allposters International BV*, Case C-419/13, ECLI:EU:C:2015:27

[14] https://creativecommons.org/faq/#when-is-my-use-considered-an-adaptation

[15] https://creativecommons.org/licenses/by-nd/4.0/legalcode

```
#<objec#|C
#acter>#|C
consisted...xx-PW_CTYPE|C
xxxxx...about-NW_CTYPE|C
many-NW|C
nowadays-WORD|C
housed-PW|C
law-PSEQW|CpC
Planet-Wide-PSEQpW|CpC
essentially-still-PSEQW2|CpC
```

Figure 2: Feature excerpt from a CoreNLP NER model.

5. Scenario III: Corpora with unclear licence statements or restrictively licensed

In the previous scenarios we discussed texts and annotations under licences which explicitly allow reproduction. However, it is much more common that only annotations are under such a licence. Obtaining corpora under similar licences is much more difficult. Most of the texts that can be found online do not carry any licence at all or are part of commercial offers which do not permit reproduction. Thus, in the present scenario we investigate if such texts can still be used for training models. We do not investigate the relationship between corpora and annotation.

5.1. Scenario description

Corpus. An example of a text corpus obtained from the web and enriched with annotations under a CC licence is the English part of the Universal Dependency Treebank (UDT-EN).[16] While the annotations are provided under CC BY-SA 4.0, the texts come from the English Web Treebank[17] which has been collected by Google from online weblogs, newsgroups, emails, reviews and question-answering websites. As the UDT-EN website states, the copyright of portions of the texts may reside with "*Google Inc., Yahoo! Inc., Trustees of the University of Pennsylvania and/or other original authors*." Thus, the licensing status of the individual texts is not entirely transparent, and should be prudently considered to be under an "*all rights reserved*" status.

Training. Again, we consider the same process of training a NER model described in Scenarios I and II. However, this time we take a closer look at the training process in order to assess whether the model *reproduces* significant parts of the original document, an activity reserved to the copyright owner.

During the training process, the NER training tool extracts so-called *features* from the input data. This is the critical step in the training process, as it determines how much of the original text and annotations is retained. Figure 2 shows a sample of feature values generated by the Stanford NER. The Stanford NER uses a sequence classifier based on conditional random fields (CRFs). The tool runs through the text token-by-token and consumes features that have been extracted for each token, such as the token string, a configurable number of characters forming the prefix/suffix of the token, the left and right context of each token, e.g. the fact that token *X* appears left of token *Y*, and similar information. The context captured in the features is very limited and usually includes only the current word and a preceding and following word. E.g. "*essentially-still-PSEQW2—CpC*" indicates that the sequence "*essentially still*" was included in the training corpus. Additionally, the CRF learns a set of weights encoding the probability of an NE occurring in the presence of the specific features (Finkel et al., 2005). It is important that the features and weights capture a *limited* information about the tokens and their annotations. They *discard* the context details, because the ML algorithm needs to learn a *generalised model* from the training data which does not overfit on the training data.

The features and algorithm for NER-like tasks are designed in such a way that the trained model represents an abstraction of the training data. It is generally not possible to reconstruct the original text from this abstraction.

5.2. Scenario analysis

Is reproduction permitted? As briefly observed above, the right of reproduction, defined as "*any direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part*" is reserved to the right holder of copyright works in all EU countries by Art. 2 InfoSoc Directive and its national implementations.[18]

The CJEU had the opportunity to clarify that certain acts of temporary reproduction carried out during a "*data capture*" process fulfil the requirements of the exception for temporary copies (Art. 5(1) InfoSoc) under the cumulative conditions that those acts:

1. Must constitute an integral and essential part of a technological process. This condition is satisfied notwithstanding the fact that initiating and terminating that process involves human intervention;
2. Must pursue a sole purpose, namely to enable the lawful use of a protected work; and
3. Must not have an independent economic significance provided:
   (a) that the implementation of those acts does not enable the generation of an additional profit going beyond that derived from the lawful use of the protected work;
   (b) that the acts of temporary reproduction do not lead to a modification of that work (Case C-5/08 Infopaq I and C-302/10 Infopaq II).[19]

Under these conditions, temporary acts of reproduction are permitted by EU law.

A brief description of the facts of the Infopaq case may be helpful. The decision, referring to the compilation, extraction, indexing and printing of newspaper articles and keywords, identifies five phases relevant in the process of

[16] https://github.com/UniversalDependencies/UD_English

[17] https://catalog.ldc.upenn.edu/LDC2012T13

[18] Directive 2001/29/EC, OJ L 167, 22.6.2001

[19] Judgement of 16 Jul 2009, Infopaq International A/S v Danske Dagblades Forening, Case C-5/08 I, ECLI:EU:C:2009:465 and Judgement of 17 Jan 2012, Infopaq International A/S v Danske Dagblades Forening, C-302/10, ECLI:EU:C:2012:16

data capture: (1) newspaper publications are registered manually in an electronic registration database; (2) sections of the publications are selectively scanned, allowing the creation of a Tagged Image File Format (TIFF) file for each page of the publication and transferring it to an Optical Character Recognition (OCR) server; (3) the OCR server processes this TIFF file digitally and translates the image of each letter into a character code recognisable by the computer and all data are saved as a text file, while the TIFF file is then deleted; (4) the text file is processed to find a search word defined beforehand, identifying possible matches and capturing five words before and after the search word (i.e. a snippet of 11 words, before the text file is deleted; (5) at the end of the data capture process, a cover sheet is printed out containing all the matching pages as well as the text snippets extracted from these pages.

The Court found that the exception of Art. 5(1), which covers acts of *temporary* reproduction, only exempts the activities listed in points 1) to 4) above, whereas the activity of point 5), i.e. printing, constitutes a *permanent* act of reproduction which is therefore not covered by the exception for temporary copies.

It should further be noted that, in point 5), what is printed is not the entire literary text, but only 11 consecutive words. Only if these 11 consecutive words constitute a "*reproduction in part*" of the original work, copyright would be infringed. In this regard, the EUCJ found that "*it cannot be excluded*" that 11 consecutive words constitute the author's own intellectual creation and therefore represent a partial (and thus infringing) permanent reproduction. The 11 words threshold should not be taken as a strict parameter. The real test is that of the author's own intellectual creation. Accordingly, there will be shorter extracts that meet such a condition, and longer extracts that do not meet it.

As a result, it can be argued that current EU law permits the temporary copy of non-licensed copyright works for purposes such as "*data capturing*" as long as the cumulative conditions of Art. 5(1) as interpreted by the Court are met. However, when the results of the data capturing process leads to the *permanent* reproduction of the author's own intellectual creation this constitutes an infringement of the right of reproduction.

Nevertheless, it must be stressed that the conditions of Art. 5(1) are not only cumulative (i.e. all must be met) but also partially unclear (especially regarding the exact meaning of "independent economic significance") and must be interpreted *strictly*. These considerations have led many commentators to the conclusions that Art. 5(1) is not suitable as a general solution for TDM purposes (Triaille et al., 2014). This conclusion is certainly correct, nevertheless, until when a proper TDM exception is introduced at the EU level, the suitability of Art. 5(1) for unlicensed corpora should be explored further for specific ML/NLP cases, as in the present scenario. The remainder of this section will attempt such an exploration.

It seems that the the ML/NLP steps described in scenario three are substantially similar to those described by the EUCJ in the reported case law. In particular:

1. transforming the text corpus and the annotations into the input format of the Stanford NER tool is arguably equivalent to converting a TIFF image into text using OCR but much less sophisticated;

2. inspecting each word in the text in turn in order to create a ML feature representation capturing from the word and its immediate left and right neighbours, and from the annotation on the word is arguably equivalent to extracting the search term and the words before and after it, although extracting only one word before/after instead of 5, for a total of 3 words instead of 11 words extracted;

3. creating a probabilistic report (i.e. model) about the data obtained in this way is arguably equivalent to printing a cover sheet containing the matching pages, although the report consists of a numeric matrix encoding probabilities of observed features and correlations, as well as a feature and label dictionary containing encoded features covering words, word pairs, parts of words, and word shapes observed in the text, and the labels.

It seems plausible that the temporary copies created in points 1. and 2. are transient or incidental if they are only kept for the amount of time justified by the proper completion of the technological process and are automatically destroyed at the and of the process. It is also arguable that the act of reproduction is an integral and essential part of a technological process (the conversion of the text into data) which is necessary to enable a lawful use (statistical analysis is arguably as lawful as the preparation of summaries and is not a right reserved to the right holder by EU copyright law, however if the right holder contractually limits this operation and domestic law allows it, probably this condition would not be met). The requirement of absence of independent economic significance is probably harder to assess. Independent economic significance is present if the author of the reproduction is likely to make a profit out of the economic exploitation of the temporary copy. This profit has to be distinct from the efficiency gains that the technological process allows (see Infopaq II, 51).

Point 3. above refers to the results of the training process (the creation of a model) which are *permanent* by definition. Therefore, point 3. cannot be exempted on the basis of acts of *temporary* reproduction. It must be assessed however, whether the model constitutes a "*reproduction in part*" within the meaning of Art. 2 InfoSoc. If it does not, there is simply no copyright relevant activity and thus no need to rely on an exception.

In the present scenario, the trained model contains three consecutive words of the original "*all rights reserved*" corpora. While the test to be applied is not 11 vs. 3 consecutive words, but that of the "*author's own intellectual creation*", it seems plausible that three consecutive words are too insubstantial to constitute a "*reproduction in part*" of the original corpora. Therefore, the trained model does not reproduce in part the original corpora.

In conclusion, it can be argued that Art. 5(1) has the potential to be useful when the technological process is similar to

the one described in this scenario. However, given the cumulative, strict and partially unclear conditions that qualify it, a very careful case-by-case assessment should be performed before deciding to rely on this exception given the unavoidable degree of risk involved.

Are the results a derivative work? The CJEU in Allposters[20] clarified that the right of adaptation has not been object of EU harmonisation. Nevertheless, it must be observed that cases where adaptation does not require the reproduction at least in part of the original work may be of difficult conceptualisation (illustratively, jurisdictions such as France and the Netherlands classify the right of adaptation as a type of reproduction). That said, there are situations where there is adaptation without reproduction. An obvious example is the translation of a literary work into a different language. Technically speaking, there is no direct reproduction of the sentences in the original language. Consequently, it should not come as a surprise that the right of translation was the first right to be included in the minimum standard of protection in the oldest copyright international treaty, the Berne Convention.[21] However, training a model does not seem to possess the characteristics be considered a translation. Therefore, excluding the cases when this activity constitutes a reproduction (see above), it should be ascertained whether and under which circumstances training a model creates an adaptation.

Definition of derivative works in legislation At the international level, the Berne Convention grants copyright protection to translations, adaptations, arrangements of music and other alterations without prejudice to the original work to which they refer.[22]

There is no explicit definition in the Convention of adaptation, arrangement or other alterations of a work, a definitory lacuna that has stimulated some debate over the possible meanings of their specific wording (cf. (Ricketson and Ginsburg, 2006), p. 480). As some have suggested ((Goldstein and Hugenholtz, 2001), p. 252) adaptation means recasting a work from one format to another, whereas arrangement means modification within the same format, while others have underlined how the effort of defining them may be even unnecessary, if the law treats them essentially the same in terms of protection ((Chow and Lee, 2006), p. 181).

Illustratively, the notion of derivative work is instead explicitly defined by US law, which under the Copyright Act (17 U.S. Code) at §106(2) regulates the exclusive right to prepare derivative works.

As already pointed out, EU law does not harmonise the right of adaptation (except in the case of software and databases), therefore a proper analysis should look at how this right is regulated at the national level, thereby introducing an additional layer of complexity especially for scientific initiatives which are often international.

The Court of Justice, in Allposters, avoided to define derivative works by substantially referring to the right of reproduction, purporting that a new work incorporating a pre-existing protected work is "*an alteration of the copy of the protected work, which provides a result closer to the original*" and so constitutes "*a new reproduction of that work*" that remains in the exclusive rights of its right holder.[23] Furthermore, what seems determinant is that the new work (often identified as secondary work) reproduces, adapts or alters what constitutes the intellectual creation of the pre-existing (primary) work adding however an "*authorial contribution*" (Margoni, 2014).

Given the lack of EU harmonisation for the right of adaptation, the analysis should focus on the domestic law of EU Member States (MS). This type of inquiry would need to be done with a depth of analysis in 28 jurisdictions that is not be possible in the present paper. Nevertheless, it seem arguable that, as opposed to the broader US notion of derivative works, the EU counterparts tend to define adaptation in a narrower way, "*even narrower than the original Berne formulation*" (Bently and Sherman, 2014), p. 170, ft 220. Domestic courts have held that adaptations must show "*some quality or character which the raw material did not possess and which differentiates the product from the raw material*" ((Bently and Sherman, 2014), pp. 112-113 and Interlego v. Tyco Industries 1989 AC 217), and that in order for the right to be infringed, the elaboration should reveal the pre-existing work in its own individuality (ex multis, Corte Cassazione, 29 maggio 2003, n. 8597).

In the light of the above elements, which (it must be restated) constitute a mere superficial exploration of EU MS domestic law orientation on adaptations, it seems at least arguable that when the elaboration (trained model) does not reproduce the original (corpora) nor reveals "*its individuality*", no infringement should be found. From a policy point of view, training a model should be considered a free use. Future work should concentrate on this aspect.

6. Conclusion

This paper underlines the complexities in the relationship between concerning copyright and science in the context of ML/NLP. The legal analysis has been based on three specific scenarios which are all evolving around the task of training models for NER from annotated texts. The same legal principles can be applied to training models for other ML/NLP tasks (e.g. POS tagging, etc.), but depending on the specific variables the conclusions may differ. The conclusions of the three case scenarios presented in the present paper can be summarised as follows. The use of corpora licensed under TM/NLP friendly licences such as CC BY 4.0 guarantees that activities such as model training are lawful. In the case when no TM/NLP friendly licences are present, the operation of certain exceptions to copyright (e.g. Art. 5(1) InfoSoc) can represent the only proper legal basis for proceeding with ML activities. Nevertheless, a considerable level of uncertainty surrounds the applicability of the exception for temporary uses of Art. 5(1) InfoSoc and a proper analysis of each case should be performed

[20] Judgement of 22 January 2015, Art and Allposters International BV, Case C-419/13, ECLI:EU:C:2015:27

[21] Berne Convention for the Protection of Literary and Artistic Works, 1886

[22] Berne Convention for the Protection of Literary and Artistic Works, Art. 2, 8 and 12

[23] Judgement of 22 Jan 2015, Art and Allposters International BV, Case C-419/13, ECLI:EU:C:2015:27, paragraph 43.

before relying on it. Still, once the aspect of reproduction is properly addressed, we suggest refraining from defining training models in terms of derivative/adapted works, with the consequence that licensing restrictions (e.g. all rights reserved, ND or SA) imposed on the input training resources may not find application in the resulting output. At the same time, we acknowledge that the scope of TM/NLP is too broad to be handled homogeneously and that different types of algorithms and parametrisations require dedicated legal analysis, for example based on the level of abstraction they attain over the input data and the type of original material that is reproduced in the trained model.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-EINFRA-2014-2) under grant agreement No. 654021 (OpenMinTeD). It reflects only the authors' views and the EU is not liable for any use that may be made of the information contained therein. It was further supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1816B (CEDIFOR) and by the German Research Foundation as part of the RTG *Adaptive Preparation of Information from Heterogeneous Sources* (AIPHES) under grant No. GRK 1994/1. Additional thanks go to Mark Perry for review and comments.

Bibliographical References

Agić, v. and Ljubešić, N. (2014). The SETimes.HR Linguistically Annotated Corpus of Croatian. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2014*, pages 1724–1727, Reykjavik, Iceland. ELRA.

Benikova, D., Biemann, C., and Reznicek, M. (2014). Nosta-d named entity annotation for german: Guidelines and dataset. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2014*, Reykjavik, Iceland. ELRA.

Bently, L. and Sherman, B. (2014). *Intellectual Property Law*. Oxford University Press, 4th edition.

Borghi, M. and Karapapa, S. (2013). *Copyright and Mass Digitization: A cross-jurisdictional perspective*. Oxford University Press.

Chow, D. and Lee, E. (2006). *International Intellectual Property. Problems, Cases and Materials*. West Academic Publishing.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL 2005*, pages 363–370, Stroudsburg, PA, USA. ACL.

Firdhous, M. (2012). Automating legal research through data mining. *Int. Journal of Advanced Computer Science and Applications*, 1(6):9–16.

Goldstein, P. and Hugenholtz, P. B. (2001). *International Copyright: Principles, Law, and Practice*. Oxford University Press.

Lucie Guibault et al., editors. (2013). *Safe to be open. Study on the protection of research data and recommendations for access and usage*. Universitätsverlag Göttingen.

Handke, C., Guibault, L., and Vallbé, J.-J. (2015). Is Europe Falling Behind? Copyright's Impact on Data Mining in Academic Research. In Birgit Schmidt et al., editors, *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust – Proceedings of Elpub 2015*, pages 120–130.

Hugenholtz, P., (2016). *Something Completely Different: Europe's Sui Generis Database Right*, volume 37 of *Information Law Series*, pages 205–222. Kluwer Law Int.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL 2014: System Demonstrations*, pages 55–60, Baltimore, Maryland. ACL.

Margoni, T. (2014). The Digitisation of Cultural Heritage: Originality, Derivative Works & (Non) Original Photographs. *IVIR Institute for Inf. Law*, pages 18–25.

Margoni, T. (2016). The harmonisation of EU copyright law: The originality standard. In Mark Perry, editor, *Global Governance of Intellectual Property in the 21st Century*, pages 85–105. Springer, Switzerland.

Marimon, M., Bel, N., Fisas, B., Arias, B., Vázquez, S., Vivaldi, J., Morell, C., and Lorente, M. (2014). The IULA Spanish LSP Treebank. In Nicoletta Calzolari, et al., editors, *Proceedings of LREC 2014*, Reykjavik, Iceland. ELRA.

Payne, D. and Landry, B. J. L. (2012). A Composite Strategy for the Legal and Ethical Use of Data Mining. *Int. Journal of Management, Knowledge and Learning*, 1(1):27–43.

Ricketson, S. and Ginsburg, J. C. (2006). *International Copyright and Neighbouring Rights. The Berne Convention and Beyond*. Oxford University Press.

Schroeder, C. T. and Zeldes, A. (2016). Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

Irina Stamatoudi et al., editors. (2000). *Copyright in the New Digital Environment: The Need to Redesign Copyright*, volume 8 of *Perspectives on Intellectual Property Law*. Sweet & Maxwell.

Triaille, J.-P., de Meeûs d'Argenteuil, J., and de Francquen, A. (2014). Study on the legal framework of text and data mining (tdm). *Luxembourg: Publications Office*.

Truyens, M. and Van Eecke, P. (2014). Legal aspects of text mining. *Computer Law & Security Rev.*, 30(2):153–170.

Tsiavos, P., Piperidis, S., Gavrilidou, M., Labropoulou, P., and Patrikakos, T. (2014). *Legal framework of textual data processing for Machine Translation and Language Technology research and development activities*. QTLP report & wikibook.

Verspoor, K., Cohen, K. B., Lanfranchi, A., Warner, C., Johnson, H. L., Roeder, C., Choi, J. D., Funk, C., Malenkiy, Y., Eckert, M., Xue, N., Baumgartner, W. A., Bada, M., Palmer, M., and Hunter, L. E. (2012). A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(1):207.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Lang. Resour. Eval.*, 51(3):581–612.