# Machine-Assisted Extraction of Metadata in Large Collections of Documents

**Natalie Widmann, HURIDOCS**

**Aaron Swartz Fellow, 9th April - 9th July 2018**

## MOTIVATION

Through technological advancements, vast amounts of data become accessible: documents, images, recordings, audio-visual materials, etc. However, the sole access to data is not enough to draw knowledge from it. It requires an in-depth study to connect information and to reveal the bigger picture.
In this process of structuring and organising a collection, metadata is very useful. It represents a structured summary of the most important information and allows to quickly filter, access and analyse the collection.

However, extracting metadata from a large and unorganised collection is a very tedious and time consuming process. It requires the manual processing of each collection item at least once. If, due to a change in the research interest, other type of information becomes interesting, another iteration of the manual processing of the collection is required. Especially, small non-profit organisations struggle with organising and analysing data due to time constraints and lack of resources.

The goal of this project is to support human rights defenders in extracting relevant information from collections of documents by using machine learning. Machine learning is a powerful tool that can help to detect patterns of corruption, predict poverty and analyse evidence of human rights violations.

The Aaron Swartz fellowship is conducted in close collaboration with HURIDCOS[1], a non-governmental organisation that supports human rights defenders with their information management challenges. HURIDOCS ensures that the project meets the needs of human rights defenders and provides the underlying infrastructure.

## UWAZI

To make machine-assisted metadata extraction accessible to human rights defenders and researchers, a framework to intuitively work with collections of documents is required. We use a free and open-source software called Uwazi[2].

Uwazi is Swahili and means “openness”. Huridocs build Uwazi based on the needs of human rights organisations to manage, organise and publish their documents and audio-visual materials. Its goal is to make human rights information more open and accessible to the defenders who need it.

The big advantages of Uwazi are its flexible adaptation of the metadata schema to the specific needs of an organisation, its strong emphasis on connecting documents, paragraphs or persons, and its intuitive design.

Different human rights organisations use Uwazi: CEJIL, the Center for Justice and international Law, makes the knowledge of causes of the Commission and the Inter-American Court of Human Rights accessible with Uwazi[3]. Memoria y Ciudadanía, an Uwazi collection of 1500 official reports and financial documents, was used to tell the story of corruption in Peru[4].

Adding machine-assisted extraction of metadata to Uwazi speeds up the tedious process of manual labeling and draws a researcher’s attention to relevant documents in the collection.

Before starting the Aaron Swartz fellowship, we already worked on integrating a machine learning algorithm. We will refer to it as sentence classifier, as it learns the relevant sentences specific to a research purpose and when fed with enough examples makes suggestions about related phrases in the collection. For the Aaron Swartz fellowship we build upon the implemented sentence classifier in order to automatically identify and extract project specific metadata.

[1] HURIDOCS is an NGO that supports human rights organisations in the use of information and technology.
[2] Uwazi is an free and open-source project. This means that everyone can download the code, use the software, point out problems, make changes and contribute to its development.
[3] Browse through the collection at https://cejil.uwazi.io/
[4] The collection is publicly available at https://japiqay.uwazi.io/ and a HURIDOCS blog post on telling the story of corruption can be found here.

## ACCOMPLISHMENTS

The integration of machine-assisted extraction of metadata from collections of documents into Uwazi consists of different tasks which include the preprocessing of texts, the integration of a machine learning algorithm, its optimisation and testing, as well as the communication with users. The following sections describe in detail the components that have been accomplished with support of the Aaron Swartz fellowship.

### Sentence Splitting

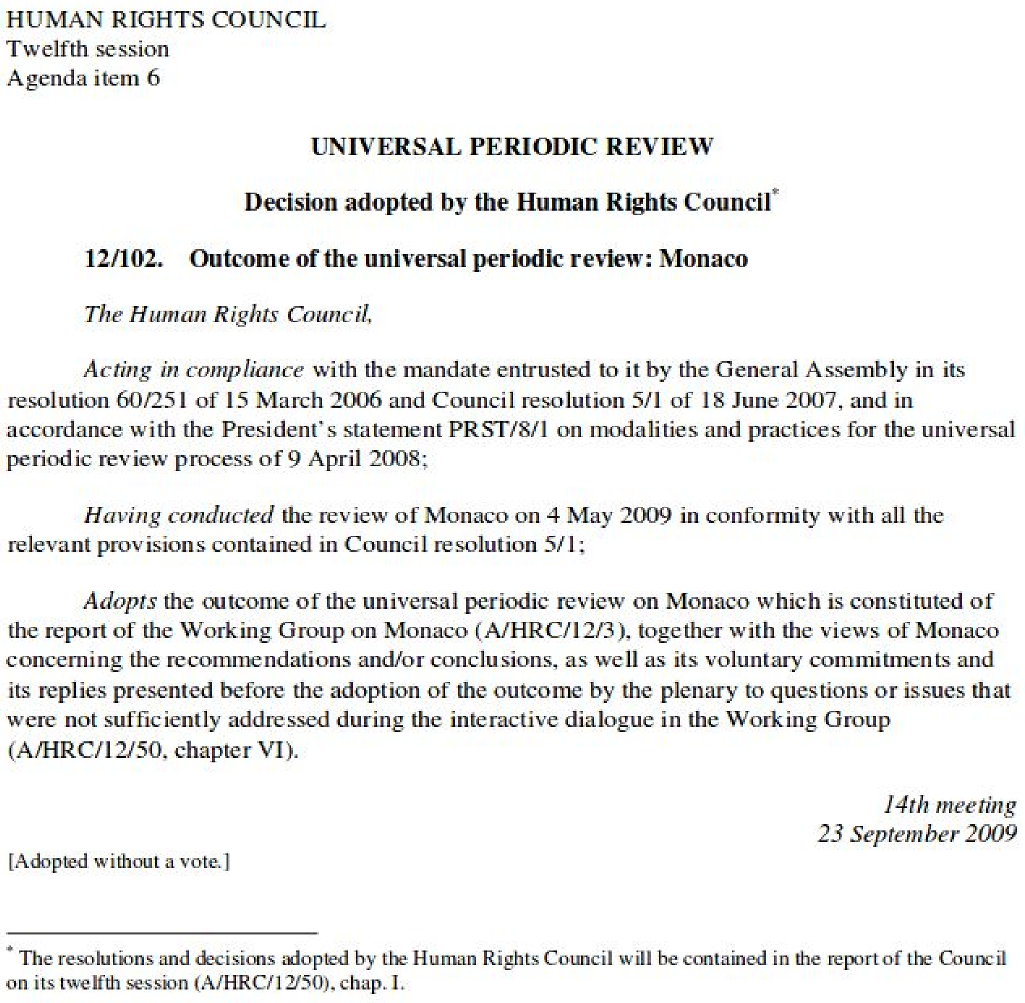
Sentences are at the core of our approach to extract metadata from collections of documents. They are more informative than single words or word combinations while narrowing down the diverse content of an entire document to one specific aspect. For example, a report on the human rights situation in Hungary might contain many different topics, such as the freedom of expression, discrimination, right to social security and protection, etc. However, one particular sentence will most likely focus on one of these issues. This makes it easier to identify and to connect it to similar content. Furthermore, sentences are self-contained while still reflecting the context of previous and subsequent information and references.

Whereas it is an easy task for humans to split a text into sentences, algorithms struggle to find a generic rule for it. Different document types, domain-specific abbreviations, the usage of uppercase letters, direct and indirect speech and other stylistic elements require a sophisticated approach.

A good example of  these irregularities are the resolutions of the Human Rights Council (see Figure 1 for a sample document). They often use semicolons instead of a full stops to indicate the end of a semantic unit. Furthermore, sentences are very long, nested and contain several information aspects.

We tackle these challenges by building upon the *NLTK*[5] sentence tokenizer, a pre-trained open source model to split a text into sentences. We adapt it to the domain of legal human rights documents by manually adding common abbreviations such as *art., para., crim.,* etc. Furthermore, a minimum (4) and maximum (40) number of words per sentence is defined to keep them in a certain length range and to avoid either too short and maybe meaningless phrases, as well as too long, nested sentences. This form of standardisation is necessary for the training and prediction of the sentence classifier and the user experience when displaying sentences.

[5] The *Natural Language Toolkit* (NLTK) is a *Python* open source library containing tools for natural language processing, such as text classification, tokenisation, stemming, tagging, parsing, and semantic reasoning.



HUMAN RIGHTS COUNCIL
Twelfth session
Agenda item 6

**UNIVERSAL PERIODIC REVIEW**

**Decision adopted by the Human Rights Council**[*]

**12/102. Outcome of the universal periodic review: Monaco**

*The Human Rights Council,*

*Acting in compliance* with the mandate entrusted to it by the General Assembly in its resolution 60/251 of 15 March 2006 and Council resolution 5/1 of 18 June 2007, and in accordance with the President's statement PRST/8/1 on modalities and practices for the universal periodic review process of 9 April 2008;

*Having conducted* the review of Monaco on 4 May 2009 in conformity with all the relevant provisions contained in Council resolution 5/1;

*Adopts* the outcome of the universal periodic review on Monaco which is constituted of the report of the Working Group on Monaco (A/HRC/12/3), together with the views of Monaco concerning the recommendations and/or conclusions, as well as its voluntary commitments and its replies presented before the adoption of the outcome by the plenary to questions or issues that were not sufficiently addressed during the interactive dialogue in the Working Group (A/HRC/12/50, chapter VI).

*14th meeting*
*23 September 2009*

[Adopted without a vote.]

[*] The resolutions and decisions adopted by the Human Rights Council will be contained in the report of the Council on its twelfth session (A/HRC/12/50), chap. I.

**Figure 1:** Sample document of the Human Rights Council  illustrates the challenges of sentence splitting.

Sentences that are longer than the maximum sentence length are split into semantic parts by searching for punctuation that indicates semantic breaks such as  *; : , “ …* . If none of these is found, the sentence is split in the middle.

This approach significantly improves the performance of the *NLTK sentence tokenizer* for the type of human rights related legal documents with which HURIDOCS' partner organisations work.

## Integration of the Universal Sentence Encoder

In Uwazi a user highlights a sentence that is relevant for her research question or categorisation task. This sentence is then passed to the already implemented sentence classifier which learns to distinguish relevant from irrelevant content, in respect to this specific research purpose. It can identify and extract similar phrases in documents that have not been processed. This way users can find other relevant documents and get a quick insight into the topic-specific aspects of long text.

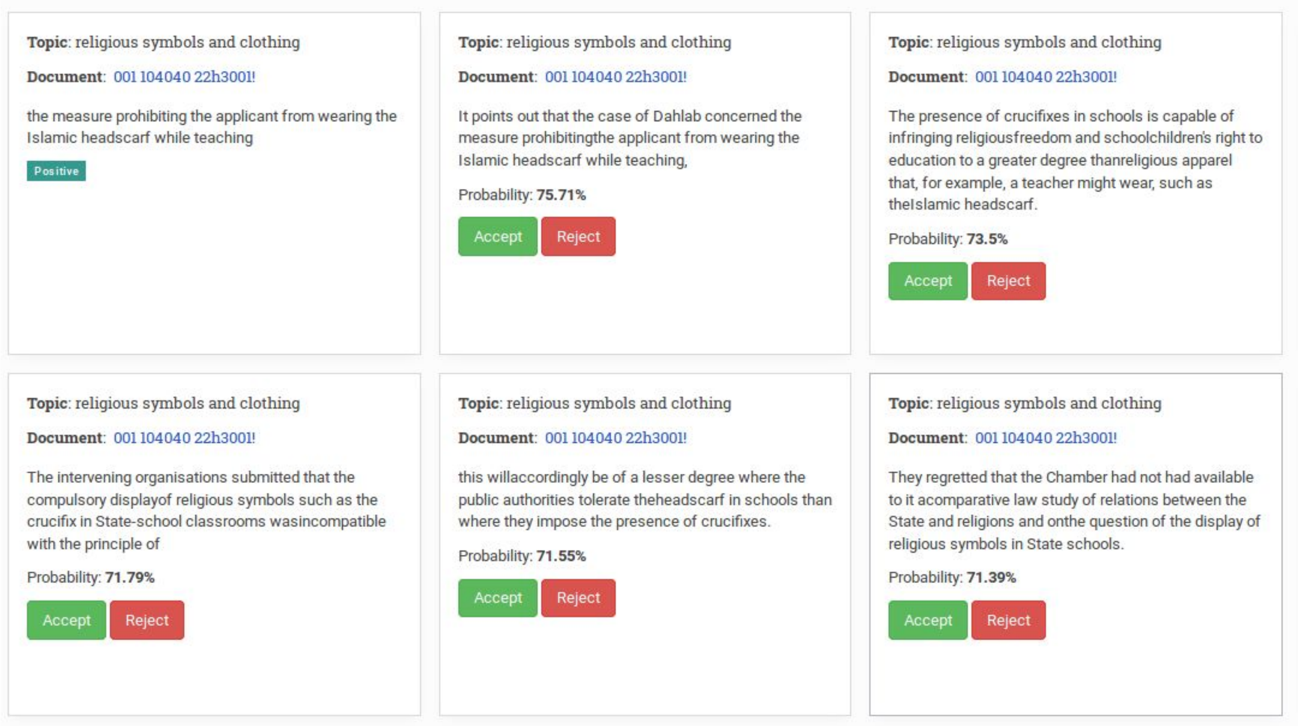
However, the main drawback in training a machine learning classifier is that it requires data to learn from. The more, the better. The scarcity if data leads to wrong, imprecise and misleading results that might discourage human rights defenders or even increase the amount of time they spend on organising their documents.

To overcome this challenge, we use *Google's Universal Sentence Encoder*[6] (Cer et. al., 2018). It is a machine learning model that has been trained on a variety of different text data and optimized to solve multiple natural language understanding tasks. As it represents the content of a sentence in a high-dimensional vector, it is a powerful tool to compare the semantic similarity of texts. This way, highlighting only one sample sentence is enough to gather more data to properly train the sentence classifier.

The extraction of semantically similar sentences goes beyond a string comparison of words. The high-dimensional representation of sentences enables the integration of morphological, contextual and semantic associations.

Consider the example scenario in Figure 2. The sentence in the upper left corner has been highlighted by a user as relevant for the category *religious symbols and clothing.* With this single sentence the Universal Sentence Encoder is able to identify phrases with a similar content even though a different wording is used. It is able to associate the word *teaching* with *schools* and *Islamic headscarf* with *religious symbols* and even identifies *crucifixes* as related symbols in Christian religion.

[6] More information on the module and its way of using can be found here.



**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

the measure prohibiting the applicant from wearing the Islamic headscarf while teaching

**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

It points out that the case of Dahlab concerned the measure prohibitingthe applicant from wearing the Islamic headscarf while teaching,

Probability: **75.71%**

**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

The presence of crucifixes in schools is capable of infringing religiousfreedom and schoolchildren's right to education to a greater degree thanreligious apparel that, for example, a teacher might wear, such as theIslamic headscarf.

Probability: **73.5%**

**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

The intervening organisations submitted that the compulsory displayof religious symbols such as the crucifix in State-school classrooms wasincompatible with the principle of

Probability: **71.79%**

**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

this willaccordingly be of a lesser degree where the public authorities tolerate theheadscarf in schools than where they impose the presence of crucifixes.

Probability: **71.55%**

**Topic**: religious symbols and clothing

**Document**: 001 104040 22h3001!

They regretted that the Chamber had not had available to it acomparative law study of relations between the State and religions and onthe question of the display of religious symbols in State schools.

Probability: **71.39%**

**Figure 2:** Sentence suggestions provided by the Universal Sentence Encoder when searching for content similar to the manually selected sentence (upper left corner).

Users can very quickly accept or reject suggestions made by the sentence encoder. This is valuable because it allows to generate a big dataset in a very short amount of time to reliably train the sentence classifier. The sentence classifier itself is needed to refine and specify the results of the universal sentence encoder to a particular category.

## Sentence Classifier - Optimisation and Evaluation of the Convolutional Neural Network

So far we referred to the machine learning model as sentence classifier. In this section we will give more details about the underlying convolutional neural network in order to understand how it has been optimised and evaluated during the Aaron Swartz fellowship.

In 2014, Kim Yoon showed that convolutional neural networks are a good baseline for sentence classification[7]. The input is a sentence in which every word is represented as a vector based on a pre-trained word embedding model. The output is the probability that it belongs to a predefined category, e.g. *religious symbols and clothing.*

[7] For more information on Understanding Convolutional Neural Networks for NLP and Implementing a CNN for Text Classification in Tensorflow see the linked blog posts by Denny Britz (2015).

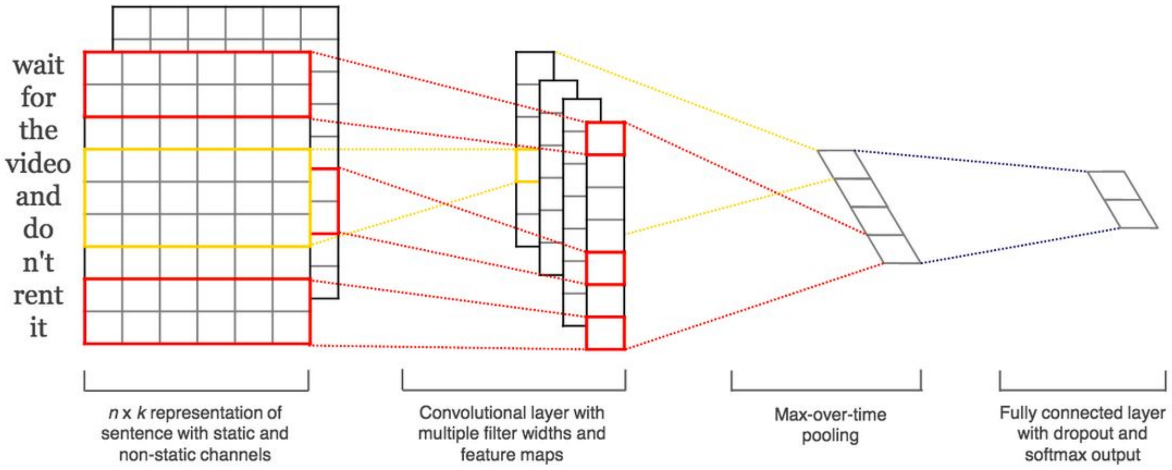

wait
for
the
video
and
do
n't
rent
it
*n* x *k* representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

**Figure 3:** Architecture of the convolutional neural network used for sentence classification (Yoon, 2014).

The advantage of a shallow convolutional neural network (see Figure 3) for sentence classification is that little preprocessing, e.g. considering sequences of words that belong together such as *Human Rights Council,* is needed. Compared to other deep learning approaches with a similar performance the shallow neural networks need less training data and only little parameter tuning.

However, one big challenge of this project is its aim for generality. So, independent of the collection of documents, as well as the specific research question, the trained algorithms should be as reliable and precise as possible when suggesting relevant information.

Machine learning algorithms contain lots of parameters that are optimised to perfectly fit its application. For example, Figure 3 displays filters (the yellow and red squares in the first column) of different size. These filters increase the level of natural language understanding by combining several words into semantic units. While the size and number of these filters can be set arbitrarily, they have a significant impact on the accuracy and computation time of the sentence classifier (Zhang and Wallace, 2016).
Other influential parameters are the batch size which controls how many sentences are fed to the classifier at once, and the epoch size which determines how often the entire training dataset is shown to the classifier. It's important to find good values for these parameters.
However, as we have only a very small number of training sentences available, an extensive parameter optimisation of the convolutional neural network is not possible.

In order to find a parameter setting that is as accurate and generic as possible we set up a test database containing different types of metadata from representative collections of documents

HURIDOCS is working with: resolutions and reports of the Human Rights Council[8], sentencing decisions involving sexual and gender-based violence cases in the Pacific Islands[9] and Grand Chamber Judgments of the European Court of Human Rights[10]. The different collections resemble the challenges of human rights organisations when analysing or publishing documents: assigning labels regarding the content (e.g. freedom of speech, women's rights, data protection, etc.), analysing one aspect related to a specific research purpose (e.g. the victim age in sexual assault cases) and extracting metadata to make the collections easier accessible (e.g. document type, publication date, involved persons and organisations). These different types of metadata and collections ensure that the neural network leads to reasonable results in different real world use cases.

The parameters of the convolutional neural network are adjusted such that they fit all of these cases reasonably well. The performance a machine learning model is measured by:

- accuracy: the ratio of sentences that are correctly labelled
- precision: the ratio of all suggestions that are correct
- recall: the ratio of all relevant sentences that are identified correctly

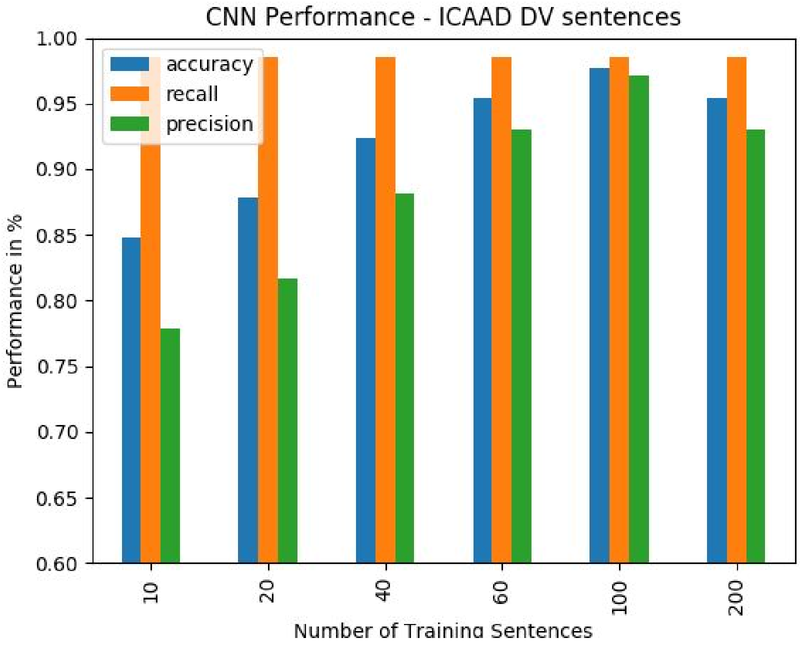


**Figure 4:** Evaluation of the performance (accuracy, precision, and recall) of a convolutional neural network trained with different numbers of training sentences.

[8] Available at https://www.right-docs.org/

[9] Taken from the Pacific Islands Legal Information Institute database http://www.paclii.org/

[10] The European Court of Human Rights publishes their judgements, decisions, resolutions and other documents with their HUDOC database.

Figure 4 shows the performance of a model which was trained to find sentences related to *domestic violence* cases within the ICAAD dataset. Even when trained with only 10 documents the model performs reasonably well.
Please note, that the recall is constantly very high, which means that almost all *domestic violence* related sentences in the dataset are found. This is crucial, as it is easier for users to reject wrongly extracted sentences, than to manually browse through the collection to find relevant information which the algorithm has missed.

Another important factor, especially when integrating machine learning algorithms into a software product, is the time needed for training the classifier and making predictions. Figure 5 shows the influence of the number of training sentences on computation time. To avoid that users have to wait for too long, the number of iterations can be adapted.

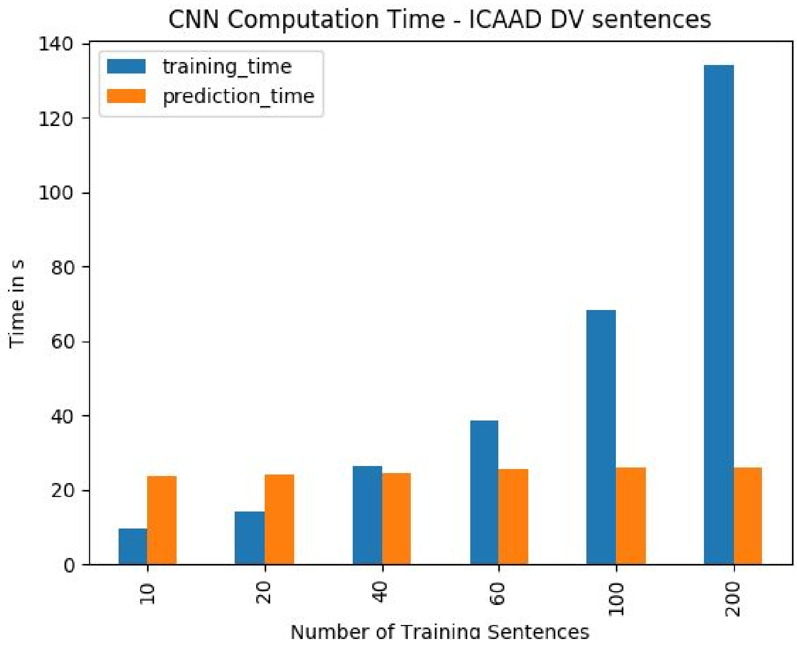


**Figure 5:** Evaluation of the computation time needed for training the convolutional neural network and making predictions with it.

The adjustment of parameters and the evaluation of models is an ongoing process which over time will include more datasets and an increasing number of different types of metadata.
Finally, the real value and performance of machine-assisted metadata extraction will be measured by the human rights defenders who use the tool.

## User Manual

An important goal of this project is to make machine learning algorithms accessible and intuitively usable for human rights defenders and researchers. However, to ensure that their self-trained models are really supportive and beneficial, it is necessary that users understand the basic concepts of machine learning. Users have to recognise that machine learning algorithms need enough examples to be train reliably and the data they feed into them has a huge impact on the results.

A user manual is a way to communicate these underlying concepts, and to demonstrates how in Uwazi a customised algorithm is trained and how its suggestions can be further improved. While the manual serves as an accessible introduction to certain machine learning approaches such as supervised and unsupervised learning, word embeddings, etc., it also points interested readers to more detailed resources.

As the workflow in Uwazi is still under development, the user manual is not yet publicly available.

## Human Rights and Machine Learning Discussion

Besides a manual which is focused on the user interaction with the machine learning algorithms in Uwazi, it is important to initiate an open discussion around the topic of human rights and machine learning.

We have seen that machine learning is a powerful tool that while offering tremendous opportunities in the field of human rights, also raises significant ethical issues. Algorithmic biases have the potential to completely change the lives of individuals, as well as to reinforce and even accelerate existing social and economic inequalities.

Our goal as human rights defenders is to distinguish beneficial machine learning systems from harmful automated decision-making processes in order to minimise the risks and maximise the impact of new technologies in human rights work. For this it is necessary to bring together human rights defenders, tool developers and machine learning practitioners to share their knowledge and experience with machine learning techniques in this field.

Therefore, from 18th till 22nd of June, HURIDOCS and the Center for Human Rights Science at Carnegie Mellon University held a New Tactics online conversation on this topic[11].
We talked about the very basic concepts of machine learning, tutorial and tools on how to get started, as well as already existing projects in the field of human rights and the impact of this technology on society and legal systems. Due to limited resources in non-profit organizations, it is

[11] The online conversation on Machine Learning and Human Rights is online available at the New Tactics forum.

often a very small team or even a single person who works on applying data science or machine learning. Therefore, a platform for machine learning practitioners to exchange experience and to talk about struggles in implementing and communicating their work is even more important.

About 360 people from 60 countries followed the discussion, which shows that there is great interest in this intersection of fields. The online discussion is a very good start for building up such a community of human rights and machine learning. We will continue these efforts in the long run.

## NEXT STEPS

Within the three months of the Aaron Swartz Fellowship we were able to implement and significantly improve the machine-assisted extraction of metadata in Uwazi. However, every improvement also widens the horizon of possibilities. This sections describes next steps in the project.

### Information extraction layer

To improve the user experience, a second layer of information extraction is needed. After highlighting or accepting a machine-suggested sentence, its relevant content should be automatically extracted in the metadata specific format.
For example, in an ideal scenario, from the sentence '*At the 14th of August 2011, the perpetrator was sentenced to 2 years imprisonment.'* the algorithm automatically extracts and standardises the date of the judgement to *14/08/2011*, as well as sets *2 years* as sentence length.

### Adaptation of the user workflow

The implemented two-step process consisting of the sentence encoder to extract similar sentences and a convolutional neural network to learn user specific research interests requires an adaption of the workflow.

While researchers or investigative journalists strive to connect persons or organisations with evidences or to browse through relevant snippets of documents, archivists or curators with the aim to publish a collection, have a different workflow. For them it's important to assess all possibly interesting categories and metadata of a document in a structured and complete way.

A detailed user testing with different groups and aims is needed to refine the workflow of machine-assisted metadata extraction in Uwazi.

## CONCLUSION

The three months I spent as Aaron Swartz fellow at the Open Society Archives in Budapest were a very inspiring and productive time.

By implementing a machine-assisted extraction of metadata human rights defenders and researchers are supported in their work of organising and analysing large collections of documents. With minimal effort users can extract semantically similar content and efficiently label the documents accordingly. Integrated in Uwazi this tool becomes intuitively usable such that human rights defenders can adapt it to their specific research interests.
Even though there is a long way ahead, step-by-step non-profit organisations can access and make use of the benefits of machine learning to help advance their cause.
Furthermore, human rights activists, AI experts and NGOs have laid the groundwork for building a strong machine learning and human rights community to come together, discuss and learn from each other. The efforts to use machine learning for good and to strengthen this community will continue.

The open exchange with artists, historians, researchers and archivists at the Open Society Archives in Budapest gave valuable insights into their personal way of transforming unstructured text to knowledge. Their view on technology and the changes it brings to the field they work in underlines the challenges of researchers in the digital age, but also paves the way for technological support.

In the spirit of Aaron Swartz, this project is a tribute to open source technology.
Without people who implement, document and share their knowledge about machine learning and natural language processing methods in an open and accessible way, this project would not have been possible. With this in mind,  Uwazi and the implemented machine-assisted extraction of metadata are available as open-source code.

## REFERENCES

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal Sentence Encoder. arXiv:1803.11175, 2018.

[2] Yoon Kim. Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882, 2014.

[3] Ye Zhang and Byron C. Wallace. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. arXiv:1510.03820, 2017.