ALMA MATER STUDIORUM

Second cycle degree in Artificial Intelligence

Advanced techniques for cross-language annotation projection in legal texts

Relatore: Chiar.mo
Prof. **Paolo Torroni**
Correlatore: Chiar.mo
Prof. **Andrea Galassi**

Tesi di Laurea di:
Antici Francesco

Academic year 2020/2021

Contents

Abstract

Advanced techniques for cross-language annotation projection in legal texts

Nowadays, the majority of the services we benefit from, are provided online and their use is regulated by the acceptance to the terms of service by the users. All our data are handled accordingly with the clauses of such document and all our behaviours must comply with it. Given so, it would be very useful to find automated techniques to ensure fairness of the document or inform the users about possible threats. The focus of this work, is to create resources aimed to the development of such tools in languages other than English, which may lack in linguistic resources and annotated corpus. The enormous breakthroughs of the last years in Natural Language Processing techniques made it possible the creation of such tools through automated and unsupervised process. One of the means to achieve that is through the annotation projection between two parallel corpora. The difficulties and costs of creating ad hoc resource for every language has brought the need to find another way for achieving the goal.

This work investigates the cross language annotation projection technique based on sentence embedding and similarity metrics to find matches between sentences. Several combination of methods and algorithms are compared, among which there are monolingual and multilingual embedding neural models. The experiments are conducted on two datasets, where the reference language is always English and the projection are evaluated on Italian, German and Polish. The results obtained provide a robust and reliable technique for the task and a good starting point to build multilingual tools.

1 Introduction

Being able to inscribe human knowledge into machineries, has been the main quest of computer scientist, in the field of Artificial Intelligence, for the last 50 years. The recent development of the technologies we have access to everyday, has brought to enormous breakthroughs in Artificial Intelligence and all of its subfields, among which, a special mention goes to Natural Language Processing. This discipline concerns the study of natural language and the capabilities of creating tools which can understand and process text like us, humans, would do. Natural Language Processing has provided us with high-efficiency tools that can help us in many ways. The downside of this technology, is that the development of each tools require a lot of resources and labelled data, which are not always easy to obtain, especially for languages other than English. The focus of this work is to investigates techniques aimed to automatically create such resources and annotated data, in order to help and increase the development of linguistic tools even for low-resources languages. I am going to present the annotation projection method to transfer labels between two parallel corpus. This technique relies on a series of steps to find matches, in terms of similarity of content, between the sentences of two documents and then transfer the knowledge. Different embedding techniques will be studied aiming to find the best way to represent sentences and different combination of algorithm and similarity metrics will be tested to come up with a robust, efficient and performing method to achieve the goal.

2 Background

2.1 Artificial Intelligence

The last decades have been marked by enormous breakthroughs in the field of information technologies, due to the development of the ICT industry and the accessibility to computing power in computers, leading us to having several branches of informatics developing in parallel and autonomously. Among those there's Artificial Intelligence, which is one of the most studied field in the last years, despite its invention goes way back to the middle of the 20th century.

2.1.1 History of AI

During and after World War 2, technology started playing a significant role in everyday's life, making more and more scientists and engineers start to question themselves about how to turn machines into intelligent tools. Computer science defines AI research as the study of intelligent agents [1], i.e. any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals. A more elaborate definition characterizes AI as "a system's ability to correctly interpret external data, to learn from such data, and to use this "experience" to achieve specific goals and tasks through flexible adaptation. One of the pioneer of AI was Alan Turing who proposed in his paper "Computing Machinery and Intelligence"[2], to switch the quest for inferring intelligence, to the one for making the machines mimic what us(thinking entities) would do. The field of AI research was born at a workshop at Dartmouth College in 1956[3], where the term "Artificial Intelligence" was coined by John McCarthy to distinguish the field from cybernetics. Attendees Allen Newell (CMU), Herbert Simon (CMU), John McCarthy (MIT), Marvin Minsky (MIT) and Arthur Samuel (IBM) became the founders and leaders of AI research. They and their students produced pro-

grams that the press described as "astonishing": computers were learning checkers strategies and by 1959 were reportedly playing better than the average human. Computers could solve word problems in algebra, proving logical theorems and speaking English. By the middle '60s, research in the U.S. was heavily funded by the Department of Defense and laboratories had been established around the world. AI's founders were optimistic about the future: Herbert Simon predicted, "machines will be capable, within twenty years, of doing any work a man can do". Marvin Minsky agreed, writing, "within a generation the problem of creating 'artificial intelligence' will substantially be solved"[4]. They failed to recognize the difficulty of some of the remaining tasks, resulting in a harsh hinder in progress and research, due to ongoing pressure from the US Congress to fund more productive projects, both the U.S. and British governments cut off exploratory research in AI. The next few years would later be called an "AI winter", a period when obtaining funding for AI projects was difficult. In the late 1990s and early 21st century, AI began to rise again, being used for logistics, data mining, medical diagnosis and other areas. The success was due to increasing computational power, greater emphasis on solving specific problems, new ties between AI and other fields (such as statistics, economics and mathematics) and a commitment by researchers to mathematical methods and scientific standards. Deep Blue[5] became the first computer chess-playing system to beat a reigning world chess champion, Garry Kasparov, on 11 May 1997. In the same year, Hochreiter and Jürgen Schmidhuber proposed a neural architecture called Long Short-Term Memory (LSTM)[6], a type of a recurrent neural network used today in handwriting recognition and speech recognition and one year later LeCun, Yoshua Bengio and others started publishing papers on the application of neural networks to handwriting recognition and on optimizing backpropagation[7]. The start of the 21th century AI began popular even in masses' culture with the release of several films about robots

and intelligent agents, this too helped arousing incredible interest around the field all over the world and in science. In 2006 Hinton publishes “Learning Multiple Layers of Representation,”[8] summarizing the ideas that have led to “multilayer neural networks that contain top-down connections and training them to generate sensory data rather than to classify it,” i.e., the new approaches to deep learning, revolutionizing the subject for ever. In 2010 the first AI based competition was launched: The ImageNet Large Scale Visual Recognition Challenge (ILSVCR), an annual AI object recognition competition. Faster computers, algorithmic improvements, and access to large amounts of data enabled advances in machine learning and perception; data-hungry deep learning methods started to dominate accuracy benchmarks around 2012. The Kinect, which provides a 3D body–motion interface for the Xbox 360 and the Xbox One, uses algorithms that emerged from lengthy AI research as do intelligent personal assistants in smartphones. In March 2016, AlphaGo[**alphago**] won 4 out of 5 games of Go in a match with Go champion Lee Sedol, becoming the first computer Go-playing system to beat a professional Go player without handicaps. In the 2017 Future of Go Summit, AlphaGo won a three-game match with Ke Jie, who at the time continuously held the world No. 1 ranking for two years, which marked the completion of a significant milestone in the development of Artificial Intelligence as Go is a relatively complex game, more so than Chess.

2.1.2 Ai in today’s society

According to Bloomberg’s Jack Clark[9], 2015 was a landmark year for artificial intelligence, with the number of software projects that use AI within Google increased from a "sporadic usage" in 2012 to more than 2,700 projects. Clark also presents factual data indicating the improvements of AI since 2012 supported by lower error rates in image processing tasks. He attributes this to an increase

in affordable neural networks, due to a rise in cloud computing infrastructure and to an increase in research tools and datasets. Other cited examples include Microsoft's development of a Skype system that can automatically translate from one language to another and Facebook's system that can describe images to blind people. In a 2017 survey, one in five companies reported they had "incorporated AI in some offerings or processes". Around 2016, China greatly accelerated its government funding; given its large supply of data and its rapidly increasing research output, some observers believe it may be on track to becoming an "AI superpower". Nowadays AI has become an important part of our society and more and more institutions are starting to come up with regulamentations for the use of it. The EU, for example, has a Commission's Communication on AI and other commitees to supervise the use and the outcomes of AI. The definition proposed within the European Commission's Communication on AI is the following[10]:

"Artificial intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions, with some degree of autonomy, to achieve specific goals. AI based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)."

2.2 Natural Language Processing

One of the most important subfield of Artificial Intelligence is doubtlessly Natural Language Processing(NLP).

NLP is not related only to AI, but also to linguistics and computer science in general, since it concerns giving computers the ability to understand text and spoken words ideally in the same way human beings can. The goal is having

tools capable of analysing the content of documents in order to extract information and insights contained in the document or categorize the document itself. Challenges in natural language processing frequently involve speech recognition, natural language understanding, and natural-language generation. Since the so-called "statistical revolution" in the late 1980s and mid-1990s, much natural language processing researches have relied heavily on machine learning. This technique relies on learning pattern and rules through the analysis of large corpora of typical real-world examples, instead of using statistical inference.

Many different classes of machine-learning algorithms have been applied to natural-language-processing tasks. These algorithms take as input a large set of "features" that are generated from the input data. Increasingly, however, research has focused on statistical models, which make soft, probabilistic decisions based on attaching real-valued weights to each input feature. Such models have the advantage of expressing the relative certainty of many different possible answers rather than only one, producing more reliable results when such a model is included as a component of a larger system.

2.2.1 Neural revolution

A major drawback of statistical methods is that they require elaborate feature engineering. In the last decade the field has thus largely abandoned statistical methods and shifted to neural networks. In some areas, this shift has entailed substantial changes in how NLP systems are designed, such that deep neural network-based approaches may be viewed as a new paradigm distinct from statistical natural language processing. Since the neural turn, statistical methods in NLP research have been largely replaced by neural networks, continuing, though, to be relevant for contexts in which statistical interpretability and transparency is required[11].

2.3 Word and sentence embeddings

Popular techniques include the use of word/sentence embeddings to capture semantic properties of words, and an increase in end-to-end learning of a higher-level task (e.g., question answering) instead of relying on a pipeline of separate intermediate tasks (e.g., part-of-speech tagging and dependency parsing). These models assign a vectorized representation of words/sentences dependently on the context or the word as seen in other examples or in a corpus. Methods like these gained a lot of popularity in the last years, concurrently with an exponential growth of the state of the art of them. One of the first neural method used to achieve word/sentence embeddings was the word2vec[12]. It was created, patented, and published in 2013 by a team of researchers led by Tomas Mikolov at Google over two papers. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such model can detect synonymous words or suggest additional words for a partial sentence, since it is trained to assign a similar embedding to similar word/sentence. The similarity is given by simple mathematical functions like the cosine distance between the vectors provided by the algorithm. A similar solution is provided by Glove[13]. In both cases the algorithms can utilize either of two model architectures to produce a distributed representation of words: continuous bag-of-words or continuous skip-gram. In the continuous bag-of-words architecture, the model predicts the current word from a window of surrounding context words. In the continuous skip-gram architecture, the model uses the current word to predict the surrounding window of context words. The skip-gram architecture weighs nearby context words more heavily than more distant context words. The drawback of these solution is that they provide context-free embedding, meaning that the context in which is used that particular word/sentence won't affect the embedding. Contextual models have been heavily developed in the last years, one of firsts was Elmo[14]. Elmo is a pre-trained

model based on a deep neural architecture that can provide word vectors based on context and able to model complex characteristics of the language. It uses a deep, bi-directional LSTM model to create word representations. Rather than a dictionary of words and their corresponding vectors, ELMo analyses words within the context that they are used. It is also character based, allowing the model to form representations of out-of-vocabulary words. This therefore means that the way ELMo is used is quite different to word2vec or glove. Rather than having a dictionary 'look-up' of words and their corresponding vectors, ELMo instead creates vectors on-the-fly by passing text through the deep learning model. This network is very popular and performs very well, but, the state of the art performances up to this date are achieved by the Bert encoder[15], which is one of the most suitable example of the neural revolution in NLP. Bidirectional Encoder Representations from Transformers(BERT) is a state of the art word embedding technique which achieved astonishing results and it has been widely used for every kind of tasks in the field. BERT makes use of Transformer[16], an attention[17] mechanism that learns contextual relations between words in a text. Transformers include two separate mechanisms: an encoder that reads the text input and a decoder that produces a prediction for the task. The architecture is pretty different from the other neural methods which relied on recurrent networks to predict the output. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary. BERT has its origins from pre-training contextual representations including Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit[18]. Unlike previous models, BERT is a deeply bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. BERT takes into account the context for each occurrence of a given word. For instance, whereas the vector for "running" will have the same word2vec vector representation for both of its occurrences in the sentences "He is running a company" and "He is running a

marathon", BERT will provide a contextualized embedding that will be different according to the sentence. In October 2020, almost every single English-based query was processed by BERT[19]. The original English-language BERT has two models:

1. the BERTBASE: 12 Encoders with 12 bidirectional self-attention heads;
2. the BERTLARGE: 24 Encoders with 16 bidirectional self-attention heads.

Both models are pre-trained from unlabeled data extracted from the Google Books Corpus with 800M words and English Wikipedia with 2,500M words.

2.4 Ai/Nlp for legal systems

Given the endless applications of Ai and Nlp, in the recent years, it has taken place the concept of LegalAI. Legal Artificial Intelligence (LegalAI) focuses on applying the technology of artificial intelligence, especially natural language processing, to benefit tasks in the legal domain. The typical applications are Legal Judgment Prediction, Similar Case Matching and Legal Question Answering.

2.4.1 Legal Judgement Prediction

Legal Judgment Prediction (LJP)[20] is one of the most critical tasks in LegalAI, especially in the Civil Law system. In the Civil Law system, the judgment results are decided according to the facts and the statutory articles. One will receive legal sanctions only after he or she has violated the prohibited acts prescribed by law. The task LJP mainly concerns how to predict the judgment results from both the fact description of a case and the contents of the statutory articles in the Civil Law system, among the approaches that constitute the state of the art, we must rank[21]. While most existing works only focus on a specific subtask of judgment prediction and ignore the dependencies among subtasks, the authors of [21] formalize the

dependencies among subtasks as a Directed Acyclic Graph (DAG) and propose a topological multi-task learning framework, TopJudge, which incorporates multiple subtasks and DAG dependencies into judgment prediction.

2.4.2 Similar Case Matching

In those countries with the Common Law system like the United States, Canada, and India, judicial decisions are made according to similar and representative cases in the past. As a result, how to identify the most similar case is the primary concern in the judgment of the Common Law system. In this field, where trying to find similar content is essential, the aforementioned embedding encoder Bert gained optimal results, as well as [22].

2.4.3 Legal Question-Answering

Another typical application of LegalAI is Legal Question Answering (LQA) which aims at answering questions in the legal domain. One of the most important parts of legal professionals' work is to provide reliable and high-quality legal consulting services for non-professionals. In this field the research is still challenging, since the results are not comparable to the advice of a professional, once again Bert is the most used model for this task[23].

3 Contracts

We live in a society where pretty much all of our trading interactions are safeguarded by some kind of contracts. These need to be fair for both the parties who sign them and every form of unfairness in the terms should be punished severely. The developing of internet services and applications has brought to the attention the need for regulamentations to protect both the consumer and the provider.

3.1 ToS and PP

The publishing of the GDPR[24] forced the providers to ask for the consent of the user to use data accordingly with the manners explicited in the Privacy Policy document, while the users must comply with the behaviours listed in the Terms of Service.

3.1.1 Terms of Service

Terms of service[24] are the legal agreements between a service provider and a user, who is to agree to abide by the terms of service in order to use the offered service. The document tells the customers what will be legally required of them if they subscribe to the service and provides the company with a legal leg to stand on in the event of abuse or litigation.

3.1.2 Privacy Policy

A privacy policy[24] is a statement or legal document that discloses some or all of the ways a party handles a customer or client's data. This requirement has come from regulations like the European Union's GDPR and California's CalOPPA that aim to protect personal information. Privacy regulations have been created by

governmental bodies, like the aforementioned California and EU, to protect their citizens' privacy.

3.2 Unfairness in contracts

3.2.1 EU Regulamentation about unfairness in contracts

In 1993, the EU published the Unfair Contract Terms Directive[25] which protects consumers against unfair standard contract terms imposed by traders. It applies to all kinds of contracts on the purchase of goods and services, for instance online or off-line-purchases of consumer goods, gym subscriptions or contracts on financial services, such as loans. The definition of unfair terms provided by the third article of the document are the following:

- A contractual term which has not been individually negotiated shall be regarded as unfair if, contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer.

- A term shall always be regarded as not individually negotiated where it has been drafted in advance and the consumer has therefore not been able to influence the substance of the term, particularly in the context of a pre-formulated standard contract. The fact that certain aspects of a term or one specific term have been individually negotiated shall not exclude the application of this Article to the rest of a contract if an overall assessment of the contract indicates that it is nevertheless a pre-formulated standard contract. Where any seller or supplier claims that a standard term has been individually negotiated, the burden of proof in this respect shall be incumbent on him.

This means that there are some types of clauses that traders are prohibited from using in the contracts and there is unfairness whenever a term cause significant imbalance in the parties' rights and obligations to the detriment of the consumer. The Directive has been amended by Directive (EU) 2019/2161 of 27 November 2019 on better enforcement and modernisation of Union consumer protection rules, part of the 'Review of EU consumer law - New Deal for Consumers". The amendment introduces an obligation for Member States to provide for effective penalties in case of infringements. It has to be transposed by 28 November 2021 and applied from 28 May 2022. Standard contract terms have to be drafted in plain intelligible language and ambiguities are to be interpreted in favour of consumers. Examples of unfair terms are listed in the annex to the document:

(a) excluding or limiting the legal liability of a seller or supplier in the event of the death of a consumer or personal injury to the latter resulting from an act or omission of that seller or supplier;

(b) inappropriately excluding or limiting the legal rights of the consumer vis-a-vis the seller or supplier or another party in the event of total or partial non-performance or inadequate performance by the seller or supplier of any of the contractual obligations, including the option of offsetting a debt owed to the seller or supplier against any claim which the consumer may have against him;

(c) making an agreement binding on the consumer whereas provision of services by the seller or supplier is subject to a condition whose realization depends on his own will alone;

(d) permitting the seller or supplier to retain sums paid by the consumer where the latter decides not to conclude or perform the contract, without providing

for the consumer to receive compensation of an equivalent amount from the seller or supplier where the latter is the party cancelling the contract;

(e) requiring any consumer who fails to fulfil his obligation to pay a disproportionately high sum in compensation;

(f) authorizing the seller or supplier to dissolve the contract on a discretionary basis where the same facility is not granted to the consumer, or permitting the seller or supplier to retain the sums paid for services not yet supplied by him where it is the seller or supplier himself who dissolves the contract;

(g) enabling the seller or supplier to terminate a contract of indeterminate duration without reasonable notice except where there are serious grounds for doing so;

(h) automatically extending a contract of fixed duration where the consumer does not indicate otherwise, when the deadline fixed for the consumer to express this desire not to extend the contract is unreasonably early;

(i) irrevocably binding the consumer to terms with which he had no real opportunity of becoming acquainted before the conclusion of the contract;

(j) enabling the seller or supplier to alter the terms of the contract unilaterally without a valid reason which is specified in the contract ;

(k) enabling the seller or supplier to alter unilaterally without a valid reason any characteristics of the product or service to be provided;

(l) providing for the price of goods to be determined at the time of delivery or allowing a seller of goods or supplier of services to increase their price without in both cases giving the consumer the corresponding right to cancel

the contract if the final price is too high in relation to the price agreed when the contract was concluded;

(m) giving the seller or supplier the right to determine whether the goods or services supplied are in conformity with the contract, or giving him the exclusive right to interpret any term of the contract;

(n) limiting the seller's or supplier's obligation to respect commitments undertaken by his agents or making his commitments subject to compliance with a particular formality;

(o) obliging the consumer to fulfil all his obligations where the seller or supplier does not perform his;

(p) giving the seller or supplier the possibility of transferring his rights and obligations under the contract, where this may serve to reduce the guarantees for the consumer, without the latter s agreement ;

(q) excluding or hindering the consumer s right to take legal action or exercise any other legal remedy, particularly by requiring the consumer to take disputes exclusively to arbitration not covered by legal provisions, unduly restricting the evidence available to him or imposing on him a burden of proof which, according to the applicable law, should lie with another party to the contract.

4 Automated unfairness detection system

4.1 Risks of unfair contracts

In our everyday life in order to use web services or applications we are required to agree upon a huge amount of contracts regulating the way we may or may not behave and how our data are handled. Despite the importance of them, several studies[26] proved that consumers hardly ever read the terms of what they are agreeing upon, exposing themselves to possible risks and threats. Whenever we mark the "i have read and agree to the terms and conditions" without actually reading the content of the *Terms of service*, we could be signing unfair contracts and in case of unwanted aftermaths we could be at risk. There are reasons why many consumers do not read or understand Terms of service, as well as privacy policies or end-user license agreements[27]. Reports indicate that such documents can be overwhelming to the few consumers who actually venture to read them[28]. It has been estimated that actually reading the privacy policies alone would carry costs in time of over 200 hours a year per Internet user[29]. Another problem is that even if consumers did read the ToS thoroughly, they would have no means to influence their content: the choice is to either agree to the terms offered by a web app or simply not use the service at all.

4.2 Prevention and countermeasures

It is important to highligth that once a contract is agreed upon and something undesired happen, it is very difficult, if not impossible, to prove the provider guilty, since the consumer gave his consent in the first place. The only way to fight this situations is prevention. The 93/13 Directive depict two mechanisms of prevention: individual and abstract control of fairness. The former requires the consumer to go to court and only after it is found that a clauses is unfair there is

no more binding on the consumer. However, most consumers do not take their disputes to courts. That is why abstract fairness control has been created. In each EU Member State, consumer protection organizations have the competence to initiate judicial administrative proceedings, to obtain the declaration that clauses in consumer contracts are unfair. Each state has its own ways of: applying abstract control, involving competent parties in the control process and punish providers who propose unfair contracts. As reported in[30] and in [31] the practice of placing unfair clauses in contracts is still widely used.

4.3 Claudette

In order to solve this problem it has been proposed in [31], an automated system based on machine-learning techniques to detect potentially unfair clauses in Terms of service and Privacy Policy.

4.3.1 Labels

In their work[30], the authors defined five main categories of potentially unfair clauses, which are often present in the contracts aforementioned:

1. establishing jurisdiction for disputes in a country different than consumers residence;
2. choice of a foreign law governing the contract;
3. limitation of liability;
4. the provider's right to unilaterally terminate the contract/access to the service;
5. the provider's right to unilaterally modify the contract/the service.

The authors of Claudette proposed 3 additional categories:

6. requiring a consumer to undertake arbitration before the court proceedings can commence;

7. the provider retaining the right to unilaterally remove consumer content from the service, including in-app purchases;

8. having a consumer accept the agreement simply by using the service, not only without reading it, but even without having to click on "I agree/I accept." The final set of labels is summed up in table 1.

Type of clause	**Symbol**
Arbritation	<a>
Unilateral change	<ch>
Content removal	<cr>
Jurisdiction	<j>
Choice of law	<law>
Limitation of liability	<ltd>
Unilateral termination	<ter>
Contract by using	<use>

Table 1: Macro groups of tags

4.3.2 Corpus annotation

The corpus is composed of 50 Terms of service provided by several on-line platforms, manually annotated accordingly with the labels in table 1. After the annotation draft each sentence of each document is tagged and formatted as an xml document where each tag represent an unfairness label. The final corpus contains 12,011 sentences overall, 1,032 of which (8.6%) were labeled as positive, thus

containing a potentially unfair clause. Arbitration clauses are the least common, being present in 28 documents only, whereas all the other categories appear in at least 40 out of 50 documents. Limitation of liability and unilateral termination categories represent more than half of the total potentially unfair clauses. The percentage of potentially unfair clauses in each document is quite heterogeneous, ranging from 3.3% (Microsoft) up to 16.2% (TrueCaller).

4.3.3 Automated detection techniques

In order to build a fully automated system, there's the need for a classification model. After segmenting, tokenizing and removing fragments shorter than 5 words, several Machine Learning/Deep Learning models were trained and then tuned on the validation set derived from the whole dataset. The models used are:

(a) a single SVM exploiting BoW (unigrams and bigrams for words and part-of-speech tags);

(b) a combination of eight SVMs (same features as above), each considering a single unfairness category as the positive class, whereby a sentence is predicted as potentially unfair if at least one of the SVMs predicts it as such;

(c) a single SVM exploiting TK[32] for sentence representation;

(d) a CNN trained from plain word sequences;

(e) an LSTM trained from plain word sequences;

(f) an SVM-HMM performing collective classification of sentences in a document (same features as a));

(g) a combination of eight SVM-HMMs, each performing collective classification of sentences in a document on a single unfairness category as the positive class (same setting as b));

(h) an ensemble method, that combines the output of all the other with a voting procedure (sentence predictive as positive if at least 3 systems out of 5 classify it as such).

The bad results of a-g led the authors to propose the ensemble method h), the results are listed in Fig.1.

Classifier	Method	P	R	F_1
C1	SVM – Single Model	0.729	0.830	0.769
C2	SVM – Combined Model	0.806	0.779	0.784
C3	Tree Kernels	0.777	0.718	0.739
C4	Convolutional Neural Networks	0.729	0.739	0.722
C5	Long Short-Term Memory Networks	0.696	0.723	0.698
C6	SVM-HMM – Single Model	0.759	0.778	0.758
C7	SVM-HMM – Combined Model	0.848	0.720	0.772
C8	Ensemble (C1+C2+C3+C6+C7)	0.828	0.798	0.806
	Random Baseline	0.125	0.125	0.125
	Always Positive Baseline	0.123	1.000	0.217

Figure 1: Results of the different techniques for automated unfairness detection, the table is taken from [31]

5 Cross lingual annotation

5.1 The linguistic problem

The universalizability, democratization and accessibility of tools and resources is one of the main aim of technological development and AI in general. This goal, though, is not always easily achievable in every field. One of the milestones of the EU fundamentals rights is the cultural diversity, including the linguistic one. All the documents and laws published by the European Parliament are published in all of EU's official languages with the same content, aiming to make the statements as clear as possible for every citizen.
As scientist, we wish that even technological tools could be available in as many language as possible. Concerning Natural Language Processing solutions, for example, this is very hard to achieve, since a lot of data, resource and works need to be collected for each specific language, making the task very costly, both in an economic and time-consuming meaning. In light of these alternative roads need to be found to achieve universalizability of tools.

5.2 Annotation projection

Given such premises, the annotation projection techniques are gaining more and more popularity in the last years. The main idea behind this approach is to have two sets of documents with the same content but in two different languages. If we had annotations on just one of the two we need to find a way to project them from the source(annotated) language to the target language.

The projection is achieved evaluating in some way the similarity between two sentences/words and projecting the knowledge of the most similar annotated data to the unlabelled one. These unsupervised methods don't require the creation of new datasets or ad-hoc models, just a method to match the information of sentences

in different languages.

The sentence alignment task is presented for the first time in the work of Simard and Plamondon[33] by defining a "corridor of alignment" based on global information. Basically, a candidate matching between sentences takes into account the position of the sentences inside each document. One of the most referenced work in this field is the one of Yarowsky et al.[34], which introduced a technique to project POS tagging in multilingual sentences between two corpora thanks to n-grams and statistical NLP methods. With the development of the NLP tools at our disposal and the improving of the state of the art of words' embedding, nowadays we can achieve very reliable and satisfying results in a task like the projection of the knowledge in different languages. Projection has been used also for argumentation mining[35], to create training data for machine learning models for low-resource languages, portuguese in their case. In particular, the authors argue against the necessity of human-translated parallel corpora as a resource, since they obtain comparable results using machine translated parallel documents. The projection of structural information between parallel documents is tackled by Bamman et al.[36], where alignment is performed firstly sentence-wise (1-1) and then word-wise. To address the task of aligning documents in which sentences do not appear in the same order, Zamani et al.[37] presented an approach based on Integer Linear Programming, which is the approach we are interested in, since, as we are going to see later on, our documents will suffer from asymmetry.

6 Cross lingual annotation projection in contracts

6.1 Problem definition

The aim of this work is to present a stable technique to perform cross lingual projection of annotation in the context of legal document. The focus, more specifically, is on on-line contracts such as Terms of Service and Privacy Policy.
The problem is defined as the task of transferring the knowledge, provided by legal experts and encoded in the form of annotations, into any target language. This knowledge in this particular case is the fairness/unfairness of clauses of the terms of the aforementioned on-line contracts, the labels and the unfairness matter are explained in section 3. Given the fact that these documents are translated in a lot of languages, it would be very useful, for people from different countries, to be able to have a direct access to the fairness of each clause in their native language too and not only in English.
Especially because it has been observed[38] that, too often, translated documents don't report the content of the original ones as correctly as desired. This can lead to have totally fair clauses in English turned into unfair in another language. Providing a tool to analyse those cases is vital. Several experiments have been performed in order to find the best transferring technique for the annotations, from new multilingual sentence embedding methods to using automated translating methods to have the target document in the same language as the labelled one.

6.2 DTW techinque

Dynamic Time Warping(DTW)[39] is an algorithm designed to compute similarity between temporal sequence which may vary in time and intensity. It measures the dissimilarity between pairs of elements of the two series to create a matrix. Each element of the matrix represents a matching between these elements, and

its value represents their dissimilarity, or cost, of the matching. The algorithm computes the cheapest path from one end to the other of the cost matrix. The alignment between the two series is given by the cells in the path, while the dissimilarity measure is the cost of the path. DTW has been applied to temporal sequences of video, audio, and graphics data, indeed, any data that can be turned into a linear sequence can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. It can also be used in partial shape matching applications. It guarantees to find an optimal alignment with quadratic complexity with respect to the length of the time series[REF TO PAPER]. The algorithm can be combined to any kind of distance metric to evaluate the sample of the sequences.

6.3 Automated translation process

The automated translation is achieved through an open source tool called apache joshua[40], which is a statistical machine translation toolkit for both phrase-based and syntax-based decoding.

6.4 Bray-Curtis dissimilarity metric

The Bray-Curtis dissimilarity[41], named after J. Roger Bray and John T. Curtis, between two numerical vectors a and b is defined as a normalized version of the Manhattan distance, since it is computed as the sum over the absolute differences between elements a_j and b_j, divided by the sum over the elements computed for each vector, separately.

$$D^{BC} = 1 - \frac{\sum_j^n |a_j - b_j|}{\sum_j^n |a_j + b_j|} \tag{1}$$

6.5 Projection steps

6.5.1 Data involved

The input, given a generic language L is defined by three resources:

- The original annotated English version of the document D_E.
- The original non annotated version of the document D_L.
- The automated translation of D_L in English: ${D_L}^t$.

The goal is to find a correspondence between the sentences in D_L and the sentences in D_E via the automatically translated sentences of ${D_L}^t$. In this way, the original an- notations associated with the sentences in D_E can be transferred from the English document into a sequence of corresponding labels in the target. All the correspondences are thus evaluated among pairs of English sentences. The choice of English as reference language is merely due to nature of our datasets but the techniques are independent from the annotated documents' language. The annotation projection algorithm is based on two main steps. The former is the computation of a set of matches between each sentence of the translated target document ${D_L}^t$ and one or more sentences of the source document D_E. While the latter is the straightforward projection of tags from ${D_L}^t$ to D_L, which have a 1 to 1 perfect match, given the construction of the documents.

6.5.2 Matches finding

The matches finding step consists of finding the most similar sentences in D_E and ${D_L}^t$, to do so, we used the most recent sentence embedding pre-trained models. This step allow us to have a numeric representation of sentences and make the match finding just a minimum dissimilarity search problem among all the possible combinations. The dissimilarity metric used in our experiments is the Bray-Curtis.

6.5.3 Tag projection

Given the construction of the documents, once every sentence in $D_L{}^t$ is tagged with the same label of its most similar sentence in D_E, every sentence in D_L is annotated accordingly.

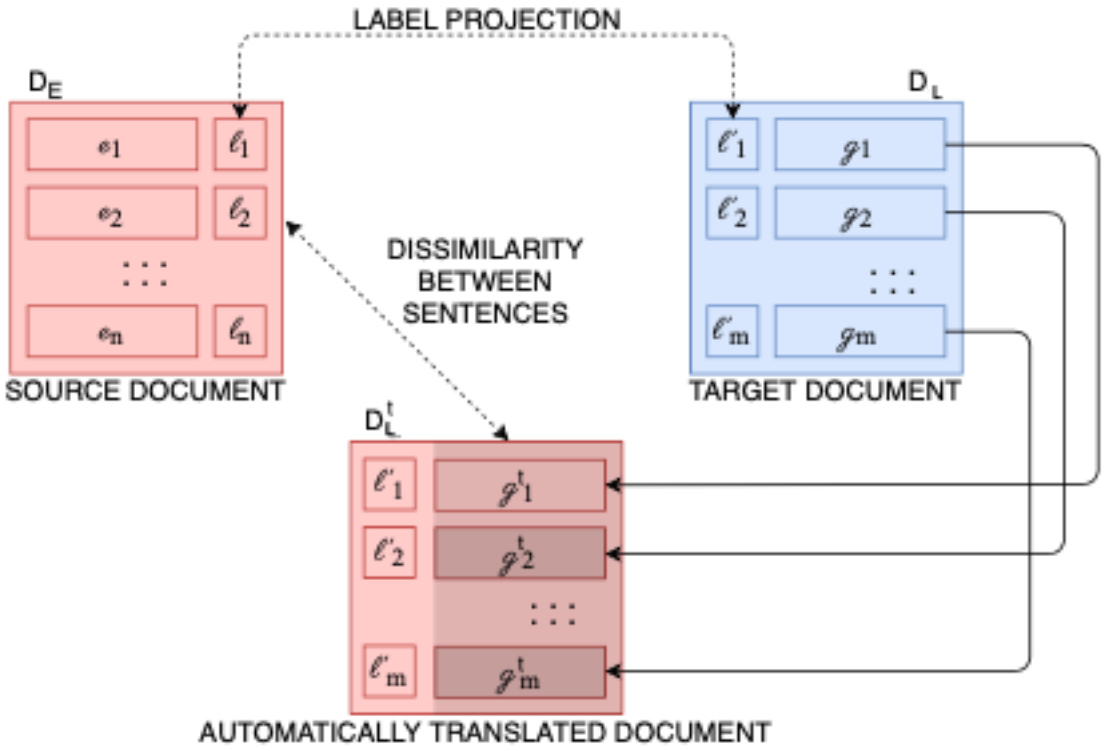

Figure 2: Projection architecture, originally provided by [42].

7 Datasets

The datasets used for our experiments are two, which are similar in the context but different in the composition.

7.1 Dataset 1

The first is an English-German corpus created for the task including ToS and PP. The documents were sourced from the CLAUDETTE training corpus[31] and the German versions were annotated by a legal expert fluent in English and German. The dataset composition and stats are described in table 3. The sources of the documents are listed in table 2 The terms of service (ToS) set consists of 5 contracts used by online service providers: Box.com, Garmin, Grindr, Linkedin and MyHeritage. It includes 2,808 sentences and 342 tags identifying 27 classes, divided into 9 categories, as described by [43]. The data privacy set (PP) comprises privacy policies from Dropbox, Facebook, Supercell, Tumblr and Twitter. The composition of the subsets of documents is reported in Tables 4 and 5. Annotations also indicate the degree of unfairness: for example, ltd3 means high degree of unfairness on grounds of limitation of liability, whereas ltd1 indicates a fair clause, i.e. it does not exclude the provider's liability.
Below are reported some tables showing the composition of the datasets, both in their whole and divided in their subsets(PP and ToS).

ToS Documents	PP Documents
Garmin	Dropbox
Box	Facebook
Grindr	Supercell
Linkedin	Tumblr
MyHeritage	Twitter

Table 2: "Source of ToS documents(left) and PP(right)"

Language	Sentences	Tags	a1	a2	a3	ad1	ad2	ad3	basis1	basis2	cat1	cat2	ch2	cr2	cr3	j1	j3	law1	law2	ltd1	ltd2	ltd3	pinc2	source1	source2	ta1	ta3	tc1	tc2	tc3	ter2	ter3	tpr1	tpr2	tu1	tu3	use2
English	2091	419	0	4	2	2	9	14	27	18	28	69	34	4	14	2	12	4	11	1	24	1	2	16	13	4	1	6	15	15	14	14	6	6	14	3	10
German	2336	540	1	6	3	3	12	15	33	22	38	89	37	7	16	3	12	4	11	1	38	3	2	22	15	6	1	8	22	20	15	17	9	12	21	3	13

Table 3: Composition of the whole dataset.

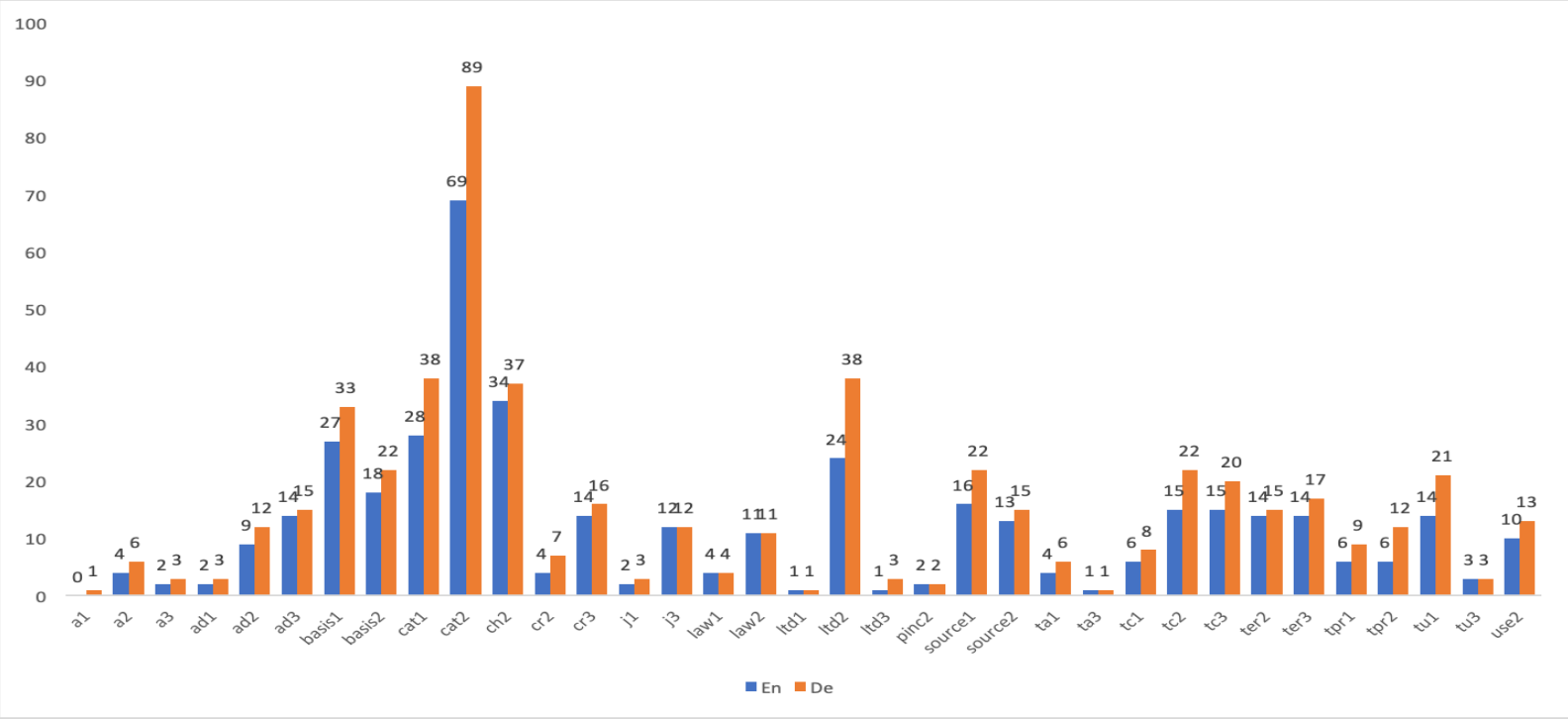

Figure 3: Comparison of the labels' distribution over the two languages.

	Total	Garmin	Box	Grindr	Linkedin	MyHeritage
Sentences	1323	87	198	503	165	370
Tags	153	16	16	54	24	43
a2	4	0	0	4	0	0
a3	2	0	0	2	0	0
ch2	34	3	7	9	7	8
cr2	4	0	0	0	1	3
cr3	14	0	0	5	3	6
j1	2	0	0	2	0	0
j3	12	3	2	2	3	2
law1	4	0	0	4	0	0
law2	11	2	2	3	3	1
ltd1	1	0	0	1	0	0
ltd2	24	4	1	7	1	11
ltd3	1	0	0	0	0	1
pinc2	2	0	0	1	0	1
ter2	14	2	1	6	1	4
ter3	14	0	2	5	3	4
use2	10	2	1	3	2	2

	Total	Garmin	Box	Grindr	Linkedin	MyHeritage
Sentences	1481	102	235	529	186	429
Tags	189	20	23	74	27	45
a1	1	0	0	1	0	0
a2	6	0	0	6	0	0
a3	3	0	0	3	0	0
ch2	37	3	8	11	7	8
cr2	7	2	1	0	1	3
cr3	16	0	0	7	3	6
j1	3	0	0	3	0	0
j3	12	3	2	3	2	2
law1	4	0	0	4	0	0
law2	11	2	2	4	2	1
ltd1	1	0	0	1	0	0
ltd2	38	6	3	11	6	12
ltd3	3	0	0	2	0	1
pinc2	2	0	0	1	0	1
ter2	15	2	2	5	2	4
ter3	17	0	2	8	3	4
use2	13	2	3	4	1	3

Table 4: Composition of ToS documents in English(left) and German(right)

	Total	**Dropbox**	**Facebook**	**Supercell**	**Tumblr**	**Twitter**
Sentences	768	122	177	66	228	175
Tags	266	33	57	30	68	78
ad1	2	0	0	0	0	2
ad2	9	2	0	3	4	0
ad3	14	0	3	2	2	7
basis1	27	3	3	9	4	8
basis2	18	2	0	2	9	5
cat1	28	2	7	2	6	11
cat2	69	12	16	5	15	21
source1	16	3	3	0	5	5
source2	13	0	4	2	1	6
ta1	4	0	2	1	0	1
ta3	1	0	1	0	0	0
tc1	6	2	2	0	1	1
tc2	15	2	7	1	0	5
tc3	15	1	5	1	6	2
tpr1	6	2	1	1	0	2
tpr2	6	0	2	0	3	1
tu1	14	2	0	0	11	1
tu3	3	0	1	1	1	0

	Total	**Dropbox**	**Facebook**	**Supercell**	**Tumblr**	**Twitter**
Sentences	855	137	216	68	253	181
Tags	351	48	84	34	97	88
ad1	3	0	0	0	0	3
ad2	12	2	0	4	5	1
ad3	15	0	4	2	2	7
basis1	33	4	3	10	8	8
basis2	22	3	1	2	11	5
cat1	38	3	10	2	9	14
cat2	89	17	25	6	19	22
source1	22	3	6	1	6	6
source2	15	0	6	2	1	6
ta1	6	1	2	1	1	1
ta3	1	0	1	0	0	0
tc1	8	2	3	0	2	1
tc2	22	2	12	1	0	7
tc3	20	2	6	1	9	2
tpr1	9	4	1	1	1	2
tpr2	12	0	3	0	7	2
tu1	21	5	0	0	15	1
tu3	3	0	1	1	1	0

Table 5: Composition of PP documents in English(left) and German(right)

As it is noticeable from figure 3, despite the content of the documents should be the same, the labels are very differently distributed when the language changes. This behaviour is even more highlighted in tables 4 and 5, where the distribution of the labels has a focus on single documents, showing big differences even with regard to the same exact contract.

7.2 Dataset 2

The second dataset is a collection of 25 ToS, listed in 6 which we'll use to validate the results obtained on the first dataset.

Document
Booking
Dropbox
Electronic Arts
Evernote
Facebook
Garmin
Google
Linkedin
Grindr
Mozilla
Pinterest
Quora
Ryanair
Skype
Skyscanner
Spotify
Snap
Terravision
Tinder
Tripadvisor
Tumblr
Uber
Weebly
Yelp
Zynga

Table 6: Source of documents for the dataset.

Differently from the first dataset the number of tags is higher, since there was

an addition of almost 50 tags, for a total amount of 100. The languages in the corpus are 4: Italian, English, German and Polish and each of the 25 document is present in both the original language and the automatically translate(to English) one.

The main differences between the two datasets are:

- the number and the nature of documents.
- The tagset.
- The addition of different languages other than English and German.
- The lack of the division in ToS and PP.

Once again, the language difference of the contracts makes the labels' distribution vary a lot.

Language	Sentences	Tags	a1	a2	a3	ch2	ch3	cr2	cr3	j1	j3	law1	law2	ltd1	ltd2	ltd3	pinc2	ter2	ter3	use2
English	6831	753	3	29	4	100	1	27	24	15	49	16	39	17	229	1	21	71	49	58
German	5911	707	3	22	3	98	1	25	23	14	46	18	33	19	212	1	17	69	49	54
Italian	6295	739	4	29	4	103	0	28	24	15	48	16	36	16	216	1	20	71	50	58
Polish	6892	771	4	35	4	103	0	26	26	15	50	19	36	17	229	1	21	75	49	61

Table 7: Composition of the whole dataset.

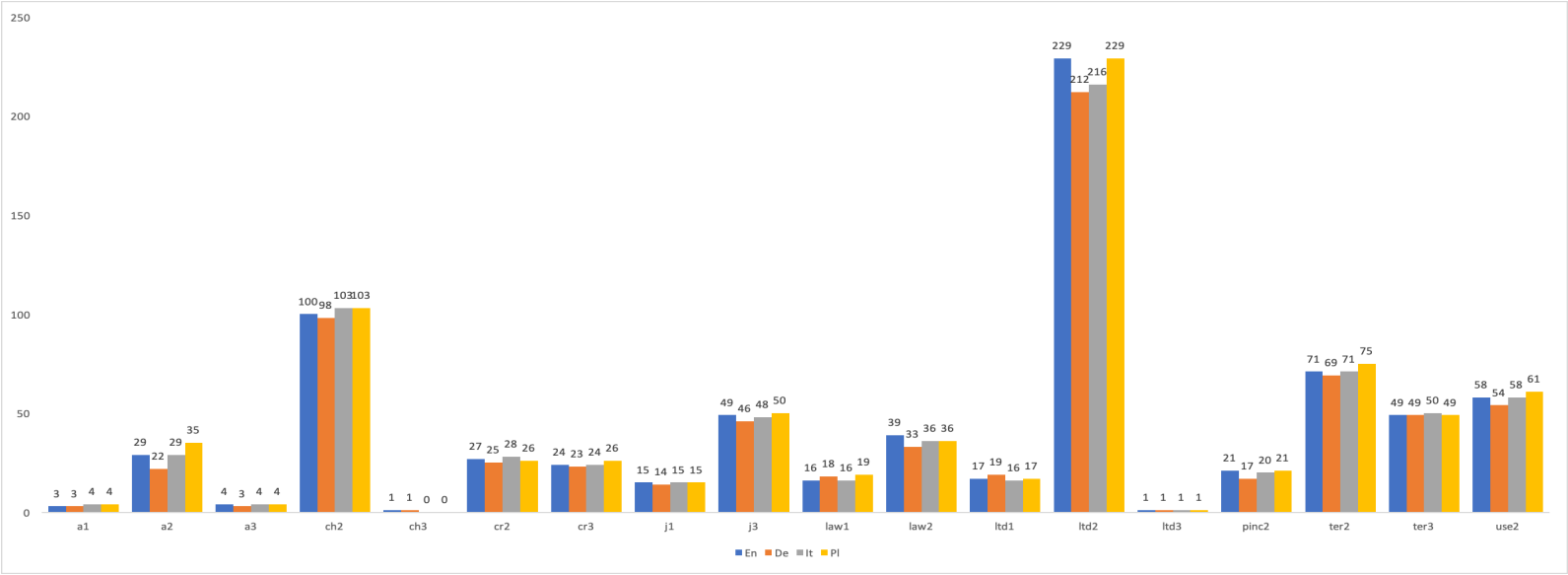

Figure 4: Comparison of the labels' distribution over the four languages.

One thing to notice from figure 4 and 3, is that the distributions are much more homogeneous in this second dataset rather than in the first. No patterns emerge from these data, for each class there is a different combination of most similar languages in terms of distribution of labels. For instance, the label ltd2, which is the most common, is equally present both in English and in Polish, while the second most used, ch2, is equally counted in Polish and Italian. The label ch3, is present just in English and in German, meaning that the clause is either safe in the other languages or presents a different kind of unfairness. In this specific case the clause in English is reported as: *"Skype reserves the right to remove or amend the available payment methods at its sole discretion."*, while in Italian is: *"Skype si riserva il diritto di rimuovere o modificare i metodi di pagamento disponibili a suo insindacabile giudizio."* and it is labelled as ch2. The translation phase, in this situation brought to a different level of unfairness in the clause.

8 Embedding techniques

Word embeddings are a way to associate a numerical vector to each word in a corpus, typically computed through sub-symbolic techniques. Usually, these embeddings are learned through a computationally demanding training process based on a very large corpus. Such learned representations embed many different aspects of the entity they represent, that can be used as features for other tasks. Aditionally, pre-trained embeddings yield a lightweight computational footprint, which makes them particularly suitable when the available computational resources are limited.

All the embeddings techniques used for experiments are based on high level pre-trained neural network. All the methods are contextual, meaning that the embedding of the single word is dependent on the other words used in the sentence and not only by the word itself. All but the last architecture are monolingual oriented, meaning that the two sentences are required to be in the same language to have reliable results.

8.1 Elmo embedding with translated document

ELMo[14] is a deep contextualized word representation that models both complex characteristics of word use (e.g., syntax and semantics), and how these uses vary across linguistic contexts (i.e., to model polysemy). These word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. They can be easily added to existing models and significantly improve the state of the art across a broad range of challenging NLP problems, including question answering, textual entailment and sentiment analysis.

8.2 Bert embedding with translated documents

The Bert embedding technique is widely discussed in 2 and it is used the English **bert base uncased**[ref] as standalone to embed each sentence.

8.3 Sentence Bert embedding with translated documents

Sentence-BERT(SBERT)[44], is a modification of the pretrained BERT network that use siamese and triplet network structures to derive semantically meaningful sentence embeddings. This particular architecture has been proved very efficient for tasks like similarity computation. For the purposes of our tests we use the **paraphrase-mpnet-base-v2** pre-trained model.

8.4 Multilingual embedding with original documents

This last model is the most interesting, because it would allow us to work directly with original documents, without the need for translation. The training is based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence, since the semantic content of a sentence should not change when it is translated. The creators of the model used original monolingual models to generate sentence embeddings for the source language and then trained a new system on translated sentences to mimic the original model. Compared to other methods for training multilingual sentence embeddings, this approach has several advantages: It is easy to extend existing models with relatively few samples to new languages, it is easier to ensure desired properties for the vector space, and the hardware requirements for training are lower. The models used in these experiments is the **paraphrase-xlm-r-multilingual-v1**, which has been trained on over 50 languages.

9 Experiments and results

In this section are reported the experimental results of the use of the different embedding/projection techniques described in section 8. The tests are divided into two phases, the former concerns experiments on the first dataset introduced in 7. The latter is performed on the second as a validation of the best techniques found so far.
All the documents treated in the first part are translated from German to English for the projection part. To evaluate the matches we used the Bray-Curtis distance by itself and then integrated in the DTW algorithm, both the procedures are explained in 6.

9.1 Experiments on Dataset 1

For the sake of these experiments, we'll be showing the performances of all the techniques described in 8. The first tests we are going to investigate are the ones on the two subsets separately, then we'll evaluate the results on the whole corpus.

9.1.1 Performances on the ToS subset

In table 8 are shown the results of the different embedding techniques. As it is noticeable, the multilingual method outperforms all the others, ELMo embedding reaches good scores while the basic BERT seems not to be adapt for the task, especially without the use of the dtw algorithm.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	0.70	0.76	0.75	0.81	0.71
paraphrase-mpnet-base-v2	0.66	0.77	0.76	0.81	0.75
paraphrase-xlm-r-multilingual-v1	**0.74**	**0.80**	**0.79**	**0.84**	**0.77**
bert-en-uncased	0.16	0.20	0.20	0.21	0.19

Table 8: Results of the different embedding techniques on the ToS subset.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	**0.75**	0.82	**0.82**	0.87	**0.78**
paraphrase-mpnet-base-v2	**0.75**	0.82	**0.82**	0.87	**0.78**
paraphrase-xlm-r-multilingual-v1	**0.75**	**0.83**	**0.82**	**0.88**	**0.78**
bert-en-uncased	0.43	0.49	0.49	0.52	0.46

Table 9: Results of the different embedding techniques on the ToS subset, with the application of the DTW algorithm.

As it is clear from the tables, all the methods improves significantly with the use of the dtw algorithm, the paraphrase-xlm-r-multilingual-v1, though, has a very slight improvement, while it is astonishing in the case of the paraphrase-mpnet-base-v2 and bert-en-uncased. Below, in table 10 are reported the performances of the best model, the paraphrase-xlm-r-multiningual-v1 model on the single labels.

	precision	recall	f1-score	support
a1	0.00	0.00	0.00	1
a2	1.00	0.50	0.67	6
a3	1.00	0.67	0.80	3
ch2	0.94	0.89	0.92	37
cr2	0.80	0.57	0.67	7
cr3	0.70	0.88	0.78	16
j1	1.00	0.67	0.80	3
j3	0.77	0.83	0.80	12
law1	1.00	1.00	1.00	4
law2	0.71	0.91	0.80	11
ltd1	1.00	1.00	1.00	1
ltd2	0.88	0.61	0.72	38
ltd3	1.00	0.33	0.50	3
pinc2	0.50	1.00	0.67	2
ter2	0.87	0.87	0.87	15
ter3	0.82	0.82	0.82	17
use2	0.82	0.69	0.75	13

	precision	recall	f1-score	support
a1	0.00	0.00	0.00	1
a2	1.00	0.67	0.80	6
a3	1.00	0.67	0.80	3
ch2	0.94	0.89	0.92	37
cr2	1.00	0.57	0.73	7
cr3	0.93	0.88	0.90	16
j1	1.00	0.67	0.80	3
j3	0.69	0.92	0.79	12
law1	1.00	1.00	1.00	4
law2	0.62	0.91	0.74	11
ltd1	1.00	1.00	1.00	1
ltd2	0.89	0.63	0.74	38
ltd3	1.00	0.33	0.50	3
pinc2	0.50	0.50	0.50	2
ter2	0.93	0.87	0.90	15
ter3	0.94	0.88	0.91	17
use2	0.90	0.69	0.78	13

Table 10: Performances on single labels of the paraphrase-xlm-r-multilingual, to the right with the use of DTW algorithm.

9.1.2 Performances on the PP subset

On this subset the scores obtained are way higher than the one in the ToS subset. Differently from the ToS case, the best model without the use of the dtw algorithm is the ELMo, while the paraphrase obtain an higher f1-weighted score. On the other hand, with the use of the dtw, the situation changes and ELMo, paraphrase-mpnet-base-v2 and paraphrase-xlm-r-multilingual-v1 have pretty much the same performances. Because of that, there are reported the score on single labels, with and without the use of dtw, of both the paraphrase-xlm-r-multilingual-v1 and ELMo embedding technique.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	**0.84**	**0.83**	0.83	0.89	0.77
paraphrase-mpnet-base-v2	0.79	0.82	0.82	0.87	0.78
paraphrase-xlm-r-multilingual-v1	0.82	0.82	**0.85**	**0.92**	**0.80**
bert-en-uncased	0.20	0.23	0.23	0.26	0.21

Table 11: Results of the different embedding techniques on the PP subset.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	**0.88**	**0.88**	**0.88**	**0.96**	0.81
paraphrase-mpnet-base-v2	**0.88**	**0.88**	**0.88**	0.94	**0.82**
paraphrase-xlm-r-multilingual-v1	**0.88**	**0.88**	**0.88**	**0.96**	0.81
bert-en-uncased	0.76	0.79	0.79	0.86	0.73

Table 12: Results of the different embedding techniques on the PP subset, with the application of the DTW algorithm.

	precision	recall	f1-score	support
ad1	1.00	0.67	0.80	3
ad2	0.89	0.67	0.76	12
ad3	0.83	1.00	0.91	15
basis1	1.00	0.85	0.92	33
basis2	0.85	0.77	0.81	22
cat1	0.91	0.76	0.83	38
cat2	0.87	0.75	0.81	89
ource1	0.84	0.73	0.78	22
ource2	0.93	0.87	0.90	15
ta1	1.00	0.67	0.80	6
ta3	1.00	1.00	1.00	1
tc1	0.78	0.88	0.82	8
tc2	0.94	0.77	0.85	22
tc3	0.93	0.70	0.80	20
tpr1	0.78	0.78	0.78	9
tpr2	1.00	0.67	0.80	12
tu1	0.94	0.76	0.84	21
tu3	0.75	1.00	0.86	3

	precision	recall	f1-score	support
ad1	1.00	0.67	0.80	3
ad2	1.00	0.75	0.86	12
ad3	1.00	1.00	1.00	15
basis1	1.00	0.85	0.92	33
basis2	1.00	0.86	0.93	22
cat1	0.91	0.76	0.83	38
cat2	0.93	0.78	0.85	89
ource1	1.00	0.73	0.84	22
ource2	1.00	0.87	0.93	15
ta1	1.00	0.67	0.80	6
ta3	0.50	1.00	0.67	1
tc1	1.00	0.88	0.93	8
tc2	0.90	0.86	0.88	22
tc3	0.94	0.85	0.89	20
tpr1	1.00	0.78	0.88	9
tpr2	1.00	0.67	0.80	12
tu1	1.00	0.95	0.98	21
tu3	1.00	1.00	1.00	3

Table 13: Performances on single labels of ELMo, to the right with the use of DTW algorithm.

	precision	recall	f1-score	support
ad1	1.00	0.67	0.80	3
ad2	0.90	0.75	0.82	12
ad3	1.00	0.93	0.97	15
basis1	1.00	0.82	0.90	33
basis2	0.86	0.86	0.86	22
cat1	0.91	0.76	0.83	38
cat2	0.90	0.78	0.83	89
ource1	0.89	0.73	0.80	22
ource2	0.93	0.87	0.90	15
ta1	1.00	0.67	0.80	6
ta3	0.00	0.00	0.00	1
tc1	0.88	0.88	0.88	8
tc2	1.00	0.82	0.90	22
tc3	1.00	0.80	0.89	20
tpr1	0.88	0.78	0.82	9
tpr2	1.00	0.67	0.80	12
tu1	0.86	0.90	0.88	21
tu3	1.00	1.00	1.00	3

	precision	recall	f1-score	support
ad1	1.00	0.67	0.80	3
ad2	1.00	0.75	0.86	12
ad3	1.00	1.00	1.00	15
basis1	1.00	0.85	0.92	33
basis2	1.00	0.86	0.93	22
cat1	0.91	0.76	0.83	38
cat2	0.95	0.78	0.85	89
ource1	1.00	0.73	0.84	22
ource2	1.00	0.87	0.93	15
ta1	1.00	0.67	0.80	6
ta3	0.50	1.00	0.67	1
tc1	1.00	0.88	0.93	8
tc2	0.90	0.86	0.88	22
tc3	0.94	0.85	0.89	20
tpr1	1.00	0.78	0.88	9
tpr2	1.00	0.67	0.80	12
tu1	1.00	0.95	0.98	21
tu3	1.00	1.00	1.00	3

Table 14: Performances on single labels of the paraphrase-xlm-r-multilingual, to the right with the use of DTW algorithm.

9.1.3 Performances on the whole corpus

To conclude with the first batch of experiments, we evaluate the performances on the whole dataset. Once again the paraphrase-xlm-r-multilingual-v1 shows to be most solid technique, since it performs well under every circumstances. One thing to notice is the incredible difference in the performances of the bert-en-uncased with and without the use of the dtw algorithm, as shown in Fig.5. Of couse, the model still performs pretty poorly compared to all the others, but it's the one that benefits the most from the application of the dtw technique.

	f1-macro	f1-micro	f1-weighted	Precision	Recall
ELMo	0.77	0.81	0.80	0.86	0.75
paraphrase-mpnet-base-v2	0.73	0.81	0.80	0.85	0.77
paraphrase-xlm-r-multilingual-v1	**0.78**	**0.84**	**0.83**	**0.89**	**0.79**
bert-en-uncased	0.18	0.22	0.22	0.24	0.20

Table 15: Results of the different embedding techniques on the whole dataset.

	f1-macro	f1-micro	f1-weighted	Precision	Recall
ELMo	**0.82**	**0.86**	**0.86**	**0.93**	0.80
paraphrase-mpnet-base-v2	**0.82**	**0.86**	**0.86**	0.92	**0.81**
paraphrase-xlm-r-multilingual-v1	**0.82**	**0.86**	**0.86**	**0.93**	0.80
bert-en-uncased	0.60	0.68	0.68	0.74	0.64

Table 16: Results of the different embedding techniques on the whole dataset, with the application of the DTW algorithm.

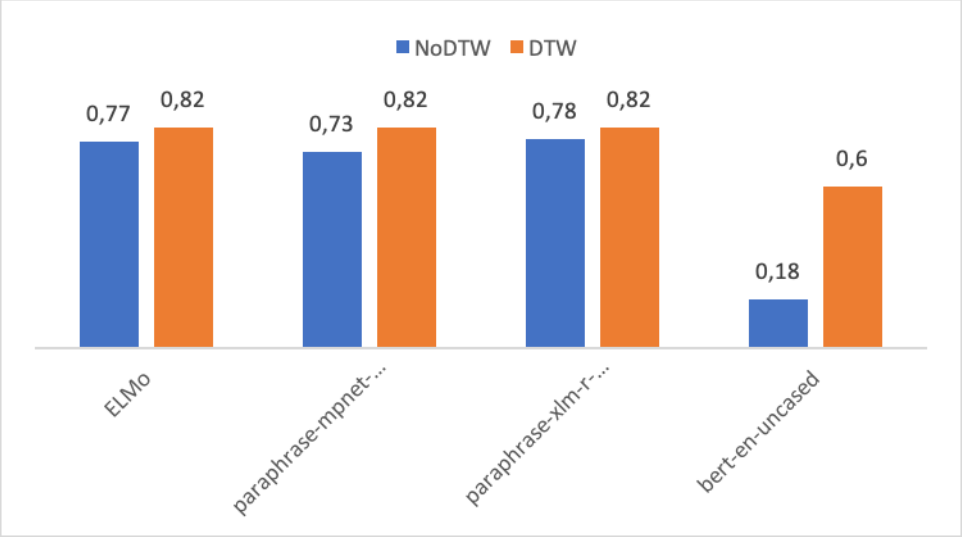

Figure 5: Values of f1-macro score of the models with and without the dtw algorithm.

	precision	recall	f1-score	support
a1	0.00	0.00	0.00	1
a2	1.00	0.67	0.80	6
a3	1.00	0.67	0.80	3
ad1	1.00	0.67	0.80	3
ad2	1.00	0.75	0.86	12
ad3	1.00	1.00	1.00	15
basis1	1.00	0.85	0.92	33
basis2	1.00	0.86	0.93	22
cat1	0.91	0.76	0.83	38
cat2	0.91	0.79	0.84	89
ch2	0.92	0.89	0.90	37
cr2	1.00	0.57	0.73	7
cr3	0.93	0.88	0.90	16
j1	1.00	0.67	0.80	3
j3	0.69	0.92	0.79	12
law1	1.00	1.00	1.00	4
law2	0.62	0.91	0.74	11
ltd1	1.00	1.00	1.00	1
ltd2	0.86	0.63	0.73	38
ltd3	1.00	0.33	0.50	3
pinc2	0.50	0.50	0.50	2
source1	0.94	0.73	0.82	22
source2	1.00	0.93	0.97	15
ta1	1.00	0.67	0.80	6
ta3	0.50	1.00	0.67	1
tc1	1.00	0.88	0.93	8
tc2	0.86	0.86	0.86	22
tc3	0.94	0.85	0.89	20
ter2	0.93	0.87	0.90	15
ter3	0.94	0.88	0.91	17
tpr1	1.00	0.78	0.88	9
tpr2	0.90	0.75	0.82	12
tu1	1.00	0.95	0.98	21
tu3	1.00	1.00	1.00	3
use2	0.90	0.69	0.78	13

	precision	recall	f1-score	support
a1	0.00	0.00	0.00	1
a2	1.00	0.67	0.80	6
a3	1.00	0.67	0.80	3
ad1	1.00	0.67	0.80	3
ad2	1.00	0.75	0.86	12
ad3	1.00	1.00	1.00	15
basis1	1.00	0.85	0.92	33
basis2	1.00	0.86	0.93	22
cat1	0.91	0.76	0.83	38
cat2	0.95	0.78	0.85	89
ch2	0.94	0.89	0.92	37
cr2	1.00	0.57	0.73	7
cr3	0.93	0.88	0.90	16
j1	1.00	0.67	0.80	3
j3	0.69	0.92	0.79	12
law1	1.00	1.00	1.00	4
law2	0.62	0.91	0.74	11
ltd1	1.00	1.00	1.00	1
ltd2	0.89	0.63	0.74	38
ltd3	1.00	0.33	0.50	3
pinc2	0.50	0.50	0.50	2
source1	1.00	0.73	0.84	22
source2	1.00	0.87	0.93	15
ta1	1.00	0.67	0.80	6
ta3	0.50	1.00	0.67	1
tc1	1.00	0.88	0.93	8
tc2	0.90	0.86	0.88	22
tc3	0.94	0.85	0.89	20
ter2	0.93	0.87	0.90	15
ter3	0.94	0.88	0.91	17
tpr1	1.00	0.78	0.88	9
tpr2	1.00	0.67	0.80	12
tu1	1.00	0.95	0.98	21
tu3	1.00	1.00	1.00	3
use2	0.90	0.69	0.78	13

Table 17: Performances on single labels of the Elmo and paraphrase-xlm-r-multilingual with the use of DTW algorithm.

9.2 Experiments on Dataset 2

As already discussed in section 7, the second dataset is used to validate the best methods emerged from the experiments on the first dataset, listed in previous section.

In light of the results presented before, the embedding techniques we decided to evaluate are two: ElMo and paraphrase-xlm-r-multilingual-v1. The former due to the widely available weights and data and its computational lightweightness. The latter, due to the best scores obtained in all the situation and, especially, due to the independence from a translation phase, since it accepts sentences in different languages as input. The languages of the documents in the datasets are 4, English, Italian, German and Polish. The experiments are conducted on the four language separately and the results are listed in tables 18, 19 and 20.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	0.93	0.94	0.94	0.92	0.96
paraphrase-xlm-r-multilingual-v1	**0.94**	0.95	0.95	0.93	0.97
ELMo DTW	0.91	0.93	0.93	0.90	0.97
paraphrase-xlm-r-multilingual-v1 DTW	**0.94**	**0.96**	**0.96**	**0.94**	**0.98**

Table 18: Results of the different embedding techniques on the whole dataset it-en.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	**0.93**	**0.91**	**0.91**	0.88	**0.95**
paraphrase-xlm-r-multilingual-v1	0.91	**0.91**	**0.91**	**0.89**	0.94
ELMo DTW	0.91	**0.91**	**0.91**	0.87	**0.95**
paraphrase-xlm-r-multilingual-v1 DTW	0.91	0.90	0.90	0.87	0.94

Table 19: Results of the different embedding techniques on the whole dataset de-en.

	f1-macro	**f1-micro**	**f1-weighted**	**Precision**	**Recall**
ELMo	0.92	0.93	0.93	0.91	0.94
paraphrase-xlm-r-multilingual-v1	0.91	0.93	0.93	0.90	0.97
ELMo DTW	**0.93**	0.93	0.94	**0.92**	0.95
paraphrase-xlm-r-multilingual-v1 DTW	0.92	**0.94**	**0.94**	0.90	**0.98**

Table 20: Results of the different embedding techniques on the whole dataset pl-en.

From tables 18, 19 and 20, we can state that for the Italian and Polish subset of documents, the paraphrase-xlm-r-multilingual-v1 model generally performs better than the ELMo in all of the use cases, confirming the trend observed in the experiments on the first dataset. For the German subset instead, the ELMo embedding without the use of the dtw obtain the best scores in terms of f1-macro, micro and weighted. One thing to notice, though, is that on this second batch of tests, the application of the dtw algorithm has a very small influence on the benchmarks, differently from the results showed in figure 5. The use of the dtw algorithm, in most of the cases, improves the recall of the models, but worsen the precision, which could result in an lower f1 score, as in Fig. 6.

Figure 6: Values of f1-macro score of the models with and without the dtw algorithm on the different projections.

The further results and experiments will be regarding just with the best of all our model in terms of f1-macro, i.e. the paraphrase-xlm-r-multilingual-v1 with the use of dtw in the projection English-Italian.

In table 21, we can see the performances of the models on the single labels. Once again it is pretty evident that the use of the dtw algorithm improves recall and worsen precision, even at a single label level. The compared scores of the f1 metric on single labels are showed in figure 7.

	precision	**recall**	**f1-score**	**support**
a1	1.00	0.75	0.86	4
a2	0.97	1.00	0.98	29
a3	0.80	1.00	0.89	4
ch2	0.95	0.94	0.95	103
cr2	0.93	0.93	0.93	28
cr3	0.92	1.00	0.96	24
j1	0.94	1.00	0.97	15
j3	0.96	0.96	0.96	48
law1	0.89	1.00	0.94	16
law2	0.95	0.97	0.96	36
ltd1	0.75	0.94	0.83	16
ltd2	0.95	0.97	0.96	216
ltd3	1.00	1.00	1.00	1
pinc2	0.90	0.95	0.93	20
ter2	0.91	0.99	0.95	71
ter3	0.98	0.96	0.97	50
use2	0.88	0.98	0.93	58

	precision	**recall**	**f1-score**	**support**
a1	1.00	0.75	0.86	4
a2	0.97	1.00	0.98	29
a3	0.80	1.00	0.89	4
ch2	0.95	0.97	0.96	103
cr2	1.00	0.96	0.98	28
cr3	0.86	1.00	0.92	24
j1	0.94	1.00	0.97	15
j3	0.94	0.98	0.96	48
law1	0.94	1.00	0.97	16
law2	0.92	1.00	0.96	36
ltd1	0.79	0.94	0.86	16
ltd2	0.97	0.99	0.98	216
ltd3	1.00	1.00	1.00	1
pinc2	0.83	0.95	0.88	20
ter2	0.99	1.00	0.99	71
ter3	0.96	0.98	0.97	50
use2	0.92	0.98	0.95	58

Table 21: Performances on single labels of the paraphrase-xlm-r-multilingual-v1 with(right) and without(left) the use of the dtw algorithm.

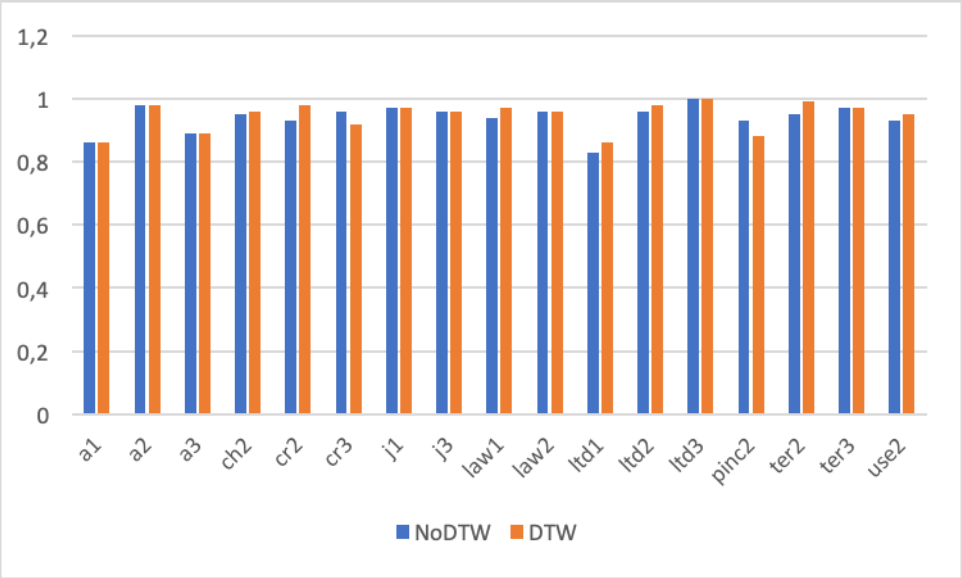

Figure 7: Values of f1-macro score of the paraphrase-xlm-r-multilingual-v1 embedding on the single labels, with and without the dtw algorithm.

9.3 Error analysis

In this section it will be analysed the errors made by the best model which resulted to be the best on the second dataset, in terms of f1-macro. The case taken into consideration is the projection from english to italian. The report on misclassification is provided by Fig.8, where for each document are listed the number of false positives and false negatives.

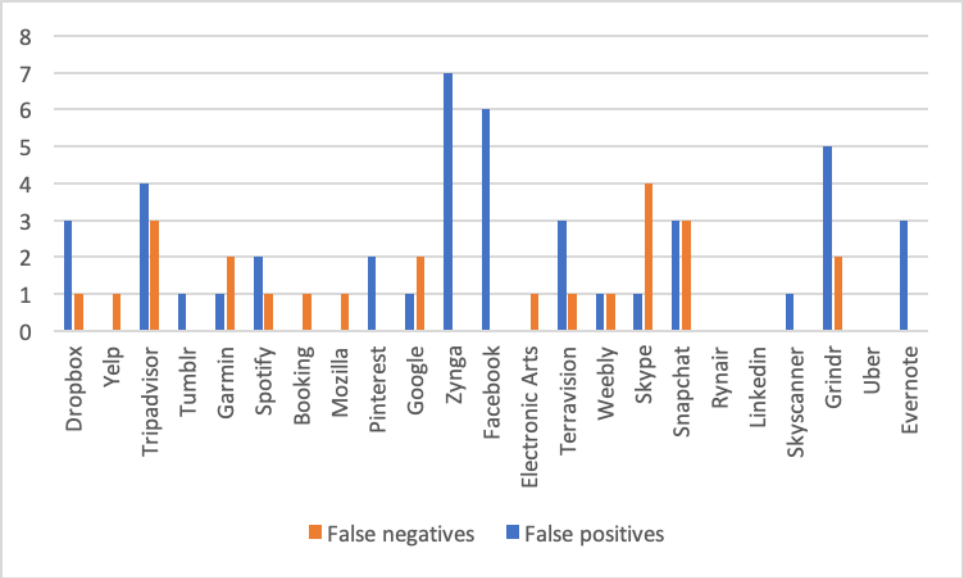

Figure 8: Distribution of false positives and false negatives during the projection en-it on the whole dataset.

After analysing the misclassification cases, it is possible to observe that most of the errors can be classified into 3 categories: Concerning the false positives, they can be of two kind:

1. False positve: incomplete or de-contextualized sentences in the target language which is linked to a complete or more explicit sentence in the source language.

2. False positive: well written terms, which in the source laguage could be bad written or ambiguous, leaving space for possible unfariness.

3. False negative: badly written term in the target language that can be unfair, while in the source language is safer and fair, even though the content of the clause is pretty similar

An example of the first category of false positives is reported in Tumblr contracts:

- *"Disputes concerning any use of or action taken using your Account by you or a third party.".*

- *"tutti collettivamente con il Sito, i " Servizi") (Tumblr, Inc.,".*

This clause is a false positive, misclassified as ltd2. The sentence by itself is incomplete and can't be dangerous de-contextualized, but it was matched with a longer sentence in the orignal english document which made far more dangerous assumption on the use of the service, such that it was labelled as unfair.

The second example, (i.e. the second case of false positives) is here presented:

- *"If you continue to use the Services after the changes are posted, you are agreeing that the changes apply to your continued use of the Services."*

- *"Se l'Utente non intende accettare dette modifiche, potrà scegliere di recedere dai Servizi ai sensi dei presenti Termini."*

In this case the term in Italian and the one in English report similar concepts, but under completely different points of view. The English clause binds the use of the service to the automatic acceptance of a change in the terms, while the Italian states that if you don't accept the term you can stop using the services accordingly to the terms of the contracts. As it is pretty clear, the source sentence is labelled as unfair, while the Italian is not, resulting in a false positive.

An example of the opposite(i.e. false negative category), the one of these two sentences matched by the algorithm:

- *"You hereby irrevocably waive, to the fullest extent permitted by law, any objection which you may now or hereafter have to the laying of the venue of any such proceeding brought in such a court and any claim that any such proceeding brought in such a court has been brought in an inconvenient forum.".*

- *"Con il presente documento, l'utente rinuncia irrevocabilmente, nella misura massima consentita dalla legge, a sollevare eventuali obiezioni che potrebbero insorgere ora o in seguito in tale sede, nonché a eventuali richieste di risarcimento derivanti da tale procedimento."*.

The sentence is classified as fair, but if analysed, in Italian, it makes very strong statement about the impossibility of objection from the user. In english the term is written in a clearer and more lawful way, which are not as ambiguous as in Italian.

10 Conclusions

The aim of this work was to investigate the best combination of techniques and tools to achieve reliable results in the field of cross language annotation projection. There were presented the metrics used to measure the similarity of the sentences and it was given an introduction on the dtw algorithm to better find matches between two parallel documents. There were introduced two corpora thoroughly annotated for unfairness detection in multiple languages, several sentence embedding techniques based on pre-trained neural architectures were presented and tested on the datasets. The focus was on the performances at corpus, document and label level with and without the use of the dtw algorithm. The results obtained are very important for several reasons. First of all, the scores obtained are very high, both in the first and in the second datasets, showing robustness of the models and of the whole method in general. These results validate a very powerful resource in the field of legal annotation projection for multiple languages, which can be applied potentially to all kinds of contracts. The steps of the algorithm don't rely on additional model to classify unfairness in clauses, freeing the process from heavy computation and making the tools easily exploitable in lots of applications. Moreover, all the models and tools described in this work are open source and accessible by anyone.

One thing to bear in mind, is that the best embedding model which emerged from the experiments is the paraphrase-xlm-r-multilingual-v1, which is a multi-language embedding model. This means that it is possible to could work directly with documents in their original language, both the target and the source, without the need to rely on translation processes of any kind, which may make the data noisy. In addition, the model is trained on 50 languages but it is easily extensible to other language by providing appropriate resources in the desired language.

The applications of this work are countless, from an integrated plugin in browsers to a standalone script. The models used for the embeddings can always be updated

and improved, thus providing for this task even better outcomes.
The results on the four languages of the second dataset are encouraging, since they're all over 90% of f1-macro and they are four very different languages, proving once again the robustness of the method. This work extended and improved the results obtained in [42], presenting a new state-of-the-art on the subject.

10.1 Future works

Future developments for this work could be tests on other languages, even changing the direction of the projection(e.g. instead of projecting from English to Italian, try projecting from Italian to English) to see if it yields better results. It is possible to test the use of other comparison method for the projection, like a cross-encoder, a pre-trained neural network computing the similarity between two sentences, thus skipping the embedding step. The next step of this research is the creation of labelled datasets for languages other than English, with the aim of creating automated unfariness detection systems like CLAUDETTE[43], ideally, for every language. This could be achieved thanks to the annotation projection technique investigated in this work, thus replacing the costly hand labelling of data.

References

[1] Stuart J. Russell and Peter Norvig. *Artificial Intelligence A Modern Approach*. 2003.

[2] Alan Turing. "Computing machinery and intelligence". In: (1950), pp. 433–460.

[3] John McCarthy et al. "A proposal for the Darthmout summer research project on Artificial Intelligence". In: (1955).

[4] Marvin Minsky. "Computation: finite and infinite machines". In: (1967).

[5] Murray Campbell, A.Joseph Hoane, and Feng-hsiung Hsu. "Deep Blue". In: *Artificial Intelligence* 134.1 (2002), pp. 57–83. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(01)00129-1. URL: https://www.sciencedirect.com/science/article/pii/S0004370201001291.

[6] Sepp Hochreiter and Jürgen Schmidhuber. "Long-short term memory". In: (1997).

[7] Yann LeCun et al. "Gradient-Based Learning Applied to Document Recognition". In: (1998).

[8] Geoffrey Hinton. "Learning multiple layers of representation". In: (2007).

[9] Jack Clark. "Why 2015 was a breakthrough year in artificial intelligence". In: (2015).

[10] High-Level Expert Group on Artificial Intelligence. *A definition of AI: Main capabilities and scientific disciplines*. 2018.

[11] Florentina T. Hristea. “Statistical Natural Language Processing”. In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1452–1453. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_82. URL: https://doi.org/10.1007/978-3-642-04898-2_82.

[12] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: (2013). arXiv: 1301.3781 [cs.CL].

[13] Jeffrey Pennington, Richard Socher, and Christopher Manning. “Glove: Global Vectors for Word Representation”. In: vol. 14. Jan. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

[14] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[15] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT (1)*. ACL, 2019, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

[16] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[17] Andrea Galassi, Marco Lippi, and Paolo Torroni. “Attention in Natural Language Processing”. In: *IEEE Transactions on Neural Networks and Learning Systems* (2020), pp. 1–18. ISSN: 2162-2388. DOI: 10.1109/tnnls.2020.3019893. URL: http://dx.doi.org/10.1109/TNNLS.2020.3019893.

[18] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. 2018. arXiv: 1801.06146 [cs.CL].

[19] Barry Schwartz. “Google: BERT now used on almost every English query”. In: (2020).

[20] Fred Kort. “Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the “Right to Counsel” Cases”. In: *American Political Science Review* 51.1 (1957), pp. 1–12. DOI: 10.2307/1951767.

[21] Haoxi Zhong et al. “Legal Judgment Prediction via Topological Learning”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3540–3549. DOI: 10.18653/v1/D18-1390. URL: https://aclanthology.org/D18-1390.

[22] Wenpeng Yin et al. *ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs*. 2018. arXiv: 1512.05193 [cs.CL].

[23] Na-Na Zhang and Yinan Xing. “Questions and Answers on Legal Texts Based on BERT-BiGRU”. In: *Journal of Physics: Conference Series* 1828 (Feb. 2021), p. 012035. DOI: 10.1088/1742-6596/1828/1/012035.

[24] Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 1st. Springer Publishing Company, Incorporated, 2017. ISBN: 3319579584.

[25] Council of European Communities. *Unfair terms in consumer contracts*. 1993.

[26] Jonathan Obar and Anne Oeldorf-Hirsch. “The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services”. In: *Information, Communication Society* 23 (July 2018), pp. 1–20. DOI: 10.1080/1369118X.2018.1486870.

[27] Yannis Bakos, Florencia Marotta-Wurgler, and David Trossen. “Does Anyone Read the Fine Print? Consumer Attention to Standard-Form Contracts”. In: *The Journal of Legal Studies* 43 (Jan. 2014), pp. 1–35. DOI: 10.1086/674424.

[28] The Department of Commerce Internet Policy Task Force. “Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework”. In: *J. Priv. Confidentiality* 3 (2011).

[29] A. M. McDonald and L. Cranor. “The Cost of Reading Privacy Policies”. In: 2009.

[30] Marco Loos and Joasia Luzak. “Wanted: a Bigger Stick. On Unfair Terms in Consumer Contracts with Online Service Providers”. In: *Journal of Consumer Policy* (Mar. 2016). DOI: 10.1007/s10603-015-9303-7.

[31] Marco Lippi et al. “Automated Detection of Unfair Clauses in Online Consumer Contracts”. In: *JURIX*. 2017.

[32] Alessandro Moschitti. “Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees”. In: Sept. 2006, pp. 318–329. ISBN: 978-3-540-45375-8. DOI: 10.1007/11871842_32.

[33] Michel Simard and Pierre Plamondon. “Bilingual Sentence Alignment: Balancing Robustness And Accuracy”. In: *Machine Translation* 13 (Nov. 1997). DOI: 10.1023/A:1008010319408.

[34] David Yarowsky, Grace Ngai, and Richard Wicentowski. “Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora”. In: (Aug. 2001). DOI: 10.3115/1072133.1072187.

[35] Steffen Eger et al. *Cross-lingual Argumentation Mining: Machine Translation (and a bit of Projection) is All You Need!* July 2018.

[36] David Bamman, Alison Babeu, and G. Crane. “Transferring structural markup across translations using multilingual alignment and projection”. In: *JCDL ’10*. 2010.

[37] Hamed Zamani, Heshaam Faili, and Azadeh Shakery. “Sentence Alignment Using Local and Global Information”. In: *Computer Speech Language* 39 (Apr. 2016). DOI: 10.1016/j.csl.2016.03.002.

[38] Kinga Klaudy and Krisztina Károly. “Implicitation in Translation: Empirical Evidence for Operational Asymmetry in Translation”. In: *Across Languages and Cultures - ACROSS LANG CULT* 6 (Apr. 2005), pp. 13–28. DOI: 10.1556/Acr.6.2005.1.2.

[39] JB Kruskal and Mark Liberman. “The symmetric time-warping problem: From continuous to discrete”. In: *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison* (Jan. 1983).

[40] Matt Post, Yuan Cao, and Gaurav Kumar. “Joshua 6: A phrase-based and hierarchical statistical machine translation system”. In: *The Prague Bulletin of Mathematical Linguistics* 104 (Oct. 2015). DOI: 10.1515/pralin-2015-0009.

[41] J Roger Bray and John T Curtis. “An ordination of the upland forest communities of southern Wisconsin”. In: *Ecological monographs* 27.4 (1957), pp. 326–349.

[42] Andrea Galassi et al. “Cross-lingual Annotation Projection in Legal Texts”. In: *COLING*. 2020.

[43] Marco Lippi et al. “CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service”. In: *Artificial Intelligence and Law* 27.2 (2019), pp. 117–139.

[44] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: 1908.10084 [cs.CL].