AI for Patent Essentiality Review

Katie Atkinson and Danushka Bollegala
Department of Computer Science, University of Liverpool, UK

November 2022

1 Introduction

In the process of developing novel standards for Information and Communication Technologies (ICT), an important step is to determine whether a patent held by a company is, or might be, required to practice the concepts covered in a given ICT specification. The patents that claim inventions that are necessary to practice a particular ICT standard are called Standard Essential Patents (SEPs) [Baron and Pohlman, 2021]. Existing approaches for automatically detecting SEPs for a given specification rely on textual similarity measures such as Latent Semantic Analysis (LSA) [Landauer and Dumais, 1997, Deerwester et al., 1990]. In this report, we first give an overview of the task of manually detecting SEPs as conducted by patent lawyers and subject matter experts, in section 2. Next, we will discuss the associated challenges from the view point of the state-of-the-art (SoTA) in Artificial Intelligence (AI), in section 3. We provide a general overview of how AI has been applied to the domain of Law in section 4, including specifically the applications of AI in Patent Law. We discuss existing tools that purport to facilitate patents that are essential for a given ICT specification in section 5. We summarise in section 6 SoTA developments in the Machine Learning (ML) and Natural Language Processing (NLP) communities that can potentially address the challenges discussed in section 3. Finally, we conclude this report in section 7 by providing a set of recommendations from a technological perspective and we list requirements that must be satisfied by future solutions to the SEP detection problem such that more accurate and explainable tools can be developed.

2 Overview of Manual Task of Patent Essentiality Review

According to the World Intellectual Property Organisation (WIPO[1]), a United Nations agency, a patent is "an exclusive right granted for an invention, which is a product or a process that provides, in general, a new way of doing something, or offers a new

[1] https://www.wipo.int/portal/en/

technical solution to a problem. To get a patent, technical information about the invention must be disclosed to the public in a patent application."[2] The rights of patent owners can be enforced in a court of law and when a case is brought before a court with a claim of infringement of the owner's exclusive rights, the infringement claim is assessed according to the intellectual property legislation of the country (or region) in which the patent in question was filed and granted.

Specifically in relation to standardised Information and Communication Technologies, such as WiFi and cellular standards like 4G LTE and 5G New Radio, is the notion of a *Standard Essential Patent*, which is a patent that is infringed when one complies with a technical standard. Litigation may ensue when when essentiality of a patent is called into question with respect to a particular standard, and the device maker refuses to licence the patent in question. When determining the outcome of such cases, solicitors and judges involved in the cases are required to consider detailed, technical information that is very specific to the products for which the patent is relevant.

To take one characteristic example, consider the 2021 case of *Interdigital v. Lenovo*, from the High Court of Justice Business and Property Courts of England and Wales, March 2021. The case proceeded through a series of trials concerning five patents. The case concerned inventions in the area of 3G and 4G telecommunications technology and all patents were asserted to be standard essential patents. The claimant, InterDigital, alleged that the Lenovo group of companies did not have (and would not commit to) licence InterDigital's patents and had infringed by importing and marketing 4G devices in the UK. Lenovo admitted the actions but denied that they were an infringement and counter-claimed for a declaration that the patent was invalid. Expert witnesses were provided by each side, who contributed to detailed discussions covering a significant range of technical aspects of the technology. In the judgement of the case, these technical aspects are set out in lengthy detail. An insight into the intricacies of the case can be gleaned by considering some of the key summary aspects regarding essentiality, about which the judge states "[t]he short point is whether in an LTE network the configuration allocated by the eNB indicates which sub-carrier resource of the NCB uplink control channel is to be used by the UE for transmitting the scheduling requests." (Quote from Lenovo Trial A at 283).

Considering the arguments set out, InterDigital argued that "because the sr-PUCCH-ResourceIndex is used to calculate the value which represents the position of the resource blocks to be used in LTE, the sr-PUCCH-ResourceIndex indicates which sub-carrier resource of the NCB uplink control channel is to be used." (Quote from Lenovo Trial A at 292).

The counter position set out by Lenovo "was that since the sr-PUCCH-ResourceIndex is used for this purpose only in combination with five other parameters, it does not indicate within the meaning of claim 1." (Quote from Lenovo Trial A at 292).

From the above quotes alone, it can be seen that resolution of the essentiality issue requires intricate understanding of both the technical context around the technologies under consideration and the legal impact of different patent language in order to inform a judgement about a patent's essentiality. The conclusion of the case was that "The Patent is valid, essential to Release 8 of LTE and is infringed. InterDigital's conditional

[2]Quote taken from: https://www.wipo.int/patents/en/

application to amend the Patent falls away." (Quote from Lenovo Trial A at 298).

This example case is a characteristic one demonstrating the complexity involved in legal cases concerning patent essentiality determinations.

Looking at the topic more broadly, the European Commission (EC) published a Pilot Study for Essentiality Assessment of Standard Essential Patents in 2020 [Bekkers et al., 2020], in which it "investigate[d] the technical and institutional feasibility of a system that ensures better essentiality scrutiny for Standard Essential Patents (SEPs)" (p12). In that study, the EC found that the primary means of determining patent essentiality is through claim charts. For example, in patent pools, "[i]ndividual companies prepare claim charts [...] for their own, standard-based licensing programs" (p14). These claim charts are reviewed by "independent, specialist third parties" (p14) to determine whether or not each element of the claims is satisfied by a device performing the portions of the standard that are identified for that claim element. "The assessors [...] are usually technical engineers (both senior and supervised junior), patent attorneys, and patent lawyers" (p15). Assessing a single European patent can cost up to €5000 – €10,000 and take up to three working days (p30). With more than 7,000 patent families declared essential to the LTE standards alone, AI-tools are desired to reduce the cost and burden associated with essentiality analysis (p35). As we discuss in further detail below and as the EC confirmed in its study, AI-based systems where "essentiality assessments would be performed fully based on automated systems," are judged to be "not yet [feasible]" but it is recognised that this objective may be "possible in the (distant) future" (p106, Table 21).

We now go on to consider how AI technologies are being developed to assist with the automation of tasks related to legal matters generally, and patent-related matters specifically, covering the state-of-the-art in techniques and tools available, and challenges that are yet to be addressed.

3 Challenges from the State-of-the-Art in AI

Existing industry-leading solutions for the automatic detection of SEPs model this problem as an instance of measuring the semantic similarity between a given standard's specification and a set of patents or classifying a patent as to whether it is essential or otherwise. However, similarity and essentiality are not equivalent concepts. A patent might be essential to a standard but might not necessarily have a high similarity in terms of textual overlap. On the other hand, between two patents that are highly similar to a given standard, one could be essential while the other might not.

To visualise the relationship between similarity and essentiality further consider the Venn diagram shown in Figure 1. If essential patents to a given standard are all similar to that standard, we would be able to retrieve essential patents purely based on similarity as shown in Figure 1 (a). However, when essentiality and similarity are different concepts, we will be able to retrieve only a subset of the essential patents using similarity as shown in Figure 1 (b). This has two important implications. First, the patents that are essential but not similar to the standard would not be retrieved by a similarity-based essentiality score. Second, if we wanted to retrieve all essential patents, we must reduce the cutoff threshold for similarity, thereby retrieving a potentially large set of

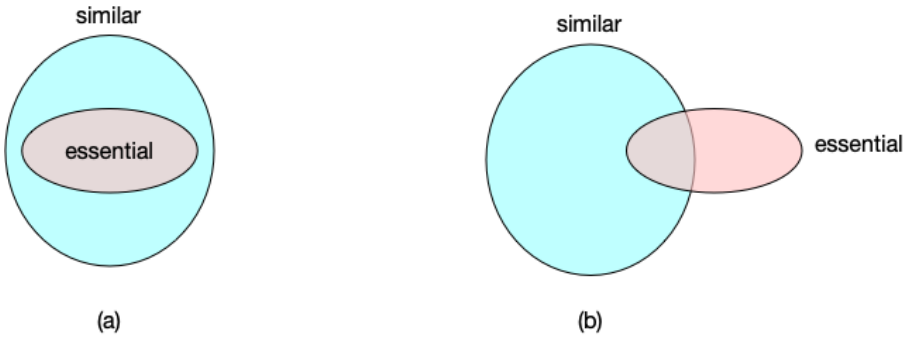

Figure 1: The relationship between *similarity* and *essentiality*. Figure 1(a) on the left is the status-quo assumption in SEP detection tools that assume essential patents can be retrieved using the similarity to a given standard. However, we argue that the relationship is more like the one shown on Figure 1(b) on the right, where there is indeed some overlap between patents similar to standard specifications and essential patents but complete subsumption does not hold.

non-essential patents in the process. This increases – rather than reduces – the manual effort of going through each retrieved patent and deciding for its essentiality.

The similarity-based approach can be classified as an *unsupervised* one because no human labelled data is required in the similarity computation. Specifically, patents are ranked in the descending order of their semantic similarity scores to a target specification. Top-ranked patents still require a human user to manually read them one-by-one and determine whether they are indeed essential.

On the other hand, the classification approach can be seen as a *supervised* one because one must first annotate a set of patents as being essential or otherwise regarding a particular standard. Although we must invest time and effort to annotate patents in the supervised approach, it enables us to build classification models that are aware of signals useful for predicting essentiality.

Although both approaches are supposed to reduce the burden of going through a large set of patents manually, and limit it to a task of reading a few top-ranked ones, which is indeed a massive time saver in theory, we identify below several challenges of these modelling and scoring methods while highlighting pitfalls that render these approaches unreliable.

3.1 Reducing Essentiality to a One-Dimensional Score is Suboptimal

A first critical problem with using semantic similarity and other current methods for essentiality analysis is that they cannot properly account for all the dimensions that a patent claim must be analysed on. For example, a target specification S can be related to different patents P_1 and P_2 for different reasons. For example, let us assume S to be a specification related to both modulation and communication channel initialisation,

whereas P_1 a patent claiming for a novel frequency-domain modulation, and P_2 a patent claiming for a novel communication channel initialisation method. P_1 and P_2 might be very different from each other because they are covering distinct technologies addressing diverse problems. S will be similar to P_1 only if we consider the modulation aspect, whereas it will be similar to P_2 if we consider the communication channel initialisation aspect. In short, S is similar to P_1 and P_2 not in all aspects but with respect to a subset of the aspects covered in the specification. However, *similarity score* is a one-dimensional (scalar) value and does *not* distinguish among the multiple aspects that are shared among documents under comparison.

An alternative approach is to model essentiality prediction as a classification problem. Alium [Alium] trained a classifier using the fastText [Bojanowski et al., 2017] embeddings for Open Radio Access Network (RAN)[3] using manually labelled patents. fastText represents a given document as a set of subword units and learns embeddings for representing those subwords. This approach addresses the out-of-vocabulary problem and enables us to learn more generalisable classification models because even when words do not overlap between training and test data, subwords would. However, as noted by the Alium developers [Alium], manually annotating patents for their essentiality with a particular standard is a laborious, costly and time-consuming task, which has low inter-annotator agreement due to the subjectivity of the task.

3.2 Insensitivity to Polysemous and Temporal Semantic Variations

Under the similarity-based detection of essentiality of a patent to a given standard, a popular approach that is used in existing tools for SEP detection is Latent Semantic Analysis (LSA). In this approach, co-occurrence matrices are built from a collection of patents and standards specifications where rows represent each document and columns represent the words occurring in the documents. The elements of these matrices are set to the number of times a particular word is occurring in a particular document. However, such co-occurrence matrices tend to be *sparse* and contain many zero-valued elements because only a handful of words of the entire vocabulary will occur in an any given document. To address this problem, LSA uses Singular Value Decomposition (SVD) to project the patents into a lower dimensional latent space determined by the left singular vectors of each co-occurrence matrix. Finally, the similarity between a patent and a standards specification is computed by the cosine similarity between the corresponding low-dimensional vectors for those two documents.

LSA represents a document using a single vector, thereby conflating multiple aspects into a single representation. On the other hand, methods have been developed in the field of NLP for representing multiple aspects of meaning, which are known as *multi prototype embeddings* [Reisinger and Mooney, 2010]. For example, the word *Apple* can mean either the fruit or the company (Apple Inc.). Its *fruit sense* is similar to *Banana*, whereas its *organisation sense* could be similar to *Microsoft*, which is also a large-scale Information Technology (IT) company. Representing both senses using the same vector (i.e. embedding) is clearly suboptimal in this case because by doing so we will be conflating two distinct meanings together.

[3] https://www.o-ran.org/

A technical term that is later used in a standard to refer to an innovation made by a particular patent might not necessarily be used in the original patent. However, subsequent patents might assign a technical term to the innovation made by a previous patent to differentiate any novel technical significance. Due to this reason, if we simply use a technical term to search for the original patent, we might not necessarily be able to find it. This is a particularly challenging problem when measuring the similarity between a patent and a given specification or standard.

Assigning novel meanings to existing terms poses an additional challenge when processing technical documents because semantic representations created using methods such as LSA are not sensitive to temporal semantic variations. More recently in NLP, researchers have proposed *Dynamic Word Embedding* [Hofmann et al., 2021] methods that can accurately encode temporal semantic variations of words.

3.3 Lack of Interpretability in Essentiality Scores

What does it actually mean to say that the essentiality score of a patent is 0.8 to a given specification? As stated in the Pilot Study for Essentiality Assessment of Standard Essential Patents, published by the European Commission [Bekkers et al., 2020], *Essentiality is a binary concept, but an essentiality assessment is a complex process*. This view aligns well with our point argued above that essentiality is a multi-faceted concept, which cannot be compressed into a single number. A related problem here is the difficulty of interpreting an essentiality score. Whether a particular essentiality score is high or low depends not only on the similarity between a patent and a specification, but also on what other patents and specifications we must consider, and what technical or legal attributes are of interest to us. Moreover, it is not readily clear whether the essentiality scores for different patents and specifications are comparable.

This lack of interpretability is a problem in real-world essentiality detection systems because once the patents are ranked according to their essentiality to a given specification, we must decide a cut-off point. We would like to keep the set of patents we wish to further manually inspect (i.e. true positives) to a manageable level, while not ignoring essential patents (i.e. false negatives). There is a trade-off between these two objectives. Of course, we could decide not to filter out any patents from the manual inspection, which might not be a viable option except when the number of patents is extremely small. On the other hand, setting a very high threshold on the essentiality score would significantly reduce the number of patents that we must manually inspect but also increases the risk of overlooking an essential patent, which could be a costly error. Striking a fine-balance between these two objectives likely makes the "proper" use of essentiality scoring tools too subjective to provide useful information to parties negotiating a patent license. Each side could correctly argue that the cutoffs are improperly applied. Such discussions are further complicated by the lack of interpretability of essentiality scores.

3.4 Insensitivity to the Word and Sentence Ordering in Documents

As already described in subsection 3.2, LSA uses a *bag-of-words* representation of a document where the frequency of a word in a document is used to represent that

document. For example, the text "John killed Mary" is represented by a vector [('John', 1), ('killed', 1), ('Mary', 1)], where each dimension corresponds to a word in the text and its value is set to the number of times that word occurs in the text. Although this representation is sufficient for tasks such as document classification, it does not retain the relative ordering of words within the text, which is especially problematic in legal contexts. For example, the text "Mary killed John" also contains the same set of words and will be represented by the same vector as "John killed Mary", thereby making the two indistinguishable by any subsequent machine learning components, even though the order has legal significance in determining the aggressor and the victim.

A partial solution to this problem is provided by considering continuous spans of words (aka n-grams). For example, bi-grams extracted from "John killed Mary" are *John+killed* and *killed+Mary*. Here, we use '+' to indicate the two individual words (i.e. *unigrams*) forming a bi-gram. Likewise, the set of bi-grams for "Mary killed John" contains *Mary+killed* and *killed+John*. Hence, the two sets of bi-grams are different for the two sentences, thereby preserving some of the word order information in the textual representation. However, bi-grams do not capture long-range dependencies between words in a sentence, nor do they preserve the relative ordering among the sentences in a document. Moreover, including higher-order n-grams in text representations results in high-dimensional and sparse representations because many of those bi-grams will not be repetitive.

It is often the case that novel technical terms, such as those found in a new technical standard, are created by reusing one or more existing terms, standalone or as a compound noun. For example, *Support Vector Machines* [Vapnik, 1998] refers to a specific classification algorithm in machine learning, and its name is formed by combining three words that have their own meanings: *support*, *vector* and *machine*. The compound noun inherits its meaning from its constituent components, but corresponds to a novel concept, different to the individual components. Under a bag-of-words representation, such technical terms will be split into individual unigrams and the meaning of the original term will not be preserved in that representation space.

There are more modern innovations in the NLP community such as Recurrent Neural Networks- (RNNs) and Transformer-based sentence/document representations that can accurately capture long-range dependencies among words in a sentence [Peters et al., 2018, Liu et al., 2019, Devlin et al., 2019]. Moreover, these methods represent texts using fixed-dimensional dense vectors, which do not increase in dimensionality with the length (i.e. number of words) in the text. We further discuss these modern text representation methods in section 6.

3.5 Domain Insensitivity of the Textual Representations

Text representations that are used in SEP detection must be aware of the language usage in Law, but currently technology still cannot achieve this. For example, the term *sentence* has very different meanings in Linguistics (i.e. *a string of words satisfying the grammatical rules of a language*) vs. Law (i.e. *a final judgment of guilty in a criminal case and the punishment that is imposed*). The text representations we are interested in using in SEP detection must be aware of such domain-specific usage of language. As discussed later in section 6, contextualised word representations produced by Masked

Language Models (MLMs) have reported superior performance compared to static (i.e. context-agnostic) word representations such as the ones produced by LSA on a broad range of NLP applications. However, it is important to train MLMs on the data from the domain in which they will be later used in order to obtain word/text representations that capture domain-specific word usages.

For example, bi-directional transformer (BERT) [Devlin et al., 2019] is a general purpose MLM that was originally trained and released by Google on Web texts. However, versions of BERT that were trained on scientific (Sci-BERT) [Beltagy et al., 2019] or bio-medical texts (Bio-BERT) [Lee et al., 2020], such as PubMed articles[4] or legal texts (Legal-BERT) [Chalkidis et al., 2020], have outperformed the original BERT model on applications in Medicine and Law, respectively. In evaluating SEPs, text representations must account for at least three domains: (1) patent law, (2) the broad technology of the standard (e.g., cellular or video coding standards), and (3) the sub-domains within those standard (e.g., channel estimations, partitioning procedures, etc.). Therefore, we emphasise the importance of using text representations that are learnt from multiple domains and technical fields of the standard to accurately capture the meaning of patents in the SEP detection task.

3.6 Lack of Large-Scale Training Datasets

Lack of large-scale datasets for training automated SEP detection has been pointed out as a significant challenge in the Pilot Study for Essentiality Assessment of Standard Essential Patents, published by the European Commission (EC) in 2020 [Bekkers et al., 2020], where it states "*An AI system would require a reference training set, with a sufficiently large number of assessments, both positive and negative, of a very high confidence level. Such a perfect training set does not (yet) exist*" (page 61). From a machine learning point-of-view, the problem of finding essential patents to a given standard can be seen as an instance of *binary classification*, where we must categorise patents as being relevant or irrelevant to the given standard. If we had access to a large training dataset which consists of positive training instances (i.e. patents P^+ that are annotated as being relevant to a standard S) and negative training instances (i.e. patents P^- that are annotated as being irrelevant to S), we could use it to train a binary classification model. For example, we can represent a patent by a feature vector (or an embedding) using some text representation method and concatenate it with the representation of S, represented also by a feature vector. Then we could feed a binary classification algorithm such as a neural network with a Sigmoid output layer to predict the probability of a given (patent, standard) pair being positive – i.e. the patent being relevant to the standard. Using this trained binary classifier, given a standard as the input query, we can induce a total ordering among all patents in the descending order of their relevance (i.e. the probability of forming a positive example with the given standard). This approach transforms the problem of SEP detection from one where we must measure the similarity between a given standard and a patent to one of classification, where we can rank based on relevance instead of similarity.

The above-described classification-based approach for SEP detection has several

[4] www.nlm.nih.gov/databases/download/pubmed_medline.html

attractive features. First, it enables us to focus on relevant similarities between a standard and a patent. For example, the classifier-based approach is able to assign different weights to the different words shared between a standard and a patent depending on whether that word is relevant or otherwise to discriminate positive examples from the negative ones. This is preferable to the similarity-based approach, which although unsupervised, is unable to make such distinctions, thereby treating all shared words equally. Second, we can fine-tune our classifier to reflect the notion of essentiality of a patent as held by the human experts (i.e. patent lawyers and technical experts) in the classifier-based approach by annotating (labelling) positive and negative examples to train with. The similarity-based approach does not have such tuneable parameters, which makes it difficult to customise for a particular domain.

Due to these reasons, typically in machine learning tasks where there are high quality and large-scale training data, supervised approaches outperform unsupervised approaches. However, the challenge when applying such a supervised solution to the SEP detection problem is the unavailability of large-scale training datasets to train classification models. The problem is exacerbated by the need for training datasets for each relevant subdomain within a given standard to address the problems discussed above in subsection 3.1 through to subsection 3.5. And, even with human expert based training sets, experts often disagree on the binary essentiality question, meaning that only coarse tuning may ultimately be achievable.

That said, training signals could be obtained indirectly from existing retrieval systems similar to *clickthrough* [Joachims, 2002] data in Web search engines. For example, we could record the standard queried by the users of a SEP detection system, and the patents they actually read (considered as positive examples) vs. skip over (considered as negative examples) among the results presented by the system. However, the amount of training data points and the quality of those data points depend on the accuracy of the current SEP detection systems and the number of users using those systems. For example, if the accuracy of the existing systems are low and only a handful of patents returned for a given standard is considered to be relevant by the users, this semi-automated training data aggregation process will result in a highly imbalanced training dataset with most of the examples annotated as negative. It is noteworthy however to mention that there are methods for learning accurate models of imbalanced datasets already proposed by the ML community [He and Garcia, 2009, Li et al., 2012, Wu and Chang, 2003, Provost, 2000, Cieslak and Chawla, 2008, Kubat and Matwin, 1997]. Another challenge could be that the training signal obtained via clickthrough could be very noisy and unreliable. For example, a large number of patents that are returned for a user query might be clicked by a user but they might later decide some of the clicked patents to be non-essential. On the other hand, if we can obtain behavioural data such as dwell time on the patents by individual users, that could provide a more reliable training dataset, but even this approach would require to take into consideration that different experts may perform these tasks at different speeds in the abstract and in the sub-domain of their individual expertise. Thus, it remains an open question whether such data could replace the need to manually annotate patents for essentiality as required to train supervised methods.

3.7 Lack of Well-Defined Success Criteria

The exact definition of what is an *essential* patent for a given standard remains a subjective one. Two different patent lawyers often disagree on the same set of patents being essential, as evidenced by litigation. In the ML community, this is known as *inter-annotator agreement* computed by a set of annotators annotating the same set of examples. There are established measures such as the Cohen's κ for measuring the inter-annotator agreement for subjective annotating tasks such as annotating for the sentiment expressed in a customer review of a product, for the purpose of building sentiment classifiers. However, to the best knowledge of the authors of this article, we are not aware of such inter-annotator agreement measurements for SEP detection. Inter-annotator agreement provides an estimate on the upper bound of the performance an ML method can hope to obtain because even two humans would still have a certain level of disagreement between themselves for the same task.

Given this backdrop, it is questionable what is a useful success criterion or a metric for evaluating a SEP detection system. The focus so far in this document has been to maximise *relevance* of a retrieval system that returns patents based on their essentiality score against a given input standard. However, if the *determination* of essentiality remains ambiguous and subjective, it begs the question of the validity of any relevance metrics based on essentiality. On the other hand, from the end-user's point-of-view there could be multiple valid definitions of *success* such as (a) was the system able to return a patent that a human expert (e.g. a patent lawyer or a technical expert familiar with the technical domain) would consider to be important or potentially essential to the standard being investigated?, (b) did the system miss any relevant patents to the standard being queried?, (c) what was the saving in time/effort/cost provided by using the system as opposed to manually reading all patents?, and (d) if the system returned irrelevant patents what was the extra time/effort/cost incurred by the end users in reading through and deciding (the patents were indeed irrelevant) by themselves? These are all important yet open questions that require answers in order to define success criteria that are not only valid as an optimisation function for ML models, but also meaningful for the end-user.

4 Overview of the Field of AI and Law

Whether a patent is standard essential presents a legal question. There is a longstanding community of academics who have been working specifically on development of research in AI for legal matters, with an early formal gathering being the First International Conference on Artificial Intelligence and Law (ICAIL) that was held in Boston, USA in 1987, establishing an event that is held every two years and hosted at different international institutions. The community's journal[5] started up in 1992 and since then has been publishing the latest research results of ongoing research covering a wide range of topics that span the field of AI and Law.

The list of focus topics being investigated within the community has evolved over time, reflecting developments in the wider field of artificial intelligence. Still, there are

[5] https://www.springer.com/journal/10506

a number of topics that have been staple features of interest and for which effective techniques have matured over time. In fact, looking back at the titles of papers that appeared in the first edition of the ICAIL conference proceedings, numerous concerns are still relevant for modern AI, demonstrating the importance of comprehensively accounting for legal theory and practice within the development of AI technologies for supporting legal work. For instance, the 1987 proceedings contain papers covering topics of modelling legal data, conceptual information retrieval, case based reasoning and explainable decision support systems. Of course, there have been significant leaps forward with the development of techniques to tackle these topics as the fundamental research bears fruit, computer hardware becomes more sophisticated, and applications of the research become viable. Below is a set of summary highlights of key topics of study within the field of AI and Law (though the list is intended to serve as an exemplar of highlights rather than a comprehensive review).

4.1 Ontologies and Legal Knowledge Representation

In computer science terms, an *ontology* is a "conceptualisation of a domain." Within the general field of AI there has been a significant volume of research conducted on the development of techniques to represent taxonomies and structured relations about a particular domain of knowledge. Many such taxonomies have been defined manually – two simple examples could be the animal kingdom and the human body – but within AI, the challenge is to represent the relations between domain concepts and define techniques to automate reasoning about these. For example, in the animal kingdom taxonomy, when given the features of a particular type of animal, it should be possible to determine whether or not it is a mammal being described and if so what particular category or individual type. There are mature techniques available for ontology engineering and extensive applications of the use of ontologies. A particular domain for which ontologies have seen widespread application and deployment is the medical domain; SNOMED-CT (Systematized Nomenclature of Medicine Clinical Terms)[6] is a visible and mature example of such an application in widespread use.

Legal knowledge, in different spheres, has similarly been captured in ontologies for AI applications in scoped settings under specific jurisdictions. Such ontologies are often built to capture knowledge that is needed for automation of reasoning tasks when that task needs to be repeated frequently, drawing on that domain knowledge. With the proliferation of ontologies for specific tasks in specific jurisdictions, this raises questions as to how ontologies can interface with one another, be merged or be identified as specialised instances of more general concepts relevant to or across domains.

For an overview of the range and development of legal ontologies over time, see the study reported in [de Oliveira Rodrigues et al., 2019]. The study provides a categorisation of legal ontologies according to their purpose, level of generality and underlying legal theories, with a view to identifying legal ontologies that can be reused and assist with practical applications in law.

[6] https://www.snomed.org/

4.2 Argumentation for Legal Reasoning

A sub-field of AI that has been flourishing since the 1990s is the community working on the topic of *computational models of argument*. Within this field, researchers work on the development of computational techniques that emulate how humans exchange arguments within a debate where information brought into the discussion may be incomplete and/or inconsistent. Arguments and counter-arguments are exchanged to advance a position within a debate, then the arguments can be evaluated to determine which are the winning arguments, and why. Computational techniques have been developed to enable arguments to be represented and reasoned about by automated software. These techniques can be applied in a wide range of real world domains[7] and the legal domain is a natural one for application given the everyday use of (human) argumentation within a variety of legal settings and tasks.

A significant body of research has been developed over recent decades to build models of how legal cases are argued (in common law jurisdictions). A key early line of this research was the HYPO system [Rissland and Ashley, 1987], and later the CATO system [Aleven, 1997] [Ashley, 1990], in which an argumentation model was developed to represent legal domains as abstract factors that apply across sets of cases whereby the factors favour one of the parties in the dispute. The model was developed into a tool to support law students in identifying arguments to present within case disputes. More recently, the ANGELIC methodology [Al-Abdulkarim et al., 2016] was developed using recent advances from the field of computational models of argument to capture legal domains as a knowledge base from which arguments can be generated to automatically decide, given the facts of a case, which legal issues and factors can be accepted and thus decide the case. The formal models have been transformed into an implemented tool and applied in a variety of settings, including a real world setting supplied by a law firm, as reported in [Al-Abdulkarim et al., 2019], where the tool was applied to cases considering claims of individuals' hearing loss attributed to negligence on the part of an employer. The methodology is thus used to build a model of each scoped legal domain to which it is applied, in order to capture the relevant factors for that domain.

A crucial feature of the argumentation approaches to modelling legal reasoning is that they can easily provide *explanations* of the automated reasoning that has been carried out by the software that implements the models. Argumentation-based approaches explicitly encode both legal concepts and legal reasoning within the models produced, thus going beyond capture of data points only. The drawback of capturing fine-grained legal expertise within an AI model is that this can be resource intensive, leading to what is known as *the knowledge acquisition bottleneck*, whereby domain knowledge has to be identified and captured meticulously within the model. However, for scoped domains where an argumentation-based decision-support tool can be shown to consistently speed up processing of cases over a period of time, the initial investment made in the knowledge acquisition task is not necessarily prohibitive, as demonstrated through recent exercises [Al-Abdulkarim et al., 2019].

[7] See [Atkinson et al., 2017] for a high level overview of the field of computational models of argument and tools arising from it, plus applications in legal, medical, and e-government domains.

4.3 Automated Legal Information Extraction and Natural Language Processing

Recent advances in general AI techniques for information extraction are also being developed for legal AI models to determine case outcomes, similar to the developments discussed in subsection 4.2.

Looking first at academic research in this area, a characteristic example tackling the issue of how to use AI methods to move through a natural language description of a case to get an outcome decision is the SMILE system [Ashley and Brüninghaus, 2009], developed within the CATO line of work. The key task that the SMILE system was developed for was identifying factors in a case, within a textual description, to determine whether a factor in the defined set for a specific domain is present or absent in the particular case. To achieve this task, a mixture of information extraction and machine learning techniques were deployed in the SMILE system. The output factors extracted from the source texts are then able to be used in an Issue-Based Prediction (IBP) system [Bruninghaus and Ashley, 2003], which delivers a prediction on the outcome of the particular case under consideration. Experiments showed that IBP had a 91% accuracy in its case prediction.

Looking beyond a specific example, legal information extraction can be conducted for a wide variety of purposes. In recognition of the growing volume of research being carried out on this topic, a number of projects and workshop series have sprung up in recent years to gather together researchers interested in this sub-topic of AI and Law.

In 2015 a workshop series and interest group was set up on the topic of "Automated Semantic Analysis of Information in Legal Text" (ASAIL). The aim of the group is to "serves as a platform for researchers and practitioners working on natural language processing of legal text."[8]. The workshop has run every two years since 2015 and has showcased work on the application of natural language processing and machine learning to the semantic analysis of legal texts. The focus on semantic analysis recognises a shift towards automated processing of *meanings* of textual elements in the texts within a given legal domain. The most recent editions of the workshop have featured papers demonstrating increasing sophistication of semantic techniques for automatically analysing legal texts, for example, evidence extraction from court judgments, annotation of texts to identify legislative content and identification of persuasive features in legal texts. However, there is not yet a suite of highly mature techniques available for widespread deployment in practice across all these tasks.

4.4 Machine Learning for Legal Tasks

In recent years there have been many academic papers published reporting on the application of mature machine learning algorithms to processing of various legal tasks. A task that has been the focus of a significant proportion of this work is predicting outcomes of legal cases.

A notable paper that received attention at the time it was published (2016) is [Aletras et al., 2016]. The work reported is a machine learning-based binary classification

[8] https://sites.google.com/view/asail/asail-home

task whereby natural language processing is undertaken to extract textual content from a legal case and provide this to the classifier to predict whether the case constitutes or not a violation of an article of the European Convention on Human Rights (ECHR). Experiments reported in the study demonstrate a success rate of 79% on average in the classifier producing a correct prediction (violation or no violation). However, as discussed in [Atkinson et al., 2020], the outcomes provided by such classifiers are not accompanied by suitable explanations that provide reasons for the prediction that are grounded in legal terms. In [Aletras et al., 2016], the authors provide a list of the 20 most frequent words from the dataset that give weight to the prediction outcome produced. Examples from a list given for Article 6 cases include words such as "court", "case", "January", "human". As well as being a mixture of words that might be expected to appear across cases, the list also contains more surprising words such as months of the year. Furthermore, these words are not built into reasoned explanations of the outcomes, akin to what would be produced, and expected, of a human judge undertaking this task.

The ECHR domain has served as a useful testbed for other work that has investigated the application of machine learning algorithms to legal outcome prediction. To enable experimental evaluation of the efficacy of the machine learning algorithms in predicting case outcomes, past cases decided by the courts already are used to determine whether the AI-based approaches can replicate the human decisions. In the work of [Medvedeva et al., 2020], the authors consider the task of predicting decisions for future cases, based on learning from past cases. Their results demonstrate that accuracy performance of this task deteriorates, with their results having an average accuracy range from 58% to 68%. An additional line of investigation that they report on is the classification performance rate when predicting outcomes based only on the surname of the judge who hears the case. For this task, an accuracy result of 65% is achieved, but a host of ethical questions are raised about the potential use of such approaches in practice whereby judges' names are used to make predictions on case outcomes; indeed, due to these concerns, some countries such as France have banned the use of 'judge analytics'[9]

The work in [Medvedeva et al., 2020] has been developed into an online tool, JURI SAYS[10], that predicts upcoming cases in the European Court of Human Rights and reports ongoing metrics on it performance, though the authors make clear that they do not advocate the use of such tools to replace human judges.

4.5 AI and Patents

Research papers have been appearing on the topic of AI and Patents for quite some time, with earlier work considering, for example, text analysis for rating the innovativeness of a patent [Hasan and Spangler, 2007], paraphrasing and summarization of patent claims [Bouayad-Agha et al., 2009] and search strategies for prior art search [Bouadjenek et al., 2015].

[9] As reported in a number of articles in the legal press such as:
https://www.artificiallawyer.com/2019/06/04/france-bans-judge-analytics-5-years-in-prison-for-rule-breakers/

[10] https://www.jurisays.com/

More recently, in 2021 a workshop[11] was organised, co-located with the International Conference on AI and Law, to bring together researchers interested in the development of new AI techniques for automating the processing of patents and tackling issues of relevance for the use of AI technologies within international patent systems. The workshop was organised in association with the Center for AI and Patent Analysis[12] and the remit of the workshop covered "Machine Learning and Natural Language Processing in patent examination, extracting meaning and information from the text of patents, evaluating patent portfolios, patent litigation analytics, patent citation analysis, and evaluating patent licenses."

As can be seen from the contributions to this workshop, there is interest in developing and deploying AI to tackle various different aspects of patent-related work while research to advance computational support for different tasks is advancing. However, it is widely acknowledged that there is a high level of complexity involved in assessment of patent matters, as discussed in section 2, so the extent to which AI is being deployed is currently limited to tasks substantially less complex than the task of evaluating a patent's standard essentiality.

Contributors to the aforementioned workshop make clear in their talks that as of 2021 there is promise for the use of AI for supporting processing of patent matters, but a high level of manual analysis is still currently required, and is likely to be so for some time, for the most challenging tasks involving interpretation and application of the law. Multiple examples were given from Patent Offices around the world (e.g. USA, UK, Japan, Australia) as to how AI is now being used to assist with different tasks within patent-related work, such as search and information retrieval; language translation; image similarity processing. It is also clear that structured data is becoming more accessible and thus is paving the way for increased analytics to support decision making. However, contributors to the workshop also emphasised how important the issues of interpretability and explainability are in examination of patents and that AI tools are not sufficiently mature to be able to perform such tasks in a fully automated and robust manner. Thus, whilst the state-of-the-art in AI for patent processing is advancing, judgement by humans is expected to remain centre stage in the domain, at least in the medium term, but with increasing support from AI tools.

More broadly, the aforementioned CMU Center for AI and Patent Analysis, provides a host of resources that includes: articulation of a long-term research agenda for developing patent-specific AI capabilities; a set of initiatives and projects aimed at fulfilling the objectives; white papers on matters related to AI patent research; standards being developed to assist with auto-generation of legal documents; API development to patent analytics software; metrics and benchmarks for AI patent search technologies. Such research centres focused specifically on AI and patents serve to demonstrate the swell of interest in the development of AI technologies for patent work and the increasing momentum for advancing the emerging innovative tools.

[11] https://www.cmu.edu/epp/patents/events/icail21/index.html

[12] https://www.cmu.edu/epp/patents/index.html

5 Tools for Patent Essentiality Review

We now briefly review five commercial tools available for patent essentiality reviews. Our analysis is based on publicly available information about those tools, which might be different than the actual proprietary implementations. There is no underpinning literature we are aware of that evaluates the efficacy of these commercial tools for the binary determination of essentiality and/or essentiality within a particular confidence level. Furthermore, as mentioned above and uniformly recognised, explainability is a core requirement for implementation of AI decision support tools. As discussed in detail in [Atkinson et al., 2020], if AI tools are to be trusted by the end users for whom they are being developed, then any conclusions drawn by the AI tool, and any decisions or recommendations made, must be accompanied by a suitable explanation. This explanation must give, in terms understandable to the human end user, reasons as to why the conclusion provided is justified in the particular case being considered, and also why other outcomes or conclusions were rejected. Qualitative explanations need to accompany quantitative or probabilistic outputs to explain how and why numeric scores have been derived by the AI tool for a particular case. Furthermore, for applications for law, explanations for automated outcomes must be grounded in legal terms to ensure that the law is applied as intended and expected. We see this notion of explainability as a key requirement yet to be captured in the tools available for conducting patent essentiality review.

5.1 IPlytics

IPlytics (https://www.iplytics.com/) is a tool supporting the essentiality reviewing process, which ranks and displays essential patents for a query issued by a user. In particular, it computes a Semantic Essentiality Score (SES) that indicates "how likely essential a patent is to the standard it is declared for", using a score that ranges from 1 to 100 "with 100 being the most likely essential."[13] Two approaches have been discussed in [Baron and Pohlman, 2021] for computing semantic essentiality scores: (a) a sampling-based approach, and (b) a supervised predictive modelling approach. They recommend sampling for the estimation of essentiality ratios in large firm portfolios of declared SEPs, whereas for smaller datasets the predictive modeling approach is reported to be more accurate. As with the other tools described below, the precise reliability and feasibility of these approaches warrants further review and investigation. At present, it is unclear how effective, for example, the Semantic Essentiality Score is at arbitrating the decision of whether a patent is essential to a standard or not.

5.2 Alium

Alium is a joint venture between MPEG LA and Unified Patents to provide an "OPEN RAN Patent Portfolio License."[14] Alium provides a tool for royalty allocation [Alium]. This tool measures the potential essentiality of granted patents to the Open RAN Standard for the purposes of royalty allocation among Licensors. They built a text classifier

[13] https://www.iplytics.com/platform/semantic-essentiality-score

[14] https://www.alium-llc.com/blog/alium-introduces-open-ran-patent-portfolio-license

for this purpose using the fastText library. They used a dataset of 6.6K manually classified patent families of self-declared in ETSI for 3GPP to train their classifier. The probability of a given patent being classified into the positive (essential) class is used for royalty allocation.

5.3 Amplified

Amplified (https://www.amplified.ai) has developed a tool that can search and annotate patents collaboratively. It is able to find similarity between terms in user queries and documents not only based on keyword matching, but also by analysing the entire content of the document. In particular, the platform provides a *classic* mode and a *neural* mode to sort the search results based on the relevance to the queries. In the classic mode, the keywords mentioned in the query are matched in the documents, whereas in the neural mode the matching is done at a semantic-level, where the entire text in a document will be matched against a query. For example, the word-order in the query will induce different rankings among search results in the neural mode.

5.4 Apex Standards

Apex Standards (https://www.apexstandards.com) provides a platform for analysing the status and relations among patent claims and technical specifications. A distinguishing feature of this platform from its competitors is the visualisation of essential patents to a given standard as a navigable knowledge graph. This may help users to view the relationships among different patents easily and provides a holistic viewpoint. Moreover, different sorting and filtering criteria can be applied on the set of matching patents for a user query.

5.5 AI Patents

AI Patents (https://www.aipatents.com/seps) has developed a tool that helps implementors and licensors to identify potential SEPs for licensing negotiatons or litigations. Their proposed solution is claimed to be generic and can identify potential SEPs from a given database or portfolio of patents. Moreover, it does not appear that training data is required in their solution, and thus it can be classified as an unsupervised approach for SEP detection.

6 Machine Learning for Essentiality Review

As we discussed in previous sections, essentiality score prediction has been modelled as a similarity measurement problem or a text classification problem in existing tools. In both similarity measurement and text classification, a fundamental task is to represent a given text by a vector (aka *embedding*) such that it could be used to train a machine learning algorithm [Mikolov et al., 2013, Pennington et al., 2014]. The bag-of-words approach and the associated n-gram variants discussed in subsection 3.4 result in high-dimensional and sparse text representations, which are problematic because

the overlap of features between train and test data can be small. This is problematic because for a machine learning model to accurately predict test instances it must have observed the features that occur in test data during the training phase. This overlap between the features in train and test data is essential for a machine learning model to generalise to unseen test data. Consequently, dimensionality reduction methods such as Singular Value Decomposition (SVD) have been used in text representation methods such as LSA/LSI to reduce the dimensionality of the feature space, and represent texts in low-dimensional dense spaces. Obtaining low-dimensional dense text representations via this approach is known as *top-down* text representation. This approach improves performance in text similarity measurement because two similar texts that did not share any n-gram features, would otherwise have returned a zero similarity score due to the feature sparseness in high-dimensional feature spaces.

An alternative approach to the top-down method is to first assign (for example randomly initialised) embeddings to all words in a vocabulary, and then update those embeddings such that some supervised training objective is optimised. Given a sentence written by a native speaker (thus assumed to be grammatically correct and meaningful), we can mask out a single word in the sentence and require a neural network to predict the masked out word using the embeddings for the remainder of the words in the sentence. This is known as masked language modelling (MLM) [Devlin et al., 2019, Liu et al., 2019] and provides a *bottom-up* approach to learning low-dimensional dense word embeddings without requiring the application of dimensionality reduction methods as a post-processing step as used in LSA. For example, given the sentence *I had bread and butter for breakfast*, we could mask out *butter* and update the embeddings of other words such that we can predict *butter* as the answer to the MASK in *I had bread and MASK for breakfast*. This approach has produced contextualised word embedding models that are sensitive to the context in which a target word occurs, which are both word-order sensitive and word-sense aware. Moreover, it can be seen as a form of *self-supervised* learning, where no explicit human annotation is required to train the word embeddings. Thus it can be scaled up to learn from large amounts of electronically available texts in different domains and languages. Consequently, contextualised embeddings obtained from large-scale MLMs such as BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019], XL-Net [Yang et al., 2019] etc., have comprehensively outperformed word embeddings produced by top-down approaches in numerous NLP tasks. However, to the best knowledge of the authors of this article, contextualised word embeddings have not yet been used in SEP detection tasks. It remains an interesting research direction to evaluate the possibility of using contextualised word embeddings in SEP detection tasks.

7 Conclusions

This paper has considered the topic of AI for patent essentiality review, starting with a conceptual view of how the legal task is currently conducted manually, then progressing on to a survey of the landscape of research developments aimed at using AI to assist with the task. The survey covered the specific challenges associated with the automatic detection of Standard Essential Patents, considering the extent to which state-of-the-art

AI techniques can be applied successfully to this task. The challenges were further contextualised in relation to broader recent developments within the field of AI and Law. We also considered the development of commercial tools that are now becoming available for essentiality reviews.

We conclude by summarising some key points from our survey:

- Recent research in AI and Law has demonstrated the viability for automated support for an increasingly wide range of legal tasks, but with strength of results being heavily dependant upon the complexity of the particular legal task.
- While the body of research on AI for patent analysis has recently been expanded, with research and tools focusing in on the task of essentiality assessments, a range of challenges remain to be addressed to ensure both precision in technical performance and end user acceptance for AI tools being built to address patent essentiality. Due to the complexity of standard essentiality determinations, current techniques and tools cannot replicate or replace human expert review.
- Reducing an essentiality 'score' to a scalar number is problematic because it conflates multiple aspects into a single value and techniques such as LSA fail to account for the legal significance of the ordering of terms.
- Similarity and essentiality are not equivalent concepts. A patent might be essential to a standard but might not necessarily have a high similarity in terms of textual overlap. On the other hand, between two highly similar patents to a given standard, one could be essential whilst the other might not be.
- As with various concepts in law, essentiality scores must take into consideration terminology that evolves over time. Simply using technical terms from a standard to evaluate essentiality is insufficient because earlier patents may describe or assign a different technical term to the functionality under consideration.
- Patent essentiality review is a complex task for humans to complete and reach agreement upon, as seen in the plethora of multi-faceted technical discussions comprising legal cases involving essentiality questions. Such tasks are prone to error and are subjective determinations, which present a challenging task for the current state-of-the-art in AI in automation exercises.
- High-quality training sets for patent essentiality review are not currently available, reflecting in part the lack of consensus among subject-matter experts as to the binary determination of essentiality. Multiple training sets would likely be required to take into account the multiple domains for which an AI-based essentiality determination may be desired given the disparate subject matter with a given standard.
- To enable full automation of the task of patent essentiality review, and realisation in tools that perform this task, requires progress in demonstration of a variety of criteria being met, including the level of overall precision, capture of crucial contextual information and explainability of quantitative assessments as expressed in essentiality scores.

AI is now enabling increasing levels of support for automation of legal tasks, but there remain many complex tasks that present challenges in reaching expert-level performance. Patent essentiality review is one such challenging task and our survey has identified the aspects to be tackled to advance capabilities of AI tools being developed for this task.

Acknowledgements

We would like to thank Talbot Hansum from Norton Rose Fulbright for bringing this interesting topic to our attention and providing insights on standard essential patents. We also acknowledge partial funding of the project from Qualcomm Inc. The authors' review and opinions on the state-of-the-art in AI for patent essentiality as expressed in this paper are entirely their own.

References

L. Al-Abdulkarim, K. Atkinson, and T. Bench-Capon. A methodology for designing systems to reason with legal cases using ADFs. *AI and Law*, 24(1):1–49, 2016.

L. Al-Abdulkarim, K. Atkinson, T. Bench-Capon, S. Whittle, R. Williams, and C. Wolfenden. Noise induced hearing loss: Building an application using the ANGELIC methodology. *Argument & Computation*, 10(1):5–22, 2019.

N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, and V. Lampos. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Comput. Sci.*, 2:e93, 2016. doi: 10.7717/peerj-cs.93. URL https://doi.org/10.7717/peerj-cs.93.

V. Aleven. *Teaching case-based argumentation through a model and examples*. Ph.d. thesis, University of Pittsburgh, 1997.

Alium. Alium open ran ip analytics training methodology. URL https://www.alium-llc.com/blog/hnybywxnf3y9wf1tngjuwif8ybp9w8.

K. D. Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, Cambridge, Mass., 1990.

K. D. Ashley and S. Brüninghaus. Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law*, 17(2):125–165, 2009. doi: 10.1007/s10506-009-9077-9. URL https://doi.org/10.1007/s10506-009-9077-9.

K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, and S. Villata. Towards artificial argumentation. *AI Mag.*, 38(3):25–36, 2017. doi: 10.1609/aimag.v38i3.2704. URL https://doi.org/10.1609/aimag.v38i3.2704.

K. Atkinson, T. J. M. Bench-Capon, and D. Bollegala. Explanation in AI and law: Past, present and future. *Artif. Intell.*, 289:103387, 2020. doi: 10.1016/j.artint.2020.103387. URL https://doi.org/10.1016/j.artint.2020.103387.

J. Baron and T. Pohlman. Precision and bias in the assessment of essentiality rates in firms' protfolios of declared seps. Technical report, IPlytics, 2021.

R. Bekkers, J. Henkel, E. M. Tur, T. V. D. Vorst, M. Driesse, B. Kang, A. Martinelli, W. Maas, B. Nijhof, E. Raiteri, L. Teubner, and N. Thumm. Pilot study for essentiality assessment of standard essential patents. Scientific analysis or review KJ-NA-30111-EN-N (online),KJ-NA-30111-EN-C (print), Luxembourg (Luxembourg), 2020.

I. Beltagy, K. Lo, and A. Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://aclanthology.org/D19-1371.pdf.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, Dec. 2017.

M. R. Bouadjenek, S. Sanner, and G. Ferraro. A study of query reformulation for patent prior art search with partial patent applications. In T. Sichelman and K. Atkinson, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Law, ICAIL 2015, San Diego, CA, USA, June 8-12, 2015*, pages 23–32. ACM, 2015. doi: 10.1145/2746090.2746092. URL https://doi.org/10.1145/2746090.2746092.

N. Bouayad-Agha, G. Casamayor, G. Ferraro, S. Mille, V. Vidal, and L. Wanner. Improving the comprehension of legal documentation: the case of patent claims. In *The 12th International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 8-12, 2009, Barcelona, Spain*, pages 78–87. ACM, 2009. doi: 10.1145/1568234.1568244. URL https://doi.org/10.1145/1568234.1568244.

S. Bruninghaus and K. Ashley. Predicting outcomes of case based legal arguments. In *Proceedings of the 9th ICAIL*, pages 233–242. ACM, 2003.

I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.pdf.

D. A. Cieslak and N. V. Chawla. Learning decision trees for unbalanced data. In *ECML 2008*, 2008.

C. M. de Oliveira Rodrigues, F. L. G. de Freitas, E. F. S. Barreiros, R. R. de Azevedo, and A. T. de Almeida Filho. Legal ontologies over time: A systematic mapping study. *Expert Syst. Appl.*, 130:12–30, 2019. doi: 10.1016/j.eswa.2019.04.009. URL https://doi.org/10.1016/j.eswa.2019.04.009.

S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

M. A. Hasan and W. S. Spangler. Assessing patent value through advanced text analysis. In A. Gardner and R. Winkels, editors, *The Eleventh International Conference on Artificial Intelligence and Law, Proceedings of the Conference, June 4-8, 2007, Stanford Law School, Stanford, California, USA*, pages 191–192. ACM, 2007. doi: 10.1145/1276318.1276354. URL https://doi.org/10.1145/1276318.1276354.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

V. Hofmann, J. Pierrehumbert, and H. Schütze. Dynamic contextualized word embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6970–6984, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.542. URL https://aclanthology.org/2021.acl-long.542.pdf.

T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 133–142, 2002.

M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML 1997*, pages 179 – 186, 1997.

T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(211 – 240), 1997.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

S. Li, S. Ju, G. Zhou, and X. Li. Active learning for imbalanced sentiment classification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural*

Language Processing and Computational Natural Language Learning, pages 139–148, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D12-1013.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

M. Medvedeva, M. Vols, and M. Wieling. Using machine learning to predict decisions of the european court of human rights. *Artif. Intell. Law*, 28(2):237–266, 2020. doi: 10.1007/s10506-019-09255-y. URL https://doi.org/10.1007/s10506-019-09255-y.

T. Mikolov, K. Chen, and J. Dean. Efficient estimation of word representation in vector space. In *Proc. of International Conference on Learning Representations*, 2013.

J. Pennington, R. Socher, and C. D. Manning. Glove: global vectors for word representation. In *Proc. of EMNLP*, pages 1532–1543, 2014.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.pdf.

F. Provost. Machine learning from imbalanced data sets. In *AAAI 2000 Workshop on Imbalanced Data Sets*, 2000.

J. Reisinger and R. J. Mooney. Multi-prototype vector-space models of word meaning. In *Proc. of HLT-NAACL*, pages 109–117, 2010.

E. Rissland and K. Ashley. A case-based system for trade secrets law. In *Proceedings of the 1st ICAIL*, pages 60–66. ACM, 1987.

V. Vapnik. *Statistical Learning Theory*. Wiley, Chichester, GB, 1998.

G. Wu and E. Y. Chang. Adaptive feature-space conformal transformation for imbalanced-data learning. In *ICML'03*, 2003.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.