Artificial Intelligence and Law (2023) 31:293–324
https://doi.org/10.1007/s10506-022-09313-y

ORIGINAL RESEARCH

Using machine learning to create a repository of judgments concerning a new practice area: a case study in animal protection law

Joe Watson[1] · Guy Aglionby[2] · Samuel March[3]

Accepted: 5 April 2022 / Published online: 8 May 2022

Abstract

Judgments concerning animals have arisen across a variety of established practice areas. There is, however, no publicly available repository of judgments concerning the emerging practice area of animal protection law. This has hindered the identification of individual animal protection law judgments and comprehension of the scale of animal protection law made by courts. Thus, we detail the creation of an initial animal protection law repository using natural language processing and machine learning techniques. This involved domain expert classification of 500 judgments according to whether or not they were concerned with animal protection law. 400 of these judgments were used to train various models, each of which was used to predict the classification of the remaining 100 judgments. The predictions of each model were superior to a baseline measure intended to mimic current searching practice, with the best performing model being a support vector machine (SVM) approach that classified judgments according to term frequency—inverse document frequency (TF-IDF) values. Investigation of this model consisted of considering its most influential features and conducting an error analysis of all incorrectly predicted judgments. This showed the features indicative of animal protection law judgments to include terms such as ‘welfare’, ‘hunt’ and ‘cull’, and that incorrectly predicted judgments were often deemed marginal decisions by the domain expert. The TF-IDF SVM was then used to classify non-labelled judgments, resulting in an initial animal protection law repository. Inspection of this repository suggested that there were 175 animal protection judgments between January 2000 and December 2020 from the Privy Council, House of Lords, Supreme Court and upper England and Wales courts.

Keywords Machine learning · Court judgments · Document classification · TF-IDF · Support vector machine

✉ Joe Watson
j.watson@jbs.cam.ac.uk

Extended author information available on the last page of the article

1 Introduction

Case law has long impacted on the protection of animals. However, judgments concerned with animal protection have typically been assigned to established areas of legal practice such as veterinary negligence, defamation, criminal, regulatory and public law. Indeed, the recognition of animal protection law as a distinct area of professional legal practice is a relatively new phenomenon. This is highlighted by the recent formation of the UK's first dedicated animal protection law firm, Advocates for Animals (founded in 2019: Advocates for Animals, n.d). The nascency of animal protection law as a distinct practice area means that there has been no publicly available repository of UK animal protection law judgments. With no repository, those seeking to identify animal protection law judgments might have to consider the relevance of individual judgments found through keyword searches. Such an approach is time-consuming, prone to human error and potentially hindered by flawed search tools (see Sect. 3.1). We therefore use computational techniques to create an initial repository of judgments meeting our adopted animal protection case law definition. Under this definition, an animal protection law judgment is one that substantially concerns, relates to, or affects the welfare or protection of one or more animals (following Overcash, 2012).

Beyond the creation of a repository, this research makes several subsidiary contributions. The suitability of different machine learning (ML, see Sect. 2) models to practice area classification is considered by comparing their performance (against one another and a baseline indicative of current judgment searching practice). Criticism of the limited interpretability of ML in law is addressed by exploring the most influential features in judgment classification. Further, this research is intended to promote ML understanding among non-technical legal researchers and practitioners through the employment of a straightforward structure.

This paragraph lays out this structure for the remainder of the paper. In Sect. 2, we describe the concepts underpinning this work, identify related work and consider why the application of ML in law remains contentious. Section 3 presents our repository creation process, with dedicated subsections explaining how we:

- Searched judgments for those containing a keyword
- Labelled a selection of these judgments,
- Trained models using these data,
- Calculated results and chose the final model,
- Evaluated the final model,
- Used the model to classify unlabelled judgments.

Repository creation information is followed by a discussion of findings, limitations and potential future work in Sect. 4. Lastly, Sect. 5 concludes the paper by summarising the creation process and contents of our law repository.

2 Background

This paper details the creation of an animal protection law repository using ML. ML is the practice of using algorithmic methods to learn from data and make predictions. It is often used as a tool for natural language processing (NLP), which is concerned with using computational techniques to process and analyse human language data. Both ML and NLP have been employed in the legal space for some time (Nay, 2018). Westlaw, for example, has offered legal search tools drawing on both techniques since the 1990s (Thomson Reuters, n.d.). As the availability of law-related documents has heightened, modelling and feature extraction approaches originating from the fields of ML and NLP (see Sect. 3) have become ever more important to legal research.

The (increasing) importance of ML and NLP techniques is reflected in their application to a broad range of legal tasks. These tasks include computing judgment similarity (Mandal et al., 2021), predicting violations of the European Convention on Human Rights (Aletras et al., 2016; Medvedeva et al., 2020) and gauging the influence of demographics characteristics on judicial decision making (Rachlinski & Wistrich, 2017). The application of ML and NLP has even extended to the task underlying this research—legal text classification. Prior law-centred classification endeavours have involved the identification of: documents 'relevant' and 'not relevant' to legal claims during e-discovery processes (discussed by Ashley, 2017); whether a statutory provision applies to a legal issue (Savelka et al., 2015); and, which *well-established* practice area a court judgment falls into (Lei et al., 2017; de Araujo et al., 2020).[1] In creating a judgment repository for a *recently recognised* practice area that is partially automatically constructed, this research therefore widens the range of legal classification tasks addressed through ML and NLP.

There are patterns in how previous practice area classification efforts have both represented legal documents and used these representations during modelling. Judgments have typically been represented using methods based on term frequency. De Araujo and colleagues (2020) classified Brazilian lawsuits by established 'themes' using a tuned term frequency—inverse document frequency approach (TF-IDF: see Sect. 3.3.1). Similarly, Lei and colleagues (2017) used a TF-IDF method as their sole document representation method when categorising Chinese judgments by 'industry divisions'.[2] We find no evidence that neural representation approaches such as Bidirectional Encoder Representations from Transformers (BERT) have been used for judgment practice area categorisation. However, the viability of such representation approaches is suggested by their employment in other legal document

[1] Research by Sulea and colleagues (2017a; 2017b) also involved the prediction of the 'law area' of French court cases. These law areas could not, however, be considered equivalent to practice areas. Instead, the areas represented different divisions (or, *chambres*), such as the *Chambre Sociale*, *Chambre Civile 1*, *Chambre Civile 2* and *Chambre Civile 3*.

[2] Term frequency methods have also been employed in related legal document classification work beyond practice area categorisation. French court cases were classified by division (or, chambre) using a simple term frequency method (bag of words: Sulea et al., 2017a), before better model performance was achieved on the same task with TF-IDF (Sulea et al., 2017b).

classification work. Undavia et al. (2018) found the automatic classification of legal court opinions into Supreme Court Database categories to be best achieved using a neural representation method (in conjunction with a convolutional neural network model). Embeddings from a BERT-based model were found to provide an effective foundation for multi-label classification of a dataset of legal opinions (Song et al., 2021). Additionally, Longformer achieved state-of-the-art performance on a case outcome prediction task (Bhambhoria, Dahan & Zhu, 2021).[3]

Both Lei and colleagues (2017) and de Araujo and colleagues (2020) trialled multiple ML models when classifying court judgments using TF-IDF representations. In the former, this trialling showed a linear support vector machine (SVM: see Sect. 3.3.2) model to outperform naive Bayes (NB), decision tree and random forest models. In the latter, the results achieved by an XGBoost approach surpassed those from SVM and NB approaches. When using a smaller dataset, however, XGBoost and SVM results were comparable (de Araujo et al., 2020). To our knowledge, there exists no published research in which multi-layer perceptrons (MLPs, see Sect. 3.3.2) have been applied to judgment practice area tasks. Yet, MLPs have been used elsewhere in legal research to, for example, recognise named entities (such as the date of judgment) within court judgments (Vardhan et al., 2020). MLPs also have aided a diverse range of document classification efforts outside the legal sector, spanning from the categorisation of German business letters (Wenzel et al., 1998) to identifying the author of plays (Merriam & Matthews, 1994).

While ML has been used in various legal settings, its application remains contentious. Three key reasons for this contentiousness are briefly considered, each of which is pertinent to our research. Firstly, methods involving ML are considered backward-looking in relation to legal reasoning (Markou and Deakin, 2020). 'ML's effectiveness is diminished in direct relation to the novelty of the cases it must process' and, connectedly, the rate of change in the context where it is applied (Markou and Deakin, 2020, p. 63). This concern is clearly consequential to certain legal tasks, such as the prediction of future European Court of Human Rights case outcomes (see Medvedeva et al., 2020). The backward-looking nature of ML might also affect our ML endeavours. This is because our selected model could conceivably be applied to judgments made in the years following 2020. However, the primary intention of this research was inherently retrospective—to create a repository of *existing* judgments between 2000 and 2020. As such, we ultimately specified our model in a manner that prioritised predictive performance on past judgments over future judgments (Sect. 3.5).

A second concern about the use of ML in law stems from the potential threat that these methods pose to the autonomy of judicial processes. This threat would be clearly manifested should ML methods be used to conduct adjudication without

[3] Embeddings have also been applied to document similarity research. Universal Sentence Encoder (USE) embeddings were successfully employed in a model that identified semantically similar sentences in legal documents (to aid human annotation of sentences: Westermann et al., 2020). Further, BERT was one of multiple feature extraction approaches trialled on a court case similarity task (Mandal et al., 2021).

human participation (Markou & Deakin, 2020). In such an instance, ML approaches would be contributing directly to law creation. In contrast with a model that carries out adjudication, the ML models in our paper are merely intended to aid the identification of animal protection law judgments. They could not, therefore, be seen to contribute to the law directly. Yet, the case law repository created through machine-based approaches might still contribute to the indirect creation of law (see further, Burri, 2017).[4] This could occur if, for example, lawyers' arguments were influenced by animal protection judgments that they would not have identified without a repository made using ML predictions. Still, the tangential nature of such a contribution to law creation means that the models considered in this paper need not be considered to pose a serious threat to judicial autonomy. Indeed, lawyers concerned with animal protection law are likely already using machine-based (albeit not ML-based) approaches to identify judgments (such as the 'Case Law Search' tool discussed below: Sect. 3.1).

Thirdly, arguments against the employment of ML in law have focused on the limited extent to which these techniques permit human understanding. Deakin and Markou (forthcoming, p. 15) contended that the findings of ML models cannot 'be adequately explained using the types of arguments which lawyers are accustomed to making'. However, this research makes efforts to address the comprehension-based concerns voiced in the law and technology literature. The process through which ML techniques are applied is detailed in a step-by-step manner. Papers with such a format have arisen infrequently, with these instead more 'focused on the implications of the running model' (Lehr & Ohm, 2017, p. 655). Further, the features most influential in determining whether or not a judgment is classified as concerned with animal protection law by our selected model are presented graphically and considered qualitatively (Sect. 3.5). By providing intelligible explanations of the manner in which our final model classifies judgments and the process through which it was created, it is hoped that this paper contributes to efforts to demystify the use of NLP and ML in law (Lehr & Ohm, 2017).

3 Repository creation

3.1 Searching judgments

Creating an animal law judgment repository began by identifying court judgments that contained the word 'animal'. We adopted the assumption that this term would be present in every animal protection law judgment through discussion with the domain expert, who felt it highly improbable that any such judgment would not contain this term (Sect. 4). In fact, it is feasible that prior attempts to identify relevant judgments would have treated any containing 'animal' as relating to animal protection law until

[4] Burri (2017) previously recognised how ML might lead to the emergence of 'soft law' through 'behavioural interaction among [ML] systems' (p. 3). This paragraph suggests that ML might produce law in a different manner.

proven otherwise. We therefore use this strategy as a baseline measure for comparison against ML models (Sect. 3.4).[5]

The judgments searched were all those available from the British and Irish Legal Information Institute (BAILII) made by the Privy Council, House of Lords, Supreme Court and upper England and Wales courts between 2000 and 2020.[6] BAILII was used as the basis for this search as it provides the most extensive collection of British legal materials freely available online (BAILII, n.d.). Searching involved opening each judgment and recording its URL when ‘animal’ was found in the text. Implementation of the search in late December 2020 found that 1637 of the 55,202 judgments by upper courts from January 2000 to December 2020 contained the word ‘animal’.

Those judgments that contained ‘animal’ were typically longer than those that did not. The median word length for judgments with ‘animal’ was 10,785, while those without had a median of 5869. This corresponds with findings on the number of sentences in judgments containing ‘animal’. Judgments containing ‘animal’ had a median of 420 sentences, while those without had a median of 231. These statistics are presented below, alongside minimum, maximum, 5th percentile and 95th percentile values for both the number of words and sentences across each group (Table 1). The difference in length is potentially unsurprising: the likelihood of any given term occurring in a document increases with document length.

An attempt was made to streamline the search for relevant judgments by using BAILII’s inbuilt ‘Case Law Search’ tool. This tool appeared as if it should be able to identify judgments containing a user-specified word. However, trialling demonstrated that the ‘Case Law Search’ misreported the number of judgments identified, reported certain judgments twice, picked up judgments that did not contain ‘animal’ and missed other judgments that did. Specifically, a BAILII search of the same courts over the same period as used previously claimed to provide 1810 judgments containing ‘animal’, but actually gave a list of 1809 judgments of which 1800 were unique and that only contained the term ‘animal’ on 1568 occasions. What is more, BAILII’s Boolean search tool did not provide a count of how many judgments were searched. Using the ‘Case Law Search’ would therefore also have obstructed the identification of the proportion of judgments that contained the term, ‘animal’, and were concerned with animal protection law (Sect. 3.6). The identification of these limitations suggested the judgment-by-judgment search method initially employed to be better for creating an animal protection law repository. Additionally, the shortcomings of BAILII’s ‘Case Law Search’ mean that our repository of 1637 judgments is a more precise collection of judgments containing ‘animal’ than that which could be created through BAILII alone.

[5] The creation of a baseline measure through assigning all judgments to a singular class also follows previous judgment classification research (de Araujo et al., 2020).

[6] Specifically, these courts were the House of Lords (2000 to 2009), UK Supreme Court (2009 to 2020), Privy Council, Court of Appeal (Civil and Criminal divisions) and High Court (Administrative, Admiralty, Chancery, Commercial, Senior Court Costs Office, Family, Technology and Construction, Mercantile, Patents and Queens’ Bench Division).

Table 1 Descriptive statistics for all judgments

	Number of judgments	Minimum words	5th percentile words	50th percentile words	95th percentile words	Maximum words	Minimum sentences	5th percentile sentences	50th percentile sentences	95th percentile sentences	Maximum sentences
Without animal	53,492	238	1318	5869	21,498	301,438	8	52	231	866	16,837
With animal all	1637	462	2781	10,785	42,711	360,301	23	107	420	1772	18,421

Note. All values are provided to 0 decimal places

The ‘Without animal’ judgments exclude 73 judgments for reasons including that they were reported to have been withdrawn (with no opinion or explanation), moved to a different court (with no opinion or explanation), removed due to transcription mistakes or contained no sentence punctuation. No judgments containing the term ‘animal’ were excluded

3.2 Labelling judgments

500 judgments were randomly sampled for human labelling from the 1637 judgments found to contain ‘animal’. Labelling was carried out by the domain expert alone, following guidance written by them and the lead author (available here: https://github.com/JoeMarkWatson/animal_law_classifier/blob/main/animal_protection_law_labelling_guidance.docx). After human labelling was completed, stratified random sampling was used to create a training set of 400 labelled judgments and a testing set of 100 labelled judgments, each with the same proportion of positively-tagged judgments. Accordingly, the 400 training set judgments included 66 concerning animal protection law, while the 100 test set judgments included 17 (after the correction of one human labelling error that did not affect stratification: Sect. 3.5.2). Word and sentence length information is provided for those judgments containing ‘animal’ that: were not labelled; were labelled; were labelled and assigned to the training set; and, were labelled sentences and assigned to the test set (Table 2).

3.3 Model training

Each implemented ML approach depended on both feature extraction and modelling. Feature extraction involved transforming the text of each judgment into a format suitable for use in a ML model. Modelling entailed using the extracted features for each judgment, combined with their label, for ML training.

3.3.1 Extracting features

Five feature extraction methods were used: TF-IDF vectors; USE (Cer et al., 2018); sentence-BERT (s-BERT: Reimers & Gurevych, 2019); Longformer (Beltagy, Peters & Cohan, 2020); and BigBird (Zaheer et al, 2020) embeddings. Each method takes the text of a judgment as input, and returns a vector representation of the text. The first 200 words of each judgment were excluded, as this contained judges’ and party’s names that, if employed in a classification model, could lead to overfitting. Explicitly identifying and removing this information would have been preferable, yet such an approach was obstructed by the (variable) structure of judgments on BAILII (see Sect. 4).

TF-IDF is a well-established approach in the field of NLP (Jones, 1972) that has been frequently applied to judgment classification tasks (see Sect. 2). The approach is based on a bag-of-words assumption, which takes no account of the relationship between terms. TF-IDF is defined as follows (Eq. 1):

$$TF - IDF(t, d) = TF(t, d) * IDF(t) \tag{1}$$

The TF-IDF value for a term t in a document d is a function of its frequency within that document (term frequency, TF) and its overall frequency in the corpus (as inverse document frequency, IDF). IDF is defined in Eq. 2:

$$IDF(t) = log\left[(1 + N)/(1 + DF(t))\right] + 1 \tag{2}$$

Table 2 Descriptive statistics for judgments containing ‘animal’

	Number of judgments	Minimum words	5th percentile words	50th percentile words	95th percentile words	Maximum words	Minimum sentences	5th percentile sentences	50th percentile sentences	95th percentile sentences	Maximum sentences
With animal unlabelled	1137	462	2788	10,743	43,584	273,811	23	107	412	1857	18,421
With animal labelled	500	1014	2781	10,824	40,220	360,301	33	106	444	1615	16,090
With animal train	400	1014	2782	10,967	41,359	187,320	33	106	446	1667	9605
With animal test	100	1264	2729	9885	29,863	360,301	58	109	418	1358	16,090

Note. All values are provided to 0 decimal places

In the above, N is the number of documents in the corpus, and DF(t) is the number of these that contain *t*.[7]

Various pre-processing methods were tested when creating TF-IDF vectors; the use or not of each method was controlled via a parameter. One such parameter controlled whether terms were lemmatised or not before term and document frequencies were calculated. Lemmatisation involves reducing terms that have been inflected in accordance with tense or number, inter alia, to a root term (lemma). For example, the lemmas of ‘culling’ and ‘culled’ are both ‘cull’. Additionally, the terms themselves could be single words (unigrams), or multiple contiguous words grouped to act as a single term (n-grams; bigrams when two contiguous words are used). Minimum and maximum term document frequency thresholds were also set, with the terms appearing above or below a certain frequency excluded (as these might not aid classification efforts). Lastly, the vector size was controlled by only counting the top *n* most frequent terms. All parameters were optimised during training (see Sect. 3.3.3).

When the TF-IDF feature extraction method was used, text underwent additional pre-processing before creating representations. This additional pre-processing was only carried out when creating TF-IDF vectors, as neither USE nor s-BERT embeddings directly represent individual terms in the same way as TF-IDF vectors. The intention of this TF-IDF pre-processing step was to improve the generalisability of the model to data outside of the training set. Pre-processing therefore involved removing aspects of the text which were assumed not to be indicative of a judgment’s classification, but could nonetheless be used by the model because of a chance correlation with one of the classes. This entailed:

- Removing URLs and HTML tags from the text,
- Transforming words to lowercase,
- Deleting digits and punctuation from the text,
- Retaining only English words.

While transformer-based embedding approaches have been used previously for numerous legal tasks (outside judgment practice area classification: Sect. 2), these remain relatively new developments in NLP. USE and s-BERT sentence-embedding approaches were presented, respectively, in 2018 by Cer and colleagues, and 2019 by Reimers and Gurevych in 2019 (following the previous development of BERT: Devlin et al., 2019). The Longformer and BigBird transformer approaches to long text sequence embedding were developed more recently still (Beltagy, Peters & Cohan, 2020; and Zaheer et al, 2020, respectively). These models have all been trained on a large amount of text; by enhancing the investigation of newly available data through models trained on previously available data, this research employed a

[7] The ‘1’ at the end of the IDF formula (Eq. 2) means that terms occurring in all documents need not be wholly disregarded, yet the maximum document frequency values ultimately selected through hyperparameter tuning (Sect. 3.3.3) resulted in the exclusion of those words occurring in at least 60 percent (TF-IDF SVM) and 70 percent (TF-IDF MLP) of documents.

form of transfer learning (Devlin et al., 2019). This instilled our judgment classifier with natural language understanding derived from text data outside training set judgments. As a result, embeddings-based models might be able to function adequately with less labelled data than that required by more traditional approaches (like TF-IDF: Asgari-Chenaghlu et al., 2020). Achieving acceptable performance with limited training data was important in our case, as human labelling was both time consuming (taking approximately 10 hours per 100 judgments) and could conceivably have been expensive (see further, Muller et al., 2021).[8]

There are multiple types of USE, s-BERT, Longformer and BigBird models. In this research, the large USE, base s-BERT, Longformer and BigBird models were used, returning 512-dimensional embeddings for USE and 768-dimensional embeddings for all other models. As all models are pre-trained, no hyperparameters need to be tuned. Longformer and BigBird were used because they are designed to embed text sequences longer than just sentences, which USE and s-BERT are designed for. One or more embeddings are computed for each document using each model; for USE and s-BERT by embedding each sentence, and for Longformer and BigBird by splitting judgments into chunks up to the maximum length allowed by the models, and embedding these.

As it is desirable for a judgment to be represented by only one embedding, we consider two methods for obtaining these from the (potentially) multiple that exist. The first is to select the embedding of the first sentence or chunk, and the second is to take a mean average over all embeddings. Prior work has found averaging sentence embeddings to work well for document retrieval (Yang et al, 2019), aspect-extraction (Verma et al, 2021), and fake news identification (Slovikovskaya & Attardi, 2020). The second approach is therefore taken for the sentence-based models. Another reason for this choice is that it is unlikely that the first sentence of a judgment would be sufficient to identify whether the case concerned animal welfare according to our definition. Two steps were taken to limit the computational demands of the sentence embedding process. First, sentence embeddings were not created in the rare case when a sentence was over 1000 words. Second, 5000 sentences were randomly sampled for embedding from any judgment that exceeded 5000 sentences in length.

Both approaches were trialled for the document-level models. It was found that using the first chunk gave better results than averaging all embeddings. This is likely due to the large amount of text represented by each embedding. Indeed, 49 of the 500 labelled judgments could be entirely represented by one embedding.

3.3.2 Modelling

This research trialled two modelling approaches. One was a linear SVM (Cortes & Vapnik, 1995) specified using the scikit-learn library (Pedregosa et al., 2011). The other was a MLP (Rumelhart et al., 1986) created through the scikit-learn wrapper for Keras (Chollet, 2015). The trialling of a linear SVM approach follows their

[8] The domain expert who classified judgments for this project did so free of charge.

strong performance in previous judgment practice area classification tasks (Sect. 2). Linear SVMs output a single weight per input feature based on 'a boundary in a vector space between the positive and negative instances of a category or class that is maximally distant from all of the instances' (Ashley, 2017, p. 251). These models are trained to maximise the margin of the decision boundary and classify as many training points correctly as possible (Boser, Guyon & Vapnik, 1992). These two training aims can conflict, with the extent to which priority is given to either objective controlled through the regularisation hyperparameter, C (see below).

MLPs have long been applied to a broad range of document classification tasks, although this has hitherto not extended to practice area classification (Sect. 2). In contrast with the SVM, which separates data with a linear boundary, the MLP includes an activation function which makes a nonlinear boundary possible (see Sect. 3.3.3). This function is applied in a so-called 'hidden layer' between the input layer, where the features are provided to the model, and the output layer, where the classification is given. If our judgment classification task had such complexity that a linear boundary could not adequately separate classes, the application of a nonlinear boundary could increase model performance. However, it is also possible that using a non-linear multi-layer method could lead to overfitting that limits generalisability to new data.

Creating MLP models involved setting various hyperparameters. One of these controlled whether dropout (Srivastava et al., 2014) was employed. Dropout is a regularisation technique that limits potential overfit by randomly dropping neurons and their connections during training, which has been shown to improve ML performance on tasks including document classification (Maaten et al., 2013; Srivastava et al., 2014). Where a non-zero dropout value was used, it was applied to both the input and hidden layer and used in conjunction with max-norm regularisation (set at a value of 3, as using dropout with max-norm regularisation is likely to produce better results than dropout alone: Srivastava et al., 2014). Hyperparameter settings also affected the model's learning rate and number of epochs. The learning rate controls the magnitude of changes made to model weights during each update, with higher learning rates producing larger changes. An epoch denotes a full pass through the training data. These two settings are interdependent, as models trained with low learning rates will generally require more epochs to train and vice versa.

3.3.3 Implementation

Ten ML systems were established by combining the two modelling approaches (SVM and MLP) and five feature extraction approaches (TF-IDF, USE, s-BERT, Longformer and BigBird). The complexity of these systems differs in accordance with two factors. Firstly, a MLP is more complex than a SVM. SVMs are only able to separate data with a linear boundary, while MLPs can use a non-linear boundary. Secondly, TF-IDF embeddings are simpler than USE, s-BERT, Longformer and BigBird embeddings. TF-IDF embeddings are derived from word frequency counts, whereas embeddings approaches use large neural networks trained on external data.

A grid search was performed to identify the optimal hyperparameters for each ML approach, by choosing the set of hyperparameters that gave the highest

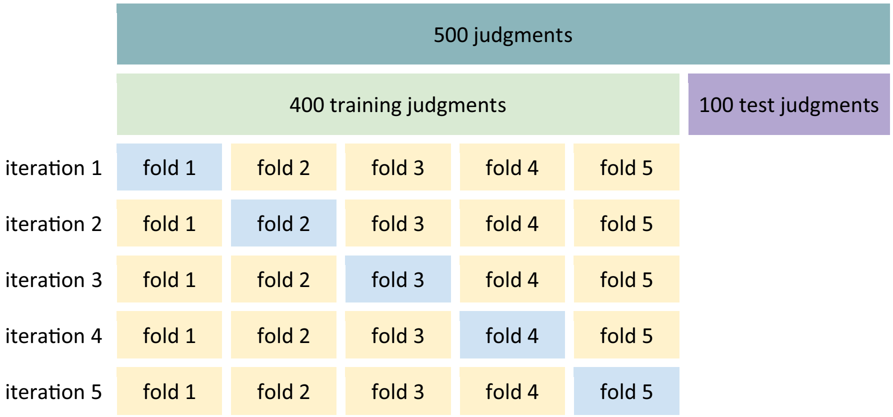

Fig. 1 Five-fold cross validation. *Note*. In each row, folds used for training are in yellow and the validation fold is blue. (Color figure online)

average macro-F1 score in five-fold cross validation. Stratified five-fold cross validation splits the dataset into five equally sized sets, or ‘folds’, each with the same proportion of judgments related to animal protection law. In each iteration, four folds are used together for training while the remaining one is used for validation. This process is depicted below (see Fig. 1).

We considered the optimal hyperparameters to be those that achieved the greatest mean macro-F1 score across validation folds. Macro-F1 is a simple arithmetic mean of per-category F1 scores. F1 is given by the following equation (Eq. 3):

$$F1 = 2(P * R)/P + R) \tag{3}$$

In the above formula, the F1 score is the harmonic mean of precision (P) and recall (R). Precision and recall are defined below (Eqs. 4 and 5).

$$R = TP/(TP + FN) \tag{4}$$

$$P = TP/(TP + FP) \tag{5}$$

In Eqs. 4 and 5, TP refers to a true positive, FN to a false negative and FP to a false positive classification. Macro-averaged F1 is preferable to accuracy when there is a class imbalance as it more greatly reflects poor performance on the minority class.

For all SVM systems, the following hyperparameters were optimised:

- Loss function: squared hinge or hinge,

- Regularisation value, C: 0.1, 1, 2, 5, or 10.[9]

For MLP systems, different hyperparameters were tuned:

- Number of neurons in the hidden layer: NF/16, NF/8, NF/4, and NF/2,[10]
- Dropout on both the input layer and hidden layer: 0, 0.2,
- Learning rate: 0.1, 0.01 and 0.001,
- Number of epochs: 20, 50, 100 and 200.

For MLP systems, 'NF' in the first bullet point refers to the number of input features; for example, in USE embeddings this was 512. All models employed one hidden layer with ReLU activation and an output layer of one neuron with sigmoid activation. Each MLP system was also trained using the Adam optimiser (Kingma & Ba, 2017) and binary cross-entropy loss with a batch size of 32.

In SVM or MLP systems that used TF-IDF vectors, we also optimised the following feature extraction hyperparameters:

- Term lemmatisation: using lemmatised or unlemmatised terms,
- Single terms or multiple contiguous terms: employing unigrams, or unigrams and bigrams,
- Upper document frequency (DF) threshold: ignoring terms featuring in over 60 percent, 70 percent, or 80 percent of all documents,
- Lower DF threshold: ignoring terms featuring in less than one, two or three documents,
- Vector size: Employing a maximum number of TF-IDF features of 500 or 1000.

3.4 Results and final model selection

The optimal hyperparameter combinations for each system are detailed below (Tables 3 and 4), with the macro-F1 values achieved by these combinations provided later (alongside test set results: Table 5).

After establishing the best-performing hyperparameters for each system, we next considered these models against a baseline measure and one another. In the baseline measure, all 100 judgments in the test set were assumed to concern animal protection law (in a manner intended to mimic existing attempts to identify animal protection law judgments: Sect. 3.1). For the systems, the versions that achieved the best average score across validation folds were used to predict the classifications of the test set judgments before macro-F1 scores were recorded (Table 5). While average validation fold and test set results are provided together, only the test set judgments

[9] A C value of 10 was permitted for USE and s-BERT embeddings models but not the TF-IDF model. This decision was taken to limit the training time of the TF-IDF model, which was otherwise heightened by the tuning of multiple TF-IDF-specific hyperparameters.

[10] The shape of the TF-IDF vectors used during MLP model training meant that the options for this hyperparameter had to be altered (becoming NF/20, NF/10, NF/4, NF/2).

Table 3 Optimal hyperparameter combinations for TF-IDF systems

		SVM	MLP
TF-IDF hyperparameters	*Lemmatisation*	True	True
	Unigrams	True	True
	Upper DF	60	70
	Lower DF	1	3
	Number of features	1000	1000
SVM hyperparameters	*C value*	1	
	Loss function	Squared hinge	
MLP hyperparameters	*Neurons*		250
	Dropout		0.2
	Learning rate		0.01
	Epochs		50

Note. For models using TF-IDF vectors: a Lemmatisation value of 'True' means that terms were lemmatised; and, a Unigrams value of 'True' means that the model only considered unigrams

The selected TF-IDF SVM lower DF value (1) was considered surprising, as retaining terms featuring in just one document can lead to overfit. However, trialling suggested that the influence of this hyperparameter was limited as TF-IDF features remained highly consistent when setting this value to different integers over one

Table 4 Optimal hyperparameter combinations for embeddings systems

		USE	s-BERT	Longformer	BigBird
SVM hyperparameters	*C value*	10	1	10	10
	Loss function	Squared hinge	Hinge	Hinge	Squared Hinge
MLP hyperparameters	*Neurons*	256	96	192	48
	Dropout	0.2	0.2	0.2	0.2
	Learning rate	0.01	0.001	0.001	0.1
	Epochs	10	50	100	10

had not been used for any part of model training. As such, the macro-F1 testing findings provide a more unbiased estimate of model performance.

All ML-based systems significantly outperformed the baseline on the test set ($p = 0.001$).[11] Amongst these ML systems, the best-performing was the SVM model with TF-IDF features. The difference between the TF-IDF SVM and other systems was significant ($p = 0.05$) except in cases where USE embeddings were employed. As the TF-IDF SVM outperformed the less complex baseline measure while clearly

[11] A monte carlo permutation test was used to determine statistical significance in all cases.

Table 5 Comparison of system macro-F1

	Baseline	TF-IDF SVM	USE SVM	s-BERT SVM	Longformer SVM	BigBird SVM	TF-IDF MLP	USE MLP	s-BERT MLP	Longformer MLP	BigBird MLP
macro-F1 validation	0.142	0.780	0.796	0.820	0.766	0.660	0.814	0.812	0.813	0.469	0.455
macro-F1 testing	0.145	0.866	0.828	0.784*	0.765*	0.554*	0.764*	0.801	0.759*	0.457*	0.457*

Note. An asterisk (‘*’) denotes model test scores that are significantly worse than the selected TF-IDF SVM model score ($p = 0.05$)

not being outperformed by any more complex system, our results suggested that it should be used to predict the classification of the remaining (unlabelled) judgments.

Before concluding system selection, we considered whether the relative testing performance of the TF-IDF SVM seemed logical. Various reasons could suggest this not to be the case. Firstly, all feature extraction methods other than TF-IDF might have allowed models to benefit from transfer learning (Sect. 3.3.1). Secondly, as TF-IDF makes a bag of words assumption there is some loss of information in the embedding (Sect. 3.3.1). Thirdly, unlike a SVM,[12] a MLP can learn nonlinear decision boundaries, which is plausibly necessary to model the intricacies of judgment classification.

However, we identified a number of factors that did support the relative strength of the TF-IDF SVM approach. The tuning process for the TF-IDF SVM optimised feature extraction hyperparameters. This is in contrast with transformers, which were not tuned for the task. Additionally, it is quite possible that useful information was lost in the creation of our embeddings features. The averaged USE and s-BERT embeddings were simple means of the many sentence vectors created for each judgment. Similarly, the Longformer and BigBird embeddings did not take account of the latter part of longer judgments. The superiority of the TF-IDF SVM over the TF-IDF MLP could also be due to SVMs being insensitive to class imbalance relative to MLPs (Japkowicz & Stephen, 2002) and possible MLP overfit.[13] Moreover, the finding that a TF-IDF feature extraction technique and SVM architecture performed strongly on a judgment classification task is consistent with previous literature (Sulea et al., 2017b; Lei et al., 2017). Given the existence of reasons why a TF-IDF SVM might outperform other models and congruous findings in related work, our choice to use a TF-IDF SVM was finalised.

3.5 Investigating the chosen model

3.5.1 Considering influential features

To interpret the behaviour of the model, the features (lemmas) with the highest and lowest feature coefficients (weights) were plotted (Fig. 2). Lemmas with high coefficient values were likely to be most predictive of judgments that were concerned with animal protection law while those with low coefficient values were likely most predictive of judgments that were not. These weights provided us with encouragement that the selected model was taking relevant information into account when classifying judgments. There were very few ‘meaningless’ unigrams amongst those with the highest and lowest coefficient values (cf. Medvedeva et al., 2020, p. 255). In fact, the majority of the terms with positive coefficient values in Fig. 2 were logical predictors of animal protection law judgments. ‘Welfare’, ‘hunt’ and ‘conservation’, for

[12] A linear kernel function was used.

[13] While the potential for overfit should have been limited somewhat by, inter alia, the use of regularisation and five-fold cross validation (see Sect. 3.3), the lower macro-F1 scores for both models on the test data than validation data indicates that overfitting might still have occurred.

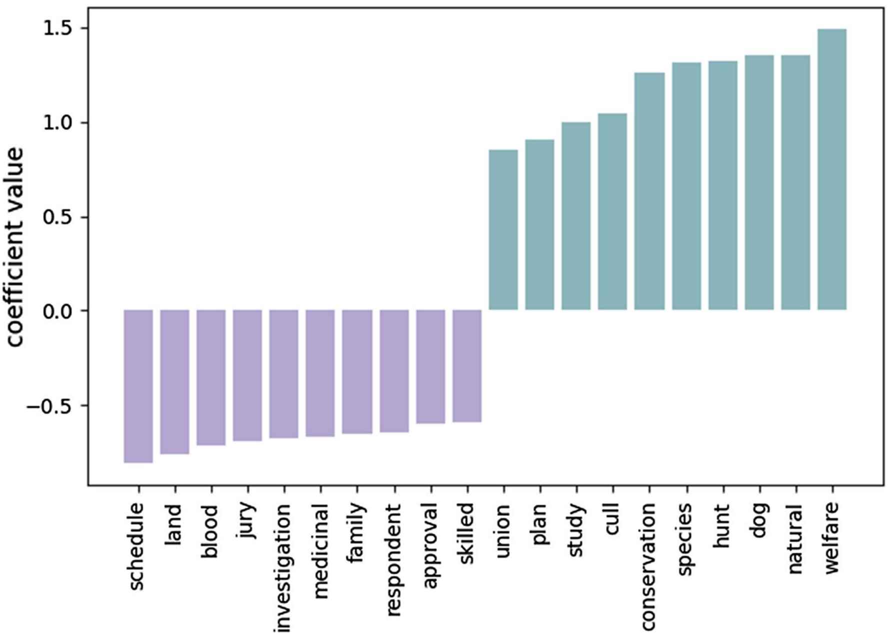

Fig. 2 Ten most positive and negative coefficient values assigned to model lemmas

example, appeared likely to correspond with the adopted definition of animal protection law (see Sect. 1; Overcash, 2012).

The lemmas with negative coefficients also typically seemed rational indicators of judgments containing the term 'animal' that were not predominantly concerned with the welfare or protection of animals. These included 'jury', the presence of which reflected the fact that most criminal animal protection law judgments from 2000 to 2020 came from summary cases (which have no jury). However, the very reason why 'jury' could be considered a rational indicator was also a source of contention. Animal offences will soon become either way offences which are triable with a jury, after the enactment of the *Animal Welfare (Sentencing) Act* 2021 later this year (2021). This could cause an increase in the use of 'jury' in animal protection law judgments, thereby reducing the extent to which the lemma is a useful negative predictor. The domain expert initially felt that the feature should therefore be removed. However, removing 'jury' reduces model performance (albeit not significantly) on the main aim of the system: creating a repository of *existing* animal protection law judgments. The feature was therefore kept, meaning that the system remained backward-looking (Markou and Deakin, 2020).[14]

On first inspection, the domain expert also voiced that a minority of features were not intuitive predictors of whether a judgment was concerned with animal protection

[14] While 'jury' was ultimately retained as a model feature, the discussion of this term highlights the importance of using interpretable approaches in legal NLP.

Table 6 Confusion matrix

	Labelled animal protection law	Labelled not animal protection law
Predicted animal protection law	12 true positives	2 false positives
Predicted not animal protection law	5 false negatives	81 true negatives

law. These included 'schedule', which had the most negative coefficient value. In apparent contrast with this value, the feature could certainly have occurred in judgments concerning animal protection law. *The Endangered Species (Import and Export) Act* 1976, for example, contains multiple schedules of relevance to animal welfare that might have been drawn on in various animal protection law judgments. In fact, schedules within this Act are referenced in a judgment in our initial repository that was classified as animal protection law by our ML model (R v. Sissen, 2000). Yet, further domain expert examination of our judgment repository uncovered many examples of judgments containing the term 'schedule' that did not satisfy our animal protection law definition. These included an EWCA judgment (*European Brand Trading Ltd v. HM Revenue and Customs*, 2016) which referred to schedules under the *Customs and Excise Management Act* 1979. The domain expert therefore ultimately accepted that the presence of 'schedule' could well have been more indicative of a judgment not concerning animal protection law.

3.5.2 Error analysis

An error analysis was carried out on each test set judgment that was mis-predicted by the ML system. A confusion matrix shows that there were relatively few incorrectly predicted test set judgments (7 of 100: Table 6). This matrix also suggests that the system might be more susceptible to false negative errors (i.e., predicting that a judgment is not concerned with animal protection law when it actually is). Conducting human investigation into false negative errors was therefore considered imperative.

Qualitative feedback provided by the domain expert showed the majority of false negative predictions to be highly marginal cases. These included a judgment involving the movement of cattle and regulations around bovine tuberculosis (*Banks v. Secretary Of State For Environment, Food & Rural Affairs*, 2004). While the judgment had clear implications for the protection of animals, these were primarily discussed in financial terms. Additionally, one judgment centred on slander relating to allegations of animal cruelty (Barkhuysen v. Hamilton, 2016). Given that the allegations under discussion were almost certainly false, the domain expert acknowledged that others could contend the protection of animals was never truly a central issue. In contrast with false negative errors, the domain expert felt that both false positive errors were unambiguous (*First Corporate Shipping Ltd t/a Bristol Port Company* *v. North Somerset Council*, 2001; *Sienkiewicz v. South Somerset District Council*, 2015). Still, it was noted that these judgments possessed terminology indicative of animal protection law (with both referencing 'wildlife' and the 'environment'). Feedback on all errors is provided as an annex ('Appendix 1').

This error analysis also led to the identification and subsequent correction of one annotation mistake, where a judgment was initially labelled as not concerning animal protection law. The judgment in question was *R (on the application of Aggregate Industries UK Ltd) v. English Nature* (2002). Re-inspection showed that this judgment clearly centred on animal protection: it concerned a challenge by industry to a Department for Environment, Food and Rural Affairs (DEFRA) decision to designate a particular area as a Site of Special Scientific Interest, which had direct implications for the protection of wild birds. Upon finding that the initial human classification for this judgment was the result of a labelling mistake, the human classification was revised and test set results for all models were re-calculated.

The correction of a labelling error and re-calculation of results could be seen as undesirable. As judgments were revisited solely in instances where the selected ML model and human classifications did not match, any labelling change would only lead to an improvement in the model's macro-F1 score. However, we felt it appropriate that reported results were adjusted for any known error.[15] Further, amending the labelling error improves the judgment repository as this includes all human-labelled judgments.

3.6 Classifying unlabelled judgments

The selected system was applied to the remaining 1137 unlabelled judgments and the results combined with the 500 labelled judgments. This gave a repository of 1637 judgments from the Privy Council, House of Lords, Supreme Court and upper England and Wales courts containing 'animal' and their classifications (available at: https://github.com/JoeMarkWatson/animal_law_classifier/blob/main/case_law_repository.csv). 175 (10.7%) of these judgments were classified as meeting our definition of animal protection law, including 92 found automatically. Following our assumption that the term 'animal' should be present in every animal protection law judgment (presented in Sect. 3.1 and further discussed in Sect. 4), this finding tentatively suggests that 0.32 percent of all 55,202 judgments from our selection of courts were substantially concerned with animal protection law.

The proportion of animal protection law judgments among human- and ML-classified judgments differed substantially. 16.6 percent (or, 83 of 500) of the judgments labelled by the domain expert were found to concern animal protection law, while just 8.09 percent (92 of 1137) of non-labelled judgments were predicted to concern animal protection law by the ML model. It is potentially concerning that the proportion of predicted animal protection law judgments is lower than the proportion

[15] Adjusted results remain similar to non-adjusted results: the macro-F1 score for the baseline measure and all ML models except the USE SVM model increased very slightly (between 0.01 and 0.02); the macro-F1 score for the USE SVM score decreased very slightly (by 0.01).

of human-labelled judgments with the same classification. This suggests that the selected ML model could have misclassified some of the non-labelled judgments and that misclassified judgments were more likely to be false negatives. This idea could be corroborated by the fact that there were more false negatives than false positives among our selected model's test set results (Sect. 3.5.2).

However, domain expert feedback on mis-predicted test set judgments did not suggest the selected model to perform poorly on judgments that concerned animal protection law. Multiple missed animal protection law judgments were highly marginal decisions ('Appendix 1'). The model also correctly classified the majority (12 of 17) of animal protection law judgments in the test set and made almost as many animal protection law predictions (14) as there were judgments (17). What is more, all model predictions on the test set and unlabelled data were made using an architecture that is relatively unsusceptible to class imbalance (Japkowicz & Stephen, 2002). All this suggests that the 500 judgments randomly selected for human labelling might simply have possessed a higher proportion of animal protection law judgments than the 1137 judgments not selected. Tentative support for such a conclusion could be provided by the findings of a brief inspection of ML-classified judgments carried out by the domain expert, which suggested both animal protection law and not animal protection law classifications to be sensical.

The domain expert also surveyed both the ML-predicted and human-labelled judgments that were classified as animal protection law. This highlighted the broad range of issues covered by judgments concerned with animal protection law between 2000 and 2020. These included conventional public law challenges to government actions or decisions (such as a decision to proceed with the culling of badgers: *R (on the application of National Farmers Union) v. Secretary of State for Environment, Food and Rural Affairs*, 2020) and interpretations of the provisions of animal welfare offences (*R (on the application of Highbury Poultry Farm Produce Ltd) v. Crown Prosecution Service*, 2020). There were also advertising standards judgments concerning the use of animal welfare-related language in advertising materials (*R (on the application of Sainsbury's Supermarkets Ltd) v. The Independent Reviewer of Advertising Standards Authority Adjudications*, 2014) and planning judgments that affected endangered animals (*R (on the application of Bizzy B Management Ltd) v. Stockton-On-Tees Borough Council,* 2011). It is hoped that these initial remarks on the breadth of topics within the judgment repository are a precursor to further reflection by other users.

4 Discussion

Users of the animal protection law repository should be aware that it is restricted in scope. Such words of caution would be relevant to any attempt to compile UK court judgments from BAILII. This collection of legal materials contains many but not all judgments (from 2000 onwards in the High Court and above). Additionally, at the time of writing, our repository only includes judgments on BAILII that were made available online before analysis began (mid-December 2020). Indeed, our selected model might not be so successfully applied to future judgments (given anticipated changes in animal protection law judgments: Sect. 3.5.1).[16] The repository also only contains judgments in which the word ‘animal’ was used. The domain expert advised that it was highly likely that all relevant judgment would contain ‘animal’, yet it remains feasible that there are animal protection law judgments which do not include the term.[17] Lastly, UK animal protection law as a whole extends beyond court judgments to ‘hard law in the form of statutes and treaties and soft law such as standards issued by international organisations’ (Peters, 2020, p. 1).

Beyond limitations in the scope of documents on which the model was trained and applied, there were also multiple ways in which the performance of different trialled systems might have been increased. For all USE, s-BERT, Longformer and BigBird systems, text embeddings were created using pre-trained transformer models. It is conceivable that fine-tuning the transformers could have produced superior results (Devlin et al., 2019; Sun et al., 2019). Additionally, the features used to represent each judgment were not derived from any specific portion of the judgment text. Working with only particular sections of the judgment such as the facts of the case might have enhanced prediction accuracy. Lastly, the classifier was trained using just 400 judgments labelled by a single domain expert. Using multiple people to label the same number of judgments might have reduced error (see further, Aslan et al., 2017; Muller et al., 2021). Alternatively, using a greater number of human-labelled judgments for model training would likely have improved the model’s predictions (and increased the proportion of judgments in the repository that were human-classified).

While one or more of these tasks could be considered necessary for the creation of a conclusive repository of judgments, each was beyond the scope of this initial repository creation project. Tuning sentence encoders and extracting sections of the text from judgments on BAILII are both more complex than the work presented here. With regards to the latter, the (inconsistent) structure of BAILII judgments obstructs the division of judgments into distinct sections (cf. Medvedeva et al., 2020). Further, additional labelling would have added time and potentially cost to the project. Each recognised limitation is therefore merely considered a pointer towards potential future research.

Indeed, we remain confident in the ML system on which our judgment repository was partially built. This performed significantly better than most alternate systems

[16] Any findings added to the repository following subsequent application of our selected model to post-2020 judgments would be clearly labelled to ensure that prospective repository users are aware that predictions could be less accurate.

[17] To investigate this possibility, the selected method was applied to the 53,565 judgments that did not contain ‘animal’ (of 55,202 judgments total). Our model predicted 1.2 percent of these judgments (688 judgments) to concern animal protection law. A random sample of 25 of these judgments, plus 25 that were predicted as the negative class, were sent to the domain expert for feedback. This feedback showed that judgments predicted as the negative class typically had nothing to do with animals in any sense. Additionally, while the judgments predicted as positive often considered environmental concerns, none was found to be substantially concerned with the welfare or protection of animals.

and a baseline measure intended to reflect current searching practice (Sect. 3.4). It was also constructed in a manner that permitted investigation into influential features. This investigation suggested many features to make rational contributions to judgment classification (Sect. 3.5.1). Amongst the (rational) negative predictors was the term 'jury', which stimulated important discussion of the backward-looking nature of our model. This consideration of influential terms highlighted the benefits of using ML systems that permit some level of human understanding.

5 Conclusion

Using animal protection law as a case study, this paper has shown that ML can be employed to create a worthwhile judgment repository concerning a new practice area. To achieve this, we outlined a judgment repository creation process that began with the identification of 1637 judgments on BAILII containing 'animal' made by the Privy Council, House of Lords, Supreme Court and upper England and Wales courts between January 2000 and December 2020. This amount contrasts with BAILII's own search tool, which only identified 1568 judgments. 500 of the judgments containing 'animal' were labelled by a domain expert and used to train and validate ten ML systems. The best performing system confirmed the merits of using NLP and ML for judgment classification by achieving a macro-F1 score of 0.87 and accuracy of 0.93 on a test set of 100 judgments. This system was used to classify the remaining (unlabelled) judgments, giving a repository of 175 animal protection law judgments of which 92 were found automatically. Preliminary examination of the repository suggests it could aid the identification of individual animal protection law judgments and enhance understanding of the breadth of animal protection law created by courts.

Appendix 1

Appendix 1: Incorrectly predicted judgments and qualitative feedback
(See Table 7).

Table 7 Incorrectly predicted judgments and qualitative feedback

Citation	Year	Link	Predicted classification	Human classification	Qualitative feedback
R (on the application of Quintavalle) v. Secretary of State for Health (2003) UKHL 13	2003	https://www.bailii.org/uk/cases/UKHL/2003/13.html	0	1	This is a very marginal judgment concerned primarily with a statute passed for the protection of live human embryos created outside the human body. The judgment confirmed that 'Parliament outlawed certain grotesque possibilities (such as placing a live animal embryo in a woman or a live human embryo in an animal)'. Interpretation of the provisions relating to animals and animal embryos arguably affects the protection of one or more animals. It is, however, debatable as to whether animal protection was a central matter in this judgment
First Corporate Shipping Ltd t/a Bristol Port Company v. North Somerset Council (2001) EWHC Admin 586	2001	https://www.bailii.org/ew/cases/EWHC/Admin/2001/586.html	1	0	This is a planning-focused judgment about a port which was an important animal feed terminal. There are references to wildlife when considering some areas of environmental significance near the port. Still, while this judgment touches on some animal protection issues, it is neither substantially concerned with the protection of areas of environmental significance nor the wildlife within them

Table 7 (continued)

Citation	Year	Link	Predicted classification	Human classification	Qualitative feedback
R (on the application of Doe) v. Secretary of State for Transport (2002) EWHC Admin 2269	2002	https://www.bailii.org/ew/cases/EWHC/Admin/2002/2269.html	0	1	This is a judgment with direct relevance to animal protection law, which concerns planning permission for a game keeper's mobile home. There are multiple criteria of relevance to obtaining planning permission for such homes, including whether the erection of the home could permit the delivery of essential care to animals at short notice. Additionally, there is some discussion about animal theft and the welfare and safety of the birds
Banks v. Secretary of State for Environment, Food & Rural Affairs (2004) EWHC Admin 416	2004	https://www.bailii.org/ew/cases/EWHC/Admin/2004/416.html	0	1	This is a marginal judgment concerning regulations around bovine tuberculosis and the movement of cattle. Arguments could be made for both the exclusion or inclusion of this judgment within animal protection law. This judgment was not explicitly about the protection of the animals, which were primarily viewed in financial terms by the applicant farmer. However, the issues addressed have clear implications for animal protection. Indeed, the notice being challenged severely disrupted farming activities, preventing the slaughter of the cattle for human consumption

Table 7 (continued)

Citation	Year	Link	Predicted classification	Human classification	Qualitative feedback
Trailer & Marina (Leven) Ltd v. Secretary of State for Environment, Food & Rural Affairs (2004) EWHC Admin 153	2004	https://www.bailii.org/ew/cases/EWHC/Admin/2004/153.html	0	1	This judgment centres on a challenge by a canal owner to a DEFRA decision to designate a particular area as a Site of Special Scientific Interest. As such, this judgment has important implications for the protection of animals and should therefore be classified as animal protection law
Sienkiewicz v. South Somerset District Council (2015) EWHC Admin 3704	2015	https://www.bailii.org/ew/cases/EWHC/Admin/2015/3704.html	1	0	This judgment should be classified as not concerned with animal protection, as it does not go into meaningful discussion of animal protection issues. It does, however, mention 'animal welfare products' when referring to the products sold by the company involved. There are also some mentions of wildlife because the judgment involves an Environmental Impact Assessment

Table 7 (continued)

Citation	Year	Link	Predicted classification	Human classification	Qualitative feedback
Barkhuysen v. Hamilton (2016), EWHC QB 2858	2016	https://www.bailii.org/ew/cases/EWHC/QB/2016/2858.html	0	1	This is a marginal judgment concerning allegations of animal cruelty and consequent action for slander. While this judgment was classified as animal protection law, an equally strong argument could be made that this judgment should not be classified otherwise. The allegations in question were almost certainly false, meaning that there might not have been a real issue about the *protection* of the animals in question. Given that the judgment considered slander against a person purporting to be concerned about animals, it is, however, clearly concerned with animal law in a broader sense

Funding This research was not funded. All authors provided their time free-of-charge and all data is publicly available on the British and Irish Legal Information Institute (BAILII: https://www.bailii.org/).

Data availability The python scripts used for data collection and analysis are available in an open-access GitHub repository (https://github.com/JoeMarkWatson/animal_law_classifier). The data for analysis was obtained (through 'scrape_label.py') from publicly available judgments hosted by the British and Irish Legal Information Institute (BAILII, https://www.bailii.org/). The labelling guidance used for judgment labelling ('animal_protection_law_labelling_guidance.docx') and judgment repository ultimately produced through this research ('case_law_repository.csv') are also accessible via the main repository.

Code availability The python scripts used for data collection and analysis are available in an open-access GitHub repository (https://github.com/JoeMarkWatson/animal_law_classifier). Data was collected through one file ('scrape_label.py') and analysed using another ('train_classify2.py').

Declarations

Conflicts of interest The authors declare that they have no conflict of interests. There is and has been no financial relationship between any author and any organisation of relevance to this work. Further, no author is currently in any negotiations regarding future paid employment with any organisation of relevance. Samuel March (the third author and domain expert in this research) works as a part-time paralegal (volunteer) at Advocates for Animals, having held this position since April 2020. Joe Watson (the first author) also works as a part-time paralegal (volunteer) at Advocates for Animals and began this position in December 2020. The voluntary positions held by Samuel March and Joe Watson are provided for reasons of transparency, yet Advocates for Animals has had no influence on the contents of this paper.

Ethical approval Numerous ethical considerations were taken into account. The manuscript has not been submitted or published anywhere else, nor will it be submitted elsewhere until completion of the editorial process. We provide access to our code and information on where the data underlying our research is available. Additionally, all authors have approved the manuscript for submission and consent to publication should this submission be successful. This research did not directly involve any human or animal participants. All humans (aside from the domain expert, who is also the third author) and animals mentioned in the document text are present in court judgments publicly available on BAILII.

Consent to participate No humans or animals beyond the listed authors participated directly in this research. As such, consent to participate was not sought.

Consent for publication No humans or animals beyond the listed authors participated directly in this research, and all humans (aside from the domain expert) and animals mentioned in the document text are present in court judgments publicly available on BAILII. Consent to publish was therefore not sought.

References

Articles, books, conference papers, Python libraries and web pages

Advocates for Animals, Advocates for Animals. https://advocates-for-animals.com/. Accessed 11 Jul 2021

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

Asgari-Chenaghlu M, Nikzad-Khasmakhi N, Minaee S (2020) Covid-transformer: detecting COVID-19 trending topics on twitter using universal sentence encoder. arXiv:200903947 [cs]

Ashley KD (2017) Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge University Press, Cambridge

Aslan S, Mete SE, Okur E et al (2017) Human expert labeling process (HELP): towards a reliable higher-order user state labeling process and tool to assess student engagement. Educ Technol 57:53–59

BAILII, About BAILII. https://www.bailii.org/bailii/. Accessed 13 Jul 2021

Beltagy I, Peters ME, Cohan A (2020) Longformer: The long-document transformer. arXiv:200405150

Bhambhoria R, Dahan S, Zhu X (2021) Investigating the state-of-the-art performance and explainability of legal judgment prediction. In: Proceedings of the Canadian conference on artificial intelligence. Vancouver, Canada

Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory. Association for Computing Machinery, New York, NY, USA, pp 144–152

Burri T (2017) Machine Learning and the Law: Five Theses. Paper accepted at NIPS 2016. Barcelona, Spain, pp 1–4. https://doi.org/10.2139/ssrn.2927625

Cer D, Yang Y, Kong S, et al (2018) Universal Sentence Encoder. arXiv:180311175 [cs]

Chollet F (2015) Keras. https://github.com/fchollet/keras

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018

de Araujo PHL, de Campos TE, Braz FA, da Silva NC (2020) VICTOR: a dataset for Brazilian legal documents classification. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France, pp 1449–1458

Deakin S, Markou C (forthcoming) Evolutionary interpretation: law and machine learning. J Cross-Discipl Res Comput Law. https://doi.org/10.2139/ssrn.3732115

Devlin J, Chang M-W, Lee K, Toutanova K (2019) 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intell Data Anal 6:429–449

Jones KS (1972) A statistical interpretation of term specificity and its application in retrieval. J Doc 28:11–21

Kingma DP, Ba J (2017) Adam: A Method for Stochastic Optimization. arXiv:14126980 [cs]

Lehr D, Ohm P (2017) Playing with the data: what legal scholars should learn about machine learning. UCDL Rev 51:653

Lei M, Ge J, Li Z et al (2017) Automatically Classify Chinese Judgment Documents Utilizing Machine Learning Algorithms. In: Bao Z, Trajcevski G, Chang L, Hua W (eds) Database Systems for Advanced Applications. Springer International Publishing, Cham, Switzerland, pp 3–17

Maaten L, Chen M, Tyree S, Weinberger K (2013) Learning with marginalized corrupted features. In: International Conference on Machine Learning. PMLR, Atlanta, USA, pp 410–418

Mandal A, Ghosh K, Ghosh S, Mandal S (2021) Unsupervised approaches for measuring textual similarity between legal court case reports. Artif Intell Law. https://doi.org/10.1007/s10506-020-09280-2

Markou C, Deakin S (2020) Ex Machina Lex: Exploring the Limits of Legal Computability. Is Law Computable? Critical Perspectives on Law and Artificial Intelligence. Hart Publishing, Oxford, United Kingdom, pp 31–66

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

Merriam TV, Matthews RA (1994) Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. Lit Linguist Comput 9:1–6

Muller M, Wolf CT, Andres J, et al (2021) Designing Ground Truth and the Social Life of Labels. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–16

Nay J (2018) Natural language processing and machine learning for law and policy texts. Soc Sci Res Netw, Rochester, NY. https://doi.org/10.2139/ssrn.3438276

Overcash EA (2012) Unwarranted discrepancies in the advancement of animal law: the growing disparity in protection between companion animals and agricultural animals comment. NC L Rev 90:837–883

Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830

Peters A (2020) Introduction. In: Studies in Global Animal Law. Springer Nature, Berlin, Germany, pp 1–16

Rachlinski JJ, Wistrich AJ (2017) Judging the judiciary by the numbers: Empirical research on judges. Ann Rev Law Soc Sci 13:203–229

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:190810084 [cs]

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536. https://doi.org/10.1038/323533a0

Šavelka J, Trivedi G, Ashley KD (2015) Applying an interactive machine learning approach to statutory analysis. In: Proceedings of the 28th Annual Conference on Legal Knowledge and Information Systems (JURIX 2015). IOS Press, Minho, Portugal

Slovikovskaya V, Attardi G (2020) Transfer learning from transformers to fake news challenge stance detection (FNC-1) Task. In: Proceedings of the 12th Conference on Language Resources and Evaluation. Marseille, France

Song D, Vold A, Madan K, Schilder F (2021) Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Inf Syst. https://doi.org/10.1016/j.is.2021.101718

Srivastava N, Hinton G, Krizhevsky A et al (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958

Sulea O-M, Zampieri M, Malmasi S, et al (2017b) Exploring the use of text classification in the legal domain. arXiv preprint arXiv:171009306

Sulea O-M, Zampieri M, Vela M, Van Genabith J (2017a) Predicting the law area and decisions of french supreme court cases. arXiv preprint arXiv:170801681

Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune BERT for text classification? China National Conference on Chinese Computational Linguistics. Springer, Cham, Switzerland, pp 194–206

Thomson Reuters, AI Timeline. https://www.thomsonreuters.com/en/artificial-intelligence/ai-timeline.html. Accessed 13 Jul 2021

Undavia S, Meyers A, Ortega JE (2018) A comparative study of classifying legal documents with neural networks. 2018 Federated conference on computer science and information systems (FedCSIS). Poznań, Poland, pp 515–522

Vardhan H, Surana N, Tripathy BK (2020) Named-entity recognition for legal documents. International conference on advanced machine learning technologies and applications. Springer, Cham, Switzerland, pp 469–479

Verma R, Shinde K, Arora H, Ghosal T (2021) Attend to Your Review: A Deep Neural Network to Extract Aspects from Peer Reviews. In: Proceedings of the International Conference on Neural Information Processing Conference. Springer, pp 761–768

Wenzel C, Baumann S, Jäger T (1998) Advances in document classification by voting of competitive approaches. Document Analysis Systems II. World Scientific, Singapore, pp 385–405

Westermann H, Šavelka J, Walker VR et al (2020) Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents. Legal Knowl Inf Syst 334:164–173. https://doi.org/10.3233/FAIA200860

Yang Y, Abrego GH, Yuan S, et al (2019) Improving Multilingual Sentence Embedding using Bi-directional Dual Encoder with Additive Margin Softmax. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence

Zaheer M, Guruganesh G, Dubey KA, et al (2020) Big Bird: Transformers for Longer Sequences. In: Larochelle H, Ranzato M, Hadsell R, et al. (eds) Proceedings of the Advances in Neural Information Processing Systems Conference. Curran Associates, Inc., pp 17283–17297

Judgments

Banks v. Secretary of State for Environment, Food & Rural Affairs (2004) EWHC Admin 416. https://www.bailii.org/ew/cases/EWHC/Admin/2004/416.html

Barkhuysen v. Hamilton (2016), EWHC QB 2858. https://www.bailii.org/ew/cases/EWHC/QB/2016/2858.html

European Brand Trading Ltd v. HM Revenue and Customs (2016), EWCA Civ 90. https://www.bailii.org/ew/cases/EWCA/Civ/2016/90.html

First Corporate Shipping Ltd t/a Bristol Port Company v. North Somerset Council (2001) EWHC Admin 586. https://www.bailii.org/ew/cases/EWHC/Admin/2001/586.html

R v. Sissen (2000), EWCA Crim 67. https://www.bailii.org/ew/cases/EWCA/Crim/2000/67.html

R (on the application of Aggregate Industries UK Ltd) v. English Nature (2002) EWHC Admin 908. https://www.bailii.org/ew/cases/EWHC/Admin/2002/908.html

R (on the application of Bizzy B Management Ltd) v. Stockton-On-Tees Borough Council (2011) EWHC Admin 2325. https://www.bailii.org/ew/cases/EWHC/Admin/2011/2325.html

R (on the application of Doe) v. Secretary of State for Transport (2002) EWHC Admin 2269. https://www.bailii.org/ew/cases/EWHC/Admin/2002/2269.html

R (on the application of Highbury Poultry Farm Produce Ltd) v. Crown Prosecution Service (2020) UKSC 39. https://www.bailii.org/uk/cases/UKSC/2020/39.html

R (on the application of National Farmers Union) v. Secretary of State for Environment, Food and Rural Affairs (2020) EWHC Admin 1192. https://www.bailii.org/ew/cases/EWHC/Admin/2020/1192.html

R (on the application of Quintavalle) v. Secretary of State for Health (2003) UKHL 13. https://www.bailii.org/uk/cases/UKHL/2003/13.html

R (on the application of Sainsbury's Supermarkets Ltd) v. The Independent Reviewer of Advertising Standards Authority Adjudications (2014) EWHC Admin 3680. https://www.bailii.org/ew/cases/EWHC/Admin/2014/3680.html

Sienkiewicz v. South Somerset District Council (2015) EWHC Admin 3704. https://www.bailii.org/ew/cases/EWHC/Admin/2015/3704.html

Trailer & Marina (Leven) Ltd v. Secretary of State for Environment, Food & Rural Affairs (2004) EWHC Admin 153. https://www.bailii.org/ew/cases/EWHC/Admin/2004/153.html

Legislation

Animal Welfare (Sentencing) Act 2021. c. 21 https://www.legislation.gov.uk/ukpga/2021/21/contents/enacted

Customs and Excise Management Act 1979. c. 2 https://www.legislation.gov.uk/ukpga/1979/2/contents

Endangered Species (Import and Export) Act 1976. c. 72 https://www.legislation.gov.uk/ukpga/1976/72

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Joe Watson[1] · **Guy Aglionby**[2] · **Samuel March**[3]

Guy Aglionby
guy.aglionby@cl.cam.ac.uk

Samuel March
paralegal@advocates-for-animals.com

[1] The Psychometrics Centre, Judge Business School, University of Cambridge, Cambridge, UK

[2] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

[3] Advocates for Animals, London, UK