Using Deep Learning to Automate Discovery of Deontic Logic Statements in Unstructured Text Data

Cooper M. Stansbury
College of Engineering and Computer Science
University of Michigan Dearborn
Dearborn, MI
Email: cstansbu@umich.edu

***Abstract*—Current biomedical information systems do not represent individual donor permissions specified during informed consent. Researchers aiming to access biospecimens or clinical data based on specific conditions of use are not able to query available records subject to the ethical constraints imposed by deontic acts prior to storage. This study shows an effective method for generalizing the latent structure of 'permissions' that may be used to facilitate greater, more ethical access to valuable biospecimens or clinical data.**

1. Introduction

Access to biospecimens and clinical data are critical to advancing medical science. The ethical and legal conditions of collection, distribution and future use of specimens and data are determined by potential donors during a process of informed consent. The conditions of use, i.e. 'permissions,' may be highly variable and context dependent. Record of permissions are often stored in information systems that cannot accurately represent the choices made by the individual donor at the time of consent [1], [2]. Hard copies of informed consent forms may be the only ground truth with granular fidelity to a donor's specifications. Physical archives of consent forms may not be regulated, making forms difficult or impossible to locate, and easily damaged or lost. Further, current information systems that support clinical data and samples are ill-equipped to handle the variability in permissions for future use, material transfer, or data transfer at the specimen level [2].

Ethical and legal review is dependent on manual human-interpretation of a complex set of intertwined policies and regulations which is both time consuming and error prone [1], [2], [3], [6]. Locating available data or samples is not scalable, and specific conditions of use are not easily available to researchers attempting to advance their fields [1], [2]. Recent efforts to digitize the informed consent process and promote controlled access to medical information systems and biorepositories have resulted in centralized access points that integrate heterogeneous networks of samples and data. While this is a promising direction for the future of ethical access, there are still large quantities of biomatter and data whose use is governed by paper documents.

Medical institutions create proprietary informed consent forms for a variety of medical purposes to meet local policy and procedure needs. Institutions conducting research involving human subjects are required to meet criteria prescribed by The Common Rule [4] and institutions handling sensitive personal data must adhere to the Health Insurance Portability and Accountability Act of 1996 [5]. Complex and sometimes conflicting regulatory landscapes results in a high degree of heterogeneity in document format, legal language, and jurisdictional requirements [3], [6]. The result is a highly variable set of unstructured text documents covering a range of medical procedures and research studies that have two fundamental goals. First, consent forms serve to inform the potential donor about the clinical procedure or study, future use of data and specimens collected, and the risks and benefits associated with the procedure or sample collection. Second, consent forms serve as a record of explicit donor consent to the actions described in the form [4].

There is a need to explore the possibility of automating the discovery of statements of explicit donor consent from informed consent documents. Statements of 'permission' are highly contextually sensitive and often conditional in nature, making them difficult to isolate and study. However, informed consent forms serve the explicit purpose of capturing a set of complex permissions, and thus offer a interesting opportunity to study the latent structure of a class of deontic concepts using current computational methods from natural language processing (NLP), artificial intelligence (AI) and machine learning.

The purpose of this study is to examine the possibility discovering statements of permission in informed consent documents. Permissions, once identified, may be categorized further using multi-class classification models or using knowledge-based techniques and logical inference. While the scope of this work is limited to the use of computational techniques for discovery, it is necessary to understand the landscape of knowledge management that makes automating discovery difficult and necessary. This work is part of a larger effort to define semantically aware information systems that can integrate heterogeneous data and represent permissions at the specimen or data-record level in order to to address the current shortcomings in management of consents. Figure 1 shows a simplified overview of a proposed

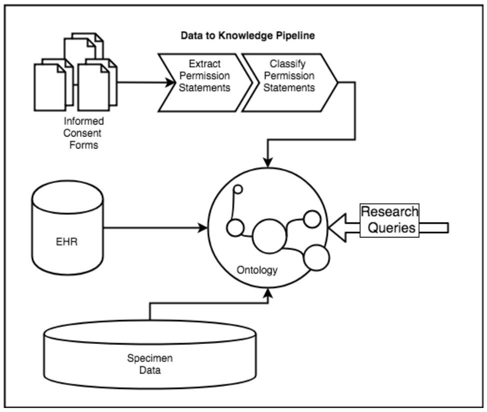

Figure 1: Proposed information architecture for facilitating research queries subject to the constraints of specific conditions of use.

system that may allow for ethical access to stored samples and data.

This study will assume a donor-centric perspective of permissions. Although deontic concepts follow in both directions during a process of informed consent, the goal is to discover what a donor is permitting in order to better protect a donor's choices and to make those choices known to researchers seeking samples or data. By using modern computational techniques in a semantically ambiguous space the author aims to show that it may be possible to facilitate informed retrieval of samples and data subject to specific constraints, i.e., for which certain permissions have been granted by a donor.

1.1. Background

There are three research areas that contribute to the possibility of automating discovery and management of permissions: (1) formal deontic logic, (2) computational semantics, and (3) machine learning.

1.1.1. Computable Deontic Logic. 'Permission' is notoriously difficult to define. Deontic logic is the philosophical field of study relating to idealized and actual states of human affairs [9]. Deontic logic is important to this study because it provides a basis for understanding computable notions of 'permission.' One such formalization of 'permission' treats 'permissability' as an attribute of an act, i.e., a given act can be: 'permitted' or 'forbidden' [7], [8], [9]. It should be noted that there are known issues with formal deontic logic. Notably, the difference between a statement that prescribes a permission and a description of the same permission may be entirely context dependent [9]; the statements themselves may be otherwise identical. The takeaway is that there are no standard, agreed upon rigorous definitions of permission that can be made computational and operational without causing violence. Still, deontic logic provides insight for the researcher attempting to train a computer to understand the difference between *describing* and *prescribing*. Much of the early work in computational deontic logic and legal automation is nicely summarized by J.J. Meyer before the turn of the century [7], [8].

I will adapt one over-simplified definition from the Stanford Encyclopedia of Philosophy as a working definition of a 'permissable action':

> *"[some action X] is permissible iff its negation is not obligatory"* [9].

This negative definition requires understanding 'permissable actions' against a set of normative obligations, for example the complex regulatory landscape that surrounds biospecimen sharing and data use. The implication of this definition is that it is not sufficient to understand the conditions of use prescribed by an informed consent form at the character-level because the permissability of a particular action depends on the regulatory context. Detecting permissions as they relate to the broader regulatory context is out of scope for the current study, but is an interesting and necessary implication of automated detection. Formalizing notions of deontic concepts is the first important step towards automating decisions and facilitating information retrieval in larger information systems.

1.1.2. Semantic Technologies. The informatics community has had some success developing semantically aware information architecture. For example, the Open Biological and Biomedical Ontology (OBO) Foundry is a community of scientists working to define universals and their relations in specific domains using a standard upper-level model of reality [10], [11], [12], [13]. Work to represent deontic concepts [13] may help to enable rich inferences should instances of permissions be discoverable. From a knowledge management perspective, information structures like ontologies that are written in OWL enable semantic queries (using SPARQL against triple-stores, for example) against sets of classes, instances of those classes, and the relations between them. Ontologies such as the Informed Consent Ontology [12], the Data Use Ontology [16], and the Document Act Ontology [13] may be critical in modelling complex notions of permission and facilitating interoperability between datastores and biobanks.

Significant efforts in developing controlled vocabularies of concepts surrounding informed consent are ongoing on multiple fronts [14], [15], [16], [17], suggesting not only the difficulty in standard definitions of 'permissions,' but also the need to make them computable. Taken together, it may be possible to leverage existing knowledge representing permission to increase the performance of computational discovery. Ontologies have shown to be useful in aiding in NLP tasks [18], [19], [20]. This possibility is not explored here. Recent efforts regarding knowledge representation of informed consent are an important step towards enabling ethical access to available biospecimens, but they leave the challenge of understanding past records of consent unsolved.

1.1.3. Natural Language Processing. Controlled vocabularies help us manage knowledge, but they don't aid it knowledge discovery from unlabeled, unstructured data. The primary aim of this study is to train a learning algorithm to classify textual data with respect to the permissions it prescribes. Traditional approaches to natural language processing tasks typically involve representing textual data numerically and calculating probabilities of occurrence based on large joint probability distributions [21], [22]. A famous publication argues that frequency counts, when large enough, can outperform sophisticated learning approaches [23]. However, these joint probability approaches are made difficult by domains with highly specialized vocabularies, such as medical or legal domains [24]. Further, word order is decidedly important when considering conditional statements, and many 'bag-of-words' approaches do not explicitly represent the syntactic structure of the text.

Many challenges such as unsupervised or semi-supervised machine translation and structured document generation have only recently shown significant progress through the use of attention models [25], [26], [27]. This study will explore the possibility of using modern machine learning techniques to leverage computational power in an ambiguous semantic space. The specifics of the training algorithms will be left for section 2.2 of this document.

Previous work to classify deontic statements use ontologies to support rule-based analysis after sentence parsing [28], [29]. The approach taken here is primarily data-driven, using NLP techniques to learn the latent structure of permissions in unstructured task for ingestion into an ontology. However, Governatori et al. have shown there is promise in using knowledge encoded in ontologies to identify deontic statements.

1.2. Definition of Permission

One of the primary challenges in automating the discovery of deontic concepts in text data is the difficulty in defining these concepts in the first place. Concrete examples of permission are needed in order to train machine learning algorithms. To the author's knowledge, there exist no standard training corpus for language tasks in this domain. The problem space is made complex by ambiguous notion of 'permission' and the highly variable semantic space which encloses permissions. Take for example the circular definitions of 'permission' in WordNet below. Further definitions of these terms can be viewed in Appendix A.

1) ***"permission*** *(2) : the act of giving a formal (usually written) authorization"* [30].
2) ***"authorization*** *(3) : official permission or approval"* [30].

Challenges in definition pose downstream challenges in recognition. Take for example the nuanced differences between the following synthetic sentences:

1) *I understand that there is a chance that my personal data may be shared with researchers at other universities.*
2) *I understand that there is a chance that my personal data may be accessed by unauthorized parties.*
3) *I understand that there is a chance that my personal data may be accessed by people with no affiliation to the university.*

The first sentence is often construed as a permission while the second is sometimes viewed as a description of risk. The third sentence is decidedly ambiguous without further context. It is uncertain if this sentence prescribes the researchers permission to share data with commercial entities or if it is a description of risk that malicious outsiders may gain access to unauthorized information.

It may not be possible to define 'permission' computationally without doing violence. Still, in order to see progress in this space a working definition is needed. A working definition of 'permission' is: *'a potential state that is realized when acts are performed by some agent, either an organization or an individual'*. Colloquially, one might say that a permission is an 'action, or set of actions that may be allowably performed by some agent.' This study is not concerned with the potential state of affairs itself, but the textual data that *prescribes* this state, should it participate in an actual process of informed consent. The notion of 'prescribes' is important to understanding the precise entities we wish to discover. The formal definition of the relation *prescribes* from the Common Core Ontologies:

> *"For all types T_1 and T_2, if T_1 prescribes T_2, then there is some instance of T_1, t_1, that serves as a rule or guide to some instance of T_2, t_2 ... or that serves as a model for some instance of T_2, t_2"* [31]

Using this definition, we might say that there is an instance of some textual data (t_1) that prescribes some potential state (t_2) that may be realized in actions that are performed by some agent. What we are interested in is discovering instances of type T_1 in unstructured text, while treading lightly on the complex deontic concepts that come to bear on t_1 in practice.

Definition: *'Statement of Permission:'* A textual instance of t_1 that prescribes a potential state t_2 of an action whose negation is not obligatory [9].

Though this definition is singular, it is appropriate to treat 'action' as a set of actions. Future work may also address the agents who participate the the actions, as they are critical to understanding the limits of permitted actions.

1.3. Problem Definition

The purpose of this study is to compare different classification models in attempt to gain insight concerning the latent structure of permissions in the domain of informed consent. The study has two primary aims:

1) Identify candidate statements of permission from the donor perspective to train text classification models.
2) Train a function $F'(\bar{\mathbf{X}})$ to predict binary classification ($t_1; !t_1$) on unlabeled of sentences based on a feature vector $\bar{\mathbf{X}}$.

The scope of this study is restricted to textual analysis for information retrieval. This study represents a first step towards building a system that can facilitate scientific queries subject to the constraints of individually specified permissions prescribed by informed consent forms. Such a system would help to integrate data from consent forms, Electronic Health Records, and biorepositories that manage biospecimens as previously shown in Figure 1.

It should be noted that treating textual data as mutually exclusive instances of $(t_1; !t_1)$ is problematic for understanding statements subject to a continuous notion of 'force of authority.' This choice is justified by the use of probabilistic classifiers, where a decision point c must be specified by a human in order to determine the predicted class.

1.3.1. Research Question. Is it possible to use machine learning to automate the discovery of statements of permission in informed consent forms? Specifically, can we avoid overt rule-based techniques to achieve reasonable performance in a complex linguistic domain? Can we train machine learning algorithms to understand the latent structure of a permission?

2. Methods

Obtaining training data needed for a learning task in the biomedical domain is a daunting task. Due to the sensitive nature of informed consent forms and potential liability issues, intuitions are generally reluctant or unwilling to share archived legal forms of any kind. Since 2015 a small research team at the University of Michigan, Ann Arbor began preliminary collection of publicly available informed consent forms from web sources. It should be noted that consent documents may be subject to sampling bias by virtue of their availability, which is a known limitation of the current study. Efforts are currently underway to strengthen the quality of training data, but acquiring training data is a persistent challenge.

The consent form data can be broadly classified into two categories: (1) informed consent forms for research studies and (2) informed consent forms for clinical procedures. At the time of this writing there are 752 (285 research, 467 clinical) consent forms in total, including template forms and a number of near and exact duplicates. Research consent forms require adherence to The Common Rule, and IRB and are often specific to a particular study. Clinical consent forms, on the other-hand, are used to grant permission for clinical procedures and may contain provisions concerning the use of excess tissue for research purposes as embedded clauses. This diversity has important semantic implications, as the language used to specify permissions (even for the same action) is governed by different regulatory contexts. The choice to include both types of consent was made to first see if a generalizeable model can be trained. The preliminary hypothesis is that the semantic or syntactic structure of permissions may be consistent, even in domains such as software terms of use.

Additionally, it has been shown that it is possible to reliably distinguish between documents of each category using a small set of seed word counts and simple classifiers[1]. For this study, I choose to include templates because they provide recommended deontic language, and may be helpful while the sample size remains small. This 'boilerplate' language is for institutions to use when crafting new consent forms specific to a study, typically with a focus on adhering to the requirements prescribed by The Common Rule. Further, because many researchers use boilerplate' language it is reasonable to expect a number of near or exact duplicate sentences in consent forms from different intuitions. Indeed, of the 752 uniquely named documents on 698 were unique at the character level. I chose to work only with unique documents for this study, but choose to include duplicate sentences as this is representative of the general population.

All forms are unsigned and can only serve as a data set of instances which prescribe permissions rather than records of historical permissions. Most forms made publicly available are in *.pdf* format and require conversion using open-source OCR tools in order to facilitate textual data analysis. An unfortunate consequence of this conversion process is information loss w.r.t check-boxes and other non-character structural components of the informed consent forms. Still, format conversion has other important implications beyond data quality and information loss. Because paper consent forms are often the primary documentation method inside medical establishments, the ability to work with low-quality textual data may have important implications as institutions attempt to create and join federated research networks in order to make historical data and samples available.

2.0.1. Unit of Analysis. Choosing the unit of analysis has important implications for this study. First, the description of actions whose potential state may be altered need not exist near the prescribing of permission themselves. In other words, permissions do not depend on structural proximity of their conditions within a text document. Take for example the clause: *"I consent to all procedures described in the document."* The specific procedures, or acts, whose state is prescribed as 'permissable' during an affirmative completion of consent may be uniformly distributed throughout the consent form. Further, there may be descriptions of action that do not bare on the the process of consent. Second, statements of permissions are often conditionals. The conditions of consent are sometimes specified separately from the actions or the 'prescribing.' While taking any unit smaller than the whole document eliminates the possibility of studying these important phenomenon, semantic relationship disambiguation has shown to be problematic [32], [33] especially across sentence boundaries [34].

Given the challenges inherent in identifying deontic statements in unstructured text, I chose to work with sentences as the unit of analysis. While missing location insensitive characteristics, sentences often contain the full

1. Earlier work on form classification can be found at the following URL: `https://github.com/CooperStansbury/document_commandline_tools`

conditions, actions, and language prescribing permission. Future models may make use of named entity recognition (NER) techniques to capture more granular character-level information related to the boundaries of permissions. NER may also prove useful to classifying other entities related to the permission that may be useful in classification of permissions. However, full sentences tend to be human-readable, which allows for human evaluation of results which is important for tasks without a gold-standard. Further, classification tasks output a probability with a prediction that can be meaningfully interpreted in this context, especially because the implications of a mistake can have dire consequences.

2.1. Candidate Identification

The first aim of this study is to develop a method for identifying candidate sentences with a high probability of belonging to class T_1. I use a set of seed words and semantic similarity measures to reduce the number of sentences that require human annotation in two passes. The first-pass identifies candidate sentences based on the explicit inclusion of *any* seed word. The second-pass includes all sentences above a predetermined cosine-similarity threshold. The intuition behind identifying candidates is that there are a finite number of words used to explicitly signal statements of permission. Working with domain experts I identified a list of 11 seed words that, if present, would yield high probability that the statement is an instance of T_1. This list was expanded using WordNet synsets based on the $lemmas$ of each lexical 'clue' identified. This resulted in 65 total new lexical clues. By manually revising this list with domain experts I identified 26 tokens likely to be included in statements of permission (see Appendix A for full list and definitions). The use of general purpose lexical resources (WordNet) for this task is appropriate because many IRBs require language is that it is sufficiently clear at an 8th grade reading level [35].

There are two important notes regarding the use of lexical resources. First, while many of these seed terms may be associated with a deontic statement, not every use of these terms will prescribe a permission. The use of seed words should result a large number of false positives, which is consistent with the findings. Because the first-pass candidate extraction is naïve, word sense or context is not considered. Second, although each seed word has associated metadata, I used only the string for pattern-matching. While the risk of incomplete seed specification is mitigated by the use of semantic similarity measures (second-pass), only exact matches will be included in the first-pass. Future work may be directed towards automating the curation of seed words for example, using an ontology. Future models will aim to reduce, and eventually eliminate dependence on hand-picked lexical resources. But at this time seed words are critical to building a labeled corpus.

By parallelizing spAcy's built-in tokenization, lemmatization, and default rule-based sentence parsing functionality each raw text file was converted into a list of tokenized sentences [36], [37]. Empirically, rule-based sentence parsing required preliminary data preprocessing such as normalizing white-space and forcing conversion to ASCII character sets in order to yield human readable sentences. Sentences were stripped if they were shorter than a specified character threshold ("I consent") in order reduce noise during human evaluation. For each document d in the corpus of 698 consent a sublist of candidates is created. Algorithm 1 shows recovery of sentences containing any of the lexical seeds.

ALGORITHM 1
First-Pass Candidate Identification

$M_d \leftarrow$ document corpus with d documents
$V \leftarrow$ vocabulary
$C \leftarrow$ [empty list]

for all $document$ in M_d **do**
  for all $sentence$ in $document$ **do**
    if $sentence.containsAny(V.item())$ **then**
    $C \leftarrow append(sentence);$ **break**
    end if
  end for
end for
return $C \leftarrow$C.unique()

In order to mitigate potential bias in the candidate selection process and to reduce the risk of false negatives a second-pass is performed to include sentences above a .95 threshold to those already identified by simple string-pattern methods in the first-pass. Pair-wise similarity is computed at the entire sentence level using GloVe vectors and cosine similarity [37], [39]. Because both passes operate at the single document level it is possible to parallelize each operation, which will carry important implications as the number of input documents scales.

ALGORITHM 2
Second-Pass Similar Candidates

$M_d \leftarrow$ document corpus with d documents
$C \leftarrow$ candidate sentences

for all $document$ in M_d **do**
  for all $sentence$ in $document$ **do**
    for all $candidate$ in C **do**
      if $sentence.similarity(candidate) > .95$ **then**
      $C \leftarrow append(candidate);$ **break**
      end if
    end for
  end for
end for
return $C \leftarrow$C.unique()

The result of both passes is a list of candidate sentences was a reduction of the original space of 65,595 sentences to 24,698 identified as candidates, or 0.376 of the total from all documents. Of these, 18,670 sentences were unique. Because language is often standard across forms, duplicates were not removed from the final list of candidates. The number of candidates is intentionally artificially high in order to mitigate the risks of false negatives. 3,923 of 20,000[2] sentences were annotated using binary indicators (1: t_1; 0: $!t_1$). Candidate sentences had a mean length of 34 ($\mu = 34.15, \sigma = 44.78$) tokens compared to the sample average for all sentences (taken after stripping short sentences) of 94 tokens ($\mu = 85.09, \sigma = 103.84$). Candidate extraction did not identify any candidates for 5 forms (.007 of the sample) because they used language not included in the seed word list. These edge cases will serve to strengthen next iterations. Overall .22 of the 3,923 annotated sentences were true positives according to human evaluation.

2.2. Text Classification

Using the the 3,923 binary annotations a number of binary classification models were trained to predict instances of $(t_1; !t_1)$. The aim is to train and evaluate different functions that approximate latent structure of statements of permission in order to classify unseen textual data. By training a classifier to discriminate between $(t_1; !t_1)$ it may be possible to facilitate a general deontic information retrieval model whose output can be used in sequential classification tasks. The ideal classifier outputs a probability so that experiments can be conducted a varying decision thresholds c. The approximation function takes the following general form:

$$\hat{y} = F'(\bar{\mathbf{X}}) + \epsilon$$

Where,

$$\hat{y} = \begin{cases} (1 : t_1) & \text{if } F'(\bar{\mathbf{X}}) \geq c \\ (0 : !t_1) & \text{if } F'(\bar{\mathbf{X}}) < c \end{cases}$$

This is based on the assumption that $F(\mathbf{X})$ is a *true* data generating function that determines whether or not the statement prescribes deontic concepts of permission. This function takes as input some theoretical latent space $(\mathbf{X})$ with unknown dimensions. A first challenge is to estimate a feature space $\bar{\mathbf{X}}$ that allows us to train a learning function $F'(\bar{\mathbf{X}})$ that produces $(t_1; !t_1)$ predictions with high fidelity to Y.

The ground-truth (Y) is defined according to the annotations. Predictions $(\hat{y})$ are evaluated on a hold-out set of data with known labels Y. In this case, the determination of $\bar{\mathbf{X}}$ has major implications for downstream processes. Notably, many estimations of a true latent feature-space do not account for word order, and may miss nuances inherent in conditional permission statements. Three different sets of features were used and the predictions $(\hat{y})$ made by a number of approximation functions $[F_1'(\bar{\mathbf{X}}), F_2'(\bar{\mathbf{X}}), ..., F_n'(\bar{\mathbf{X}})]$ were compared against known labels (Y).

2. Annotations were conducted using the free tier of DataTurks, an open-source annotation tool. The free tier limits users to 20K text classifications instances. Because the document order was not meaningful and DataTurks randomizes during annotation, I uploaded the first 20K sentences.

2.2.1. Feature Selection I. The first estimation of **X** was based on two core ideas. First, context-free pre-trained word GloVe vectors available through spaCy's language model [37], [39] were used. GloVe vectors were trained on Common Crawl [37], [40] to learn representations of words and word co-occurences [39]. The idea in using general probabilistic models of language is to apply latent English knowledge information theoretically obtained through learning on a huge training set to a more specific domain. The second constituent used was part-of-speech sequence positions along with simple total word counts and densities. The POS tagger is a multi-task statistical model (CNN) pre-trained on OntoNotes [36], [37], [41]. Again, the intuition in using pre-trained models is to effectively leverage knowledge of part-of-speech learned on more data than currently available from our domain. It is likely that customizing these training tasks (vectors and POS tags) will have performance implications in the domain of permissions.

Using pre-trained models also reduces the risk in over-fitting models to the training data, which has known issues. Additionally, they are easily computable on unseen instances, making them useful in practice. POS sequence vectors were represented as shown below. Most statistical models assume independence between column vectors, which is necessarily violated when representing language in this manner. This is partially accounted for using GloVe vector representation alongside POS positions. The attention-based models discussed in 2.2.3 address this concern specifically.

For all sequences that were shorter than the max sequence length empty elements were padded with zero. Sparse matrices create problems for many classifiers, especially when sample sizes are small. However, the approach only requires an estimate of the maximum length of a permission statement, and thus is relatively robust to unseen instances of t_1.

$$P \leftarrow \text{"I consent to ... "}$$

$$P : [PRP, VBD, TO...]$$

A downside of using POS sequence vectors is that it creates a sparse matrix which can be problematic for traditionally robust learning algorithms, especially with a small training set. Additionally, it requires knowing the length of the longest sentence, which is heavily influenced by the training data. However, because all features were represented numerically, it was possible to over-sample the under-represented instances of the positive class t_1 using SMOTE [42]. Using SMOTE, the proportion of labeled instances was brought to 1:1 on the training set only. SMOTE marginally influenced performance on simpler models, but risked making training data uninterpretable w.r.t to the text represented numerically[3].

The annotation set was split into stratified partitions based on proportions of t_1 in the sample. The majority of the exploratory classifiers were trained using this estimation

3. Random distances between POS centroids in multi-dimensional space do not retain fidelity to discrete notions of parts of speech.

of **X**. All classifiers were evaluated on a hold-out set (.3 proportion for all except BERT; .1 for BERT).

2.2.2. Feature Selection II. The second Feature Set was generated in order to test the first 100 (0 padded) token sequence without any morphological features or word embeddings. The length approximately accounts for 2σ around the mean sentence length. This Feature Set requires knowledge of the vocabulary of the sample space, which may not be a realistic assumption in the biomedical domain. Only an exploratory Convolutiuonal Neural Network (CNN) was trained using this Feature Set.

2.2.3. Training Classifiers. 12 binary classification models were trained on the task of predicting instances of t_1. Of these, 10 were trained on the same Feature Set and can be meaningfully compared against one-another. Hyperparameter specification for each model can be found in Appendix B. Most models trained on Feature Set 1 are either basic classification models, tree-based models or ensembles of weak learners used to set a performance baseline and understand the possibility of classification. Generally, these classifiers are shallow tree-based methods with large number of iterations. The intuition guiding hyperparameter specification for all models is to trade predictive accuracy for a model that can generalize well because of semantically ambiguous nature of deontic statements and the small training set. Known bias in the sample set also requires extreme caution against overfitting.

A deep neural network is a non-linear statistical model that transforms an input space ($\bar{\mathbf{X}}$) into an output ($\hat{y}$) through a series of linear combinations and binary decisions at several internal nodes. One of these models was a shallow autoencoder that performed reasonable well on hold-out testing data. This model contains 6 dense, fully connected layers whose activation functions are a mixture of ReLu and softmax. Autoencoders contain bottlenecks in their hidden layers, intuitively forcing the network to learn the most important features w.r.t to some loss function. The encoding layer compresses from 65,300 parameters to 5,010 and is approximately symmetrical. The simple ANN model uses the Keras built-in loss function for computing binary cross entropy [43], [44]. The binary case of cross-entropy can be defined as follows:

$$H = [t_1 * log(p(t_1))] + [!t_1 * log(p(!t_1))]$$

Which results in an loss increase when probability predictions $\hat{y}$ diverge from Y [45], [46]. The learning rate for this model was set to 0.001. Batch sizes were set to 3 for both the simple ANN and the CNN models based on the observations made by Tang et al. [47]. Structural choices for ANN and CNN models were relatively arbitrary to show that small DNNs can be trained in a reasonable amount of time with limited resources.

The CNN model was trained using Feature Set II with a limited vocabulary size (1,000) and uses only the first 100 tokens in the sequence (using zero padding for shorter sentences) to in order to mitigate the risk of overfitting. This model makes use of a high-drop out rate (.2) and a 1-dimensional convolutional layer (ReLu) in order to prevent overfitting. Additionally, this model contains an LTSM layer [43], [44], [48], as this has been shown to be effective in practice for text classification problems [49], [50], [51].

Finally, the final layers of BERT, a pre-trained bi-directional[4] attention model was trained on this classification task [52]. The pre-trained model used contained 12-layers (768-hidden) and 12-heads, with 110M paramaters [52], [53]. The BERT model is primarily intended for GPU/TPU environments, so training on model this model was restricted to lightweight pre-trained models and a shallower network. By leveraging theoretical latent understanding of the English language the author hoped to show promising results w.r.t permission statement classification.

The final models (CNN, BERT) cannot be meaningfully compared with the models because they are trained on a different set of features. However, they serve as important demonstrations of deep neural networks' ability to detect patterns in deontic statements.

3. Results

Overall, learning models show promise on binary classification of deontic statements. The models trained on Feature Set I show moderate variance and low-precision ($\mu = 0.576, \sigma = 0.189$). Table 1 shows the results of each classifier on a stratified test set. Importantly, all classifiers in Table 1 are compared against naïvely classifying all test examples as instances of the majority class ($!t_1$). A majority classification for all instances in the training set results in a misclassification rate of .22. Given this baseline performance we seek to train a learning algorithm that performs well w.r.t to positive class identification (t_1).

TABLE 1: Classifier Performance on Validation Data

Classifier	Feature Set	Acc.	Prec.	AUC
Naïve Majority	1	0.78	0.00	0.50
KNN	1	0.48	0.26	0.57
Logistic Regression	1	0.83	0.72	0.66
Decision Tree	1	0.69	0.32	0.57
Random Forest I	1	0.84	0.64	0.79
Random Forest II	1	0.84	0.81	0.66
Bagging I	1	0.76	0.48	0.73
Bagging II	1	0.78	0.75	0.51
AdaBoost	1	0.84	0.64	0.78
Gradient Boosting	1	0.81	0.57	0.76
Simple ANN	1	0.73	0.73	0.79

Accuracy, though reported, does not account for an asymmetric cost structure of Type I and Type II errors. This is addressed in hyperparameter specification of the decision tree model (more in Appendix B). Precision is a more suitable metric for discovery tasks, at least in the proof-of-concept space. The author notes that while precision is

4. BERT was trained on Wikipedia and BookCorpus [52]. Interested readers are encouraged to visit `https://github.com/google-research/bert`

important relative to model performance, the implications of incorrectly assuming that a statement prescribes *actual* permission may significantly outweigh the implications of *not assuming permission* in a production system. So while the precision is important w.r.t to a model's ability to facilitate discovery, AUC may be a stronger indicator of a model's overall appropriateness for a information ingestion pipeline. Precision is defined:

$$p_r = \frac{\text{true } t_1}{(\text{true } t_1 + \text{ false } t_1)}$$

Precision and accuracy depend on a decision threshold for classification c. This is a hyperparameter that requires specification and profoundly affects the performance of different classifiers. The default value, and the threshold reported on in Table 1, is .5. For instance:

$$p(t_1 \geq .5) \rightarrow t_1$$

AUC is defined as the integral (0,1) of the ROC curve. ROC represents a classifiers' ability to balance true t_1 vs. false t_1 predictions. An $AUC = 1$ indicates perfect predictive performance (no Type I or Type II errors). An $AUC = .5$ indicates a random prediction (see the Naïve Majority in Table 1). AUC is threshold (c) invariant measure of classifier performance, and thus is a good general metric for comparing binary classification models. In this case, an ideal model will perform well when considering both precision and AUC.

Selecting values for c has important implications. For this study, no experiments were conducted regarding different values of c, but heuristically higher values of c would suggest a more conservative outlook on model predictive power w.r.t. deontic statement identification. Empirically speaking, a higher value of c also represents a commitment to only extract permission statements we have a high degree of confidence are actually permissions, which is appropriate given this problem space.

Table 2 shows the same performance metrics for the fitted CNN and BERT models. Both these models appear to generalize to testing data fairly well, as indicated by high AUC and precision scores. One possible reason that the CNN and BERT perform well is that they use sequences as input features as do not learn based on synthetic train examples. In essence, the use of attention layers means that these models are word-order-aware and may suggest that this is an important attribute of $\bar{\mathbf{X}}$. Despite achieving high precision and AUC, the CNN use of the *first* 100 tokens as $\bar{\mathbf{X}}$ means that it cannot detect longer permissions, which is problematic. This can be mitigated by increasing the max sequence parameter at the expense of lower precision and AUC.

TABLE 2: Classifier Performance on Validation Data

Classifier	Feature Set	Acc.	Prec.	AUC
CNN	2	0.85	0.85	0.97
BERT	3	0.86	0.75	0.86

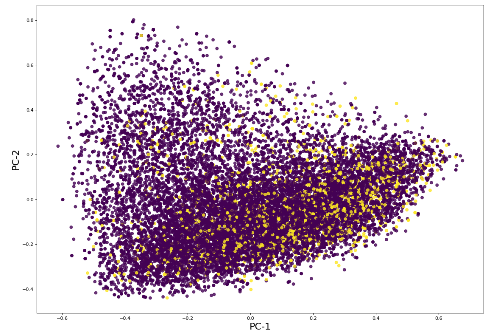

Figure 2: Separation of predictions based on the first 2 principal components of the td-idf matrix for all candidate sentences. Predictions for t_1 shown in yellow.

3.0.1. Predictions. Evaluating predictions presents a challenge because there are not ground truth labels available for the entire set of candidates. Predictions were generated by the CNN on the original corpus of 20,000 candidate sentences. Of these the CNN made 2,282 predictions of t_1 with $p(t_1) \geq c = .5$. The CNN only made predictions for 431 (0.715) of the 603 consent forms represented in the input file, making a mean of five predictions per form ($\mu = 5.29, \sigma = 6.23$).

While many of the predictions appear to be correct, the model is certainly fit to noise in the training data and may be improved using different values of c. Below are three cherry-picked examples showing predictions from a random consent form. While the first two predictions are promising, the third statement is hard to interpret as a statement of permission.

1) *"i give my permission for my xxx records to be accessed for use in this research study."*
2) *"information about you and related to this study will be kept confidential to the extent permitted or required by law."*
3) *"the alternative is to sign up for another study or to choose to complete another assignment as described in your syllabus."*

One measure of quality of predictions is to see if they tend to cluster in the feature space. By computing a td-idf matrix from the data using all n-grams through 0-5 and minimum and maximum frequency thresholds, respectively (0.1, 0.99) we can express the candidates in terms of the vocabulary that they contain. In order to visualize this in 2-dimensional space the first two principal components were computed and plotted. Figure 2 shows predictions of permission statements (t_1) in yellow and the negative class ($!t_1$) in dark purple.

Though empirical evaluation of the sentences lends itself towards congratulatory remarks, there is no observable pattern in vocabulary between the dimensions with highest variability. This may suggest that permission statements are not easily separable using vocabulary alone (preliminary classification models would attest to this), but it also may suggest that the functions approximations have not yet decoded true latent representations of permission. However, bag-of-words was not a predominant factor in feature selection.

3.1. Limitations

Although classification tasks show significant success, they are subjected to critical assumptions regarding the nature deontic statements as represented in textual data. The major limitations can be broken into three major categories: (1) representational, (2) statistical, and (3) ground-truth validity.

3.1.1. Representational Limitations. There are profound limitations of expressing deep semantics numerically that are beyond the scope of this investigation. However, beyond the issues of representation choice there exist issues in the constitution of $\bar{\mathbf{X}}$ that influence the study of deontic statements. Notably, all models except the CNN assume that pre-trained lexical resources can be leveraged in order to create $\bar{\mathbf{X}}$. While intuitively this seems appropriate, it may be the case that these resources are not specific enough to address the nuanced character of permissions in practice. Future work to address this would involve a large training corpus of permission-centric legal documents from a variety of domains.

3.1.2. Sample Limitations. While text based-informed consent forms represent a common modality, they are not the only mode of communication of permissions, even in the domain of informed consent. Documents represent the most easily computable form of communication, but even non-document communication and multi-document communication are present during informed consent processes in the real-world. The sample of documents is not only limited by virtue of availability, but also by format. Further, fears of liability prevent many intuitions from making informed consent forms public resulting in small sample sizes. It is hard to say if the documents used for this analysis are representative of the population, but this claim seems unlikely. Researchers are currently working to expand and strengthen this corpus, but the number of observable forms is a limitation of this study.

3.1.3. Quality of Annotations. Ideally, unsupervised learning will enable researchers to learn latent features of semantically ambiguous concepts without the significant hassle of unreliable human annotation. This study is limited not only by the number of annotators (1), but also by the total number of annotations produced. While it is possible that 'experts' may be able to create a larger set of annotations with higher fidelity to the latent structure of permissions, empirical evidence suggests that even experts will disagree on exactly what constitutes a statement of permission [54]. However, if statistical models are used to generalize notions of permission, those responsible for the ground-truth must also be accountable to it.

3.1.4. Directions for Future Study. First, by separating clinical and research informed consent forms it may be possible to develop more highly specialized models that capture notions of the regulatory frameworks which govern their respective informed consent processes. Second, it may be possible to combine an efficient heuristic method (A*) to reducing the search space for candidate identification. This could have the added benefit of scaling well beyond the constraints of the current study, even with limited computational resources. Lastly, as briefly mentioned, it may be possible to make use of the predictions for more granular downstream tasks. This seems largely unexplored in the literature, but may be of high value to research networks.

4. Conclusions

The goal of this study is identify a gap in current information system architecture and propose a method for revisiting computational deontic logic. We show it is possible to train moderately successful classifiers to discriminate between instances of $(t_1; !t_1)$. This data-driven approach suggests the possibility of a general purpose information retrieval model. Such a model may serve as a basis for instantiating ontological classes in a semantically aware knowledge-base whose objective is to make deontic notions available to the querying researcher. The added benefit of using a knowledge based approach is to codify rules based on existing regulatory frameworks and to provide an additional check on the predictions of a statistical model. Additionally, the outputs of such a model may be sequentially classified based on their specific content, potentially using only pre-trained statistical models, or by using encoded axioms and reasoners to enhance the query-able space.

Automating discovery of conditions of use has profound ethical implications for the medical research communities beyond those discussed here. The goal of this study is to demonstrate a data-driven proof-of-concept for discovery and does not address the full set of challenges needed to implement such a solution in practice. Still, this work shows it is possible to achieve moderate performance on semantically ambiguous space through the use of transfer learning and poses major implications for biospecimen and data access.

Acknowledgments

The author would like to thank the Michigan Institute for Data Science (MIDAS) and Michigan Integrated Center for Health Analytics and Medical Prediction (MiCHAMP) for making this research possible. The author thanks Dr. Marcy Harris, Dr. Neil J. Otte, Elizabeth Umberfield, Krishan Amin, and investigators of the NIH/NHGRI 5U01HG009454-03.

5. Supplemental Materials

Source code and data can be found at the following URL: `https://github.com/CooperStansbury/permission_statement_extraction`

6. Appendix A: Seed Words Used

The following section lists the lexical clues and all listed definitions from WordNet [30]:

1) **"authorization** : 1 : a document giving an official instruction or command 2 : the power or right to give orders or make decisions 3 : official permission or approval 4 : the act of conferring legality or sanction or formal warrant"
2) **"let** : 1 : a brutal terrorist group active in Kashmir; fights against India with the goal of restoring Islamic rule of India 2 : a serve that strikes the net before falling into the receiver's court; the ball must be served again 3 : make it possible through a specific action or lack of action for something to happen 4 : actively cause something to happen 5 : consent to, give permission 6 : cause to move; cause to be in a certain position or condition 7 : leave unchanged 8 : grant use or occupation of under a term of contract"
3) **"certify** : 1 : provide evidence for; stand as proof of; show by one's behavior, attitude, or external attributes 2 : guarantee payment on; of checks 3 : authorize officially 4 : guarantee as meeting a certain standard 5 : declare legally insane"
4) **"permit** : 1 : a legal document giving official permission to do something 2 : the act of giving a formal (usually written) authorization 3 : large game fish; found in waters of the West Indies 4 : consent to, give permission 5 : make it possible through a specific action or lack of action for something to happen 6 : allow the presence of or allow (an activity) without opposing or prohibiting"
5) **"permission** : 1 : approval to do something 2 : the act of giving a formal (usually written) authorization"
6) **"consent** : 1 : permission to do something 2 : give an affirmative reply to; respond favorably to"
7) **"grant** : 1 : any monetary aid 2 : the act of providing a subsidy 3 : (law) a transfer of property by deed of conveyance 4 : Scottish painter; cousin of Lytton Strachey and member of the Bloomsbury Group (1885-1978) 5 : United States actor (born in England) who was the elegant leading man in many films (1904-1986) 6 : 18th President of the United States; commander of the Union armies in the American Civil War (1822-1885) 7 : a contract granting the right to operate a subsidiary business 8 : a right or privilege that has been granted 9 : let have 10 : give as judged due or on the basis of merit 11 : be willing to concede 12 : allow to have 13 : bestow, especially officially 14 : give over; surrender or relinquish to the physical control of another 15 : transfer by deed"
8) **"sanction** : 1 : formal and explicit approval 2 : a mechanism of social control for enforcing a society's standards 3 : official permission or approval 4 : the act of final authorization 5 : give sanction to 6 : give authority or permission to 7 : give religious sanction to, such as through on oath"
9) **"authorize** : 1 : grant authorization or clearance for 2 : give or delegate power or authority to"
10) **"give** : 1 : the elasticity of something that can be stretched and returns to its original length 2 : cause to have, in the abstract sense or physical sense 3 : be the cause or source of 4 : transfer possession of something concrete or abstract to somebody 5 : convey or reveal information 6 : convey, as of a compliment, regards, attention, etc.; bestow 7 : organize or be responsible for 8 : convey or communicate; of a smile, a look, a physical gesture 9 : give as a present; make a gift of 10 : cause to happen or be responsible for 11 : dedicate 12 : give or supply 13 : transmit (knowledge or skills) 14 : bring about 15 : leave with; give temporarily 16 : emit or utter 17 : endure the loss of 18 : place into the hands or custody of 19 : give entirely to a specific person, activity, or cause 20 : give (as medicine) 21 : give or convey physically 22 : bestow 23 : bestow, especially officially 24 : move in order to make room for someone for something 25 : give food to 26 : contribute to some cause 27 : break down, literally or metaphorically 28 : estimate the duration or outcome of something 29 : execute and deliver 30 : deliver in exchange or recompense 31 : afford access to 32 : present to view 33 : perform for an audience 34 : be flexible under stress of physical force 35 : propose 36 : accord by verdict 37 : manifest or show 38 : offer in good faith 39 : submit for consideration, judgment, or use 40 : guide or direct, as by behavior of persuasion 41 : allow to have or take 42 : inflict as a punishment 43 : occur 44 : consent to engage in sexual intercourse with a man 45 : proffer (a body part)"
11) **"admit** : 1 : declare to be true or admit the existence or reality or truth of 2 : allow to enter; grant entry to 3 : allow participation in or the right to be part of; permit to exercise the rights, functions, and responsibilities of 4 : admit into a group or community 5 : afford possibility 6 : give access or entrance to 7 : have room for; hold without crowding 8 : serve as a means of entrance"
12) **"authorise** : 1 : give or delegate power or authority to 2 : grant authorization or clearance for"
13) **"agree** : 1 : be in accord; be in agreement 2 : consent or assent to a condition, or agree to do something 3 : be compatible, similar or consistent; coincide in their characteristics 4 : go together 5 : show grammatical agreement 6 : be agreeable or suitable 7 : achieve harmony of opinion, feeling, or purpose"
14) **"accord** : 1 : harmony of people's opinions or actions or characters 2 : concurrence of opinion 3 : a written agreement between two states or sovereigns 4 : sympathetic compatibility 5 : go together 6 :

allow to have"

15) **"provide** : 1 : give something useful or necessary to 2 : give what is desired or needed, especially support, food or sustenance 3 : determine (what is to happen in certain contingencies), especially by including a proviso condition or stipulation 4 : mount or put up 5 : make a possibility or provide opportunity for; permit to be attainable or cause to remain 6 : supply means of subsistence; earn a living 7 : take measures in preparation for"

16) **"check** : 1 : a written order directing a bank to pay money 2 : an appraisal of the state of affairs 3 : the bill in a restaurant 4 : the state of inactivity following an interruption 5 : additional proof that something that was believed (some fact or hypothesis or theory) is correct 6 : the act of inspecting or verifying 7 : a mark indicating that something has been noted or completed etc. 8 : something immaterial that interferes with or delays action or progress 9 : a mark left after a small piece has been chopped or broken off of something 10 : a textile pattern of squares or crossed lines (resembling a checkerboard) 11 : the act of restraining power or action or limiting excess 12 : obstructing an opponent in ice hockey 13 : (chess) a direct attack on an opponent's king 14 : examine so as to determine accuracy, quality, or condition 15 : make an examination or investigation 16 : be careful or certain to do something; make certain of something 17 : lessen the intensity of; temper; hold in restraint; hold or keep within limits 18 : stop for a moment, as if out of uncertainty or caution 19 : put a check mark on or near or next to 20 : slow the growth or development of 21 : be verified or confirmed; pass inspection 22 : be compatible, similar or consistent; coincide in their characteristics 23 : block or impede (a player from the opposing team) in ice hockey 24 : develop (children's) behavior by instruction and practice; especially to teach self-control 25 : consign for shipment on a vehicle 26 : hand over something to somebody as for temporary safekeeping 27 : abandon the intended prey, turn, and pursue an inferior prey 28 : stop in a chase especially when scent is lost 29 : mark into squares or draw squares on; draw crossed lines on 30 : decline to initiate betting 31 : hold back, as of a danger or an enemy; check the expansion or influence of 32 : place into check 33 : write out a check on a bank account 34 : find out, learn, or determine with certainty, usually by making an inquiry or other effort 35 : verify by consulting a source or authority 36 : arrest the motion (of something) abruptly 37 : make cracks or chinks in 38 : become fractured; break or crack on the surface only"

17) **"concur** : 1 : be in accord; be in agreement 2 : happen simultaneously"

18) **"assent** : 1 : agreement with a statement or proposal to do something 2 : to agree or express agreement"

19) **"approve** : 1 : give sanction to 2 : judge to be right or commendable; think well of"

20) **"accept** : 1 : consider or hold as true 2 : receive willingly something given or offered 3 : give an affirmative reply to; respond favorably to 4 : react favorably to; consider right and proper 5 : admit into a group or community 6 : take on as one's own the expenses or debts of another person 7 : tolerate or accommodate oneself to 8 : be designed to hold or take 9 : receive (a report) officially, as from a committee 10 : make use of or accept for some purpose 11 : be sexually responsive to, used of a female domesticated mammal"

21) **"attest** : 1 : provide evidence for; stand as proof of; show by one's behavior, attitude, or external attributes 2 : authenticate, affirm to be true, genuine, or correct, as in an official capacity 3 : give testimony in a court of law 4 : establish or verify the usage of"

22) **"cede** : 1 : give over; surrender or relinquish to the physical control of another 2 : relinquish possession or control over"

23) **"allow** : 1 : make it possible through a specific action or lack of action for something to happen 2 : consent to, give permission 3 : let have 4 : give or assign a resource to a particular person or cause 5 : make a possibility or provide opportunity for; permit to be attainable or cause to remain 6 : allow or plan for a certain possibility; concede the truth or validity of something 7 : afford possibility 8 : allow the other (baseball) team to score 9 : grant as a discount or in exchange 10 : allow the presence of or allow (an activity) without opposing or prohibiting"

24) **"accede** : 1 : yield to another's wish or opinion 2 : take on duties or office 3 : to agree or express agreement"

25) **"concede** : 1 : admit (to a wrongdoing) 2 : be willing to concede 3 : give over; surrender or relinquish to the physical control of another 4 : acknowledge defeat"

26) **"licence** : 1 : excessive freedom; lack of due restraint 2 : freedom to deviate deliberately from normally applicable rules or practices (especially in behavior or speech) 3 : a legal document giving official permission to do something 4 : authorize officially"

7. Appendix B: Classifier Hyperparameter Specification

TABLE 3: Simple Baseline Classifiers

Classifier	Model Parameters
KNN	algorithm='auto' leaf size=30 metric='minkowski' metric params=None n jobs=None n neighbors=5 p=2 weights='distance')
Logistic Regression	C=1.0 class weight=1: 0.21 dual=False fit intercept=True intercept scaling=1 max iter=1000000 multi class='warn' n jobs=None penalty='l1' random state=None solver='liblinear' tol=0.0001 verbose=0 warm start=False)

TABLE 4: Tree-Based Classifiers

Classifier	Model Parameters
Decision Tree	class weight=1: 0.2 criterion='gini' max depth=10 max features=9 max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 presort=False random state=None splitter='best')
Random Forest I	bootstrap=True class weight=1: 0.2 criterion='gini' max depth=None max features='auto' max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 n estimators=1000 n jobs=None oob score=False random state=None verbose=0 warm start=False)
Random Forest II	bootstrap=True class weight=1: 0.2 criterion='gini' max depth=10 max features='auto' max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 n estimators=1000 n jobs=None oob score=False random state=None verbose=0 warm start=False)

TABLE 5: Cont. Tree-Based Classifiers

Classifier	Model Parameters
Bagging I	base estimator=DecisionTreeClassifier criterion='gini' max depth=3 max features=None max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 presort=False random state=None splitter='best') bootstrap=True bootstrap features=False max features=9 max samples=1.0 n estimators=100 n jobs=None oob score=False random state=None verbose=0 warm start=False)
Bagging II	base estimator=DecisionTreeClassifier criterion='gini' max depth=10 max features=9 max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 presort=False random state=None splitter='best') bootstrap=True bootstrap features=False max features=9 max samples=1.0 n estimators=100 n jobs=None oob score=False random state=None verbose=0 warm start=False)

TABLE 6: Sequential Ensemble Classifiers

Classifier	Model Parameters
AdaBoost	algorithm='SAMME' base estimator=DecisionTreeClassifier(class weight=None criterion='gini' max depth=3 max features=None max leaf nodes=None min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 presort=False random state=None splitter='best') learning rate=1.0 n estimators=100 random state=None)
Gradient Boosting	criterion='friedman mse' init=None learning rate=0.1 loss='deviance' max depth=10 max features=None max leaf nodes=4 min impurity decrease=0.0 min impurity split=None min samples leaf=1 min samples split=2 min weight fraction leaf=0.0 n estimators=100 n iter no change=None presort='auto' random state=None subsample=1.0 tol=0.0001 validation fraction=0.1 verbose=0 warm start=False)

References

[1] R. R. Faden, T. L. Beauchamp, and N. E. Kass, Informed Consent, Comparative Effectiveness, and Learning Health Care, New England Journal of Medicine, vol. 370, no. 8, pp. 766768, Feb. 2014.

[2] C. Grady, Enduring and Emerging Challenges of Informed Consent, The New England Journal of Medicine; Boston, vol. 372, no. 9, pp. 855862, Feb. 2015.

[3] H. L. Harrell and M. A. Rothstein, Biobanking Research and Privacy Laws in the United States, J Law Med Ethics, vol. 44, no. 1, pp. 106127, Mar. 2016.

[4] Revised Common Rule, HHS.gov, 17-Jan-2017. [Online]. Available: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/finalized-revisions-common-rule/index.html. [Accessed: 20-Apr-2019].

[5] O. for C. Rights (OCR), Privacy, HHS.gov, 07-May-2008. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. [Accessed: 20-Apr-2019].

[6] H. L. Harrell and M. A. Rothstein, Biobanking Research and Privacy Laws in the United States, J Law Med Ethics, vol. 44, no. 1, pp. 106127, Mar. 2016.

[7] J.-J. C. Meyer, Applications of Deontic Logic in Computer Science: A Concise Overview, in Deontic Logic in Computer Science: Normative System Specification, 1993, pp. 1740.

[8] J.-J. C. Meyer, R. J. Wieringa, and F. P. M. Dignum, The Role of Deontic Logic in the Specification of Information Systems, in Logics for Databases and Information Systems, J. Chomicki and G. Saake, Eds. Boston, MA: Springer US, 1998, pp. 71115.

[9] P. McNamara, Deontic Logic, in The Stanford Encyclopedia of Philosophy, Fall 2018., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2018.

[10] R. Arp, B. Smith, and A. D. Spear, Building Ontologies with Basic Formal Ontology. MIT Press, 2015.

[11] B. Smith et al., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology, vol. 25, no. 11, pp. 12511255, Nov. 2007.

[12] Informed Consent Ontology. Contribute to ICO-Scrub-Team/ICO development by creating an account on GitHub. ICO-Scrub-Team, 2019.

[13] M. B. Almeida, L. Slaughter, and M. Brochhausen, Towards an Ontology of Document Acts: Introducing a Document Act Template for Healthcare, in On the Move to Meaningful Internet Systems: OTM 2012 Workshops, 2012, pp. 420425.

[14] Biomedical Research Integrated Domain Group —. [Online]. Available: https://bridgmodel.nci.nih.gov/. [Accessed: 20-Apr-2019].

[15] Consent - FHIR v4.0.0. [Online]. Available: https://www.hl7.org/fhir/consent.html. [Accessed: 20-Apr-2019].

[16] S. O. M. Dyke et al., Consent Codes: Upholding Standard Data Use Conditions, PLOS Genetics, vol. 12, no. 1, p. e1005772, Jan. 2016.

[17] Regulatory Ethics Toolkit. [Online]. Available: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/. [Accessed: 20-Apr-2019].

[18] M. van Gurp, M. Decoene, M. Holvoet, and M. Casella dos Santos, LinKBase, a Philosophically-inspired Ontology for NLP/NLU Applications, in CEUR Workshop Proceedings, 2006, vol. 222.

[19] J. A. Bateman, J. Hois, R. Ross, and T. Tenbrink, A Linguistic Ontology of Space for Natural Language Processing, Artif. Intell., vol. 174, no. 14, pp. 10271071, Sep. 2010.

[20] K. Liu, W. R. Hogan, and R. S. Crowley, Natural Language Processing methods and systems for biomedical ontology learning, Journal of Biomedical Informatics, vol. 44, no. 1, pp. 163179, Feb. 2011.

[21] C. D. Manning and H. Schtze, Foundations of statistical natural language processing. Cambridge, Mass: MIT Press, 1999.

[22] D. Jurafsky and J. H. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed. Upper Saddle River, N.J: Pearson Prentice Hall, 2009.

[23] A. Halevy, P. Norvig, and F. Pereira, The Unreasonable Effectiveness of Data, IEEE Intelligent Systems, vol. 24, no. 2, pp. 812, Mar. 2009.

[24] D. MAYNARD, Y. LI, and W. PETERS, NLP Techniques for Term Extraction and Ontology Population, p. 21.

[25] D. Bahdanau, K. Cho, and Y. Bengio, Neural Machine Translation by Jointly Learning to Align and Translate, Sep. 2014.

[26] A. See, P. J. Liu, and C. D. Manning, Get To The Point: Summarization with Pointer-Generator Networks, Apr. 2017.

[27] M.-T. Luong, H. Pham, and C. D. Manning, Effective Approaches to Attention-based Neural Machine Translation, Aug. 2015.

[28] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori, Combining NLP Approaches for Rule Extraction from Legal Documents, in 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016), Sophia Antipolis, France, 2016.

[29] Zhang Jiansong and El-Gohary Nora M., Semantic NLP-Based Information Extraction from Construction Regulatory Documents for Automated Compliance Checking, Journal of Computing in Civil Engineering, vol. 30, no. 2, p. 04015014, Mar. 2016.

[30] WordNet — A Lexical Database for English. [Online]. Available: https://wordnet.princeton.edu/. [Accessed: 20-Apr-2019].

[31] The Common Core Ontology Repository holds the current released version of the Common Core Ontology suite.: CommonCoreOntology/CommonCoreOntologies. CommonCoreOntology, 2019.

[32] C. Freitas, D. Santos, C. Mota, H. Gonalo Oliveira, and P. Carvalho, Relation detection between named entities: report of a shared task, in Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009), Boulder, Colorado, 2009, pp. 129137.

[33] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, Distant supervision for relation extraction without labeled data, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, 2009, pp. 10031011.

[34] K. Swampillai and M. Stevenson, Extracting Relations Within and Across Sentences, in Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, Hissar, Bulgaria, 2011, pp. 2532.

[35] M. K. Paasche-Orlow, H. A. Taylor, and F. L. Brancati, Readability Standards for Informed-Consent Forms as Compared with Actual Readability, New England Journal of Medicine, vol. 348, no. 8, pp. 721726, Feb. 2003.

[36] M. Honnibal and I. Montani, spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing, To appear, 2017.

[37] Models for the spaCy Natural Language Processing (NLP) library: explosion/spacy-models. Explosion, 2019.

[38] StanfordNLP - Python NLP Library for Many Human Languages — StanfordNLP. [Online]. Available: https://stanfordnlp.github.io/stanfordnlp/. [Accessed: 21-Apr-2019].

[39] J. Pennington, R. Socher, and C. Manning, Glove: Global Vectors for Word Representation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 15321543.

[40] Statistics of Common Crawl Monthly Archives by commoncrawl. [Online]. Available: https://commoncrawl.github.io/cc-crawl-statistics/. [Accessed: 20-Apr-2019].

[41] OntoNotes Release 5.0 - Linguistic Data Consortium. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2013T19. [Accessed: 20-Apr-2019].

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, 1, vol. 16, pp. 321357, Jun. 2002.

[43] Deep Learning for humans. Contribute to keras-team/keras development by creating an account on GitHub. Keras, 2019.

[44] FAQ - Keras Documentation. [Online]. Available: https://keras.io/getting-started/faq/. [Accessed: 20-Apr-2019].

[45] Loss Functions ML Cheatsheet documentation. [Online]. Available: https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html#cross-entropy. [Accessed: 21-Apr-2019].

[46] D. Godoy, Understanding binary cross-entropy / log loss: a visual explanation, Towards Data Science, 21-Nov-2018. [Online]. Available: https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a. [Accessed: 21-Apr-2019].

[47] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, Sep. 2016.

[48] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Comput., vol. 9, no. 8, pp. 17351780, Nov. 1997.

[49] S. Ahamed, Text Classification Using LSTM and visualize Word Embeddings: Part-1, Medium, 09-Jan-2018. .

[50] J. Brownlee, Sequence Classification with LSTM Recurrent Neural Networks in Python with Keras, Machine Learning Mastery, 25-Jul-2016. .

[51] E. R. Fonseca, Text autoencoder with LSTMs. Contribute to erickrf/autoencoder development by creating an account on GitHub. 2019.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs], Oct. 2018.

[53] TensorFlow code and pre-trained models for BERT. Contribute to google-research/bert development by creating an account on GitHub. Google AI Research, 2019.

[54] A. Dumitrache, L. Aroyo, and C. Welty, Achieving Expert-Level Annotation Quality with CrowdTruth, p. 13.