LDSI W2021 Literature Survey

Name: Syed Husain Mustafa
Gloss ID: tum_ldsi_28

Summary

In this survey report I examine 6 papers on the topic of legal norm classification. This field of research finds great application in legal database generation, in information retrieval and aims to provide legal practitioners with the tools to examine ever-changing laws and regulations at a faster pace. The papers that have been curated span 8 years and deal with legal norms originating in German, Dutch, British and EU legal texts. In this survey we observe a split in performance between methods that consider a holistic view of legal norm classification and those that instead focus on a subset of all legal norms, namely “deontic” norms. This survey also covers an evolution in the computational models used for automated legal text classification. Whereas earlier approaches relied on Knowledge Based systems, subsequent approaches deal with Machine Learning and Neural Network based models that demonstrate greater document and domain portability than the former approach.

Corpora Used

The earliest paper examined in this survey comes from Maat et al. (2010) [1], where statistical methods are used to classify sentences in 18 Dutch laws. The corpora comprises 584 sentences and 62 non-sentential phrases in the form of bulleted lists, in total constituting 13 modalities/norms. The taxonomy used in this paper is explained in a prior work by Maat et al. [2]. Subsequent works by Waltl. et al. ([4], [5]), and Glaser et al. [6] expand upon the work by Maat et al. within the context of German Civil Law, namely by classifying norms originating in tenancy law as defined in the German Civil Code (§535-§595). This constitutes 601 sentences, segmented at the ending period, and annotated by a single domain expert. The study by O’Niell [7] focuses on a subset of the modalities used by Maat et al., namely “deontic” modalities, where phrases indicate obligations and permission, which are paramount to contractual agreements. O’Niell et al. compile a corpus of 1297 sentences, with 596 obligations, 94 prohibitions, and 607 permissions that were annotated by subject matter experts. To test their results, O’Niell et al use EU and UK Anti-Money Laundering (AML) legislations. The approach of O’Niell et al. is further refined and improved upon by Chalkidis et al [8]. This is done by employing a larger dataset that comprises of 31,545 training, 8,036 development and 5,563 test sentences that were obtained from 100 randomly selected English service agreements. Chalkidis et al. used the same taxonomy as O’Niell et al., only substituting the expertise of one legal annotator with 5 law students, that were later cross-marked by a paralegal expert.

Learning Methods and Features Used

Maat et al.(2010) employ Support Vector Machines. They compare the SVM against a rule-based approach based on a context-free grammar that was modelled in a prior study [3]. Maat et. al. represent each document, i.e., sentence using the bag of words model, the words are assigned a one-hot encoding, and are characterised by TF-IDF features. Each sentence is represented as a vector sum of corresponding words that fulfill minimum frequency thresholds. In the experiment by Waltl et. al (2017)[4] nine combinations of “Active Machine Learning" (AML) models are used, where the semi-supervised machine learning models are trained in rounds. In each round, a set number of labelled sentences are used to learn a classification scheme, a variable amount of unlabelled data is then introduced to the classifier, various “queries” are used to measure the information gain from classifying the unlabelled data. Here "queries" are classification schemes that are determined mathematically, the query which maximes information gain is ultimately used to update the supervised learning classifier. Waltl et al. then combine AML with a rule-based model which generally demonstrates high bias and low variance, to obtain a hybrid model that can reliably classify legal norms. Nine variants of hybrid model are compared against three conventional supervised learning (CSL) methods. In a later study, Waltl et al.[5](2019), inspired by Maat et al. [1], employ local linear model agnostic explainations (LIME) to ascertain how a statistical model classifies norms, and compare their findings against a rule-based approach in a German legal context. Citing earlier studies from Maat et al. ([1], [3]) where only 44 out of 87 observable patterns in the data lead to a successful classification, Waltl et al. chose to focus their LIME approach on how "modal-verbs" play a role in determining sentence/clause modality. In preprocessing German stop words were removed, features include word count vectors and TF-IDF metrics, in addition to the bag-of-words. The study employs the Support Vector Classifier (SVC), Random Forest (RF), Multilayer Perceptron, Multinomail Naive Bayes and Logistric Regression. The LIME library [9] was used to reconstruct the decision making process of the best performing statistical model. In a concurrent study by Glaser et al.[4], the document portability of statistical models in the context of German legal contracts is studied. The authors employ three taxonomies for legal norms originating in the tenacy law of the German Civil Code (BGB). The first differentiates between "rights" and "obligations". The next constitutes "rights", "references", "definitions" and "legal consequences", and a third taxonomy that is extended from Waltl. et al.[5](2019). Preprocessing includes line break removal, punctuation removal and umlat replacement. Like prior approaches by Waltl et. al ([4], [5]) count vectors and TF-IDF features along with a bag-of-words approach are used. Six statistical methods are trained and evaluated in a 10 fold cross validation manner. The models are applied to 169 sentences from German rental agreements. To compensate for the small dataset, the model is retrained on a set of 312 rental agreement sentences post initial round of testing. The study by O'Niell et al.[7] classifies deontic modalities within the context of financial law. They employ count based vectors trained on "Wikimedia" and "GigaWord", additionally word2vec vectors are trained via "300d Google embeddings".The paper compares the performance between Artificial Neural Network (ANN) and non-ANN statistical approaches. The later are suplemented with ensemble approaches to compensate for any loss of performance in large dimensional feature space. The ANN approaches involve a Bidirectional Long-Short-Term Memory (BiLSTM) model trained on three embeddings, a Convolutional Neural Network (CNN), LSTM, CNN-LSTM, and a vanilla ANN. The follow up work by Chalkidis et al.[8] employs pre-trained embedding vectors of 200 dimension, an 25 additional dimensions that capture Parts of Speech (POS) tags, and another 5 dimensions that capture shape embeddings, i.e., the various forms in which a word appears, capitalized, non-capitalized, etc. They add a self attention mechanism to the BiLSTM classifier, introduce a hierarchical BiLSTM to embed context awareness in classifier, and generate a BiLSTM classifier that prepends and appends 150 tokens to the sentence that is to be classified as an alternate form of context embedding dubbed "X-BILSTM-ATT".

Results Obtained and Unresolved Issues

As a consequence of having a small dataset Maat et. al used a form of cross validation, namely, the "Leave-One-Out" procedure where the same data is used for training and testing. The averaged accuracy over entire set is 93%, which improves slightly with preprocessing to a maximum of 94.69%. The best performance is achieved, as they note by using binaryily weighted word vectors, with removal of stop words and inclusion of words with a minimum frequency of two in the document. The dataset being heavily skewed resulted in under-represented classes having lower recall and precision. The model is not able to clearly discern between "obligations", "permissions" and "definitions". Waltl et. al (2017)[4] obtain an F1 score of 80% with only 35% of labeled instances with AML, which is an improvement by 6% over CSL methods that employ 65% of the labeled instances. AML models converge faster inaddition to having better F1 metrics. Waltl et. al note that the AML approach is unable to discern between “definition” and “procedure” type sentences, and attribute this to low representation of the aforementioned types in the data used for training. Though this claim is put into question by the AML models’ performance on classifying sentences of type “continutation” with an accuracy of ~80% which only has 2 more instances than that of type “definiton”. In the subsequent study by Waltl et al.[5] most of the statistical methods employed, aside from SVC and LR, fair much worse than the rule-based model. Counter-intuitively, not removing stop-words improve the performance of the statistical models in the German language setting. Waltl et al. also note that SVC and LR approaches demonstrate higher support for “permission” (F1 of 0.94) and “consequence”(F1 of 0.91) types (i.e, sentences of the deontic norm type) over the rule-based approach. Further analysis with LIME demonstrates high level of verb-type correspondence in the SVC and LR approaches comparable to the rule-based approach. The study on document portability by Glaser et al.[6], demonstrates Extra Tree Classifier (ETC) and SVM perform the best out of the statistical methods considered (0.815 and 0.828 F1 respectively). TF-IDF features infact lead to inferior performance. Additionally removing stop words in the German language context makes it difficult for the models to account for modal verbs that appear in the infinitiv form in the test set. Work by O’Niell et al.[7] in deontic norm detection demonstrates the superiority of ANN based statistical methods over non-ANN Methods, with BiLSTM fairing the best. It is noted that word embeddings trained in the legal domain in addition to the Google News Embeddings demonstrate higher F1 score than the later taken alone, thereby reaffirming the results discussed in our lecture on Zheng et al. (2021) [10], i.e., that sufficiently hard domain-specific pre-training tasks add to performance. The BiLSTM approach is flexible enough to learn key verb-type associations as well as capture the context associated with each type, going far and beyond the rule-based approach which only account for the verb-type pair. Chalkidis et al.[8] demonstrate that self-attention applied to BiLSTM lead to improved results, supporting the hypothesis that self-attention allows the classifier to learn “indicative tokens”.They observe faster convergence time, and further improved results in Heirarchical BiLSTM where the context of an entire section is used to determine type of a sentence. The alternative approach dubbed “X-BILSTM-ATT” however does not yield any improvement.

Discussion

Maat et. al.[1] being one of the first to employ statistical models in the task of automated legal classification rightly observe that a lack of data annotated by experts hamstrings the generalizability of any such model conceived at the time. The works by Waltl et al. ([4], [5]) employ improved taxonomy and semi-supervised learning to make better use of the little data that was available, and yet still the models are unable to discern the difference between certain types (“definition” and “procedure”). They employ LIME to provide some insight into how verb-type relationships come about in deontic norms, yet there is no discernable pattern in the non-deontic norms. This, the authors note may be a consequence of data bottlenecks, and the complex structure of definitions and procedures in the German legal context which involve multiple cascading sentences and the use of wildcards(*). On the otherhand the works by O’Niell et al.[7] and Chalkidis et al.[8] focus solely on deontic norms and demonstrate commercially viable models. The success can be attributed in large part to the many regulatary/ financial documents which allows for one to utilize the full potential of deep ANNs. With regard to the non-deontic norms which seldom appear in commercial law/ financial agreements, and are hence a rare sight, severally limit the viability of statistical methods.

References

[1] Emile de Maat, Kai Krabben, and Radboud Winkels. 2010. Machine Learning versus Knowledge Based Classification of Legal Texts. In Proceedings of the 2010 conference on Legal Knowledge and Information Systems: JURIX 2010: The Twenty-Third Annual Conference. IOS Press, NLD, 87–96.

[2] E. de Maat and R. Winkels. Categorisation of Norms, in: A. Lodder and L. Mommers (Eds.), Legal Knowledge and Information Systems. Jurix 2007. IOS Press, Amsterdam, 2007, 79-88.

[3] E. de Maat and R. Winkels. A Next Step towards Automated Modelling of Sources of Law, in: Proceedings of ICAIL 2009. ACM, New York, 2009, 31-39.

[4] Waltl, B.; Muhr, J.; Glaser, I.; Bonczek, G.; Scepankova, S.; Matthes, F.:Classifying Legal Norms with Active Machine Learning, Jurix: International Conference on Legal Knowledge and Information Systems, Luxembourg, Luxembourg, 2017

[5] Bernhard Waltl, Georg Bonczek, Elena Scepankova, and Florian Matthes. 2019. Semantic types of legal norms in German laws: classification and analysis using local linear explanations. Artif. Intell. Law 27, 1 (Mar 2019), 43–71.

[6] Glaser, Ingo & Scepankova, Elena & Matthes, Florian. (2018). Classifying Semantic Types of Legal Sentences: Portability of Machine Learning Models.

[7] O'Neill, James & Buitelaar, Paul & Robin, Cécile & O'Brien, Leona. (2017). Classifying Sentential Modality in Legal Language: A Use Case in Financial Regulations, Acts and Directives. 159-169. 10.1145/3086512.3086528.

[8] Chalkidis, Ilias & Androutsopoulos, Ion & Michos, Achilleas. (2018). Obligation and Prohibition Extraction Using Hierarchical RNNs.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778

[10] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. <i>Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law</i>. Association for Computing Machinery, New York, NY, USA, 159–168. DOI:https://doi.org/10.1145/3462757.3466088