Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210317

The Unreasonable Effectiveness of the Baseline: Discussing SVMs in Legal Text Classification

Benjamin CLAVIÉ [a,1] and Marc ALPHONSUS [a]
[a] *Jus Mundi*

Abstract. We aim to highlight an interesting trend to contribute to the ongoing debate around advances within legal Natural Language Processing. Recently, the focus for most legal text classification tasks has shifted towards large pre-trained deep learning models such as BERT. In this paper, we show that a more traditional approach based on Support Vector Machine classifiers reaches competitive performance with deep learning models. We also highlight that error reduction obtained by using specialised BERT-based models over baselines is noticeably smaller in the legal domain when compared to general language tasks. We discuss some hypotheses for these results to support future discussions.

Keywords. Natural Language Processing, Text Classification, Machine Learning

1. Introduction

Recently, the state-of-the-art in many Natural Language Processing (NLP) tasks has been achieved by large pre-trained models such as BERT and its variants [1]. Specialised BERT-based models have been developed for many fields, establishing the state-of-the-art in domain specific tasks, as evidenced in the biomedical domain [2].

In legal NLP, recent work has focused on exploring the applications of BERT-based approaches on a variety of existing tasks and how to best adapt BERT to the legal domain [3,4]. These efforts, while successful at establishing state-of-the-art on a variety of tasks, also reveal an interesting trend: the performance gain between a general language BERT and a specifically legal-language trained BERT appears to be smaller than in other specialised domains [4].

A common application of legal NLP is text classification. Text classification tasks target various kinds of legal insight, such as predicting the outcome of a ruling from a decision's body [5], whether a given clause is likely to be unfair to a customer [6] or whether a sentence indicates the overruling of a precedent [4].

Little attention has been given to comparing these new BERT-based approaches to well-optimised baselines, such as Support Vector Machine (SVM)-based classifiers, which historically perform well on text classification tasks, opting instead for comparisons with other deep learning-based baselines.

[1]Contact: Benjamin Clavié, Jus Mundi, 30 Rue de Lisbonne, 75008 Paris, France; b.clavie@jusmundi.com.

Table 1. Best performing model on all tasks.

	ECHR (Both)	Overruling	Terms of Services
Best Approach	NBSVM + bigrams	NBSVM + bigrams	Linear SVM + trigrams

In this short paper, we aim to (A) highlight the very strong performance of optimised *baseline* classifiers on multiple legal text classification tasks compared to deep learning classifiers, (B) show that the gains from BERT-based approaches is noticeably smaller on legal-domain tasks than on general tasks and (C) discuss three hypotheses to explain the previous two phenomena.

2. Experimental Setup

2.1. General Domain

For all general domain tasks, we use results from BERT-ITPT-FiT [7], which optimises BERT for text classification, on four common benchmarks. For SVM results, we report the score of the best performing variant from a large scale comparison [8].

2.2. Legal Domain Experiments & Baselines

We compare SVMs to BERT-based results on four existing legal text classification tasks.

Terms of Services (ToS) is a task aiming to determine whether or not a clause found in a contract is likely to be unfair to the customer [6].

Overruling is a binary classification task to identify if a given sentence in a US court decision represents a reversal of precedent (*overruling* a previous decision) [4].

ECHR text classification tasks use the text of the *Facts* part of decisions from the ECtHR and exists in two variants. **ECHR (Binary)** aims to detect if any article of the Charter has been violated and **ECHR (Multi)** requires identifying specifically which article has been violated. We use the *Frequent* version of this last task, meaning we remove any label with fewer than 50 training examples from the data [5].

On each of the legal tasks, we train and evaluate SVM classifiers with modest optimisation. We also experiment with NBSVM[2], an SVM classifier using Naïve Bayes features to represent words [9]. The best approach for each task is listed in Table 1.

For BERT-based models, we report results from the literature. Results for BERT on both **ECHR** tasks are from the paper introducing the task [5] and results from Legal-BERT from the paper introducing it [3]. Results for all BERT-based models on **Overruling** and **Terms of Services** are from the paper introducing Legal-Bert-Custom [4].

2.3. Metrics and Evaluation

In line with the nature of this paper, all metrics reported follow the existing literature. For all General Domain tasks, the metric used is accuracy over the test set.

[2]Implementation available at https://gitlab.com/jusmundi/Legal-svm-baselines/

Table 2. Results for the best performing model of each kind on a variety of General Domain (GD) and Legal text classification tasks. *Error reduction is calculated between the Best SVM and the best BERT variant.*

	General Domain				Legal			
Model	**AGNews**	**IMDB**	**Yelp!**	**DBPedia**	**ECHR (Binary)**	**ECHR (Multi)**	**Over-ruling**	**ToS**
Best SVM	75.3	80.7	84.0	87.1	82.2	61.1	94.9	79.3
BERT	95.2	95.6	98.1	99.3	82.0	60.8	95.8	72.2
Legal-BERT [3,4]	n/a	n/a	n/a	n/a	88.3	65.2	97.4	78.7
Error Reduction	80.6%	77.2%	88.1%	94.8%	34.3%	10.5%	49%	-1.8%

We report macro-averaged F1 score for **ECHR (binary)**, micro-averaged F1 score over all classes for **ECHR (multi)** [5] and average F1 score over 10-fold cross-validation on both Overruling and ToS [4].

In all cases, we report the error reduction between BERT models and SVMs as the percentage decrease in error rate between models to simplify evaluating the impact of using a different model over multiple tasks. The error rate is calculated as $100 - Score$.

3. Classification Results

Table 2 gives an overview of the various classification results and presents the error reduction obtained by using the best BERT-based model over the relevant SVM classifier.

The error reduction between SVM and BERT models in the general domain is high, at **85.175%** on average over the four tasks, with the lowest reduction being 77.2%.

The difference is much less stark within the legal domain: on all but one of the tasks, the performance of the SVM classifier exceeds that of a general domain BERT[3] .

Legal-BERT models, optimised for legal texts, do reach the best performance on three out of four tasks and only slightly fall short of it on the fourth. However, in all cases, the performance increase is much less pronounced, with an average error reduction across all four tasks of **23%**.

4. Discussion

The results highlight an interesting phenomenon: despite impressive performance in both the general domain and other specialised domains without the need for domain adaptation [2], BERT falls short within the legal domain. Even after domain adaptation is performed to train specialised Legal-BERT models, the performance improvement remains modest and does not reproduce the very notable gains found in other applications.

On **ECHR (Multi)**, potentially the most complex task due to being multi-label, there is remarkably only an **10.5%** error reduction between SVM and Legal-BERT.

There is no clear explanation for this phenomenon, but we discuss multiple hypotheses. The first, initially proposed by Zheng et al. [4] to explain the mild improvements from Legal-BERT, is that the tasks on which we evaluate legal NLP algorithms are not suit-

[3]The results on ECHR are considerably better than the SVM approach reported in the original paper [5] as they use tf-idf weighting for feature generation, which performs notably worse than other methods.

able, either due to them being too simple or their language not being sufficiently domain-specific to take advantage of the models' pretraining. However, this does not provide an explanation for the overall weak improvement from deep learning over SVM classifiers.

A similar potential explanation could be that *simple* mono-lingual text classification is not enough to truly take advantage of the possibilities offered by more powerful BERT-based models. This would indicate that the powerful language representation of Legal-BERT models could be key to tackling more complex tasks which have started being explored, such as textual entailment [4] or legal rationale extraction [10].

However, this still does not fully address the weak performance gains on text classification. A final hypothesis we propose is that large language models, even when trained on legal language, still lack the ability to capture the depth of legal language and its specific vocabulary. These models could also fail to properly weigh the meaning of multiple legal concepts being mentioned together. This hypothesis would suggest the need to develop a way to integrate sources of legal information, such as knowledge-bases or ontologies, within deep learning models to truly take advantage of their potential.

5. Conclusion

We show that SVM classifiers perform well on multiple legal text classification benchmarks. SVM models can outperform general domain BERT models, but perform worse than BERT-based models adapted for legal text. We also show that the relative performance improvement between the BERT-based models and the SVM models is considerably smaller within the legal domain than on general domain classification tasks.

We propose and discuss three potential explanations for these results. Future work will focus on exploring the limits of BERT models within the legal field, both by exploring more complex tasks and integrating existing knowledge bases with them.

References

[1] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL 2019;. .

[2] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019 09.

[3] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of EMNLP 2020;. .

[4] Zheng L, Guha N, Anderson BR, Henderson P, Ho DE. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. In: Proceedings of ICAIL2021. ACM;. .

[5] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. In: Proceedings of ACL; 2019. p. 4317-23.

[6] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artif Intell Law. 2019;27(2):117-39.

[7] Sun C, Qiu X, Xu Y, Huang X. How to fine-tune bert for text classification? In: China National Conference on Chinese Computational Linguistics. Springer; 2019. p. 194-206.

[8] Riekert M, Riekert M, Klein A. Simple Baseline Machine Learning Text Classifiers for Small Datasets. SN Computer Science. 2021;2(3):1-16.

[9] Wang S, Manning CD. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In: Proceedings of ACL 2021;. .

[10] Chalkidis I, Fergadiotis M, Tsarapatsanis D, Aletras N, Androutsopoulos I, Malakasiotis P. Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases. In: Proceedings of NAACL 2021;. .