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

Benjamin Clavié and Marc Alphonsus

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 surprisingly competitive performance with BERT-based models on the classification tasks in the LexGLUE benchmark. 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 present and discuss three hypotheses as potential explanations 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 identifying the topic of a contract close [7]. Recently, the LexGLUE benchmark has been released to allow for easier, more transparent benchmarking in legal NLP [8].

Little attention has been given to comparing these new BERT-based approaches to well-optimised baselines, such as Support Vector Machine (SVM)-based classifiers,

---

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

which historically perform well on text classification tasks, opting instead for comparisons with other deep learning-based baselines.

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.

1.1. General Domain

For all general domain tasks, we use results from BERT-ITPT-FiT [9], 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 [10].

1.2. Legal Domain Experiments & Baselines

For ease of comparison and reproducibility, we benchmark SVM classifiers on LexGLUE [8]. LexGLUE, is a benchmark suite comprising of commonly studied legal NLP tasks. For the purpose of this paper, we evaluate our models on all six classification tasks and leave the QA task aside and focus on the six text classification tasks:

ECtHR A and B [5,11] are multi-label tasks which use *facts* part of European Court of Human Rights (ECtHR) rulings and the goal is to predict which article(s) were found by the court to have been violated (A) and which article(s) were considered by the court as allegedly violated (B).

SCOTUS [8] is a multi-class task on opinions from the Supreme Court of the United States which need to be classified into one of 14 issue areas.

EUR-LEX [12] is a multi-label task on EU law documents. The labels are comprised of 100 common EuroVoc legal concepts. The aim of this task is to identify all relevant EuroVoc label linked with a document.

LEDGAR [7] is a multi-class task on contract provisions (paragraph) from the US Securities and Exchange Commission (SEC) where the goal is to detect the main topic of a given provision.

Unfair Terms of Service (ToS) [6] is a multi-label task with clauses from the Terms of Services of 50 online plateforms. The aim of this task is to detect if a clause is likely to violate consumer rights and exactly which right is at risk of being violated.

Further information about the LexGLUE tasks is provided by Chalkidis et al. [8].

On each of the legal tasks, we train and evaluate SVM classifiers with modest optimisation. Results are reported as the average of ten runs. We also experiment with NBSVM, an SVM classifier using Naïve Bayes features to represent words [13]. We release the code used to train all models. [2].

For BERT-based models, we report the benchmark results from the LexGLUE paper [8]. Following their nomenclature, we refer to the legal Legal-BERT model from Chalkidis et al. [3] as Legal-BERT and the Legal-BERT model from Zheng et al. [4], trained on US case law documents, as CaseLaw-BERT.

---

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

1.3. 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. As these tasks can be multi-class but not multi-label classification tasks, accuracy is identical to micro-averaged F1 score. [14]

For all legal tasks, we report both micro-averaged (μF1) and macro-averaged (mF1) F1 score for all tasks to allow for easier comparison with both other models evaluated on LexGLUE and results from general domain tasks.

In all cases, we report the error reduction between the best performing BERT-based model 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*.

2. Classification Results

Table 1 gives an overview of the various classification results on General Domain tasks and presents the error reduction obtained by using the a fine-tuned BERT model over the best SVM classifier in the literature.

Table 2 presents the classification results on the 6 LexGLUE classification tasks. We report the results for our best performing SVM as well as the results from BERT-Base, Legal-BERT and CaseLaw-BERT from the existing LexGLUE benchmark. We also compute the error reduction between the SVM classifier and the best performing BERT-based model for each task.

The error reduction between SVM and BERT models in the general domain is high, at **85.1%** on average over the four tasks, with the lowest reduction being 77.2%.

The difference is much less stark within the legal domain. BERT-based models do obtain the best results on all six takss, with Legal BERT models reaching the best performance on five out of the six tasks and a general domain BERT slightly outperforming them on **ECtHR (A)**. However, in all cases, the performance increase is much less pronounced than in other domains, with an average micro-F1 error reduction across all six tasks of just **18.1%**.

The average reduction in macro-F1 is overall similar, with an average reduction of **18.3%**. However, is noticeable that this reduction in is considerably more pronounced on both *ECtHR* tasks. The ECtHR dataset is rather imbalanced, with all both tasks containing labels with very few examples and others covering a large proportion of the data [8]. This seems to point towards the BERT models' ability t oget better performance with lower quantities of data.

3. Discussion

The results highlight an interesting phenomenon: while the average micro-F1 error reduction on general domain benchmarks is **85.1%**, it is only **18.1%** on legal text classification tasks.
It is worth noting that the performance increases for BERT-based models is larger on

Table 1. Results for the best performing model of each kind on a variety of General Domain (GD). *Error reduction is calculated between the Best SVM and the best BERT variant.*

Model	AGNews	IMDB	Yelp!	DBPedia	Average
Best SVM	75.3	80.7	84.0	87.1	81.78
BERT	95.2	95.6	98.1	99.3	97.0
Error Reduction	80.6%	77.2%	88.1%	94.6%	85.1%

Table 2. Results for the various models on the 6 LexGLUE classification tasks. μF1 refers to micro-averaged F1-score and mF1 to macro-averaged F1 score.*Error reduction is calculated between the Best SVM and the best BERT variant.*

Model	ECtHR				SCOTUS		EUR-LEX		LEDGAR		Unfair Tos		Average	
	A		**B**											
	μF1	mF1	μF1	mF1	μF1	mF1	μF1	mF1	μF1	mF1	μF1	mF1	μF1	mF1
SVM	66.3	55.0	76.0	65.4	74.4	64.5	65.7	49.0	88	82.6	95.1	75.9	77.6	65.4
BERT	**71.4**	64	79.6	**78.3**	70.5	60.9	71.6	55.6	87.7	82.2	97.3	80.4	**79.7**	70.2
Legal BERT	71.2	**64.9**	**80.6**	77.2	76.2	65.8	**72.2**	**56.2**	**88.1**	**82.7**	97.4	**83.4**	**81.0**	**71.7**
CaseLaw BERT	71.2	64.2	79.7	76.8	**76.4**	**66.2**	71	55.9	88	82.3	**97.4**	82.4	80.6	71.3
Error Red. %	15.1	22	19.2	37.3	7.8	4.8	19.0	14.1	0.8	0.6	46.9	31.1	18.1	18.3

the LexGLUE tasks with more pronounced label imbalance, with just a few labels representing the vast majority of a dataset and certain classes having very few examples. This is also noticeable with a much larger increase in macro-F1 score when compared to the increase in micro-F1 score on the same tasks. This highlights the increased robustness of the BERT-based models, whose use of pre-training and transfer learning techniques allow them to reach certain levels of performance performance with much fewer examples than traditional methods such as SVM. As training in the legal domain can be difficult to acquire and label, this is an important step towards democratising legal NLP.

Despite this and their impressive performance in both the general domain and other specialised domains, it is still apparent that in the legal domain, BERT-based models, even with specific domain pre-training, produce only a modest improvement on the 6 evaluated LexGLUE tasks.

There is no clear explanation for this phenomenon, but we discuss multiple hypotheses (a, b, c). The first (**a**), 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 suitable, 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 a clear explanation for the overall weak improvement from deep learning over SVM classifiers.

A similar potential explanation (**b**) 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. Such tasks have now

started being explored, such as legal rationale extraction [11] or textual entailment in the form of a multiple choice QA task [4]. This QA task, *CaseHold*, is the final task of the LexGLUE benchmark, and shows a noticeably better increase in performance of Legal BERT models over General Domain BERT models, which supports the interest of further studies towards validating this hypothesis.

However, this explanation does not fully address the weak performance gains on text classification. A final hypothesis (**c**) 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.

4. Conclusion and Future Work

We experimentally confirm the intuition that SVM classifiers perform remarkably well on multiple legal text classification benchmarks. We notably highlight that the relative performance improvement between BERT-based models and SVM models is considerably smaller within the legal domain than on general domain classification tasks, even with BERT models specifically trained for the legal domain.

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.

We believe our results do not indicate the unsuitability of BERT-based approaches, but rather show that they have shortcomings and that they are perhaps better suited to more complex tasks. We also show that in the case of imbalanced datasets with very few examples for some of their classes, BERT-based models result in bigger increases in macro F1-score than in micro-F1 score, showcasing their ability to reach better results in low-data downstream tasks.

Future work will focus on exploring and pushing the limits of BERT models (and variants) within the legal domain, both by exploring more complex tasks and attempting to integrate external knowledge bases within them to improve performance on tasks such as text classification.

We hope our work will help support future work on Legal NLP focusing on exploring the specificities of legal text and better taking them into account. We make our experiments' code available to support future work.

5. Acknowledgements

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] Tuggener D, von Däniken P, Peetz T, Cieliebak M. LEDGAR: A Large-Scale Multi-label Corpus for Text Classification of Legal Provisions in Contracts. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. .

[8] Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz DM, et al.. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English; 2021. arXiv cs.CL 2110.00976.

[9] 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.

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

[11] 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;. .

[12] Chalkidis I, Fergadiotis M, Androutsopoulos I. MultiEURLEX–A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In: EMNLP 2021;. .

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

[14] Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview; 2020. arXiv stat.ML 2008.05756.