https://doi.org/10.1007/s42979-025-04300-x

ORIGINAL RESEARCH

Evaluating Shallow and Deep Learning Strategies for Legal Text Classification of Clauses in Non-Disclosure Agreements

Niall McCarroll[1] · Philip McShane[3] · Eoin O'Connell[3] · Kevin Curran[1] · Muskaan Singh[1] · Eugene McNamee[2] · Angela Clist[3] · Andrew Brammer[3]

Received: 3 March 2024 / Accepted: 3 August 2025

Abstract

A non-disclosure agreement (NDA) is a legal contract between at least two business parties that restricts the disclosure of confidential and sensitive information. As part of the NDA document negotiation and review workflows, there is a need to identify, extract and track clauses to ensure they comply with the companies' policies. The legal Natural Language Processing (NLP) landscape is rapidly evolving and offers a range of technologies and Machine Learning tools, such as text classification, that enable users to automate various key stages of the contract life cycle, thus easing the administrative burden. The proposed system was developed to classify individual clauses within an NDA contract. Automated legal text classification is challenging due to the high dimensionality of a word-based feature space. Additionally, corporate privacy concerns have limited the availability of publicly accessible legal corpora. To leverage the power of deep learning neural architectures, pre-trained word embeddings have attempted to resolve these issues by reusing an input feature space trained on a large, general-purpose dataset, and then fine-tuning it to adapt to a downstream classification task on a smaller annotated legal dataset. In this paper, we evaluate several shallow and deep supervised learning strategies for the classification of 26 individual mutually exclusive clause classes within an NDA contract. The impact of using pre-trained word embeddings on a small legal NDA contract dataset is also evaluated. The shallow SVM and XGBoost classification methods outperform the deep learning long short-term memory neural network (LSTM) approach in a small imbalanced dataset, even when supported by pre-trained embeddings. The potential of using novel transfer learning techniques that allow reuse of legal-domain specific NLP models and feature representation schemes is explored.

Keywords Legal text classification · eDiscovery · Predictive coding · Technology assisted review · Natural language processing · Machine learning · Deep learning · Word embeddings · Bi-LSTM · ELMo

Introduction

Information management has become one of the most significant challenges facing all businesses. Electronically stored information is growing at a rapid pace, with the Global DataSphere forecast predicting that the amount of data that will be created in the next two years will be more than all of the data created in past 30 years [1].

Lawyers and legal support teams can face overwhelming eDiscovery and eDisclosure review and evidence gathering tasks in large-scale litigation scenarios where documents can number in the millions. This ultimately requires the dedication of significant levels of resourcing, incurring huge costs that are passed on to the client. In a move to alleviate the administrative burden from legal professionals and deliver value for clients, there is increased momentum towards the adoption and evolution of technology and data-driven strategies that are designed to enhance the overall quality, speed and cost efficiency of document review processes [2, 3].

Law is built on a framework of language and most legal information used for activities such as contract review; document analysis and legal research is stored as unstructured,

✉ Kevin Curran
kj.curran@ulster.ac.uk

[1] School of Computing, Engineering and Intelligent Systems, Ulster University, Derry, Northern Ireland

[2] School of Law, Ulster University, Jordanstown, Newtownabbey, Northern Ireland

[3] Allen & Overy, Donegall Quay, Belfast, Northern Ireland

Published online: 29 August 2025

natural, free form text that is not readily accessible for computers. Natural Language Processing (NLP) offers a range of methods for exploiting patterns and regularities within in this unstructured data that can convert legal text into structured representations and document features that computers and algorithms can analyse and understand [4].

NLP and Machine Learning are at the core of Predictive Coding, which is one of the more recent developments in Technology Assisted Review (TAR) and is an umbrella term for several technologies and workflows that apply text classification algorithms to quickly cull through large volumes of text data to quickly locate target documents or, at a lower level, relevant sub-sections or clauses within a document. As these technologies have matured in recent years, they have begun to change the way the legal sector operates, signalling an era of increased automation with NLP being the key component to reviewing and understanding legal documents at scale [5, 6].

A non-disclosure agreement (NDA) is a legal contract between at least two business parties which restricts the disclosure of confidential and sensitive information that has been shared for a specific purpose. NDAs can be used to share commercial or trading information, share intellectual property or to formalise a relationship (e.g. between an employer and employee). Such contracts play a vital role in building relationships and create obligations among different parties. Usually, contracts can be complex and full of legal jargon. As part of the NDA contract workflows, there is a need to identify and track clauses to address any violation of the contract terms. An automated contract review process involves decomposing the documents into individual clauses for further assessment through key information extraction and then mapping and comparison against a playbook of gold standard terms [5].

Early contract review methods exploit the presence of key terms and headings to guide information extraction, making use of code-driven approaches such as regular expressions (regex), decision logic and other pre-programmed rule-based technology. While many of the modern offerings still avail of these methods, there has been a shift of focus towards sophisticated data-driven Machine Learning strategies that can identify patterns in complex data and develop their own logic and rules which are often too numerous and complicated to pre-programme by hand [7].

Performance on text classification tasks varies greatly and certain algorithms yield greater accuracy depending on factors such as the length or volume of text or characteristics of the textual content. However, many of the off-the-shelf solutions available on the market offer a black box system that is restricted to a single Machine Learning algorithm and static pre-processing parameters for developing a predictive model. In addition, they offer few guarantees of the performance of these models, meaning the bulk of the risk is transferred to the client. Realising the full benefits of understanding the best approaches to training algorithms and establishing robust pre-processing parameters is the best way to empower legal teams with the knowledge they need to implement efficient and reliable predictive coding as part of a wider eDiscovery or eDisclosure process [8].

Within this context, this paper outlines several different algorithms and pre-processing parameters applied to a novel, real-world dataset of NDAs containing various combinations of 26 different clauses. The comparative study assesses the performance of recent state-of-the-art models against that of traditional statistical models for the legal text classification challenge.

Related Research

Legal Text Classification

Text classification is a Supervised Machine Learning technique which assigns predefined categories to unstructured, natural language text. In the context of law, text classification is becoming increasingly important as the core functionality behind Technology Assisted Review (TAR), also known as Predictive Coding, which provides a suite of automated techniques including document classification, document clustering, document relevancy ranking and topic labelling, that enable lawyers to efficiently glean valuable insights from massive volumes of text data [9, 5].

Traditionally, popular shallow processing techniques such as rule-based systems and Machine Learning algorithms such as Logistic Regression, Support Vector Machines (SVMs), Decision Trees and Naive Bayes, have been considered robust baselines for a range of text classification tasks [10–15]. Despite their simplicity, shallow processing techniques often obtain state-of-the-art performances if the right features are adopted and also have the potential to scale to very large corpora [16].

Applying some of these techniques to legal text classification tasks, Chhatwal et al. (2016) evaluated SVM and Logistic Regression models on three real-world legal datasets. The aim was to gain a better understanding of the underlying techniques and pre-processing parameters required to make predictive coding more effective in the task of determining document relevancy in an eDiscovery task. Results showed that Logistic Regression outperformed the SVM model across all three datasets, finding that predictive models generated using SVM would require review of 1.6–3.3% more documents to achieve an equivalent average recall rate than the Logistic Regression models. Using the scenario of a one million document legal matter, the authors point out that the inefficiencies of a Predictive Coding system built from a SVM model would require an extra 16,000 to 33,000 additional documents for review before achieving an equivalent performance rate to a Logistic Regression system [8].

Shallow learning methods are largely restricted by the quality of sample features that they employ. Having considered the use of different complex features for traditional text classification models, the majority of researchers finally settled on simple bag-of-word unigrams or bigrams [17]. The main disadvantage of these methods is that they overlook contextual information and sequential text structure thus disregarding the semantic information contained within text [18]. This has the potential to limit their generalisability to large output feature spaces where data sparseness can lead to certain classes having a small number of examples [14, 19].

To address this, recent approaches to NLP have centred around the use of deep learning architectures to achieve Language Modelling (LM), which utilises various probabilistic and statistical techniques to infer the probability of a given sequence of words occurring in a sentence given rules for language context. Therefore, these methods automatically provide semantically meaningful representations to facilitate text analysis and mining. Zhang, et al. (2016) present a Convolutional Neural Network (CNN) model for text classification and adopt a granular labelling approach that jointly utilises labels of both documents and their constituent sentences that add further support to the overall document classification. This two-level learning approach was applied to five datasets that had document-level labels and the requisite sentence-level labels. The authors found that the CNN model consistently outperforms strong SVM variant baselines across five different datasets [20].

Assessing the effectiveness of deep learning in a legal context, Wei et al. (2018) train a CNN that uses GloVe embedding inputs to the neural network on a text classification task on four separate datasets of real legal data. The authors found that the deep learning system outperformed a baseline SVM algorithm but only when trained on the datasets that contained larger volumes of legal data. There was a difference of 62,155 training samples on the largest dataset compared to 11,935 training samples on the smallest dataset [21].

The application of Recurrent Neural Network (RNN) language models for learning sequential context has been widely researched [22, 23]. However, parallelisation is not possible with RNN architectures, making scaling to large corpora or longer sequence length texts challenging. Transformer and Bi-LSTM architectures facilitate significantly more parallelisation along with bi-directional context awareness and are thus more suited to a wide variety of tasks including text classification [24, 25].

Legal text classification remains a relatively unexplored area of deep learning research. Classification tasks, particularly in the deep learning domain, require large quantities of training data. However, the construction of large training sets is very costly and time consuming, and requires the involvement of legal practitioner experts, who have a thorough knowledge of the text and terminology, to label data [6]. Privacy concerns are another barrier to building legal text classification systems as it has led to a dearth of publicly available annotated legal data. This low data scenario has stifled the big data revolution in legal Machine Learning compared to other areas of research. To address this, there has been a shift of focus to transfer learning methods which reuse pre-trained models as a starting point to be repurposed for a second related task [26, 27].

Transfer Learning

Within the NLP frameworks, one approach to developing deep architectures for specific language tasks, has been to exploit unsupervised feature representation transfer learning from large datasets of general-purpose data such as Wikipedia or Google News corpus, and reapply them to smaller annotated datasets [28, 29].

Fixed *feature-based* and *fine tuning* are the two main strategies for applying pre-trained deep neural language models to downstream classification tasks. In Machine Learning, a representation scheme transforms data from its original representation (e.g. unstructured text) to a new representation (e.g. a term-document frequency matrix) that retains essential information that is necessary for the task at hand. Feature-based approaches reuse a fixed input representation scheme that has been pre-trained on a large generic text corpus or multiple related corpora. However, these 'out-of-the-box' models have been reported to underperform in specialised domains as they are highly skewed towards generic language [30, 31].

More recently, fine-tuning approaches have been introduced which reuse a feature space trained on a larger dataset and adapt them for classification tasks on smaller annotated datasets by simply re-learning the weights in single or multiple layers of the deep neural architecture. This has been successfully demonstrated in the ELMo [32], BERT [33], GPT [34] and ULMFiT [35] models. These models have achieved impressive performance gains across many natural language processing tasks, such as question answering, natural language inference, sequence labelling and text classification. Models leveraging these pre-trained embeddings have achieved competitive results on all nine of the GLUE natural language understanding benchmarks [36], the SQUAD [37] and RACE [38] generic benchmark datasets and have also surpassed human baselines in certain tasks [39].

Fine-tuning strategies have had varying success within legal text classification tasks. Duan et al. (2019) developed a multi-class legal text classification model for a legal question and answering system using the BERT-Base Chinese pre-trained model [40] that is then fined tuned on training data sourced from the Chinese Judicial Reading Comprehension (CRJC) that was also created for the project. Whilst the authors report that the BERT model is a useful baseline, they do concede that it significantly underperforms compared to human annotator performance. Performance deficits can be attributed to a relatively small dataset of approximately 10,000 documents and the limitations of a generic BERT-Base Chinese model being deployed in specialised domain of a legal use case [41].

Bartolo et al. (2019) demonstrate the successful transfer of feature-based pre-trained language models to a legal domain task of litigation code classification. The authors built upon the BERT framework [33] and developed it further by fine-tuning the pre-trained parameters to identify and extract context from short legal narratives before labelling with appropriate Judicial Codes (J-Codes). The feature-based model achieved significant performance gains over the next best performing SVM classifier [42].

Within specialised domains, such as law or biomedical science where language can be nuanced, pre-trained models have had varying degrees of success depending on the additional effort employed to adapt them to low coverage scenarios such as unusual terminology or phrase structures [43, 44]. Attempts to address this have proposed using sub-word information to enrich vectors with additional morphological context to ‘fill in the gaps’ and extrapolate from root words and stems [45, 46].

These pre-trained approaches are now key components in many natural language applications. However, their performance has been challenged by human-level deep learning performance on large domain-specific datasets and by shallow approaches, such as SVM and Bayesian methods, outperforming when training on small dataset scenarios [47, 48].

Table 1 Breakdown of the number of training, validation and testing sentences contained within the 66,567 sentence real-world NDA dataset

# of Training Examples	# of Validation Examples	# of Testing Examples
49,393	7,973	9,201
Overall total of examples		**66,567**

Research Challenge

Text classification is challenging due to the high dimensionality of word-based feature space. Ideally this can be resolved through training complex models to perform feature selection and feature extraction on large datasets. Solving this problem in the legal text domain has been difficult due to a lack of sharing of large publicly available corpora over privacy concerns. As training of deep learning neural architectures on small datasets may lead to overfitting, pre-trained word embeddings might help to resolve these issues by reusing an input feature space trained on a large general-use dataset and fine-tuning it to adapt to a downstream classification task on a smaller annotated legal dataset.

Therefore, given (1) the recent success of training deep neural models, supported by pre-trained contextualised embeddings, in small data experiments and (2) the competitive performance of shallow statistical methods against deep learning approaches in these scenarios, this paper will compare robust shallow baseline standards against deep neural learning methods in the task of multi-class legal text classification of NDA clauses. To the best of our knowledge, this is the first study to evaluate shallow and deep learning text classifiers and the impact of pre-trained embeddings with subword information in a legal text classification context using a novel real-world dataset of NDA clauses.

Materials and Methods

Data

An in-house team of expert legal professionals manually annotated a dataset of 675 NDA contracts - split into training (495), validation (93) and testing (87) subsets for the supervised learning task. These documents contained a total of 66,567 sentences, each of which could be part of one of 26 different clauses to be extracted in the multi-class categorisation task.

Table 1 gives a breakdown of the number of training, validation and testing sentences within the NDA dataset.

Figure 1 shows the percentage breakdown of training, validation and testing sentences within each clause category. They are ordered from highest to lowest number of training samples across the clauses. The ‘Other’ or non-clause class category contains a significantly greater number of sentences compared to the clause class categories. The breakdown of sentences for the ‘Other’ category is training (29,290 (59.3%), validation (5,490 (68.9%)) and testing (5,108 (55.5%). To aid visualisation of the breakdown of train, validation and test samples for each clause, the ‘Other’ category is removed from Fig. 1.

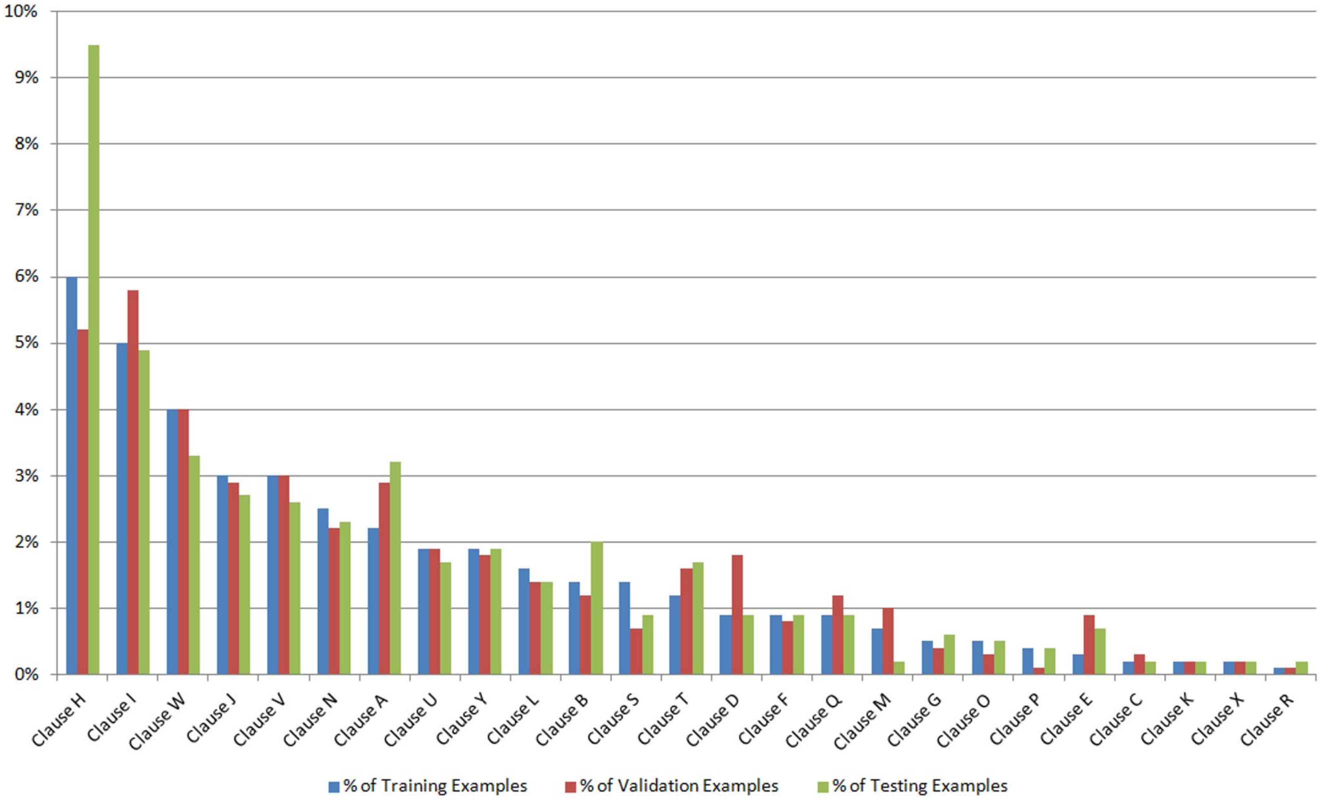

Fig. 1 Class categories and percentage of labelled example sentences of the 26 NDA clauses identified by the legal professional team during the manual annotation task. Each category is broken down into colour coded percentages of training (blue), validation (red) and testing (green) samples. Clauses are ordered from highest to lowest number of training samples across the categories. Note that clauses are anonymised for commercial privacy

Preprocessing

Figure 2 gives a high-level overview of the text classification pipeline. To begin with, input text was pre-processed and segmented using a customised rule-based regex algorithm that deals with sentence delimiter tokens and common abbreviations. Standardised text cleansing techniques of white space removal, lowercasing, tokenisation, stop word removal and punctuation removal were also applied.

Label Imbalance

Label imbalance was present in the dataset. To address this issue, several steps were taken. The ‘Other’ class contains a significantly greater number of samples than the rest of the classes. As these contracts make use of standard language a certain amount of the samples are duplicates which could be removed without removing any information from the dataset. A combined oversampling and undersampling strategy was applied to address imbalances within the dataset. For those classes with the least number of samples, new examples of those clauses were drafted by the in-house Legal Professional team to allow for more of them to be seen at training time. Additionally, these classes were oversampled so that each example of them is seen more often during an epoch. A weighting was also applied to the classes to equalize their importance, with the exception being the ‘Other’ class which was weighted lower to reduce its importance during training.

Models

To demonstrate any improved performance from the use of pre-trained contextual embeddings on this domain specific task we perform benchmark comparisons against a variety of different baseline models. Three models were evaluated for the shallow learning strategies - Support Vector Machine, Naive Bayes and XGBoost. All three classification models were trained on n-gram (n ranges from 1 to 3) bag-of-words representations of the input documents using Term Frequency - Inverse Document Frequency (tf-idf) to convert the text into feature vectors. We prefer n-gram frequencies to single-token frequencies because combinations of words are more distinct between different contexts/classes.

The deep learning approach adopts an ELMo-LSTM framework. This framework is a stacked recurrent neural

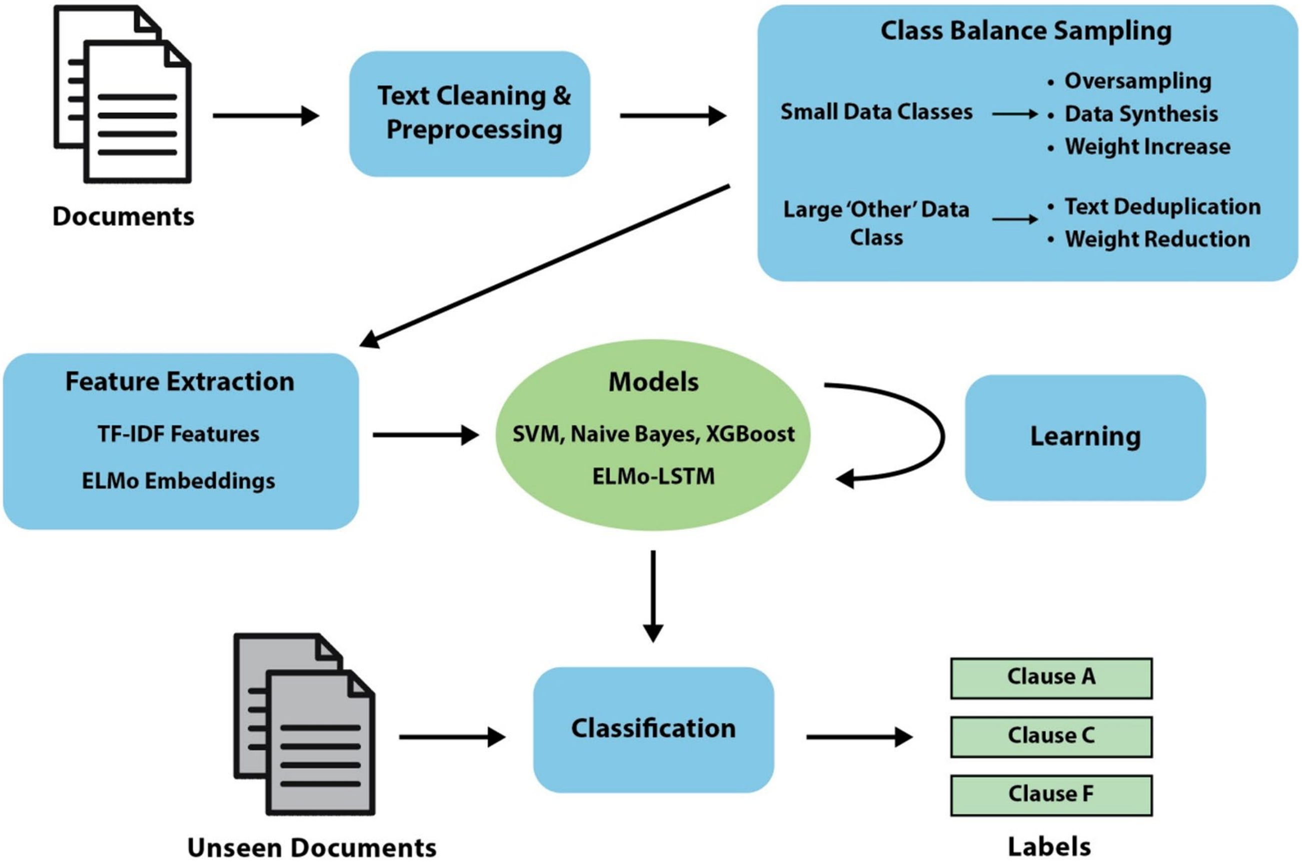

Fig. 2 Overview of the key processes of the NDA clause classification pipeline

network (RNN) architecture - bidirectional LSTM layer, a LSTM layer followed by a feedforward output layer for prediction. The input text was represented as a sequence of pre-trained contextualised ELMo embeddings.

ELMo (Embeddings from Language Models) is a NLP framework developed by AllenNLP. ELMo uses a deep, bi-directional LSTM which is pretrained on the 1 billion Word Benchmark news crawl data from WMT 2011. ELMo models polysemy by analysing words within the different linguistic contexts that they are used. It is also character based, enabling the model to leverage morphological clues to form representations of out-of-vocabulary words not seen in training [32].

Adopting a fine-tuning strategy, the ELMo representation schemes that have been pre-trained on a larger generic dataset are reused and fine-tuned in the downstream training phase using the smaller NDA clause dataset.

Table 2 A comparison of accuracy performance between the different shallow and deep learning classification models on the NDA contract clause extraction task

	SVM	Naive Bayes	XGBoost	ELMo-LSTM
Train Accuracy	0.865	0.737	0.803	0.767
Val. Accuracy	0.803	0.709	0.777	0.759
Test Accuracy	0.810	0.712	0.763	0.760

Results and Discussion

The SVM classifier showed the highest overall accuracy and both the SVM and XGBoost shallow machine learning strategies outperformed the deep learning LSTM approach using the small imbalanced dataset, even when supported by contextualised pre-trained embeddings (See Table 2).

The Deep learning strategy failed to capture the term diversity due to the small training dataset. Again, the weakest performance by the Naive Bayes model can be attributed to the smaller amount of available data which tends to keep precision and recall low.

Figure 3 shows a breakdown of the Precision, Recall and F1 metrics for the top performing SVM model across each individual clause of the Test NDA dataset. The results are ordered from highest to lowest F1 performance across the

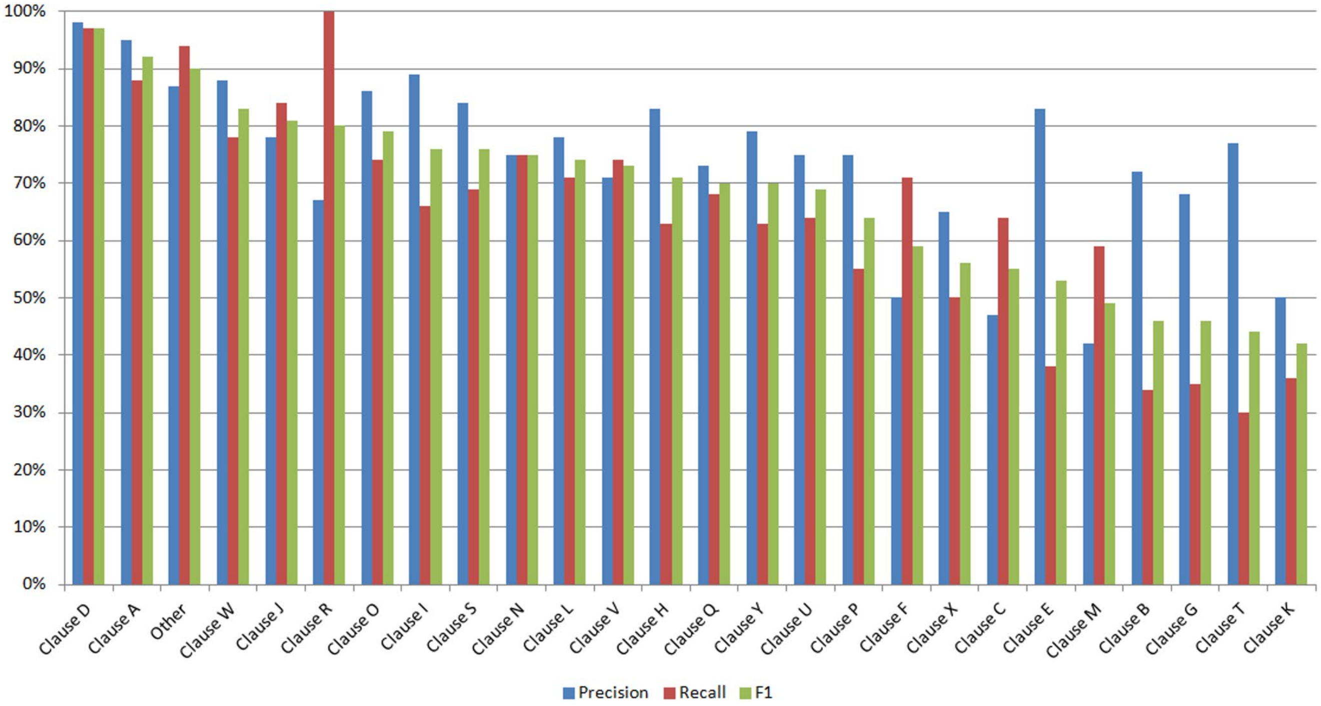

Fig. 3 Percentage precision (blue), Recall (red) and F1 (green) metrics for the top performing SVM model across each individual clause of the test NDA dataset. Clauses are ordered by highest to lowest F1 performance across the class categories

range of clauses. The variation in F1 predictive performance ranges from a 97% highest performance (Clause D) to a 42% lowest performance (Clause K).

Overall, the weighted average performance across all clause categories was Precision (84%), Recall (82%) and F1 (82%). The similar Precision and Recall scores indicate that the model is performing well with a good balance between correctly identifying positive cases and finding all relevant positive instances. This suggests the model is both accurate in its positive predictions and comprehensive in its ability to capture all positives.

Shallow ML models can still outperform deep learning methods, even when augmented with pre-trained embeddings. A high textual diversity compared with the amount of available balanced training data contributed to some labels having less associated data and thus decreasing accuracy. The effectiveness of pre-trained methods to incorporate prior knowledge and learn on low resource data has not been realised in this study compared to other empirical investigations. An important contributing factor to this lower performance compared to the baseline standards may be attributed to the linguistic differences between the source text data for the pre-trained embeddings and legal narrative of the target NDA contract use case.

Conclusion and Future Work

The ELMo-LSTM framework, using a representation scheme pre-trained on a larger generic corpus and fine-tuned on the smaller legal dataset failed to achieve performance gains over the best performing baseline.

There are challenges when it comes to adapting deep learning neural classifiers for legal domain tasks. Legal text differs in structure from the common, everyday language that is used in Wikipedia and other mass sources of text data that are used to train these state-of-the-art deep learning neural models. Indeed, the developers of the LawGeex Contract Review Platform indentify this complex, counter-intuitive and technical “legalese” as being the major barrier to Machine Learning fully understanding contracts, as no existing NLP computational language models or off-the-shelf solutions can read legalese coherently [49]. They go so far as to develop research around this legal ‘language’ and have created proprietary Legal Language Processing (LLP) and Legal Language Understanding (LLU) models for contract processing tasks, mirroring the success of similar approaches to text mining and information retrieval in the biomedical domain with the development of Biomedical Natural Language Processing (BioNLP) [46, 48, 49].

Novel transfer learning techniques that allow reuse of legal-domain specific NLP models and feature representation schemes may help bridge the gap between the performance of deep learning neural models on large and small datasets. Thus, the potential way forward is the training of bespoke task dependent embedding models, a direction that has achieved some success in the medical domain [26, 30] and which could be applied across a variety of tasks in the legal domain.

In an attempt to capture legal domain specific feature representations, Chalkidis and Kampas (2019) were the first researchers to make publicly available a word embedding resource trained exclusively on large legal corpora. The Law2Vec model utilises corpora from a range of public legal sources amounting to a total of 123,066 documents consisting of 492 million individual word tokens [44]. And most recently, Chalkidis et al. (2020) have introduced LEGAL-BERT, which is a family of BERT models pre-trained on 12GB of diverse, publicly available legal text from several fields to assist in computational law and the application of legal NLP to technology applications [50]. As the authors have successfully demonstrated that integrating domain knowledge into pre-trained models enhances the reasoning ability between legal concepts, these dedicated legal frameworks have accelerated progress in this research area.

To fully exploit the power of deep learning systems, they must be trained on significant volumes of data. Given the justified concerns over the privacy of resources, such as contracts within the commercial law setting, access to sufficiently large quantities of this type of data is not likely to happen any time soon. Possible solutions for dealing with low data scenarios include text augmentation strategies to synthesise data to overcome privacy restrictions and artificially enlarge the amount of available training data [51].

Additional effort will be required to address small data scenarios and nuanced legal terminology, particularly as we seek to adapt and evolve solutions from an experimental setup to commercially viable solutions. However, the robust performance of traditional statistical models indicates that there is sufficient exploitable information within the legal texts for both shallow and deep learning methods to perform legal NLP tasks as well as, if not better than, natural language.

References

1. Reinsel JF, Rydning D, Grantz J. Worldwide Global DataSphere Forecast, 2020–2024: The COVID-19 Data Bump and the Future of Data Growth [Internet]. 2020 [cited 2025 Aug 20]. Available from: https://www.idc.com/getdoc.jsp?containerId=US44797920
2. Nicholas LZ, Pace M. Where the Money Goes Understanding Litigant Expenditures for Producing Electronic Discovery, 2011. [Online]. Available: http://www.rand.org/content/dam/rand/pubs/monographs/2011/RAND_MG996.pdf
3. Chhatwal R, Gronvall P, Huber-Fliflet N, Keeling R, Zhang J, Zhao H. Explainable text classification in legal document review A case study of explainable predictive coding. Proc - 2018 IEEE Int Conf Big Data. 2018;Big Data 2018:1905–11. https://doi.org/10.1109/BigData.2018.8622073.
4. Mustafi J. Stud Big Data. 2016;17:53–74. https://doi.org/10.1007/978-3-319-27520-8_4. Natural Language Processing and Machine Learning for Big Data.
5. Dale R. Law and word order: NLP in legal tech. Nat Lang Eng. 2019;25(1):211–7. https://doi.org/10.1017/S1351324918000475.
6. Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M. How does NLP benefit legal system: A summary of legal artificial intelligence. Proc Annu Meet Assoc Comput Linguist. 2020;5218–30. https://doi.org/10.18653/v1/2020.acl-main.466.
7. Hildebrandt M. In: Deakin M, editor. Code-driven law: freezing the future and scaling the past, in is law computable?? No. August 2018. Ed. Hart Publishing; 2020.
8. Chhatwal R, Huber-Fliflet N, Keeling R, Zhang J, Zhao H. Empirical evaluations of preprocessing parameters' impact on predictive coding's effectiveness. Proc – 2016 IEEE Int Conf Big Data. 2016;Big Data 2016(no i):1394–401. https://doi.org/10.1109/BigData.2016.7840747.
9. Roitblat HL, Kershaw A, Oot P. Document categorization in legal electronic discovery: Computer classification vs. manual review. J Am Soc Inf Sci Technol. 2010 Jan;61(1):70–80 https://doi.org/10.1002/asi.21233
10. McCallum A, Nigam K. A comparison of event models for naive Bayes text classification. In: AAAI/ICML-98 Workshop on Learning for Text Categorization. 1998; p. 41–8. doi:10.1.1.46.1529.
11. Joachims T. Text categorization with support vector machines: Learning with many relevant features, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 1998;1398:137–142. https://doi.org/10.1007/s13928716
12. Wilcox A, Hripcsak G. Classification algorithms applied to narrative reports. Proc AMIA Symp. 1999:455–9.
13. Yang Y, Liu X. A re-examination of text categorization methods. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999). 1999:p. 42–9. https://doi.org/10.1145/312624.312647
14. Nallapati R, Manning CD. Legal docket-entry classification: Where machine learning stumbles. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008). 2008 Oct;438–46 [Online]. Available: https://aclanthology.org/D08-1046.
15. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! Long live rule-based information extraction systems! in EMNLP 2013–2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, Oct. 2013; pp. 827–832, [Online]. Available: https://aclanthology.org/D13-1079
16. Wang S, Manning CD. Baselines and bigrams: Simple, good sentiment and topic classification, in 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012

- Proceedings of the Conference. Jul. 2012;2:pp. 90–94, [Online]. Available: https://aclanthology.org/P12-2018

17. Dumais S, Piatt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Proceedings of the International Conference on Information and Knowledge Management (CIKM 1998). 1998 Jan;148–55. https://doi.org/10.1145/288627.288651
18. Li Q, et al. A survey on text classification: From shallow to deep learning. arXiv. 2020; abs/2008.0 [Online]. Available: http://arxiv.org/abs/2008.00364
19. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification, in 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference. Apr. 2017;2:427–431. https://doi.org/10.18653/v1/e17-2068
20. Zhang Y, Marshall I, Wallace BC. Rationale-augmented convolutional neural networks for text classification, in EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings. Nov. 2016:795–804. https://doi.org/10.18653/v1/d16-1076
21. Wei F, Qin H, Ye S, Zhao H. Empirical study of deep learning for text classification in legal document review. Proc - 2018 IEEE Int Conf Big Data Big Data 2018. 2018;3317–20. https://doi.org/10.1109/BigData.2018.8622157.
22. Mikolov T, Kombrink S, Deoras A, Burget L, Černocký J, Autom D. RNNLM --- Recurrent Neural Network Language Modeling Toolkit, in Proceedings of ASRU. pp. 1–4, [Online]. Available: https://www.microsoft.com/en-us/research/publication/rnnlm-recurrent-neural-network-language-modeling-toolkit/
23. Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of Recurrent Network architectures, in 32nd International Conference on Machine Learning, ICML 2015. 2015;3:2332–2340.
24. Vaswani A et al. Attention is all you need, in Advances in Neural Information Processing Systems, 2017;2017-Decem, pp. 5999–6009.
25. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: Attentive language models beyond a fixed-length context, ACL 2019–57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. pp. 2978–2988, 2020, https://doi.org/10.18653/v1/p19-1285
26. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc. 2019 Nov;26(11):1247–54. https://doi.org/10.1093/jamia/ocz149
27. Soh J, Lim HK, Chai IE. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In: Proceedings of the Natural Legal Language Processing Workshop. 2019 Jun;67–77. https://doi.org/10.18653/v1/w19-2208
28. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic Language model. J Mach Learn Res. 2003;3(6):1137–55. https://doi.org/10.1162/153244303322533223.
29. Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59. https://doi.org/10.1109/TKDE.2009.191.
30. Lee J, et al. BioBERT: A pre-trained biomedical Language representation model for biomedical text mining. Bioinformatics. Feb. 2020;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.
31. Beltagy I, Lo K, Cohan A. SCIBERT: A pretrained language model for scientific text, in EMNLP-IJCNLP 2019–2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, Nov. 2019; pp. 3615–3620. https://doi.org/10.18653/v1/d19-1371
32. Peters ME, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018). 2018 Jun;1:2227–37. https://doi.org/10.18653/v1/n18-1202
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for Language Understanding. NAACL HLT 2019–2019 Conf North Am Chapter Association Comput Linguistics: Hum Lang Technol - Proc Conf. 2019;1:4171–86.
34. Radford A, Sutskever I. Improving Language Understanding by generative Pre-Training. Homology Homotopy Appl. 2007;9(1):399–438.
35. oward J, Ruder S. Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Long Papers. 2018;1:328–39. https://doi.org/10.18653/v1/p18-1031
36. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR. GLUE: a multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 1st Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (EMNLP 2018). 2018 Nov;353–5. https://doi.org/10.18653/v1/w18-5446
37. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2016). 2016 Nov;2383–92. https://doi.org/10.18653/v1/d16-1264
38. Lai G, Xie Q, Liu H, Yang Y, Hovy E. RACE: Large-scale ReAding comprehension dataset from examinations. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). 2017. p. 785–94 https://doi.org/10.18653/v1/d17-1082
39. He P, Liu X, Gao J, Chen W. Deberta: Decoding-Enhanced Bert With Disentangled Attention, ICLR 2021–9th Int. Conf Learn Represent. 2021;vol. abs/2006.0, [Online]. Available: https://arxiv.org/abs/2006.03654
40. He W, et al. DuReader: A Chinese machine reading comprehension dataset from real-world applications. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. 2018 Jul;37–46 https://doi.org/10.18653/v1/w18-2605
41. Duan X, Liu Y, Sun Y, Zhou Q, Xie H, Wang S, et al. CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2019;11856:439–51. https://doi.org/10.1007/978-3-030-32381-3_36
42. Bartolo M, Tylinski K, Moore A. Pre-trained contextual embeddings for litigation code classification. CEUR Workshop Proc. 2019;2484:38–45.
43. Arnold S, Gers FA, Kilias T, Löser A. Robust named entity recognition in idiosyncratic domains, Aug. 2016, [Online]. Available: http://arxiv.org/abs/1608.06757
44. Chalkidis I, Kampas D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artif Intell Law. 2019;27(2):171–98. https://doi.org/10.1007/s10506-018-9238-9.
45. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–46. https://doi.org/10.1162/tacl_a_00051.
46. Chen Q, Peng Y, Lu Z. BioSentVec: Creating sentence embeddings for biomedical texts. In: 2019 IEEE International Conference on Healthcare Informatics (ICHI). 2019 Jun. https://doi.org/10.1109/ICHI.2019.8904728
47. yoshua bengio. Deep learning. MIT Press. 2019;29:7553.

48. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and mesh. Sci Data. 2019;6(1):1–9. https://doi.org/10.1038/s41597-019-0055-0.

49. blog Lawgeex, Comparing the performance of artificial intelligence to human lawyers in the review of standard business contracts, 2018. [Online]. Available: www.lawgeex.com.

50. Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020. 2020 Nov;2898–904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

51. Bayer M, Kaufhold MA, Reuter C. A survey on data augmentation for text classification. ACM Comput Surv. 2022;55(7). https://doi.org/10.1145/3544558.