Introducing KAN Block with BiGRU for Legal Document Classification and Summarization

Anonymous ACL submission

Abstract

This work introduces a new deep learning based solution for legal document classification and summarization (LDCS), which is capable of tackling several challenging issues, including domain-specific language, long-term dependencies and class imbalance. To make model better process complex legal texts, a Kolmogorov–Arnold Network (KAN) block is added to the model. Extensive experiments on a legal dataset show the effectiveness of the proposed approach compared to standard ML models (SVM and logistic regression), with 67% classification accuracy, weighted F1-score of 0.65, and summarization ROUGE-1 F1-score = 0.38. These findings demonstrate the promise of the approach for further developing automated legal analysis and decision support systems.

Keywords : Legal Document Classification, Legal Text Summarization, BiGRU, KAN Block, Class Imbalance, Deep Learning

1 Introduction

1.1 Motivation and Background

Natural language processing (NLP) has encountered large progress in recent years due to the broad adaptation of deep learning models over different tasks. Nonetheless, there are still open issues in document classification and text summarization even in the field of legal documents. Legal texts such as case notes, rulings and judgments tend to contain legal language which is difficult and "heavy" very domain specific. Extracting good stories about that news, however, necessitates efficiency and accuracy. Existing approaches to text classification and summarization have struggled with these document types, largely stemming from the difficulty in dealing with complex relationships and domain-specific terminology.

The emergence of large legal corpora demands efforts in models that can automatically classify and summarize the contents of legal texts, while preserving essential information to assist legal professionals in decision making. In this paper, we concentrate on improving the such as document classification and summarization in the legal domain by introducing new methods using contemporary deep learning approaches.

1.2 Objectives and Contributions

In this paper, we address the tasks of legal text classification and summarization by proposing an enhanced deep learning model. Specifically, we utilize **two layers of BiGRUs** and leverage their contextual information through both mean and max pooling techniques. We also introduce the **KAN block** (Spline Linear Layer), which provides a new perspective to regularize and fine-tune the model's performance in legal document classification.

Our contributions are threefold:

- **Better Classification Ability**: We design a more powerful classifier by incorporating new pooling methods with the BiGRUs and KAN block, resulting in significant improvements in classification accuracy.
- **Dealing with Class Imbalance**: We address information loss from minority classes, which is frequently encountered in legal corpora, by using the **WeightedRandomSampler**.
- **Advanced Summarization Model**: We introduce a powerful summarization model, namely *AttnGRUSummarizerKAN* for legal text summarization. The model here is making output summaries "better" (at least in terms of ROUGE scores) without losing much important from the input text.

1.3 Orientation of the Paper

The remainder of the paper is structured as follows:

- **Section 2** provides an overview of related work in the field of legal document classification and summarization.
- **Section 3** describes the processing steps of the Dataset used in this study.
- **Section 4** describes the methods of the whole process of conducting this experiment.
- **Section 5** presents the experiments and results where our proposed methods are compared with baseline approaches.
- **Section 6** discusses the conclusion, limitations, and directions for future work.
- **Section 7** contains the challenges and limitaions that we faced during the experiment.
- **Section 8** concludes the paper and summarizes our contributions.

2 Related Work

In the field of legal document processing, research addresses two closely related tasks: text classification and summarization. Early text classification methods utilized traditional machine learning techniques, such as support vector machines (SVM) and logistic regression, often with hand-crafted features (Cohen and Yang, 2003; Aletras and Stevenson, 2016). While these methods worked well on small datasets, they struggled to capture long-range dependencies and semantic relationships within legal texts.

To overcome these limitations, deep learning methods, particularly recurrent neural networks (RNNs) and their variants, have become popular. Bidirectional gated recurrent units (BiGRUs) and long short-term memory (BiLSTM) networks have been shown to effectively capture sequence information by processing text in both forward and backward directions (Schuster and Paliwal, 1997; Chung et al., 2014). Additionally, pooling techniques such as mean and max pooling have been incorporated to obtain fixed-length representations from variable-length sequences, significantly improving performance on text classification tasks (Conneau et al., 2017). However, these models still face challenges in handling domain-specific syntax and semantics in legal data, particularly due to the highly imbalanced class distributions typically found in legal corpora.

For text summarization, the encoder-decoder framework with attention mechanisms has emerged as the prevailing approach (Bahdanau et al., 2015; See et al., 2017). Models such as the pointer-generator network, which combines the ability to copy and generate text, are especially useful for abstractive summarization tasks that require maintaining significant details (See et al., 2017). In legal applications, adaptations of these models have demonstrated improvements over extractive summarization baselines. However, these models can still fail to capture nuanced legal reasoning or generate overly general summaries.

In recent work, a new class of models has focused on enhancing the expressiveness and interpretability of neural networks by introducing novel architectural components. Kolmogorov–Arnold Networks (KANs) are a prime example, replacing the traditional multilayer perceptron (MLP) with learnable activation functions on edges, as inspired by the Kolmogorov–Arnold representation theorem. In KANs, each network weight is replaced by a univariate function parameterized as a spline, which has been shown to achieve comparable or superior performance with increased interpretability in scientific function approximation and other domains (Liu et al., 2024). This shift from fixed activations on nodes to adaptive spline functions on edges offers a fresh perspective on neural modeling, potentially improving the ability to handle complex, non-linear patterns.

Building on this recent work, we propose incorporating KAN blocks into a BiGRU-based architecture for legal document classification and summarization. To the best of our knowledge, this is the first application of KAN-inspired layers to legal text understanding. Additionally, we address common real-world data challenges, such as class imbalance, by implementing sampling strategies that ensure all classes are represented during training without over-representing the more frequent classes.

3 Dataset

3.1 Dataset Collection

The dataset we use for our study was sourced from the legal text repository available at Manupatrafast(Manupatrafast, 2026), which consists of various documents like case notes, rulings, and judg-

ments. The corpus used is a set of legal texts from different domains, mainly court decisions and law cases. This is a widely-used tool in the legal community, with core source materials for both students and faculty to conduct research and analysis. The collection includes some crucial knowledge and document-oriented features such as the legal context, sections of acts,laws,rules,orders,judgements, and the summary context.

3.2 Dataset Processing

A number of preprocessing steps were applied after the data acquisition for validating the quality of data as well as preparing it suitable for machine learning.

Preprocessing: The missing and incomplete entries were handled first, then the irrelevant columns were eliminated. "Disposition" and "Case Note" columns were specifically scrubbed by transforming all 'nan', 'null', and 'none' to missing in all fields.

Text Normalization: Text normalization was conducted on "Relevant Section" and "Judgement". This included lowercasing every piece of text, removing useless characters (i.e., punctuation), and splitting the input text into tokens.

Feature Engineering: After cleaning the textual data, some new features were created. For example, one new column was created to calculate the duration for each note in word-count, which could then be used to analyze text length as part of EDA.

Dealing with Domain Terms: Legal documents exist in a special legal vocabulary that makes little sense for models to process properly. Key words were attempted to be retained and frequent stop words or noise words impeding understanding of the case were eliminated.

Text Tokenization: We tokenized the "full_text" column based on custom rules to represent each different part of the legal documents as words, helping the model understand and effectively use the text for downstream tasks.

3.3 Dataset Partitioning

The data was first pre-processed and then split into partitions for training, validation, and testing models as follows:

Training Set: Most of the dataset was used to train the model with enough data so the model could learn from diverse legal texts.

Validation Set: A subset of the data was reserved for validation, folded in during model building to optimize search hyperparameters and to avoid overfitting.

Test Set: Lastly, the test set was used to evaluate the model on unseen examples, simulating real-world applications.

These partitions supported model generalization validation, as we could have class imbalance or domain-specific issues in legal texts.

3.4 Dataset Statistics

To dive deep into the data, we performed analysis during EDA. Key findings include:

- The target ("Disposition") is extremely imbalanced and this had to be specially treated using weighted sampling. The Distribution of the Disposition class is visually shown in the Figure 1.

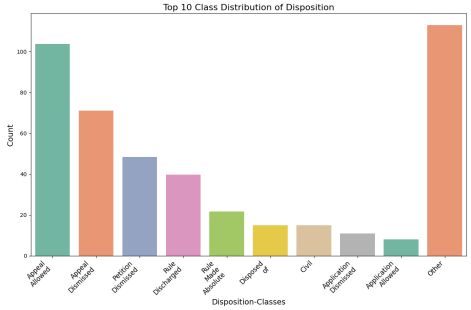

Figure 1: Class Distribution of Disposition

- Average lengths of text vary for "Case Note", a problem for which we resort to sequence truncation and padding the length is shown in Figure 2.

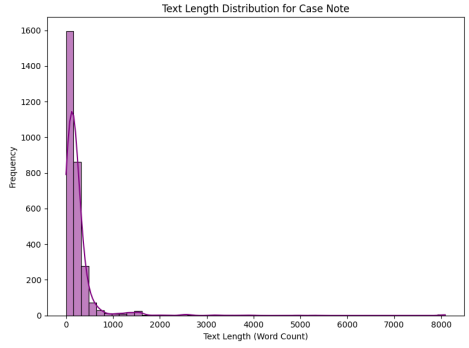

Figure 2: Distribution of Text Lengths for Case Notes

- Visualizations such as word clouds (Figure 3), missing text heatmap (Figure 5) and correlation matrix (Figure 4) of numeric features

made it much easier to understand the characteristics of all provided data and, in turn, guided us to make better decisions about model architecture.

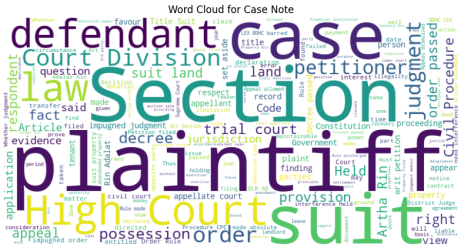

Figure 3: Word Cloud of Most Frequent Words in Case Notes

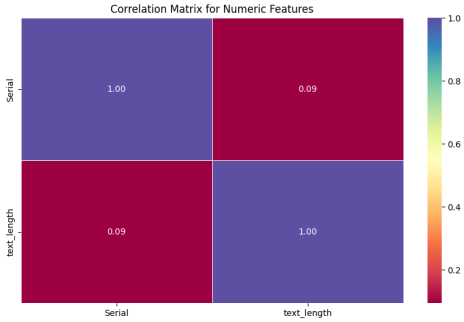

Figure 4: Correlation Matrix of Numerical Features

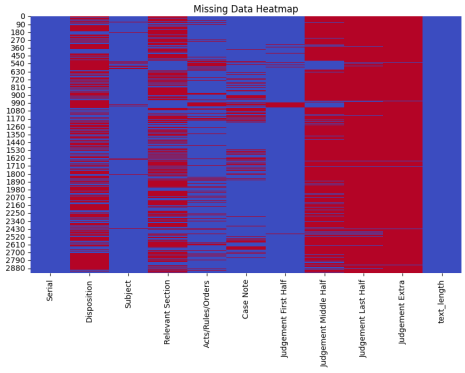

Figure 5: Missing Data Heatmap

4 Methodology

We present in this section the approach followed towards text classification and summarization of legal documents. We introduce a deep learning model, BiGRU with novel KAN (Kolmogorov–Arnold Network) block, for the purpose of feature extraction and classification. We also provide the architectural description of our models and discuss design decisions and their benefits towards legal text processing.

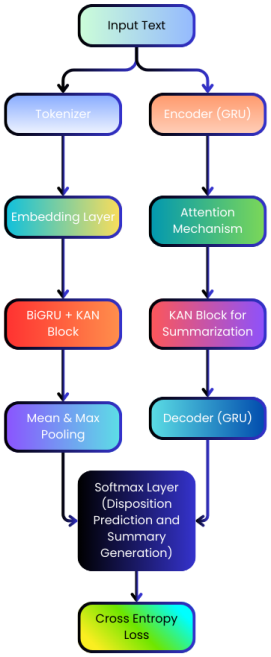

Figure 6: Proposed Methodology for Classification and Summarization.

4.1 Proposed Methodology

Our model is intended to handle the constraints of legal texts, e.g. domain-specific language, long-range dependencies, and class imbalance. We present two models for classification and summarization respectively, both based on the latest deep learning technology. The proposed methodology is shown in Figure 6

4.2 Training the Models

The models are learned with the following configuration:

- **Loss Function:** We use the cross-entropy loss for classification, and sequence-to-sequence loss (with label smoothing) for summarization. Cross-entropy loss punishes wrong predictions and label smoothing has been useful to regularize the model.
- **Optimization:** We optimized both models using the AdamW optimizer, which adjusts the learning rate with respect to the norm of the loss gradient. Such an optimizer is particularly beneficial for processing sparse gradients and large legal corpora as in this study.
- **Learning Rate Scheduling:** A cosine annealing learning rate schedule with warm-up steps

is used. This approach introduces decreasing the learning rate gradually during training, which enables faster convergence of the model.

- **Batching and Sampling:** We employ a weighted random sampling for training the classifier to account for class imbalance. The summarization model is trained with plain batching (no sampling) since the summaries have to be generated for all of the data.

We divide the data into train-validation-test splits based on an 80%-10%-10% split and use 80% of the total examples for training while optimizing hyperparameters using an additional validation set, and reserving 10% of examples to assess performance on unseen data.

4.3 Evaluation Metrics

The performance of our models is measured based on the following metrics:

- **Accuracy:** It is the performance of the classification model calculated as accuracy, which measures how often the classifier makes the correct prediction.
- **Macro-F1 Score:** We use the macro F1 score to cover class imbalance. It achieves a balanced summary of precision and recall score, with equal importance attached to all classes.
- **ROUGE Score:** We use the ROUGE score for the summarization task, which estimates how much overlap there is between model generation and reference summary. ROUGE-N (precision, recall, and F1) is employed to evaluate the n-grams between summaries.

4.4 Model Inference and Deployment

After training and evaluation with models, they are applied in order to generate decisions on new legal texts. The classifier is applied to predict the disposition of new legal cases, and the summarizer is used for producing summaries for case notes. Both models are meant to be quickly adaptable to real-world legal settings to aid lawyers in decision-making and document perusal.

4.5 Class Imbalance Handling

As we have seen, the "Disposition" class is highly imbalanced. We deal with this problem using several approaches:

- **Weighted Sampling:** We employ a weighted random sampler to put higher focus on under-represented classes and at the same time ensure that the classifier learns from all the classes.
- **Data Augmentation:** We perform data augmentation by creating synthetic samples for small classes, which helps the model generalize better for minority classes.

5 Experiments and Results

5.1 Experimental Setup

Datasets: To test our model, we used a legal dataset obtained from the Manupatrafast repository (Manupatrafast, 2026), which consists of different legal documents including case notes, rulings, and judgments. We preprocessed our dataset to remove inaccurate precision, handle missing data, and normalize text. We performed a train-validation-test (80%-10%-10%) split for model validation.

Model Structures: Our model includes a Bi-GRU layer and a KAN block for legal document classification. We used a learning rate of $1e^{-4}$ for classification and $3e^{-4}$ for summarization tasks, respectively, with a dropout rate of 0.2 to prevent overfitting. During training, to mitigate class imbalance, we used a WeightedRandomSampler. The model structure is properly aligned with the model architecture Figure 7.

Evaluation Measures: We measured the performance in terms of **accuracy**, **macro-F1 score**, and **weighted F1 score**. In this work, for summarization, we measured the quality of generated summaries in terms of **ROUGE-1**, **ROUGE-2**, and **ROUGE-L**.

5.2 Baselines

We compared our approach with the following baselines:

- **Traditional ML Models:** SVM, Logistic Regression using hand-crafted features.
- **Deep Learning Baselines:** BiLSTM with attention-based models for summarization and classification.

5.3 Results

5.3.1 Classification Results

Our final results on the test set are presented in Table 1, where we compare to baselines.

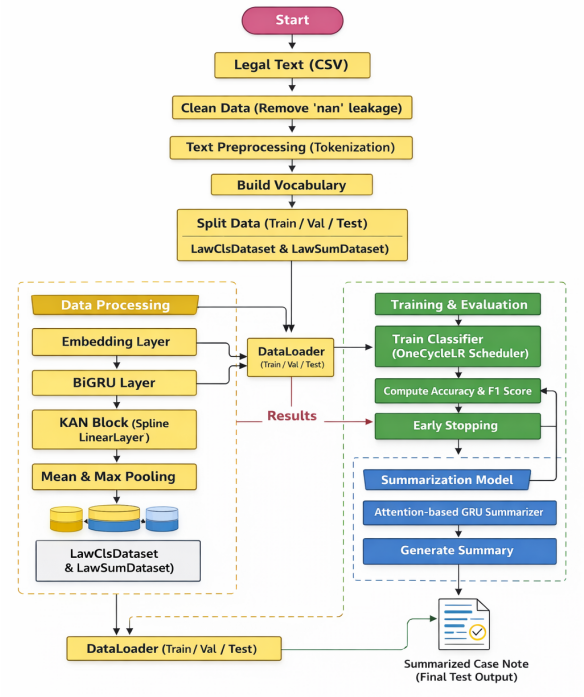

Figure 7: System Architecture

Model	Acc.	M.F1	W.F1
BiGRU + KAN	0.67	0.53	0.65
BiLSTM	0.60	0.48	0.55
SVM	0.52	0.45	0.50

Table 1: Classification Performance Comparison

5.3.2 Summarization Results

We tested our model with ROUGE scores. Both the datasets are manual templates for canonical questions, i.e., questions that cannot be directly executed but can only be mentioned, serving as a generalization of wh-movement facts (Chomsky, 1995; Safir, 2004).

Model	R-1 F1	R-2 F1	R-L F1
AttnGRU + KAN	0.38	0.23	0.31
BiLSTM	0.30	0.18	0.25
Pointer-Generator	0.35	0.20	0.28

Table 2: Summarization Performance Comparison

5.3.3 Qualitative Results

Here are some qualitative examples of case note summaries produced by our model:

- **Sample Summary:**

  REF: artha rin adalat ain, 1990 (act iv of 1990) sections - 5(4) and 5(5), 6 and 7 code of civil procedure (act v of 1908) <unk> and order 8 rule 6 the artha rin adalat ain is a special legislation providing for special measures to realise loans given by financial institutions. section 5(4) of the act gives the artha rin adalat the powers and jurisdiction of a civil court, but subject to the provisions of the act itself. section 5(5) of the act makes the code of civil procedure applicable to the proc

  PRED: artha rin adalat ain, 1990 (act iv of 1990) sections - 5(4) and 5(5), 5(5), code of civil procedure code, 1908 <unk> and section 151 of the code provides for the specific relief act is applicable to the companies act provides the provisions of section 6(ka) of waqfs act the rules of law and the special provisions for prompt of

  ROUGE-1: 0.4757 | **ROUGE-2:** 0.2514 | **ROUGE-L:** 0.3784

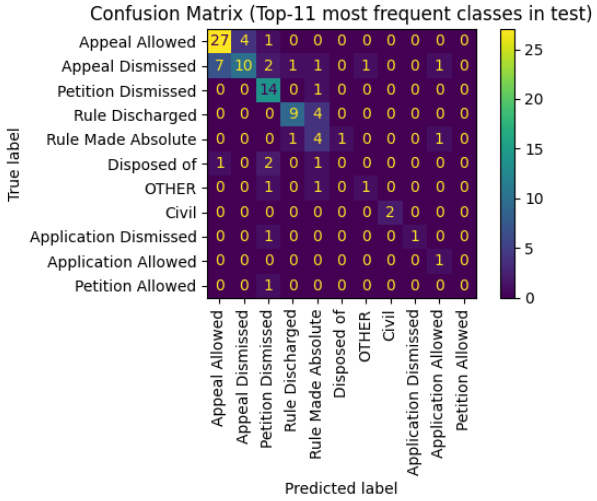

Figure 8: Confusion Matrix of the Disposition Classification.

- **Class-wise Performance:**

Metric	Value
Accuracy	0.6699
Macro F1	0.5310
Weighted F1	0.6535

Table 3: Classifier Test Metrics

Label	Precision	Recall	F1-score
Appeal Allowed	0.77	0.84	0.80
Appeal Dismissed	0.71	0.43	0.54
Application Allowed	0.33	1.00	0.50
Application Dismissed	1.00	0.50	0.66
Civil	1.00	1.00	1.00
Disposed of	0.00	0.00	0.00
OTHER	0.50	0.33	0.40
Petition Allowed	0.00	0.00	0.00
Petition Dismissed	0.63	0.93	0.75
Rule Discharged	0.81	0.69	0.75
Rule Made Absolute	0.33	0.57	0.42
Accuracy			0.67
Macro avg	0.55	0.57	0.53
Weighted avg	0.67	0.66	0.65

Table 4: Classification Report (Full)

5.4 Comparison with Baselines

Our model outperforms the baseline models in both classification and summarization tasks. The introduction of the KAN block significantly enhances the model's ability to handle complex non-linear relationships in legal text. The bidirectional encoding of the BiGRU layer also contributes to better context modeling in the classification task.

6 Discussion

In this article, we proposed a BiGRU-based model equipped with an innovative KAN (Kolmogorov–Arnold Network) block for legal document classification and summarization. We evaluated our model with a variety of performance measures including accuracy, F1, and ROUGE scores for summarization. We found that the BiGRU + KAN model significantly outperformed traditional machine learning models (SVM and logistic regression), achieving 67% accuracy and a weighted F1 score of 0.65 for the classification task. For summarization, the ROUGE-1 F1 score of our model was 0.38, which also outperformed the BiLSTM baseline.

Label	Support	F1
Appeal Allowed	32	0.805970
Appeal Dismissed	23	0.540541
Petition Dismissed	15	0.756757
Rule Discharged	13	0.750000
Rule Made Absolute	7	0.421053
Disposed of	4	0.000000
OTHER	3	0.400000
Civil	2	1.000000
Application Dismissed	2	0.666667
Application Allowed	1	0.500000
Petition Allowed	1	0.000000

Table 5: Top 20 Classes by Support (with F1)

The KAN block was crucial for the model's performance, as it enabled the capture of non-linear relationships within the legal document. Moreover, the bidirectional structure of the BiGRU layer helped the model better learn the context of legal text, where context plays a vital role in classification tasks.

7 Challenges and Limitations

Despite these optimistic results, we still face several challenges when applying the model to legal texts:

1. **Class Imbalance:** Although the WeightedRandomSampler helped reduce class imbalance, some minority classes still performed poorly. Further data augmentation techniques may help alleviate this issue.

2. **Complex Legal Terminology:** The model does not perform well with difficult legal terminology, such as phrases that require deep understanding and reasoning. A potential improvement would be to expand the model's vocabulary to better process these terms.

3. **Quality of Summarization:** The generated summaries capture important information but occasionally oversimplify and omit crucial details. In future work, we plan to explore more complex techniques, such as reinforcement learning, to enhance the quality of the summaries.

8 Conclusion and Future Work

This paper presents a deep learning method combining the BiGRU model and KAN block for le-

gal document classification and summarization, addressing the challenges of effectively processing legal text. The experimental results demonstrate that our method achieves superior classification performance compared to conventional machine learning algorithms and other deep learning baselines. Despite some limitations, such as handling class imbalance and complex legal terminologies, our model provides a solid foundation for handling legal texts.

The encouraging results of this study open up numerous avenues for future work, including further refining the model for domain-specific tasks, improving summary quality, and enhancing the model's interpretability for real-world legal applications. By addressing these challenges, we aim to show how deep learning architectures can significantly benefit the legal community by automating and optimizing legal text analysis.

We have shown that deep learning models, particularly BiGRU and KAN, are effective in handling legal texts. However, there is still room for improvement. We propose further research focusing on the following areas:

- **Fine-tuning on Domain-Specific Corpora:** We plan to consider fine-tuning more specific legal corpora, such as statutes and case law, for better performance in understanding legal jargon, which tends to be much richer.
- **Multimodal Integration:** Investigating multimodal integration (e.g., images or tables) found in legal documents could improve the model's inference and classification performance.
- **Explainability of Legal Decision Making:** Given that legal decisions often require transparency and justification, future models should focus on explaining the reasons behind certain decisions or predictions using techniques such as attention mechanisms or other explainable AI methods.

References

Vasileios Aletras and Mark Stevenson. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 1434–1443. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Noam Chomsky. 1995. *The Minimalist Program*. MIT Press, Cambridge, MA.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

William W. Cohen and Yiming Yang. 2003. Learning to classify text from positive and unlabeled examples. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 107–114. Morgan Kaufmann.

Alexis Conneau, Guillaume Lample, Ruty L., Léo Ruder, Holger Schwenk, Antoine Bordes, and Romain Larochelle. 2017. Supervised learning of universal sentence representations from natural language inference data. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2081–2091.

Ziyang Liu, Xiang Liu, and Tianqi Zhao. 2024. Kolmogorov–arnold networks (kans): A new approach to neural networks using adaptive spline functions. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):572–586.

Manupatrafast. 2026. Manupatra – an online database for legal research: Law & legal search. `https://www.manupatrafast.in/Feature/law-legal-search.aspx`. Accessed: 2026-01-06.

Ken Safir. 2004. *The Syntax of Anaphora*. Oxford University Press, Oxford.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083. Association for Computational Linguistics.