Empowering Prior to Court Legal Analysis: A Transparent and Accessible Dataset for Defensive Statement Classification and Interpretation

Yannis Spyridis, Jean-Paul Younes
Kingston University London
London, United Kingdom
{yannis.spyridis, j.younes}@kingston.ac.uk

Haneen Deeb
University of Portsmouth
Portsmouth, United Kingdom
haneen.deeb@port.ac.uk

Vasileios Argyriou
Kingston University London
London, United Kingdom
vasileios.argyriou@kingston.ac.uk

***Abstract*—The classification of statements provided by individuals during police interviews is a complex and significant task within the domain of natural language processing (NLP) and legal informatics. The lack of extensive domain-specific datasets raises challenges to the advancement of NLP methods in the field. This paper aims to address some of the present challenges by introducing a novel dataset tailored for classification of statements made during police interviews, prior to court proceedings. Utilising the curated dataset for training and evaluation, we introduce a fine-tuned DistilBERT model that achieves state-of-the-art performance in distinguishing truthful from deceptive statements. To enhance interpretability, we employ explainable artificial intelligence (XAI) methods to offer explainability through saliency maps, that interpret the model's decision-making process. Lastly, we present an XAI interface that empowers both legal professionals and non-specialists to interact with and benefit from our system. Our model achieves an accuracy of 86%, and is shown to outperform a custom transformer architecture in a comparative study. This holistic approach advances the accessibility, transparency, and effectiveness of statement analysis, with promising implications for both legal practice and research.**

sectionIntroduction The legal system is built upon the principles of justice and the rule of law, and the settlement of legal disputes often depends on the strength of the presented arguments. Individuals facing accusations typically put forth statements and arguments to either defend themselves or mitigate their liability. The classification of statements made in this context involves their automated identification and categorisation into distinct classes. It plays a vital role in ensuring a fair process by enabling legal professionals to better understand and evaluate their case. This classification is essential for legal practitioners to navigate the complexities of criminal cases, enabling them to distinguish between various argument strategies, assess their validity, and ultimately facilitate the pursuit of justice.

Classifying statements made by accused individuals presents a series of distinctive challenges. Firstly, the language used in these statements may vary widely, from straightforward admissions or denials to subtle explanations, justifications, or even appeals to sympathy. The diversity in language and argumentation strategies necessitates a refined approach to classification, which is often complicated by the emotional and psychological factors at play in criminal cases [1]. Furthermore, the need to protect the rights and dignity of the accused imposes an ethical dimension on this classification task, requiring sensitivity and respect for privacy. The task fundamentally involves distinguishing between credible statements and false or misleading claims, and presents an inherent difficulty. In addition, the need for interpretability and transparency in legal decision-making [2], calls for methods that not only classify arguments but also provide insights into the reasoning behind the classification, which adds an additional layer of complexity to the task [3].

Various methodologies can be employed for the automated classification of statements made by accused individuals. These may include rule-based systems, supervised machine learning models, or sentiment analysis [4]. While human professionals also engage in this task, their involvement can introduce subjective biases and inconsistencies into the classification process [5]. The applications of this process extend to legal research, case assessment, and criminal justice system improvement. Additionally, the incorporation of legal domain knowledge and contextual information is often important to enhance the accuracy of classification. Accurate categorisation of arguments can aid legal professionals in evaluating the credibility and relevance of statements made by accused individuals, assisting in plea bargaining, and ensuring that due process is adhered to [6]. Furthermore, it can contribute to a more efficient and fair criminal justice system by streamlining the review of statements and enhancing the understanding of the defendant's perspectives. Inaccurate assessments of statements can have profound legal and societal implications, potentially leading to unjust legal outcomes, misallocated resources, and undermined public trust in the legal system. Therefore, the need for robust, interpretable, and accessible tools in this domain is significant [7].

This paper represents a novel effort in addressing the critical issue of classifying statements made during police interviews through a holistic approach that leverages state-of-the-art natural language processing (NLP) techniques and user-centric design principles. At its core, our primary objective is to significantly enhance the accuracy and transparency of statement

analysis, thereby equipping legal professionals, researchers, and stakeholders with a powerful tool for making more informed and equitable legal decisions. To achieve this goal, we introduce a novel dataset of transcripts from police interviews, which provides a valuable resource for the advancement of NLP research in the legal domain. Additionally, we develop a fine-tuned DistilBERT model tailored specifically for the classification of these arguments, ensuring the highest levels of accuracy in categorisation. Notably, our research goes beyond mere classification, as we also incorporate explainable artificial intelligence (XAI) techniques to shed light on the model's decision-making process, promoting transparency and trust in the system. Lastly, in an effort to make our research accessible and practical, we develop a user-friendly XAI interface, bridging the gap between sophisticated NLP technology and end-users in the legal field. Through this holistic approach, our research aims to contribute to the advancement of NLP and legal informatics, while fostering a more just and efficient legal system.

In summary, the key contributions of this paper are the following:

- It introduces a novel and carefully curated dataset of statements made during police interviews. This dataset serves as the foundation for the research and enables the development and evaluation of the classification model.
- It develops a fine-tuned DistilBERT model, refined to the task of statement classification. The model achieves high performance in distinguishing truthful from deceptive statements.
- It employs state-of-the-art explainability visualisation techniques to enhance the interpretability of the model's decisions. These visualisations provide insight into the rationale behind the model's predictions, allowing users to understand and trust the model's outputs.
- It presents an advanced user-friendly interface to cover the need for XAI in both legal practice and research. This interface enables users of any background to interact with the classification system, making statement analysis more accessible and intuitive.

The remainder of this paper is structured as follows: Section I provides NLP background and reviews relevant research conducted in the legal domain on statement classification. Section II presents the methodology that was followed. including background on the dataset and the explainability methods. Section III discusses the results of this work and presents the developed XAI interface. Section IV provides the conclusion of the paper.

I. Related Work

Like in many areas, NLP has emerged as a strong tool within the legal domain, revolutionising the way professionals and researchers interact with text-based statements and legal documents [8]. NLP methods enable the automated analysis of legal texts, including contracts, court cases, or statements prior to court proceedings. These techniques facilitate tasks such as information extraction, fake news detection [9], argument classification, and legal document summarisation [10]. Therefore, NLP can significantly accelerate the research process and enhance the efficiency of legal practitioners by providing tools for information retrieval, knowledge discovery, and evidence assessment [6]. Furthermore, NLP-powered applications are instrumental in ensuring legal compliance, contract management, and the automation of routine legal tasks, reshaping the landscape of the legal field in novel ways.

A. Transformer architectures

The transformer model [11] is a deep learning architecture that relies on self-attention mechanisms to capture contextual relationships between words in a text. This architecture has achieved state-of-the-art performance on a wide range of NLP tasks, including text classification in the legal domain. One of the key advantages of the transformer architecture is its ability to capture long-range dependencies in text, which is particularly important in legal arguments. The transformer's self-attention mechanism allows it to attend to different parts of the input text and weigh the importance of each word in the context of the entire text [12]. This enables the model to effectively capture the semantic meaning and context of statements, leading to improved classification accuracy [13].

In the context of binary classification, transformers can be used to classify input data into one of two classes. The input data is typically represented as a sequence of tokens, such as words or characters, and each token is embedded into a continuous vector representation. The transformer model then processes these embeddings through a series of self-attention layers and feed-forward layers, allowing it to capture the relevant information for the classification task [14], [15]. During training, the transformer model learns to optimise its parameters by minimising a loss function, such as binary cross-entropy, which measures the discrepancy between the predicted probabilities and the true labels. The model updates its parameters using backpropagation and gradient descent, iteratively improving its performance on the training data [16]. The key advantage of transformers in this context is their ability to handle variable-length input sequences. The self-attention mechanism allows the model to attend to different parts of the input sequence, regardless of their position, enabling it to capture long-range dependencies and contextual information [17].

B. Classification in the legal domain

Several studies have investigated AI-assisted classification in the legal domain across various subjects. The study in [18] addresses challenges in legal multi-label document classification by introducing a new dataset of legal opinions with manually labeled legal procedural postures. A domain-specific pre-trained deep learning architecture is presented, with a label-attention mechanism that showcases efficacy in overcoming data scarcity and class imbalance issues.

Court case transcripts are analysed in [19] based on discourse and argumentative properties. The study investigates the utilisation of discourse relationships and sentence properties to

extract relevant information from the transcripts, addressing the significance of such data for the legal domain. It proposes a classification framework that employs a combination of machine learning models and rule-based approaches to categorise sentence pairs based on their observed relationship types and distinguish them as contributing to legal arguments or not.

The task of legal judgment prediction, which involves automatically predicting the outcome of a court case based on the text describing the case's facts, is investigated in [20]. The study introduces a new legal judgment prediction dataset, comprising cases from the European Court of Human Rights. An evaluation of several neural models is conducted on this dataset, leading to the establishment of robust performance benchmarks that outperform previous feature-based models across binary and multi-label classification and case importance prediction.

C. Model interpretability in NLP

Despite the success of state-of-the-art language models, interpretability remains a significant challenge within this area of research. Language models are often considered "black boxes" because it is arduous to discern the internal mechanisms and feature representations that lead to specific predictions. Some approaches focus on local explanations, which aim to identify the most important features or words that contribute to a model's prediction. For instance, input saliency methods have been used to explain predictions of deep learning models in NLP [21]. These methods highlight the words or features that have the most influence on the model's output.

Moreover, various studies have delved into the influence of model architecture and design decisions on interpretability. For instance, approaches that focus explicit word interaction graph layers have been proposed to enhance interpretability by capturing intricate word interactions [22]. Furthermore, studies have explored the incorporation of likelihood considerations in NLP classification explanations, aiming to offer more insightful and easily understandable explanations [23].

II. Methodology

A. Produced Dataset

In our study, the employed dataset was curated from a series of transcripts of statements provided to police prior to court proceedings. The interviews were conducted in the context of analysing verbal veracity cues. Participants were asked to report on a truthful or false event in experiments that took place within the same laboratory located in the United Kingdom and were granted ethical approval by the institutional ethics committee.

TABLE I. Dataset preview

ID	Text	GT
1	Yes sure. It was on a Sunday... Sunday evenin...	1
2	Well, on that day I felt like death, so I did ...	1
3	So yeah, like a couple of months back , thirt...	0
4	OK. So basically it was months ago, this alre...	0
5	OK, I sang in Lichfield Cathedral for a weeken...	1

The initial dataset underwent a data cleansing process, involving the elimination of irrelevant elements such as fillers (e.g., 'uhm,' 'err'), indicators of participants' non-verbal behaviors (e.g., pauses, smiles), and any contributions from the interviewer. The dataset comprises a total of 687 statements, each accompanied by its corresponding ground truth. The calculated Gini Index of 0.49 indicates a notably balanced distribution of the binary labels. A preview of the data is depicted in Table I. The last column lists the ground truth, whereby 1 indicates a deceptive statement, and 0 corresponds to the truth.

Comprising a diverse array of statements made by individuals during these preliminary investigative interactions, the dataset aims to capture the varied expressions and complexities inherent in the language used in such interviews. The content encompasses a series of expressions, ranging from straightforward admissions or denials to more subtle explanations, justifications, and appeals. The dataset has been carefully curated to reflect the varied linguistic and argumentative strategies employed in the context of police interviews, mirroring the real-world challenges faced by law enforcement professionals in discerning the credibility of statements.

B. Custom transformer

To address the task of classifying statements in police interviews, we initially explored the development of a custom transformer model tailored to meet the specific demands of the statement classification domain. The architecture of the model is illustrated in Table II and incorporates the following elements:

- Token and Position Embedding Layer: This layer combines token embeddings and positional embeddings to provide a comprehensive representation of the input sequences.
- Transformer Block: The transformer block is employed to capture contextual dependencies within the input data. This block encompasses multi-head self-attention and feed-forward layers, enabling the model to discern the relevance of different tokens and understand complex relationships.
- Global Average Pooling: The global average pooling layer is applied to summarise the model's output across the entire sequence, consolidating essential information for downstream processing.
- Dropout: Dropout layers are introduced to enhance model robustness and mitigate overfitting.

The custom transformer model aimed to provide insight to the context of statement classification and to the expected accuracy performance in this task, given the curated dataset. A preliminary evaluation suggested domain-specific advantages, indicating the model's ability to capture subtle language usage

TABLE II. CUSTOM TRANSFORMER ARCHITECTURE

Layer Type	Output Shape	Param #
Input	(None, 200)	0
Token and Position Embedding	(None, 200, 32)	223,616
Transformer Block	(None, 200, 32)	10,656
Global Average Pooling1D	(None, 32)	0
Dropout 2	(None, 32)	0
Dense 2	(None, 16)	528
Dropout 3	(None, 16)	0
Dense 3	(None, 1)	17

pertaining to the data. Nevertheless, the results of this evaluation motivated the decision to explore the possibility of fine-tuning state-of-the-art NLP models, capable of capturing contextual information more effectively and providing enhanced performance and generalisation capabilities. The DistilBERT model was selected for this purpose.

C. DistilBERT Model

As mentioned above, the transition to DistilBERT was driven by the anticipation of higher performance in light of the significance of accurately classifying statements in the legal context. DistilBERT is a variant of the Bidirectional Encoder Representations from Transformers (BERT) model [12]. It is specifically designed to provide a more computationally efficient but highly effective solution for a wide range of NLP tasks. Its architectural design adheres to the transformer architecture, known for its ability to capture contextual relationships within textual data. DistilBERT inherits the bidirectional context-awareness of BERT while reducing its complexity by distillation techniques, resulting in a more lightweight model, with an approximately 40% reduced size.

1) Applicability: DistilBERT was selected due to its ability to maintain a competitive performance in NLP tasks, while significantly reducing computational and memory requirements. This efficiency was particularly useful in the case of the curated statement dataset, and facilitated a more accessible model training and deployment process. In addition to its efficiency, the model is pre-trained on a significant amount of text, having a solid foundation on linguistic context and understanding. Finally, the structure of DistilBERT allows straightforward fine-tuning on domain-specific datasets, allowing the model to specialise in this specific task while retaining the benefits of pre-trained knowledge.

2) Fine-tuning: The fine-tuning process was executed with careful consideration of the model configuration and hyperparameter settings. The aim was to adapt the pre-trained DistilBERT model to the specific requirements of statement classification while optimising its performance. The Hugging Face Transformers [24] library was employed to facilitate the implementation and fine-tuning of the model. The library offers a valuable collection of pre-trained transformer models, fine-tuning pipelines, and utilities for NLP tasks.

TABLE III. MODEL CONFIGURATION

Parameter	Value
Base Model	distilbert-base-uncased
Activation Function	GELU
Attention Dropout	0.1
Hidden Dimension (dim)	768
Global Dropout Rate	0.1
Fine-Tuning Task	”sst-2” (Binary Classification)
Hidden Layer Dimension	3072
Attention Heads	12
Transformer Layers	6
Dropout	0.2

The configuration of the DistilBERT model is outlined in Table III. It employs the Gaussian error linear units (GELU) activation function, commonly used in Transformers to introduce non-linearity and improve the model's capacity to learn complex patterns [25]. The GELU function can be approximated as follows:

$$\mathrm{GELU}(x) = 0.5x(1 + \tanh[\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)]) \quad (1)$$

The GELU function exhibits linear behavior for positive inputs, similar to ReLU activation, but it introduces a smooth, differentiable transition for negative inputs. This transition is achieved by applying a Gaussian distribution to the negative inputs, which helps to prevent the vanishing gradient problem that can occur with ReLU when the input is negative.

The model was optimised using the binary cross-entropy loss function, which effectively quantifies the divergence between predicted and actual binary labels as follows:

$$\mathrm{L}(y, p) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \quad (2)$$

TABLE IV. TRAINING HYPERPARAMETERS

Hyperparameter	Value
Loss Function	binary cross-entropy
Optimiser	AdamW
Learning Rate	0.0002
Per Device Batch Size	4
Weight Decay	0.01
Gradient Accumulation Steps	2
Epochs	5

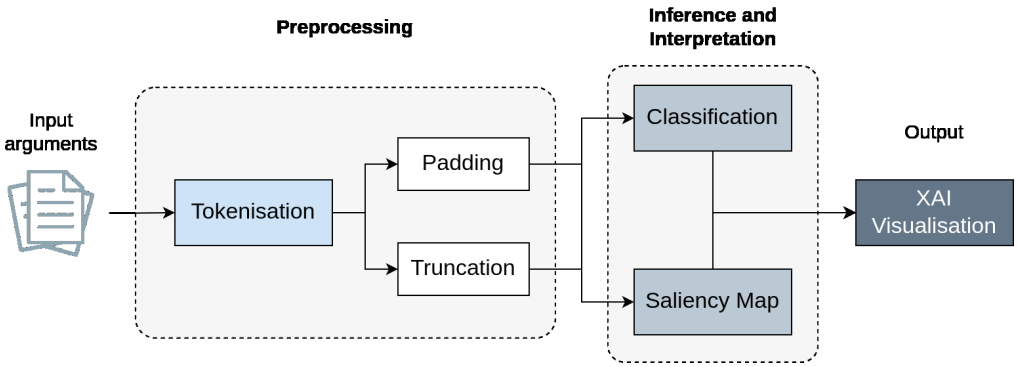

Fig. 1. The end-to-end system pipeline for defensive statement classification and explainability

where N represents the number of samples, y_i is the true label for the i_{th} sample, and p_i is the predicted probability that the i_{th} sample belongs to the positive class.

To promote regularisation during training, the AdamW optimiser was adopted, utilising weight decay to penalise excessive model weights and avoid overfitting. The low selection of per device batch size was made in light of optimising the training process in alignment with the available computational resources, while ensuring the model's efficient convergence and enhancing the overall training performance. The training hyperparameters are outlined in Table IV.

The dataset used for training and evaluating the DistilBERT model was divided into training (70%) and testing (30%) sets, to assess the model's generalisation performance on unseen data. The training set was used to train the model during the fine-tuning process, while the testing set served as an independent evaluation to determine the model's performance on previously unseen samples and ensure robust insights of the model's effectiveness in real-world scenarios.

D. Explainability through saliency maps

A crucial aspect of XAI and model interpretability lies in the ability to discern the significance of individual features within the input data. Saliency maps can serve this purpose by quantifying the influence each token in the input sequence has on the model prediction. The central concept in generating a saliency map is determining the sensitivity of the prediction to small perturbations in the input token values. This sensitivity is examined through gradient computations:

$$\text{Saliency}(\mathbf{t}) = \nabla_{\mathbf{t}} P(y|\mathbf{x}), \tag{3}$$

where t represents a token within the input text $\mathbf{x}$, $P(y|\mathbf{x})$ is the probability distribution over model predictions for the classification task, and $\nabla_{\mathbf{t}}$ denotes the gradient operator with respect to token t.

The gradients are captured in a two-step process utilising registered hooks on the model's embeddings. To make the saliency scores more interpretable, we apply L1 normalisation to the computed gradient. This normalisation ensures that the saliency scores sum to 1, enabling direct comparison of the relative importance of each token in the input. The normalised saliency gradient, is calculated as:

$$\text{Saliency}(\mathbf{t}) = \frac{\nabla_{\mathbf{t}} P(y|\mathbf{x})}{\|\nabla_{\mathbf{t}} P(y|\mathbf{x})\|_1} \tag{4}$$

where $\|\cdot\|_1$ denotes the L1 norm.

The saliency scores are then mapped to words in the tokenised input text, to generate a saliency map that indicates the most influential w ords i n t he i nput. T his mapping provides insights to the model's behaviour and is used in a custom function that highlights the respective words in the text for interpretability. The complete pipeline of our system is illustrated in Fig. 1.

III. Evaluation

A. Setup and metrics

The fine-tuned DistilBERT model was trained and evaluated on an NVIDIA T4 Tensor Core GPU. To comprehensively assess and provide a holistic understanding of the model's capabilities in the context of classifying statements made in police interviews, a range of evaluation metrics were employed, including Accuracy, Precision, Recall, and F1 Score:

1) Precision:
$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$
2) Recall:
$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$
3) Accuracy:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$
where TP, TN, FP, and FN are the true positive, true negative, false positive, and false negative samples respectively.
4) F1 Score:
$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

To gain further insights to the DistilBERT performance, two additional metrics were used:

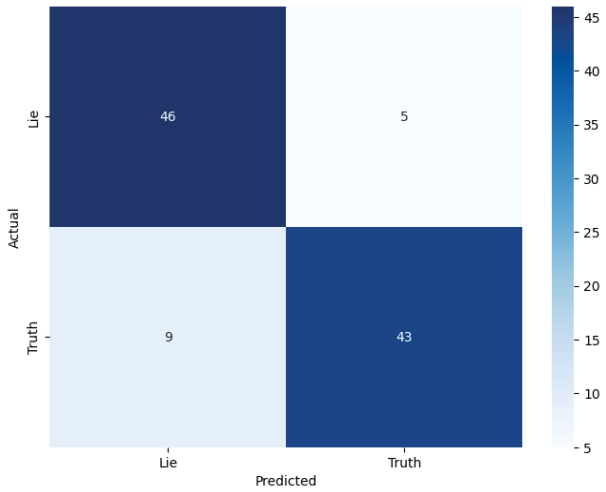

Fig. 2. Confusion matrix of the fine-tuned model

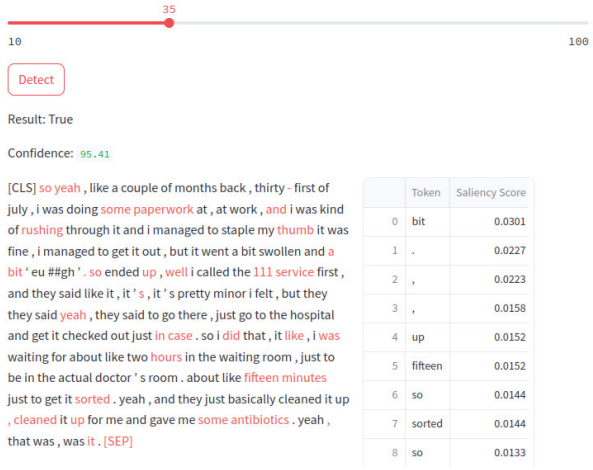

Fig. 3. The XAI interface for defensive statement classification and inter-pretability

1) Receiver Operating Characteristic Area Under the Curve (ROC AUC):

$$\text{ROC - AUC} = \int_0^1 TPR(FPR^{-1}(t))\,dt \qquad (9)$$

where TPR is the true positive rate and FPR is the false positive rate. The ROC - AUC score is used to quantify the model's ability to distinguish between positive and negative instances.

2) Average Precision:

$$\text{Average Precision} = \sum_{i=1}^{n}(R_n - R_{n-1})P_n \qquad (10)$$

where R is the Recall and P is Precision. Average Precision is used to offer insights into the model's precision-recall trade-offs.

B. Results

The results of the two models across the main metrics are presented in Table V. The DistilBERT model achieved a score of 86% in Accuracy, 90% in Precision, 82% in Recall, and 86% in F1 Score, suggesting the model's robustness and effectiveness in distinguishing between truthful and deceptive claims. The custom transformer's scores of approximately 80%, while respectable, imply that it may struggle with making accurate predictions in several cases. Figure 2 presents the confusion matrix on the fine-tuned model.

TABLE V.
CLASSIFICATION RESULTS

Model	Accuracy	Precision	Recall	F1 Score
Fine-tuned distilbert	0.8666	0.9006	0.8265	0.8664
Custom transformer	0.8005	0.7936	0.8126	0.8074

The DistilBERT model provided high performance as also evidenced by the ROC-AUC score of 0.8671 and the area under the Average Precision of 0.8320. The ROC-AUC score indicates that the model effectively discriminates between positive and negative instances over a range of threshold values in the dataset, while the Average Precision score highlights its ability to correctly classify positive instances while minimising false positives. These results reflect the model's proficiency in capturing the complexity of statements and its potential to make informed and precise predictions in the legal domain.

The results also depict the DistilBERT model's harmonious balance between precision and recall. This is particularly valuable for the legal domain, where both false positives and false negatives can have significant consequences. Ultimately, the results highlight the benefits of leveraging pre-trained transformer models and fine-tuning them for domain-specific tasks, as opposed to utilising custom architectures, thus demonstrating the potential of pre-trained state-of-the-art models in legal defensive statement analysis.

C. Explainable artificial intelligence integration

The XAI interface was developed with the primary objective of bridging the gap between theoretical research and its practical application, with a particular emphasis on democratising AI, especially in the legal domain. It provides an accessible and user-friendly platform designed for a diverse audience, including legal practitioners and researchers. Users can submit statements through the interface, and select the number of tokens for interpretability. The deployed model then provides real-time predictions, accompanied by the presentation of the argument with visual emphasis on influential w ords, as indicated by the saliency map. The XAI interface is presented in Figure 3. We also provide an example of visualising the attention in each layer in Fig. 4.

IV. CONCLUSION

In the evolving landscape of the legal domain, the integration of NLP methods has become increasingly impor-

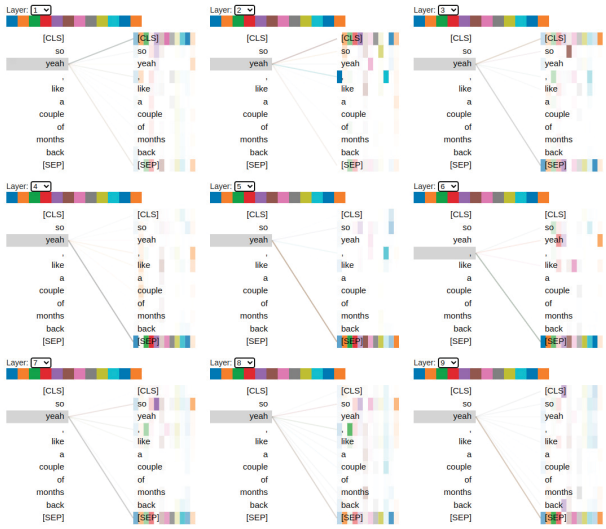

Fig. 4. Visualising the attention in each layer

tant, offering innovative solutions to the challenges that legal professionals and researchers encounter. This paper makes significant contributions to the field of argument analysis in statements provided by individuals during interrogations, by introducing an NLP dataset of transcripts from police interviews, allowing for the development and evaluation of our fine-tuned DistilBERT model. The model's high performance in distinguishing truthful from deceptive statements, as well as its capacity to achieve a balance between precision and recall, underscores its potential in making informed and precise predictions in the legal domain.

Furthermore, the utilisation of state-of-the-art explainability visualisation techniques enhances the model's interpretability, fostering trust and transparency in its decision-making process. The accompanying user-friendly XAI interface further democratises access to our developed tool, bridging the gap between legal practice and research, and making statement analysis more accessible for users of diverse backgrounds.

The experimental results collectively underscore the utility of leveraging pre-trained transformer models and fine-tuning them for domain-specific tasks, as opposed to utilising custom architectures. The success of this approach reinforces the potential of pre-trained state-of-the-art models in the realm of statement analysis in the context of police interrogations, and offers a holistic solution that advances the field, serving as a valuable resource for legal professionals and researchers.

REFERENCES

[1] J. M. Salerno and B. L. Bottoms, "Emotional evidence and jurors' judgments: The promise of neuroscience for informing psychology and law," *Behavioral sciences & the law*, vol. 27, no. 2, pp. 273–296, 2009.

[2] A. Deeks, "The judicial demand for explainable artificial intelligence," *Columbia Law Review*, vol. 119, no. 7, pp. 1829–1850, 2019.

[3] G. Vilone and L. Longo, "Explainable artificial intelligence: a systematic review," *arXiv preprint arXiv:2006.00093*, 2020.

[4] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 45, no. 4, pp. 765–818, 2020.

[5] W. C. Thompson and E. L. Schumann, "Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy," in *Expert Evidence and Scientific Proof in Criminal Trials*. Routledge, 2017, pp. 371–391.

[6] R. M. Re and A. Solow-Niederman, "Developing artificially intelligent justice," *Stan. Tech. L. Rev.*, vol. 22, p. 242, 2019.

[7] J. Cui, X. Shen, and S. Wen, "A survey on legal judgment prediction: Datasets, metrics, models and challenges," *IEEE Access*, 2023.

[8] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does nlp benefit legal system: A summary of legal artificial intelligence," *arXiv preprint arXiv:2004.12158*, 2020.

[9] E. Shushkevich and J. Cardiff, "Detecting fake news about covid-19 on small datasets with machine learning algorithms," in *2021 30th Conference of Open Innovations Association FRUCT*. IEEE, 2021, pp. 253–258.

[10] K. Merchant and Y. Pande, "Nlp based latent semantic analysis for legal text summarization," in *2018 international conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2018, pp. 1803–1807.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[14] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *arXiv preprint arXiv:2005.00928*, 2020.

[15] Y. Tay, D. Bahri, D. Metzler, D.-C. Juan, Z. Zhao, and C. Zheng, "Synthesizer: Rethinking self-attention for transformer models," in *International conference on machine learning*. PMLR, 2021, pp. 10 183–10 192.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[17] Y. Tay, M. Dehghani, S. Abnar, Y. Shen, D. Bahri, P. Pham, J. Rao, L. Yang, S. Ruder, and D. Metzler, "Long range arena: A benchmark for efficient transformers," *arXiv preprint arXiv:2011.04006*, 2020.

[18] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, p. 101718, 2022.

[19] G. Ratnayaka, T. Rupasinghe, N. de Silva, M. Warushavithana, V. S. Gamage, M. Perera, and A. S. Perera, "Classifying sentences in court case transcripts using discourse and argumentative properties," *The International Journal on Advances in ICT for Emerging Regions*, vol. 12, no. 1, 2019.

[20] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in english," *arXiv preprint arXiv:1906.02059*, 2019.

[21] S. Rönnqvist, A.-J. Kyröläinen, A. Myntti, F. Ginter, and V. Laippala, "Explaining classes through stable word attributions," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1063–1074.

[22] A. Sekhon, H. Chen, A. Shrivastava, Z. Wang, Y. Ji, and Y. Qi, "Improving interpretability via explicit word interaction graph layer," *arXiv preprint arXiv:2302.02016*, 2023.

[23] D. Harbecke and C. Alt, "Considering likelihood in nlp classification explanations with occlusion and language modeling," *arXiv preprint arXiv:2004.09890*, 2020.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[25] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023.