IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 March 2023, accepted 14 May 2023, date of publication 22 May 2023, date of current version 8 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279034

RESEARCH ARTICLE

Classifying European Court of Human Rights Cases Using Transformer-Based Techniques

ALI SHARIQ IMRAN[1], (Member, IEEE), HENRIK HODNEFJELD[1], ZENUN KASTRATI[2], NOUREEN FATIMA[3], SHER MUHAMMAD DAUDPOTA[3], AND MUDASIR AHMAD WANI[4]

[1]Department of Computer Science, Norwegian University of Science and Technology (NTNU), 2815 Gjøvik, Norway
[2]Department of Informatics, Linnaeus University, 351 95 Växjö, Sweden
[3]Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan
[4]EIAS Data Science Laboratory, College of Computer and Information Sciences, Prince Sultan University, Riyadh 11586, Saudi Arabia

Corresponding author: Ali Shariq Imran (ali.imran@ntnu.no)

This work was supported in part by the Department of Computer Science (IDI), Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Gjøvik, Norway; and in part by Direktoratet for Internasjonalisering og Kvalitetsutvikling i høyere utdanning (DIKU) through the Curricula Development and Capacity Building in Applied Computer Science for Pakistani Higher Education Institutions (CONNECT) Project under Grant NORPART-2021/10502.

ABSTRACT In the field of text classification, researchers have repeatedly shown the value of transformer-based models such as Bidirectional Encoder Representation from Transformers (BERT) and its variants. Nonetheless, these models are expensive in terms of memory and computational power but have not been utilized to classify long documents of several domains. In addition, transformer models are also often pre-trained on generalized languages, making them less effective in language-specific domains, such as legal documents. In the natural language processing (NLP) domain, there is a growing interest in creating newer models that can handle more complex input sequences and domain-specific languages. Keeping the power of NLP in mind, this study proposes a legal documentation classifier that classifies the legal document by using the sliding window approach to increase the maximum sequence length of the model. We used the ECHR (European Court of Human Rights) publicly available dataset which to a large extent is imbalanced. Therefore, to balance the dataset we have scrapped the case articles from the web and extracted the data. Then, we employed conventional machine learning techniques such as SVM, DT, NB, AdaBoost, and transformer-based neural networks models including BERT, Legal-BERT, RoBERTa, BigBird, ELECTRA, and XLNet for the classification task. The experimental findings show that RoBERTa outperformed all the mentioned BERT versions by obtaining precision, recall, and F1-score of 89.1%, 86.2%, and 86.7%, respectively. While from conventional machine learning techniques, AdaBoost outclasses SVM, DT, and NB by achieving scores of 81.9%, 81.5%, and 81.7% for precision, recall, and F1-score, respectively.

INDEX TERMS Legal documents classification, European court of human rights (ECHR) dataset, natural language processing, transformers, BERT, BigBird, ELECTRA, XLNet, legal-BERT.

I. INTRODUCTION

We live in a society increasingly governed by legal rules and regulations [1]. The *juridification* of society increases the importance of defending and/or upholding one's legal rights. High-quality legal representation is often expensive, so those who cannot afford it rely on public legal aid programs [2]. Studies of several countries' legal systems indicate that legal aid is not being provided to all those in need. An analysis of the legal system in Norway [3] shows that only approximately 9% of the population is qualified for legal aid. A study of the legal system in the U.K. [4] indicates that there are several issues with providing free legal aid. At the maximum level of disposable income at which legal aid is available, many households do not have sufficient income to meet a minimum standard of living before legal fees are paid. Typically, their

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

disposable incomes are 10 to 30 percent too low to meet a minimum budget [5].

Even those with the lowest incomes, the most vulnerable individuals, are excluded from legal aid if they have savings or assets worth more than £8,000, or in some cases, £3,000. If a person has this much money in the bank, they can pay for some legal expenses without affecting their current ability to maintain a minimum standard of living [4].

A means test is used to determine whether a person or household is eligible for a particular benefit or payment [6]. However, the means test also considers the value of people's homes. As a result, homeowners who are not employed may be excluded from legal aid, despite having no realistic option for paying the legal fees [4].

However, by improving the effectiveness of the provision of legal assistance, the cost of legal aid may be reduced, and increase the number of cases assisted within the publicly funded legal aid budgets. Through this research paper, we aim to look at a solution to improve efficiency, thereby lowering the legal system's costs for private plaintiffs and defendants, by automating the classification of legal cases.

We adopted several transformer-based models to classify large legal documents regarding human rights violations. With transformer-based architectures, it has become easier to build more capable models, and pretraining has made it possible to use this capability effectively for a wide variety of tasks. This article is based on the authors' previous thesis work and the readers are advised to read the thesis for more background information [7].

A. STUDY OBJECTIVES AND RESEARCH QUESTIONS

This research investigates the possibility of classifying large legal documents using transformers: a foundation model to determine whether a human rights article has been violated and, if so, which articles. Both binary class (two-class) and multi-class methods have been investigated to determine the optimal approach for categorizing documents of this type. More specifically, we intend to address the following research questions (RQs):

- RQ1: How influential are transformers in handling long sequences of text data from legal documents?
- RQ2: Which features can be exploited to train transformers and effectively classify legal documents?
- RQ3: How viable would transformers be in enhancing the efficiency of legal assistance?

B. CONTRIBUTION

The major contributions of this research are as given below:

- Proposed a transformer-based model for classifying large legal documents.
- Demonstrated how to increase the maximum sequence length for transformer models artificially. With our approach, we have employed a sliding window technique to allow for multiple sub-sequences to enable the models to evaluate text sequences longer than their usual limit.
- Handled highly imbalanced dataset by scrapping additional samples from the ECHR portal.
- Evaluated the performance of both transformer- and conventional machine learning-based models. for long sequence text classification.

The rest of this study is arranged as follows. Section II describes the existing works on classifying the European Court of Human Rights. Section III presents methods and implementation used in this study. Section IV presents the experimental settings and results. Section IV-D presents the theoretical analysis of our obtained results in the form of a discussion. Finally, Section V concludes the article and provides future directions.

II. RELATED WORK

In this section, we examine studies that provide different approaches to our text classification problems and work that has addressed similar issues.

A. TRANSFORMERS: A FOUNDATION MODEL

In the paper ''On the Opportunities and Risks of Foundation Models,'' written by a team of researchers from Stanford, they refer to Transformers and BERT, particularly as a ''foundation model.'' They name foundation models as these are trained on broad data at scale and adaptable to a wide range of downstream tasks, as well as to emphasize their critically central yet incomplete character [8].

The team shows how transformers are vastly powerful and central in deep learning yet have not reached their full potential. Many new variations of transformers have been created, contributing to the ever-growing transformer-based model pool.

B. NEURAL LEGAL JUDGMENT PREDICTION IN ENGLISH

Our research work is inspired by the study conducted in [9], which focuses on automatically predicting a court case's outcome, given the case's details. The authors applied neural networks to English legal judgment datasets in order to make predictions [9]. It appears that this started a trend within the natural language processing field in legal judgments as more papers have been published since then, such as [10] and [11].

Together they looked at a wide variety of different neural models on binary, multi-class classification, and case importance prediction [9]. Our work builds on their binary and multi-label classification outperforming benchmark results with numerous conventional machine learning and advanced deep learning models with performance insights, as they did not ''deep-dive'' into any specific models.

Chalkidis et al. [9] proposed a hierarchical version of BERT that bypasses the length limitation of 512-word pieces. This was one way to circumvent the problem BERT faces; however, in this research, we have looked deeper into ways

to solve BERTs length limitation and further improve upon Chalkidis et al. results [9].

C. GROWING NEED FOR NLP SOLUTIONS IN THE LEGAL SYSTEM

This section discusses different approaches for e-discovery and NLP solutions within the legal field. Firstly, we determine what parts are relevant to our work. To begin with, we examine how electronic discovery is currently conducted and how it came about.

Let us first make clear what ''classic'' discovery is. In a legal case, discovery is the process by which one party (the producing party) makes available to the other (the requesting party) any relevant materials that are contained in their possession [12]. This discovery phase was traditionally practiced with pen and paper documents in most parts of the world and still is in many developing countries. At the same time, it is the primary electronic discovery used nowadays.

In 2010, Conrad stated the need for artificial intelligence as information retrieval in e-discovery and as a whole [13]. The industry of e-discovery has rapidly been growing since 2005. E-discovery market revenues are growing from over 1.8$ billion in 2015 to over 3.7$ billion by 2019-an average annual growth rate of 19% [14]. Just like the growth rate, the industry has seen different techniques and companies arise.

D. DOCUMENT BERT

The study in [15] investigated what they called a ''straightforward'' classification model using BERT to achieve state-of-the-art results on four popular datasets. There are a few characteristics of document classification that might lead one to believe BERT is not the most appropriate model. For instance, syntactic structures matter less for content categories, and documents are typically much longer than BERT input. Consequently, the team optimized the BERT model by adding an additional soft-max classifier parameter, as shown in equation 1. For both single-class and multi-class tasks, they minimized the cross-entropy and binary cross-entropy loss.

$$W \in \mathbb{R}^{K*H} \tag{1}$$

The equation 1 depicts the parameters of the softmax classifier for fine-tuning BERT for document classification. Here, H represents the dimension of the hidden state vectors, and K represents the number of classes.

To alleviate the computational burden associated with BERT, the team applied knowledge distillation [16] to transfer the knowledge from BERT to the smaller state-of-the-art Bidirectional Long Short Term Memory (BiLSTM) model.

E. DOMAIN SPECIFIC NATURAL LANGUAGE PROCESSING

The pretraining of large neural language models, such as BERT, has yielded impressive results in NLP [17]. Nevertheless, most pretraining efforts focus on general domain corpora, such as Wikipedia. There is a general assumption that even domain-specific pretraining can benefit from starting with general domain language models.

Y Gu. et al. in [18] challenges the assumption that pretraining general-domain language models continuously leads to better performance in certain domains, such as biomedicine. Instead, pretraining language models from scratch on unlabeled text in specific domains results in significant performance improvements across a range of biomedical natural language processing tasks. The researchers also explored using named entity recognition to extract structured summary-level data from unstructured scientific text to address the problem of summarizing large amounts of published literature.

Named entity recognition (NER) is the process of identifying and categorizing key information (entities) in text. Any word or set of words that refer consistently to the same thing can be considered an entity. Detected entities are classified into predetermined categories. A NER model may, for instance, detect the word 'Bus' in a text and identify it as a 'Car'.

On three materials datasets,[1] they compared the performance of four NER models including a bidirectional long short-term memory (BiLSTM) and three transformer models (BERT, SciBERT, and MatBERT) with varying levels of domain-specific materials science pre-training. MatBERT improves over the other two BERT BASE-based models by 1% to 12%, implying that domain-specific pre-training provides measurable advantages. BiLSTM consistently outperformed BERT despite its relative simplicity, perhaps due to its domain-specific pre-trained word embeddings [18].

As a result, the researchers hypothesized that the measurable advantages previously demonstrated with domain-specific pre-training could be applied to models specific to narrower scientific disciplines such as materials science. The team concluded that domain-specific pre-training for large transformer models is still an open question in the field of NLP domain-specific [18].

III. MATERIAL AND METHODS

Classifying large text datasets is a challenging task when using transformer-based algorithms. To aid our algorithms, we want high-quality data. It is essential for training neural networks and machine learning techniques. This means we need a good-sized dataset that does not have too significant an imbalance. This section looks at choosing a suitable dataset for training our algorithms and the general architecture structure they follow. The detailed research methodology is depicted in Figure 1.

A. DATASET

The detail of the dataset is given in the subsequent subsections:

[1]**Source for the datasets:** https://figshare.com/articles/dataset/NER_Datasets_DOIs_and_Entities_Doping_and_AuNP_/16864357

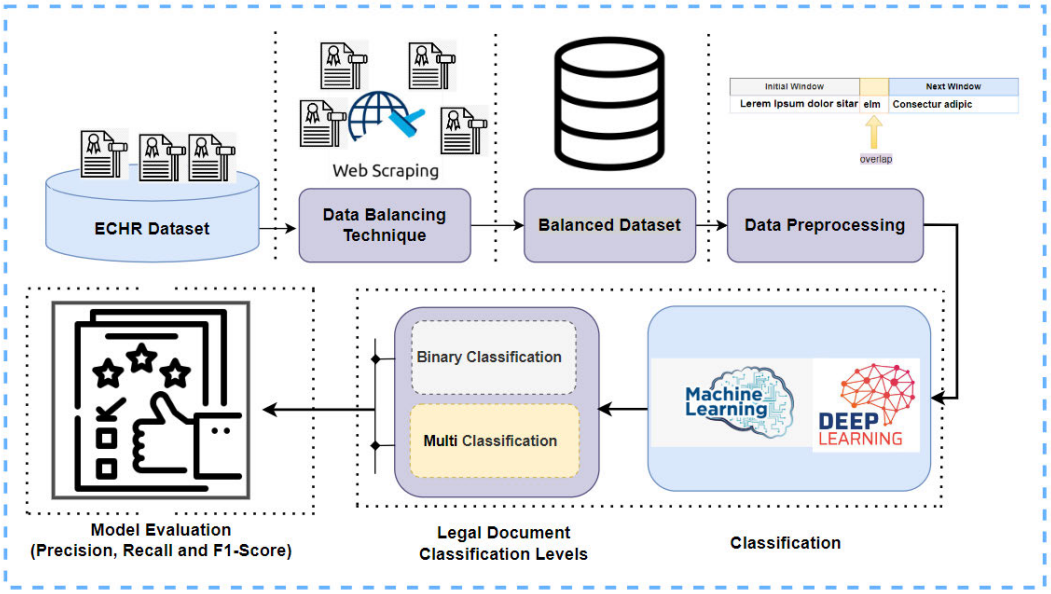


FIGURE 1. Proposed legal document classification framework.

1) CHOICE OF DATASET

The European Court of Human Rights (ECHR) dataset published publicly by Chalkidis et al. 2019 [9] bases itself on allegations of breaches of the human rights provisions. The dataset comes in two variations. The first is unaltered, while the other has anonymized identifiers of who and where the alleged violated article took place.

We chose to use the unaltered version of the dataset due to the anonymized version of the dataset showed little effect on classification results [9] and because biases in algorithms are not the focus of this research.

2) ECHR DATASET

The ECHR dataset consists of 11,500 cases from the ECHR's public database. Every case has a text describing the facts of the case. However, there are several other data features to consider:

- **BRANCH:** Which branch of the court the case was held in? Going from admissibility Court → Committee → Chamber → Grandchamber.
- **DATE:** Which year the case is from?
- **IMPORTANCE:** How important the ECHR considered the case ranging from a 1 representing a critical case contributing to the development of case law and a 4 being unimportant.
- **RESPONDENT:** Which country is accused of the human rights article breach?
- **VIOLATED ARTICLES:** Which article was violated (the classification target)?

As can be seen from Figure 2, the cases that are of greater importance than four often involve breaches of human rights

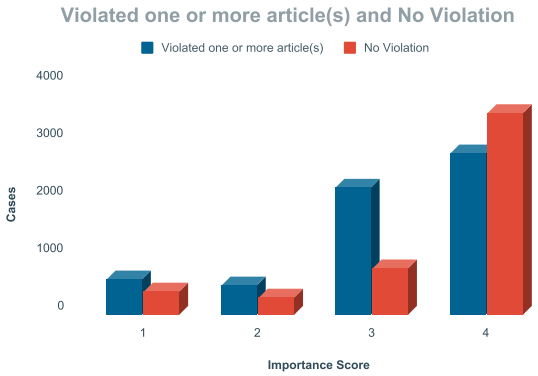


FIGURE 2. Number of violated vs no violation cases.

article(s), especially cases with an importance score of three. It will be interesting to observe what impact the addition of such information as features will have on the classification to answer RQ2.

The case details range in length from 68 to 317971 characters, with an average of 13585 characters.

3) THE SIZE OF DATASET

The size of the binary classification and the number of samples for each class in the ECHR dataset are given in Table 1. For the multi-class classification, the labels are typically more similar than the binary task of a violation versus a non-violation. It is also assumed that categorizing into which of the twenty-three articles are violated is more complex than classifying if a human rights article has been violated as shown in Table 3.

TABLE 1. Case split between Violated and Non-Violated. Note that each case can have multiple violated articles, but only count as a single violated case.

Dataset	Violated	Non-Violated
Base dataset:	6215	5263
Scraped dataset:	7320	5263

TABLE 2. Violation overview for the ECHR dataset before scraping.

Violated Article / Protocol	Occurrence
No Violation	5263
Violated Article 6: Right to a fair trial	3055
Violated Article 3: Prohibition of torture	1170
Violated Article 5: Right to liberty and security	1109
Violated Article 13: Right to an effective remedy	933
Violated Article 8: Right to respect for private and family life	734
Violated Protocol No. 1:	547
Violated Article 2: Right to life	382
Violated Article 10: Freedom of expression	355
Violated Article 14: Prohibition of discrimination	161
Violated Article 11: Freedom of assembly and association	143
Violated Article 34: Individual applications	82
Violated Article 9: Freedom of thought, conscience, and religion	60
Violated Article 38: Individual applications	36
Violated Article 7: No punishment without law	27
Violated Protocol No. 4:	23
Violated Protocol No. 7:	23
Violated Article 4: Prohibition of slavery and forced labour	9
Violated Article 12: Right to marry	9
Violated Article 18: Limitation on use of restrictions on rights	9
Violated Article 25: Plenary Court	5
Violated Protocol No. 12:	3
Violated Protocol No. 6:	2
Violated Article 46: Binding force and execution of judgments	1

B. DATASET BALANCING TECHNIQUE

A big part of the quality of a dataset is how balanced the dataset is. The distribution of violated articles in the dataset can be seen in Table 2. There are 5810 cases of non-violated articles and 5668 instances where a human rights article was violated. This is a very similar distribution between the binary case of classifying if an article is violated or not.

There is a big gap between the most violated human rights article: Article 6: Right to a fair trial with 3055 cases and Protocol No. 1 with only a single case.

To balance the dataset, we considered both scraping and data augmentation. However, due to the low number of cases ranging from one to nine in some of the classes, we did not have a large sample size for data augmentation. If we changed too few words, the newly generated sentences wouldn't have sufficient differences from the original ones to be distinguished by the classification model, which is an insignificant data augmentation outcome [19]. It is also possible that the original sentences will be transformed into entirely different ones if too many changes are made, which could result in the loss of information that contributes to the classification of the cases [20], [21]. We, therefore, went with data scraping to collect more data from the public ECHR database.

1) SCRAPED DATASET

To collect more data, we scraped the ECHR public database[2] for new cases. The points of the case scraper are:

1) **Initialize scraper** Go to database overview page.
2) **Load all cases** Loads the full webpage by scrolling to the bottom of the main page.
3) **For each case and for every accompanying case details do the following:**
   a) **Extract case name** Extract what the case name is. E.g: "In the case of X v. the Czech Republic".
   b) **Extract full article text** As it is not possible to only extract the facts about the case from the case overview, we extract all text about the case.
   c) **Go to case details** Go deeper into the specific case by going to the case details page.
   d) **Extract importance level** Extract what importance level the case was evaluated to be. This ranges from one to four.
   e) **Extract originating body** Says which court the case was judged in. E.g.: "Grand Chamber".
   f) **Extract article(s) broken** Exactly what articles where broken. E.g: "8, P1" for article eight and protocol number one.
   g) **Extract case language** Extract what language(s) the case is written up in.

The flow of the scraper can be seen in Figure 3. The figure shows how the scraper moves for every case that is scraped. Starting from the main overview page from the database that shows every case, our search is specified. After that, it loops for every case article found and extracts the necessary data from that page. Finally, coming to the end of the iteration in the case details page where the scraper extracts the final information needed.

A problem encountered when scraping the ECHR database is how it was impossible to specify just the facts needed for extraction. Because we had to extract the whole case article, we needed to use regular expressions (regex) to remove unnecessary text. Removing all texts before "THE FACTS" and keeping all text until the next section titled "RELEVANT LEGAL FRAMEWORK AND PRACTICE."

The end result of scraping is that the new dataset only has three instead of seven violations with under twenty occurrences. Table 3 shows the total amount added to each class label. With a total sum of 1979 new occurrences of violated articles across 1105 new cases.

C. DATA PREPARATION

Transformer models typically restrict the maximum length allowed for a sequence. The length is defined as the number of tokens, where a token is any of the "words" that appear in the model vocabulary. Unfortunately, each model type also

[2]ECHR public database: https://hudoc.echr.coe.int/eng

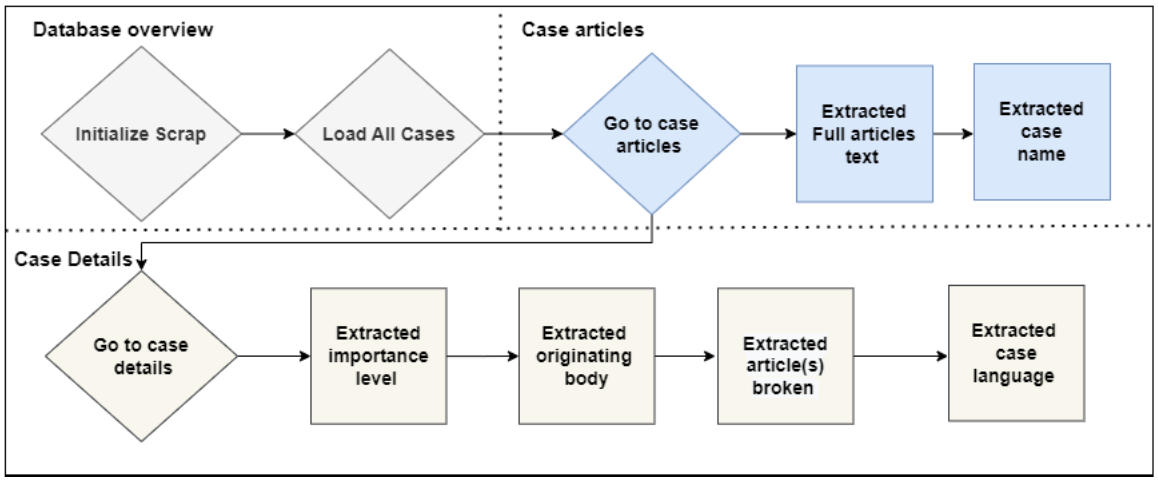


FIGURE 3. The steps followed to scrap the legal documents from the ECHR portal.

TABLE 3. Violation overview for the ECHR dataset after scraping. The increase in occurrence is shown in parentheses.

Violated Article / Protocol	Occurence
No Violation	5263 (+0)
Violated Article 6: Right to a fair trial	3502 (+ 447)
Violated Protocol No. 1:	1520 (+ 973)
Violated Article 3: Prohibition of torture	1212 (+ 42)
Violated Article 5: Right to liberty and security	1175 (+ 66)
Violated Article 13: Right to an effective remedy	1039 (+ 106)
Violated Article 8: Right to respect for private and family life	772 (+ 38)
Violated Article 2: Right to life	393 (+ 11)
Violated Article 10: Freedom of expression	370 (+ 15)
Violated Article 14: Prohibition of discrimination	178 (+ 17)
Violated Article 11: Freedom of assembly and association	151 (+ 8)
Violated Article 34: Individual applications	107 (+ 25)
Violated Protocol No. 7:	83 (+ 60)
Violated Article 9: Freedom of thought, conscience and religion	71 (+ 11)
Violated Article 7: No punishment without law	57 (+ 30)
Violated Article 46: Binding force and execution of judgments	52 (+ 51)
Violated Article 18: Limitation on use of restrictions on rights	51 (+ 42)
Violated Article 38: Examination of the case	36 (+ 0)
Violated Protocol No. 4:	29 (+ 6)
Violated Article 4: Prohibition of slavery and forced labour	20 (+ 11)
Violated Protocol No. 12:	20 (+ 17)
Violated Article 12: Right to marry	9 (+ 0)
Violated Article 25: Plenary Court	5 (+ 0)
Violated Protocol No. 6:	5 (+ 3)

has an upper bound for the token length, most commonly 512 [22].

Padding and truncation are strategies for dealing with this problem. The padding adds a special padding token to ensure shorter sequences will have the same length as either the longest sequence in a batch or the maximum size accepted by the model. Truncation works in the other direction by truncating long sequences. Truncation shortens long input

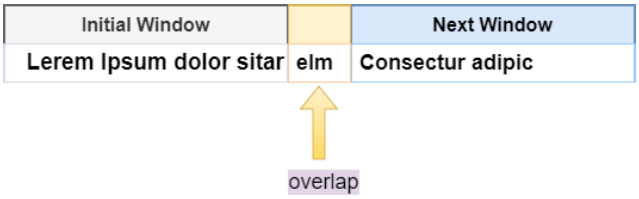


FIGURE 4. Sliding window example in a smaller scale than the larger 512 sequence windows.

text to fit the maximum length size set by models. We want to avoid this as we want as much text as possible for training our models and not to miss any text describing breached articles.

While there is currently no standard method of circumventing this issue, a plausible strategy is to use the sliding window approach. Any sequence exceeding the maximum sequence length will be split into several windows (sub-sequences). However, doing so will increase the training time of the models as all available text will now be trained on rather than shortened down from truncation.

The windows will overlap to a certain degree to minimize any information loss that hard cutoffs may cause as shown in Figure 4. The amount of overlap between the windows is determined by the stride. The stride is the distance in terms of the number of tokens that the window will be moved to obtain the next sub-sequence. We set the stride to 0.9 of the maximum sequence length resulting in about 10% overlap between the sub-sequences.

Another problem with using a sliding window is that the model has to classify several sub-sequences of the entire input text depending on the maximum length available to the model and the amount of input case text. The multiple classifications needed on all sub-sequences will further increase the training time. The total number of training samples will also grow to be higher than the number of sequences originally in the train data.

TABLE 4. Details of Machine Learning Parameters for each model.

Model	Parameter
SVM	kernel='linear', C=1, verbose=1
DT	Min_Samples_Split=2 and Min_Samples_Leaf=1
NB	Multinomial
AdaBoost	n_estimator is 50 and learning_rate=1

D. CLASSIFICATION MODELS

We have employed a variety of conventional and deep learning classification models to evaluate the proposed model and to identify the one suitable for classifying the European Court of Human Rights cases model.

1) CONVENTIONAL MACHINE LEARNING TECHNIQUES

According to Arthur Samuel, machine learning is defined as a ''computer's ability to learn without being explicitly programmed'' [23]. Conventional machine learning algorithms are based on learning from a training set to develop a trained model for further prediction [24]. The detail of each machine learning model parameter is presented in Table 4.

No single algorithm of machine learning exists that outperformed in all the areas of application, theorem given by Wolpert and Macready, named ''No free lunch'' [25]. Hence, various machine learning algorithms should be tested. Fernandez-Delgado et al. [26] evaluated the performance of 179 machine-learning classifiers on 121 different datasets. The experimental results showed that Decision Tree, SVM, AdaBoost, and Naïve based perform well on most datasets in natural language processing. Thus, this study employed four machine learning algorithms (SVM, Decision Tree, Naive Bayes, and AdaBoost).

Support Vector Machine (SVM): is a supervised machine learning algorithm. Support vectors that give SVM its namesake are the two data points closest to the hyperplane. A new hyperplane would have to be drawn if one of the two support vectors were removed from the dataset. After the hyperplane has been found, one can confidently say that the further away a dataset is from the hyperplane, the more confident we are in its class [27].

Decision Tree (DT): is a supervised machine learning algorithm that divides datasets based on rules that closely resemble human decision-making. We will focus on the classification part of decision trees; however, Decision trees can be used for regression tasks as well [28].

Naive Bayes classifier (NB): bases itself on Bayes' theorem, that is, to find the probability of an event occurring given the likelihood of another event that has already happened [29].

AdaBoost: is a meta-learning method created to increase the efficiency of binary classifiers. An AdaBoost classifier tries a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but changes the weights of incorrectly classified instances. The following classifiers focus more on the problematic cases [30].

2) TRANSFORMER-BASED NEURAL NETWORKS

We will discuss transformers, their predecessor recurrent neural networks, and the background for transformer-based neural networks that we have employed in this research as given below:

BERT: Bidirectional Encoder Representation from Transformer (BERT) is a deep contextual language representation model. BERT was developed by researchers from Google in 2018 to pre-train bidirectional representations of words from the unlabeled text. This is accomplished by jointly conditioning the left and right contexts in every layer. After this model is created, it can be fine-tuned by adding just one additional output layer to create exceptional performing models for various tasks. By conditioning both left and right context simultaneously, BERT is designed to prepare deep bidirectional representations from an unlabeled text. Therefore, the pre-trained BERT model can be fine-tuned using just one additional output layer to create state-of-the-art models for a wide range of NLP tasks. It continues to learn unsupervised from the unlabeled text and improves even when used in practical applications [31]. Its pre-training serves as a base layer of ''knowledge'' to build from there. BERT can adapt to the ever-growing body of searchable content and queries and be perfected to a user's specifications.

LEGAL-BERT: As previously discussed, BERT has demonstrated impressive results in several NLP tasks. A number of variations have developed from BERT's success, including [32]. However, only limited research has been conducted on its application to specialized domains. Reference [24] has developed a version of BERT that focuses on the legal field, known as Legal-BERT.

Legal-BERT has been pre-trained on 12 GB of diverse legal English text. During the training, the algorithm learns more about legal text than the ''regular'' BERT [24]. It was found that Legal-BERT had marginal improvements in binary classification tasks with about 1% higher accuracy than BERT and 2.5% higher accuracy in multi-class classification on legal classification tasks on the ECHR dataset.

RoBERTa: stands for Robustly Optimized BERT-pretraining Approach and was introduced by [32]. RoBERTa builds upon BERT by effectively fine-tuning BERT with further training. RoBERTa is essentially BERT retrained over 160GB of additional text. Adding to the baseline of Wikipedia and the corpus of books that BERT was trained on RoBERTa was further trained on CommonCrawl News Data, stories, and text from OpenAI GPT [32].

BigBird: was proposed by [33] and is a sparse-attention-based transformer that extends the usual Transformer based models like BERT for longer sequences.

BigBird uses sparse attention but also global attention and random attention on the input sequence. This is due to the theory of using all three attention types: sparse, global, and random attention, which equals full attention while being much more computationally efficient on longer sequences. This makes BigBird an excellent fit for modeling a network on long NLP tasks.

TABLE 5. Overview of the models used for each neural network and the datasets they were trained on. "Base" refers to BooksCorpus and English Wikipedia.

Model	Dataset pretrained on
bert-base-uncased	Base
legal-bert-base-uncased	Legal corpora
roberta-base	Base, cc-news, openwebtext, stories
bigbird-roberta-base	Base, cc-news
electra-base-discriminator	Base, Giga5, ClueWeb, Common Crawl
xlnet-base-cased	Base, Giga5, ClueWeb, Common Crawl

ELECTRA: The ELECTRA model was introduced in [34]. ELECTRA is a new approach to pretraining that trains two transformer models. One is the ''generator'', and the other is the ''discriminator''. The generator replaces tokens in a sequence trained as a masked language model. The discriminator identifies which tokens the generator replaced in the series.

XLnet: In [35], the XLNet model is proposed. XLNet is a method for learning unsupervised language representations based on generalized permutation language models. In terms of longer text language tasks, XLNet has demonstrated excellent performance using Transformer-XL as its backbone model. According to the paper from [35], BERT corrupts input with masks that remove the dependency between the masked positions and suffers from a pre-train-finetune discrepancy. They propose a generalized autoregressive pretraining method in order to counteract this corruption. In particular, XLNet is able to learn bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcoming the limitations present in BERT through its autoregressive formulation [35].

We have trained various models to make predictions on the datasets. A complete list of the exact architecture for each model can be found in Table 5.

All models except Legal-BERT are pre-trained on a joint base dataset of the BooksCorpus and English Wikipedia, totaling 13GB of plain text. However, most have some deviation in what dataset they are trained on.

Legal-BERT differs from other models in that it is developed exclusively from domain-specific data. The training data includes EU legislation, UK legislation, European Court of Justice decisions, US court cases, US contracts, and data from the ECHR database. The total amount of legal data used for pretraining Legal-BERT totals 12GB. Interestingly, Legal-BERT has been pre-trained on data from the ECHR dataset before being trained on the same dataset for classification tasks in this research.

ELECTRA is trained on the same dataset as XLNet comprised of BooksCorpus, English Wikipedia, Giga5, ClueWeb 2012-B (19GB), and Common Crawl (110GB) [35].

In comparison to the other models we test, BigBird can handle a maximum sequence length of 4096, which is 8x the maximum sequence length of 512 that the other models have. The specific parameters that change before training each model are:

- What model is to be used?
- Exactly which architecture from the model is to be used? Using architectures from Huggingface shown in Table 5.
- Which features are to be added? (We use both importance and branch).
- Maximum sequence length. For most models, the maximum sequence length is 512, except BigBird which has a maximum sequence length of 4096.
- What extra facts are to be added to the facts about the case? This can be what branch of court the case heard, what importance score the case was given, or both.

Evaluation of Models

The dataset was split into training and evaluating sets (70%-30%), respectively. The split between Violated and Non-Violated articles can be seen in Table 1. In this step, we evaluated the performance of two approaches, i.e., conventional machine learning and transformer-based techniques. In addition to this, we used three performance metrics, namely, precision, recall, and F1-score to measure the performance of the proposed model. The definitions of each performance metric are listed below:

Precision and Recall:

The precision score gets the number of relevant predictions from the model's total number of positive predictions. The precision score is calculated by dividing the number of true positives by the number of positive predictions [36]. With this, one can tell how many of the predictions are relevant. A 100% precision score would mean no false positive predictions, and every positive prediction from the model was correct. Whereas recall shows the number of cases the model correctly classified as relevant out of all relevant cases.

The recall score is calculated by true positives divided by the number of predictions that should have been positive. A 100% recall score would mean no false negatives are predicted. Therefore, all negative predictions are correct.

F1-score:

F1-score is the cumulative result of combining the precision and recall score into a single metric [37]. F1-score is primarily used for comparison between classifiers. This is because having a good recall score does not guarantee a good precision score and vice versa. F1-score results in a harmonic mean between the two measuring units [38].

IV. EXPERIMENTAL SETTINGS AND RESULTS

In this section, we look at the experimental setup. In this work, we have performed two settings; binary and multi-large legal documentation classification. We have employed conventional machine learning algorithms (SVM, DT, NB, and AdaBoost) and Transformer based neural networks (BERT, Legal-BERT, BigBird, ELECTRA, and Xlnet). In addition to this, we evaluated the large legal document classification results in terms of precision, recall, and F1 score. We apply the K-fold validation technique to the dataset and the K value was 5.

TABLE 6. Results from neural network models without any added features (± std. dev).

Model	Precision	Recall	F1-score
BERT	86.6±0.1	85.0±0.0	85.9±0.1
Legal-BERT	86.3±0.1	85.4±0.2	85.6±0.1
RoBERTa	**89.1**±0.1	**86.2**±0.1	**86.7**±0.1
BigBird	86.9±0.2	86.1±0.0	86.3±0.1
ELECTRA	86.1±0.3	85.1±0.1	85.3±0.1
XLNet	86.7±0.1	86.0±0.1	86.2±0.0

TABLE 7. Results from Chalkidis et al. [9].

Model	Precision	Recall	F1-score
COIN-TOSS	50.4±0.7	50.5±0.8	49.1±0.0
BOW-SVM	71.6±0.0	72.0±0.0	87.8±0.0
BERT	24.0±0.2	50.0±0.0	17.0±0.5
HIER-BERT	**90.4**±0.3	**79.3**±0.9	**82.0**±0.9
HAN	88.2±0.4	78.3 ±0.2	80.5±0.2
BIGRU-ATT	87.0±1.0	77.2±3.4	79.5±2.7

TABLE 8. Results from conventional machine learning techniques without any added features(± std. dev).

Model	Precision	Recall	F1-score
SVM	80.0±0.01	79.4±0.01	79.6±0.01
Decision Tree	78.6 ±0.01	78.5±0.01	78.5±0.01
Naïve Bayes	71.4±0.03	69.0±0.03	69.4±0.03
AdaBoost	**81.9**±0.05	**81.5**±0.05	**81.7**±0.04

A. EXPERIMENT 1: BINARY CLASSIFICATION

In this section, we are focusing on binary classification. The dataset has two classes, and each case must belong to one. We also look at the results of adding additional features to the dataset i.e., the branch and the importance. Every document present in the datasets is labeled with one class only when doing binary classification. Due to the large size of the raw text in the dataset, we had to train on a very small batch size of four. Combined with the small batch size, we trained the algorithms at a learning rate of 4e-5 over ten epochs.

Comparing the results from our models seen in Table 6 to the original paper from [9] in which the best performing model was HIER-BERT. This BERT model allowed for a bigger length sequence that had an F1-score of 82%. We can see there has already been a big improvement.

The improvements are likely due to the stride applied to allow for learning beyond the transformer's usual limited length. This study [9] had an abysmal performance from BERT with an F1-score of 17.0% (Table 7). The obtained F1-score is worse than just randomly guessing. The likely culprit is the truncation of the case facts BERT was trained on [9]. As in our case, using a striding window allowed BERT to train on the whole case description, which in turn resulted in an F1-score of 85.9% for the BERT (Table 6).

The best-performing model on the base dataset without any added features was RoBERTa. The performance makes sense as RoBERTa in essence modified to improve on BERT. The modifications were done by increasing the amount of pre-training data and hyperparameter tuning.

TABLE 9. Results from multi-classification with additional scraped data showing the weighted average results for each model (± std. dev).

Model	Precision	Recall	F1-score
BERT	77.4±0.1	74.2±0.0	75.2±0.0
Legal-BERT	78.4±0.1	74.5±0.2	76.1±0.2
RoBERTa	78.9±0.3	77.0±0.2	77.7±0.1
BigBird	**80.0**±0.2	**77.1**±0.2	**78.1**±0.1
Electra	77.1±0.1	74.3±0.0	75.2±0.1
XLNet	79.3±0.2	77.7±0.1	77.9±0.3

Using the results from multiple conventional machine learning techniques shown in Table 8 as a comparison for the transformer-based models, it is clear that neural models are superior. Yet, the results achieved by AdaBoost and SVM are at par with those reported in Table 7 by [9].

B. EXPERIMENT 2: MULTI-CLASS CLASSIFICATION

In multi-class classification, the labels are typically more similar than the binary task of a violation versus a non-violation. It is also assumed that categorizing into which of the twenty-three articles are violated is more complex than classifying if a human rights article has been violated.

To be able to directly compare our results with the results reported in [9], we trained the models on the original dataset and included no violation as a possible label.

After scraping additional data, we retrained all the models on the new dataset. The results are shown in Table 9. With a weighted F1-score of 78.1%, BigBird had the best results. It outperformed other models that have lower maximum sequence lengths. Having fewer sub-sequences is believed to be the reason for BigBird's superior performance. Sub-sequences did not pose a significant problem for the other models in the binary classification task. As in the binary classification task, classifying each sub-sequence as either a violated or non-violated article resulted in the final full sequence prediction only being affected by the sum of these two labels.

However, in the case of multi-class classification, an example of a single case with a sequence length of 5000 will result in ten sub-sequences. Within each sub-sequence, there can be 23 different ''sub-predictions'' on which labels are violated. With a maximum sequence length of 4096, there would only be two subs sequences for BigBird. Compared to the previous unscraped dataset, BigBird and most of the other models had a much higher true positive result on the lower violated article counts. The confusion matrix for each individual label prediction from the best performing multi-class classification model BigBird is illustrated in Figure 5. Please note that in the figure, the TN is in position (0,0) and the TP is in position (1,1).

As a result, some of the lower values, such as Article 18 and Protocol 12, had over 50% of their cases categorized correctly, even though they had only 15 and 6 cases in the evaluation set, respectively.

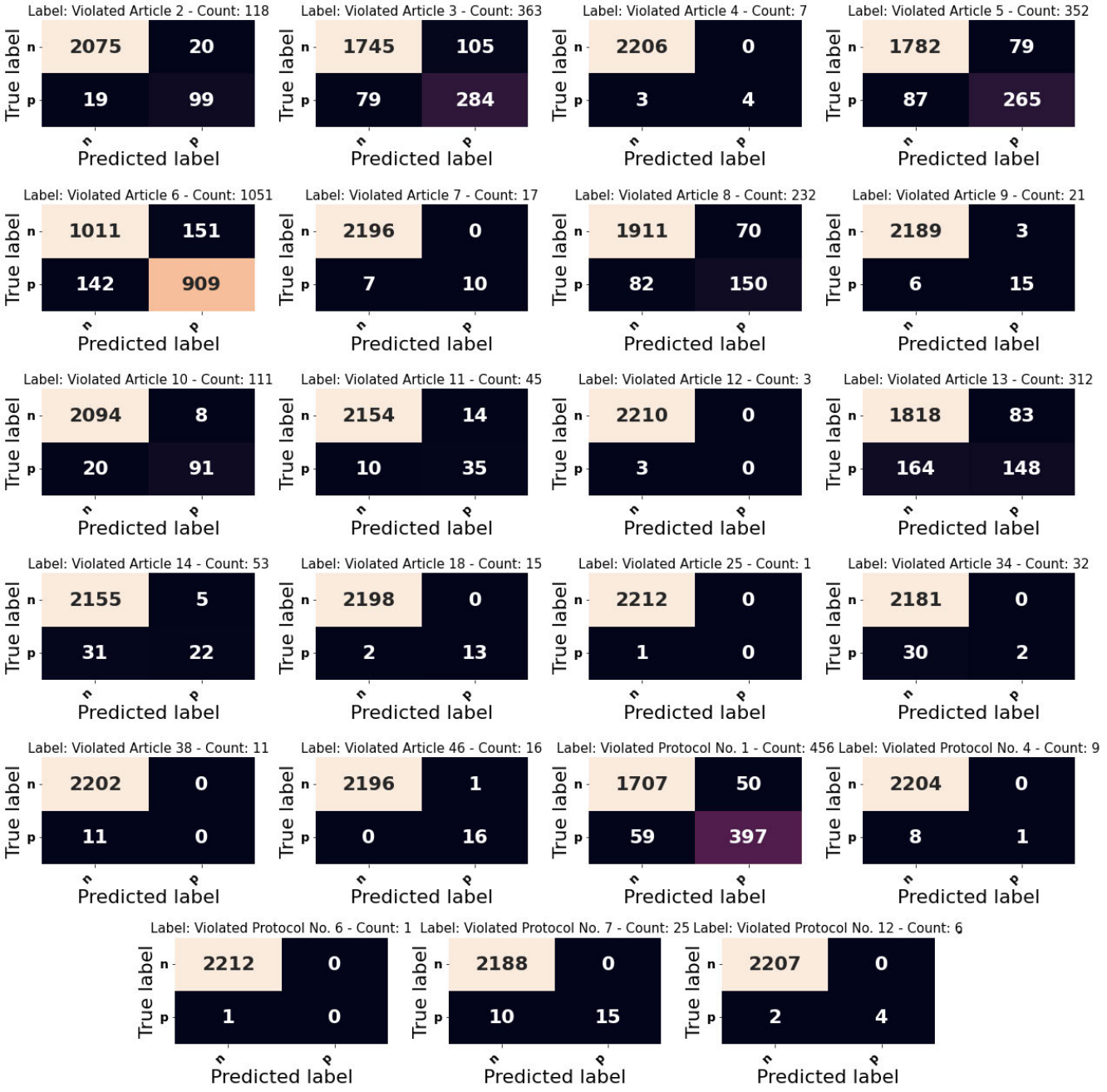


FIGURE 5. **Individual confusion matrices for each category based on the BigBird model's results.**

C. EXPERIMENT 3: ADDED FEATURES

We further expanded the feature set to test both binary (violated vs non-violated) and multi-class (violated articles) with Branch code and importance from the ECHR dataset specified in III-A2. Testing RoBERTa and BigBird yielded no significant results concerning accuracy as already reported in Table 6 and Table 7. This is mostly due to the fact that the dataset didn't contain enough samples with uniquely important cases in specific branches for specific article breaches. Having said that, we do believe taking additional meta-data case information and features may improve model performance with the addition of newer cases on the ECHR portal in years to come.

D. DISCUSSION

RoBERTa produced the most significant result on the binary task, and BigBird performed the best on the multi-class classification task. This shows the weakness of using a sliding window when multiple classification options exist. As the options become so numerous, it becomes increasingly difficult for the

models to correctly classify all the generated subsequences, which led to BigBird, which had fewer subsequences, outperforming the others [39]. Despite the shortcomings outlined above, the addition of a sliding window had a positive effect. However, the addition of extra facts to the sequences yielded little to no impact. The little impact is most likely due to the already large text base to classify from, and we would probably have seen a more significant effect of the additional facts on smaller text sequences.

We were limited to testing the models on the ECHR dataset alone due to the lack of publicly available anonymized legal datasets of this nature and to compare the results to existing works. Unfortunately, there are only a handful of examples of ECHR-based text classification applications to compare. The one by [9] reported the best results on this dataset to our knowledge to which we compared our results. Additionally, our research scraped additional data for the multi-class classification task, the results may differ from those of others who use the base dataset from [9]. However, for the binary classification task, we outperformed the benchmark results on the same dataset as reported in [9].

In search to our questions posed earlier, *for RQ1: How effective are transformers in handling long sequences of text data for legal documents?*, we found that using a sliding window approach (as discussed in the section III-C) to handle the long data sequences resulted in state-of-the-art results. Its downside is the large computational load and the potential for overfitting in the overlap between the steps. We further examined that the addition of what should be impactful features had little effect due to the length of information already available to the models, in response to *RQ2*. However, we summarize that the impact would be more significant on smaller text sizes, which is in line with the findings reported in [21]. Also, a possible ensembling model which could use both the transformers on the full-length text and the individual extra feature facts in another model might yield better results.

Additionally, we show the possibility of classifying cases not only as violated or not but specifically which ones are possible via multilabel classification for article breaches to check how viable transformer-based solutions be on legal documents *(RQ3)*. Accordingly, we conclude that using models such as the ones used in this research can help to reduce the cost of legal aid.

V. CONCLUSION AND FUTURE DIRECTIONS

We have carried out experiments using six different neural models on two text classification tasks. Within the text classification tasks, we have looked at the results of adding additional data features to the initial data sequence resulting in various different workflows. According to these experiments, adding more information to the extensive text documents had little impact on the classification outcome. As the amount of text data available to the models was already so large, there was little impact expected. Nevertheless, we expected at least some effect, because looking at the dataset, it was possible to draw some conclusions based on the importance given and from what branch cases came. Moreover, the solution of a sliding window over the input text resulted in state-of-the-art results from various models. As the primary contribution of our paper, we present improved baselines that can form the basis for future research.

While state-of-the-art transformer models have shown success on large legal text classification tasks, there is still room for improvement, as indicated by our experiments. In addition to the multi-class tasks, where the models are far from perfect, there is also much room for improvement in the binary classification tasks. In the following text, we suggest potential areas for further improvement of the results of this research.

Domain Specific Pretraining: A potential model that could achieve good results in binary and multi-class classification could be a combination of Legal-BERT and BigBird. Increasing the maximum sequence length, such as BigBird, can help reduce the amount of overlapping and overfitting issues that may arise due to the sliding window technique. This new model could also be enhanced by being pre-trained on domain-specific corpora, such as Legal-BERT.

Ensemble Model: Another possible model would be to create a transformer ensemble model that would emphasize the extra features we tried to introduce. It is possible that the model would allow for better use of these features than the transformer models and together make for a more successful classification model as a whole. Ensemble modeling could eliminate the noise of introducing new features to the long text. The transformer models currently rely too heavily on the case details and do not consider the additional features added.

Dataset Quality: Datasets are essential to the performance of transformer models. Obtaining high-quality datasets is a time-consuming process, and while we explored scraping additional data, further work exists to be done, and we propose that this be our focus. Additionally, there is exciting work being done on dataset augmentation, which could be further explored, such as Shaikh et al. [21], Shorten et al. [40] and Karras et al. [41].

REFERENCES

[1] J. P. Olsen, "Sentraladministrasjonen I en utfordrende æra: Tid for ettertanke," *Norsk Statsvitenskapelig Tidsskrift*, vol. 35, no. 1, pp. 4–27, Apr. 2019.
[2] J. E. Carlin and J. Howard, "Legal representation and class justice," *UCLA L. Rev.*, vol. 12, p. 381, Jan. 1964.
[3] I. Tønnessen, "Likhet for loven—Lov om støtte til rettshjelp," Regjeringen.no, Norway, Tech. Rep. 5, 2020, pp. 136–137.
[4] D. Hirsch, "Priced out of justice? Means testing legal aid and making ends meet," Loughborough Univ., Loughborough, U.K., Tech. Rep. 3, 2018.
[5] H. Donald, "Establishing a national standard: The role of the UK's minimum income standard in policy and practice," in *Minimum Income Standards Reference Budgets*. Bristol, U.K.: Policy Press, 2020, pp. 307–318.
[6] A. K. Chaturvedi and R. Koul, "Legal aid and legislative initiatives," *Think India J.*, vol. 22, no. 14, pp. 11256–11266, 2019.
[7] H. Henrik. (2022). *Classifying European Court of Human Rights Cases Using Transformer Based Models*. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/3033966
[8] R. Bommasani, "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.
[9] I. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural legal judgment prediction in english," 2019, *arXiv:1906.02059*.

[10] I. Chalkidis, M. Fergadiotis, S. Kotitsas, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "An empirical study on large-scale multi-label text classification including few and zero-shot labels," 2020, *arXiv:2010.01653*.

[11] A. Lage-Freitas, H. Allende-Cid, O. Santana, and L. Oliveira-Lage, "Predicting Brazilian court decisions," *PeerJ Comput. Sci.*, vol. 8, p. e904, Mar. 2022.

[12] D. W. Oard, "Information retrieval for E-discovery," *Found. Trends Inf. Retr.*, vol. 7, nos. 2–3, pp. 99–237, 2013.

[13] J. G. Conrad, "E-discovery revisited: The need for artificial intelligence beyond information retrieval," *Artif. Intell. Law*, vol. 18, no. 4, pp. 321–345, Dec. 2010.

[14] *Ediscovery Market*, I. Radicati Group, Palo Alto, CA, USA, 2019, pp. 3–4.

[15] A. Adhikari, A. Ram, R. Tang, and J. Lin, "DocBERT: BERT for document classification," 2019, *arXiv:1904.08398*.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[17] X. Dai, S. Karimi, B. Hachey, and C. Paris, "Cost-effective selection of pretraining data: A case study of pretraining BERT on social media," 2020, *arXiv:2010.01150*.

[18] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–23, Jan. 2022.

[19] A. S. Imran, R. Yang, Z. Kastrati, S. M. Daudpota, and S. Shaikh, "The impact of synthetic text generation for sentiment analysis using GAN based models," *Egyptian Informat. J.*, vol. 23, no. 3, pp. 547–557, Sep. 2022.

[20] J. Gao, "Data augmentation in solving data imbalance problems," M.S. thesis, KTH Roy. Inst. Technol. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:1521110/FULLTEXT01.pdf

[21] S. Shaikh, S. M. Daudpota, A. S. Imran, and Z. Kastrati, "Towards improved classification accuracy on highly imbalanced text dataset using deep neural language models," *Appl. Sci.*, vol. 11, no. 2, p. 869, Jan. 2021.

[22] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med. Res. Methodol.*, vol. 22, no. 1, pp. 1–12, Dec. 2022.

[23] K. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, Aug. 2018.

[24] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," 2020, *arXiv:2010.02559*.

[25] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, Apr. 1997.

[26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[27] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[28] S. B. Kotsiantis, "Decision trees: A recent overview," *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013.

[29] K. M. Leung, "Naive Bayesian classifier," *Polytech. Univ. Dept. Comput. Sci./Finance Risk Eng.*, vol. 2007, pp. 123–156, Nov. 2007.

[30] R. E. Schapire, "Explaining adaboost," in *Empirical Inference*. Cham, Switzerland: Springer, 2013, pp. 37–52.

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

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[33] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2020, *arXiv:2007.14062*.

[34] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.

[35] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," 2019, *arXiv:1906.08237*.

[36] M. Buckland and F. Gey, "The relationship between recall and precision," *J. Amer. Soc. for Inf. Sci.*, vol. 45, no. 1, pp. 12–19, Jan. 1994.

[37] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 233–240.

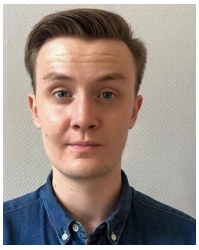
[38] J. Opitz and S. Burst, "Macro F1 and macro F1," 2019, *arXiv:1911.03347*.

[39] G. Michalopoulos, "Innovations in domain knowledge augmentation of contextual models," Ph.D. thesis, Dept. Comput. Sci., Univ. Waterloo, Waterloo, ON, USA, 2022.

[40] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text data augmentation for deep learning," *J. Big Data*, vol. 8, no. 1, pp. 1–34, Dec. 2021.

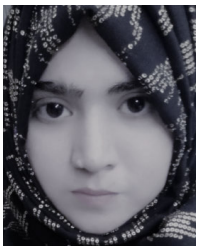
[41] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," Oct. 2020, *arXiv:2006.06676*.

ALI SHARIQ IMRAN (Member, IEEE) received the master's degree in software engineering and computing from the National University of Sciences and Technology (NUST), Pakistan, in 2008, and the Ph.D. degree in computer science from the University of Oslo (UiO), Norway, in 2013. He is currently an Associate Professor with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway. He is a member of the Norwegian Colour and Visual Computing Laboratory (Colourlab). He specializes in applied research, with a focus on deep learning technology and its application to signal processing, natural language processing, and multi-modality media analysis. He has over 100 peer-reviewed journals and conference publications to his name. He served as a Reviewer for many reputed journals over the years, including IEEE Access, as an Associate Editor.



HENRIK HODNEFJELD received the master's degree from the Norwegian University of Science and Technology (NTNU), in 2022. He was with Vima Consulting after his master's thesis. He is currently with twoday as a Consultant and a Project Manager.

ZENUN KASTRATI received the master's degree in computer science through the EU TEMPUS Program developed and implemented jointly by the University of Pristina, Kosovo, Universite de La Rochelle, France, and the Institute of Technology Carlow, Ireland, and the Ph.D. degree in computer science from the Norwegian University of Science and Technology (NTNU), Norway, in 2018. He is currently with the Department of Informatics, Linnaeus University, Sweden. He is the author of more than 50 peer-reviewed journals and conferences. His research interests include artificial intelligence with a special focus on NLP, machine/deep learning, and sentiment analysis. He has served as a reviewer for many reputed journals over the years.



NOUREEN FATIMA received the master's degree in computer science from Sukkur IBA University, Sukkur, Pakistan, in 2021. She is currently a Research Assistant with the Center of Excellence for Robotics, Artificial Intelligence, and Blockchain, Department of Computer Science, Sukkur IBA University. She is the author of several articles published in international journals. Her research interests include applied research in the field of artificial intelligence and its application to signal processing, and natural language processing. She served as a Reviewer for IEEE Access.

SHER MUHAMMAD DAUDPOTA received the master's and Ph.D. degrees from the Asian Institute of Technology, Thailand, in 2008 and 2012, respectively. He is currently a Professor of computer science with Sukkur IBA University, Pakistan. Alongside his computer science contribution, he is also a Quality Assurance Expert in higher education. He has reviewed more than 50 universities in Pakistan for quality assurance on behalf of the Higher Education Commission in the role of an Educational Quality Reviewer. He is the author of more than 35 peer-reviewed journals and conference publications. His research interests include deep learning, natural language processing, video, and signal processing.

MUDASIR AHMAD WANI received the Master of Computer Applications (M.C.A.) and M.Phil. degrees in data mining from the University of Kashmir (UoK), in 2012 and 2014, respectively, and the Ph.D. degree in computer science from Jamia Millia Islamia (A Central University), New Delhi, India, in 2019. He pursued his postdoctoral research with the Norwegian Biometrics Laboratory, Norwegian University of Science and Technology (NTNU), Norway. He was a Lecturer and a Researcher with the Department of Information Security and Communication Technology (IIK), NTNU. He is currently a Researcher in NLP with Prince Sultan University, Saudi Arabia. He is actively involved in organizing and reviewing international conferences, workshops, and journals. His research interests include the extraction and analysis of social data and the application of different statistical and machine/deep learning techniques in developing prediction models. He was a recipient of the Alain Bensoussan Fellowship Award under the European Research Consortium for Informatics and Mathematics (ERCIM), Sophia Antipolis Cedex, France.