2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) | 978-1-6654-9810-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/INISTA55318.2022.9894051

Traditional Machine Learning and Deep Learning-based Text Classification for Turkish Law Documents using Transformers and Domain Adaptation

Onur AKÇA
Computer Engineering Department
Marmara University
Istanbul, Turkey
onur.akca@marun.edu.tr

Gıyaseddin BAYRAK
Computer Engineering Department
Marmara University
Istanbul, Turkey
giyaseddinalfarkh@marun.edu.tr

Abdul Majeed ISSIFU
Computer Engineering Department
Marmara University
Istanbul, Turkey
abdul.majeed@marun.edu.tr

Murat Can GANİZ
Computer Engineering Department
Marmara University
Istanbul, Turkey
murat.ganiz@marmara.edu.tr

***Abstract*—Natural Language Processing (NLP) is an interdisciplinary field between linguistics and computer science. Its main aim is to process natural (human) language using computer programs. Text classification is one of the main tasks of this field, and they are widely used in many different applications such as spam filtering, sentiment analysis, and document categorization. Nonetheless, there is only very little text classification work in the law domain and even less for the Turkish language. This may be attributed to the complexity within the domain. The length, complexity of documents, and use of extensive technical jargon are some of the reasons that distinguish this domain from others. Similar to the medical domain, understanding these documents requires extensive specialization. Another reason can be the scarcity of publicly available datasets. In this study, we compile sizeable unsupervised and supervised datasets from publicly available sources and experiment with several classification algorithms ranging from traditional classifiers to much more complicated deep learning and transformer-based models along with different text representations. We focus on classifying Court of Cassation decisions for their crime labels. Interestingly, the majority of the models we experiment with could be able to obtain good results. This suggests that although understanding the documents in the legal domain is complicated and requires expertise from humans, it may be relatively easier for machine learning models despite the extensive presence of the technical terms. This seems to be especially the case for transformer-based pre-trained neural language models which can be adapted to the law domain, showing high potential for future real-world applications.**

***Index Terms*—Legal document classification, Natural Language Processing, Domain-specific language models**

I. Introduction

Text classification is one of the downstream tasks in Natural Language Processing (NLP) and widely used in many different applications such as spam filtering, sentiment analysis, and document categorization [1]. Interestingly, there are only a very few text classification work in law domain and even less for Turkish language [2]. This may be attributed to the complexity of the domain. The length, complexity of documents and use of extensive technical jargon are some of the reasons separates this domain from others. Similar to the medical domain, understanding these documents requires extensive specialization [3]. Another reason can be the scarcity of publicly available datasets. Although there are relatively small amount of academic work in law domain, computer systems are extensively used by legal authorities around the world, and there is an abundance of text documents as almost all stages of the legal processes generate free-style text documents. There are potentially billions of cases in almost all of the court houses in the world. These eventually end up with tons of court decisions stored in databases. One of the obvious applications is to organize these large amounts of legal documents into their respective categories such as crime types. This task can be formulated as a classification problem. This classification problem is mostly done by human experts who are specialized in the domain, therefore it is expensive.

Machine Learning, specifically, NLP techniques can help in automating this process. Traditional algorithms have been used in an attempt to solve this problem. The recent advancements in deep learning allow them to achieve usually higher performance compared to the traditional machine learning algorithms for classification tasks. Neural Network (NN) architectures such as Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM) and most recently transformer [4] based models hold the current state-of-the-art results for text classification, as well as several other downstream NLP

978-1-6654-9810-4/22/$31.00 ©2022 IEEE

tasks. However, they require a much larger amount of training data compare to their traditional machine learning-based counterparts. Availability of pre-trained models for transformer-based large neural language models contribute the popularity of these models. One interesting feature of transformer-based models is that pre-trained models can be fine-tuned to a specific downstream NLP task or domain to increase the performance on a specific application domain such as medical domain or law domain. In this study, we try to answer the question "Can we use a machine learning to classify Turkish legal documents?" and more specifically "Can a transformer trained on Turkish legal corpora outperform other machine learning algorithms?".

We aim to see how well machine learning methods perform in classifying legal documents in Turkish. We also investigate both traditional machine learning methods and deep learning-based models to shed light into this understudied topic. We also want to examine the performance of contextual representations by pre-trained models of transformers as one of the state-of-the-art methods and if their performance can be increased by fine-tuning, adapting into the legal domain.

To the best of our knowledge, this is one of the early works in Turkish legal document classification task. Also, found no clue about the presence of a standard benchmark dataset available for this very task yet, so we collect and use real world data to create datasets that we could conduct our experiments on. Also, we collect a much larger unsupervised law-related corpus to observe the effect of domain-specific fine-tuning on the overall classification performance. Our contributions can be summarized as follows: 1) We present the results for a Turkish law document classification application using real-world data, 2) We provide a comparison between the baseline model and transformer models' performance. 3) One of the first studies to approach court decisions classification in Turkish.

The following sections in this work will be organized as such: In Section II, we present an overview about the current state-of-the-art in text classification and the work of classifying legal documents, Section III explains our data collection and the creation processes of the datasets. In Section IV we go through our approach for and the followed methodology for solving the problem. Section V presents the results of the experiments and highlight the best and worst performing models and the differences between them. Finally we conclude the findings of this study in Section VI with a quick summary of the potential directions and improvements that could be done in the future.

II. Related Work

Text classification being one of the early downstream tasks of NLP, gets lots of research attention for years. In the literature, many studies tackled text classification where text is categorized into various tags, labels or classes [1]. The approaches employ many different machine learning and deep learning algorithms. Even though there are massive advancements and achievements in text classification, they are mostly using the benchmark datasets in English. There are relatively few studies in low-resource languages such as the Turkish language. We dive into performing text classification in Turkish, specifically, in legal documents. Regarding the technical methods, general-purpose text classification approaches also apply for the Turkish language, with some nuances. These nuances exist because of the syntactic and morphological features that are unique to the Turkish language [2].

In the early stages of NLP, the techniques used were largely frequency-based methods. These methods are named Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF) which calculates the frequency for the whole vocabulary in a given document. These techniques perform well in distinct topics, but fall short in finding the variations between synonyms. The order of relationships between words are not taken into consideration in the early proposed techniques. Machines do not understand and can not process text sequences in their raw format. The meaning, contextual dependency, semantic similarity etc. of the sequence needs to be taken into consideration, so, the idea of introducing vectors with fixed size could represent any word within the vocabulary overcomes the limitations and enables machines to process text sequences. This approach is commonly known as word embeddings. Word2Vec [5] is one of the first embeddings models proposed in this regard. This algorithm uses modern neural networks to train on large amount of text data to learn the relationships between words in the corpus. Mikolov et al. [5] in their work compute a continuous vector representations of words in a given corpus, which results in finding similarities and semantic relationships between the words in the corpus.

Classification of legal texts is a relatively new field with its challenges. Firstly, legal texts are tended to have very long sentence lengths which can be a difficulty for many NLP algorithms. Another challenge is the fact that legal text is highly complicated and contains lots of technical jargon. Clean and labeled data scarcity in the Turkish legal domain is another challenge not to be left out. Researchers develop answers to each of these problems in other domains of NLP which may also prove useful in the legal domain.

Despite all of the difficulties, machine learning methods seem to be plausible tools in the legal domain. Orosz et al. [6] in their work, studied Hungarian legal documents for classification. Şulea et al. [7] show that Support Vector Machines (SVM), a traditional machine learning classifier can be used to classify rulings of the French Supreme Court. Howe et al. investigates machine learning methods for classification of Singaporean Supreme Court decisions [8]. In a recent study, Chen et al. [9] showed that embeddings created using information extraction and feature engineering techniques give better results than neural network models, even though they used a traditional machine learning classifier, random forest. The work of Mumcuoğlu et al. [10] is the only study we could find on Turkish legal NLP tasks. Their study shows the performance of the traditional machine learning algorithms such as Decision Trees, Random Forests, Support Vector Machines and deep learning models like Bidirectional Long

Short-Term Memory (biLSTM) and Gated Recurrent Units (GRU). They predict the rulings of the Turkish Constitutional Court and Courts of Appeal. We also need to look at a more specific task under text classification, namely extreme multi-label text classification (XMTC) where there may be hundreds or thousands of distinct classes and one document can have multiple labels. Liu et al. [11] shows the performance of Convolutional Neural Networks (CNN). They argued that CNNs can be used in this domain with a hidden bottleneck layer for better representations of documents, and that binary cross-entropy loss is more suitable for this multi-label classification. Gargiulo et al. [12] proposes a combination of different word embeddings to better capture the grammatical and syntactic features of the text. They call their method Hierarchical Label Set Expansion. Chalkidis et al. [13] defends replacing CNNs with bi-directional GRUs to get better results on European Union Legislature. As mentioned, transformer-based models are currently the state-of-the-art in many NLP tasks. These models can also be trained specifically for legal text. Two recent studies investigate this area. Chalkidis et al. [14] fine-tunes the Bidirectional Encoder Representations (BERT) model on English legal corpora which they assembled using datasets such as European Court of Human Rights, European Union Legislature, and Supreme Court of United States. Zheng et al. [15] also works on a similar study with Harvard Law case corpus. Both studies shows that a BERT based model fine-tuned on legal corpora achieves best results in legal text classification. Transformers are also used for XMTC task. Chang et al. [16] proposes first transformer model for XMTC. Same team also trained BERT for XMTC task which they named X-BERT [17].

Both of the studies of Chalkidis et al. [14] and Zheng et al. [15] shows that fine-tuning transformer-based models increase the accuracy of the model for classification of legal texts in English. Orosz et al. [6] and Şulea et al. [7] shows that using traditional methods can achieve results on par with deep learning models. Mumcuoğlu et al. [10] sets the baseline both for traditional and neural network-based algorithms however does not cover transformer-based methods. In our work, we train transformer-based models, which are the state-of-the-art for many NLP tasks including classification. We compare our model with both deep learning-based and traditional machine learning-based methods for classification of Turkish legal texts.

III. The Dataset

Data is always the fuel to any machine learning or deep learning project. In our work, we attempt to find domain-specific datasets in the Turkish legal domain. To the best of our knowledge, there is no large-scale benchmark dataset in the Turkish legal domain. So we compiled our datasets from publicly accessible resources. We need two datasets; a supervised single-label multi-class dataset for classification, and an unsupervised Turkish legal domain corpus for neural language model training. As one of our main sources, we use search engine of the Court of Cassation of Turkey [1] which serves 6,384,952 decisions (içtihat) at the writing of this proceeding to the public. In addition to this, we also use Turkish legislature and Ph.D. dissertations on the Law field in Turkish. Both of these sources are also publicly available.

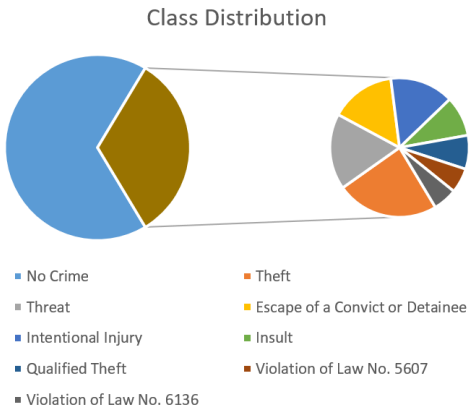

Fig. 1. Class distribution of the supervised dataset.

Supervised Dataset : While building this dataset, court decisions for the first 6 months of 2021 are downloaded from the search engine of the Turkish Court of Cassation. Court of Cassation reviews judgements of justice and criminal courts of Turkey and renders a verdict upon appeal. Ideally, first instance courts take the opinions rendered by the Court of Cassation as precedents to form a uniform application throughout the country. Court decisions from criminal courts can contain several crime labels while those from justice courts contain no crime labels. It is a multi-label dataset. In order to simplify our classification experiments, we reduce this dataset to a single-label dataset by taking the first crime label only, if there are more than one label in a document. The justice court decisions are labeled as 'No Crime'. Most of the crime labels are self-explanatory apart from violation of the law no. 5607 and 6136, which are the law against smuggling and the law that regulates carrying firearms, respectively. We could be able to download around 200,000 court decisions. From these 200,000 we decide to use documents, that belong to the most frequent nine labels to ensure enough training data for each class. We include the remaining documents in our unsupervised dataset. Table II shows the distribution of classes in our supervised dataset. Please be aware that this class distribution may not reflect the actual crime distribution since we only used the data for the first 6 months of 2021. We split the data into training, evaluation, and test sets with the ratio of 80%, 10%, and 10% respectively in a stratified manner preserving class distribution. Table I shows an example text from this dataset.

Unsupervised Dataset : For pre-training or fine-tuning transformer models, we need a Turkish Legal Corpus. To assemble this corpus we start with decisions of the Court of Cassation which is not used in the supervised dataset.

[1] https://karararama.yargitay.gov.tr/

TABLE I
AN EXAMPLE FROM THE SUPERVISED DATASET

Crime Label: Threat
K A R A R Yerel Mahkemece verilen hüküm temyiz edilmekle, başvurunun süresi ve kararın niteliği ile suç tarihine göre dosya görüşüldü; Temyiz isteğinin reddi nedenleri bulunmadığından işin esasına geçildi. Vicdani kanının oluştuğu duruşma sürecini yansıtan tutanaklar, belgeler ve gerekçe içeriğine göre yapılan incelemede: Eyleme ve yükletilen suça yönelik katılan ... vekilinin temyiz nedenleri yerinde görülmediğinden tebliğnameye uygun olarak, TEMYİZ DAVASININ ESASTAN REDDİYLE HÜKMÜN ONANMASINA, 13/04/2021 tarihinde oy birliğiyle karar verildi.

TABLE II
CLASS DISTRIBUTION OF THE SUPERVISED DATASET.

Label	**Training Set**	**Validation Set**	**Test Set**
No Crime	67960	8495	8496
Theft	7894	987	987
Threat	5846	730	731
Escape of a Convict	5012	626	627
Intentional Injury	4914	615	614
Insult	3097	387	387
Qualified Theft	2606	326	326
Violation of Law No. 5607	1914	239	239
Violation of Law No. 6136	1853	232	231

From court decisions, we use 195,376 documents. We also use 14,207 documents from Turkish Legislation and 2,245 Turkish Doctoral Thesis' on Law field. The average document length is relatively high in doctoral thesis with an average length of 102,062 per document. We collect these legal documents in a single text file without any pre-processing to assemble the unsupervised corpus. In this corpus, each sentence is a new line. This corpus is 2.6 GB in size and has 107 million tokens.

IV. APPROACH

We formulate the predicting the crime label of a court decision as a single-label multi-class classification problem. Since this is one of the first studies on this domain for Turkish, we want to establish a baseline for Turkish legal document classification. We use traditional machine learning algorithms and deep learning architectures exploiting static and contextual word embeddings throughout the experiments. We use Multinomial Naïve Bayes (NB), Logistic Regression (LR), and Support Vector Machine (SVM) with Radial Basis Function (RBF) Kernel. We run these methods using Bag of Words (BoW) representations, with binary and term frequency-inverse document frequency (TF-IDF) term weighting schemes [2]. Binary term weighting can be as efficient on certain cases [18]. BoW model cannot capture the position of a word in the text or the semantics of the words in detail.

For deep learning methods, we experiment with bi-directional Long Short Term Memory (biLSTM) classifier with fastText [19] as the word representation method. Word embeddings are considered as the vector representation of the words or tokens. Static word embeddings such as fastText are powerful tools that can capture the semantics by representing the relation between different words (words with similar meanings position closer in the vector space). Since static word embeddings learn a global vector for a word, they cannot distinguish between different meanings of the same word. They do not take into account the context each word is used in. FastText is a static word embedding method by Facebook AI Research (FAIR) lab. It maps each word into a multi-dimensional Euclidean space by creating a relatively short (e.g. 100 or 300) and dense vector. FAIR team released fastText pre-trained models for 157 languages. We use Turkish fastText vectors in the embedding layer in our network, fed to a single biLSTM layer [20] and followed by a single feed-forward classification layer. BiLSTM architecture consists of two long short-term Memory (LSTM) layers. The former of these layers pass the input forward while the latter passes backward. LSTM itself is a recurrent neural network (RNN) that has feedback connections so that it can represent sequential data [21]. Feeding forward and backward makes biLSTM a powerful tool because it can understand the relation between the words i.e. it can understand which words are followed or preceded by another word [22].

Transformers are deep learning architectures introduced in 2017 [4] that are designed to handle sequential data like text. They can also be fine-tuned for domain-specific corpus to increase their performance. This method is proven to work in English Legal Document classification [14]. Bidirectional Encoder Representations (BERT) is a transformer-based model that is published in 2018 [23] and has since become the baseline for many NLP tasks since [24]. DistilBERT is a distilled version of BERT, which is 60% faster than BERT but retains 97% of its language learning capacity [25]. It is a cheaper alternative that can achieve similar results. We use these two models.

MDZ Digital Library team from Bavarian State Library has pre-trained Turkish BERT and DistilBERT models on OSCAR corpus [26] and OPUS corpora [27] and made them publicly available. We use their models which are pre-trained on a general common crawl corpus in Turkish. We than fine-tune this general domain model using our corpus that consists of a collection of Turkish legal texts to leverage the power of transfer learning. For fine-tuning we use Masked Language Model (MLM) task. MLM is the task of filling the blanks in a sentence. We think it is a good choice for fine-tuning models for domain-specific data. As suggested in the original paper [23], we fine-tune for 3 epochs with a learning rate of 5e-5 and batch size of 32.

To compare our transformer models, we use them as the embeddings in the classification task by adding a dense layer. Unlike fastText + biLSTM where we freeze the embedding layer, we do not freeze the transformer layers while training for the classification task. To find the optimal hyperparameters we use Grid Search technique within the following search space: Epoch count = {1,2,3} as suggested in the original BERT paper, and learning rates = {1e-4, 5e-5, 7.5e-5} to see how the models react to different learning rates.

We use accuracy, precision, recall, and F-measure (F1) as our evaluation metrics. Accuracy alone is not a good indicator of performance especially if the class distribution is skewed. We also calculate precision, recall, and f1 score for each

TABLE III
MACRO AVERAGE RESULTS FOR TF-IDF AND BINARY WEIGHTING.

Method	Binary	Accuracy	Precision	Recall	F1 Score
Naive Bayes	true	0.89	0.83	0.62	0.61
	false	0.88	0.83	0.59	0.58
Logistic Regression	true	0.95	0.86	0.86	0.85
	false	0.95	0.86	0.86	0.86
SVM	true	0.95	0.87	0.87	0.86
	false	0.95	0.87	0.87	0.86

class and their macro average. Macro averaging mitigates the dominance of large classes in the results compared to the micro averaging. Since we have a skewed class distribution that can be seen in figure 1 we report macro average values.

V. EXPERIMENT RESULTS AND DISCUSSION

To establish a baseline, we start with traditional machine learning methods, namely, Naive Bayes (NB), Support Vector Machines (SVM), and Logistic Regression (LR) applied on BoW representation with binary weighting. We chose to use Multinomial NB and radial basis function (RBF) kernel for SVM as they are commonly used in text classification [18], [1]. Table III shows that the binary weighting results are comparable or better with TF-IDF.

We build our deep learning model with 3 layers. An embedding layer, a single layer of biLSTM, and a dense layer for classification. We use the Turkish pre-trained fastText model for embeddings. During the training phase, we freeze the embedding layer. We use softmax as the activation function. The Model has following parameters: Loss function: "categorical cross-entropy", optimizer: "Adam", learning rate: 1e-3. With experimentation, we determine the optimal results are achieved in 32 epochs with a batch size of 256.

Transformer models with added softmax layer for classification are implemented using HuggingFace library [28]. Using grid search for hyper-parameter optimization, we set the optimal parameters for BERT, DistilBERT, and our fine-tuned BERT model. BERT and DistilBERT models achieve their best results with 3 epochs of training and 5e-5 as the learning rate. Fine-tuned BERT achieves its best score in 1 epoch only with same learning rate. For all of our models, we use a batch size of four and weight decay of 1e-3. Figure 2 show the training losses of our models.

We compare the models according to their macro average F1 score. As we see in table IV, MNB is the worst-performing model. However, both SVM and LR achieve better scores than transformer models without fine-tuning. BiLSTM model achieves the best recall among all models with a better F1 score than BERT and DistilBERT. Comparing the resource requirements, SVM, LR, and biLSTM models are easier to train and achieve better results than general domain transformer models. Lastly, our fine-tuned BERT model achieves the best scores and shows that fine-tuning transformer models with domain-specific corpus can increase the performance of the model. It is also important to note that fine-tuned BERT

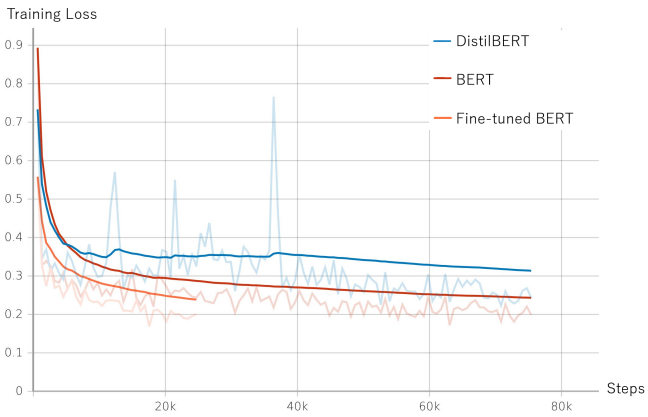

Fig. 2. Graph of training loss versus training steps for transformer models.

TABLE IV
PERFORMANCE COMPARISONS OF CRIME CLASSIFICATION MODELS ON TURKISH LAW TEXT

Method	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.89	0.83	0.62	0.61
Logistic Regression	**0.95**	0.86	0.86	0.86
SVM	**0.95**	**0.87**	0.87	0.86
FastText + biLSTM	**0.95**	**0.87**	0.86	0.86
DistilBERT	0.94	0.84	**0.88**	0.85
BERT	0.94	0.85	0.85	0.84
Fine-tuned BERT	**0.95**	0.86	**0.88**	**0.87**

achieves this result with only 1 epoch of training while other transformer models are trained for 3 epochs.

VI. CONCLUSION AND FUTURE WORK

There are only a few text classification studies in the law domain in Turkish. One of the reasons for this can be the scarcity of available supervised or unsupervised datasets, especially in the Turkish law domain. Turkish can be considered a low-resource language. In this study, we compile sizeable supervised and unsupervised Turkish law datasets from publicly available sources. Another reason for the lack of NLP studies in this domain may due to the complexity of the legal documents. The length, complexity of documents, and use of extensive technical terms are some of the reasons that separate this domain from others. Similar to the medical domain, understanding these documents requires extensive specialization in the domain.

We observed the performance of classification algorithms ranging from traditional classifiers to much more complicated deep learning and transformer-based models. These models use, again, a range of text representations of different complexity from BoW to word embeddings (fastText) and contextual embeddings by transformers. Interestingly, the majority of the models could be able to achieve high accuracy values ranging from 89% to 95%. However, we have a skewed class distribution in our dataset and if we take a closer look at the performance of the classifiers using macro averaged F1 scores we see a wider range (between 61% to 87%), showing

the value of using more complicated text representations and deep learning algorithms such as transformers. Furthermore, we analyze the effect of domain adaption in these more advanced deep learning models, more specifically transformer-based models such as the popular BERT [23]. We first use a general domain pre-trained neural language models for Turkish. These are trained with quite large but general corpora such as Wikipedia articles or web pages. We fine-tune the general model using the domain-specific corpus. As expected we observe performance improvements. We report promising results for the advancement of NLP studies in the Turkish law domain.

In future work, we intend to increase the breadth and depth of our classification experiments, text representation experiments, fine-tuning, and pre-training of neural language models for learning the domain-specific characteristics of the law domain. This can be quite useful not only for classification but also for other downstream tasks. Studies show that data augmentation techniques can improve the accuracy of deep neural network models [29]. We plan to improve the dataset using other sources and data augmentation. We also plan to distill and share our datasets publicly in order to contribute to the advancement of Turkish NLP studies in this domain.

REFERENCES

[1] B. Altınel and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances," *Information Processing Management*, vol. 54, no. 6, pp. 1129–1153, 2018.

[2] D. Torunoğlu, E. Çakirman, M. C. Ganiz, S. Akyokuş, and M. Z. Gürbüz, "Analysis of preprocessing methods on classification of turkish texts," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 112–117, IEEE, 2011.

[3] G. Bayrak, M. Ş. Toprak, M. C. Ganiz, H. Kodaz, and U. Koç, "Deep learning-based brain hemorrhage detection in ct reports," *Studies in health technology and informatics*, vol. 294, pp. 866–867, 2022.

[4] 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.

[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[6] T. Orosz, R. Vági, G. M. Csányi, D. Nagy, I. Üveges, J. P. Vadász, and A. Megyeri, "Evaluating human versus machine learning performance in a legaltech problem," *Applied Sciences*, vol. 12, no. 1, p. 297, 2021.

[7] O.-M. Sulea, M. Zampieri, S. Malmasi, M. Vela, L. P. Dinu, and J. Van Genabith, "Exploring the use of text classification in the legal domain," *arXiv preprint arXiv:1710.09306*, 2017.

[8] J. S. T. Howe, L. H. Khang, and I. E. Chai, "Legal area classification: A comparative study of text classifiers on singapore supreme court judgments," *arXiv preprint arXiv:1904.06470*, 2019.

[9] H. Chen, L. Wu, J. Chen, W. Lu, and J. Ding, "A comparative study of automated legal text classification using random forests and deep learning," *Information Processing & Management*, vol. 59, no. 2, p. 102798, 2022.

[10] E. Mumcuoğlu, C. E. Öztürk, H. M. Ozaktas, and A. Koç, "Natural language processing in law: Prediction of outcomes in the higher courts of turkey," *Information Processing & Management*, vol. 58, no. 5, p. 102684, 2021.

[11] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pp. 115–124, 2017.

[12] F. Gargiulo, S. Silvestri, M. Ciampi, and G. De Pietro, "Deep neural network for hierarchical extreme multi-label text classification," *Applied Soft Computing*, vol. 79, pp. 125–138, 2019.

[13] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Extreme multi-label legal text classification: A case study in eu legislation," *arXiv preprint arXiv:1905.10892*, 2019.

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

[15] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 159–168, 2021.

[16] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. S. Dhillon, "Taming pretrained transformers for extreme multi-label text classification," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3163–3171, 2020.

[17] W.-C. Chang, H.-F. Yu, K. Zhong, Y. Yang, and I. Dhillon, "X-bert: extreme multi-label text classification with using bidirectional encoder representations from transformers," *arXiv preprint arXiv:1905.02331*, 2019.

[18] M. Poyraz, Z. H. Kilimci, and M. C. Ganiz, "Higher-order smoothing: a novel semantic smoothing method for text classification," *Journal of Computer Science and Technology*, vol. 29, no. 3, pp. 376–391, 2014.

[19] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[20] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pp. 207–212, 2016.

[21] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[22] A. Pogiatzis and G. Samakovitis, "Using bilstm networks for context-aware deep sensitivity labelling on conversational data," *Applied Sciences*, vol. 10, no. 24, p. 8924, 2020.

[23] 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.

[24] M. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[26] P. J. O. Suárez, B. Sagot, and L. Romary, "Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures," in *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Leibniz-Institut für Deutsche Sprache, 2019.

[27] M. Aulamo, J. Tiedemann, *et al.*, "The opus resource repository: An open package for creating parallel corpora and machine translation services," in *22nd Nordic Conference on Computational Linguistics (NoDaLiDa) Proceedings of the Conference*, Linköping University Electronic Press, 2019.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[29] A. M. Issifu and M. C. Ganiz, "A simple data augmentation method to improve the performance of named entity recognition models in medical domain," in *2021 6th International Conference on Computer Science and Engineering (UBMK)*, pp. 763–768, IEEE, 2021.