Danish Asylum Adjudication using Deep Neural Networks and Natural Language Processing

Muddamsetty, Satya Mahesh; Jahromi, Mohammad Naser Sabet; Moeslund, Thomas B.; Gammeltoft-Hansen, Thomas

Published in:
Proceedings of the Seventeenth International Workshop on Juris-Informatics 2023 (JURISIN 2023)

Publication date:
2023

Document Version
Accepted author manuscript, peer reviewed version

Danish Asylum Adjudication using Deep Neural Networks and Natural Language Processing

Satya M. Muddamsetty[1 [0000−0003−0935−4609]], Mohammad N. S. Jahromi[1 [0000−0002−6332−7567]], Thomas B. Moeslund[1 [0000−0001−7584−5209]], and Thomas Gammeltoft-Hansen[2 [0000−0003−1518−137X]]

[1] Visual Analysis and Perception Laboratory (VAP), Aalborg University, Rendsburggade 14, 9000 Aalborg
[2] Faculty of Law, Center of Excellence for Global Mobility Law, University of Copenhagen, Karen Blixens Plads 16 2300 København S, Denmark
{smmu,mosa,tbm}@create.aau.dk, tgh@jur.ku.dk

Abstract. The Danish asylum adjudication procedure is a two-tiered system, with the Immigration Service making initial determinations and the Danish Refugee Appeals Board (RAB) automatically appealing cases that are rejected. This study aims to employ a deep neural network(DNN)-based Natural Language Processing (NLP) pipeline to predict asylum decision-making outcomes using a dataset of over 15,515 Danish asylum decisions provided by the Danish Refugee Appeals Board (RAB) between January 1995 and January 2021. This research seeks to improve the performance and effectiveness of decision-making in asylum cases by addressing key challenges, such as modeling the asylum decision-making problem using NLP-based DNNs and dealing with class imbalance issues. Our preliminary results indicate that DNN-based NLP predictive models are capable of learning meaningful representations of asylum cases with high precision and recall, particularly when class weights are considered than the baseline DNN model.

Keywords: Danish Asylum adjudication · Deep Neural Networks (DNN) · Natural Language Processing (NLP) · CNN · Predictive model

1 Introduction

Europe has been at the forefront of efforts to harmonize national asylum law [10]. In particular, the Common European Asylum System and EU directives were established to standardize and streamline the procedural assessment and legal criteria for asylum claims within the European Union. Despite these efforts, however, legal outcomes for similarly situated asylum seekers continue to differ widely across individual countries. A key problem for this type of legal decision-making is that outcomes often depend on how authorities assess the credibility of asylum claims, and this assessment is prone to subjectivity and bias [3][27]. Previous studies have thus shown that the applicant's level of education, gender, and religion, as well as the decision-maker's gender, experience,

and background, can all play a role impacting decision-making [16][30][31]. In addition, several European countries have begun experimenting with AI-driven solutions to asylum decision-making [25][26][28]. While artificial intelligence (AI) models cannot yet replace the qualitative components of legal judgments, they can perform tasks such as categorization [4], identity verification [24], and data entry, and in some situations even take the role of human decision-makers in the asylum application process. For authorities, the goal may both be to reduce discrepancies, but also bolster bureaucratic efficiency in an area of ever-shifting case loads. Yet, the lack of good training data have equally generated concerns that AI models may replicate and exacerbate pre-existing human bias [6][17].
As decision-making models that rely on artificial intelligence (AI) and machine learning (ML) are becoming increasingly important in the legal domain [9][17], it's important to analyze them thoroughly. Because of their superior performance in dealing with complex data patterns and extracting meaningful insights [29], advanced ML methods like deep neural networks (DNN) are becoming increasingly plausible as large-scale datasets become available within the asylum domain. It can be extremely helpful to gain a thorough comprehension of DNN models for such cases through various evaluation metrics in order to aid in the identification of specific data patterns, improve the post-interpretation of results, and ultimately determine the potential of these tools as viable solutions for legal decision-making. If these AI-based systems are effective, they can be important resources for the legal profession, aiding in the decision-making process and improving the quality of legal decisions. This in-depth analysis allows domain experts to make informed decisions on whether to adopt and further develop AI and ML technologies, such as DNNs, in the legal domain or explore alternative approaches.

DNN has demonstrated unprecedented superior performance in a number of disciplines, including computer vision [15] and natural language processing (NLP) [20] for a number of difficult problems in these fields. However, applying DNN-NLP to the asylum domain is still in the preliminary stage. The primary objective of this study is to examine the applicability of deep learning models in the legal domain, particularly in the context of asylum decision-making. This study aims to provide valuable insight into the performance of DNN by concentrating on NLP pipelines built with DNN. Specifically, this study seeks to address the following important research questions:

1. How can DNN-NLP-based can be used to model the asylum decision-making problem?
2. What are effective methods for addressing class imbalance issues, and what is the workable strategy for developing a stable NLP-based DNN predictive model?
3. How can we judge the effectiveness of the predictive model when there is a class imbalance?

We hope to contribute to a better understanding of the application and effectiveness of deep learning techniques in the legal domain, particularly in asylum decision-making, by addressing these research questions and investigating both

general methods and a specific optimal approach. The rest of the paper is structured as follows. We review some of the recent literature where AI has been employed in the legal sphere in section 2. In Section 3 we discuss our Danish asylum dataset. Section 4 describes the proposed DNN-NLP-based asylum decision predictive modeling. Section 5 presents the experimental evaluation. Finally, Section 6 provides concluding remarks and future work.

2 Related Work

Traditional machine-learning algorithms have played an important role in the early phases of research in the field of judicial judgment prediction. In several areas of law, statistical methods such as support vector machines, decision trees, and naive Bayes classifiers have been utilized to assess and predict outcomes [6]. The study presented in [2] used support vector machines to predict rulings by the European Court of Human Rights, whereas the study in [18] relied on more conventional machine learning techniques such as decision trees and random forests to predict decisions by the United States Supreme Court. Logistic regression, naïve Bayes, and support vector machines were only some of the machine learning techniques that researchers relied on in predicting decisions in the legal domain[2] [18] [32].

In recent years, deep learning methods, especially when combined with natural language processing (NLP) methods, have gained popularity in the area of judicial decision prediction. Deep neural networks (DNN) such as convolutional neural networks (CNN), and recurrent neural networks (RNN) are just a few of the state-of-the-art techniques that have shown effectiveness in tackling complex data patterns and obtaining useful insights from legal documents. The authors in [7] employed NLP-based techniques to predict the outcome of legal decisions in English using a combination of CNNs and LSTMs. The study in [23] compares classical machine learning approaches to deep learning techniques like CNNs combined with NLP for predicting European Court of Human Rights rulings. In criminal case prediction, the authors in [22] employed deep learning techniques, such as hierarchical attention networks, in combination with NLP for predicting charges in criminal cases. This growing interest in applying deep learning methods with NLP to legal judgment prediction demonstrates their potential in offering improved performance and better capturing the complexities of legal language and case-specific information. However, the use of deep learning technologies in asylum situations is still in its early stages. As a result, the current work investigates the application of deep learning approaches for forecasting asylum decision-making outcomes in order to better understand their potential in this domain.

3 Danish Asylum dataset

The asylum decision-making procedure varies from country to country. For instance, the asylum process in Denmark is two-tiered. First-instance decisions

are made by the Immigration Service. Decisions rejected at the first instance are automatically appealed to the Refugee Appeals Board (RAB), which is a quasi-judicial body with full legal competence to assess questions of both fact and law. Denmark moreover maintains a legal opt-out to EU law, which means that it is only partly bound by common EU asylum rules [11]. In this work, we employed Danish asylum decisions provided by the Refugee Appeals Board (RAB), Denmark. The dataset used in this work consists of approximately 15,515 Danish asylum decision summaries, provided by the Danish Refugee Appeals Board (RAB), spanning from January 1995 to January 2021. The case file provides information about the applicant. The information includes country of origin, gender, religion, year of applicant entry to Denmark, ethnicity, detected divergence, previous asylum, and torture cases. It also provides information where relevant to the claim about the involvement in political parties, marital status, and military service. The case file provides a brief narrative story about the applicant's reason for seeking asylum status. Finally, the case files are closed with the legal reasoning and outcome of the RAB. This information provides the reasoning that supports the decision in the case. Furthermore, the case files also contain the candidate's interview and motivation for seeking asylum, the administrative events and the asylum process that took place since the candidate entered Denmark, and the reasoned decision by the RAB. Most documents, particularly the more recent ones, include additional documents such as the initial asylum application form and/or the interview transcript from the first instance decision-maker, the Immigration Service. As it was obtained through an agreement with the Danish Refugee Council (a Danish NGO), which regularly receives case files from the RAB, our dataset does not contain the totality of decisions by the RAB. However, our dataset is statistically representative, as the yearly recognition rates published by the RAB are consistent with the yearly recognition rates calculated on our dataset. The collected asylum decisions are presented as Word or PDF documents, and some of them are printed documents that are scanned to create PDF files. To transform these texts into a machine-readable format, optical character recognition (OCR) is used. In the aforementioned dataset, the case file consists of the decision of the applicant and legal reasoning provided by decision-makers for the reason to be granted, rejected, or sent back for further evaluation. We automatically removed the legal reasoning text with the headings "Flygtningenævnet udtaler" from the case file using a regular expression. The documents that do have the legal reasoning headings are removed. We finally have 14,987 asylum case files. Table 1 summarizes the number of case files for each individual class.

Data samples	**Granted**	**Rejected**	**SentBack**
No of Samples	2,373	12,543	71
Percentage	15%	83%	0.004%

Table 1: Total number of cases for each classes.

The dataset offers an abundance of details on Danish asylum adjudication, comprising specifics of Danish asylum legislation and practice, information on each applicant's administrative procedure, and additional data including interviews and outside evidence that aren't often available to researchers. These salient features made this dataset highly suitable for modeling asylum adjudication using NLP-based deep learning methods.

4 Methodology

The topic of asylum adjudication prediction was recently characterized by Chen et al. [8], as a binary classification challenge. In their study, 441 different judges presided over 492,903 asylum proceedings held in 336 different hearing venues throughout a 32-year span from 1981 to 2013. Similar to this, Katsikouli et al. [17] assessed the predictability of the case results based on a variety of application data, including the applicant's nationality, gender, and religion. However, all of these algorithms are modeled using manually derived features, making them unable to exploit the contextual information of case files. To address these issues, we introduce an NLP-Deep Neural Network-based asylum adjudication. Data pre-processing and algorithm development are the two primary stages of the pipeline.

4.1 Pre-processing the text

We recognize that our data contains text-based asylum case files. NLP-based AI systems aims to process and predict the outcome of an asylum judgment from textual content. To train the models, large-scale datasets are being applied to asylum-decision making. Text preprocessing involves preparing and cleaning the text into a form that is predictable and analyzable for a specific task. Preprocessing the text makes the data usable and draws attention to textual elements that an algorithm can make use of. [13]. Being initial step in any pipeline of Natural Language Processing (NLP), preprocessing and refining text can have significant impact on overall accuracy of the trained model. There are numerous steps that need to be taken, including removing punctuation, deleting whitespaces, lowering case, removing stop words, lemmatization, stemming, and tokenization.

Removing punctuation and whitespace, such as commas and full stops from the comments, eliminates redundant information from text file and maintain the size of training set lower. Stop word removal (common words are removed from the text so unique words that offer the most information about the text remain). Lemmatization and stemming are then performed to normalize the text and prepare it for further processing. Normalization is the process of converting the token into its basic form (morpheme). Inflection is removed from the token to get the base form of the word. This process helps reduce the number of unique tokens and redundancy in the data. It reduces the data's dimensionality and removes the variation of a word from the text. Finally, tokenization is applied to break the text into individual tokens to feed the classifier.

4.2 DNN-NLP-based predictive model

The problem of predicting the decisions of the Danish asylum adjudications is defined as a classification task. Our goal is to predict if the information provided in a particular case is credible enough to grant the refugee status to the applicant, based on the information provided during the interview, or if there is a violation in relation to specific Articles to reject the refugee status. We model asylum adjudication using deep neural networks (DNNs) [12]. DNNs, which are based on artificial neurons and coupled in many layers to form a network, are used to predict whether asylum status should be granted to refugees.

In recent years, Convolution Neural Networks (CNN) have demonstrated ground-breaking performance in a number of NLP tasks, including text categorization [14, 19]. The embedding layer comes first in the CNN model for natural language processing, and followed by convolutional layers. In the vector space learnt by the embedding layer, words that are similar or those appearing in related situations are closer together than words that appear in unrelated settings. The embedding layer enables us to convert each word into a fixed-length vector of a defined size. The resulting vector is dense, containing real values instead of just 0's and 1's. Word vectors' fixed length and decreased dimensions enable us to express words more effectively [21]. The embedding layer has three parameters: vocabulary size, vector length for each word, and maximum sequence length.

We propose a simple CNN-NLP model for Danish asylum adjudication. Our CNN model consists of a total of five layers. Among these, we have one embedding layer, one convolution layer, and one dense layer. A ReLU non-linearity activation function is used for every convolution layer. A global average pooling layer (GAP) is added after the high-level feature extraction convolution layer, followed by a fully connected (FC) dense layer. The input layer size for this network is equal to the maximum sequence length from the training data. The number of filters used in our network is 128 and 64, respectively. The convolution kernel size used in the model is 4. A global average pooling is applied to the last convolution layer and the training procedures are described in the following section 4.3.

4.3 Training procedure of a CNN-NLP-based asylum model

The proposed methodology is trained on our RAB dataset described in Section 3. In order to train the model, the dataset is split into 80% training, 10% validation, and 10% testing subsets of a total of 14987 cases with three classes of asylum decisions. Data pre-processing steps such as removing punctuation, white spaces, unnecessary symbols, and lemmatizing text are applied to the training, validation, and test samples. Tokenization is performed on the text after noise removal and normalization. We didn't consider removing stop-words from our data. In general, stop-words can be removed, as they are considered noise that can reduce vocabulary size. However, in our case, we didn't consider removing stopwords, as they can provide some contextual information that can affect the

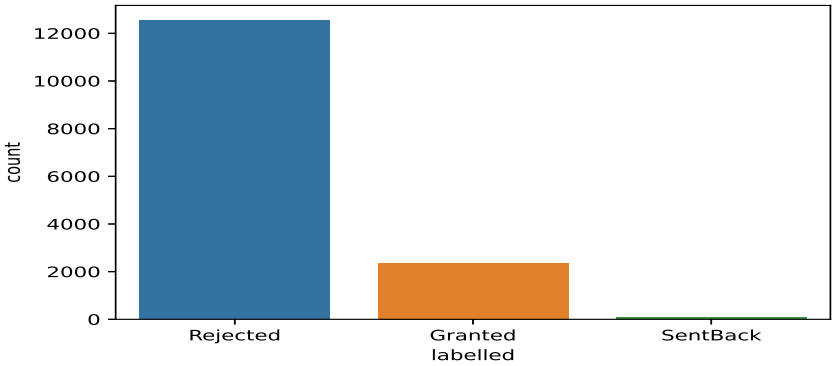


Fig. 1: Data distribution of all classes.

performance of our model in predicting the outcome of cases. The embedding layer is trained with input parameters from the full vocabulary of the trained dataset and 300 output feature vector dimensions. The CNN-NLP-based model is trained over 30 epochs with a batch size of 8 to avoid memory errors while training. We use cross entropy as a loss function and *Adam* as an optimizer, with a learning rate of 10^{-2} and momentum of 0.9. We trained our model until convergence, using an early stopping strategy that monitored the validation loss. This is a good strategy to prevent over-fitting and to save some computational time. We also added dropouts after every hidden layer to further reduce the chances of over-fitting. The framework is implemented on TensorFlow Keras with $11GB$ of GPU memory on an Nvidia, RTX 2080Ti. The predictive model is evaluated, and the results are presented in the experimental section 5.

5 Experiments & Results

In this section, we evaluate the performance of the CNN-NLP-based asylum predictive model. The proposed model is evaluated on our novel RAB dataset described in Section 3, which has three classes 'Granted',' Rejected', and 'Sentback'. We first analyzed the class distribution in our aforementioned RAB dataset. Fig 1 illustrates the distribution of cases per class and Table 1 in P.4, summarizes the number of samples for each individual class. We can clearly observe that there is a class imbalance problem in our dataset. The 'Rejected' class has higher samples followed by the 'Granted' and 'SentBack' classes with 83%, 15%, and 0.004% of total samples, respectively. In general, machine learning algorithms are not immune to unbalanced classes and generate models that are biased and less accurate. For instance, deep neural networks are trained using back-propagation, which treats each class equally when calculating the loss. If the data is not balanced, that makes the model biased for one class over another.

There are several approaches to tackle the imbalanced data problems in the literature [1]. We choose three different strategies for handling the class imbalance problem in a text dataset that contains text information. First, introducing the class weights. Class weights alter the loss function directly by penalizing the classes with varying weights. The minority class may be assigned more weight, while the majority class may be given less weight. In this manner, balance between the various classes can be achieved [5]. Second undersampling for the Majority Classes. Undersampling for the majority Class, we essentially remove a certain number of samples associated with the Majority classes for balancing the classes. Third, oversampling for minority classes, on the other hand, entails the repetition of samples associated with the minority classes. We choose three different measures recall, precision, and F1-score to evaluate the performance of our CNN-NLP-based asylum predictive model. The recall (or true positive rate) gives the proportion of the actual class correctly predicted. The precision gives the proportion of positively predicted cases that are the correct class. The F1-score gives the harmonic mean of precision and recall. We didn't consider computing the overall accuracy of the model as it can mislead when there is a class imbalance issue in the dataset.

We conducted the class imbalance experiments using the above-mentioned strategies. First, we removed the minor class, which has 71 cases which are 0.004% of the total samples, and trained the CNN-NLP model as a binary classification problem described in Section 4 on RAB dataset. Since this is the first work on the RAB dataset where a truly DNN-NLP-based model is suggested, we cannot directly compare it with the work of others. Therefore, we chose to compare the results with the proposed CNN baseline model together with different imbalance strategies. We represent the CNN model without data balancing approach as "CNN(Baseline)" in the tables. Table 2 summarizes the performance of the CNN model after applying the aforementioned class imbalanced strategies.

Binary Class Predictive Model			
Models	**Precision**	**Recall**	**F1-Score**
CNN (Baseline)	0.82183	0.84718	0.82477
CNN+ class weights	**0.82948**	**0.85188**	**0.83184**
CNN+ oversampling	0.80880	0.83847	0.81451
CNN+ downsampling	0.81126	0.75201	0.77396

Table 2: Precision and Recall for the Binary Class asylum predictions

From the tables, we can clearly observe the dealing class imbalance by introducing the class weights can significantly improve the performance of the CNN model, resulting in high precision and recall, followed by downsampling. Table 3, 4 summarizes the result on each class obtained for different imablancing strategies.

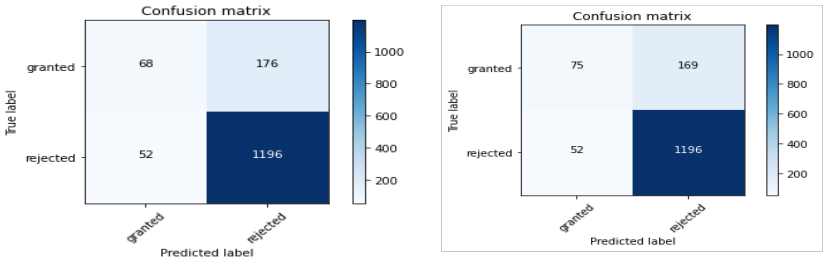
Precision for each Class	**Granted**	**Rejected**
CNN (Baseline)	0.57	0.87
CNN + class weights	**0.59**	**0.88**
CNN + oversampling	0.51	0.87
CNN + downsampling	0.34	**0.90**

Table 3: Precision table for all the multi-class experiments:

Recall for each Class	**Granted**	**Rejected**
CNN (Baseline)	0.28	0.96
CNN+ Class weights	**0.31**	**0.96**
CNN + Oversampling	0.25	0.95
CNN + Downsamping	**0.57**	0.79

Table 4: Recall for each individual class for different imbalance strategies

Analyzing the precision-recall for each individual class, we can clearly observe that the class weight strategy and downsampling strategy both demonstrate equally better precision and recall for both classes. However, downsampling strategy can balance the classes by removing the samples from the majority which might prevent the model from learning crucial information that could have been gained from the removed samples.



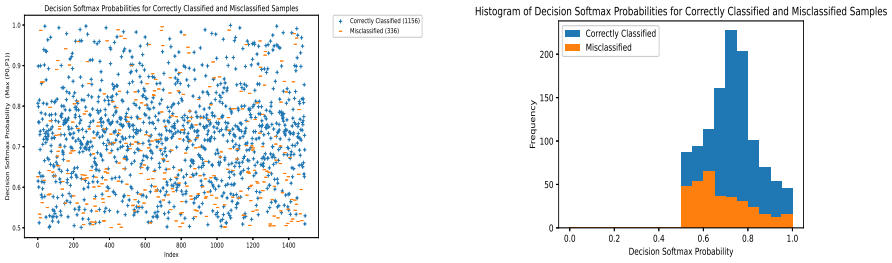
(a) (b)

Fig. 2: Confusion matrix with .(a) class weights, (b) without class weights.

The confusion matrices of both CNN models, with and without class weights, are illustrated in Fig 2. From the figure, we can clearly see the number of samples from both classes improves in correctly predicting the cases. From the tables, we conclude that addressing the class imbalance problem using class weights can improve performance in terms of precision and recall. Although either of the two strategies balances out the dataset, they do not directly tackle the issues caused by class imbalance.

We further analyzed the binary classification results of the CNN model, which performed better using the class weighting strategy. We plotted the predicted

results in terms of corrected and misclassified samples from both classes. Fig 3 shows the scatter plot and histogram distribution of correctly classified versus misclassified sample predictions of the CNN model on the test dataset.



(a) (b)

Fig. 3: Misclassified vs Classified plot. (a) Soft-max values plot, (b) Histogram distribution.

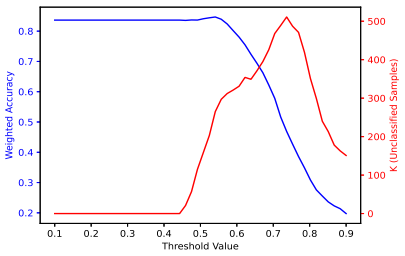
This plot illustrates the distribution of correctly classified and misclassified samples (error) across the range of decision soft-max probabilities. From this graph, we can observe that the majority of correct classes tend to have a soft-max probability greater than 0.7 (denser samples between 0.7 and 0.8). We also can observe an overlap between the correctly classified and misclassified samples of around 0.5 to 0.6 which suggests the model struggles to distinguish between classes in some cases, resulting in errors. This plot clearly explains that selecting the threshold influences the performance of the predictive model to take decisions. Hence, this suggests that traditional accuracy calculation lacks a mechanism for avoiding classifying uncertain samples, which can result in inaccurate classifications and poor performance. In addition, it implies that with a class imbalance dataset, a classifier may obtain high overall accuracy simply by predicting the majority class for all samples. In real-world situations, not all the samples need to be classified, particularly if the classifier is unsure of their class assignment. Therefore, by adjusting the threshold, when the classifier is unsure, we can choose to emphasize accuracy on classified data with higher confidence while leaving a certain number of samples unclassified, resulting in less bias in overall accuracy. This can lead to a more realistic performance calculation by further refining the accuracy of the classifier that leaves out uncertain samples and offers a great tool to control the trade-off between accuracy and uncertainty in the model.

To motivate this point, we choose to define a small epsilon region around possible range of thresholds, within which the model is undecided, as a unclassified region. Next, we compute the weighted accuracy as portion of correcty classified sample over total number of classified samples (excluding the unclassified

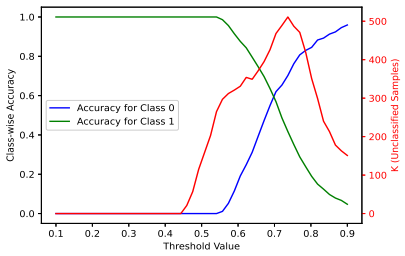
samples). Hence, the weighted accuracy of the model is defined in eq.

$$ACC = \frac{CP}{N - K} \tag{1}$$

Where CP is the total number of correct predictions for both classes "granted" and "rejected." N is the total number of input samples, and K is the number of unclassified samples (i.e., how many samples are left out after choosing the threshold value). We plotted the weighted accuracy-unclassified samples K for the two classes against the different threshold values τ ranging from 0 to 1. Weighted accuracy class-wise means computing accuracy for each class individually while taking into account the proportion of each class in the dataset. This can aid in understanding the classifier's performance in each class separately, especially when there is a class imbalance. Fig 4 (a) illustrates the weighted accuracy vs threshold values for the correct predictions. We similarly plotted each individual class in Figure 4 (b).



(a) Weighted accuracy (overall) and K against the threshold value.



(b) Class-wise accuracies for the two classes and K against the threshold value.

Fig. 4: Trade-off between weighted accuracy and class-wise accuracies for different threshold values.

From the figure, we can observe that for weighted accuracy (4. a) threshold values between 0.1 and 0.55, the weighted accuracy remains pretty stable (about 0.83). This indicates that the classifier's performance is constant within this range of threshold values. However, as the threshold value exceeds 0.55, the weighted accuracy rapidly decreases, reaching 0.2 at a threshold value of 0.9. This means that as the threshold becomes higher, the classifier's performance on classified data drops, and the number of unclassified samples mainly increases within this range. It's possible that the classifier is unsure about the class assignments for data in this range and will struggle to appropriately classify them. The class-wise accuracy case (4. b) plot shows that the classifier is very accurate for class 1 (i.e., rejected class) samples within a narrow range of threshold

values (0.1–0.55) but struggles to classify class 0 (i.e., granted class) samples. The classifier becomes more accurate in classifying class 0 (i.e., granted class) samples but less accurate in class 1 (i.e., rejected class) samples as the threshold value increases. This trend could imply a performance imbalance in the classifier between the two classes.

In practice, a threshold value should be chosen that strikes a balance between the required classification performance and the allowed number of unclassified samples while taking into account the specific needs of the target application.

6 Concluding Remarks & Future Work

In this paper, we proposed deep neural network (DNN)-NLP-based methods for modeling the Danish asylum adjudication based on textual information from asylum cases. In particular, we used the CNN model that is trained on Danish Asylum Dataset comprising approximately 15515 Danish asylum decisions provided by the Danish Refugee Appeals Board (RAB). Three distinct approaches were investigated in our research to address the class imbalance problem. We also analyzed the performance of CNN with the influence of choosing the threshold. Experimental results show that the performance of the CNN model with class weights is significantly better than the baseline CNN model when there is a class imbalance problem. Furthermore, choosing the optimal threshold for the model is crucial to lowering the misclassification rate when the dataset has a class imbalance problem. Overall, our experimental findings in this work points to the existence of further key issues that are yet to be addressed in this application context. This is mainly due to the limitation of accessing large target domain datasets to train with deep learning models that require millions of parameters. This will also limit the model's capacity to create domain-specific pre-trained word embeddings, which act as a pillar for training the model to deliver improved accuracy. However, we believe that by demonstrating the great potential of using deep learning in the legal domain, specifically in their application to decision-making or similar prediction cases, we can encourage further application and innovation of these models in the legal domain that provides a great value to the research community. To further improve performance, as future work, we intend to collect large data multilingual asylum dataset with focus on training with pre-trained word embeddings. In addition, we aim to investigate interpretibility of the predictive models using explainable AI (XAI) framework, which can help to gain deeper insight of their decision-making process.

7 Acknlowledgements

This work is part of the 'Explainable Artificial Intelligence and Fairness in Asylum Law (XAIfair)' interdisciplinary project, funded by The Villum fonden, Denmark.

References

1. Abd Elrahman, S.M., Abraham, A.: A review of class imbalance problem. Journal of Network and Innovative Computing **1**(2013), 332–340 (2013)
2. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ Computer Science **2**, e93 (2016)
3. Anker, D.E., Müller, M.: Explaining credibility assessment in the asylum procedure. International Journal of Refugee Law **19**(3) (2007)
4. Bayer, M., Kaufhold, M.A., Reuter, C.: A survey on data augmentation for text classification. ACM Computing Surveys **55**(7), 1–39 (2022)
5. Byrd, J., Lipton, Z.: What is the effect of importance weighting in deep learning? In: International conference on machine learning. pp. 872–881. PMLR (2019)
6. Byrne, W.H., Gammeltoft-Hansen, T., Piccolo, S., Møller, N.L.H., Slaats, T., Katsikouli, P., et al.: Data driven futures of international refugee law (2021)
7. Chalkidis, I., Kampas, D.: Deep learning in law: early adaptation and legal word embeddings trained on large corpora **27**, 171–198 (2019). https://doi.org/10.1007/s10506-018-9238-9, https://doi.org/10.1007/s10506-018-9238-9
8. Chen, D.L.: Judicial analytics and the great transformation of American Law **27**, 15–42 (2019). https://doi.org/10.1007/s10506-018-9237-x, https://doi.org/10.1007/s10506-018-9237-x
9. Chen, D.L., Eagel, J.: Can machine learning help predict the outcome of asylum adjudications? In: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law. pp. 237–240 (2017)
10. Fry, J.D.: European asylum law: Race-to-the bottom harmonization. J. Transnat'l L. & Pol'y **15**, 97 (2005)
11. Gammeltoft-Hansen, T., Scott Ford, S.: An Introduction to Danish Immigration Law. SSRN Electronic Journal (234) (2021). https://doi.org/10.2139/ssrn.3769962
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
13. Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., Pambudi, R.A.: An experimental study of text preprocessing techniques for automatic short answer grading in indonesian. In: 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE). pp. 230–234. IEEE (2018)
14. Hassan, A., Mahmood, A.: Convolutional recurrent deep learning model for sentence classification. Ieee Access **6**, 13949–13957 (2018)
15. He, K., Zhang, X., Ren, S., et al.: Sun., j.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Johannesson, L.: In courts we trust: Administrative justice in Swedish migration courts. Ph.D. thesis, Department of Political Science, Stockholm University (2017)
17. Katsikouli, P., Byrne, W.H., Gammeltoft-Hansen, T., Høgenhaug, A.H., Møller, N.H., Nielsen, T.R., Olsen, H.P., Slaats, T.: Machine learning and asylum adjudications: From analysis of variations to outcome predictions. IEEE Access **10**, 130955–130967 (2022)
18. Katz, D.M., Bommarito, M.J., Blackman, J.: Predicting the behavior of the supreme court of the united states: A general approach. SSRN Electronic Journal (2014)
19. Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: A survey. Information **10**(4), 150 (2019)

20. Lee, J., Toutanova, K.: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
21. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 302–308 (2014)
22. Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2727–2736 (2017)
23. Medvedeva, M., Vols, M., Wieling, M.: Using machine learning to predict decisions of the european court of human rights. Artificial Intelligence and Law **28**(2), 237–266 (2020)
24. Mohammed, I.A.: Artificial intelligence for cybersecurity: A systematic mapping of literature. ARTIFICIAL INTELLIGENCE **7**(9) (2020)
25. Molnar, P., Gill, L.: Bots at the gate: A human rights analysis of automated decision-making in canada's immigration and refugee system (2018)
26. Nielsen, T.R.: Confronting asylum decision-making through prototyping sensemaking of data and participation. In: Proceedings of 19th European Conference on Computer-Supported Cooperative Work. European Society for Socially Embedded Technologies (EUSSET) (2021)
27. Noll, G.: Salvation by the grace of state? explaining credibility assessment in the asylum procedure. In: Proof, evidentiary assessment and credibility in asylum procedures, pp. 197–214. Brill Nijhoff (2005)
28. Ozkul, D.: Automating Immigration and Asylum: The Uses of New Technologies in Migration and Asylum Governance in Europe. Oxford University Press (2023)
29. Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M.P., Shyu, M.L., Chen, S.C., Iyengar, S.S.: A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR) **51**(5), 1–36 (2018)
30. Rottman, A.J., Fariss, C.J., Poe, S.C.: The path to asylum in the us and the determinants for who gets in and why. International Migration Review **43**(1), 3–34 (2009)
31. Spirig, J.: Like cases alike or asylum lottery? Inconsistency in judicial decision making at the Swiss Federal Administrative Court. Ph.D. thesis, University of Zurich (2018)
32. Sulea, O.M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. arXiv preprint arXiv:1710.09306 (2017)