On predicting and explaining asylum adjudication

Piccolo, Sebastiano Antonio; Gammeltoft-Hansen, Thomas; Katsikouli, Panagiota; Slaats, Tijs

Published in:
ICAIL: International Conference on Artificial Intelligence and Law

DOI:
10.1145/3594536.3595155

Publication date:
2023

Document version
Early version, also known as pre-print

Document license:
Unspecified

Citation for published version (APA):
MOBILE Working Paper Series, no. 5, 2023

On predicting and explaining asylum adjudication

Sebastiano Antonio Piccolo, Thomas Gammeltoft-Hansen,
Panagiota Katsikouli, Tijs Slaats

This research is funded by the Danish National Research Foundation Grant no. DNRF169 and conducted under the auspices of the Danish National Research Foundation’s Center of Excellence for Global Mobility Law (MOBILE).
Abstract:

Asylum is a legal protection granted by a state to individuals who demonstrate a well-founded fear of persecution or who face real risk of being subjected to torture in their country. However, asylum adjudication often depends on the decision maker’s subjective assessment of the applicant’s credibility. To investigate potential sources of bias in asylum adjudication practices researchers have used statistics and machine learning models, finding significant sources of variation with respect to a number of extra-legal variables. In this paper, we analyse an original dataset of Danish asylum decisions from the Refugee Appeals Board to understand the variables that explain Danish Adjudication. We train a number of classifiers and, while all classifiers agree that candidate credibility is the single most important variable, we find that performance and variable importance change significantly depending on whether data imbalance and temporality are taken into account. We discuss the implications of our findings with respect to the theory and practice of predicting and explaining asylum adjudication.

Author: Sebastiano Antonio Piccolo, Thomas Gammeltoft-Hansen, Panagiota Katsikouli, Tijs Slaats
E-mail: Sebastiano.piccolo@jur.ku.dk

MOBILE – Center of Excellence for Global Mobility Law – focuses on systematically studying the legal infrastructures of human mobility across geographies, social divides, travel patterns and time.

This research is funded by the Danish National Research Foundation Grant no. DNRF169.
On predicting and explaining asylum adjudication

Sebastiano Antonio Piccolo  
sebastiano.piccolo@jur.ku.dk  
Faculty of Law  
University of Copenhagen, Denmark

Panagiota Katsikouli  
pk@di.ku.dk  
Department of Computer Science  
University of Copenhagen, Denmark

Thomas Gammeltoft-Hansen  
tgh@jur.ku.dk  
Faculty of Law  
University of Copenhagen, Denmark

Tijjs Slaats  
slaats@di.ku.dk  
Department of Computer Science  
University of Copenhagen, Denmark

ABSTRACT
Asylum is a legal protection granted by a state to individuals who demonstrate a well-founded fear of persecution or who face real risk of being subjected to torture in their country. However, asylum adjudication often depends on the decision maker’s subjective assessment of the applicant’s credibility. To investigate potential sources of bias in asylum adjudication practices researchers have used statistics and machine learning models, finding significant sources of variation with respect to a number of extra-legal variables. In this paper, we analyse an original dataset of Danish asylum decisions from the Refugee Appeals Board to understand the variables that explain Danish Adjudication. We train a number of classifiers and, while all classifiers agree that candidate credibility is the single most important variable, we find that performance and variable importance change significantly depending on whether data imbalance and temporality are taken into account. We discuss the implications of our findings with respect to the theory and practice of predicting and explaining asylum adjudication.

CCS CONCEPTS
• Information systems → Data analytics; Data mining; • Computing methodologies → Machine learning; Model development and analysis.

KEYWORDS
Asylum adjudication, Explanatory Modelling, Predictive Modelling, Data Imbalance

1 INTRODUCTION
Asylum is a legal protection granted by a state to individuals who demonstrate a well-founded fear of persecution or who face real risk of being subjected to torture in their country. Although international treaties such as the 1951 Refugee Convention or the 1950 European Convention for Human Rights establish fundamental human rights and define the notion of refugee, they do not stipulate how refugee claims should be assessed. As such, national authorities are left with the task to determine if an individual is eligible for asylum. This is not an easy task, as in most cases applicants cannot corroborate their story with objective evidence and their case rests on their testimony [11]. Authorities, therefore, have to take their decisions by assessing the credibility of the applicant’s narrative [11, 22]. This assessment can leave the door open to the subjective perception of the decision makers. In fact, prior research suggests that applicant’s characteristics such as educational background, gender or religion [7, 17, 18, 20], and decision maker’s characteristics such as gender, experience level, and background [3, 4, 6, 7, 20] are potential sources of bias. However, as noted in [12] some applicant’s characteristics might actually constitute legally relevant factors and are not necessarily problematic. Furthermore, as noted in [9] some prior findings pertaining characteristics of the applicant or the decision maker do not necessarily extend beyond the jurisdictions in which they were found.

The use of statistics or machine learning to find potential sources of bias in asylum adjudication is becoming more prominent. However, it is important to avoid some pitfalls that can generate spurious results. In explanatory models, it is important to include all the legally relevant factors that can explain asylum outcome. In our literature review (section 2), we find that most of the studies did not control for the credibility of the applicant’s narrative and for the context/asylum motive. In contrast, truly predictive models that try to anticipate future decisions cannot include all the legally relevant factors as some of them can only be established by the judges as part of their decisions (such as the applicant credibility). As such, as remarked in [23], predictive models are not necessarily explanatory, and vice- versa.

In this paper, using a novel dataset containing the full text of 17300+ Danish asylum decisions from the Refugee Appeals Board (see section 3 for a description of the data), we address limitations of prior research by 1) developing explanatory models that account for both credibility and context/asylum motive, and 2) addressing the practice of developing predictive models by training a number of models in various setting to account for both temporality and data imbalance. We find that credibility is the undisputed most important factor to explain asylum adjudication in Denmark, followed by country of origin and context/asylum motive (see section 4 to read more about our results). At the same time, we find that other factors shown to be relevant to asylum adjudication in other countries do not apply in Denmark; thus pointing to underlying differences in national legal systems. Furthermore, we find that it is crucial to take into account data imbalance in order to train trustworthy predictive models with good performance. Models that ignore data imbalance perform poorly, mainly learning to classify negative cases. Models trained with strategies to correct for data imbalance exhibit good performance, learning to classify both granted and rejected cases. Finally, we show that measuring performance with accuracy and area under the curve (AUC) is misleading in presence of data imbalance. We discuss the implications of our findings in
section 5, and conclude the paper by summarising our main findings in section 6.

2 ASYLUM ADJUDICATION

Asylum adjudication has received a lot of attention by scholars who aim to develop the basis for fair asylum adjudication practices[16, 20]. A stream of literature has focused on understanding what decision makers consider in order to take their decisions. This is typically done through qualitative methods such as interviews and legal analysis of decisions. This stream of literature has unveiled the fundamental role played by credibility assessment in asylum adjudication [11, 22]. This is so because in many cases applicants cannot corroborate their story with objective evidence and therefore their case rests mainly on their personal testimony [11]. This leaves decision makers with the daunting task of comparing the applicant’s testimony with objectively known facts such as country of origin information (COI), experts reports, assessments, and documentary proof [11]. Within the social sciences, a number of studies further suggests that asylum outcomes may be impacted by a range of more interpersonal factors, e.g. the applicant’s educational background, gender or religion [17, 18], as well as cultural stereotypes, political orientation or use of medical evidence by decision makers [1, 15, 24].

A second stream of literature has focused on quantitatively analysing disparities in recognition rates based on relevant case file labels and metadata, with techniques varying from simple tabulation to machine learning. In [20] the authors found asylum recognition rates, both within and between different courts and appeals bodies in the United States, to vary significantly based on different characteristics of both applicants and decision makers (such as gender, experience level, and professional background). Since this seminal work, researchers have expanded and refined the study of disparities in recognition rates. Rehaag [21] reviewed over 23000 applications for judicial review in the Canadian Federal Court from 2005 to 2010, and highlighted differences in recognition rates, through tabulation, with respect to certain characteristics of the decision makers, such as their political affiliation and other demographic factors. Ghezelbash, Dorostkar and Walsh [5] extracted relevant metadata from 6756 Australian asylum decisions and found that recognition rates vary in relation to the gender and the workload of the judge, as well as the presence of legal representation. Here, the authors used tabulation and the \( \chi^2 \) test to assess the statistical significance of their findings. In contrast with some of these prior findings, Byrne et al. [9] analysed a dataset of 15535 decisions from the Danish Refugee Appeals Board and found that the gender of the applicant and chairing judge, the applicant’s marital status, and the age of the applicant are not significantly associated with disparities in recognition rates, suggesting that some of these factors are more country specific. Beyond tabulation, researchers have used regression models in order to control for multiple factors. Colaiacovo [4] analysed 68,000+ refugee claims adjudicated by the Canadian Immigration and Refugee Board between 2006 and 2011, finding that the probability of acceptance is associated with decision makers’ characteristics such as education, gender, and professional experience, controlling for the asylum claimant’s country of origin, gender, and the year and regional office of adjudication. Emeriau [7] analyzed a hand-curated representative sample of 4,141 French asylum decisions and found that 1) a credible narrative is associated with a higher probability of being granted asylum, that 2) Christian applicants are more likely than Muslim applicants to be granted asylum, and that 3) applicants with higher education and higher levels of skills are more likely to be granted asylum. Emeriau also controlled for demographic characteristics of the applicant (age, gender, and marital status), country of origin, year of application, topics extracted from the text of the decisions, and whether the applicant spent more than a year in France before applying.

Predictive approaches based on machine learning techniques have recently been proposed as well. Chen and Eagel [3] trained a random forest algorithm on 137 features including case information, court information, judge information, news trends and weather information to predict the outcome of a decision in the US system. They reported an average accuracy of 80% and suggested that recognition rates vary in relation to judge characteristics and case information, such as the country of origin of the applicant. In [6], the authors built a similar random forest model, again on data coming from the US system, and showed that the outcome of many asylum cases is to a large extent predictable using judges’ past grant rates, and that a fraction of judges exhibits a tendency to either reject or grant asylum regardless of the individual case before them. Katsikouli et al. [12], using a publicly available database of ~8000 decisions from the Danish Refugee Appeals Board, trained a number of models to show that certain features can be used as predictors for asylum claims. They found that religious belief of the applicant, the type of claim, the year of the application, and the ethnicity of the applicant are the top predictors. The authors noted that these features are not necessarily sources of bias, as they may also constitute legally relevant features.

To summarize, empirical studies of asylum adjudication have shown variations of recognition rates with respect to the characteristics of both applicants and decision makers. While some variables such as applicant’s religious belief, ethnicity, and country of origin are often related to the applicant’s claim and the legal assessment of the case, other variables related to legally irrelevant facts, such as the gender or the political affiliation of the decision maker may be indicative of decision making bias. The small number of studies along with conflicting findings suggests that certain variables are more jurisdiction/country specific. In addition, most of the studies we reviewed here did not control for all the legally relevant variables; for instance, only [7] controlled for the credibility of the applicant’s narrative. Failing to control for all the relevant variables increases the risk of facing the Simpson’s paradox[19] or finding spurious correlations.

3 DATA AND METHODS

The dataset employed in this work consists of 17300+ Danish asylum decisions, tried by the Danish Refugee Appeals Board (RAB), spanning from January 1995 until January 2021. As background, the asylum process in Denmark is two-tiered. First instance decisions are made by the Immigration Service. Decisions rejected at first instance are automatically appealed to the RAB, which is a quasi-judicial body that has full legal competence to assess the questions of fact and law. The case files come in the form of PDF or
Word documents and 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 asylum application form and/or the interview transcript from the first instance decision maker, the Immigration Service. Our dataset comes with its challenges. Being 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. Most of the PDF files were obtained by scanning paper documents. Therefore, we had to apply optical character recognition (OCR) to convert the documents in a machine-readable format. As no metadata was available, we had to extract all the features of interest from the text of the case files (see section 3.1 for our feature extraction methods). Finally, the availability of NLP models for the Danish language is still quite limited [13]. Despite the aforementioned challenges, the dataset offers a great amount of information on Danish asylum adjudication, including aspects of Danish asylum law and its practice, details of the administrative process followed by each applicant, and additional data such as interviews and external evidence not normally available to researchers. All this richness of information is not present in the publicly available dataset of decision summaries analysed in [12]. Finally, compared to most studies reviewed in section 2, our dataset contains the full text of the decisions, as well as interview transcripts of the oral hearing with the applicant, so our analysis can go beyond the analysis of metadata. In particular, as anticipated in section 2, we can account for both credibility and context/asylum motives.

Fig. 1 shows the yearly distribution of the decisions in our dataset. The dataset is not complete in terms of representing the totality of decisions by the RAB, but was obtained through an agreement with a Danish NGO, the Danish Refugee Council, . For example, we observe that for the years 1997 and 1999 we have no data, and only two cases for the year 2000. This happens because either our dataset did not contain any file for these years or because the files we received were corrupt and unreadable. This, however, does not negatively affect our analysis, since the missing data is located only in the earlier part of the dataset. In the following section, we present the methods we employed to extract the features of interest.

### 3.1 Feature Extraction

Guided by our literature review (section 2), we extracted from the text of the decisions 17 features related to country of origin, applicant personal characteristics, characteristics of the case, chairing judge, credibility, and context/asylum motive. The features extracted are presented in Table 1.

We used regular expressions in order to extract the year of the case (Y), the entry year (EY), and the decision year (DY). The file size (FS), in bytes, is computed using python’s built-in function `getsize`. To extract the country of origin (C), we constructed regular expressions to extract the parts of the text where the nationality or ethnicity of the applicant(s) is mentioned and applied the named entity recognition (NER) tool provided by a pre-trained and publicly available Danish language model¹.

Regular expressions were also constructed to extract information about the type of applicant (Ap). The possible outcomes for the type of applicant are male, female or couple, depending on (i) the applicant names mentioned in the introduction of the case file (where the applicant(s) are presented) and which are not referred to as children, and (ii) the frequency of gender specific words (i.e., he/she, his/hers) in the case file. We also constructed regular expressions to extract information on the number of children (Num) accompanied and included in an application. For the visual exploration of the dataset, we combined the two previous features (type of applicant and number of accompanied minors) into a “family category” feature.

Whenever available, the name of the chairing judge (J), located at the end of the decision, was extracted. Due to misspellings introduced by OCR and/or minor differences in the judge signature, we normalised the judge names through string similarity. Finally, we substituted the names with anonymous unique identifiers.

With regards to information on the applicant’s interview, whenever that was available, we extracted the following three features. First, we extracted the interview’s duration in minutes (ID), by isolating time instances related to the beginning and ending of the interview. We also extracted the number of times the word “asked” (“spurgt” in Danish) appears in the text, as a proxy for the interview length (IL). The third feature is the duration of the interview’s translation in minutes (TD), based on the given time of the translation’s ending, and by using the ending of the interview as starting time, since the time a translation started is not present in the text.

As direct information on the applicant education level and occupation are not always provided and when present they vary substantially from case to case, we decided to measure a proxy for how much space is given to education (Ed) and occupation (W) in the decision. After investigating a random sample of the text files, we gathered the most frequently appearing words associated with the two subjects, and based on their frequency of appearance, we categorised each case as having little/none, some or significant mention on the applicant’s education or occupation. We remark that these two features are not directly associated with the applicant’s level of education or type of occupation.

We also extracted whether the applicant knows Danish (Dan). If the text of the decision mentions that the applicant has some

¹We used spacy’s model da_core_news_lg. More information can be found here: https://spacy.io/models/da
knowledge of Danish, no matter how little this knowledge may be, we code this variable as true. For this, little knowledge of danish, or knowledge of just speaking but not writing and vice-versa, were assigned a positive outcome.

In order to account for credibility, after investigating the files and consulting with our legal colleagues, we collected words used by the RAB which express whether the applicant’s narrative is credible (i.e., words expressing positive credibility) or not (i.e., words expressing negative credibility). In their decisions, the RAB reports if the candidate gave contradictory/inconsistent or coherent/consistent answers during the interview/ hearing. We computed the frequency of appearance of such words in each asylum application in our dataset, and based on the distribution of positive and negative words: (a) we labelled a case as Credible/not applicable if the frequency of positive words was higher than the frequency of negative words, or if no credibility related words were present, (b) we labelled a case as Non credible with small indication, if the frequency of negative words was relatively small (i.e., no more than 20 instances in a case file) but higher than the frequency of positive words, and (c) we labelled a case as Non credible with significant indication if the frequency of negative words was large (i.e., at least 20 instances in a case file) and higher than the frequency of positive words. The rationale for constructing such a proxy is showed in Fig. 2. In this Figure, the x-axis (the y-axis respectively) shows the number of Granted (Rejected, respectively) asylum applications which contained the considered credibility-associated words. The red lines show the number of times words associated with negative credibility were found in Granted, Rejected cases. Similarly, the green lines correspond to the words associated with positive credibility. We observe (i) an obvious distinction of the two classes (positive vs negative credibility words) as well as (ii) the proximity of the negative words and the positive words to the Rejected and the Granted axis respectively.

We note that the credibility proxy we computed relates in part to the RAB’s post hoc legal reasoning motivating the decision outcome. Accentuated differently, the credibility indicator we compute here captures what the decision makers established regarding the applicant credibility. This means that this variable is not a predictor sensu strictu as the applicant credibility cannot be known before the decision outcome. As facts such as applicant credibility are legally established only by the decision makers, an objective credibility cannot be computed. In fact, any attempt to do so would risk to substitute the judges’ subjective assessment of credibility with another subjective assessment of credibility, by the researchers. As such, we use the applicant credibility as an explanatory variable, with the understanding that this credibility indicator captures the judges’ belief about how credible the candidate narrative is. The inclusion of credibility in our models will, therefore, avoid the problem of omitted variable bias. In fact, failing to include a relevant variable, such as credibility, might lead to finding spurious correlations between the outcome and other irrelevant variables, thus affecting the whole analysis [23].

In order to account for the context of the decisions, we extracted the most salient topics from the corpus. We used the non-negative matrix factorization (NMF), which decomposes a term-document matrix $V \in \mathbb{R}^{m \times n}$ into the product of two matrices of rank $p \ll \min(m, n)$ as follows: $V \approx WH$ where $W \in \mathbb{R}^{m \times p}$ and $H \in \mathbb{R}^{p \times n}$. The values of $W$ and $H$ are normalised in $[0, 1]$ and can be interpreted as probabilities. The $p$ components of $W$ and $H$ can, thus, be understood as topics. The matrix $W$ can be thought of as a term-topic matrix where an entry represents the probability that a given term is associated with a given topic and the matrix $H$ can be thought of as a topic-document matrix where an entry

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>Y</td>
</tr>
<tr>
<td>file size</td>
<td>FS</td>
</tr>
<tr>
<td>country of origin/ethnicity</td>
<td>C</td>
</tr>
<tr>
<td>entry year</td>
<td>EY</td>
</tr>
<tr>
<td>decision year</td>
<td>DY</td>
</tr>
<tr>
<td>applicant</td>
<td>Ap</td>
</tr>
<tr>
<td>number of accompanied minors</td>
<td>Num</td>
</tr>
<tr>
<td>chairman judge</td>
<td>J</td>
</tr>
<tr>
<td>interviews total duration</td>
<td>ID</td>
</tr>
<tr>
<td>interview length</td>
<td>IL</td>
</tr>
<tr>
<td>translation duration</td>
<td>TD</td>
</tr>
<tr>
<td>education</td>
<td>Ed</td>
</tr>
<tr>
<td>occupation</td>
<td>W</td>
</tr>
<tr>
<td>speaks danish</td>
<td>Dan</td>
</tr>
<tr>
<td>credibility</td>
<td>Cre</td>
</tr>
<tr>
<td>primary topic</td>
<td>PT</td>
</tr>
<tr>
<td>secondary topic</td>
<td>ST</td>
</tr>
</tbody>
</table>

Table 1: Extracted features along with their brief descriptions and short notation.
On predicting and explaining asylum adjudication

Figure 2: Naive word vector representation for words associated with negative and positive credibility, plotted along Granted-Rejected axes.

represents the probability that each document is associated with a given topic. We empirically determined, with the judgment of our legal colleagues, that the most coherent topics were obtained with \( p = 16 \). The sixteen topics are illustrated in Table 2. We can see that we have two groups of topics: one group is related to geographical areas (mainly capturing the country of origin of the applicant) and another group of four topics is related to the asylum motivation claimed by the applicant (topics labelled as persecution, religious conversion, homosexuals, politics-demonstration). Using the matrix \( H \), we can map each document to its primary and secondary topics (i.e. the two topics with the highest associated probabilities). We note that the first two topics account for more than 50% of the topic composition for more than 93% of documents. Table 2 reports also the number of documents associated with a given primary topic. Finally, we extracted the outcome of the decision using regular expressions that match specific sentences that are present in the text, in a mutually exclusive way, in case the board grants asylum (Granted), when the board rejects the applicant (Rejected), or if the case is returned to the first instance (the immigration service) for a new decision on the basis that some aspects of the case have not been evaluated in the first instance (Returned). With our strategy, we were able to label 15647 decisions (\( \sim 91\% \) of the whole dataset) distributed as follows: 2445 granted applications, 12983 rejected applications, and 219 returned cases. The remaining unlabelled cases have been excluded from our analysis.

### 3.2 Exploratory data analysis

Before modelling the outcome of asylum claims as a function of the features we extracted, we perform an exploratory data analysis to understand the correlation structure of our features. In Fig. 3 we plot the correlation matrix.

We observe that the mention of occupation related words (\( W \)), file size (\( FS \)), case year (\( Y \)), interview length (\( IL \)), interview duration (\( ID \)), translation duration (\( TD \)), and the mention of education related words (\( Ed \)) are strongly and positively correlated. This means that as one of these variables increases the other increase as well. This is evidence that over time case files are becoming longer because they include more details about the applicant narrative and also include interview transcripts with higher frequency. These variables exhibit a weak positive correlation with the outcome of the case.

Credibility (\( Cre \)) is positively correlated with the outcome. This means that the more the candidate’s narrative is judged as credible, the higher is the likelihood that the applicant will be granted asylum. \( W \), \( FS \), \( Y \), \( IL \), and \( ID \) exhibit a weak negative correlation with credibility. This shows that, when the claim is rejected on the ground of credibility, the Danish RAB supports the argument by listing the circumstances in which the applicant’s narrative appeared contradictory (e.g. when the applicant explained something in divergent ways during the interviews). This observation echoes previous findings on the centrality of credibility assessment in asylum adjudication (see section 2).

In figures 4–6 we use the term probability in its statistical sense: given a random case, we show the probability that such a case is Granted, Rejected, Returned or Unlabelled, once we know the value of a given feature. Fig. 4 shows on the left the probability of the decision outcome for every extracted country of origin/ethnicity. Countries with a frequency less than 100 across all cases in our dataset, are omitted. In the middle of Fig. 4, we present the probability of the decision outcome for every reported year of entry in Denmark. Apart from some spikes in a few years, we do not observe any significant variations here. Year 2021 contains only one case, which is Granted. Fig. 4 on the right shows the probability of the decision outcomes for every possible family category. Marginally higher Granted probability can be observed in the categories where minors are included or where the main applicant is female.
Table 2: Topics discovered with non-negative matrix factorization (NMF). Topic labels have been assigned by the authors upon inspection of the topics. The top 5 words sorted in decreasing order of probability, as assigned by the algorithm. The number of decisions is computed by assigning each decision to the most probable topic.

<table>
<thead>
<tr>
<th>Topic label</th>
<th>Top 5 words</th>
<th># Decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Afghan, Afghanistan, Hazara, Jalalabad, Taliban</td>
<td>3472</td>
</tr>
<tr>
<td>Persecution</td>
<td>Prison, Organization, Political, Interrogation, Military service, Weapons</td>
<td>2961</td>
</tr>
<tr>
<td>Iraq</td>
<td>Iraqi, Kirkuk, Mosul, Baghdad, Kurdish</td>
<td>1274</td>
</tr>
<tr>
<td>Religious Conversion</td>
<td>Church, Christian, Christianity, Religion, Conversion</td>
<td>1180</td>
</tr>
<tr>
<td>Iran</td>
<td>Iranian, Kurdish, Peshmerga, Kermanshah, Altash</td>
<td>1173</td>
</tr>
<tr>
<td>Syria</td>
<td>Syrian, Damascus, Kurdish, Aleppo, Kurds</td>
<td>1136</td>
</tr>
<tr>
<td>Somalia</td>
<td>Al-Shabaab, Somalia, Somali, Mogadishu, Circumcision</td>
<td>1074</td>
</tr>
<tr>
<td>Russia-Chechnya</td>
<td>Russian, Chechnya, Chechen, Moscow, Armenian</td>
<td>875</td>
</tr>
<tr>
<td>Homosexuals</td>
<td>Homosexual, Sexual, Sexuality, Boyfriend, Homosexuality</td>
<td>871</td>
</tr>
<tr>
<td>Balcan</td>
<td>Kosovo, Serbian, Serbia, Albanian, Yugoslavian</td>
<td>705</td>
</tr>
<tr>
<td>Lebanon-Palestine</td>
<td>Lebanese, Stateless, Beirut, Palestinians</td>
<td>705</td>
</tr>
<tr>
<td>Sri Lanka</td>
<td>Colombo, Sri Lankan, Tamil, Jaffna, Vavuniya</td>
<td>342</td>
</tr>
<tr>
<td>Eritrea-Etiopia</td>
<td>Eritrea, Ethiopia, Ethiopian, Eritrean, Somali</td>
<td>551</td>
</tr>
<tr>
<td>Politics-Demonstration</td>
<td>Demonstration, Protester, Participation, University, Political</td>
<td>531</td>
</tr>
<tr>
<td>Kuwait-Bidoons</td>
<td>Kuwait, Kuwaiti, Bidoon, Nawaf, Sulabiya</td>
<td>368</td>
</tr>
<tr>
<td>Myanmar-Rohingya</td>
<td>Burma, Rohingya, Burmese, Bangladesh, Chittagong</td>
<td>161</td>
</tr>
</tbody>
</table>

Fig. 4 shows on the left the probability of the decision outcomes for the three credibility indicators (as explained earlier), and in the middle and the right the probabilities of the decision outcomes for the education and occupation proxies (as explained earlier) respectively. With regards to the credibility, we observe an obvious increase of the Granted probability as the level of credibility increases. With regards to the applicant’s education, we observe a marginal higher probability for granting a case whenever there is significant mention of the applicant’s education and slightly less in cases with some mention, as opposed to cases with little or no mention. A similar trend is observed with the occupation proxy, where the probability of the Granted cases seems to grow linear with the level of mention of the applicant’s occupation in the case.

In Fig. 6 we observe that the interview duration does not vary significantly between the possible outcomes of a case. As for the topics, we observe that religious conversion cases and Iranian cases are the ones with the higher probability to be granted asylum. This is consistent with our prior knowledge of the Danish asylum system.

4 MODELLING ASYLUM ADJUDICATION

Here, we model the outcome of asylum decisions as a function of the features we extracted before. We opt for a binary prediction\(^2\) where we focus on the two main outcome classes: Rejected and Granted. Consequently, we ignore all cases where the outcome does not belong to these two classes. We use the following algorithms in the implementations provided by Sklearn: Decision Tree, Linear SVM, Logistic Regression, Naive Bayes and Random Forest.

Following our literature review and the exploratory data analysis, we expect that credibility and country of origin will turn out to be important factors in the Danish asylum adjudication system. Therefore, we begin by training the models on all the features and computing the feature importance. Feature importance is calculated through random permutations. That is, a feature is randomly permuted and the loss of predictive accuracy of the model is measured. The higher the loss of accuracy the higher the importance of the feature. The resulting feature importance scores shown in

\(^2\)In general, we use the term prediction in the mathematical sense of modelling the value of a given variable.
On predicting and explaining asylum adjudication

Figure 5: Probability of decision per credibility indicator (left), per education proxy (middle) and per occupation proxy (right).

Figure 6: (Left) Length of interviews (in minutes) against the case's decision. Probability of decision outcome per primary topic (middle), and secondary topic (right).

Fig. 7 have been normalized across all classifiers using the min-max technique. Similar results were obtained with the use of mean normalization, omitted here due to lack of space. It is evident that credibility is ranked as the most important predictor, by far, by all the algorithms. The other variables appear to be ranked quite differently by each model, although the country of origin seems to be the overall second most important variable. One pattern that emerges from Fig. 7 is that the linear models (SVM, logistic regression, and Naive Bayes) ascribe some degree of importance to almost every variable, while the non linear models (Decision Tree and Random Forest) select only few variables. This happens because non linear models can account for both non linearity and interactions between the features, being more robust to the presence of correlation in the features (see the correlation matrix in Fig. 3). We then take the exhaustive path of training models for every possible combination of features, for all the algorithms considered here. In this way, we can evaluate the accuracy with which the outcome of a case can be predicted, depending on the features available to the modeller.

For each combination of features and each employed classifier, we trained four different models. The first model uses the dataset as-is, heavily skewed towards the Rejected cases class, without accounting for class imbalance. The dataset was split into training and validation sets randomly, for every combination of features and employed classifier, following the 80-20 rule. The second model uses the same as-is dataset, and the same 80-20 split rule, but it splits the data chronologically. That is, we sort the dataset based on the year of the case and use the first (oldest) 80% of the cases for training and the most recent 20% for validation. This temporal strategy avoids the use of information from the future to predict the past. The third model is trained as the first one, and the forth model is trained as the second one, but both of them are trained on a balanced version of the dataset, with equal number of Rejected and Granted cases. The best performing models for each classifier and each of the four training strategies discussed above are reported in Table 3, along with the selected features.
From Table 3 we observe that the unbalanced models exhibit an apparently high accuracy (higher than 80%) and ROC AUC. However, when we look at the values of precision, recall, and $F_1$ score we can clearly see that such an accuracy is an illusion and the models have learnt only to correctly predict/classify the negative cases (i.e. the dominant class). In contrast, although scoring lower on the accuracy metric, balanced models exhibit consistently better performance. Precision, recall, and $F_1$ score for the balanced models confirm that these models are learning to correctly classify both positive and negative cases. Let us take as an example the random forest model trained without balanced and temporal strategies (second row of Table 3) which exhibit an accuracy of 87% but an $F_1$ score of 33% and compare it with the random forest model trained on balanced data without temporal strategy (fourth row of Table 3) with both accuracy and $F_1$ score of 75%. Taking into account the data imbalance offers a dramatic increase in performance.

Fig. 8 shows the distribution of accuracy for balanced models in the form of box-plots. For every classifier we aggregate the results for all the possible feature combinations (more than 100,000 combinations). The maximum accuracy for balanced models is around 74% and is achieved with a random forest classifier. The temporal versions of the models typically achieve higher median accuracy compared to their non-temporal version, as more models achieve higher accuracy in these cases. Credibility appears in all top-ranking balanced models (see Table 3). Other features that typically appear in the top-ranking balanced models are country of origin, primary and secondary topic, education, number of accompanied minors, occupation and entry and decision years. To remark the importance of credibility in asylum adjudication, in

Fig. 9 we show that random forest models that include credibility as a feature perform consistently and decisively better than random forest models that do not include credibility as one of the feature. Similar trends are found for the other classifiers as well. Finally, Fig. 9 visually shows the centrality of the assessment of candidate credibility in asylum adjudication, confirming prior findings from traditional legal research [11, 22].

5 DISCUSSION

Asylum adjudication research has sought to uncover biases and/or sources of disparities in recognition rates among similar cases by modelling the outcome of decisions as a function of metadata in the case file, of the decision maker, and of the applicant. Our literature review (see Section 2) has shown that 1) the credibility of the applicant’s narrative, his/her country of origin, and indicators of social integration are the most consistent factors to explain/predict asylum adjudication; 2) factors such as the decision maker, his/her gender, the applicant gender, and the applicant marital status represent sources of disparity only in certain jurisdictions; and 3) due to lack of access to the text of the case files, most of the prior quantitative research in asylum adjudication did not account for the credibility of the applicant’s narrative and for the context/asylum motive.

Therefore, in this paper we used a novel dataset containing the full text of 17300+ Danish asylum decisions from the Refugee Appeals Board to address limitations of prior research by 1), taking into account credibility of the applicant’s narrative and the context/asylum motive of each application; and 2) training and evaluating our models in different settings by considering both temporality and data imbalance. To correctly evaluate models in presence of data imbalance, next to accuracy and ROC AUC, typically used to evaluate models in asylum adjudication research, we report precision, recall, and $F_1$ score.

We find that the most important variables (see Table 3 and Fig. 7) to explain asylum adjudication are the credibility of the applicant’s narrative, the country of origin and primary and secondary topics (which represent context/asylum motive). We note that credibility is consistently the single most important factor for all the models, as shown in Fig. 7. The general trend in Fig. 7 is the following: the linear classifiers assign some relevance to almost all the variables, while the non linear classifiers tend to exclude more variables, with random forest excluding more variables than the single decision tree.
we trained models by under-sampling the majority class (i.e. the

Table 3: Top performing models under different training strategies. We report the following measures: Accuracy (Acc), Area Under Curve (AUC), Precision (Pre), Recall (Rec), and $F_1$ score. T and F stand for True and False respectively in the cases of Balanced and Temporal models.

<table>
<thead>
<tr>
<th>Classifier-Setup</th>
<th>Balanced</th>
<th>Temporal</th>
<th>Features in the Best Model</th>
<th>Acc</th>
<th>AUC</th>
<th>Pre</th>
<th>Rec</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>F</td>
<td>T</td>
<td>Y, FS, C, TD, Ed, J, Cre, PT, EY, DY</td>
<td>0.84</td>
<td>0.71</td>
<td>0.72</td>
<td>0.14</td>
<td>0.23</td>
</tr>
<tr>
<td>Random Forest</td>
<td>F</td>
<td>F</td>
<td>Y, FS, C, Ap, Num, IL, ID, Cre, PT, ST, EY, DY</td>
<td>0.87</td>
<td>0.80</td>
<td>0.72</td>
<td>0.21</td>
<td>0.33</td>
</tr>
<tr>
<td>Random Forest</td>
<td>T</td>
<td>T</td>
<td>C, Ap, TD, Ed, J, Cre, PT, ST, EY</td>
<td>0.70</td>
<td>0.74</td>
<td>0.71</td>
<td>0.72</td>
<td>0.72</td>
</tr>
<tr>
<td>Random Forest</td>
<td>T</td>
<td>F</td>
<td>Y, FS, C, Ap, Num, TD, ID, Ed, Cre, PT, EY</td>
<td>0.75</td>
<td>0.8</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>F</td>
<td>T</td>
<td>Ap, Num, Ed, Cre</td>
<td>0.83</td>
<td>0.6</td>
<td>0.59</td>
<td>0.08</td>
<td>0.13</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>F</td>
<td>F</td>
<td>C, ID, Ed, ST</td>
<td>0.86</td>
<td>0.56</td>
<td>1</td>
<td>0.02</td>
<td>0.04</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>T</td>
<td>T</td>
<td>Num, Cre</td>
<td>0.64</td>
<td>0.62</td>
<td>0.71</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>T</td>
<td>F</td>
<td>Y, Ap, Num, TD, Cre, PT, ST</td>
<td>0.65</td>
<td>0.64</td>
<td>0.65</td>
<td>0.59</td>
<td>0.61</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>F</td>
<td>T</td>
<td>FS, Num, Cre, PT, EY</td>
<td>0.83</td>
<td>0.61</td>
<td>0.73</td>
<td>0.03</td>
<td>0.05</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>F</td>
<td>F</td>
<td>FS, C, Ap, TD, ID, Cre, PT, EY</td>
<td>0.85</td>
<td>0.66</td>
<td>1</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>T</td>
<td>T</td>
<td>W, Cre, ST, EY</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.76</td>
<td>0.69</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>T</td>
<td>F</td>
<td>Cre</td>
<td>0.62</td>
<td>0.59</td>
<td>0.64</td>
<td>0.57</td>
<td>0.6</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>F</td>
<td>F</td>
<td>Y, C, Ap, Num, Ed, SD, Cre, ST, DY</td>
<td>0.85</td>
<td>0.71</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>F</td>
<td>T</td>
<td>Num, IL, TD, Ed, J, Cre, ST</td>
<td>0.83</td>
<td>0.68</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>T</td>
<td>T</td>
<td>W, Cre</td>
<td>0.66</td>
<td>0.67</td>
<td>0.69</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>T</td>
<td>F</td>
<td>W, Cre, PT, ST</td>
<td>0.64</td>
<td>0.66</td>
<td>0.64</td>
<td>0.65</td>
<td>0.64</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>F</td>
<td>T</td>
<td>Cre, PT</td>
<td>0.84</td>
<td>0.73</td>
<td>0.71</td>
<td>0.09</td>
<td>0.16</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>F</td>
<td>F</td>
<td>Cre, PT</td>
<td>0.85</td>
<td>0.73</td>
<td>0.60</td>
<td>0.12</td>
<td>0.21</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>T</td>
<td>T</td>
<td>C, Ed, SD, Cre</td>
<td>0.70</td>
<td>0.72</td>
<td>0.74</td>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>T</td>
<td>F</td>
<td>Cre, PT, ST, DY</td>
<td>0.7</td>
<td>0.7</td>
<td>0.73</td>
<td>0.68</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Table 3: Top performing models under different training strategies. We report the following measures: Accuracy (Acc), Area Under Curve (AUC), Precision (Pre), Recall (Rec), and $F_1$ score. T and F stand for True and False respectively in the cases of Balanced and Temporal models.

From Fig. 7 and Table 3 it is clear that the judge is not a relevant factor. Only few models include the chairing judge as a predictor, and these models are not among the best performing ones. The fact that the chairing judge is not so predictive of the outcome of a decision in the Danish case, as compared to other countries, may be explained by several factors. First, the Refugee Appeals Board uses a panel of 3 or 5 judges (depending on the year) which, by virtue of the Condorcet jury theorem [2], might reduce the importance (and a possible bias) of the single judge. Second, while in some countries judges are tied to a political party and even appointed by politicians, in Denmark judges remain more independent. This suggests that national legal traditions and practices also play a strong role in shaping asylum outcomes and hence that not all findings from country specific studies are likely to extend to other countries. As such, it is important to consider the legal system and culture of the country under analysis.

Other factors such as education, occupation, or the interview length appear to have only a very small effect and to mostly correlate with credibility. This is consistent with the findings from the stream of legal research [8, 11, 22]. In summary, our analysis confirms the importance of credibility and country of origin, while also showing the importance of the context/asylum motive. Other factors such as the characteristics of the applicant or the importance of the judge, instead, seem to be country/jurisdiction specific.

We also find that accounting for the imbalance in the data is crucial in order to develop trustworthy predictive models with good performance. In order to counter the class imbalance in our data, we trained models by under-sampling the majority class (i.e. the rejected cases) to balance the training dataset. This is one possibility but other strategies exist and we refer the reader to specialised literature [10, 14]. The models that do not account for data imbalance in Table 3 exhibit high values of accuracy and AUC, but when we analyse precision and recall we discover that these models are just very good at predicting negative outcomes (i.e. rejected cases). However, in our dataset the rejected cases are ~ 80%; as such, a model that always predicts a case to be rejected would get an accuracy of ~ 0.8. When we consider precision, recall, and $F_1$ score it becomes clear that the balanced models perform considerably better, even when predicting future cases (i.e., the temporal models in Table 3). Prior research with predictive models of asylum adjudication has used measures like accuracy and AUC to evaluate models [3, 6, 12]. As we find that accuracy and AUC by themselves are misleading, we recommend researchers to evaluate their models using metrics that are robust to data imbalance (such as precision, recall and $F_1$). In short, our analysis shows that using training strategies and evaluation metrics that account for data imbalance is of paramount importance.

Finally, our analysis shows that different models might substantially disagree on the importance of the variables and even on their relative rank. As such, the practice of fitting one model to data and evaluating the importance of predictors, in an attempt to find what is significant, is misinformed. We believe that comparing across different models is a more informative strategy. In our case, this shows the importance of credibility as the key factor driving asylum outcomes. Finally, in our case linear models have been unable to filter out the non important predictors; therefore, if researchers do not have a prior theory on what features to expect as significant
and/or do not have strong theoretical expectations on the functional form of their features, we find that machine learning models that can deal with a large number of features and are able to model non-linear features and their interactions are better options. That said, we remark that being guided by theory is more likely to result in more robust findings and better models.

6 CONCLUSIONS

In this paper, we considered the problem of modelling asylum adjudication. Using an original dataset of asylum decisions from the Danish Refugee Appeals Board, we trained and evaluated a large number of models for predicting and explaining asylum adjudication in Denmark. We show that the single most important factor in asylum adjudication is how decision makers consider the credibility of the applicant’s narrative; a finding that reinforces findings in qualitative studies in legal research. We find that other factors, shown to impact asylum outcomes in other countries, do not have a significant impact on Danish adjudication, pointing to underlying differences in national legal systems.

Considering the practice of developing predictive models of asylum adjudication, we have shown that it is of paramount importance to account for imbalanced data, since most of asylum decisions are rejected. In fact, models that do not take into account the data imbalance may perform quite misleading. As such, our paper restates the paramount importance of training and evaluating models with strategies to account for data imbalance.

REFERENCES


Author(s): Sebastiano Antonio Piccolo, Thomas Gammeltoft-Hansen, Panagiota Katsikouli, Tijs Slaats

Title: On predicting and explaining asylum adjudication

MOBILE Working Paper, No. 5, 2023

Publication date: 01/06/2023

© Author

MOBILE Working Paper Series

Sebastiano Antonio Piccolo, Postdoctoral Researcher, University of Copenhagen

E-mail: Sebastiano.piccolo@jur.ku.dk

MOBILE (Center of Excellence for Global Mobility Law) was established in 2023 based on a grant from the Danish National Research Foundation. The center’s mission is to systematically study the legal infrastructures of human mobility across geographies, social divides, travel patterns and time.

Papers are available at: https://mobilitylaw.ku.dk/working-papers/

MOBILE

Danish National Research Foundation’s Center of Excellence for Global Mobility Law

The Faculty of Law, University of Copenhagen

Karen Blixens Plads 16, 2300 Copenhagen S