Semi-automatic Triage of Requests for Free Legal Assistance

Meladel Mistica[♦♥] **Jey Han Lau**[♦]
Brayden Merrifield[♣] **Kate Fazio**[♣] **Timothy Baldwin**[♦]
[♦] Computing and Information Systems, The University of Melbourne
[♥] School of Languages and Cultures, The University of Queensland
[♣] Innovation and Engagement, Justice Connect
{m.mistica, jeyhan.lau}@unimelb.edu.au, tb@ldwin.net
{brayden.merrifield, kate.fazio}@justiceconnect.org.au

Abstract

Free legal assistance is critically under-resourced, and many of those who seek legal help have their needs unmet. A major bottleneck in the provision of free legal assistance to those most in need is the determination of the precise nature of the legal problem. This paper describes a collaboration with a major provider of free legal assistance, and the deployment of natural language processing models to assign area-of-law categories to real-world requests for legal assistance. In particular, we focus on an investigation of models to generate efficiencies in the triage process, but also the risks associated with naive use of model predictions, including fairness across different user demographics.

1 Introduction

The number of Australians with unmet legal needs is estimated to be over 4 million people per year and growing (out of a total population of around 25 million): each year approximately 8.5 million Australians will have a legal problem and only around 4.5 million will access any assistance (Coumarelos et al., 2012; The Department of Justice and Regulation, 2012) — an indication that free legal assistance services are critically under-resourced. A bottleneck for free legal assistance providers is the determination of what (if any) specific legal needs the individual has. We investigate the viability of semi-automating this step by building a model that suggests how to categorise lay descriptions of problems/incidents into legal areas. It is critical that we develop models which will perform equally well for users of all backgrounds, generalise well from small amounts of curated data, and potentially dynamically interact with the help-seeker to clarify the nature of the case. However, in this preliminary work, our aim is to develop initial models as a means to ascertain what biases manifest in our given data, and to have a workable model upon which we can make incremental measurable improvements.

Text classification of any real-world data can be a challenge for many reasons. In the case of legal text classification, the classes themselves or the legal categorisation of a possible case, can vary from organisation to organisation, and also from court to court; there is no universally agreed-upon set of areas of law neatly defined into a taxonomy (Goncalves and Quaresma, 2005; Sulea et al., 2017a; Soh et al., 2019; Tuggener et al., 2020). Furthermore, a case can span multiple areas of law — for example, a FAMILY LAW matter could also fall under the umbrella of GUARDIANSHIP AND ADMINISTRATION, or a CHARITIES LAW issue may also have aspects regarding EMPLOYEES AND VOLUNTEERS. In addition to the issues surrounding the inherent fuzziness of legal categories, the descriptions of legal issues themselves exhibit a range of language styles: those who seek free legal help are not versed in the legal domain, and may have varying linguistic styles, reflecting their social, cultural, and educational background.

We report on an ongoing collaboration between Justice Connect, a public benevolent institution, as defined by the Australian government,[1] that aims to ameliorate social inequalities through legal assistance and community engagement, and Melbourne University whose aim is to alleviate the help-seeker intake bottleneck. In Section 2, we outline the importance of accessible legal assistance to those most in need, and the barriers to be overcome in providing this service to the community. Section 3 details the data collection and corpus creation process. We designed and developed an annotation platform exclusively for volunteer lawyers to annotate online requests for help from the public through the Justice Connect intake portal. Our experiments and results are outlined in Section 4

[1] https://www.acnc.gov.au/charity/232d6dcbcaa1550da90f825fe6fab643#history

Natural Legal Language Processing Workshop 2021, pages 217–227

and 5, which describe the initial fine-tuned BERT classifiers (Devlin et al., 2019) on the small curated help-seeker data informally describing issues in their own words on matters they believe require legal assistance. As a starting point, we wanted to leverage the patterns of language usage encoded in BERT given our relatively small data set. The main risk is that while robust results can be achieved by fine-tuning over relatively little labelled data in this manner, the data used in developing the pre-trained models can lead to these models implicitly capturing a variety of biases about the world (Bender et al., 2021). In Section 6, we reveal how these biases manifest for our given data set, not only in terms of which areas of law the models can, or cannot, reliably predict, but also which demographic groups the model has inherent difficulty in representing. Finally, in Section 8, we discuss how we can overcome these biases for future iterations of the model while keeping in mind the protection and privacy of the help-seekers who are most vulnerable.

2 Background

Unresolved legal problems have been shown to lead to significant life impacts at high levels of frequency, including financial strain (29%), stress-related illness (20%), physical ill health (19%), relationship breakdown (10%), and having to move home (5%) (Coumarelos et al., 2012).

Even when a person is eligible for free legal assistance, there are various ecosystem-level barriers that increase the difficulty of finding and engaging with a legal service. One such barrier is the disconnection between a person who has recognised that their problem may have a legal dimension but does not yet know the technical terminology around their issue, and a legal service that can assist that person with the problem they have. This barrier is exacerbated in online settings, as people search via search engines and directories for legal services without the right search terms and technical language required to successfully reach relevant services.

In a follow-up survey by Justice Connect, many applicants, in requesting help online, self-identify this issue: e.g. "[there are] too many seperate courts and unclear what laws do what (sic)", "it's complex and i am not an expert!" and "[I'm unsure of the category] because of the family relationship together with financial issue".

Legal services also experience this issue as a bottleneck in their processes, where significant resources are required to assist applicants to determine the nature of the problem that they have (especially difficult given that many users of such services have little or no formal experience with the legal system), whether it is legal in nature, and what specific legal services should be provided. Lack of knowledge and capability often results in "failing to identify a legal problem, consulting non-legal advisers instead of legal experts, or taking no action to resolve the problem" (The Department of Justice and Regulation, 2016, p120).

Legal area classification (Goncalves and Quaresma, 2005; Sulea et al., 2017a,b; Soh et al., 2019; Tuggener et al., 2020) can potentially help to alleviate this bottleneck, in providing semi-automatic legal triaging of user-supplied textual descriptions of the issue.

3 Data Set Development

The corpus upon which the data set is derived comes from requests for help via the Justice Connect web-based intake tool.[2] After a series of eligibility questions, the help-seeker is asked to provide more information about the issue that has brought them to the Justice Connect portal.

The description entered by the help-seeker is manually de-identified for names, dates, locations, and any other sensitive information in preparation for annotation. This is then presented to an annotator via an interface which was developed by Justice Connect in consultation with Melbourne University. Based on the description, the annotator selects one or more areas of law that they deem appropriate from the 33 options (including NOT A LEGAL ISSUE) shown in Table 1. The annotation task is a two stage process: first the areas of law are chosen including specifying certainty levels which reflect how well the text supports the area of law they have chosen, as shown in Figure 1. In the second stage, the annotators are asked to highlight the relevant passages that support their decision to label the document as belonging to a particular legal area.[3]

The annotations were collected over a period

[2] For the privacy and protection of the help seekers, we are not able to share the intake took data of real-world examples. Although the data has been anonymised, the risk too high given the sensitivity of the material and its recency.

[3] We do not use the span-level information from Step 2 in this work, but the highlighting of passages can be seen in Figure 3 in the Appendix.

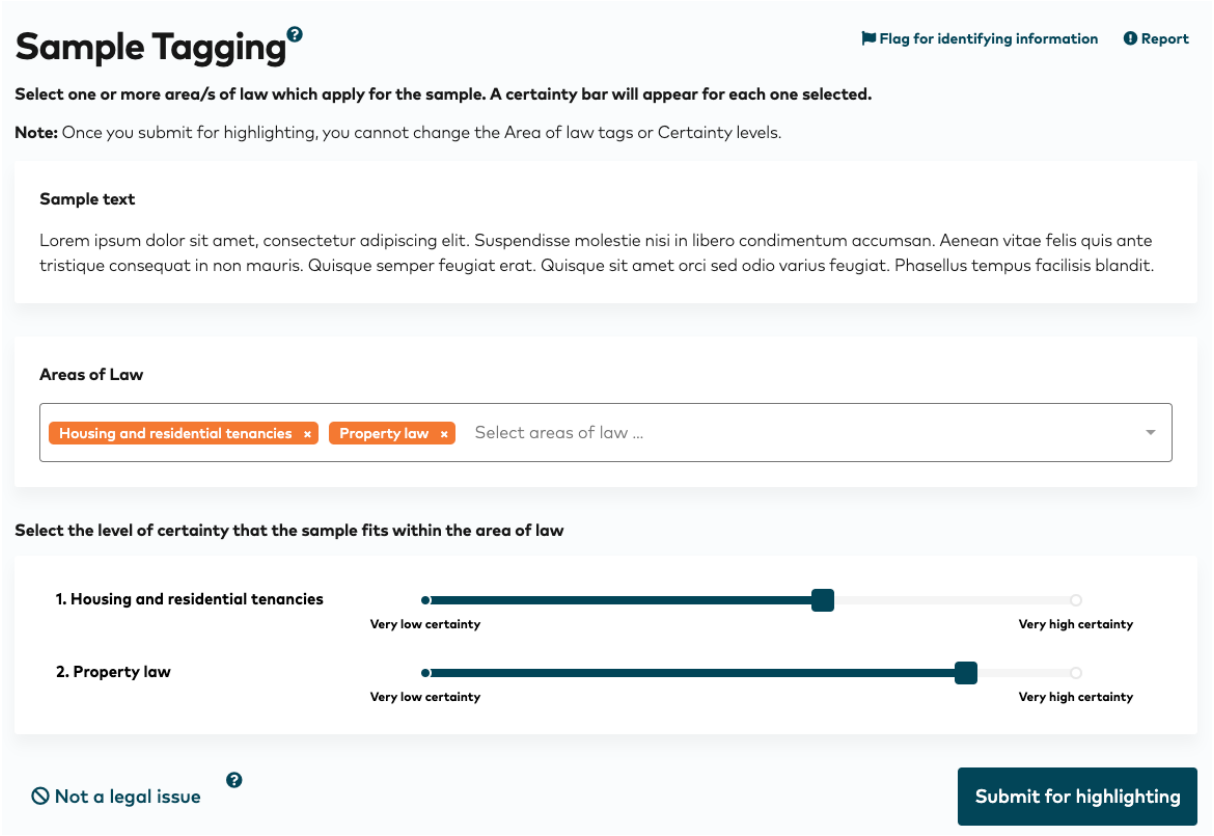

Figure 1: Step 1: Choosing the areas of law and certainty levels

of five and half months,[4] and are based on 4,062 unique descriptions. Each text sample was annotated by up to 7 different lawyers, noting that a single sample could be annotated as falling under multiple areas of law, making our task a multi-label classification problem.

Our annotators are lawyers from firms that were approached based on their level of engagement with Justice Connect. A number of firms had previously shown interest in pilot and project opportunities through Justice Connect's subscription model. Annotators were self-elected, or chosen by each firm's pro bono coordinators. They were asked to disclose how many years they had been practicing in total. Of these firms, there were 231 lawyers – all admitted to practice law in Australia – who were signed up for the annotation task. In addition, there were 12 Justice Connect-based lawyers who self-elected to participate as annotators, taking the total number of lawyers to 243 over 9 firms throughout Australia.

We derive three different labellings of the data from these annotations:

Majority-vote: majority-vote labels for each text sample based on a per-class majority vote over the annotators. That is, in order for a label to apply it must have been assigned by at least half of the annotators.

Confidence-weighted: the weighted mean of the 'certainty' or 'confidence' score for each label, as self-assessed by the annotator on a scale of 1–100 (according to the placement of a slider), by averaging over the confidence scores for all annotators who assigned a given label.

Annotator-weighted: the proportion of annotators who assigned a given area of law to the instance, divided by the number of annotators (constrained to be at least 3).

As an illustration of how confidence- and annotator-weighting work, one text sample (Entry ID 3085) had 3 annotators: the first annotator labelled it with the tags LITIGATION AND DISPUTE RESOLUTION and EMPLOYEES AND VOLUNTEERS, giving a score of 42% and 100%, respectively; the second annotator applied the same labels but rated them both 100; and the third annotator tagged this excerpt with EMPLOYEES AND VOLUNTEERS and PERSONAL SAFETY with confidence ratings 79 and 67, re-

[4] From 16 November 2020 to 30 April 2021

Area of Law	M/C	A
Banking & finance	137	96
Building & construction	229	61
Charities Law	40	32
Consumer law	145	107
Corporate & commercial law	82	53
Criminal law	458	375
Elder Law	45	29
Employees & volunteers	725	559
Environment	9	7
Family law	442	375
Fines & infringements	115	89
Fundraising law	1	1
Guardianship & administration	82	61
Health	129	94
Housing & residential tenancies	577	468
IT	2	1
Inquiries	62	4
Insurance	62	39
Intellectual Property	21	20
Litigation & dispute resolution	1258	732
Migration	121	95
Native title	1	1
Neighbourhood disputes	55	42
Not a legal issue	259	195
Personal safety	190	168
Planning & local government	62	50
Privacy	43	34
Property law	182	131
Public & administrative law	249	165
Tax	37	26
Torts & compensation	226	180
Trusts/equity	22	14
Wills, estates & probate	90	63

Table 1: Areas of Law, and the number of text samples belonging to each in the majority-vote (M), confidence-weighted (C), and annotator-weighted (A) data sets.

spectively. Under confidence-weighting, therefore, the overall weight for Litigation and dispute resolution is (42/100 + 100/100)/2 = 0.71, whereas for annotator-weighting, the score is 2/3 = 0.67.

One difference between the annotator-weighted labelling as compared to the confidence-weighted or majority-vote labelling is that it has fewer instances: the annotator-weighted data set has 3,154 instances while the other two have 4,062. This is because of the constraint that at least 3 annotators tag the text for annotator-weighting.

Table 1 shows the number of instances categorised under each area of law for the majority-vote (and confidence-weighted) data set and the annotator-weighted data set, noting that multiple areas of law can apply to the one text sample.

4 Experiments

We performed fine-tuning experiments using BERT, with early stopping based on the validation loss. In addition, we experimented with various values for the dropout rate during training, with the final value being set to 0.001. The batch size was set to 32 and the number of epochs was set to 50. All experiments are based on 20-fold cross-validation, and were run 10 times and averaged, with each run having the data split into 20 folds randomly. The non-testing portion of each fold was split into 90/10 for training/validation. For experiments over the confidence-weighted and annotator-weighted data sets, we train over the given label representation, but evaluate on the majority-vote data. We do this because the majority-vote labelled data is our approximation of a manually curated gold-standard data set, also for direct model comparability.

Given that the number of labels is quite large (33 areas of law, including Not a legal issue), we wanted to see if grouping the tags into small thematic groups would increase accuracy. We experimented with 2 grouping structures: (1) "legal", based on legal specialisations; and (2) "theme", based on topics or themes that may be shared between the areas of law. These groupings were agreed upon by trained lawyers at Justice Connect, where the first group was determined by answering the question, *In general, if a lawyer specialises in area X, do they often specialise in area Y too?*, and for the latter, *What areas of law have common narratives or topics shared between them, when people describe issues pertaining to these areas of law?*

We include three baselines to gauge how difficult the task is:

1. "random": choose 0 to 7 labels at random (based on uniform sampling, without replacement) for each instance, noting that there were between 0 and 7 areas of law for each instance in the majority-vote data set[5]

2. "shuffle": select N labels at random, where N is the number of assigned labels to the instance in either the majority-vote or annotator-weighted data set (recognising that this information would not be available for a genuinely "unseen" instance)

3. "majority": label using the most popular area(s) of law (ranked by how many instances they have been assigned in the training data).

[5]Note that in the case of 0 areas of law, Not a legal issue is assigned, and in the case of >0 areas of law, they are drawn from the remainder of the labels.

For the majority-vote data set, on average 1.5 labels are applied to each text sample, and therefore the first version of majority, we assign the 2 most popular areas of law. The annotator-weighted data set has an average of 1.3 labels per instance, so for the second version of majority, we assign the single highest-occurring label.

5 Results

Table 2 shows the micro-averaged precision, recall, and f1-score for the BERT models trained over the three data sets (majority-vote, confidence-weighted, and annotator-weighted) but evaluated over majority-vote for direct comparability. For the baselines, the two numbers in each cell indicate the results obtained based on majority-vote vs. annotator-weighted.

The results show that while the confidence-weighted-trained system outperforms the other systems with respect to precision, recall is low and misses out on correctly classifying instances well over half the time. Although, when it does label instances, it gets them correct 83.8% of the time. The annotator-weighted-trained model obtains the highest recall, and also obtains the best trade-off between precision and recall, as reflected in the best f1-score. All models perform well above all three baselines, for all of precision, recall, and f1-score.

For all 3 BERT models, recall is rather low. In the original experiments, if the model predicted a given label with a probability ≥50%, we output it as a prediction, but it is also possible to adjust the probability threshold to a value other than 50% (with higher values expected to lead to higher precision and lower recall, and lower values expected to lead to higher recall and lower precision). In a follow-up experiment, we learned a threshold per label (area of law) based on the training data. While these experiments, shown in Table 3 generally led to improvements in recall, it degraded precision substantially, with the net effect of an overall drop in f1-score. As such, in the remainder of our experiments, we maintain a fixed threshold of 0.50.

In the previous experiments, the validation data was the same type as the training data (e.g. for 'annotator-weighted', the training and validation data were both labelled with the annotator-weighted approach), but final evaluation was based on the majority-vote labelling. This means that for the models trained on 'annotator-weighted' and 'confidence-weighted', we optimise hyper-parameters on the basis of one labelling strategy, and perform our final evaluation based on a separate labelling strategy ('majority-vote'). In the next experiment, we seek to rectify this mismatch by also validating on 'majority-vote' data.

The results in Table 4 show a slight boost in recall in both cases, but overall validating on majority-vote data has relatively little impact.

Table 5 shows the results of the groupings experiment, in which we both label and evaluate all instances based on the coarser-grained 'theme' (groupings of the areas of law by topic) or 'legal' (groupings according to legal specialisation) label set,[6] using majority-vote labelling. Overall, recall improves slightly with grouped labels, in comparison to the original results over the fine-grained label set. However, precision does not improve, meaning it is difficult to justify employing the 'theme' (or the 'legal') system over the 'annotator-weighted' system, because the loss in the granularity of distinctions between the specific areas of law does not justify the small gain in recall.

Summarising the findings of these experiments, the best of the models is trained on 'annotator-weighted' labels, without modifying the probability threshold or majority-vote data validation. It is this model that we experiment with in the remainder of the paper, in terms of scoping out its viability for live deployment in semi-automatically triaging of incoming requests for legal assistance.

6 Analysis

As shown in Table 1, not all areas of law are distributed equally. For example, there are far more instances tagged as EMPLOYEES AND VOLUNTEERS than PRIVACY. It would be natural to expect that the predictive performance over PRIVACY would be lower, given the sparsity of labelled instances in the training data.

Table 6 shows the breakdown of precision, recall, and f1-score (along with the raw count of true negatives, false positives, false negatives, and true positives) for each label. There were no classes where the number of false positives was greater than the number of true positives, resulting in relatively high precision scores. Recall, on the other hand, is rather low, meaning that the model is conservative in its predictions. However, a more conservative low re-

[6] See Figure 2 in Appendix for details.

	System	p	r	f
Baselines	shuffle	.059/.065	.059/.065	.059/.065
	random	.041/.046	.074/.075	.052/.057
	majority	.232/.244	.168/.322	.195/.278
BERT	majority-vote	0.770	0.545	0.638
	confidence-weighted	**0.838**	0.460	0.594
	annotator-weighted	0.781	**0.588**	**0.671**

Table 2: Results for the three baselines, and BERT trained on the three data sets (majority-vote, confidence-weighted, and annotator-weighted); evaluation in each case is against the majority-vote test set.

System	p	r	f
majority-vote	0.565	0.555	0.560
confidence-weighted	0.622	0.521	0.567
annotator-weighted	0.743	0.569	0.645

Table 3: Results for dynamic probability thresholding of the BERT models

System	p	r	f
confidence-weighted	0.839	0.462	0.596
annotator-weighted	0.772	0.608	0.680

Table 4: Using majority-vote data for validation

call, high precision, system provides greater utility than a high recall, low precision, system because we are able to trust the predictions from the system in assigning lawyers with speciality in different areas of law. That is, we want to be confident that a pro bono lawyer who is assigned to a client is suitably credentialled to provide assistance relevant to the specifics of the request, noting that they will quickly pick up on any aspects of the case which they are not qualified to deal with (i.e. areas of law the classifier has missed) and be able to potentially bring in extra expertise without extra overhead. That is, the cost of a lawyer getting up to speed with a particular case is very much higher than the cost of that lawyer identifying extra dimensions of legal expertise that need to be brought on board, such that precision is more important than recall.

To assist in the interpretation of the model after deployment, we further categorised the areas of law according to 4 tiers, as shown in Table 6. The determining of these tiers is roughly guided by the f0.5-score ($\beta = 0.5$) of each area of law, shown in the column *f0.5*. In constructing the tiers, we place

System	p	r	f
theme	0.776	0.629	0.695
legal	0.777	0.612	0.684

Table 5: Groups by themes legal specialisation

a greater importance on precision rather than recall. Tier III and IV classes, with an f0.5-score of < 0.55, are those that are least 'trustworthy' if the system was to emit them as a prediction. Even though some classes have a precision of 1.000 (e.g. Intellectual Property and Trusts/equity), are still be treated as lower-tier classes for a number of reasons — these classes have far fewer instances and therefore the ability of the model to learn features for these classes is comparatively degraded (and the high precision is perhaps more luck than a reproducible trend). This is reflected in the very poor results in the recall and thus overall f1-score of these classes. The classes in Tiers I and II have higher precision and an f0.5 that range between 0.55 and 0.925. For the areas of law in these tiers with a higher precision, we expect that when a model predicts a text sample as one of these classes, that we can be fairly confident of that prediction.

7 Fairness

From the metadata provided by Justice Connect, we analyse 6 sub-groups of help-seeker:

1. Seniors (sen) = 102 instances;
2. Aboriginal/Torres Strait Islanders (ats) = 54 instances;
3. public housing tenants (pub) = 91 instances;
4. those who do not identify as heterosexual or cisgender (lgb) = 104 instances;
5. those who are homeless or as risk of becoming homeless (hom) = 175; and

Area of Law	p	r	f1	f0.5	TN	FP	FN	TP
TIER I								
Migration	0.959	0.737	0.833	0.904	3056	3	25	70
Insurance	0.941	0.410	0.571	0.747	3114	1	23	16
Employees and volunteers	0.931	0.896	0.913	0.924	2558	37	58	501
Fines and infringements	0.922	0.528	0.671	0.802	3061	4	42	47
Wills, estates and probate	0.919	0.540	0.680	0.806	3088	3	29	34
Building and construction	0.891	0.672	0.766	0.836	3088	5	20	41
Housing and residential tenancies	0.897	0.835	0.884	0.865	2641	45	77	391
Family law	0.828	0.760	0.793	0.813	2720	59	90	285
TIER II								
Planning and local government	0.864	0.380	0.528	0.689	3101	3	31	19
Consumer law	0.758	0.467	0.578	0.674	3031	16	57	50
Not a legal issue	0.745	0.554	0.635	0.697	2922	37	87	108
Criminal law	0.721	0.592	0.650	0.691	2693	86	153	222
Litigation and dispute resolution	0.709	0.552	0.621	0.671	2256	166	328	404
Neighbourhood disputes	0.737	0.333	0.459	0.593	3107	5	28	14
Guardianship and administration	0.719	0.377	0.495	0.609	3084	9	38	23
TIER III								
Charities Law	0.875	0.219	0.350	0.547	3121	1	25	7
Tax	0.857	0.231	0.364	0.556	3127	1	20	6
Elder Law	0.750	0.103	0.182	0.332	3124	1	26	3
Property law	0.682	0.443	0.537	0.616	2996	27	73	58
Torts and compensation	0.650	0.372	0.473	0.565	2938	36	113	67
Health	0.625	0.372	0.467	0.550	3039	21	59	35
Personal safety	0.575	0.387	0.463	0.524	2938	48	103	65
Corporate and commercial law	0.562	0.340	0.424	0.497	3087	14	35	18
Banking and finance	0.512	0.427	0.466	0.492	3019	39	55	41
TIER IV								
Intellectual Property	1.000	0.150	0.261	0.469	3134	0	17	3
Trusts/equity	1.000	0.071	0.133	0.279	3140	0	13	1
Privacy	0.500	0.059	0.105	0.200	3118	2	32	2
Public and administrative law	0.519	0.339	0.410	0.469	2937	52	109	56
Fundraising law	0.000	0.000	0.000	0.000	3153	0	1	0
IT	0.000	0.000	0.000	0.000	3153	0	1	0
Inquiries	0.000	0.000	0.000	0.000	3150	0	4	0
Environment	0.000	0.000	0.000	0.000	3147	0	7	0
Native title	0.000	0.000	0.000	0.000	3153	0	1	0

Table 6: Breakdown of the performance for the 'annotator-weighted' system per class, as well as the TN (true negative), FP (false positive), FN (false negative), and TP (true positive) counts.

6. those who disclosed their household income as being less than $50K group (low) = 1,686 instances.

Of these, the sen group fared the worst with an overall f1-score of 0.540, as shown in Table 7. This table shows that the ats and sen groups fared below the average and the worst for sen.

Aboriginal and Torres Strait Islanders make up 1.7% of the total data, and make up a comparable percentage of Family law cases (1.8%) and Criminal law cases (1.9%), however at a far lower performance than the average when compared to data for all demographics. For example Criminal law cases have p = 0.571, r = 0.571, and f1 = 0.571 (vs.

SELF-IDENTIFIED	p	r	f
Aboriginal and Torres Strait Islander ATS	**0.723**	**0.573**	**0.638**
Public housing tenants PUB	0.793	0.635	0.706
LGBTQI LGB	0.794	0.612	0.693
Homeless (or at risk) HOM	0.801	0.591	0.681
Low income LOW	0.796	0.610	0.691
Seniors SEN	**0.714**	**0.439**	**0.543**
ALL DATA	0.781	0.588	0.671

Table 7: Results for self-identified groups in the data

p = 0.721, r = 0.592, and f = 0.650 when evaluating over all the data). For LITIGATIONS AND DISPUTE RESOLUTIONS, the largest overall class, the performance for ATS was p = 0.333, r = 0.600, and f = 0.429 (vs. p = 0.709, r = 0.552, and f1 = 0.621 when evaluating over all the data). However only 0.7% of documents in this class were submitted by those who identified as ATS.

The number of submitted requests by seniors (SEN: those who identify as being over 65 years of age) are almost double the number of Aboriginal and Torres Strait Islanders, making up slightly over 3% of the total number of samples in the data (3.2% of the data used in the annotator-weighted system). Our initial hypothesis for the poor performance of the SEN data was that perhaps they made enquiries in certain areas of law that inherently had low performance. For example, if most of the seniors' enquiries fell in the Tier III and IV categories, then this would help explain the overall poorer performance of the SEN group. However, our analysis showed otherwise: the vast majority of classes with at least 5 instances from the SEN group are Tier I or Tier II, and yet for the majority of these, the relative performance is below the overall performance for that class, as seen in Table 8. This table shows the breakdown of the areas of law for SEN where the number of true instances is at least 5.

Many of the enquiries by seniors fall in Tier I and II categories, which make up the best-performing classes in general. Even though the SEN group only makes up around 3% of the total number of instances, they do however make up almost 21% of all enquires regarding ELDER LAW, over 6% of all BANKING AND FINANCE, almost 6% of CONSUMER LAW, and 10% of PLANNING AND LOCAL GOVERNMENT. While PROPERTY LAW and PLANNING AND LOCAL GOVERNMENT perform well in comparison to the overall results for these classes, the other classes within Tier I and II categories underperform. In particular, EMPLOYEES AND VOLUNTEERS and FAMILY LAW are overall top-performing classes as shown in Table 6, yet perform poorly for the SEN group, obtaining an f1-score of 0.571 and 0.545, respectively. Furthermore, in instances where precision is high for a class, recall often pulls down the f1-score for the SEN group.[7] This points to the model being systematically biased against this particular demographic, pointing to the need for explicit model debiasing.

8 Conclusion

The findings of our paper presents a preliminary exploration of the application of NLP for social good. The results of this paper shows the importance of building fairer models. One approach as a future avenue in this endeavour is by incorporating adversarial learning (Li et al., 2018) or null space projection (Ravfogel et al., 2020) to learn representations that are invariant to subgroups, so as to limit the model from learning undesirable correlations between the legal categories and sub-group features. The challenge though, is that subgroup information is not always available in the data, particularly in the Justice Connect intake tool to request help where all demographic information is volunteered and not mandatory. This means that possibly identifying *proxy attributes* (e.g. postcode and education level as a potential means to identify income status) or areas of law that are highly associated with certain sub-groups (e.g. MIGRATION issues may be likely submitted by persons whose main language is not English, and ELDER LAW issues may likely be submitted by seniors). In addition to the subgroups already presented in this study, in our next iteration

[7] This is the case for areas of law such as NEIGHBOURHOOD DISPUTES and CONSUMER LAW, which can be seen in Table 8 (in the Appendix).

Area of Law	p	r	f	Tier
Employees and volunteers	0.500	0.667	0.571	Tier I
Family law	0.600	0.500	0.545	Tier I
Housing and residential tenancies	0.750	0.462	0.571	Tier I
Litigation and dispute resolution	0.556	0.652	0.600	Tier II
Not a legal issue	0.667	0.750	0.706	Tier II
Criminal law	0.714	0.500	0.588	Tier II
Neighbourhood disputes	1.000	0.500	0.667	Tier II
Planning and local government	1.000	0.600	0.750	Tier II
Banking and finance	0.400	0.333	0.364	Tier III
Consumer law	1.000	0.167	0.286	Tier II
Elder Law	0.000	0.000	0.000	Tier III
Torts and compensation	0.000	0.000	0.000	Tier III
Property law	1.000	0.700	0.824	Tier III

Table 8: Breakdown of areas of law for SEN for where number of true instances ≥ 5

we will expand our set of the subgroups to include people who identify as CALD (culturally and linguistically diverse), those who identify as having a disability, as well as information on education levels.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Christine Coumarelos, Deborah Macourt, Julie People, Hugh M. McDonald, Zhigang Wei, Reiny Iriana, and Stephanie Ramsey. 2012. Legal Australia-wide survey (LAW survey): Legal need in Australia. Law and Justice Foundation of NSW.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Teresa Goncalves and Paulo Quaresma. 2005. Evaluating preprocessing techniques in a text classification problem. In *Proceedings of the Conference of the Brazilian Computer Society*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online.

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. 2019. Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Octavia-Maria Sulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Sulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases.

The Department of Justice and Regulation. 2012. Access to justice review. Victorian Government Report.

The Department of Justice and Regulation. 2016. Access to justice review. Victorian Government Report.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. 2020. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1235–1241, Marseille, France. European Language Resources Association.

Appendix

Group by Similar Theme/Topic

- Group 0 - not a legal issue
  - NOT A LEGAL ISSUE
- Group 1 – protecting personal interests
  - INTELLECTUAL PROPERTY
  - PRIVACY
  - CONSUMER LAW
  - EMPLOYEES AND VOLUNTEERS
- Group 2 – common law elements
  - TRUSTS/EQUITY
  - TORTS AND COMPENSATION
- Group 3 – relating to insurance
  - INSURANCE
  - BUILDING AND CONSTRUCTION
- Group 4 - litigation
  - LITIGATION AND DISPUTE RESOLUTION
- Group 5 – organisations/corporations
  - CORPORATE AND COMMERCIAL LAW
  - BANKING AND FINANCE
  - TAX
  - CHARITIES LAW
  - FUNDRAISING LAW
  - IT
- Group 6 – land and housing
  - PROPERTY LAW
  - HOUSING AND RESIDENTIAL TENANCIES
  - NATIVE TITLE
  - NEIGHBOURHOOD DISPUTES
- Group 7 – family or personal affairs
  - FAMILY LAW
  - GUARDIANSHIP AND ADMINISTRATION
  - WILLS
  - HEALTH
  - ELDER LAW
- Group 8 – relating to government
  - PUBLIC AND ADMINISTRATIVE LAW
  - PLANNING AND LOCAL GOVERNMENT
  - ENVIRONMENT
  - MIGRATION
  - INQUIRIES
- Group 9 – relating to crimes
  - FINES AND INFRINGEMENTS
  - PERSONAL SAFETY
  - CRIMINAL LAW

Group by Legal Expertise

- Group 0 - not a legal issue
  - NOT A LEGAL ISSUE
- Group 1 – corporate
  - CORPORATE AND COMMERCIAL LAW
  - BANKING AND FINANCE
  - TAX
  - CHARITIES LAW
  - FUNDRAISING LAW
- Group 2 – IP and technology
  - IT
  - PRIVACY
  - INTELLECTUAL PROPERTY
- Group 3 – property/real estate
  - PROPERTY LAW
  - HOUSING AND RESIDENTIAL TENANCIES
  - NEIGHBOURHOOD DISPUTES
- Group 4 – family law
  - FAMILY LAW
  - GUARDIANSHIP AND ADMINISTRATION
  - ELDER LAW
- Group 5 – wills and estates
  - WILLS
  - TRUSTS/EQUITY
- Group 6 – government
  - PUBLIC AND ADMINISTRATIVE LAW
  - INQUIRIES
  - MIGRATION
- Group 7 – planning and environment
  - NATIVE TITLE
  - PLANNING AND LOCAL GOVERNMENT
  - ENVIRONMENT
- Group 8 – criminal
  - CRIMINAL LAW
  - FINES AND INFRINGEMENTS
  - PERSONAL SAFETY
- Group 9 – personal injury/compensation
  - TORTS AND COMPENSATION
  - CONSUMER LAW
  - HEALTH
- Group 10: – EMPLOYEES AND VOLUNTEERS
- Group 11: – INSURANCE
- Group 12: – BUILDING AND CONSTRUCTION
- Group 13: – LITIGATION&DISPUTE RESOLUTION

Figure 2: Groupings by theme/topic and legal specialisations

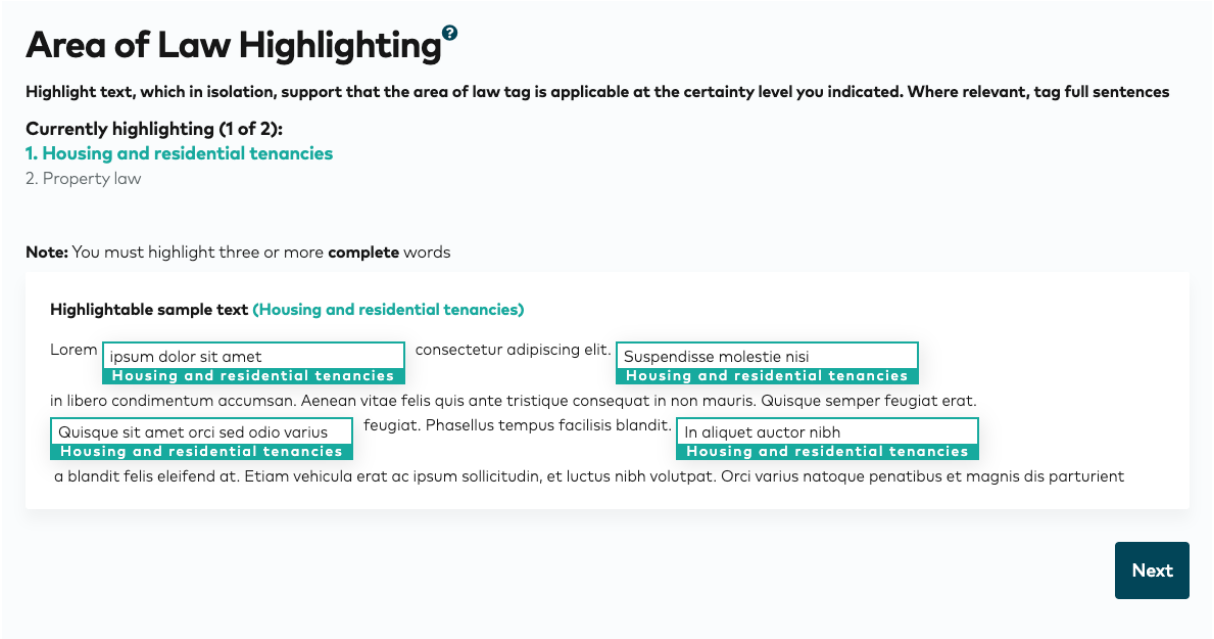

Figure 3: Step 2: Highlighting supporting text for the areas of law chosen