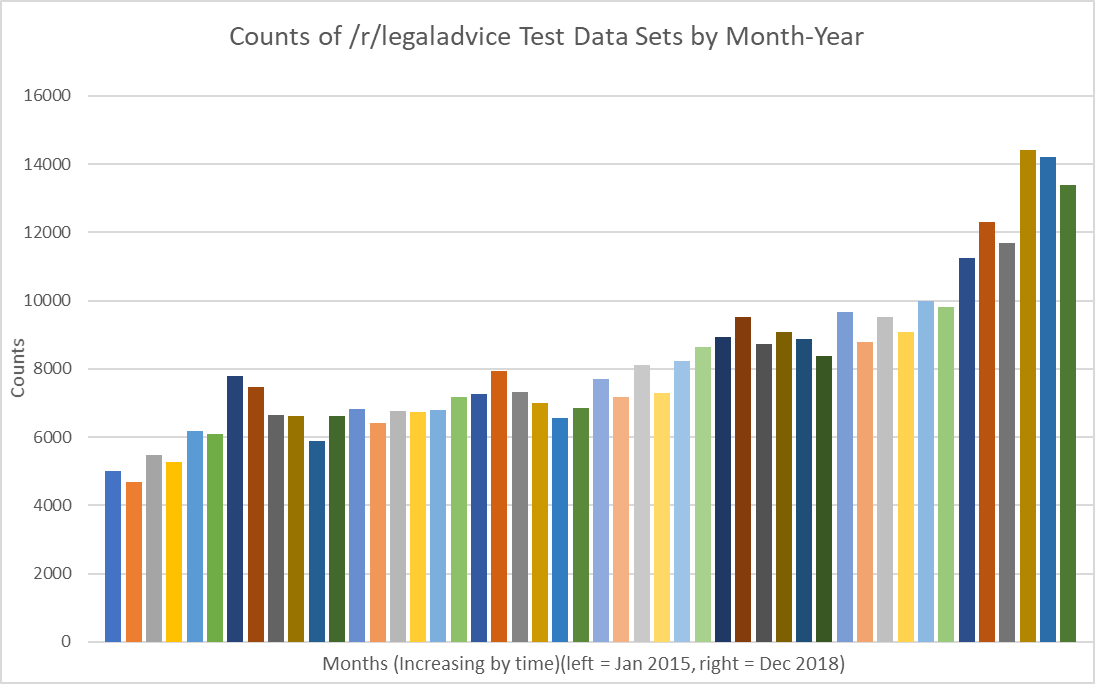
**Results Summary Doc.**

**Data Summary:**

* Training: Our training data is from the Stanford Learned Hand’s Project. More specifically in their paper report, it was stated: “our first dataset being tens of thousands of people’s problem stories from Reddit’s Legal Advice board.” Although the timeframe of those posts, were never specified (ToDo: read through whole Stanford doc again to make sure they didn’t mention anywhere, if not would be good question to ask them) they might likely to be from near the creation of their project/game so *potentially* around 2018/2019. Of that whole training set, we used their specific smaller set called ‘[2019-04-23\_best-guess\_binary.csv](https://learned-hands.github.io/project-hub/texts/2019-04-23_best-guess_binary.csv) (3,459 texts/rows, 20 categories, 35,550 labels) which contained 3,549 posts (as that is the document in this case). For their model creation, they used an 80/20% training/test split which we shall keep consistent for our earlier results but eventually change to 5-fold cross validation.

However, not all 3,549 examples could be used in our case due to lack of annotation in some of the cases where no labels were marked for that respective post. **How should we deal with poor annotations?** 🡪 approach 1: omit all rows without at least labeled category (1) and convert all Nans in that data example to 0’s so we can actually perform model fitting. (**ToDo:** **Ask how the Stanford Group handled incomplete data entries in their first model?)**

* + **Dataset used for Model after processing:** 1,826 examples (at least one 1/+ and all NaNs converted to 0’s)
  + **80/20% Training/Test Split for Stanford Data ONLY:** 
    - Created the 80/20% training/test split in the code which resulted in splits of 1,460 data examples in the training set and 366 data examples in the validation/test set.
    - These splits are used only on Stanford data to get the following information on each label: Training/True Prevalence (Prior), Accuracy, F1, and Roc\_Auc score.
* Testing:
  + We used a One versus Rest classification model using Logistic Regression on /r/legaladvice posts from Jan 2015 to Dec 2018 to obtain results on prevalence estimation metrics such as CC, PCC, and freq-e (our model of interest).
  + For our case, we trained on all of the 1,820 examples that was explained prior for the Stanford data.
  + For testing, we used posts from each month that passed our pre-processing steps. More information for pre-preprocessing can be found in the source code. The test dataset contained *516,641* posts originally, but after preprocessing it was reduced to *392,116* appropriate posts. An extended analysis of the test data is provided here:<https://github.com/slanglab/docprop_replication/blob/master/tamaspalfi/legaladvice/Legal%20Advice%20Project/input%20data%20metrics.xlsx> (\*Might be private)
  + Here is a plot showcasing the relationship between the months and # of test posts:



**Results Summary:**

**\*Note:** The columns with Avg. Pred. are averages of the predicted prevalence’s for each metric over the course of Oct 2015 to Dec 2018. Excludes the first 9 months from 2015.

(ToDo: Figure out the problem in test data at Aug 2015. Why does it usually diverge here?)

* Analysis of Aug-2015 Divergence
  + First – I tried to manually observe the posts in that month comparing it to the months prior and after it to see if anything noticeable changed in format esp. of the self\_text. After several eye observations, I did not notice anything that would cause this divergence for that month’s posts seemed quite similar to the ones before and prior.
  + Found a trend that may show the problem but can’t explain exactly why. After observing the test/inference data metrics, I found that the there was a sizable change in the number of posts expunged via pre-processing step in months following & including Sep-15 compared to those before. For example, June-15 532 posts expunged, July-15 246 posts expunged, and Aug-15 542 posts expunged compared to Sep-15 1,961 posts expunged, Oct-15 1,934 posts expunged, and Nov-15 1,591 posts expunged. This also isn’t due to change in volume of posts as July-15 had 8,031 posts Aug-15 had 8,024 posts Sep-15 had 8,603 posts and Oct-15 had 8,551 posts. Again I’m not sure how this would be causing this divergence but I have a strong inclination to think so (as it’s the only thing I’ve seen with similar effect as the divergence in plots). It could likely be due to a change in the subreddit as it got more popular to make posts more specific about what topic they are directly dealing with (?) and not be irrelevant.
  + After extensive look into reddit legal-advice posts from 2015, it seems that there was no evidence of a change in the posts there. However, after more observation there was in fact the reddit blackout on July 2 2015 which may have had a slight effect. Reddit users reacted negatively to combination of hateful subreddits getting slowed phased/banned out and firing of a popular reddit AMA employee Victoria which results in around 300 subreddits being set to private. This effectively made reddit unusable for all new members for that brief time period. It’s unlikely it has any effect on this here, the only way would be that traffic volume was very low.

**\*Note:** Some of the reported F1 scores are 0. This is due to how the predicted set (vector) on the 20% test split (cnt: 366) there are no 1/+ predictions. All the labels were predicted as 0 which resulted in an undefined metric warning.

Training metrics were computed using sklearn metrics.

My (@TamasPalfi) quick thoughts on results chart:

* After having took a look at many reddit legal advice posts in the past couple weeks, I expected the labels of Traffic & Cars, Housing, Family, Courts and Lawyers, Accidents Injuries …, Money Debt Consumers Issues to have the highest prevalence’s. The reasoning for that was how they were personal problems that would affect an every day, normal person who just wasn’t sure legally about a situation and from my bias of reading many posts. 🡪 After looking at the results this was consistent with what I anticipated with the labels of Accidents, Injuries, and Torts (Problems with Others), Money, Debt, and Consumer Issues, and Court and Lawyers performing as the three highest prevalence labels. The housing and traffic & cars labels didn’t result in prevalences as high as I anticipated but they were still in the middle range doing better than many of the other labels (that had bad training data leading to noisy classifier).

The second column is the training/true prevalence based off on all the Stanford trainings data so the 1,820 examples.  The next three columns are the acc, f1, roc based off on the 80/20% training/test split on all the Stanford training data.  The splits there were ~1,420 for training and 366 for test data.  The last four columns are the averages of prevalence predictions & difference in freq-e on the reddit legal advice data.  Thus, those are all trained on the whole Stanford set (1,820 examples) and inference/test data would be the reddit data from each month-year.

**TODO TODO TODO : ADD IN CROSS VALIDATION TO GET ACCURACY ON REDDIT DATA CLASSIFIER**

**Documentation (8/2) Closing remarks:**

* **Learned a lot about NLP this summer started from nothing.**
* **Get the cross validation accuracy on the reddit data. Try to infer any meaningful connections/insights from it as well.**
* **Look at and try to explain the difference in the Freq-e, CC, PCC in plots.**
* **Improve the Classifier.**
* **TODO Master list in book**
* **Come back and redocument everything you learned.**
* **Review everything you learned this summer 🡪 topics, papers, methods, and projects (including Henrys’ and Sirius’s)**

*---------------------Training----------------------------------------* -----------inference 2015-2018--------------------

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **Training/True Prevalence (Prior)** | **Accuracy** | **F1** | **Roc\_Auc** | **Avg. Pred. CC** | **Avg. Pred. PCC** | **Avg. Pred. Freq-e** | **Diff in Freq-e Pred. Aug-18 to Oct-15** |
| Health | 0.083 | 0.948 | 0.578 | 0.806 | 0.126 | 0.121 | 0.124 | **0.113** |
| Immigration | 0.014 | 0.981 | 0.222 | 0.837 | 0.008 | 0.041 | 0.078 | **0.111** |
| Environmental Justice | 0.004 | 0.989 | 0.000 | 0.488 | 0.002 | 0.018 | 0.075 | **0.106** |
| Disaster Relief | 0.002 | 0.997 | 0.000 | 0.175 | 0.001 | 0.027 | 0.088 | **0.106** |
| Veterans & Military | 0.002 | 0.992 | 0.000 | 0.287 | 0.001 | 0.017 | 0.083 | **0.106** |
| Education | 0.022 | 0.973 | 0.167 | 0.799 | 0.093 | 0.064 | 0.088 | **0.103** |
| Government services | 0.036 | 0.959 | 0.118 | 0.612 | 0.097 | 0.061 | 0.080 | **0.101** |
| Court and Lawyers | 0.219 | 0.801 | 0.482 | 0.708 | 0.261 | 0.261 | 0.262 | **0.085** |
| Accidents, Injuries, and Torts (Problems with Others) | 0.264 | 0.746 | 0.497 | 0.692 | 0.309 | 0.307 | 0.308 | **0.084** |
| Money, Debt, and Consumer Issues | 0.259 | 0.787 | 0.519 | 0.756 | 0.292 | 0.290 | 0.290 | **0.057** |
| Estates & Wills | 0.039 | 0.978 | 0.333 | 0.810 | 0.023 | 0.042 | 0.046 | **0.055** |
| Work and Employment | 0.171 | 0.940 | 0.800 | 0.928 | 0.205 | 0.171 | 0.172 | **0.048** |
| Small Business and IP | 0.055 | 0.964 | 0.480 | 0.872 | 0.033 | 0.061 | 0.065 | **0.047** |
| Family | 0.144 | 0.913 | 0.579 | 0.847 | 0.101 | 0.106 | 0.104 | **0.019** |
| Benefits | 0.020 | 0.981 | 0.000 | 0.820 | 0.008 | 0.012 | 0.009 | **- 0.001** |
| Civil and Human Rights | 0.073 | 0.861 | 0.164 | 0.668 | 0.043 | 0.043 | 0.043 | **- 0.007** |
| Traffic and Cars | 0.110 | 0.954 | 0.795 | 0.953 | 0.095 | 0.095 | 0.095 | **- 0.007** |
| Housing | 0.201 | 0.954 | 0.879 | 0.975 | 0.173 | 0.185 | 0.184 | **- 0.014** |
| Crime & Prisons | 0.188 | 0.852 | 0.625 | 0.854 | 0.134 | 0.134 | 0.133 | **- 0.044** |

**\*Size:** 1,820 ex.

**\*Plots…:**

* Note that the plots are ordered from increasing to decreasing on the following metric: *difference between Freq-e prevalence prediction from Aug 2018 to Oct 2015.* Chart is also ordered in this way with that metric being the last column **(bolded)**.

**Notes on different prediction methods:**

* Freq-e performs ‘**different**’ (to CC & PCC) on the following labels: Disaster Relief (pg 7) (**way different)**; Environmental Justice (pg 9) (**way different)**; Estate & Wills (pg 10); Immigration (pg 15) (**way different)**; Small Business and IP (pg 17); Veterans & Military (pg 19) (**way different)**
* Freq-e performs very similar on the following labels: Accidents, Injuries, and Torts (Problems with Others); Benefits; Civil and Human Rights; Courts and Lawyers; Crime and Prisons; Education; Family; Health; Housing; Money, Debt, and Consumer Issues; Traffic & Cars;
* CC performs better on the following labels: Government Services (pg 12); Work and Employment (pg 20) (**way different)**

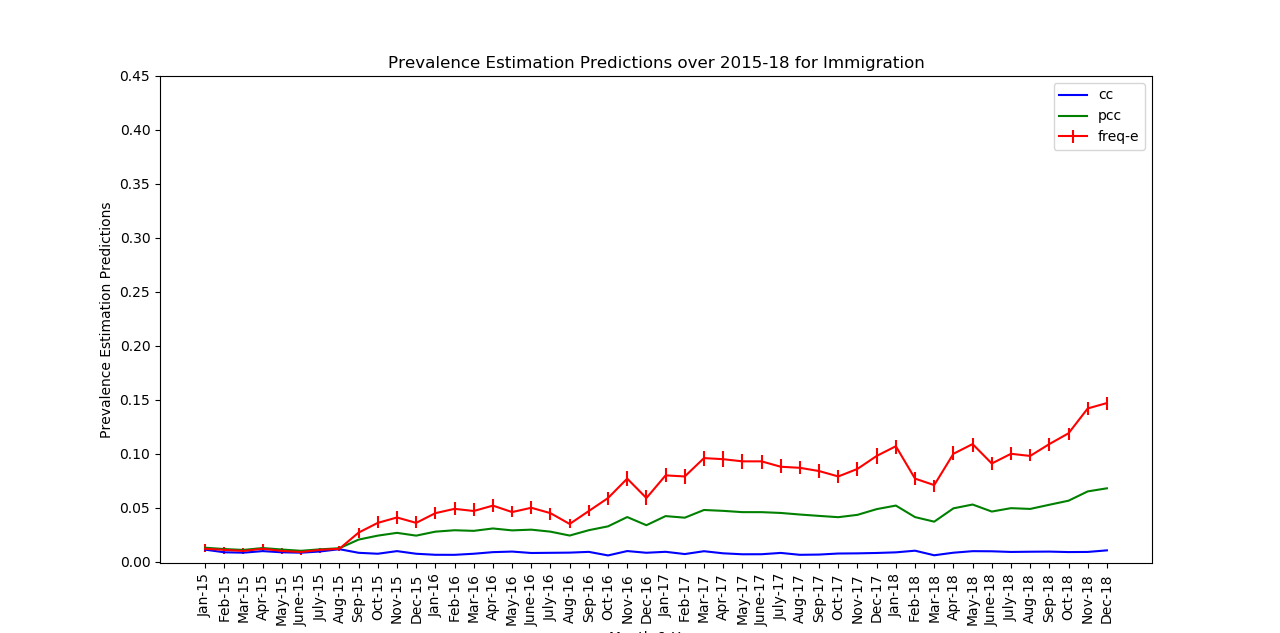
**Notes on different labels:**

* Labels that see an increase in prev. estimation: Accidents, Injuries, and Torts; Court and Lawyers; Disaster Relief (PCC & freq-e only); Education; Environmental Justice (mostly just freq-e, PCC slight shift up); Estate and Wills (PCC, freq-e); Government Services; Health; Immigration (PCC, freq-e); Money, Debt, and Consumer Issues; Small Business & IP (PCC, freq-e) (very slight shift only); Veterans and Military (big shift for freq-e, small shift for pcc); Work and Employment (CC biggest change).
* Labels with neutral change (little shift from beginning to end): Benefits; Civil and Human Rights; Family; Housing (hills/local maxima each year’s July/August); Traffic and Cars;
* Labels that see a decrease in prev. estimation: Crime and Prisons;

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.083

**

**Training Prevalence:** 0.014

*A picture containing screenshot

Description automatically generated*

**Training Prevalence:** 0.004

*A close up of a piece of paper

Description automatically generated*

**Training Prevalence:** 0.002

*A picture containing screenshot

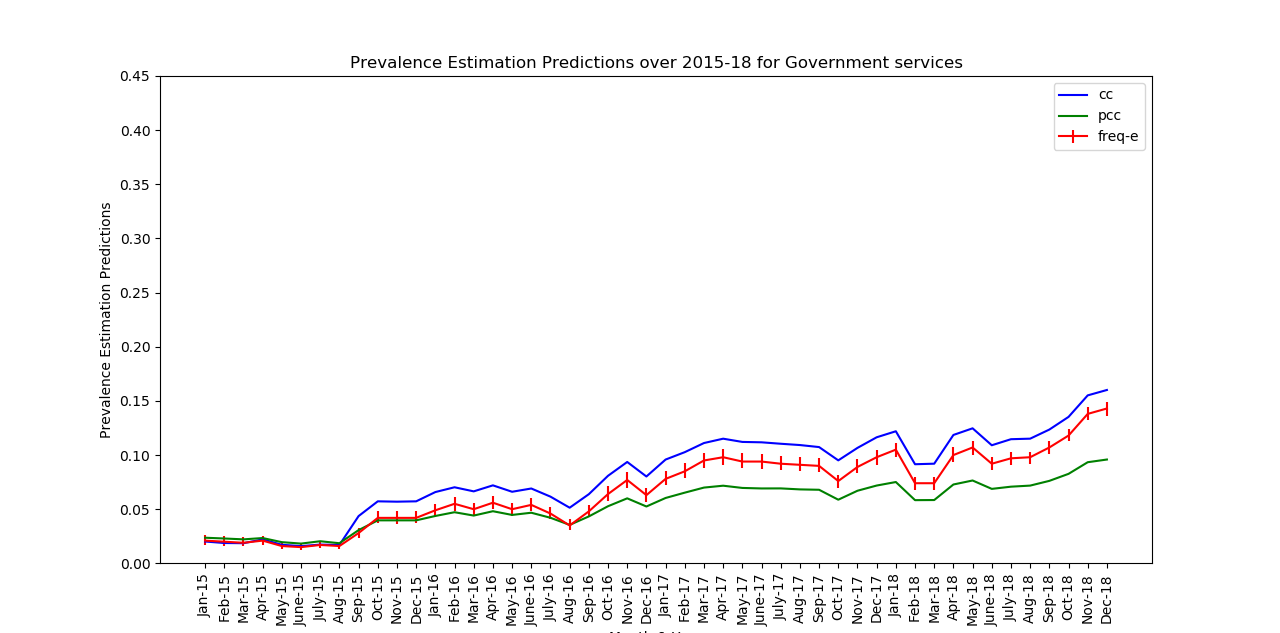
Description automatically generated*

**Training Prevalence:** 0.002

*A picture containing screenshot

Description automatically generated*

**Training Prevalence:** 0.022



**Training Prevalence:** 0.036

A screenshot of a social media post

Description automatically generated

**Training Prevalence:** 0.219

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.264

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.259

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.039

A close up of a map

Description automatically generated

**Training Prevalence:** 0.171

A picture containing screenshot

Description automatically generated

**Training Prevalence:** 0.055

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.144

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.020

0

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.073

A screenshot of a social media post

Description automatically generated

**Training Prevalence:** 0.110

A close up of a map

Description automatically generated

**Training Prevalence:** 0.201

A screenshot of a cell phone

Description automatically generated

**Training Prevalence:** 0.188

Prior Work & Questions:

* **Stanford** group wanted a better representative ML model for their work 🡪 ToDo: Haven’t really optimized model to the fullest by trying different hyperparamters. Just used what they had even excluding the second estimator of Naïve Bayes.

1. Read project description
   1. What text modeling do they use? Can we do better? (I think this is on the website?)

**Ans:** The process for them (as any research question trying to form supervised model) is (i) get labeled data (ii) text modeling (iii) final model to perform classification

1. The process used for getting the data labels is first to get annotators to ‘label’ the original data (collective group of just the text of the person’s life situation) (for best\_guess this is 3,459 examples) for each of the 20 categories needed here. The process they took for this step was different compared to normal practices (but it may have led to better annotation accuracy). They got annotators to volunteer labeling information by formatting their data as a game (<https://learnedhands.law.stanford.edu/>) and having the game being just labeling an unmarked example and saying if the category applies/was present in the person’s description. The reason that I stated that this may improve annotation accuracy is that its on volunteer basis, so they are likely taking it seriously in comparison to a person thrust into the job who may be working for speed more. They used these game results to construct a Wilson Confidence Interval and the best\_guess dataset takes the center of this interval as the label given that there’s at least one label.
2. The **text modeling** used was incorporating pre-trained word vectors via **word2vec** that created embeddings for each of the labeled texts and combine that with an ensemble model that was trained on the labeled data to perform simple binary classification for each of the available labels.
3. The final model used to perform their classification metrics such as acc, prec., recall, and F1 scores were computed using two estimators: a LogReg classifier & Gaussian NB model. Here are the parameters they used:

Model = "VotingClassifier(estimators=[('logistic', LogisticRegression(C=1000000000.0, class\_weight='balanced', dual=False,

fit\_intercept=False, intercept\_scaling=1, max\_iter=100,

multi\_class='warn', n\_jobs=None, penalty='l2', random\_state=None,

solver='warn', tol=0.0001, verbose=0, warm\_start=Fa... verbose=0, warm\_start=False)), ('GaussianNB', GaussianNB(priors=None, var\_smoothing=1e-09))],

flatten\_transform=None, n\_jobs=None, voting='soft',

weights=[1, 1, 1, 1, 1])".

**Note:** I only used LogReg as an estimator so far as sklearn’s OVR classifier takes in as a param only one estimator (ToDo: but attributes it says ‘estimators’??)

1. Train binary one-versus-rest text classifier on the data.
   * 1. What accuracy and F1 do you get? How does this differ for each label?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label/Category** | **Their Acc** | **My Acc** | **Their F1** | **My F1** |
| Accidents, Injuries, &Torts | 0.758112 | 0.7923497267759563 | 0.613208 | 0.638095238095238 |
| Money, Debt, & Consumer Issues | 0.795252 | 0.7814207650273224 | 0.682028 | 0.49367088607594933 |
| Courts & Lawyers | 0.798220 | 0.7568306010928961 | 0.626374 | 0.42580645161290326 |
| Housing | 0.961957 | 0.9371584699453552 | 0.900000 | 0.8188976377952756 |
| Crime & Prisons | 0.898256 | 0.855191256830601 | 0.740741 | 0.6344827586206897 |
| Work & Employment | 0.934466 | 0.9398907103825137 | 0.784000 | 0.7659574468085106 |
| Family | 0.916010 | 0.8989071038251366 | 0.666667 | 0.6185567010309277 |
| Traffic & Cars | 0.953608 | 0.953551912568306 | 0.808511 | 0.7384615384615385 |
| Health | 0.946835 | 0.9316939890710383 | 0.618182 | 0.4680851063829786 |
| Estate & Wills | 0.970326 | 0.9754098360655737 | 0.736842 | 0.5263157894736842 |

**NOTE:** My accuracy and F1 results seems to be slightly lower than their predictions. This is likely due to them using two estimators for their collective model while I only relied on LogReg.

Problem with data representation due to poor annotation (middle/later on of document seem to be more from random people actually playing game – few correct annotations) compared to beginning and end which were likely (?) directly annotated from the researchers at LearnedHands.

**NOTE:** Their (CC?) prev. estimation is just the proportion of affirmative examples they found in the test data based on each label. It is the third statistic in the page for <https://learned-hands.github.io/project-hub/models.html>

**\*NOTE:** Small differences in results between their prediction and mine is likely due to different train/test splits and the fact that they used two estimators (LogReg & GaussianNB) while I only utilized LogReg.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Label** | **True Prevalence** | **Their (CC?) prev. Est.** | **My CC prev. Est.** | **My PCC** | **My Freq-e prev. Est.** | **Absolute Difference from True Prev. & Freq-e point prediction** |
| Accidents, Injuries, &Torts | 0.28688524590163933 | 0.27 | 0.27049180327868855 | 0.27159703580589734 | {'point': 0.272, 'conf\_interval': (0.228, 0.318), 'conf\_level': 0.95} | 0.01488524590163931 |
| Money, Debt, & Consumer Issues | 0.2650273224043716 | 0.26 | 0.2459016393442623 | 0.24642387071851676 | {'point': 0.246, 'conf\_interval': (0.20400000000000001, 0.292), 'conf\_level': 0.95} | 0.019027322404371605 |
| Courts & Lawyers | 0.21311475409836064 | 0.23 | 0.17486338797814208 | 0.1728353119899963 | {'point': 0.17200000000000001, 'conf\_interval': (0.136, 0.213), 'conf\_level': 0.95} | 0.04111475409836063 |
| Housing | 0.1994535519125683 | 0.18 | 0.17486338797814208 | 0.17610835354661566 | {'point': 0.176, 'conf\_interval': (0.139, 0.217), 'conf\_level': 0.95} | 0.023453551912568316 |
| Crime & Prisons | 0.23224043715846995 | 0.17 | 0.15027322404371585 | 0.14924250802395023 | {'point': 0.149, 'conf\_interval': (0.115, 0.188), 'conf\_level': 0.95} | 0.08324043715846996 |
| Work & Employment | 0.17486338797814208 | 0.15 | 0.1448087431693989 | 0.1479453590580593 | {'point': 0.147, 'conf\_interval': (0.113, 0.186), 'conf\_level': 0.95} | 0.027863387978142085 |
| Family | 0.12021857923497267 | 0.11 | 0.1284153005464481 | 0.12982197410820012 | {'point': 0.129, 'conf\_interval': (0.097, 0.166), 'conf\_level': 0.95} | 0.008781420765027331 |
| Traffic & Cars | 0.12021857923497267 | 0.13 | 0.10382513661202186 | 0.10453593767301539 | {'point': 0.105, 'conf\_interval': (0.076, 0.138), 'conf\_level': 0.95} | 0.015218579234972676 |
| Health | 0.05737704918032787 | 0.08 | 0.04371584699453552 | 0.04306344558330696 | {'point': 0.069, 'conf\_interval': (0.046, 0.099), 'conf\_level': 0.95} | 0.015377049180327867 |
| Estate & Wills | 0.040983606557377046 | 0.06 | 0.01912568306010929 | 0.020595435646766205 | {'point': 0.02, 'conf\_interval': (0.009000000000000001, 0.038), 'conf\_level': 0.95} | 0.020983606557377046 |

* LearnedHands only did the modeling for 10 of the labels??? 🡪 Why is this?
  + **NO explanation** in long paper 🡪 My thoughts: They got results with poor F1 scores (like I did) and omitted those categories for the final model results. None of the low F1 scores I observed have the corresponding category included for model results, only the ones that performed reasonably well. 🡪 \*\*\*I have added results for each category possible (except Native American & Tribal Law which had to be cut out due to no training examples being labeled 1/+) in updated results at the beginning of this document.
  + After looking at metrics more it seems that they did just omit categories that didn’t have many positive/1 labels 🡪 ones that had a very low training/true prevalence.
* **Data:**
  + <https://learned-hands.github.io/project-hub/data.html>
* **Stanford Learned Hands project description:**
  + Short
    - <https://docs.google.com/document/d/1UAcghpLivHrAit7XUH4Df31GhRQXJ5EP46UZi6DJYAU/edit?usp=drivesdk>
  + Long
    - <https://docs.google.com/document/d/1PPfbA-RFN3qaXWRopE0It2jWnaiqtKsOv8V0fULHDdM/edit?usp=drivesdk>