Predicting Seasonality of Legal Needs

Hee Hwang and **Brendan O'Connor**
College of Information and Computer Sciences
University of Massachusetts Amherst
{hhwang, brenocon}@cs.umass.edu

Abstract

People come in contact with the law in a variety of ways, whether concerning finances, crime, employment, housing, or immigration. Some of these encounters are quickly resolved, while others require sustained engagement with the legal system. We focus on these unmet needs in the online legal community. In this paper, we present a classifier and prevalence estimation to predict the seasonality and impact of natural disasters on the legal community. Our results show that there is an identifiable seasonal pattern concerning the category of housing law. This study also shows the impact of natural disasters (such as the COVID-19 pandemic) on the legal systems that govern work and employment.

1 Introduction

1.1 Motivation

A significant issue is that people are often unable to identify a problem in their lives as a legal issue. In the case that a person can locate their issue as a legal one, often they do not know what kind of legal issue it may be. Lastly, they do not know how to deal with it. These unmet needs are increasing, and more people are looking for legal help. We are interested in how these needs, and other useful information, might be identified from an online community, specifically, a subreddit.

Figure 1 shows the accumulation of submissions to a legal advice subreddit. What does Reddit's legal advice data tell us about current trends in how people are seeking and identifying forms of legal assistance? What legal needs can we describe and infer through analyzing texts in an online community? People post online because they have an issue, whether it be legal or not. Some legal experts can help them and eventually resolve their issues. In current practices, a lawyer or legal advisor may not be able to identify the category of the legal problem that a person is having until they meet or speak with them. My research is to analyze and reveal the consumers' needs to connect the people who need help with those who can help them. The consumer will get more support, and the provider will get more insight into the consumer's needs. A large number of unmet legal needs exist and must be resolved. Understanding and predicting these needs will benefit both legal service consumers and providers.

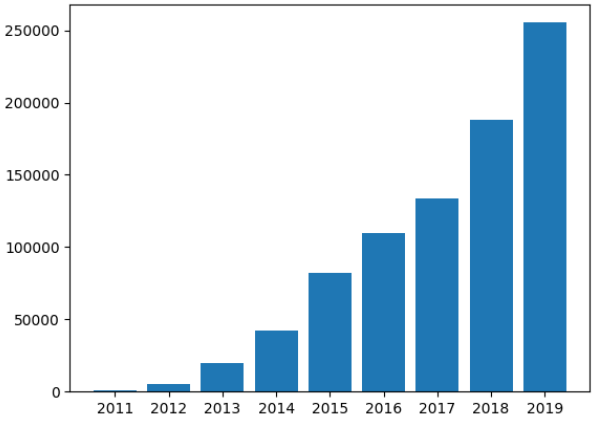


Figure 1: Number of submissions in legal advice subreddit

1.2 Problem Definition and approach

Our goal is to infer what legal needs look like and how legal trends change over time. To do this, we take a look at the prevalence of legal needs according to various legal categories. After, we analyze seasonality, which concerns recurring events throughout the year. Lastly, we measure the impact of the COVID-19 pandemic.

1.3 Our Approach

We build a legal category classifier and applied this classifier to Reddit legal advice data. We use uncertainty awareness generative models((Keith

and O'Connor, 2018) to estimate prevalence for seasonality and natural disaster. In short, we estimate prevalence of certain legal categories using the probability and prior prevalence obtained from the legal hands classification.

2 Data

2.1 NSMI v2[1]

Figure 2: Legal Taxonomy

The National Subject Matter Index(NSMI) v2 is a legal taxonomy classified by legal experts. The object of this classification is to provide an understandable and general way to categorize legal terms. Currently, we have 20 categories, including public benefit, work and employment law, health, and housing. Each group has its subclasses resulting in more than 300 subclasses.

2.2 Learned Hands[2]

Researchers from Stanford and Suffolk university created a game called Learned Hands (Hagan and Colarusso, unpublished). The game presents a text from the legal community and lawyers, law students, and others label the issue into a legal category when they reach an agreement about how to categorize them. The legal category is called NSMI v2(taxonomy.legal), which contains twenty types, including work, housing, crime, court, traffic, accidents. These categories have subterms which determine the actual legal cases defined by law experts. From this game, the extracted data contains 2777 labeled Reddit submissions collected from 7/8/2017 to 1/20/2018[3]. We use this as training data and conduct an exploratory analysis. [4]

[1] taxonomy.legal

[2] https://learnedhands.law.stanford.edu/

[3] https://spot.suffolklitlab.org/

[4] https://github.com/heeh/legal_advice/blob/master/EDA.ipynb

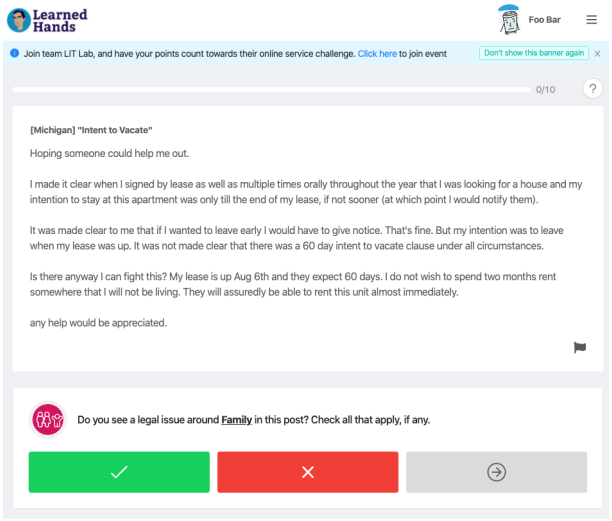
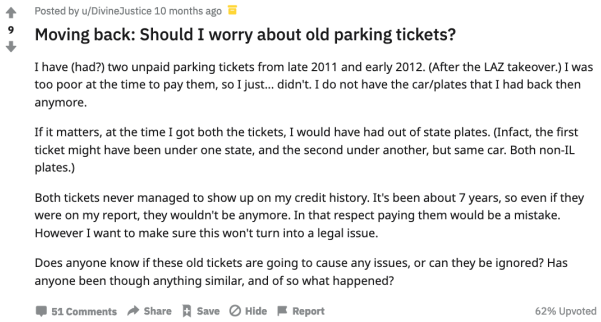


Figure 3: Number of submissions in legal advice subreddit

2.3 Reddit Legal Advice[5]



Legal advice subreddit is an online community where people can post their issues, which may or may not be legal issues. People submit their issues, location, and other users who have had similar experiences offer their advice. They are not necessarily legal experts, and we do not know if their answers are valid or not. The first submission was on April 20, 2010, and currently, there are more than 900,000 submissions. To retrieve these, we construct a downloder[6] using Pushshift Reddit API.[7]. The API allows a user to scrap 1000 submissions per request and 200 requests per minite. For example, the following API call retrieves 1000 submissions with specified fields.

```
import requests
url = "https://api.pushshift.io/reddit/search/ \
       submission/?subreddit=legaladvice&      \
       fields=id,created_utc,title,selftext&   \
       size=1000&after=2020-01-01"
subs = requests.get(url)
```

[5] https://www.reddit.com/r/legaladvice/

[6] https://github.com/heeh/subreddit_downloader/

[7] https://github.com/pushshift/api

3 Classifier

Currently, there are more than a hundred classes, and eventually, We end up with sixteen legal categories that give a consistent result. Given sixteen legal categories, we build seperate classifiers for each legal category. Each classifier predicts whether given documents are related to a certain legal category. We test TF-IDF and GloVe vector representation and decide the TF-IDF that uses the logistic regression with cross entropy loss and L1 regularization.

$$score(\lambda) = loss(\mathbf{x}^{de}, \mathbf{y}^{de}, \hat{\theta})$$
$$\hat{\theta} = \arg\max_{\theta} \log P_{\theta}(\mathbf{y}^{tr} | \mathbf{x}^{tr}) - \lambda|\theta|$$

These are the performance of classifiers. [8]

Classifier	Acc.	Prec.	Rec.	F1
TF-IDF L1	0.97	0.52	0.41	0.46
TF-IDF L2	0.97	0.55	0.22	0.28
GloVe(50) L1	0.93	0.25	0.54	0.32
GloVe(50) L2	0.92	0.24	0.56	0.31
GloVe(300)L1	0.96	0.37	0.52	0.42
GloVe(300)L2	0.97	0.40	0.51	0.44

First, we classify each submission into the sixteen legal categories. We used a simple TF-IDF logistic regression model with a grid search of the power of two. We conduct ten-fold validation on the Learned Hands data and apply this model to predict the legal category of each submission.
For example, given the previous parking ticket submission, the classifier produces following results. In short, this issue is related to traffic, court, and money.

Legal Class	Probability(%)
TR-00-00-00-00	95.61
CO-00-00-00-00	64.52
MO-00-00-00-00	37.11
BU-00-00-00-00	4.86

4 Prevalence Estimation

Prevalence estimation is a proportion of a specific class in a given time interval. For example, we want to know how prevalent the work and legal employment category is in March 2020. We use freq-e, an uncertainty awareness prediction model that utilizes the probability of a legal class, and prior prevalence of training data.

[8] https://github.com/heeh/legal_advice/blob/master/comparison.ipynb

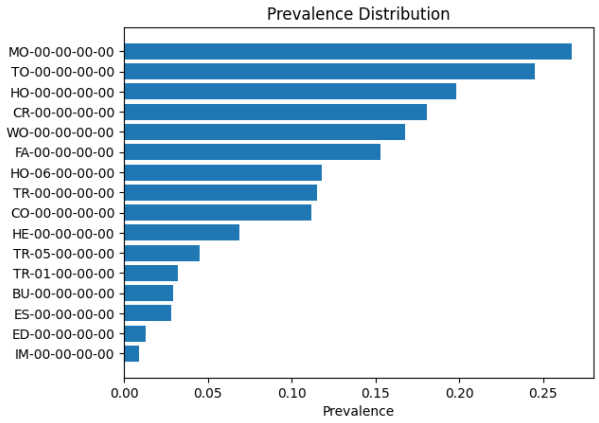


Figure 4: Prevalence Estimation(2011 - 2020)

5 Seasonality[9]

We see a cyclic pattern in the prevalence and model the pattern. To be specific, we consider trend, which is a linear function that keeps increasing and seasonality, a recurring pattern throught a year. To model this, we use following equation where t is a time step and $B^{(month)}$ a coefficient vector per month.

$$y_t = eps + B0 + C * t + (\mathbf{B}^{(month)})^T \mathbf{X}_t^{(month)}$$

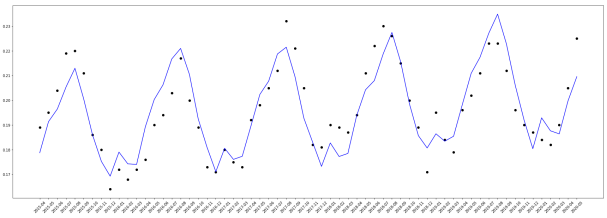


Figure 5: Seasonality Prediction(Housing, 2015-2020)

Coefficients	Values
C	0.0002887
B1	0.0038113
B2	-0.0017274
B3	-0.0032661
B4	0.0098276
B5	0.0193388
B6	0.0270501
B7	0.0369613
B8	0.0428726
B9	0.0307838
B10	0.0138951
B11	0.0012063
B12	-0.0083999
B0	0.1723536

[9] https://github.com/heeh/legal_advice/blob/master/infer_prev.ipynb

Mean absolute error: 0.0065392593
Mean squared error: 0.0000587545
Coefficient of determination: 0.80

6 Effect of Natural Disaster[10]

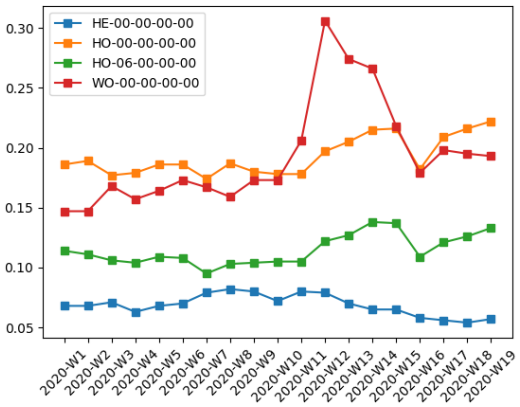


Figure 6: Prevalence Estimation(2020)

The following is the prevalence of the housing legal category from 2011 to 2020. The annual cyclic pattern is observable, and this shows that legal issues increase during the spring and summer while they are decreasing in fall and winter. Given this trend, we built a regression model for time trends. The second diagram shows the prevalence of work and employment law from 2011 to 2020. From this data, this category skyrocketed in march, and we found that the COVID-19 pandemic became a popular term. From this, we can see the relationship between the drastic change and the cause of the issue.

7 Conclusion

To understand how legal issues are represented in online community forums, we build a classifier and estimate prevalence. We see the seasonality of a legal trend and predict the prevalence of it. We also want to understand how those needs change during the COVID-19 pandemic and how the COVID-19 pandemic impacts the legal community. Thus, inferencing the legal trend will be helpful for providers to understand their consumers' needs better, and for their consumers to receive the help they need.

[10] https://github.com/heeh/legal_advice/blob/master/predictor_small.ipynb

8 Future Work

- Classifier using definition at taxonomy.legal
- Classifier using Reddit Flair
- Demographics of people by geographical location
- Prevalence with COVID-19 pandemic related text
- Major topic change within a certain a legal class

Acknowledgments

Thank you so much for all your support and for advising, Professor O'Connor. It was extremely challenging and rewarding.

References

Margaret Hagan and David Colarusso. unpublished. You say potato, we say legal issue: Adapting a digital epidemiology approach to access to justice. Unpublished.

Katherine A. Keith and Brendan O'Connor. 2018. Uncertainty-aware generative models for inferring document class prevalence. In *Proceedings of EMNLP*.