NYU Center for Data Science and Data Science Justice Collaboratory

1 Introduction

Powerful AI systems will need to be able to reason about moral and philosophical problems and ethical theories. Can AI systems learn philosophical reasoning from a diverse corpus of human text and dialogue? This would be especially helpful in domains where human values are ambiguous or underdetermined. Little work has been done to make these questions tractable.

The following are proposed as NYU capstone or natural language understanding projects.

2 Automatic legal case summaries

In this work students will focus on releasing a new benchmark data set focused on legal summaries. The work will be primarily focused on curating the data set and building baseline models to evaluate the task difficulty. Evaluation will be done using BLEU and other similar metrics leveraging SOTA Natural Language Generation (NLG) techniques to generate line by line predictions of case summaries.

1. Academic publication potential - build a task from the legal domain that could be integrated into Sam Bowman's SuperGlue benchmark or just create a separate one (the task could also be something other than summarization).
2. SOTA for summarization appears to be
   (a) https://www.aclweb.org/anthology/P19-1101
   (b) http://super-ms.mit.edu/rum-final-version.pdf

3 "Textio" for judges to monitor opinion writing

In this work students will focus their efforts on whether judicial writing can be assisted in real-time, whether to reduce error, reversal rates, or increase citation impact.

4 Transcriptions of oral arguments and a second model to do the alignment automatically. [15 years are manually aligned but another 50 years are not]

In this work students will work with the audio recordings of 50 years of oral arguments from US courts to build the largest multi-modal legal data set. The two primary tasks involved in this work will be to use a

SOTA model to convert the recordings to textual form and then subsequently train an additional model to align the converted text recording to the manually transcribed oral argument. Upon completing this work a data set consisting of 65 years of aligned and transcribed oral arguments will be generated. This data set will allow further research on understanding how tonality and other auditory characteristics of oral arguments allow researchers to better predict case outcomes at multiple levels of the US legal system.

1. Academic publication potential - challenging from a technical perspective and is impactful from a legal angle.
2. SOTA appears to be
   (a) Facebook Automatic Speech Recognition Toolkit
   (b) https://github.com/didi/delta
   (c) http://web.sas.upenn.edu/phonetics-lab/facilities/

5 Predicting court cases in real-time based upon audio of judges and maybe an image of the judge?

Students working on this project will work with oral arguments of the US Supreme and Circuit Courts to predict the final ruling using audio recordings, transcribed text, and judge information. The initial goal of this work will be to determine the upper bound of a model's predictive power using the complete oral argument information, but subsequently research will focus on what percent of the oral argument is required to predict the court ruling. Additionally, ablation studies will focus on interpreting what features determine predictions and whether models are utilizing spurious correlations to make predictions.

1. Academic publication potential - https://news.umich.edu/lie-detecting-software-uses-real-court-case-data/

6 Probe whether BERT understands legal information or is just using spurious cues to make predictions

In this work students will focus on understanding whether the BERT architecture actually understands legal documents or is simply using spurious relationships in the document to make predictions. The first step in this work will be to develop and evaluation method that utilizes features of the legal documents. Possible avenues for such a metric may be next sentence prediction or next word prediction.

1. https://bheinzerling.github.io/post/clever-hans/

7 Use documents and citations and seating graph to embed the judges, integrating other pieces of information like who was president and layer on additional data, which we could then use to cluster them

In this work students will attempt to build a graph based clustering algorithm to understand the interactions between various US courts, the judges that preside over them, and any other related data. Attempts at embedding temporal aspects into this graph, e.g., a judge presiding over different courts for different time periods may enable a temporal aspect to embedding into the graph embedding.

1. PyTorch-BigGraph

8 Originalism and doctrine platform

In this work students will attempt to assist judges and legal scholars in determining the "original meaning" of a concept. Auxiliary work intends to increase access to justice by offering users the possibility of highlighting a part of an argument in a case and then showing all the related cases.

1. Use case - someone can highlight a part of the court opinion and trace its original meaning or see the other cases that use similar arguments
2. SOTA appears to be https://blog.doctrine.fr/structuring-legal-documents-with-deep-learning/

9 Retrospective legal "clinical trial" platform

1. Use case - the random assignment of judges to create a 'retrospective clinical trial' - which relies on creating predictions of the court opinion. [The intuition is that you can use whether a case is assigned to a harsh or lenient judge (predicted to be harsh or lenient) to look at causal effects of getting that decision].

10 Datasets

1. U.S. Circuit Courts
   (a) Digitized universe since 1891 of all 380,000 cases, 1 million judge votes, across 94 hand-labeled legal topics, engineered into 2 billion N-grams of length eight, and 5 million citation edges across cases.

(b) This is merged with the 268 judges who served during this time period, 250 biographical features, a 5% random sample with 400 hand-labeled features (330-paged codebook), and 6000 cases handcoded for meaning in 25 polarized legal areas.

(c) Also merged with administrative data (date of key milestones, e.g., oral arguments, when was the last brief filed, etc.), publicly available U.S. Supreme Court datasets, U.S. District Court datasets, geocoded judge seats, biographies of judicial clerks, and oral arguments' audio files.

(d) The identities of randomly assigned judges sitting on 3-judge panels (who is authoring the opinions, writing dissents, or writing concurrences) render a random seating network among the judges.

(e) 25 polarized legal areas have in addition been collected and hand-coded: sexual harassment, eminent domain, free speech, abortion, church-state separation, affirmative action, gay rights, disability rights, campaign finance, capital punishment, criminal appeals, desegregation, sex discrimination, punitive damages, federalism, National Labor Review Board, environmental protection, National Environmental Policy Act, Federal Communications Commission, Title VII, First Amendment, Eleventh Amendment, standing, contracts, and corporate veil piercing.

2. U.S. District Courts

(a) Digitized universe of millions of criminal sentencing decisions across 94 U.S. District Courts from 1991 (with randomly assigned judges), hand-labeled biographical data of judges, and 83-paged codebook from the U.S. Sentencing Commission.

(b) Linkages to judge identity were obtained (not publicly available) and hand-labeled biographical data of judges incorporated. Data linkages have been made to daily weather and local sporting events.

(c) Text of opinions are available since 1923.

3. U.S. Supreme Court

(a) Digitized speech patterns in oral arguments since 1955–longitudinal data on speech intonation (linguistic turns) are rare.

(b) Linked to hand-labeled oral advocates' biographies, lawyers' faces, clipped identical introductory sentences*, ratings of their traits, and publicly available U.S. Supreme Court databases containing dozens of additional features and preceding U.S. Circuit Court data.

(c) *Lawyers always use the exact same sentence when they introduce themselves to the Supreme Court: "Mr. Chief Justice, (and) may it please the Court." We have clipped this data for 1955-2013 comprising over 8000 audio recordings, spoken by many different lawyers over time. Mechanical turk workers have rated 1999-2013 sample (2000+ recordings) based on whether

they sound "confident", "trustworthy", "attractive", “masculine”, etc. We also have data on the Mturk workers, the Supreme Court cases, and the Supreme Court oral advocates (including their faces).

(d) Actual analysis of speech patterns is statistically challenging, since speech is modified dynamically. A common measure for variation in speech patterns considers resonances of vowel sounds. In order to properly measure these, the starting locations of all distinct vowel sounds have been manually flagged. An algorithm then measured vowel resonances and assigned to each vowel sound a multidimensional continuous quantity. Therefore, the size of the oral data set is much larger than the size of the underlying text.

(e) Text is traditionally treated with discrete models. Speech measurements (for example resonances) by contrast are continuous.

4. U.S. State Supreme Courts

(a) Digitized universe for 1947-1994 (roughly 400,000 cases), identities of judges sitting on the panels, hand-labeled biographies, citation network, and original text.

(b) Some of these judges run for election.

5. EOIR Immigration Courts

(a) Digitized universe of administrative data on 1 million refugee asylum and 15 million hearing sessions and their time of day across 50 courthouses and 20 years (with randomly assigned judges), hand-labeled biographical data of judges, and dozens of features on the case and the defendant.

(b) We know when the asylum case was assigned, whether the hearing was an individual hearing or whether multiple individuals were scheduled in the same session, how many cases were scheduled for sessions during a day for that judge, whether this was an in person hearing or by audio or video, whether it was a written or oral order, whether there are other related applications for relief filed by the individual and the judge's ruling on each, ethnicity of the applicant, the reason for the case and the judge.

(c) Data linkages have been made to daily weather and local sporting events.

6. Vertical linkages from arrest to final sentence

(a) Digitized universe of individuals in a district attorney's office over a decade with many stages of random assignment (screener, federal prosecutor, and judge). New Orleans is the largest city and metropolitan area in the state of Louisiana. The Orleans Parish District Attorney's Office and its prosecuting attorneys are responsible for enforcing state criminal laws and local ordinances to protect and serve the citizens of New Orleans and surrounding areas.

(b) The current data set is from 1988 to 1999 and provides detailed information on approximately 430,000 charges and 280,000 cases (involving 145,000 defendants) filed or adjudicated during this timeframe. The data collected also contains detailed information regarding each individual offender, such as social security number and the corresponding prosecutor and judge.

(c) Linkages have been made to voting records, bankruptcy, and foreclosure. The dataset is rare: vertical linkages from the time of arrest, including those sent home without a trial, otherwise do not exist. There is a 594-paged codebook.

7. World War I British courts martial

   (a) Digitized World War I British archival datasets, including universe of deserters (including names and often their birthplace) reported in military diaries, police gazettes, and handwritten military trials, commuted and executed capital sentences (which historians believe was random), geocoded casualties, maps, officer lists, and order of battle.

8. US Medicare

   (a) Linked administrative Medicare data to industry-physician relationships cleaned from litigation settlements and through the Affordable Care Act, comprising 30 million payments including the date of payment and affiliated drug code.

   (b) Linked to biographical characteristics of physician and some patient demographics.

9. Other datasets include judges' financial disclosures, many social and economic datasets, and other cleaned legal datasets:

   (a) Corporate Filings: The text of disclosures, contracts, and charters submitted to the SEC by publicly listed companies for 1996-2016.

   (b) Federal and State Legislation: We have the full history of federal and state laws enacted up until 2012.

   (c) UN Parallel Texts and Hong Kong Laws.

   (d) Party Manifestos Corpus: Large corpus of political party manifestos across the world, with rich metadata.

   (e) U.S. Congressional Record: All speeches by congressman and senator for 1880-2015, with metadata.

   (f) Teacher's union contracts: Large corpus of 6000 teachers contracts for all districts in the state of Michigan since 2010.

10. Over 1000 legal databases tagged and linked including all federal (supreme, appellate, district, bankruptcy, tax, patent, trade, customs, claims, unpublished) and state (supreme, appellate, district, tax, chancery, family, labor, unpublished) court cases to the earliest available date (some as early as 1778).

    (a) Types of databases include code, statutes, bills, regulations, bulletins and notices, commission decisions, Attorney General opinions, rulings, statements, opinion letters, bill tracking, workers' compensation decisions, municipal codes, physician discipline decisions, market conduct examinations, issuances, directives, public health reports, FTC, IRS, EEOC, Department of Labor, Department of Defense, EPA, SEC, Federal Reserve, contract appeals decisions, legislative service, manuals, etc.

11. Case records collected from 24 High Courts and 3000 subordinate courts in India with details on over 8.7 million case records and 67 million hearings. Study (1) impact of court functioning on economic growth and inequality, (2) impact of economics, political, or psychological factors on court outcomes, (3) impact of court decisions or precedents on individuals' outcomes, and (4) artificial intelligence applications.

12. Similar datasets for Kenya, Philippines, Chile, and Croatia.