Language Models for Law and Social Science

ETH Zurich, Spring 2024

1. Introduction

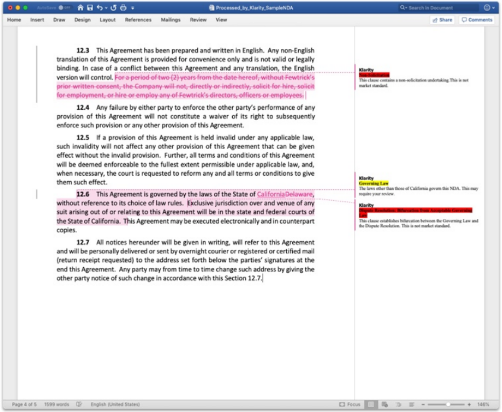
Klarity reviews NDAs under commercial market standard.

👌 Klarity highlights standard language in green.

👀 Language that requires your attention is in yellow.

😡 Non-market standard language and red-flags are in red.

Language that is not marked is boilerplate and doesn't deserve your attention.



Processed_by_Klarity_SampleNDA

Home Insert Draw Design Layout References Mailings Review View

Share Comments

12.3 This Agreement has been prepared and written in English. Any non-English translation of this Agreement is provided for convenience only and is not valid or legally binding. In case of a conflict between this Agreement and any translation, the English version will control. ~~For a period of two (2) years from the date hereof, without Fewtrick's prior written consent, the Company will not, directly or indirectly, solicit for hire, solicit for employment, or hire or employ any of Fewtrick's directors, officers or employees.~~

12.4 Any failure by either party to enforce the other party's performance of any provision of this Agreement will not constitute a waiver of its right to subsequently enforce such provision or any other provision of this Agreement.

12.5 If a provision of this Agreement is held invalid under any applicable law, such invalidity will not affect any other provision of this Agreement that can be given effect without the invalid provision. Further, all terms and conditions of this Agreement will be deemed enforceable to the fullest extent permissible under applicable law, and, when necessary, the court is requested to reform any and all terms or conditions to give them such effect.

12.6 This Agreement is governed by the laws of the State of CaliforniaDelaware, without reference to its choice of law rules. Exclusive jurisdiction over and venue of any suit arising out of or relating to this Agreement will be in the state and federal courts of the State of California. This Agreement may be executed electronically and in counterpart copies.

12.7 All notices hereunder will be given in writing, will refer to this Agreement and will be personally delivered or sent by overnight courier or registered or certified mail (return receipt requested) to the address set forth below the parties' signatures at the end this Agreement. Any party may from time to time change such address by giving the other party notice of such change in accordance with this Section 12.7.

Klarity
Non-Solicitation
This clause contains a non-solicitation undertaking.This is not market standard.

Klarity
Governing Law
The laws other than those of California govern this NDA. This may require your review.

Klarity
Dispute Resolution: Bifurcation from Acceptable Governing Law
This clause establishes bifurcation between the Governing Law and the Dispute Resolution. This is not market standard.

Page 4 of 5 1599 words English (United States) Focus 146%

The World's First Robot Lawyer

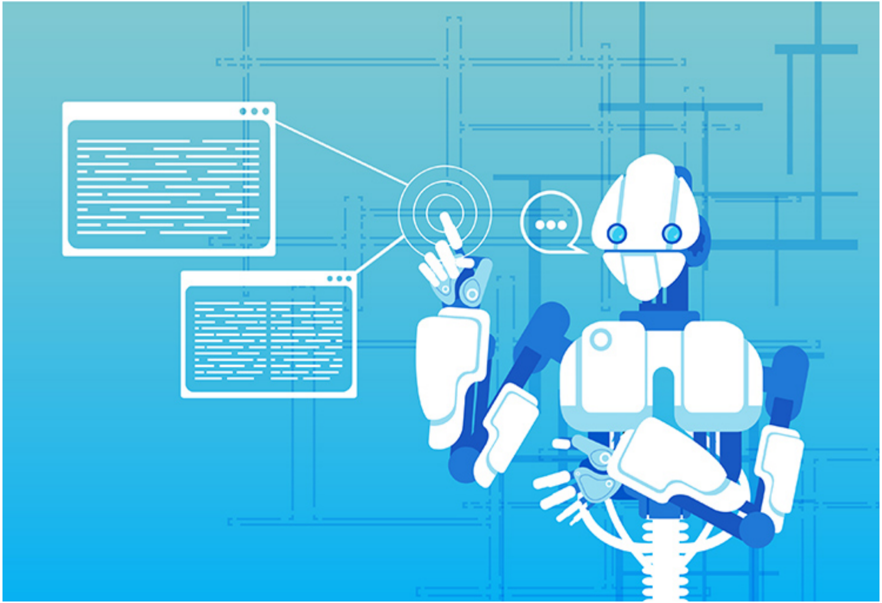
The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

Sign Up/Login

THINGS YOU CAN DO WITH DONOTPAY

- Fight Corporations
- Beat Bureaucracy
- Find Hidden Money
- Sue Anyone
- Automatically Cancel Your Free Trials

Your Court-Appointed Chatbot – Is Artificial Intelligence Threatening the Legal Profession?



https://www.cnn.com › 2023/01/26 › tech › chatgpt-passe...

ChatGPT passes exams from law and business schools - CNN

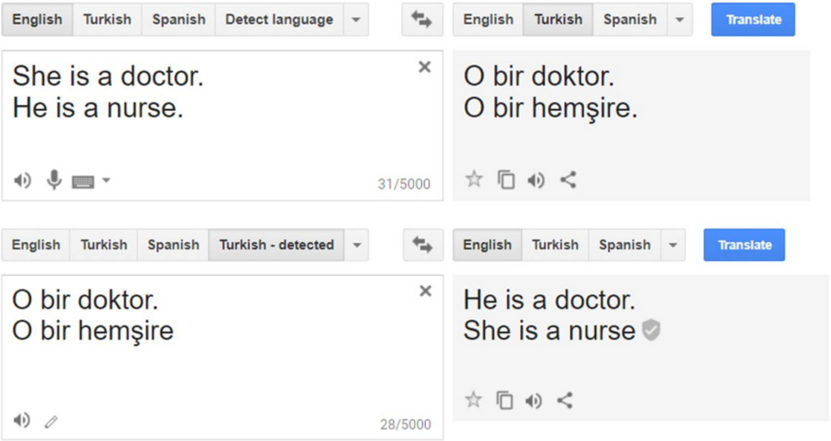
26 Jan 2023 — The powerful new AI chatbot tool recently passed **law exams** in four courses at the University of Minnesota and another exam at University of ...

https://www.reuters.com › legal › transactional › chatgpt-...

ChatGPT passes law school exams despite 'mediocre ...

25 Jan 2023 — (Reuters) - **ChatGPT** cannot yet outscore most **law** students on **exams**, new research suggests, but it can eke out a passing grade.

Language Models can be Biased



Source: fastai NLP course.

OPENAI'S NEW MULTITALENTED AI WRITES, TRANSLATES, AND SLANDERS

A step forward in AI text-generation that also spells trouble

By James Vincent | Feb 14, 2019, 12:00pm EST

Howard, co-founder of Fast.AI agrees. "I've been trying to warn people about this for a while," he says. "We have the technology to totally fill Twitter, email, and the web up with reasonable-sounding, context-appropriate prose, which would drown out all other speech and be impossible to filter."

Popular Mechanics

Artificial Intelligence Suddenly Evolves to Reach Theory of Mind

1 day ago

Discover Magazine

AI Chatbot Spontaneously Develops A Theory of Mind

2 days ago

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.
- Scientific goals:
  - Relate text data to metadata to understand social forces.
  - Understand the motivations and decisions of judges and public officials through their writings and speeches.

Welcome

- This course focuses on applications of **natural language processing** in **law** and **social science**.
- Engineering goals:
  - Develop skills in applied natural language processing
  - Apply to machine interpretation and generation of natural language texts – e.g. legal and political documents.
- Scientific goals:
  - Relate text data to metadata to understand social forces.
  - Understand the motivations and decisions of judges and public officials through their writings and speeches.
- Policy goals:
  - ask about legal/social impacts of AI that can read and write natural language.

Logistics / Learning Materials

Language Models

Corpora

Dictionary-Based Methods

Sentiment Analysis

Main Logistics

- See syllabus.
- Teaching assistant: Jingwei Ni (jingweini8@gmail.com)
    - TA sessions are recorded or on zoom – first TA session introducing python and first notebook (recorded from last year) is already posted.
    - zoom TA office hours to be scheduled
- Course communication:
    - announcements will be done on Moodle (and sent by email).
    - questions/concerns, bring up in TA office hours or post on moodle

Overview of Lectures

- Week 01: Introduction and Overview (Feb 19th)
- Week 02: Tokenization (Feb 26th)
- Week 03: Dimensionality and Distance (March 4th)
- Week 04: Machine Learning for NLP (March 11th)
- Week 05: Word Embeddings (March 18th)
- Week 06: *Guest Lecture on GTM* (March 25th)
- Week 07: Embedding Sequences (April 8th)
- Week 08: Large Language Models (April 22nd)
- Week 09: Language Model Alignment (April 29th)
- Week 10: ChatBots & Applications (May 6th)
- Week 11: *Guest Lecture, TBD* (May 13th)
- Week 12: In-Class Exam (May 27th)

Course Bibliography

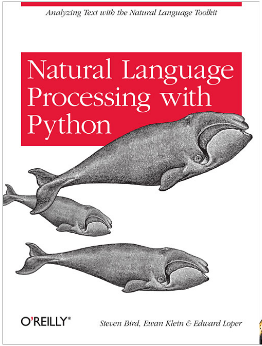
- Required readings for in-class discussions indicated in syllabus.

Course Bibliography

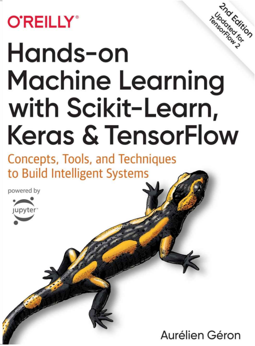
- Required readings for in-class discussions indicated in syllabus.
- Bibliography of references:
  - `https://bit.ly/NLP-reading-list`
  - reference readings on tools/methods
  - not required, but useful to complement the slides

Course Bibliography

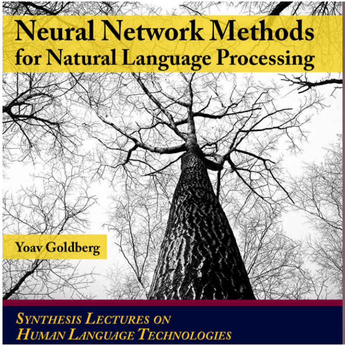
- Required readings for in-class discussions indicated in syllabus.
- Bibliography of references:
  - `https://bit.ly/NLP-reading-list`
  - reference readings on tools/methods
  - not required, but useful to complement the slides
- Bibliography of applications:
  - social science papers for response essays (more next week)



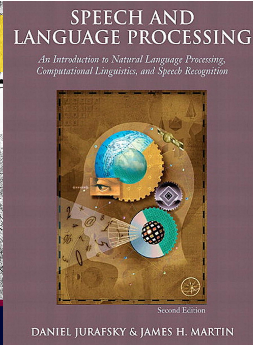
Analyzing Text with the Natural Language Toolkit
Natural Language Processing with Python
O'REILLY
Steven Bird, Ewan Klein & Edward Loper



O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron



Neural Network Methods for Natural Language Processing
Yoav Goldberg
Synthesis Lectures on Human Language Technologies



SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN

Required Readings

- Week 2: Ash and Hansen, "Text Algorithms in Economics"
- Week 6: Ash, Gauthier, and Widmer, "Generalized Topic Model"
- Week 7: Bloem, "Transformers from scratch"
- Week 9: Ouyang et al, "Training language models to following instructions with human feedback" (InstructGPT paper)

Python knowledge is a Course Pre-Requisite

- Course Repo: `https://github.com/elliottash/lm_lss_2024`
  - notebooks have examples; assignments have homeworks.
  - (some notebooks and assignments still to be updated)
- Python is ideal for text data and natural language processing.
  - Can use Anaconda or download the packages we need to a pip environment.
  - See the syllabus for list of packages we will use – especially sklearn, gensim, spacy, pytorch, and huggingface.
- First TA session video explains how to set things up.

Homework & TA Sessions

- Homeworks are due on Thursdays (first one Feb 29th)
  - Submit IPYNB file to Moodle
  - Completion grade – full credit for trying every question and submitting on time (checked programmatically and by random audit)
  - Have to submit an assignment (even if late) to see example solution.
- TA Sessions (Video Recordings)
- TA Office Hours: Thursdays, 10am-11am

Response Essays & In-Class Exam

- Response Essays (x2) (`https://eash.cc/NLP-RE`)
  - critically review one of the articles applying NLP methods
  - two drafts each; in first one, will get peer feedback.
  - more detail on this later

Response Essays & In-Class Exam

- Response Essays (x2) (`https://eash.cc/NLP-RE`)
  - critically review one of the articles applying NLP methods
  - two drafts each; in first one, will get peer feedback.
  - more detail on this later
- In-class exam on the last day of class
  - short-answer questions based on the slides.

Video Presentations

- Starting in Week 3, in most lectures we will have an in-class presentation given by a student group.
  - each student should contribute to one presentation during the term.
  - 2-3 students per presentation (groups of 3 if random, groups of 2 if paired on your own)
- Presentation version of a response essay:
  - summarizing the methods and main findings
  - identify at least one problem with the paper, or idea for improvement, per member of the group.
  - 11 minutes max.

Course Workload

3 ECTS credits $\approx$ 90 hours of work

Course Workload

3 ECTS credits $\approx$ 90 hours of work

- 12 lectures, 2 hours each = 24 hours
- 8 NLP programming homework assignments, ~1.5 hour each $\approx$ 12 hours
- In-class presentation prep, 4 hours
- 2 response essays, ~6 hours each $\approx$ 12 hours
- Other required readings $\approx$ 5 hours
- $\approx$ **55 required hours.**

Course Workload

3 ECTS credits $\approx$ 90 hours of work

- 12 lectures, 2 hours each = 24 hours
- 8 NLP programming homework assignments, ~1.5 hour each $\approx$ 12 hours
- In-class presentation prep, 4 hours
- 2 response essays, ~6 hours each $\approx$ 12 hours
- Other required readings $\approx$ 5 hours
- $\approx$ **55 required hours.**
- $\approx$ 35 hours at student discretion:
  - 8 optional TA sessions, 1 hour each $\approx$ 8 hours
  - ~27 hours for additional study time

Extra Credit: Text as Data Workshop

- On April 15-16, I am hosting an online workshop on social science papers using text as data.
- Attend for at least one hour, and ask at least one question in the zoom chat, and get 1 point added to a response essay score.

Any logistical questions about the course?

What is the endpoint of NLP?

What is the endpoint of NLP?

Machine understanding of text **discourse** across long documents and corpora.

- good summaries of long texts: extraction of relevant information, discarding of irrelevant information.
- question answering: retrieving evidence and answers from large corpora
- what else?

Four modes for NLP

- **Local:** get at linguistic information/relations from local context, e.g. sentences, paragraphs:
  - computing local sentiment
  - textual entailment
  - co-reference resolution
  - closed question answering
- **Long document**: linguistic information from long documents:
  - TF-IDF and CBOW representations $\rightarrow$ supervised learning
  - cosine distance between vectors
- **Global / knowledge base**: corpus level tasks:
  - information retrieval / search
  - open question answering / claim checking
  - knowledge graphs
- **Generative/Creative:** generate text for some purpose.
  - compose a sonnet
  - draft a legal brief to attack the opponent's brief

Objectives: Social-Science Research using Text Data

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Language modeling:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model is measuring what we want.
4. Empirical analysis
   - Produce statistics or predictions with the trained model.
   - **Answer the research question.**

Logistics / Learning Materials

Language Models

Corpora

Dictionary-Based Methods

Sentiment Analysis

Text is high-dimensional

- Text data is a sequence of characters called documents.
- The set of documents is the **corpus**, which we will call D.
- sample of documents, each n_L words long, drawn from vocabulary of n_V words.
- The unique representation of each document has dimension $n_V^{n_L}$.
  - e.g., a sample of 30-word Twitter messages using only the one thousand most common words in the English language
    - $\rightarrow$ dimensionality $= 1000^{30} = 10^{32}$

Text data is unstructured

- the information we want is mixed together with (lots of) information we don't.

Text data is unstructured

- the information we want is mixed together with (lots of) information we don't.
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.

Text data is unstructured

- the information we want is mixed together with (lots of) information we don't.
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.
- The tools from Weeks 2 (Tokenization) and 3 (Dimension Reduction) are focused on this step:
  - transforming an unstructured corpus D to a usable matrix X.

Co-Reference Resolution

The legal pressures facing 0 Michael Cohen are growing in a wide - ranging investigation of 0 his personal business affairs and 0 his work on behalf of 1 0 his former client , President Trump . In addition to 0 his work for 1 Mr. Trump , 0 he pursued 0 his own business interests , including ventures in real estate , personal loans and investments in taxi medallions .

This course is about relating documents to metadata

- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.

This course is about relating documents to metadata

- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.
- e.g., measuring positive-negative sentiment Y in judicial opinions.
  - not that meaningful by itself.

This course is about relating documents to metadata

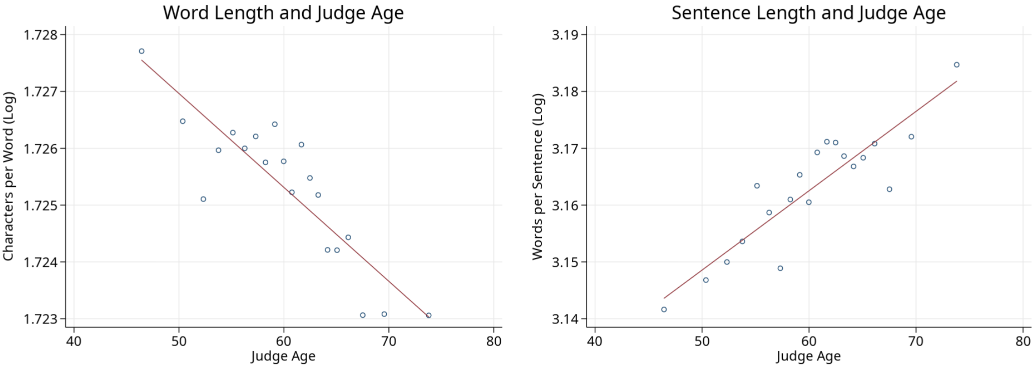
- This course is on **applied** NLP:
  - the documents are not that meaningful by themselves.
  - we want to relate **text** data to **meta**data.
- e.g., measuring positive-negative sentiment Y in judicial opinions.
  - not that meaningful by itself.
- but how about sentiment Y_{ijt} in opinion i by judge j at time t:
  - how does sentiment vary over time t?
  - does judge from party p_j express more negative sentiment toward defendants from group g_i?

e.g., Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)

e.g., Judge Age and Writing Style

Ash, Goessmann, and MacLeod (2022)



What counts as a document?

The unit of analysis (the "document") will vary depending on your question.

- needs to be fine enough to fit the relevant metadata variation
- should not be finer – would make dataset more high-dimensional without relevant empirical variation.

What counts as a document?

The unit of analysis (the "document") will vary depending on your question.

- needs to be fine enough to fit the relevant metadata variation
- should not be finer – would make dataset more high-dimensional without relevant empirical variation.

What should we use as the document in these contexts? (discuss in pairs)

1. predicting whether a judge is right-wing or left-wing in partisan ideology, from their written opinions.
2. predicting whether parliamentary speeches become more emotive in the run-up to an election
3. measuring whether newspapers use higher or lower sentiment toward different groups.

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).
- This won't be on an assignment but everyone in this class should learn how to:
  1. query REST API's
  2. run a web scraper in selenium
  3. do pre-processing on corpora, e.g. to remove HTML markup, fix errors associated with OCR.
- I also recommend everyone to become familiar with hugginface datasets (`https://huggingface.co/docs/datasets/`)

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Google Trends, Wikipedia).
- This won't be on an assignment but everyone in this class should learn how to:
  1. query REST API's
  2. run a web scraper in selenium
  3. do pre-processing on corpora, e.g. to remove HTML markup, fix errors associated with OCR.
- I also recommend everyone to become familiar with hugginface datasets (`https://huggingface.co/docs/datasets/`)
- All of the tools that we discuss in this class are available in many languages, and machine translation is now quite good (e.g. `huggingface.co/docs/transformers/master/en/model_doc/marian`).

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says "justice" vs "efficiency")

Overview of Dictionary-Based Methods

- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
  - use regular expressions for this task (see notebook)
- Corpus-specific: counting sets of words or phrases across documents
  - (e.g., number of times a judge says "justice" vs "efficiency")
- General dictionaries: WordNet, LIWC, MFD, etc.

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.

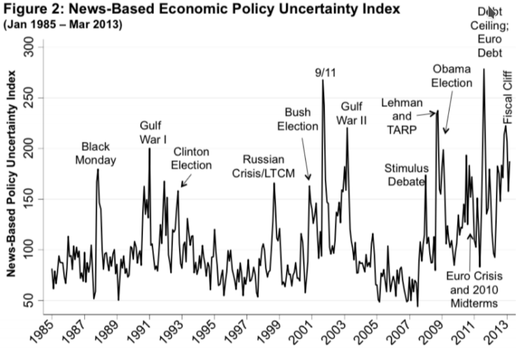
Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains "uncertain" OR "uncertainty", AND
2. Article contains "economic" OR "economy", AND
3. Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR "regulation" OR "white house"

Normalize resulting article counts by total newspaper articles that month.



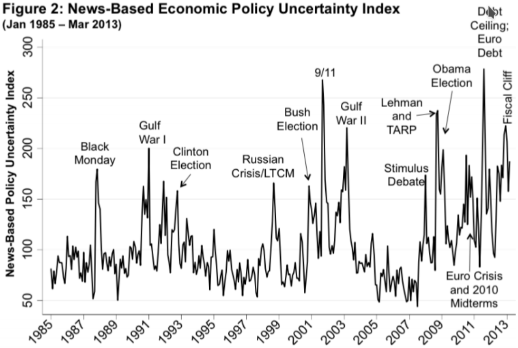
Measuring uncertainty in macroeconomy

Baker, Bloom, and Davis (QJE 2016)

For each newspaper on each day since 1985, submit the following query:

1. Article contains “uncertain” OR “uncertainty”, AND
2. Article contains “economic” OR “economy”, AND
3. Article contains “congress” OR “deficit” OR “federal reserve” OR “legislation” OR “regulation” OR “white house”

Normalize resulting article counts by total newspaper articles that month.



- but see Keith et al (2020), showing some problems with this measure (`https://arxiv.org/abs/2010.04706`).

WordNet

- English word database: 118K nouns, 12K verbs, 22K adjectives, 5K adverbs

The noun "bass" has 8 senses in WordNet.
1. bass^1 - (the lowest part of the musical range)
2. bass^2, bass part^1 - (the lowest part in polyphonic music)
3. bass^3, basso^1 - (an adult male singer with the lowest voice)
4. sea bass^1, bass^4 - (the lean flesh of a saltwater fish of the family Serranidae)
5. freshwater bass^1, bass^5 - (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
6. bass^6, bass voice^1, basso^2 - (the lowest adult male singing voice)
7. bass^7 - (the member with the lowest range of a family of musical instruments)
8. bass^8 - (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Figure 19.1 A portion of the WordNet 3.0 entry for the noun *bass*.

- Synonym sets (synsets) are a group of near-synonyms, plus a gloss (definition).
  - also contains information on antonyms (opposites), holonyms/meronyms (part-whole).
- Nouns are organized in categorical hierarchy (hence "WordNet")
  - "hypernym" – the higher category that a word is a member of.
  - "hyponyms" – members of the category identified by a word.

WordNet Supersenses (Word Categories)

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

WordNet Supersenses (Word Categories)

Category	Example	Category	Example	Category	Example
ACT	*service*	GROUP	*place*	PLANT	*tree*
ANIMAL	*dog*	LOCATION	*area*	POSSESSION	*price*
ARTIFACT	*car*	MOTIVE	*reason*	PROCESS	*process*
ATTRIBUTE	*quality*	NATURAL EVENT	*experience*	QUANTITY	*amount*
BODY	*hair*	NATURAL OBJECT	*flower*	RELATION	*portion*
COGNITION	*way*	OTHER	*stuff*	SHAPE	*square*
COMMUNICATION	*review*	PERSON	*people*	STATE	*pain*
FEELING	*discomfort*	PHENOMENON	*result*	SUBSTANCE	*oil*
FOOD	*food*			TIME	*day*

Figure 19.2 Supersenses: 26 lexicographic categories for nouns in WordNet.

Supersense	Verbs denoting ...
body	grooming, dressing and bodily care
change	size, temperature change, intensifying
cognition	thinking, judging, analyzing, doubting
communica-tion	telling, asking, ordering, singing
competition	fighting, athletic activities
consumption	eating and drinking
contact	touching, hitting, tying, digging
creation	sewing, baking, painting, performing
emotion	feeling
motion	walking, flying, swimming
perception	seeing, hearing, feeling
possession	buying, selling, owning
social	political and social activities and events
stative	being, having, spatial relations
weather	raining, snowing, thawing, thundering

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.

General Dictionaries

- Function words (e.g. *for*, *rather*, *than*)
  - also called stopwords
  - can be used to get at non-topical dimensions, identify authors.
- LIWC (pronounced "Luke"): Linguistic Inquiry and Word Counts
  - 2300 words 70 lists of category-relevant words, e.g. "emotion", "cognition", "work", "family", "positive", "negative" etc.
- Mohammad and Turney (2011):
  - code 10,000 words along four emotional dimensions: joy–sadness, anger-fear, trust-disgust, anticipation-surprise
- Warriner et al (2013):
  - code 14,000 words along three emotional dimensions: valence, arousal, dominance.

Dictionary Methods: Identifying Race-Related Research in Economics (1)

RACE-RELATED RESEARCH IN ECONOMICS AND OTHER SOCIAL SCIENCES*

ARUN ADVANI ELLIOTT ASH DAVID CAI IMRAN RASUL†

DECEMBER 2020

Abstract

How does economics compare to other social sciences in its study of race and ethnicity related issues? We assess this question using a corpus of 500,000 academic publications in economics, political science, and sociology. Using an algorithmic approach to classify race-related publications, we document that economics lags far behind the other disciplines in the volume and share of race-related research. Since 1960, there have been 13,000 race-related

Dictionary Methods: Identifying Race-Related Research in Economics (2)

Corpus. We build a corpus of publications for economics, political science, and sociology. The foundation for this corpus is the *JSTOR* database of academic journals (`jstor.org`). We consider all publications in journals that *JSTOR* characterizes as comprising the disciplines of economics, sociology, and political science. Although publication series are available back to the 1880s, our

this rises steadily over time. Our working sample from 1960 to 2020 covers nearly half a million journal publications: 224, 855 publications from 231 economics journals, 138, 188 publications from 185 sociology journals, and 110, 835 publications from 213 political science journals.

Dictionary Methods: Identifying Race-Related Research in Economics (3)

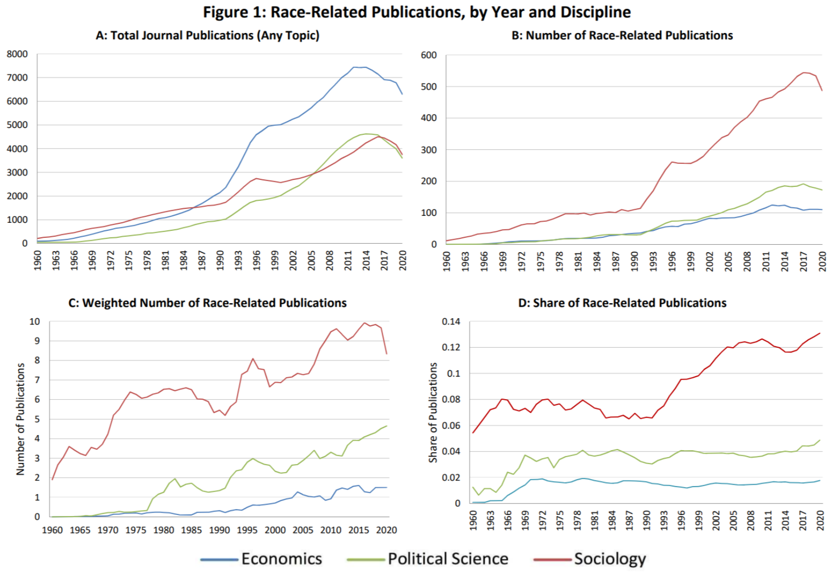
Identifying Race-Related Research. Given the volume of publications considered, it is infeasible to codify race-related research by hand. We thus take an automated approach and use an algorithm to classify race-related publications. We do so using keywords along two dimensions: (i) the racial or ethnic group being studied; and (ii) the issue being studied. Examples of (case-insensitive) keywords along the group dimension are race, african-american, person of color, and ethnicity. Examples of (case-insensitive) issue keywords include discrimination, prejudice, and stereotype.[2]

Our algorithm selects a publication as being race-related if: (i) at least one group keyword is in the title; or, (ii) at least one group keyword and at least one issue keyword are mentioned in the title or abstract. For rule (ii) we drop the last sentence of the abstract to avoid false positives from research that only mentions race parenthetically, say because it is part of some robustness check rather than the primary focus of study.

Specifically, we define three bands of group keywords that gradually expand on the racial or ethnic groups being studied. Band 0 consists of only abstract or generic keywords denoting racial and ethnic groups (e.g. race, ethnic, under represented minority). Band 1 adds group keywords relating to the main minority groups in the U.S. (African American, Latinos and Native Americans). Band 2 adds less salient group keywords (e.g. White, South Asian, Indian American, Japanese American) and other minorities based on religious beliefs (e.g. Muslim, Jewish). The full lexicon of group keywords used by Band are shown in Appendix Table A1.

The lexicon of issue keywords, shown in Appendix Table A2, are held constant and not split into bands. These words and phrases are broadly split across five broader topics: discrimination, inequality, diversity, identity, and historical issues. For example, discrimination includes prejudice and stereotypes, while inequality includes disparity and disadvantage.

Dictionary Methods: Identifying Race-Related Research in Economics (4)



Notes: We use data from JSTOR, Scopus, and the Web of Science to construct the number and shares of race related publications in economics, political science, and sociology. Panel A reports the total number of publications in each discipline. As the publication series start in the 1880s, the publication numbers do not start exactly at zero in 1960, the first year of our working sample. Panel B reports the number of articles that are determined to be race-related by our algorithm. Panel C reports a journal-weighted version of Panel B using the journal quality weights from Angrist et al. [2020]. Panel D reports the share of articles determined to be race-related by our algorithm in each discipline. All series presented are 5-year moving averages.

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily:
  - e.g., "good" versus "not good" versus "not very good"
  - what if you are analyzing court documents, and "murder" is identified as a negative sentiment term.

Sentiment Analysis

Extract a "tone" dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily:
  - e.g., "good" versus "not good" versus "not very good"
  - what if you are analyzing court documents, and "murder" is identified as a negative sentiment term.
- huggingface model hub has a number of transformer-based sentiment models, based on human annotations.
  - but again, a court document mentioning "murder" will probably get a negative-toned score

Sentiment Analysis

Extract a “tone” dimension – positive, negative, neutral

- standard approach is lexicon-based, but they fail easily:
  - e.g., “good” versus “not good” versus “not very good”
  - what if you are analyzing court documents, and “murder” is identified as a negative sentiment term.
- huggingface model hub has a number of transformer-based sentiment models, based on human annotations.
  - but again, a court document mentioning “murder” will probably get a negative-toned score
- Off-the-shelf scores are corpus specific, eg online writing – may not work for legal text, for example.
  - Hamilton et al (2016) and Zorn and Rice (2019) show how to make domain-specific sentiment lexicons using word embeddings (more on this later).

Problems with Sentiment Analyzers: NLP System Bias

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

Is this sentiment model racist?

Source: Kareem Carr slides.

NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

Example: Ash, Durante, Grebenschikova, Schwarz (2021) (1)

Classifier for Gender and Ethnicity

Example: Ash, Durante, Grebenschikova, Schwarz (2021) (2)

Table 1: IMAGE SHARES AND TEXT SENTIMENT

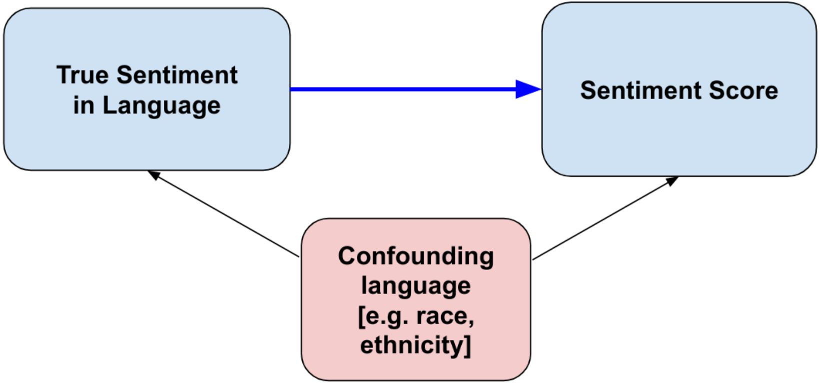
	Dep. Variable: Sentiment of Text				
	(1) Female	(2) White	(3) Black	(4) Asian	(5) Hispanic
Image Share	0.098***	0.063***	-0.072***	-0.015**	0.065***
	(0.004)	(0.004)	(0.005)	(0.007)	(0.007)
FOX × Image Share	0.001	0.055***	-0.062***	0.007	-0.024*
	(0.007)	(0.006)	(0.009)	(0.011)	(0.013)
Outlet × Section FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
outlet	Yes	Yes	Yes	Yes	Yes
Observations	404,861	404,861	404,861	404,861	404,861
Mean of DV	0.34	0.34	0.34	0.34	0.34

NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

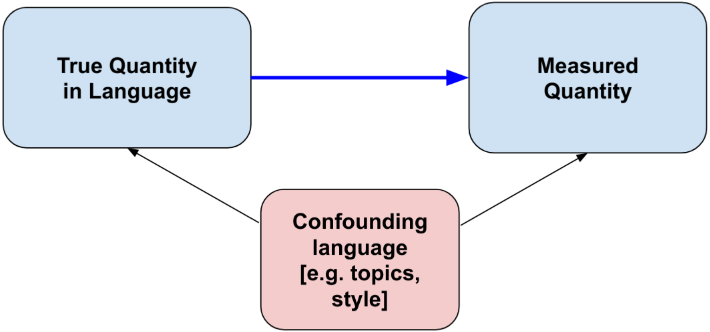
NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

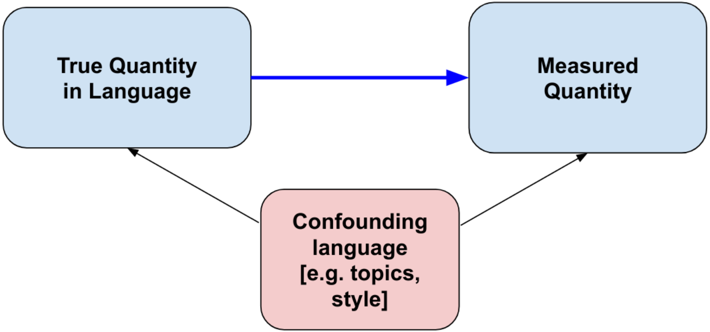


- Supervised sentiment models are confounded by correlated language factors.
  - e.g., in the training set maybe people complain about Mexican food more often than Italian food because Italian restaurants tend to be more upscale.

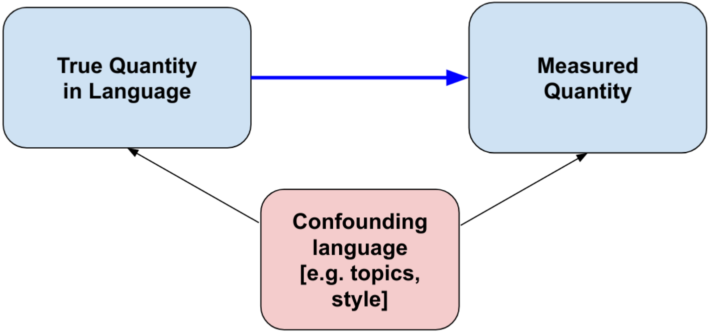
This is a universal problem



- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

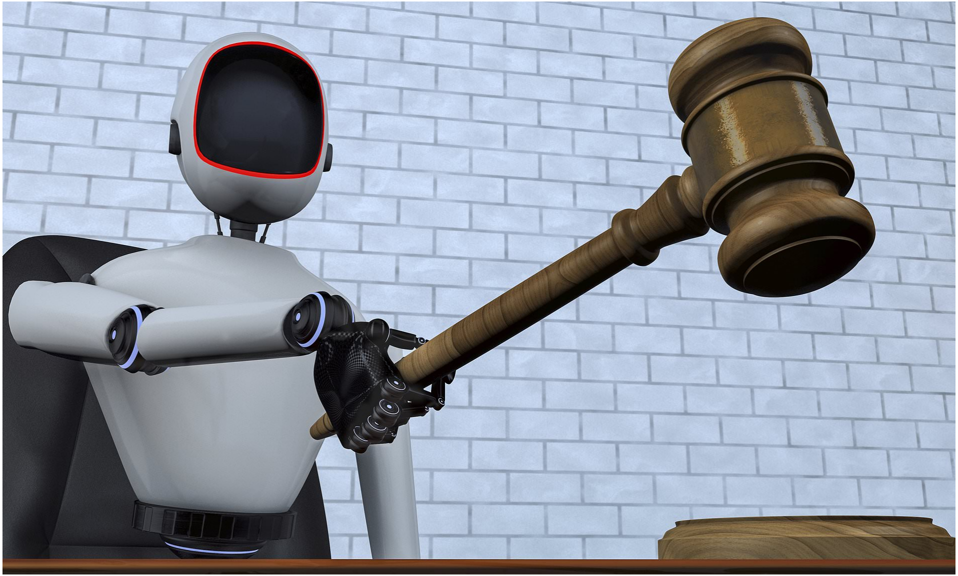


- **dictionary methods**, while having other limitations, mitigate this problem
  - the researcher intentionally “regularizes” out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.



- **dictionary methods**, while having other limitations, mitigate this problem
  - the researcher intentionally "regularizes" out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.
- but limitations of dictionaries are severe; we often cannot afford to use them.
  - so we will have to take on the considerable challenge of debiasing NLP models.

Questions/Comments?



Meeting Adjourned!