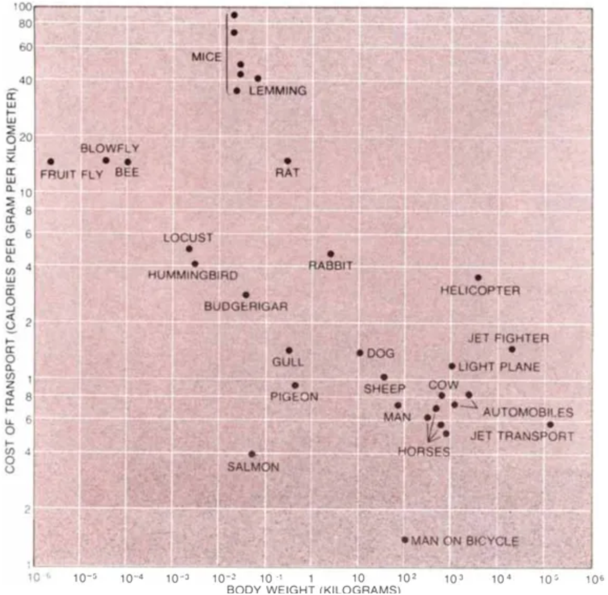
Text as Data

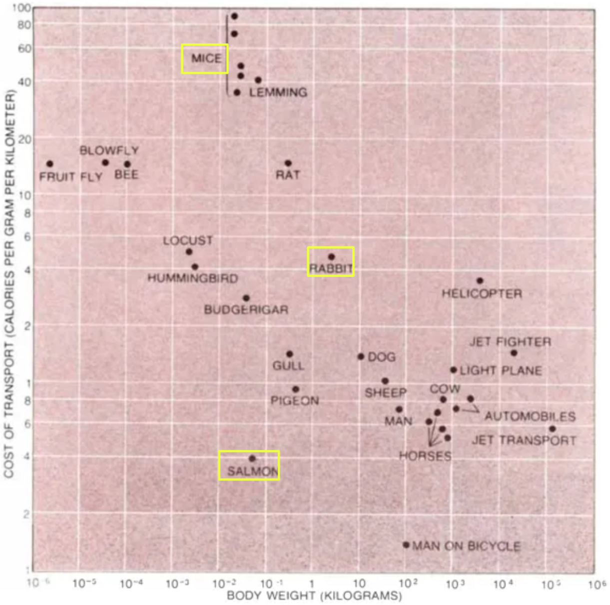
Spring 2026

1. Introduction

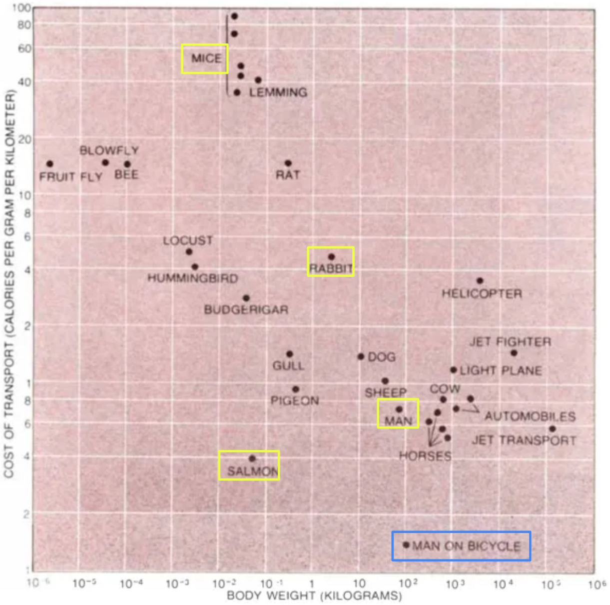


Scientific American, March 1973.

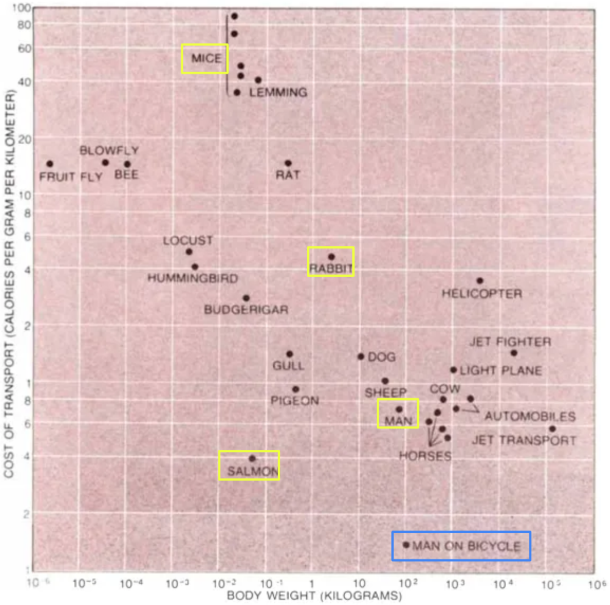
Hat tip: Sendhil Mullainathan



Scientific American, March 1973.



Scientific American, March 1973.



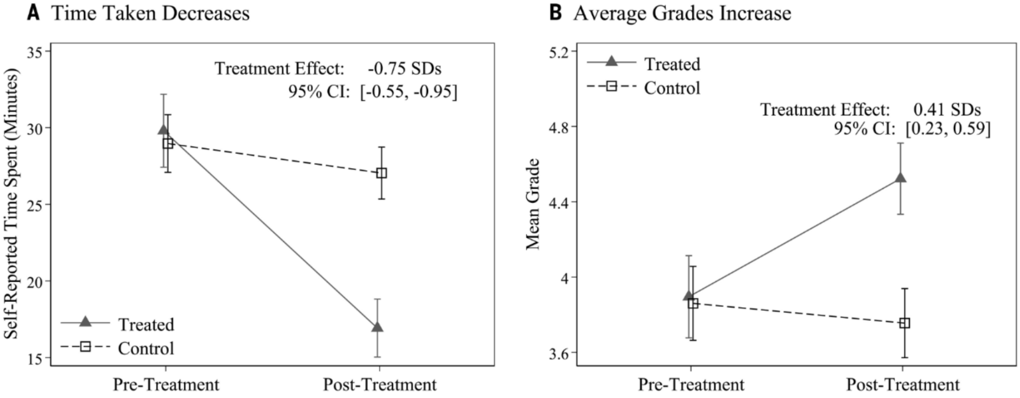
"And, a man on a bicycle, a human on a bicycle, blew the condor away, completely off the top of the charts.

And that's what a computer is to me. What a computer is to me is, it's the most remarkable tool that we've ever come up with, and it's the equivalent of a **bicycle for our minds**."

~ Steve Jobs

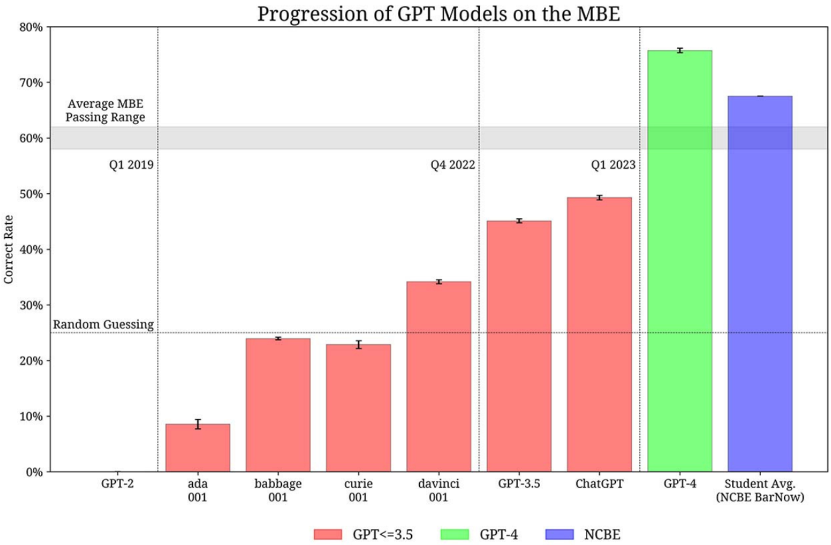
Language Models as Brain Bicycles

AI helps in professional writing tasks (Noy & Zhang 2023)



. . but other evidence is mixed (e.g. Dell’Acqua et al 2023; Cui et al 2024).

AI crushes the bar exam



… but AI still makes a lot of mistakes.

Air Canada must honor refund policy invented by airline's chatbot

Air Canada appears to have quietly killed its costly chatbot support.

ASHLEY BELANGER - 2/16/2024, 6:12 PM

On the day Jake Moffatt's grandmother died, Moffat immediately visited Air Canada's website to book a flight from Vancouver to Toronto. Unsure of how Air Canada's bereavement rates worked, Moffatt asked Air Canada's chatbot to explain.

The chatbot provided inaccurate information, encouraging Moffatt to book a flight immediately and then request a refund within 90 days. In reality, Air Canada's policy explicitly stated that the airline will not provide refunds for bereavement travel after the flight is booked. Moffatt dutifully attempted to follow the chatbot's advice and request a refund but was shocked that the request was rejected.

https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/

… but AI still makes a lot of mistakes.

FORBES > BUSINESS

BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff

I cover breaking news.

Follow

Jun 8, 2023, 02:06pm EDT

DoNotPay Has To Pay, After FTC Dings It For Lying About Its Non-Existent AI Lawyer

Legal Issues

from the *robot-lawyer-malfunctioned-when-confronted-with-the-ftc* dept

Thu, Sep 26th 2024 12:55pm - Mike Masnick

Remember "DoNotPay"? They were the company, run by Joshua Browder, claiming to be the "world's first robot lawyer." There were all sorts of sketchy things going on, some of which dated back to "DoNotPay's" **earliest days**. But things really came to a head last year when legal investigator extraordinaire, Kathryn Tewson, **started digging in** and **finding** an **awful lot** of **questionable** things **going on**.

And there are other social risks to think about

The New York Times

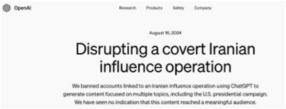
See How Easily A.I. Chatbots Can Be Taught to Spew Disinformation

By Jeremy White May 19, 2024

The New York Times

In Big Election Year, A.I.'s Architects Move Against Its Misuse

Anthropic, OpenAI, Google, Meta and other key developers are acting to prevent the technology from threatening democracies, even as their tools become more powerful.



OpenAI

Disrupting a covert Iranian influence operation

We banned accounts linked to an Iranian influence operation using ChatGPT to generate content focused on multiple topics, including the U.S. presidential campaign. We have seen no indication that this content reached a meaningful audience.

The Washington Post

Democracy Dies in Darkness

Subscribe

Tech Help Desk Artificial Intelligence Internet Culture Space Tech Policy

AI deepfakes threaten to upend global elections. No one can stop them.

As more than half the global population heads to the polls in 2024, AI-powered audio, images and videos are sowing confusion and clouding the political debate

8 min 123

How AI could manipulate voters and undermine elections, threatening democracy

The dangers and challenges of AI-powered propaganda and misinformation

Welcome

- This course focuses on applications of **text as data** in **social science**.

Welcome

- This course focuses on applications of **text as data** in **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 **text as data** in **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 **text as data** in **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 assistants
- Course communication:
  - announcements will be done on Brightspace (and sent by email).
  - questions/concerns, post on Brightspace
- Overview of Lectures

Course Bibliography

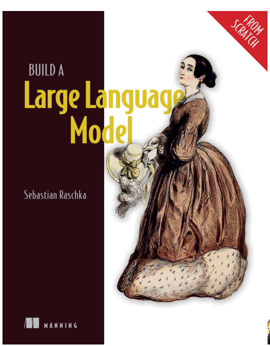
- Required readings for in-class discussions indicated in syllabus.

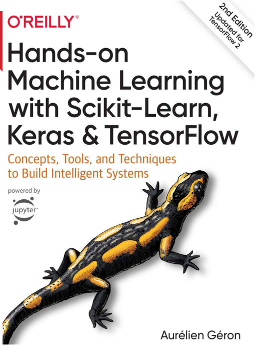
Course Bibliography

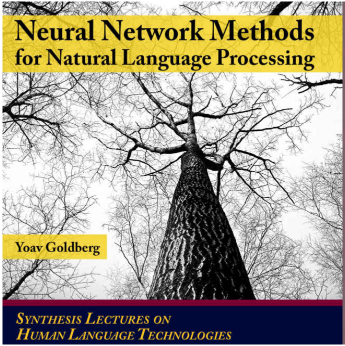
- Required readings for in-class discussions indicated in syllabus.
- Bibliography of references:
  - `https://bit.ly/NLP-reading-list`
  - reference readings on tools/methods
  - not required, but useful to complement the slides

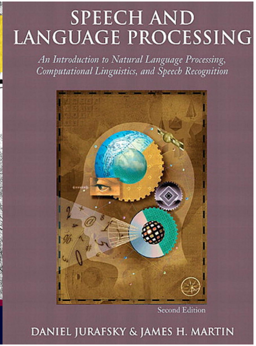
Course Bibliography

- 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)


FROM SCRATCH
BUILD A
Large Language Model
Sebastian Raschka
MANNING


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

Python knowledge is a Course Pre-Requisite

- Course Repo: `https://github.com/elliottash/tad_2026`
  - notebooks have examples; assignments have homeworks.
- Python is ideal for text data and natural language processing.
  - Can use Anaconda, google colab, 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, bertopic, huggingface.
- First TA session will help set things up.

Homework / TA Sessions / Etc

- Homeworks
- Presentations
- Exam
- Course Project
- See syllabus for most up to date info.

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 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
- are we there now?
- 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 (covering multiple topics)**: 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
- **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.**

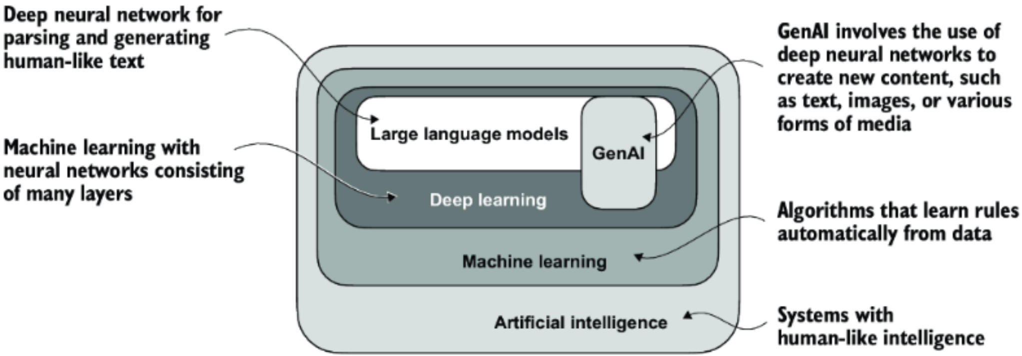


Figure 1.1 As this hierarchical depiction of the relationship between the different fields suggests, LLMs represent a specific application of deep learning techniques, using their ability to process and generate human-like text. Deep learning is a specialized branch of machine learning that focuses on using multilayer neural networks. Machine learning and deep learning are fields aimed at implementing algorithms that enable computers to learn from data and perform tasks that typically require human intelligence.

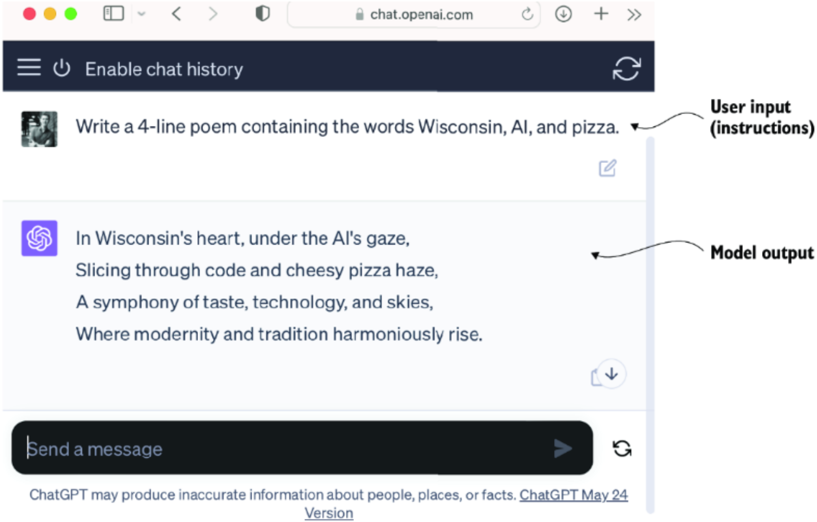


Figure 1.2 LLM interfaces enable natural language communication between users and AI systems. This screenshot shows ChatGPT writing a poem according to a user's specifications.

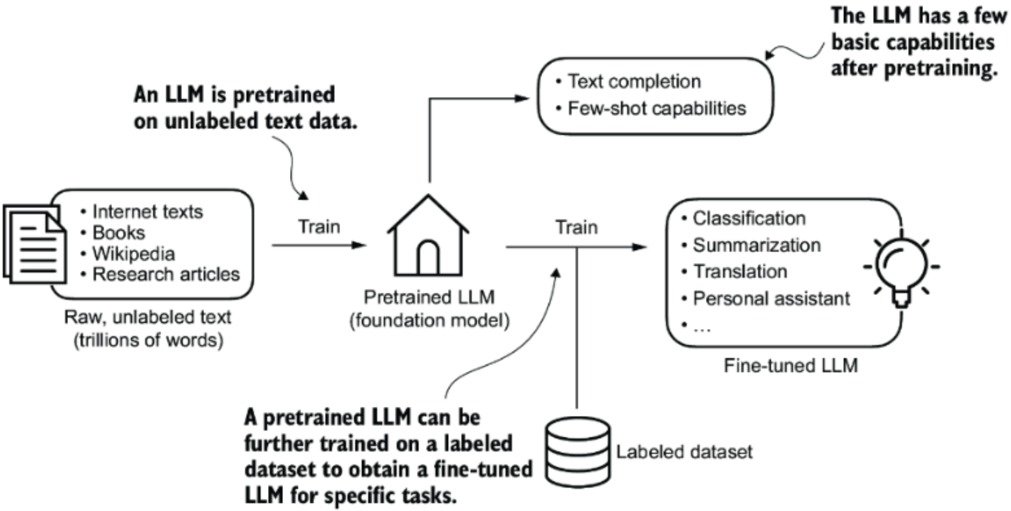


Figure 1.3 Pretraining an LLM involves next-word prediction on large text datasets. A pretrained LLM can then be fine-tuned using a smaller labeled dataset.

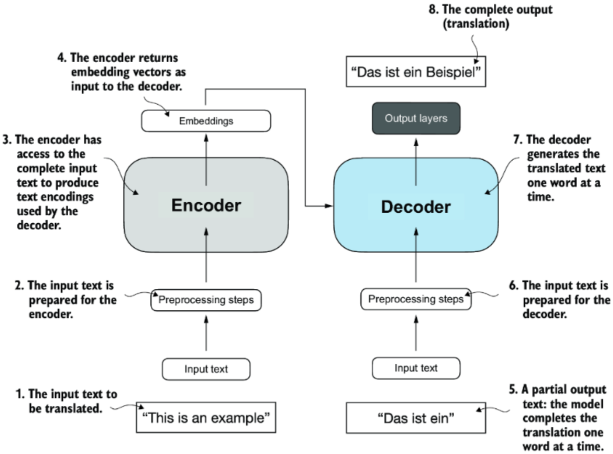


Figure 1.4 A simplified depiction of the original transformer architecture, which is a deep learning model for language translation. The transformer consists of two parts: (a) an encoder that processes the input text and produces an embedding representation (a numerical representation that captures many different factors in different dimensions) of the text that the (b) decoder can use to generate the translated text one word at a time. This figure shows the final stage of the translation process where the decoder has to generate only the final word ("Beispiel"), given the original input text ("This is an example") and a partially translated sentence ("Das ist ein"), to complete the translation.

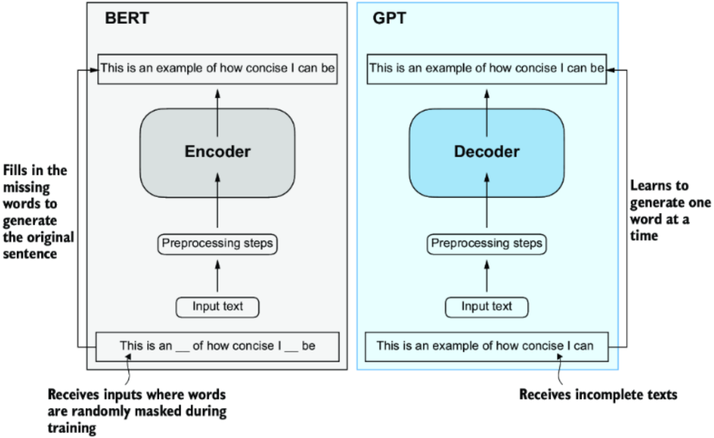


Figure 1.5 A visual representation of the transformer's encoder and decoder submodules. On the left, the encoder segment exemplifies BERT-like LLMs, which focus on masked word prediction and are primarily used for tasks like text classification. On the right, the decoder segment showcases GPT-like LLMs, designed for generative tasks and producing coherent text sequences.

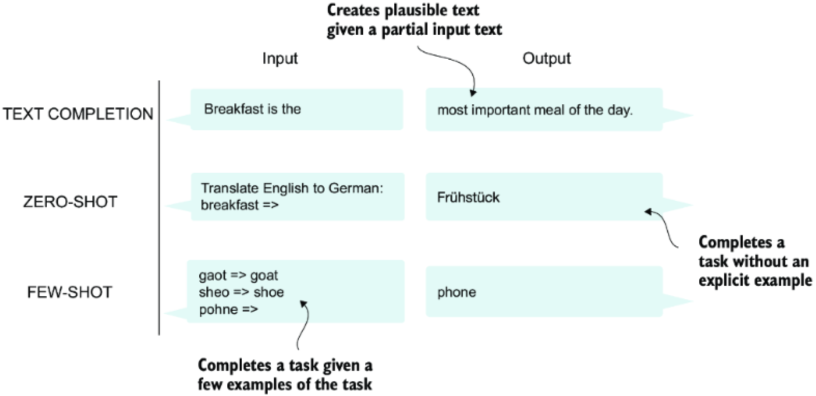


Figure 1.6 In addition to text completion, GPT-like LLMs can solve various tasks based on their inputs without needing retraining, fine-tuning, or task-specific model architecture changes. Sometimes it is helpful to provide examples of the target within the input, which is known as a few-shot setting. However, GPT-like LLMs are also capable of carrying out tasks without a specific example, which is called zero-shot setting.

Table 1.1 The pretraining dataset of the popular GPT-3 LLM

Dataset name	Dataset description	Number of tokens	Proportion in training data
CommonCrawl (filtered)	Web crawl data	410 billion	60%
WebText2	Web crawl data	19 billion	22%
Books1	Internet-based book corpus	12 billion	8%
Books2	Internet-based book corpus	55 billion	8%
Wikipedia	High-quality text	3 billion	3%

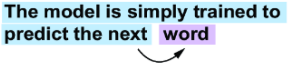


Figure 1.7 In the next-word prediction pretraining task for GPT models, the system learns to predict the upcoming word in a sentence by looking at the words that have come before it. This approach helps the model understand how words and phrases typically fit together in language, forming a foundation that can be applied to various other tasks.

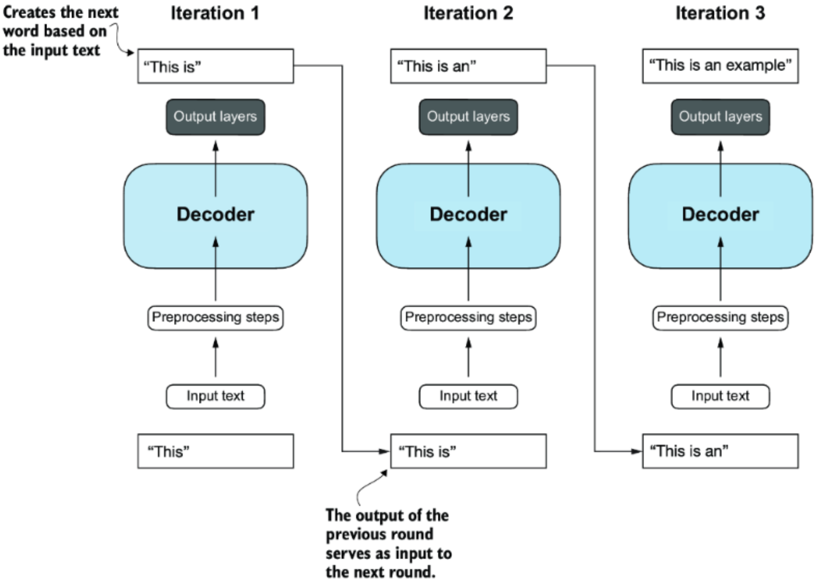


Figure 1.8 The GPT architecture employs only the decoder portion of the original transformer. It is designed for unidirectional, left-to-right processing, making it well suited for text generation and next-word prediction tasks to generate text in an iterative fashion, one word at a time.

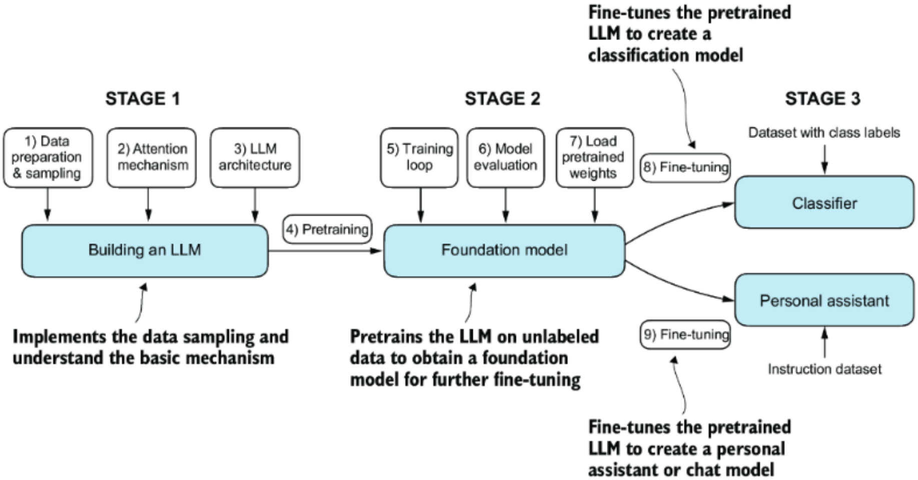


Figure 1.9 The three main stages of coding an LLM are implementing the LLM architecture and data preparation process (stage 1), pretraining an LLM to create a foundation model (stage 2), and fine-tuning the foundation model to become a personal assistant or text classifier (stage 3).

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 $n_V^{n_L}$.
  - for documents of fixed length n_L drawn from vocabulary of size n_V, the number of distinct possible documents is $n_V^{n_L}$.
    - E.g., a sample of 30-word Twitter messages using only 1,000 most common words in English:
      $\rightarrow$ distinct documents $= (10^3)^{30} = 10^{90}$

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 $n_V^{n_L}$.
    - for documents of fixed length n_L drawn from vocabulary of size n_V, the number of distinct possible documents is $n_V^{n_L}$.
        - E.g., a sample of 30-word Twitter messages using only 1,000 most common words in English:
          $\rightarrow$ distinct documents $= (10^3)^{30} = 10^{90}$
    - Represent each document as matrix of words, which are n_V-dimensional categorical variables across the vocabulary:

$$\boldsymbol{X} = [\ \boldsymbol{x}_1 \quad \dots \quad \boldsymbol{x}_t \quad \dots \quad \boldsymbol{x}_{n_L}\ ]$$

        - dimensionality of a single document $= n_L n_V$ (sparse, with n_L non-zero items)
        - for $n_L = 30$ and $n_V = 1,000$ (the Twitter example), dimensionality $= 30,000$

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.

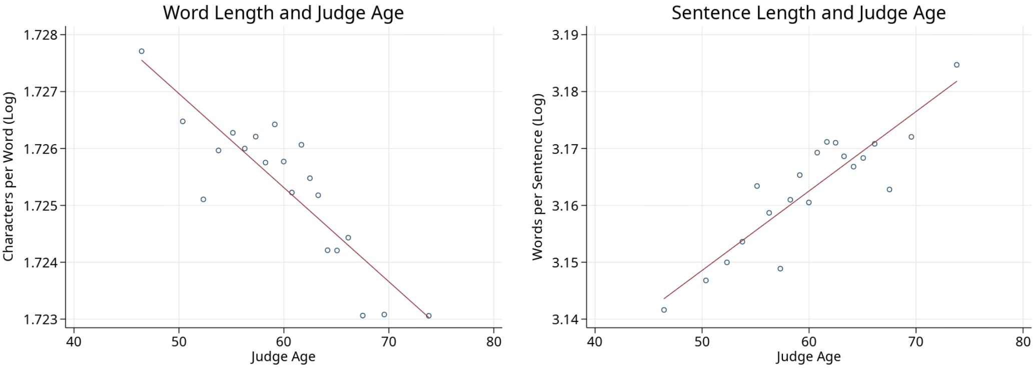
This course is about relating documents to metadata

- 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?

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. Wikipedia, CourtListener, Google Trends).

Handling Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. Wikipedia, CourtListener, Google Trends).
- 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 huggingface 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. Wikipedia, CourtListener, Google Trends).
- 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 huggingface datasets (`https://huggingface.co/docs/datasets/`)
- All of the tools that we discuss in this class are available in many languages, and machine translation with LLMs is excellent.

Logistics / Learning Materials

Language Models

Corpora

Dictionary-Based Methods

Sentiment Analysis

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

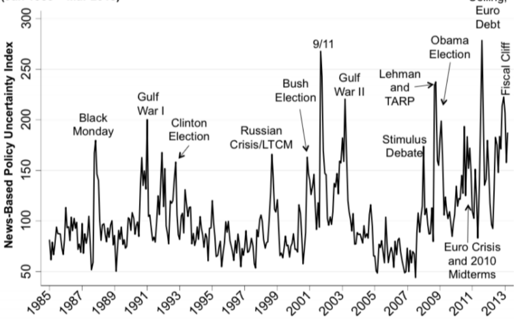
Baker, Bloom, and Davis (QJE 2016)

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.

Figure 2: News-Based Economic Policy Uncertainty Index
(Jan 1985 – Mar 2013)



Measuring uncertainty in macroeconomy

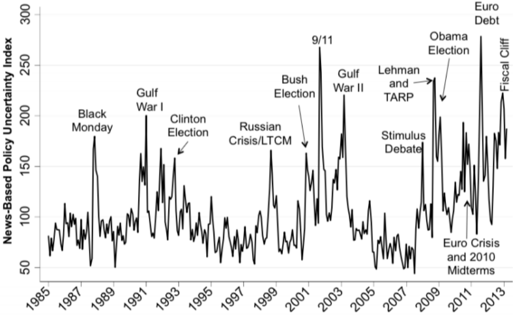
Baker, Bloom, and Davis (QJE 2016)

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.

Figure 2: News-Based Economic Policy Uncertainty Index
(Jan 1985 – Mar 2013)



- 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 in 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 in 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]

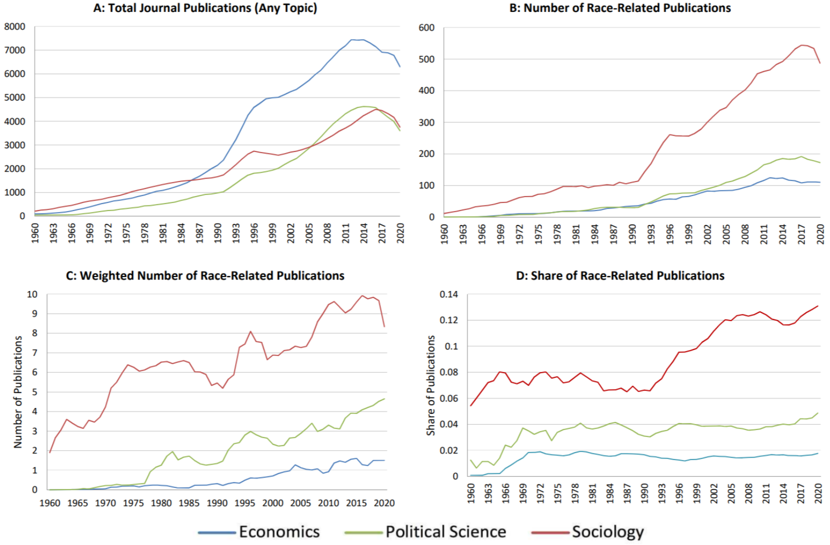
Our algorithm selects a publication as being race-related if: (i) at least one group keyword is in the title; or, (ii) at least one group keyword and at least one issue keyword are mentioned in the title or abstract. For rule (ii) we drop the last sentence of the abstract to avoid false positives from 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)

Figure 1: Race-Related Publications, by Year and Discipline



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.

Logistics / Learning Materials

Language Models

Corpora

Dictionary-Based Methods

Sentiment Analysis

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

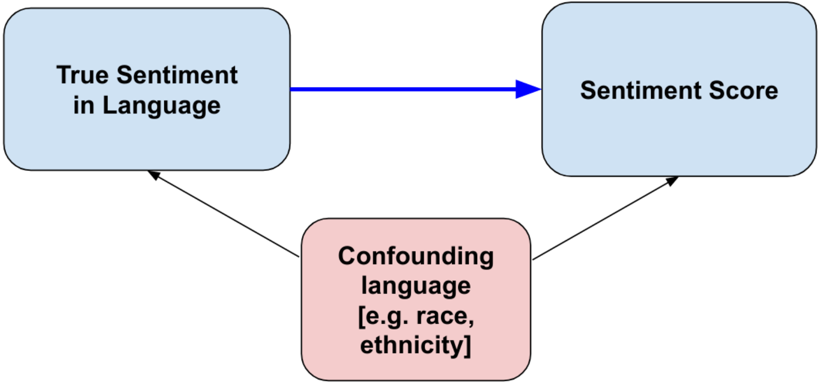
	Dep. Variable: Sentiment of Text				
	(1) Female	(2) White	(3) Black	(4) Asian	(5) Hispanic
Image Share	0.098***	0.063***	-0.072***	-0.015**	0.065***
	(0.004)	(0.004)	(0.005)	(0.007)	(0.007)
FOX × Image Share	0.001	0.055***	-0.062***	0.007	-0.024*
	(0.007)	(0.006)	(0.009)	(0.011)	(0.013)
Outlet × Section 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.

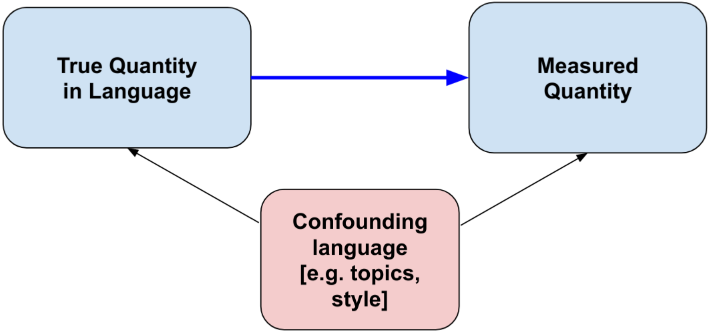
NLP "Bias" is statistical bias

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

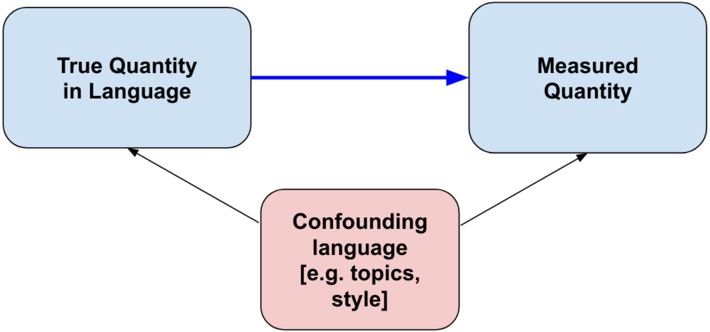


- Supervised sentiment models are confounded by correlated language factors.
  - e.g., in the training set, maybe people complain about Mexican food more often than Italian food because Italian restaurants tend to be more upscale.

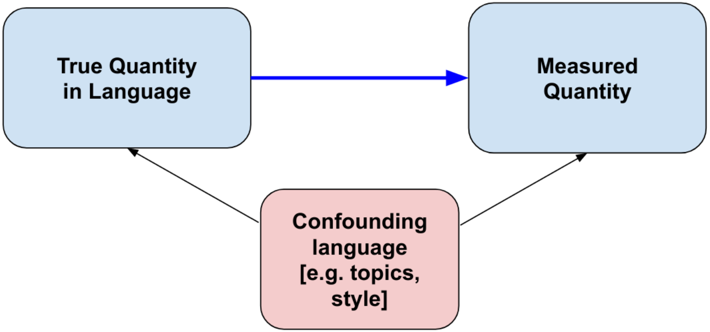
This is a universal problem



- 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.

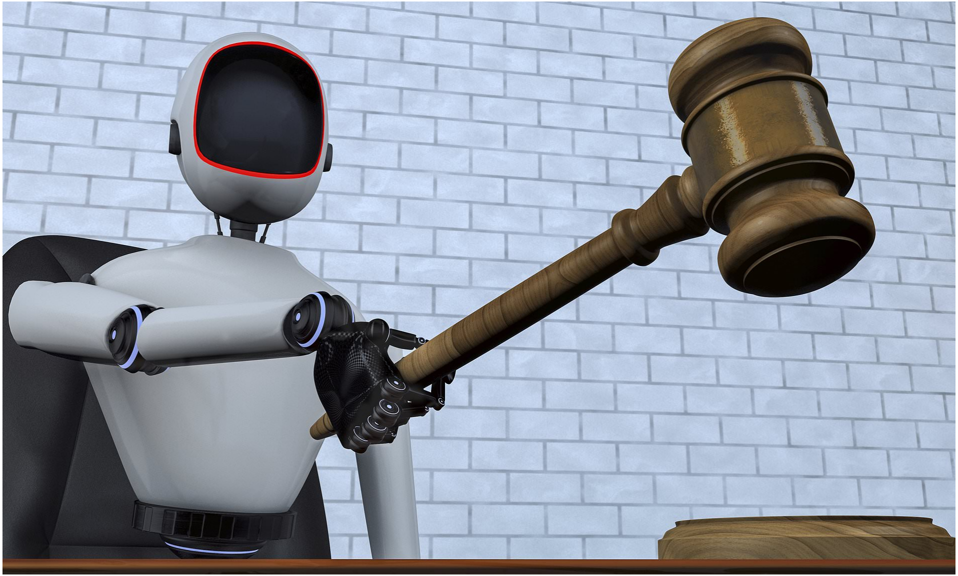


- **dictionary methods**, while having other limitations, mitigate this problem
  - the researcher intentionally "regularizes" out spurious confounders with the targeted language dimension.
  - helps explain why economists often still use dictionary methods.



- **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!