Language & AI

Lecture 2

Dr. ir. Marijn ten Thij, Nov 20st 2025

Video Lecture

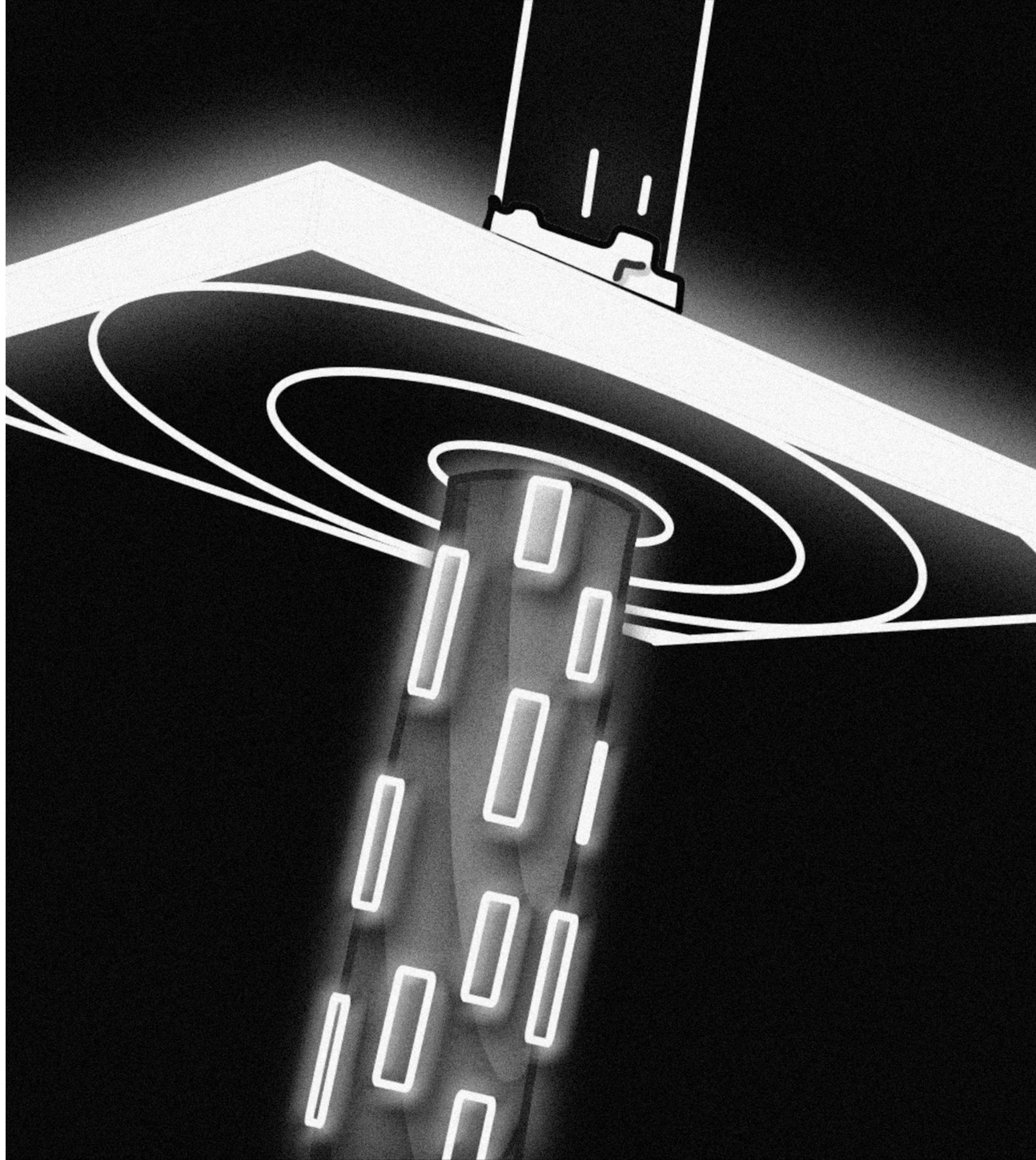
Collecting Data (etc.)

- What is / can be data?
- What is / can be noise?
- Denoising data
  - RegEx (find and replace)
  - Eval (models + corrections).
  - Normalization (spelling correction, case folding, stemming, lemmatization).
  - Encodings.

This Lecture

Thinking About Data

- Interim assignment questions.
- Data collection considerations.
- Lot about data annotation!
- Contextualizing preprocessing.



Research Components

Empirical, Quantitative, Qualitative

- Empirical questions: answers given through experiments.
- Experiments/analyses: quantitative vs. qualitative. In ML/NLP:
  - Quantitative always involves a metric:
    - We have models.
    - These are tested under certain experimental conditions (different splits, preprocessing, hyper-parameters, etc.).
    - We compare effects using evaluation metric differences.

Research Components (Cont.)
Empirical, Quantitative, Qualitative

- So what makes something a qualitative analysis in ML/NLP?
  - Usually involves actually looking at the data.
  - Can also be plotting certain things and comparing them yourself.
  - Generally this is to personally infer relations / make observations that aren't rooted in numbers; you don't have 'proof'.
  - Scribbr has some decent high level descriptions for more context.

Text Data

What They Don’t Tell You

Pre-Made
Is Unrealistic

- API’s and platforms (HuggingFace Data, Kaggle, etc.) are convenient.
- Working with .csv’s (tabular) and otherwise structured data is a breeze.
- This is (often) not what reality looks like. Even structured data can be a mess.

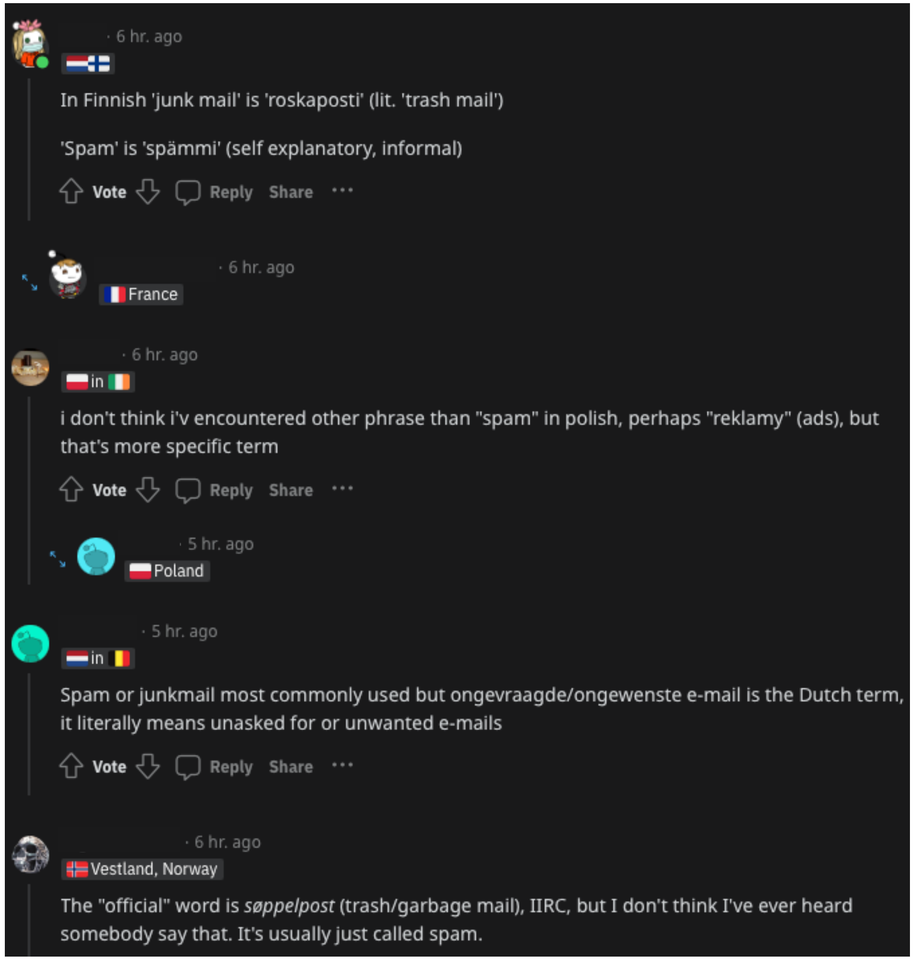
Real Text Data

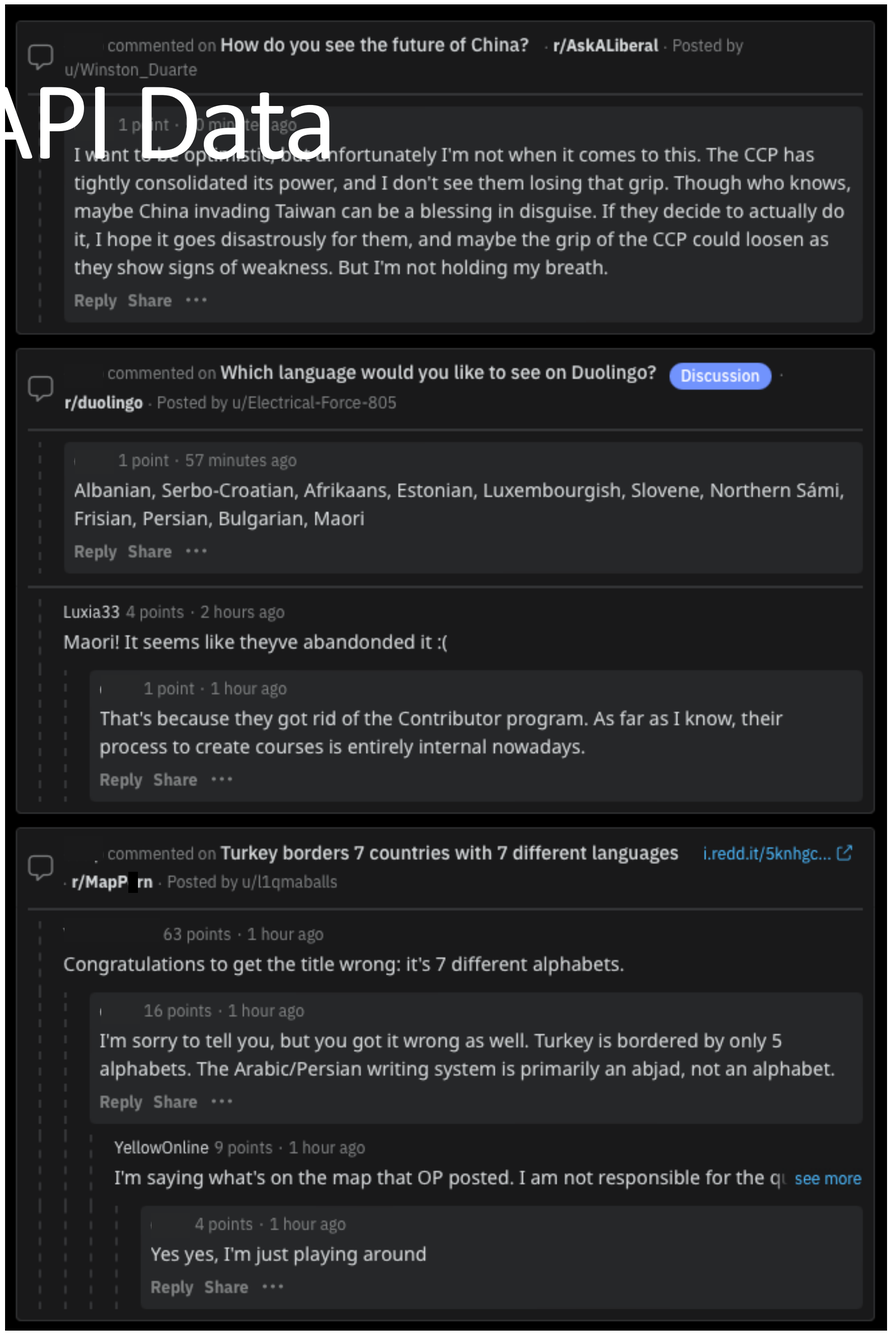
Where to Get It and Why Doesn't Everyone?

- Scraping (e.g. via Python).
- In essence very simple.
- **But**:
  - Detection Systems.
  - Scale (processing time, storage, relevancy).
  - Maintenance.

```
        <div class="...">
            <a class="..." style="color: rgb(129, 131, 132);" href="/user/ManGood2002/">
                u/ManGood2002
            </a>
            <div id="UserInfoTooltip--t3_apmsqk--lightbox"></div>
        </div>
        <a class="..." data-click-id="timestamp"
href="./r/Showerthoughts/comments/apmsqk/the_syllables_in_on_your_mark_get_set_go_are_a/
">
            3 years ago
        </a>
    </div>
    <!-- ... -->
    <div class="..." >
        <span class="..."  id="">
            <span id="PostAwardBadges--t3_apmsqk--lightbox-gid_3">
                <img alt="Platinum" class="..."  id="..."
src="https://www.redditstatic.com/gold/awards/icon/platinum_32.png">
            </span>
            <span></span>
        </span>
    <!-- ... -->
    </div>
    <div class="...">
        <div class="...">
            <div class="..."  style="--posttitletextcolor: #D7DADC;">
                <h1 class=""..." >
                    The syllables in “on your mark, get set, go” are a countdown
                </h1>
            </div>
        </div>
```

Processing API Data



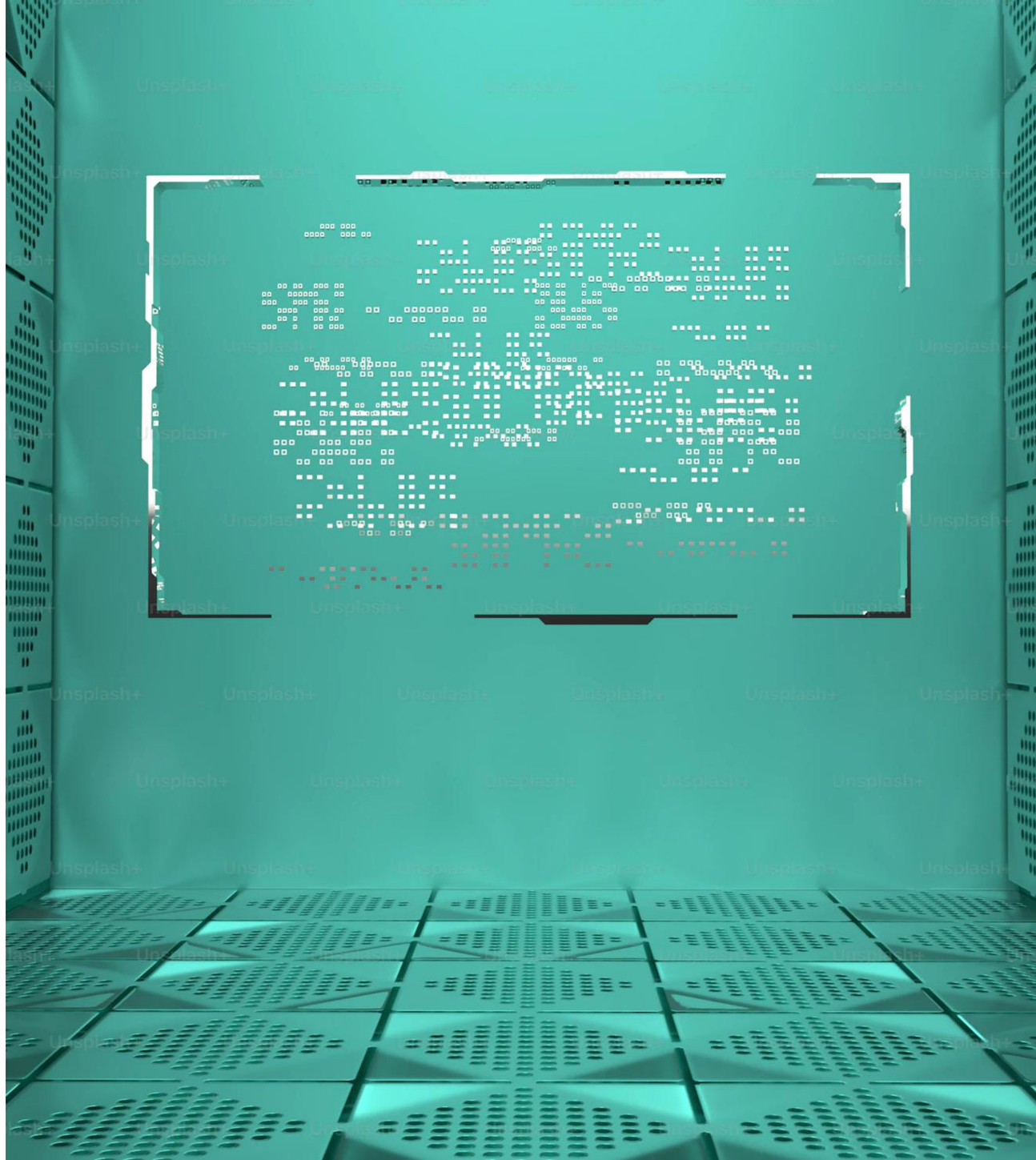


Errors?

Data Operations

After Retrieval

- Query time (training, stats).
  - What to index on?
  - What to split on?
- Redundancy (i.e., samples).
- Sharing.
- Database, dataframe, data file, output file, log file, …



Text as Value

Why the Internet is Closing

- Rising issues:
  - X has restricted research access.
  - Reddit API being restricted.
  - Data relevancy?
- Text data is monetized and access controlled.
- ‘Fair use’ often an issue.
- Don’t forget: this is your data!

Automated Content

NLP Undoing NLP

- Bots. All the bots.
- What is real and what isn’t?
  - Reviews.
  - Comments.
  - Articles.
  - Most interesting: AI trained on ‘the Internet’, generating content. And after?

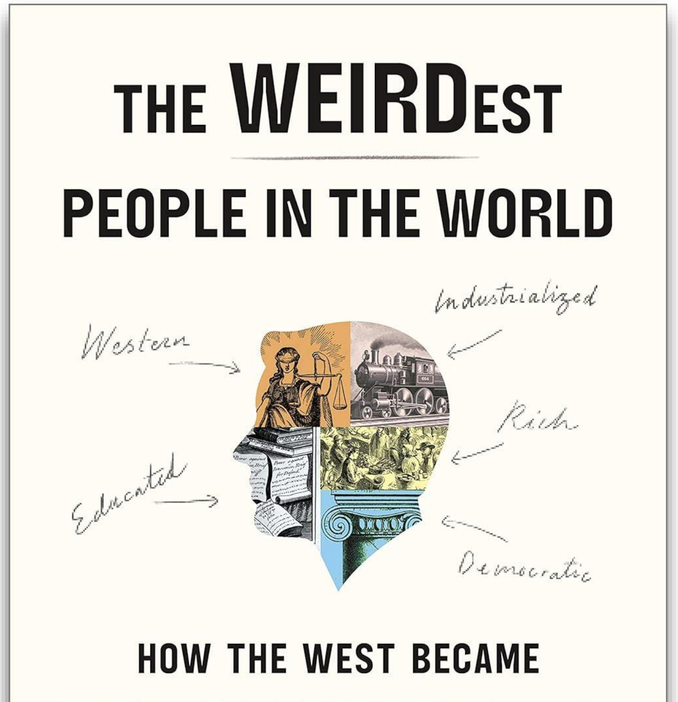
Any Questions?

(So far)

Considerations When Collecting Data

Things Other Than Computational Limitations

- Text in a structured format? Expect issues? Sensitive? Can it be shared?
- How much meta-data is required (in the future) and how can it be stored?
- Collect labels directly, using a heuristic (distant labeling), or annotators?
- Latter case: how to recruit annotators? Quality control?
  - MechanicalTurk, CrowdFlower, etc.
  - Experts. Your friends? WEIRD group.
  - Pay (how much?). What's the incentive? Inter-rater agreement?
- Sample bias (on everything). How to mitigate, and is that desirable?



Cohen's Kappa

Inter-rater Reliability Scoring

(No exercises so equation not on the exam 🥲)

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

where p_o is the relative observed agreement among raters, and p_e is:

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2},$$

where n_{k1} is the number of items (total N) classified as k by rater 1.

📄 -> 👤 👍 | 👤 👎 | 👤 👍

Annotations

And What to Pay Attention To

- How diverse is your team?
  - Not just the annotators!
- How are you instructing the annotators?
- If all fails: can we hide the information somehow?
- Might be used for post-hoc error analysis!

Don't Blame the Annotator:
Bias Already Starts in the Annotation Instructions

Mihir Parmar[1]* Swaroop Mishra[1]* Mor Geva[2]† Chitta Baral[1]
[1]Arizona State University [2]Allen Institute for AI
{mparmar3, srmishr1, chitta}@asu.edu, pipek@google.com

Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets

Mor Geva
Tel Aviv University,
Allen Institute for AI
morgeva@mail.tau.ac.

Yoav Goldberg
Bar-Ilan University,
Allen Institute for AI

Abstract

Crowdsourcing has been the prev
paradigm for creating natural language u
standing datasets in recent years. A cor
crowdsourcing practice is to recruit a
number of high-quality workers, and
them massively generate examples.
only a few workers generate the majo
examples raises concerns about data di
especially when workers freely
sentences. In this paper, we perform
of experiments showing these conc
evident in three recent NLP dataset
show that model performance impro
training with annotator identifiers as
and that models are able to recognize
productive annotators. Moreover,

cs.CL] 28 Aug 2019

Annotators with Attitudes:
How Annotator Beliefs And Identities Bias Toxic Language Detection

Maarten Sap♡♠ Swabha Swayamdipta♠ Laura Vianna♢
Xuhui Zhou♣ Yejin Choi♡♠ Noah A. Smith♡♠
♡Paul G. Allen School of Computer Science, University of Washington, Seattle, WA, USA
♠Allen Institute for AI, Seattle, WA, USA
♢Department of Psychology, University of Washington, Seattle, WA, USA
♣Georgia Institute of Technology, Atlanta, GA, USA
{maartens,swabhas}@allenai.org

Abstract

Warning: this paper discusses and contains content that is offensive or upsetting.

The perceived toxicity of language can vary based on someone's identity and beliefs, but this variation is often ignored when collecting toxic language datasets, resulting in dataset

and Tsvetkov, 2020) or backfiring against minorities (Yasin, 2018; Are, 2020, i.a.). For example, racial biases have been uncovered in toxic language detection where text written in African American English (AAE) is falsely flagged as toxic (Sap et al., 2019; Davidson et al., 2019).

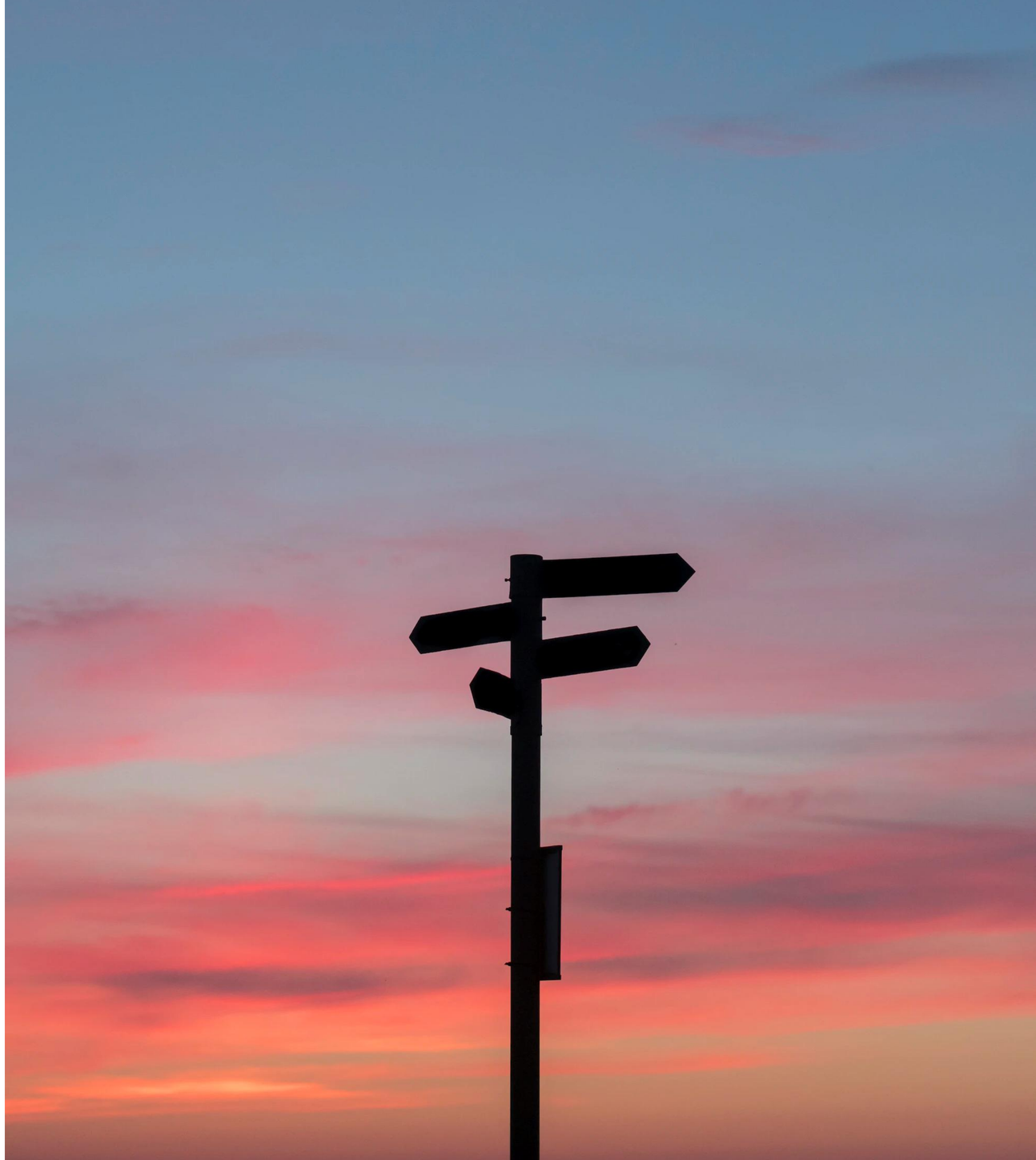
2022

Any Questions?

(So far)

What to Input?

Keep? Replace? Remove?

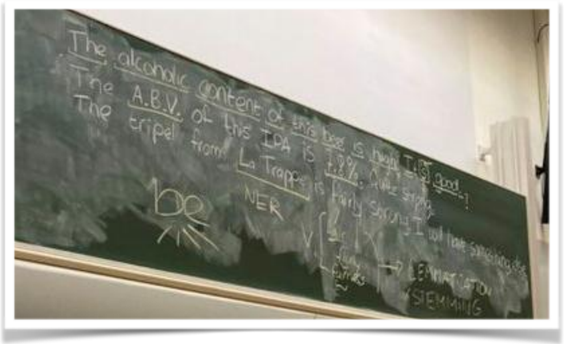


What Are Tokens?

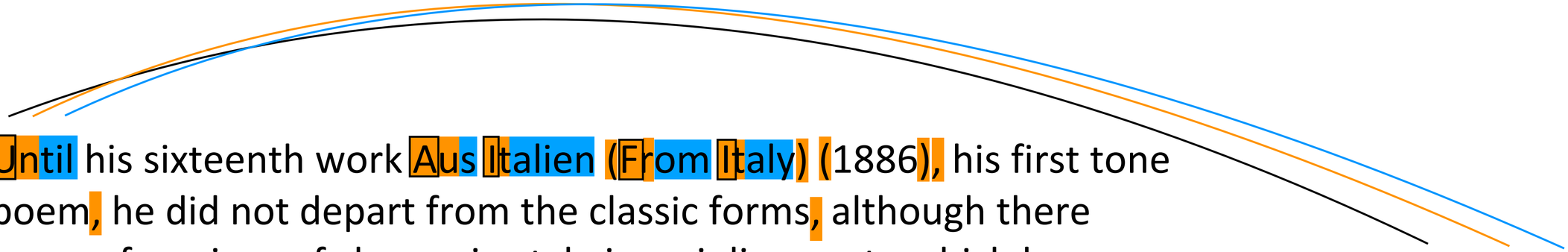
And Why Does Marijn Care So Much?

- The alcoholic content of this beer is high. It’s good.
- The A.B.V. of this IPA is 7.8%. Quite strong.
- This Triple from La Trappe is fairly strong. I will have something else.

- La Trappe’s Triple IPA. (if only)



Stylometry Example



Until his sixteenth work Aus Italien (From Italy) (1886), his first tone poem, he did not depart from the classic forms, although there were a few signs of change in style in a violin sonata which he wrote just before the tone poem. In fact, he was so much against Wagner and his innovations, that no one could have guessed that later he himself would be considered an innovator and would be accused of imitating Wagner.

During his youth, after hearing Siegfried he wrote to a friend about the music of Mime: "It would have killed 411a cat and the horror of musical dissonances would melt rocks into omelettes."

/[A-Z]\w+/g

/[^\w]+/g

Stylometry Example II

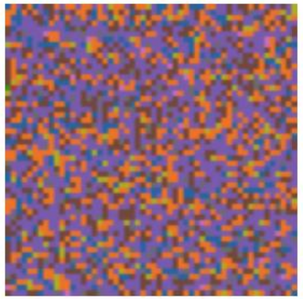
Until his sixteenth work Aus Italien (From Italy) (1886), his first tone poem, he did not depart from the classic forms, although there were a few signs of change in style in a violin sonata which he wrote just before the tone poem. In fact, he was so much against Wagner and his innovations, that no one could have guessed that later he himself would be considered an innovator and would be accused of imitating Wagner.

During his youth, after hearing Siegfried he wrote to a friend about the music of Mime: “It would have killed 411a cat and the horror of musical dissonances would melt rocks into omelettes.”

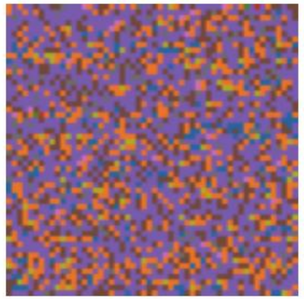
LIIII III IIIIIIIII IIII LII LIIIIII (LIII LIIII) (nnnn), III IIIII IIII IIII, II III III IIIIII IIII III IIIIII IIII, IIIIIIII IIIII IIII I III IIIII II IIIIII II IIIII II I IIIIII IIIIII IIIII II IIIII IIII IIIIII III IIII IIII. LI IIII, II III II IIII IIIIIII LIIIII III III IIIIIIIII, IIII II III IIIII IIII IIIIIII IIII IIIII II IIIIIII IIIII II IIIIIIIII II IIIIIIII III IIIII II IIIIIII II IIIIIIII LIIIII.

LIIIII III IIIII, IIIII IIIIIII LIIIIIIII II IIIII II I IIIIII IIIII III IIIII II LIII: “LI IIIII IIII IIIIII nnnI III III III IIIIII II IIIIIII IIIIIIIIII IIIII IIII IIIII IIII IIIIIIII.”

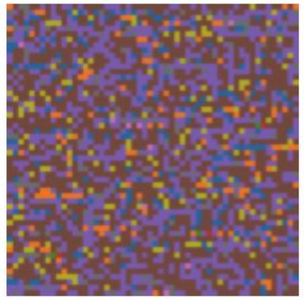
Stylometry Example III



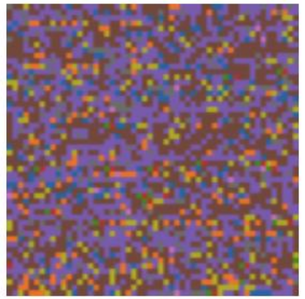
(a) *Sharing Her Crime*, M. A. Fleming



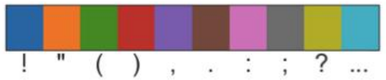
(b) *The Actress' Daughter*, M. A. Fleming

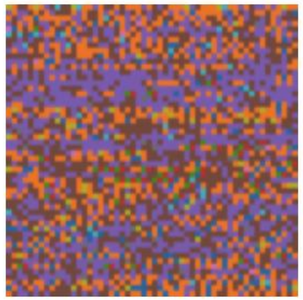


(c) *King Lear*, W. Shakespeare

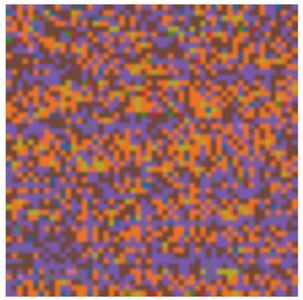


(d) *Hamlet*, W. Shakespeare





(e) *The History of Mr. Polly*, H. G. Wells



(f) *The Wheels of Chance*, H. G. Wells

Plots under MIT license (via GitHub).

Euro. Jnl of Applied Mathematics (2021), *vol.* 32, *pp.* 1069–1105
doi:10.1017/S0956792520000157

Pull out all the stops: Textual analysis via punctuation sequences

ALEXANDRA N. M. DARMON[1], MARYA BAZZI[1,2,3], SAM D. HOWISON[1] and MASON A. PORTER[1,4]

[1]*Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK*
[2]*The Alan Turing Institute, London NW1 2DB, UK*
[3]*Warwick Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK*
[4]*Department of Mathematics, University of California, Los Angeles, Los Angeles, California 90095, USA*
emails: alexandra.darmon@hotmail.fr, mbazzi@turing.ac.uk, howison@maths.ox.ac.uk, mason@math.ucla.edu

(*Received 31 December 2018; revised 16 January 2020; accepted 12 May 2020; first published online 21 September 2020*)

Is Preprocessing Essential?

A Debate for the Ages

"

The paper that is cited to back up the assumed improvement shows that it indeed increases performance for various NN setups, but decreases performance for RoBERTa. In case this requires clarification: modern LLMs use subword tokenizers that make 'classic' preprocessing steps unnecessary at best, but generally damage the performance (due to incorrect pre-tokenization).

"

The Task
Decides

- Ask yourself:
  - Do I think my preprocessing steps will add anything or remove information?
  - Consider using both types of representations!
  - Assess **quantitatively**, and diligently!

Any Questions?