Annotations

CS 485, Fall 2024
Applications of Natural Language Processing

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst

$$(x, \underset{\uparrow\uparrow\uparrow}{y})$$

- If you have labels, we know how to do:
  - Train a ML model
  - Evaluation metrics
  - Avoid overfitting

- But
  - Where do we get the labels ("annotations")?
  - Are these "gold standard" labels any good?

Tasks and getting labels

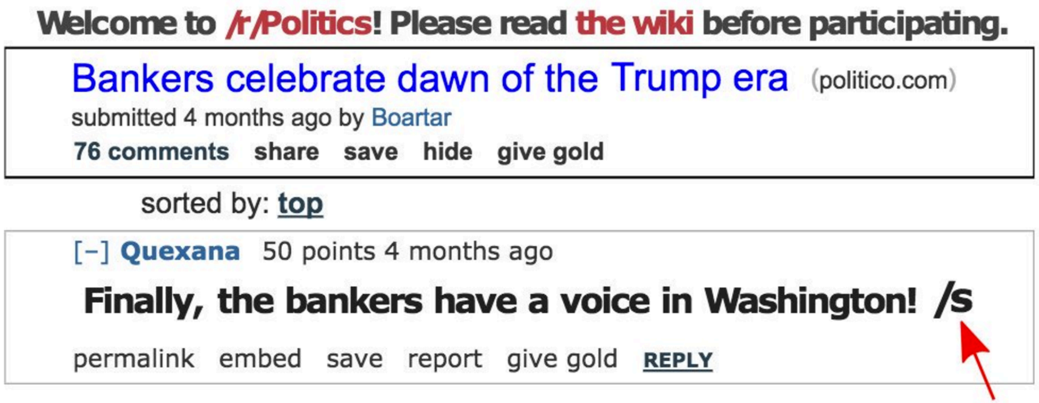
- Define a classification task that you'd like a model to do
- Then you need text and labels
  - 0. What's the text data? (upcoming)
  - 1. Natural annotations — information you can automatically retrieve about a text
  - 2. New human annotations — get people to manually create labels for a sample of texts!

- Natural annotations
    - Metadata - information associated with text document, but not in text itself
        - *Examples?*

- Stars on an online review
- Categories
- Info abt speaker
    - Gender, age, ethnicity...
- Age pred.

- # Views, Upvotes,

- Natural annotations
  - Metadata - information associated with text document, but not in text itself
  - Clever patterns from text itself



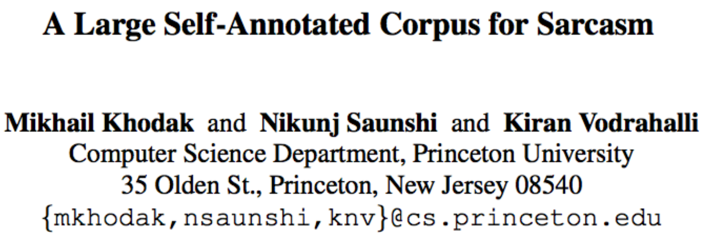
Welcome to /r/Politics! Please read the wiki before participating.

Bankers celebrate dawn of the Trump era (politico.com)

submitted 4 months ago by Boartar

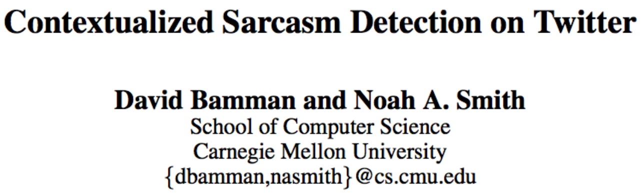
Quexana 50 points 4 months ago

Finally, the bankers have a voice in Washington! /s



A Large Self-Annotated Corpus for Sarcasm

Mikhail Khodak and Nikunj Saunshi and Kiran Vodrahalli
Computer Science Department, Princeton University
35 Olden St., Princeton, New Jersey 08540
{mkhodak,nsaunshi,knv}@cs.princeton.edu

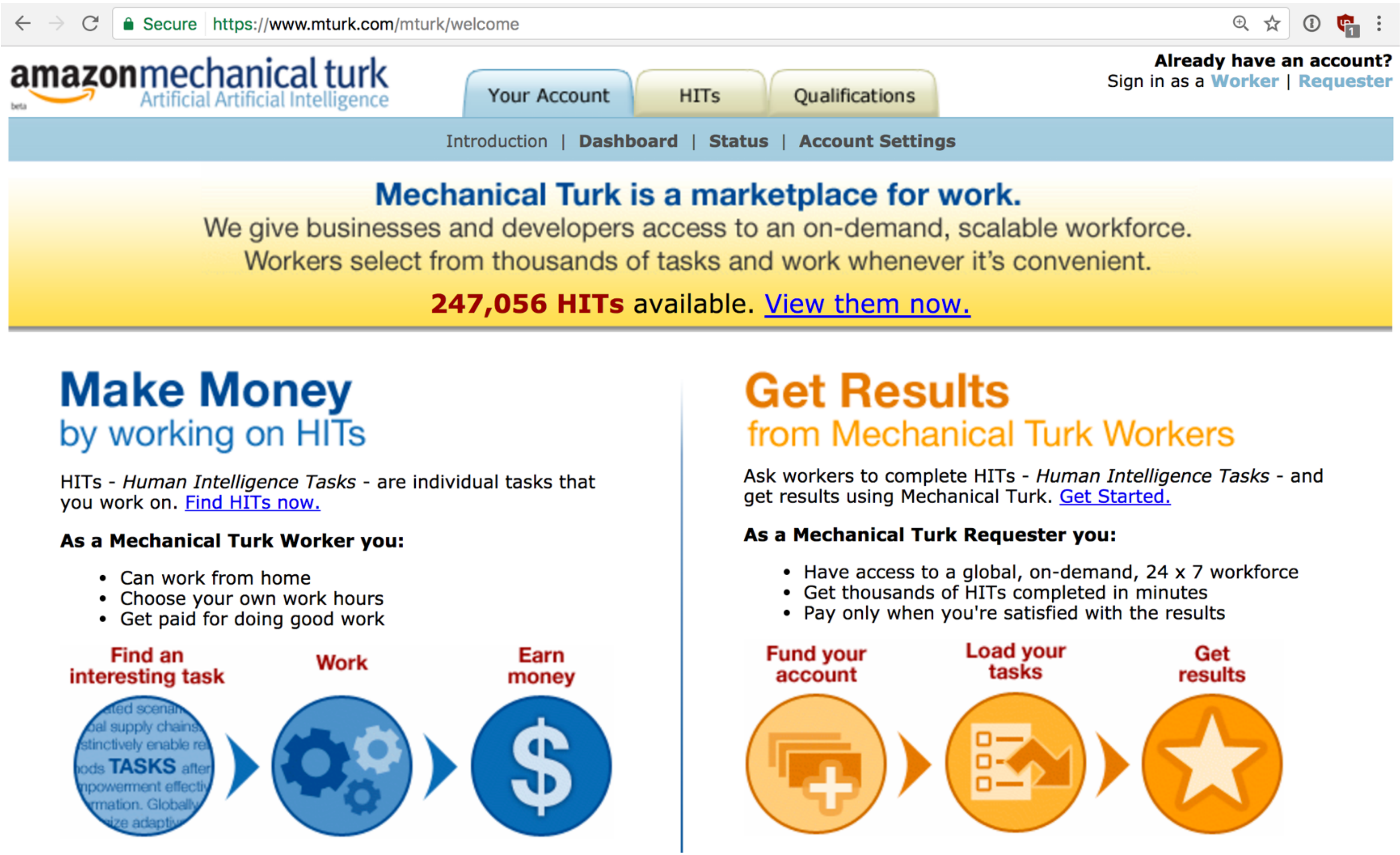


Contextualized Sarcasm Detection on Twitter

David Bamman and Noah A. Smith
School of Computer Science
Carnegie Mellon University
{dbamman,nasmith}@cs.cmu.edu

Collecting new annotations

- Steps
  1. Design a human annotation (labeling) task,
  2. Find annotators
  3. Collect the annotations
- New human annotations
  - Yourself & collaborators
  - Your friends
  - Hire people locally
  - Hire people online
    - Mechanical Turk — most commonly used crowdsourcing site
    - Many others (Prolific, Upwork, etc.)


Secure
https://www.mturk.com/mturk/welcome
amazonmechanical turk
Artificial Artificial Intelligence
beta
Your Account
HITs
Qualifications
Already have an account?
Sign in as a Worker | Requester
Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

247,056 HITs available. View them now.

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results
from Mechanical Turk Workers

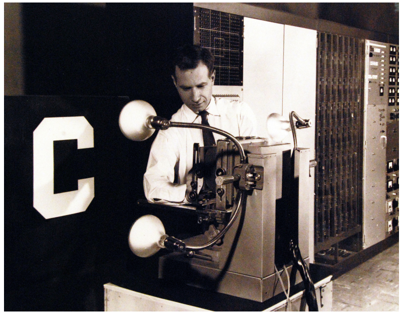
Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

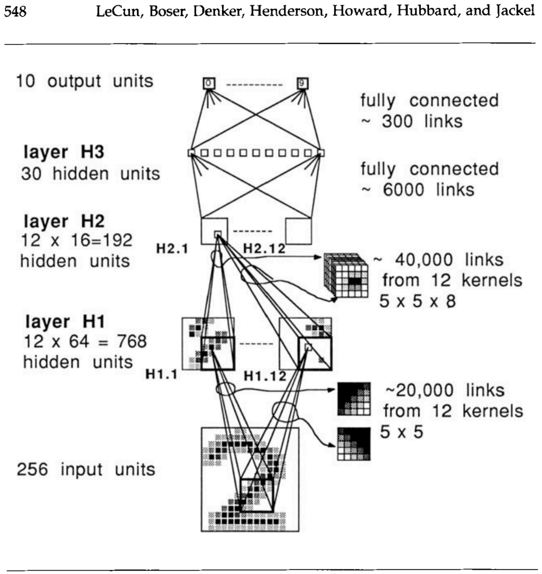
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

- Human behavioral data is the key factor in recently reviving neural network modeling (initially in computational vision)

1957
Perceptron
(~log. reg.)

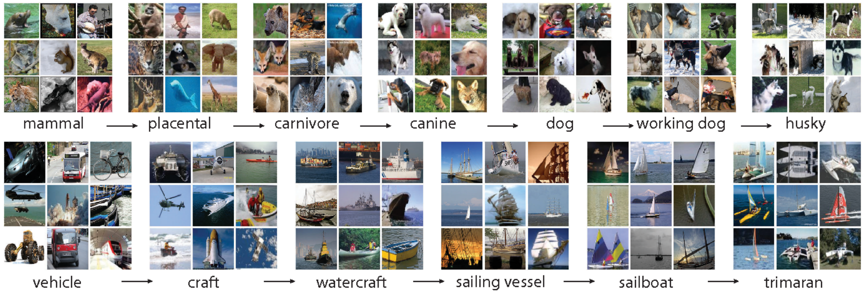


1989
Backprop &
convolutional NN



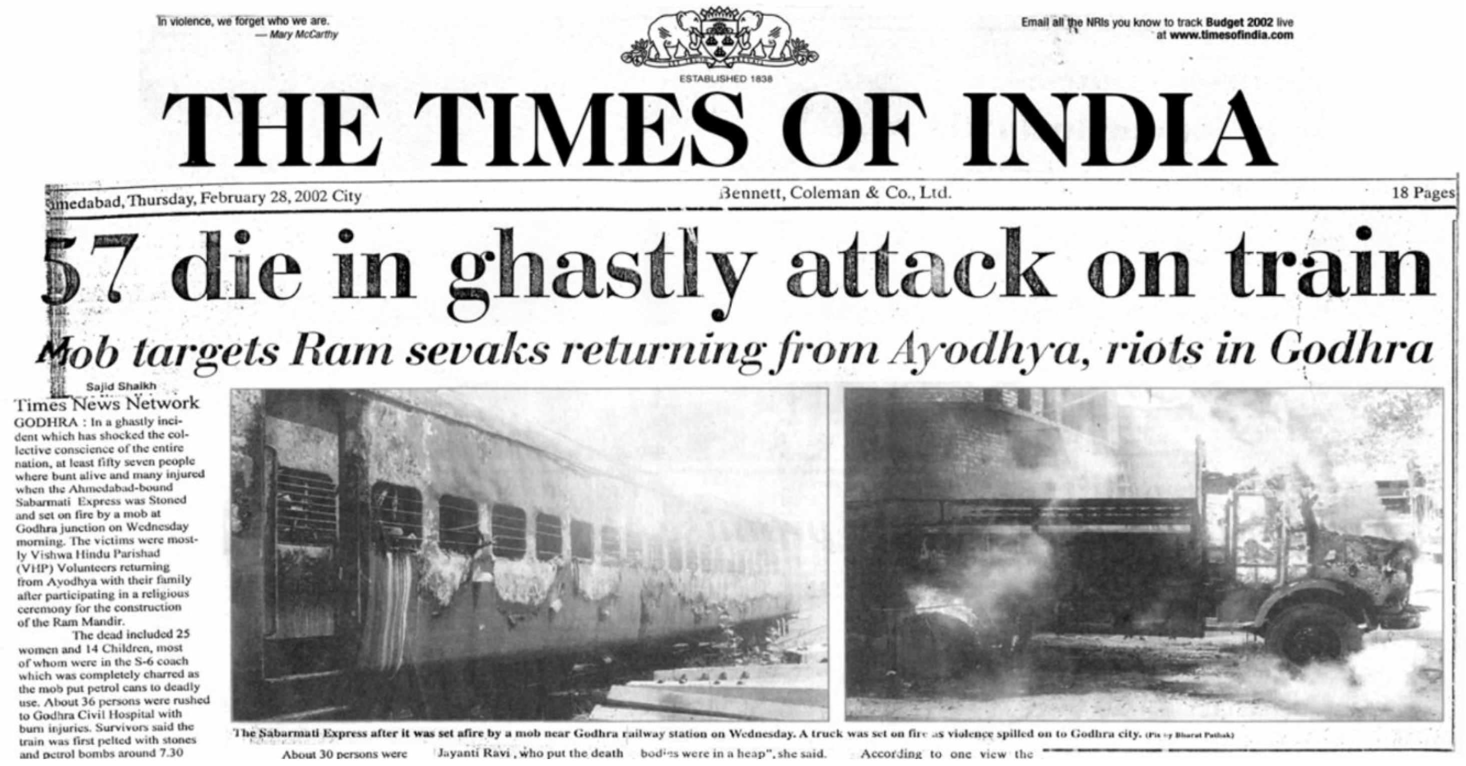
548 LeCun, Boser, Denker, Henderson, Howard, Hubbard, and Jackel

2012
ImageNet data
for CNN training



Millions of labeled objects in images, collected via crowdsourcing (MTurk)
Revolutionized CV by using nearly the same model from 1989!

- *Times of India* articles from March 2002
- Filter using place name keywords
- 1,257 stories
- 21,391 sentences



In violence, we forget who we are.
— *Mary McCarthy*

Email all the NRIs you know to track **Budget 2002** live
at **www.timesofindia.com**

ESTABLISHED 1838

THE TIMES OF INDIA

medabad, Thursday, February 28, 2002 City | Bennett, Coleman & Co., Ltd. | 18 Pages

57 die in ghastly attack on train

Mob targets Ram sevaks returning from Ayodhya, riots in Godhra

Sajid Shaikh
Times News Network

GODHRA : In a ghastly incident which has shocked the collective conscience of the entire nation, at least fifty seven people where bunt alive and many injured when the Ahmedabad-bound Sabarmati Express was Stoned and set on fire by a mob at Godhra junction on Wednesday morning. The victims were mostly Vishwa Hindu Parishad (VHP) Volunteers returning from Ayodhya with their family after participating in a religious ceremony for the construction of the Ram Mandir.

The dead included 25 women and 14 Children, most of whom were in the S-6 coach which was completely charred as the mob put petrol cans to deadly use. About 36 persons were rushed to Godhra Civil Hospital with burn injuries. Survivors said the train was first pelted with stones and petrol bombs around 7.30

The Sabarmati Express after it was set afire by a mob near Godhra railway station on Wednesday. A truck was set on fire as violence spilled on to Godhra city. (Pix by Bharat Pathak)

About 30 persons were

Jayanti Ravi , who put the death

bodies were in a heap", she said.

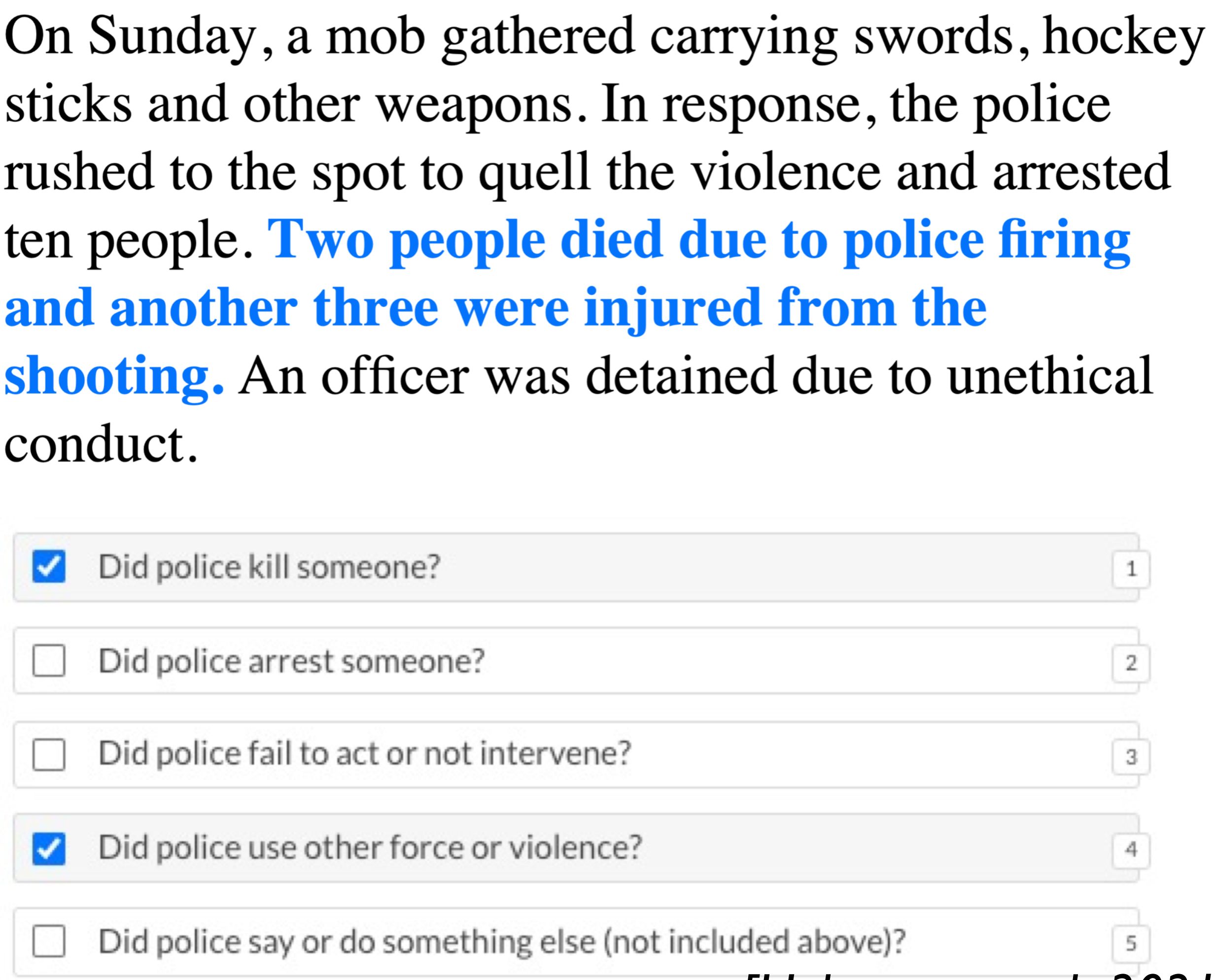
According to one view the

- Goal: evaluate new NLP methods to automatically extract violence events to assist political science analysis
- High-expertise approach: hired a team of UMass student annotators; weekly meetings over a semester

[Halterman et al., 2021]

Annotations Example: event detection as classification

- Boolean QA version of event identification: each event class is a question
  - Annotators found much easier than argument span identification
- Trained annotators
- Sentence-level annotations
- Sentences in context



On Sunday, a mob gathered carrying swords, hockey sticks and other weapons. In response, the police rushed to the spot to quell the violence and arrested ten people. **Two people died due to police firing and another three were injured from the shooting.** An officer was detained due to unethical conduct.

☑ Did police kill someone? 1

□ Did police arrest someone? 2

□ Did police fail to act or not intervene? 3

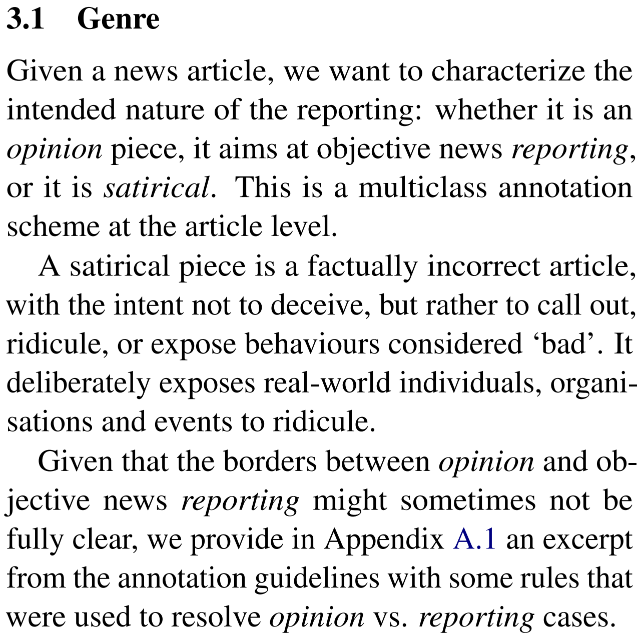
☑ Did police use other force or violence? 4

□ Did police say or do something else (not included above)? 5

[Halterman et al., 2021]

Annotation example: framing/persuasion methods

- Annotating multilingual news articles from 2020-2022

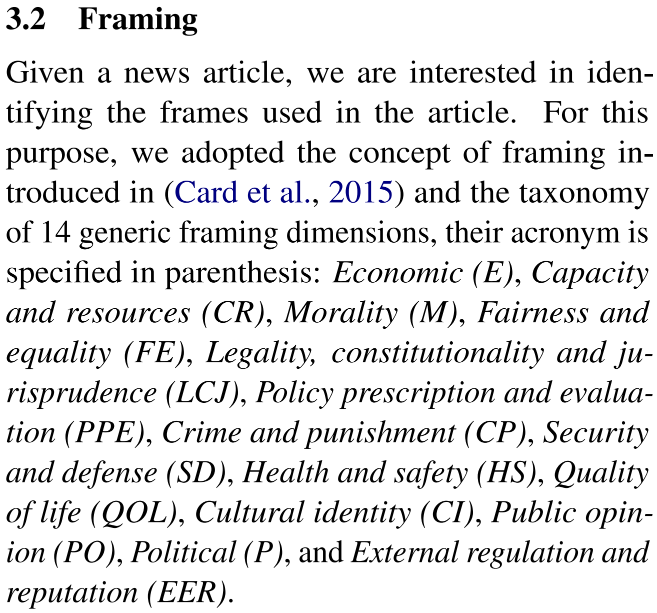


3.1 Genre

Given a news article, we want to characterize the intended nature of the reporting: whether it is an *opinion* piece, it aims at objective news *reporting*, or it is *satirical*. This is a multiclass annotation scheme at the article level.

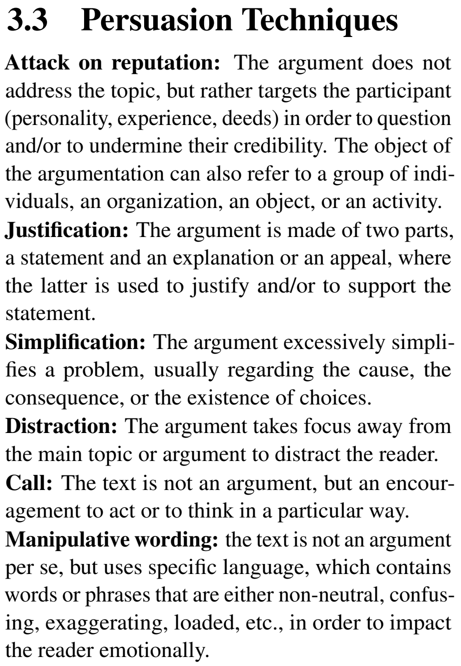
A satirical piece is a factually incorrect article, with the intent not to deceive, but rather to call out, ridicule, or expose behaviours considered 'bad'. It deliberately exposes real-world individuals, organisations and events to ridicule.

Given that the borders between *opinion* and objective news *reporting* might sometimes not be fully clear, we provide in Appendix A.1 an excerpt from the annotation guidelines with some rules that were used to resolve *opinion* vs. *reporting* cases.



3.2 Framing

Given a news article, we are interested in identifying the frames used in the article. For this purpose, we adopted the concept of framing introduced in (Card et al., 2015) and the taxonomy of 14 generic framing dimensions, their acronym is specified in parenthesis: *Economic (E)*, *Capacity and resources (CR)*, *Morality (M)*, *Fairness and equality (FE)*, *Legality, constitutionality and jurisprudence (LCJ)*, *Policy prescription and evaluation (PPE)*, *Crime and punishment (CP)*, *Security and defense (SD)*, *Health and safety (HS)*, *Quality of life (QOL)*, *Cultural identity (CI)*, *Public opinion (PO)*, *Political (P)*, and *External regulation and reputation (EER)*.

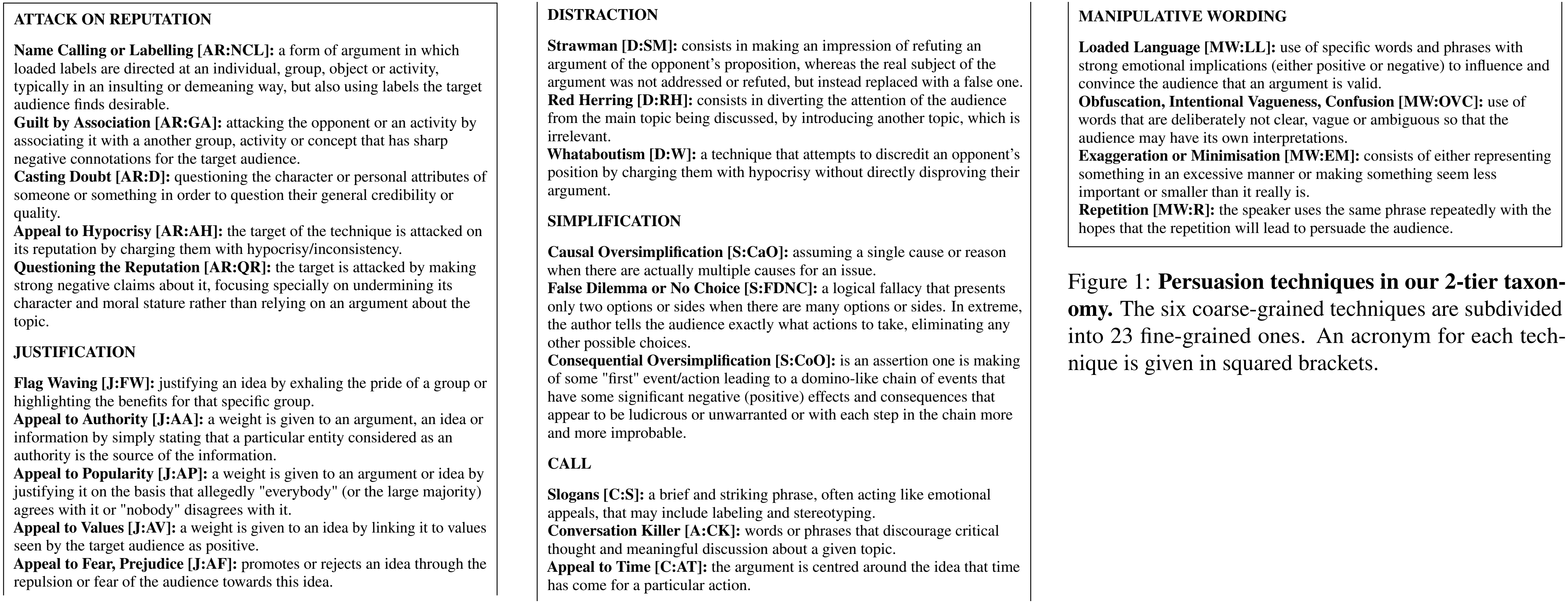


3.3 Persuasion Techniques

Attack on reputation: The argument does not address the topic, but rather targets the participant (personality, experience, deeds) in order to question and/or to undermine their credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity.
Justification: The argument is made of two parts, a statement and an explanation or an appeal, where the latter is used to justify and/or to support the statement.
Simplification: The argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices.
Distraction: The argument takes focus away from the main topic or argument to distract the reader.
Call: The text is not an argument, but an encouragement to act or to think in a particular way.
Manipulative wording: the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

[Piskorski et al. 2023]

Annotation example: framing/persuasion methods



ATTACK ON REPUTATION

Name Calling or Labelling [AR:NCL]: a form of argument in which loaded labels are directed at an individual, group, object or activity, typically in an insulting or demeaning way, but also using labels the target audience finds desirable.
Guilt by Association [AR:GA]: attacking the opponent or an activity by associating it with a another group, activity or concept that has sharp negative connotations for the target audience.
Casting Doubt [AR:D]: questioning the character or personal attributes of someone or something in order to question their general credibility or quality.
Appeal to Hypocrisy [AR:AH]: the target of the technique is attacked on its reputation by charging them with hypocrisy/inconsistency.
Questioning the Reputation [AR:QR]: the target is attacked by making strong negative claims about it, focusing specially on undermining its character and moral stature rather than relying on an argument about the topic.

JUSTIFICATION

Flag Waving [J:FW]: justifying an idea by exhaling the pride of a group or highlighting the benefits for that specific group.
Appeal to Authority [J:AA]: a weight is given to an argument, an idea or information by simply stating that a particular entity considered as an authority is the source of the information.
Appeal to Popularity [J:AP]: a weight is given to an argument or idea by justifying it on the basis that allegedly "everybody" (or the large majority) agrees with it or "nobody" disagrees with it.
Appeal to Values [J:AV]: a weight is given to an idea by linking it to values seen by the target audience as positive.
Appeal to Fear, Prejudice [J:AF]: promotes or rejects an idea through the repulsion or fear of the audience towards this idea.

DISTRACTION

Strawman [D:SM]: consists in making an impression of refuting an argument of the opponent's proposition, whereas the real subject of the argument was not addressed or refuted, but instead replaced with a false one.
Red Herring [D:RH]: consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic, which is irrelevant.
Whataboutism [D:W]: a technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

SIMPLIFICATION

Causal Oversimplification [S:CaO]: assuming a single cause or reason when there are actually multiple causes for an issue.
False Dilemma or No Choice [S:FDNC]: a logical fallacy that presents only two options or sides when there are many options or sides. In extreme, the author tells the audience exactly what actions to take, eliminating any other possible choices.
Consequential Oversimplification [S:CoO]: is an assertion one is making of some "first" event/action leading to a domino-like chain of events that have some significant negative (positive) effects and consequences that appear to be ludicrous or unwarranted or with each step in the chain more and more improbable.

CALL

Slogans [C:S]: a brief and striking phrase, often acting like emotional appeals, that may include labeling and stereotyping.
Conversation Killer [A:CK]: words or phrases that discourage critical thought and meaningful discussion about a given topic.
Appeal to Time [C:AT]: the argument is centred around the idea that time has come for a particular action.

MANIPULATIVE WORDING

Loaded Language [MW:LL]: use of specific words and phrases with strong emotional implications (either positive or negative) to influence and convince the audience that an argument is valid.
Obfuscation, Intentional Vagueness, Confusion [MW:OVC]: use of words that are deliberately not clear, vague or ambiguous so that the audience may have its own interpretations.
Exaggeration or Minimisation [MW:EM]: consists of either representing something in an excessive manner or making something seem less important or smaller than it really is.
Repetition [MW:R]: the speaker uses the same phrase repeatedly with the hopes that the repetition will lead to persuade the audience.

Figure 1: **Persuasion techniques in our 2-tier taxonomy.** The six coarse-grained techniques are subdivided into 23 fine-grained ones. An acronym for each technique is given in squared brackets.

[Piskorski et al. 2023]

High-quality annotation guidelines for complex tasks... can get complicated!

A Annotation Guidelines

This appendix provides an excerpt of the annotation guidelines (Piskorski et al., 2023a) related to news genre and persuasion techniques.

A.1 News Genre

- *opinion* versus *reporting*: in the case of news articles that contain citations and opinions of others (i.e., not of the author), the decision whether to label such article as opinion or reporting should in principle depend on what the reader thinks the intent of the author of the article was. In order to make this decision simpler, the following rules were applied:
  - articles that contain even a single sentence (could be even the title) that is an opinion of the author or suggests that the author has some opinion on the specific matter should be labelled as *opinion*,
  - articles containing a speech or an interview with a **single** politician or expert, who provides her/his opinions should be labelled as *opinion*,
  - articles that "report" what a **single** politician or expert said in an interview, conference, debate, etc. should be labelled as *opinion* as well,
  - articles that provide a comprehensive overview (spectrum) of what many different politicians and experts said on a specific matter (e.g., in a debate), including their opinions, and without any opinion of the author, should be labelled as *reporting*,
  - articles that provide a comprehensive overview (spectrum) of what many different politicians and experts said on a specific matter (e.g., in a debate), including their opinions, and with some opinion or analysis of the author (the author might try to tell a story), should be labelled as *opinion* ,
  - commentaries and analysis articles should be labelled as *opinion*.
- *satire*: A news article that contains some small text fragment, e.g., a sentence, which appears satirical **is not supposed to be annotated as** *satire*.

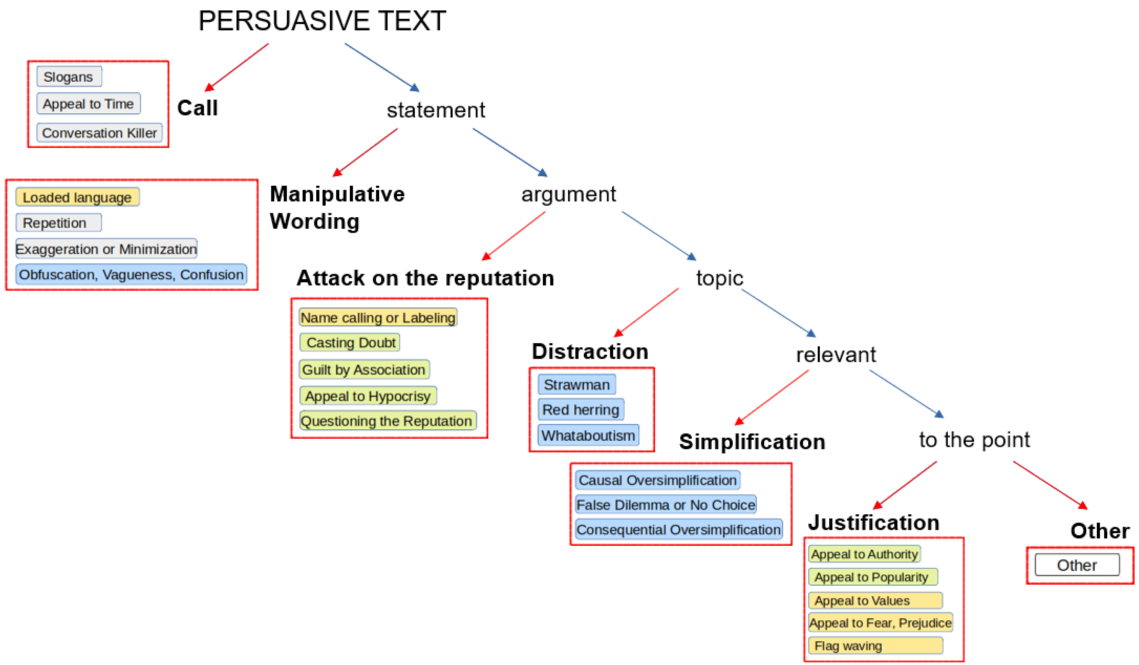
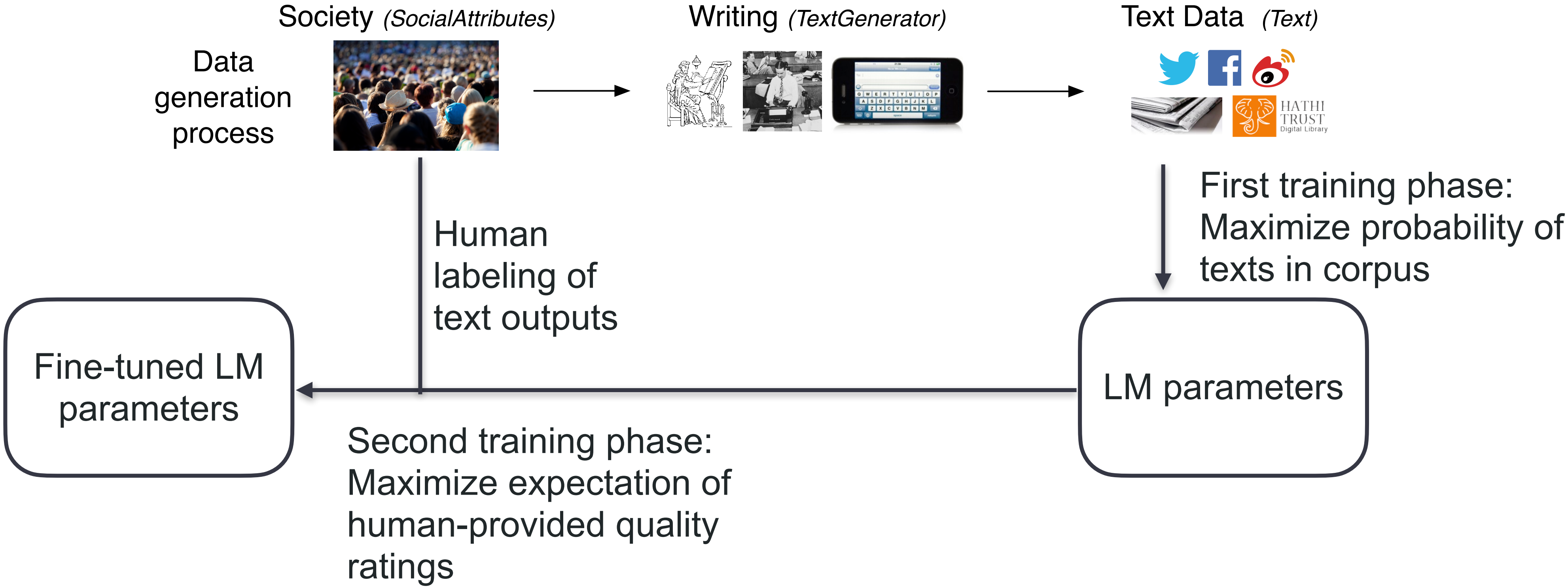


Figure 4: Decision diagram to determine which high-level approach is used in a text. The fine-grained techniques are marked in color, in an attempt to reflect the rhetorical dimension: (a) ethos, i.e., appeal to authority (green), (b) logos, i.e., appeal to logic (blue), and (c) pathos, e.e., appeal to emotions (yellow).

[Piskorski et al. 2023]

Human labeling is key to ChatGPT



Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*

Text Data *(Text)*

First training phase: Maximize probability of texts in corpus

LM parameters

Human labeling of text outputs

Fine-tuned LM parameters

Second training phase: Maximize expectation of human-provided quality ratings

[Ouyang et al., 2022, Taori et al. 2023]

Annotation process

1. Design a human annotation (labeling) task
2. Find annotators
3. Collect the annotations

- To pilot a new task, requires an iterative process
    - Look at data to see what's possible
    - Conceptualize the task, try it yourself
    - Write annotation guidelines
    - Have annotators try to do it. Where do they disagree? What feedback do they have?
    - Revise guidelines and repeat
- Checking annotation quality - do you trust your annotators?
    - Crowdsourcing sites can be tricky
- If you don't do all this, your labeled data will have lots of unclear, arbitrary, and implicit decisions inside of it

"Content Analysis"

Annotation is paramount

- Supervised learning is one of the most reliable approaches to NLP and artificial intelligence more generally.
- Alternative view: it’s *human* intelligence, through the human-supplied training labels, that’s at the heart of it. Supervised NLP merely extends a noisier, less-accurate version to more data.
- If we still want it: we need a plan to get good annotations!

Interannotator agreement

A1 y_1	A2 y_1	$\hat{y}_1$
A1 y_2	A2 y_2	$\hat{y}_2$
A1 y_3	A2 y_3	$\hat{y}_3$

IAA = 90%

BoW LR
Acc =

- How “real” is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
- Question: can an NLP system's accuracy be higher than the human agreement rate?

Lower:

NLP is hard

Humans are smart

Higher: Big train set

Humans disagree on edge cases

Interannotator agreement

- How “real” is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
- Question: can an NLP system's accuracy be higher than the human agreement rate?

- The conventional view: IAA (human performance) is the upper bound for machine performance
  - What affects IAA? Difficulty of task, human training, human motivation/effort....

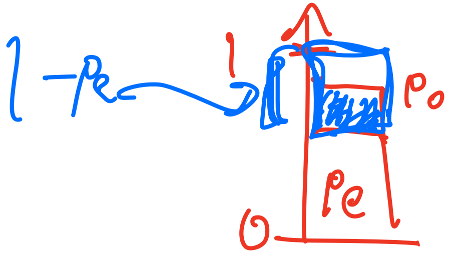
↳ inh. ambig

Knowledge

- Cultural background / Expertise
- Ages
- Lang/Comm abilities/style. Dialects, meaning

- Fatigue
- Motiv, attn.
- Pay

Cohen's Kappa for IAA



- If some classes predominate, raw agreement rate may be misleading
- Idea: normalize accuracy (agreement) rate such that answering randomly = 0.
  - From psychology / psychometrics / content analysis
- **Chance-adjusted agreement:** Take p(k) from ALL annos

Binary (k=0 or k=1): $p_e = p(k=0)\,p(k=0) + p(k=1)\,p(k=1)$

Multiclass: $p_e = \sum_{k=1}^{K} [p(k)]^2$

p_o: **o**bserved agreement rate

"IAA"

p_e: **e**xpected (by chance) rate

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

$p(0) = .1 \quad p(1) = .9$

$p_e = .01 + .81 = .82$

$p_o = .82 \Rightarrow \kappa = 0$

Other chanced-adjusted metrics: Fleiss, Krippendorff... see reading

Exercise

D ND

6 1

D 7

2 D

$$p(1) = \frac{14}{20} = .7$$

$$p(0) = .3$$

$$P_e = .3^2 + .7^2 = .58$$

$$p_o = .8$$

$$K = \frac{.8 - .58}{1 - .58}$$

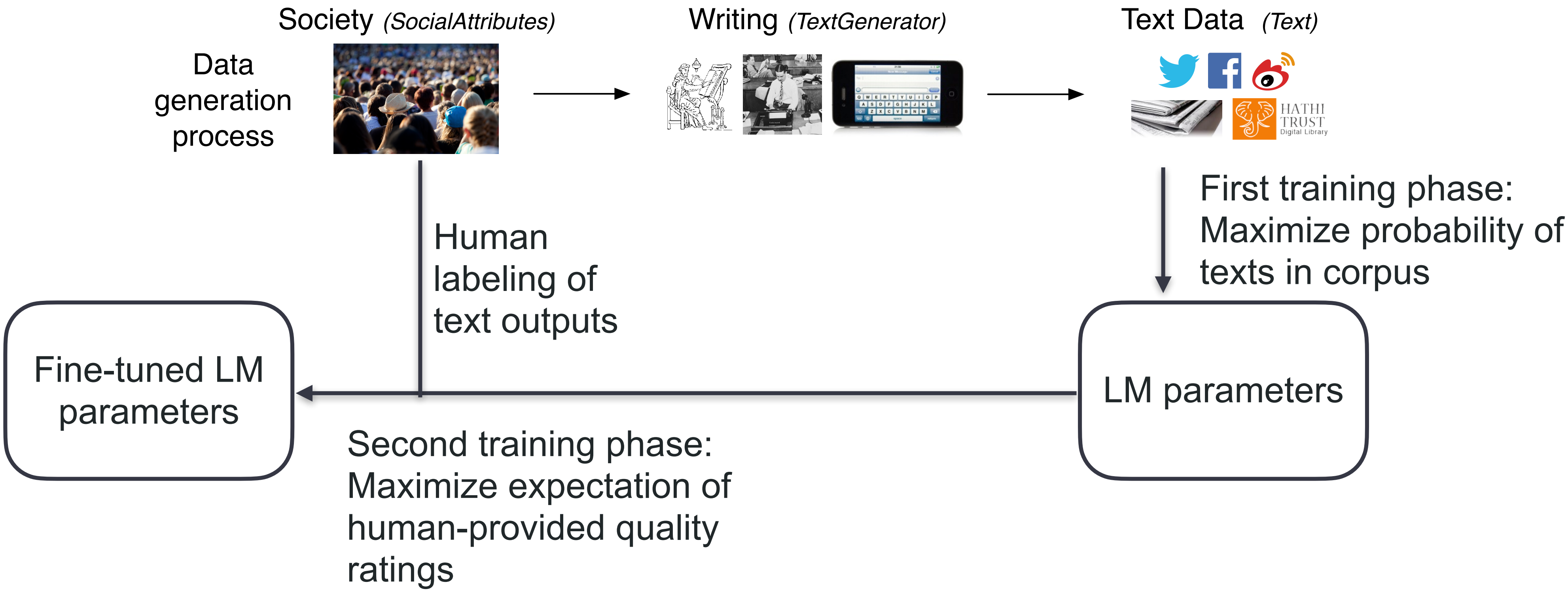
$$= .52$$

Do I have enough labels?

- For training, typically thousands of annotations are necessary for reasonable performance
  - Current work: how to usefully make NLP models with <10 or <100 training examples. "Few-shot learning"
- For evaluation, fewer is ok (but watch statistical significance! Next lecture.)
- Exact amounts are difficult to know in advance. Can do a **learning curve** to estimate if more annotations will be useful.

When is annotating ethical?

Human labeling is key to ChatGPT



Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*

Text Data *(Text)*

First training phase: Maximize probability of texts in corpus

LM parameters

Human labeling of text outputs

Fine-tuned LM parameters

Second training phase: Maximize expectation of human-provided quality ratings

[Ouyang et al., 2022, Taori et al. 2023]

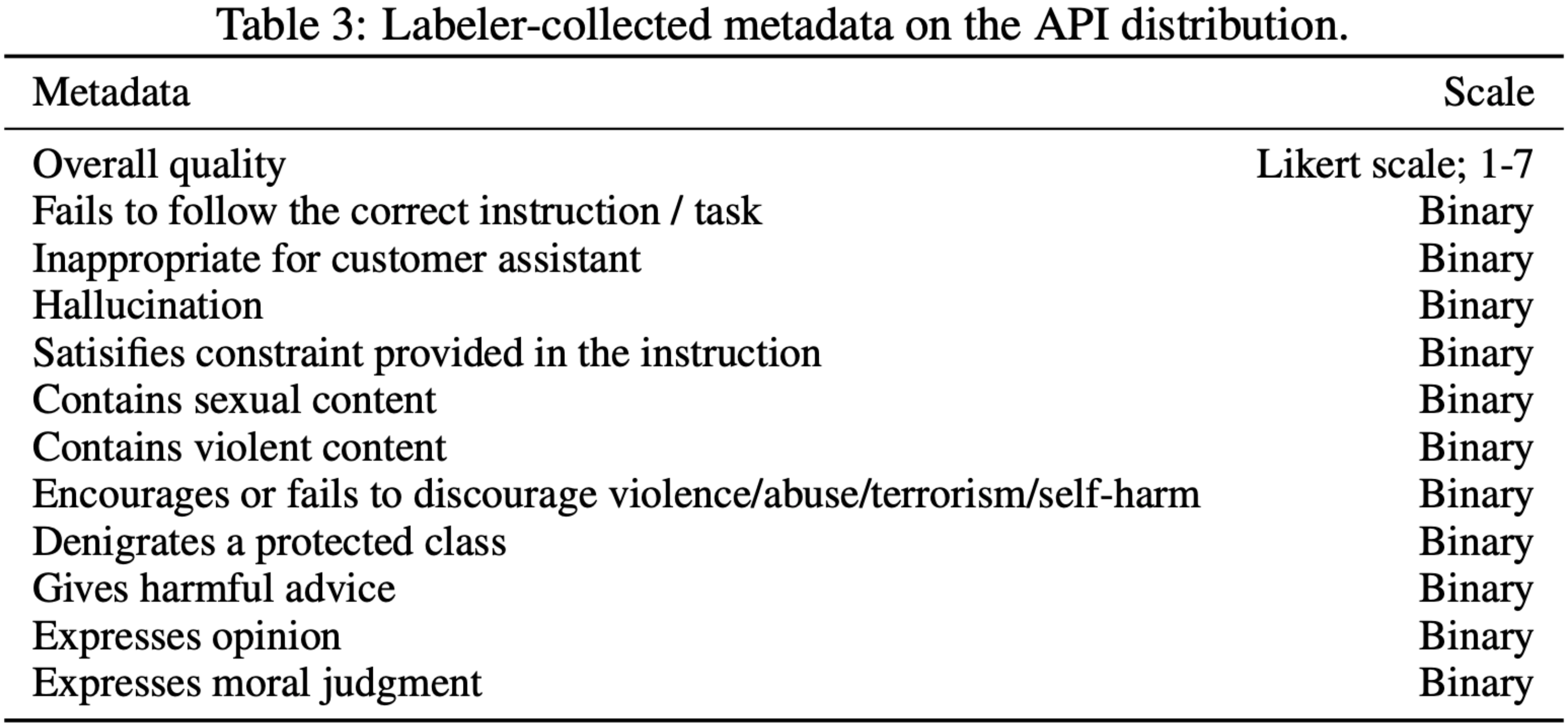
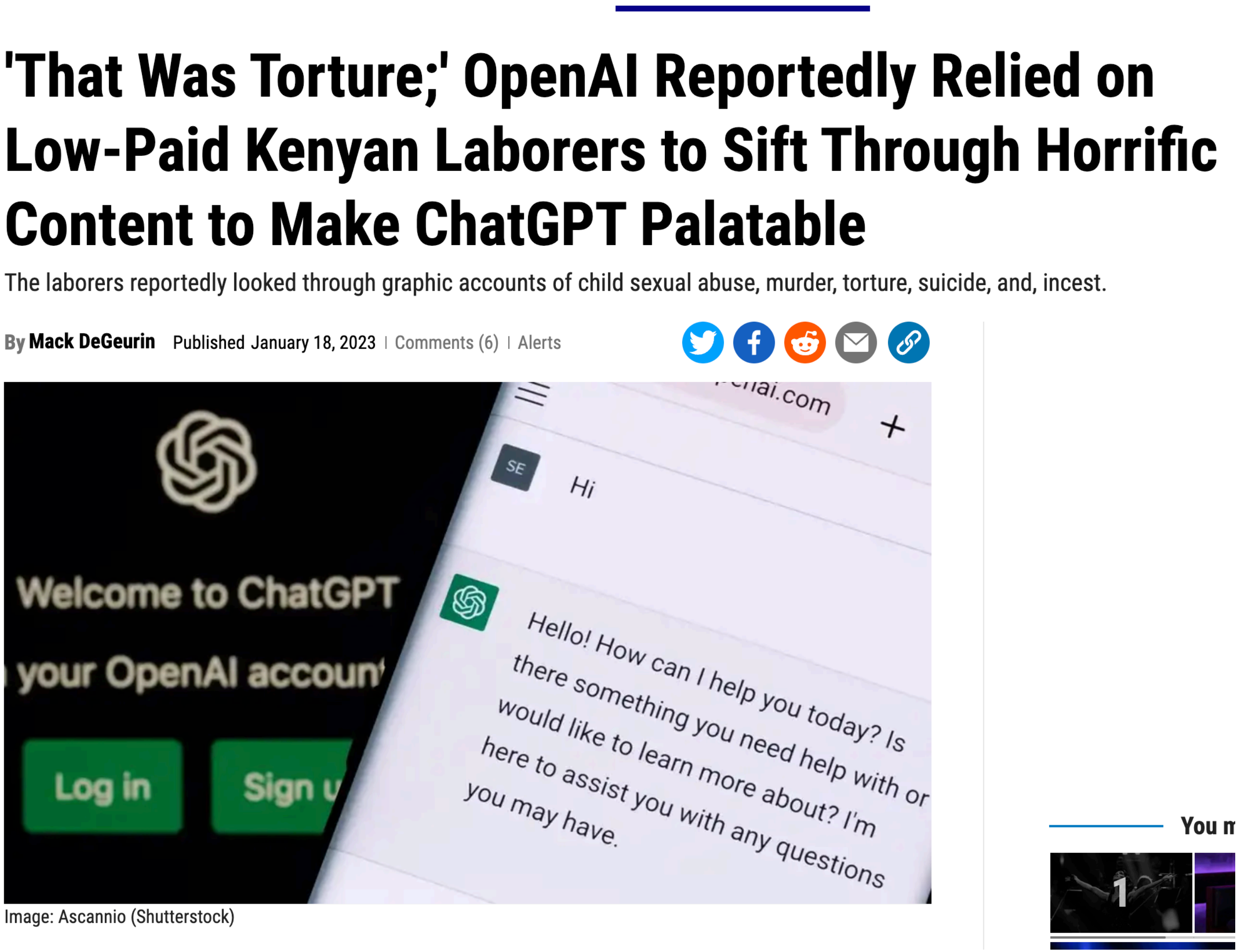


Table 3: Labeler-collected metadata on the API distribution.

Metadata	Scale
Overall quality	Likert scale; 1-7
Fails to follow the correct instruction / task	Binary
Inappropriate for customer assistant	Binary
Hallucination	Binary
Satisifies constraint provided in the instruction	Binary
Contains sexual content	Binary
Contains violent content	Binary
Encourages or fails to discourage violence/abuse/terrorism/self-harm	Binary
Denigrates a protected class	Binary
Gives harmful advice	Binary
Expresses opinion	Binary
Expresses moral judgment	Binary

[Ouyang et al., 2022]



'That Was Torture;' OpenAI Reportedly Relied on Low-Paid Kenyan Laborers to Sift Through Horrific Content to Make ChatGPT Palatable

The laborers reportedly looked through graphic accounts of child sexual abuse, murder, torture, suicide, and, incest.

By **Mack DeGeurin** Published January 18, 2023 | Comments (6) | Alerts

Welcome to ChatGPT
your OpenAI account
Log in
Sign u
enai.com
SE
Hi
Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any questions you may have.
You m
1

Image: Ascannio (Shutterstock)

Data Annotations: Conclusions

- Manual data annotation is key to many, if not most, NLP applications
  - ... because supervised learning needs it, and SL is a very effective approach for NLP
- Collecting annotations is a human process and worrying about the humans is key to high-quality annotations
  - Is the task reasonable? Well-specified? Realistic?
  - Measuring agreement as an imperfect proxy for annotation quality
  - Speed, price, feasibility of the work?
  - Is the work ethical?