



DATA QUALITY AND ANNOTATIONS

Sandeep Soni

03/28/2024

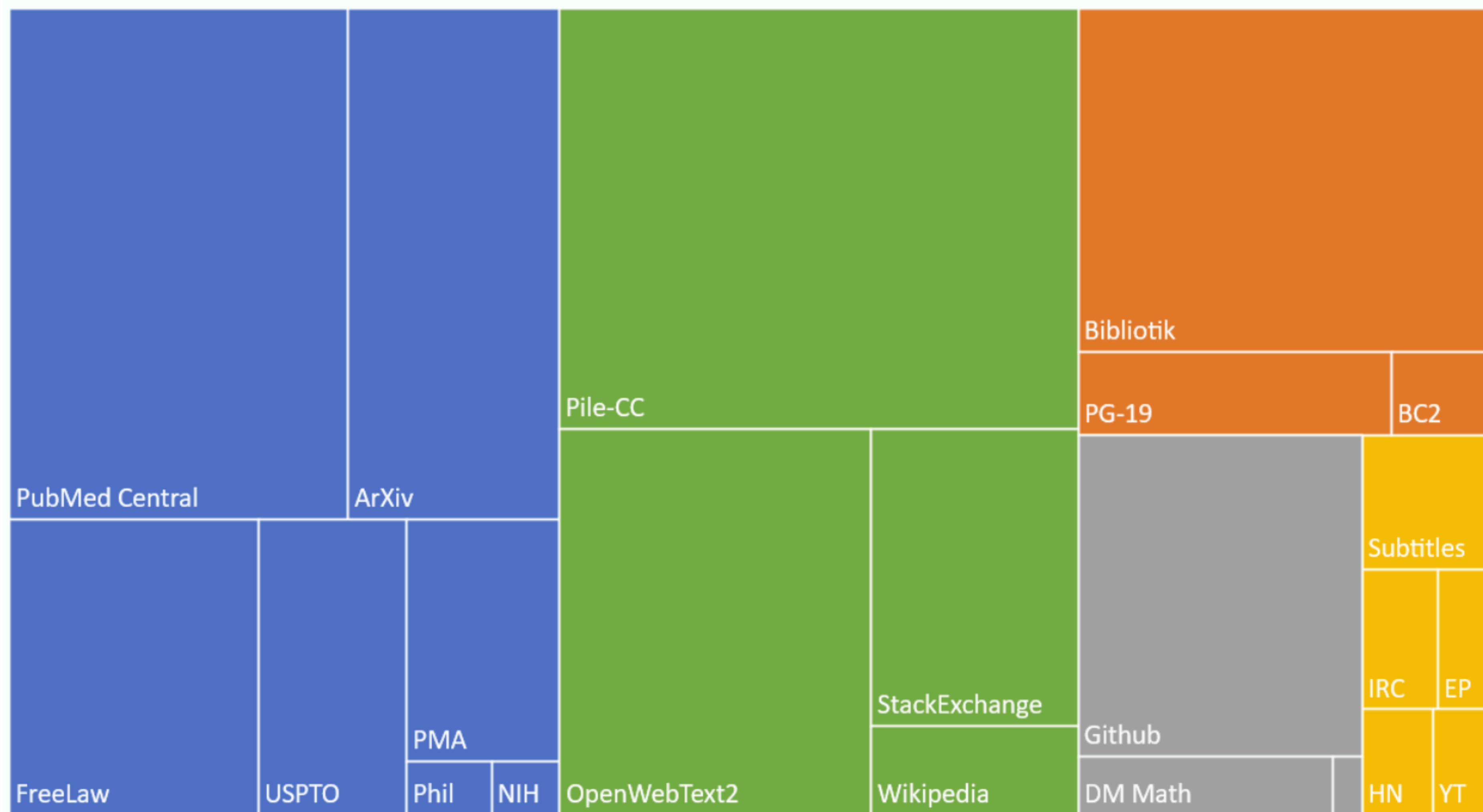
STORY SO FAR

- Large language models operate in two modes:
 - Pretrained: Support analysis using the internal representations or queries through prompts
 - Finetuned: Support model parameter updates on a small supervised task-specific data

It's important to be aware of considerations about data that we feed in both the cases

Composition of the Pile by Category

■ Academic ■ Internet ■ Prose ■ Dialogue ■ Misc



Source: Gao et. al. (2020) The Pile: An 800GB Dataset of Diverse Text for Language Modeling

DOCUMENTATION

- Pretraining corpora not always sufficiently documented
 - Provenance
 - Source composition
 - Filtering
 - Data contamination

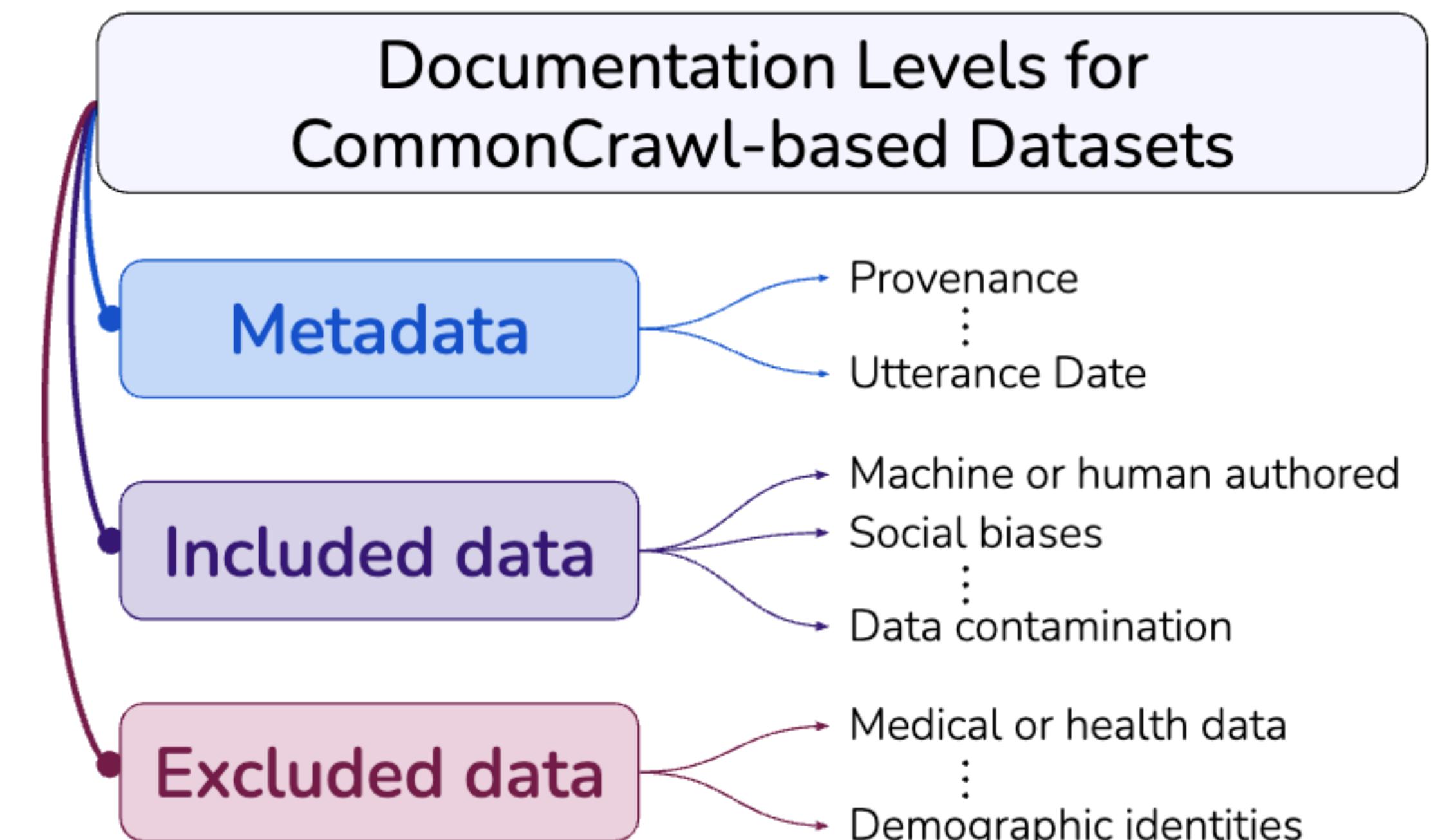


Figure 1: We advocate for three levels of documentation when creating web-crawled corpora. On the right, we include some example of types of documentation that we provide for the C4.EN dataset.

Dodge, Jesse, et al. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

UNDESIRABLE CONTENT

- Pretraining corpora can contain undesirable content such as hate speech, misogynistic and sexual content.
- Quality filters not always effective at removing such content.

Approach	Text
HateSonar	Their US/Euro plan put in your face: demonic jews hate white goyim! Such sick and twisted people, white people are.
Delimit	they are only stupid arab from wp-ar haha Yeah, dumb ass n*gger †
N-gram	nude attention whore asian bastards In America all male look like this homo

	Entire Sample	High Quality	Low Quality
1+ sexual n-grams	2.36%	1.81%	3.97%
3+ sexual n-grams	1.36%	0.42%	3.11%
10+ sexual n-grams	0.73%	0.08%	1.98%
1+ hate n-grams	17.78%	18.95%	17.19%
3+hate n-grams	6.38%	6.19%	8.26%
10+ hate n-grams	1.16%	1.17%	1.70%
Hate speech (Sonar)	4.02%	3.47%	5.09%
Hate speech (Delimit)	5.24%	5.77%	5.66%

UPSTREAM AND DOWNSTREAM FIXES

- **Upstream:** Pretraining data should be cleaned, curated, and of high quality
- **Downstream:** Finetuning, Human feedback to steer the model output.

QUALITY FILTERS

- Input data to large language models is curated by applying a quality filter
- What does quality mean?
- Variety of approaches such as heuristics, classification, stop word lists, etc
- Typically someone's values, beliefs, morality is encoded in the design of quality filters

Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection

Suchin Gururangan[†] Dallas Card[◊] Sarah K. Dreier[♡] Emily K. Gade[♣]

Leroy Z. Wang[†] Zeyu Wang[†] Luke Zettlemoyer[†] Noah A. Smith^{†♣}

[†]University of Washington [◊] University of Michigan [♡]University of New Mexico

[♣]Emory University [♠]Allen Institute for AI

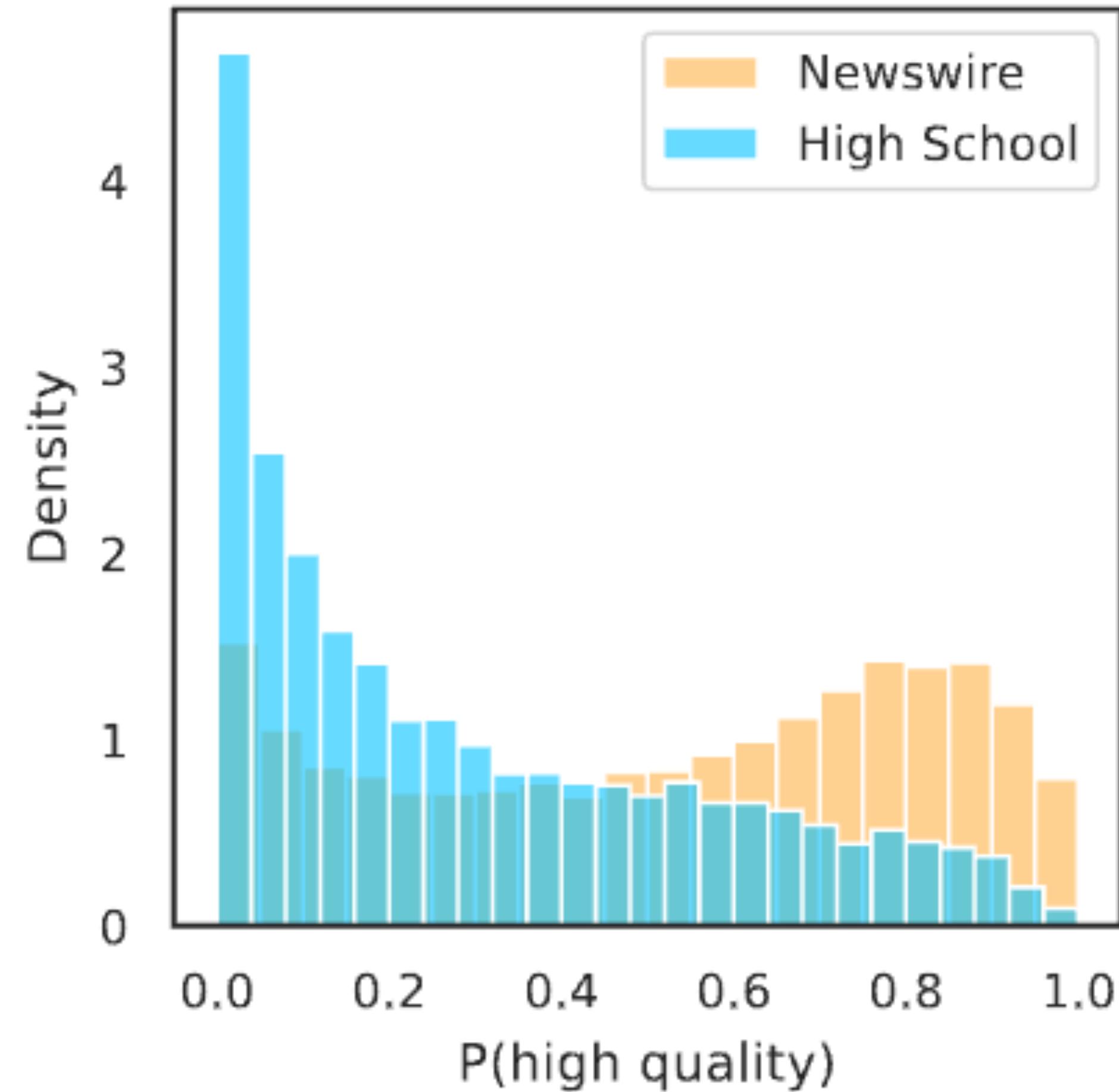
{sg01, zwan4, lsz, nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

Abstract

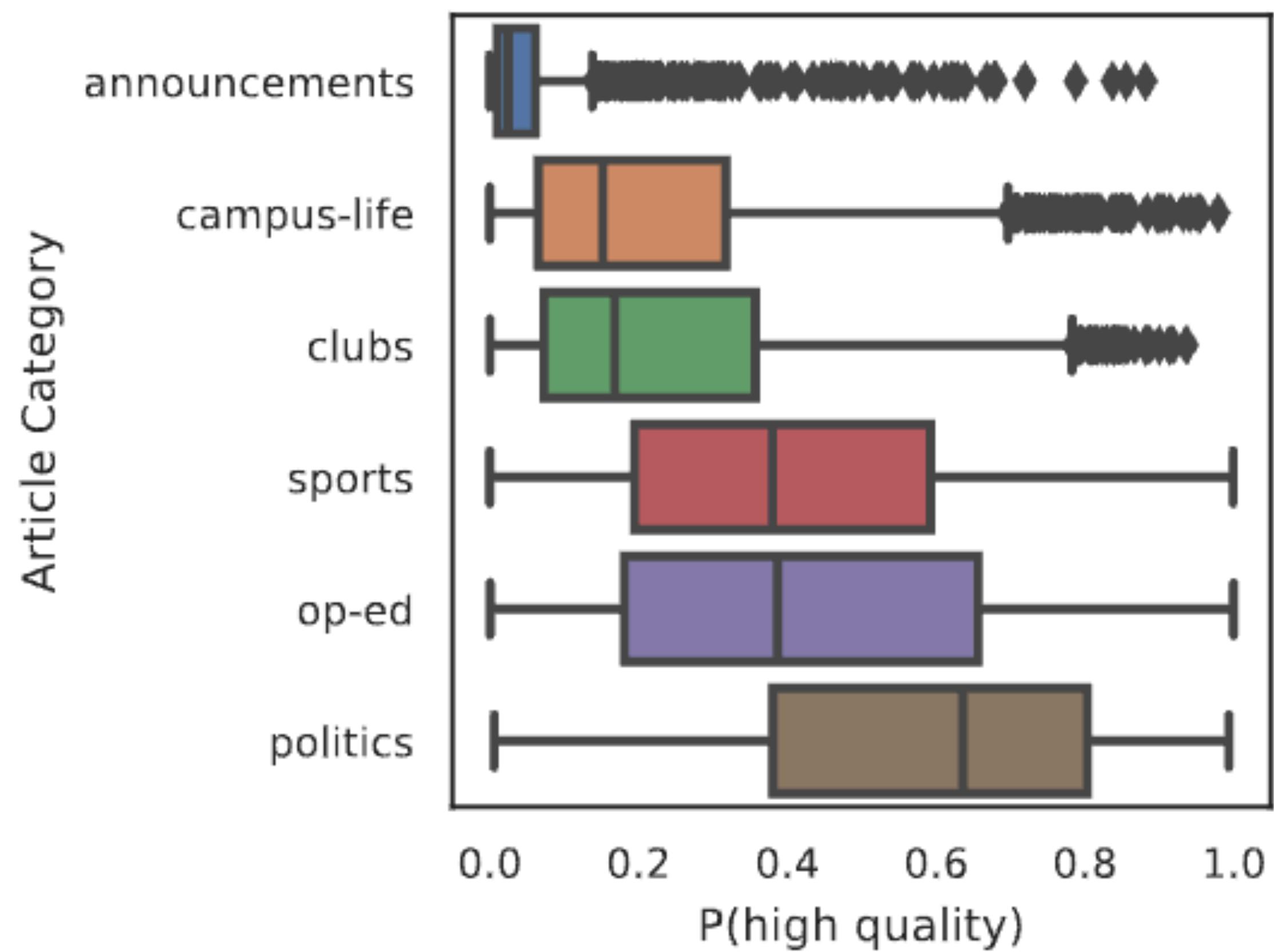
Language models increasingly rely on massive web dumps for diverse text data. However, these sources are rife with undesirable content. As such, resources like Wikipedia, books, and newswire often serve as anchors for automatically selecting web text most suitable for language modeling, a process typically referred to as *quality filtering*. Using a new dataset of U.S. high school newspaper articles—written

quality filters,¹ are often applied in an effort to remove this undesirable content from training data. These filters include code removers ([Gao et al., 2020](#)), heuristics ([Rae et al., 2021](#)), stopwords ([Raf-fel et al., 2020](#)), and classifiers ([Brown et al., 2020; Wenzek et al., 2020](#)).

Although quality filtering is often treated as a relatively neutral preprocessing step, it necessarily implies a value judgment: which data is assumed to be of sufficiently high quality to be included in the



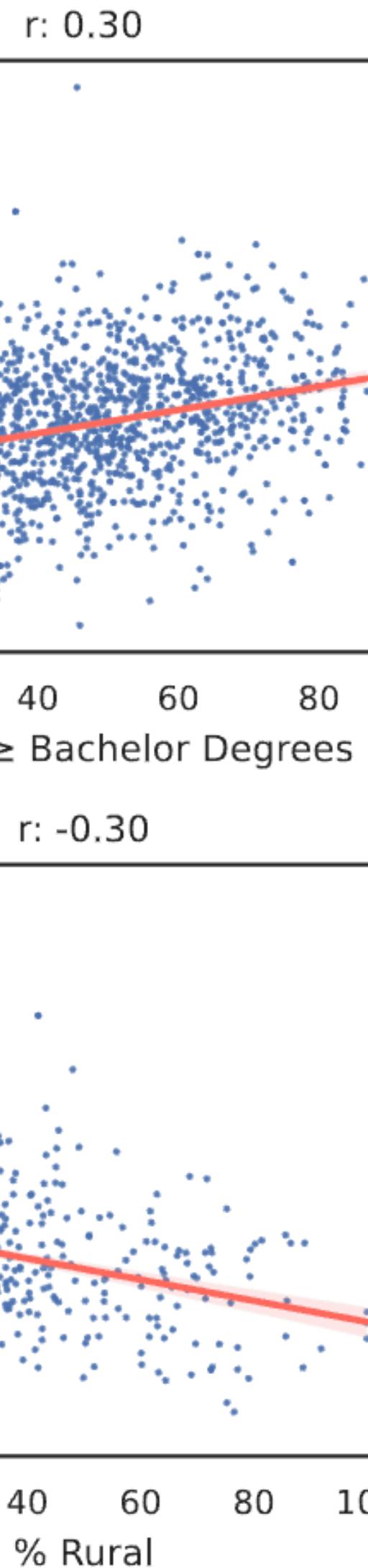
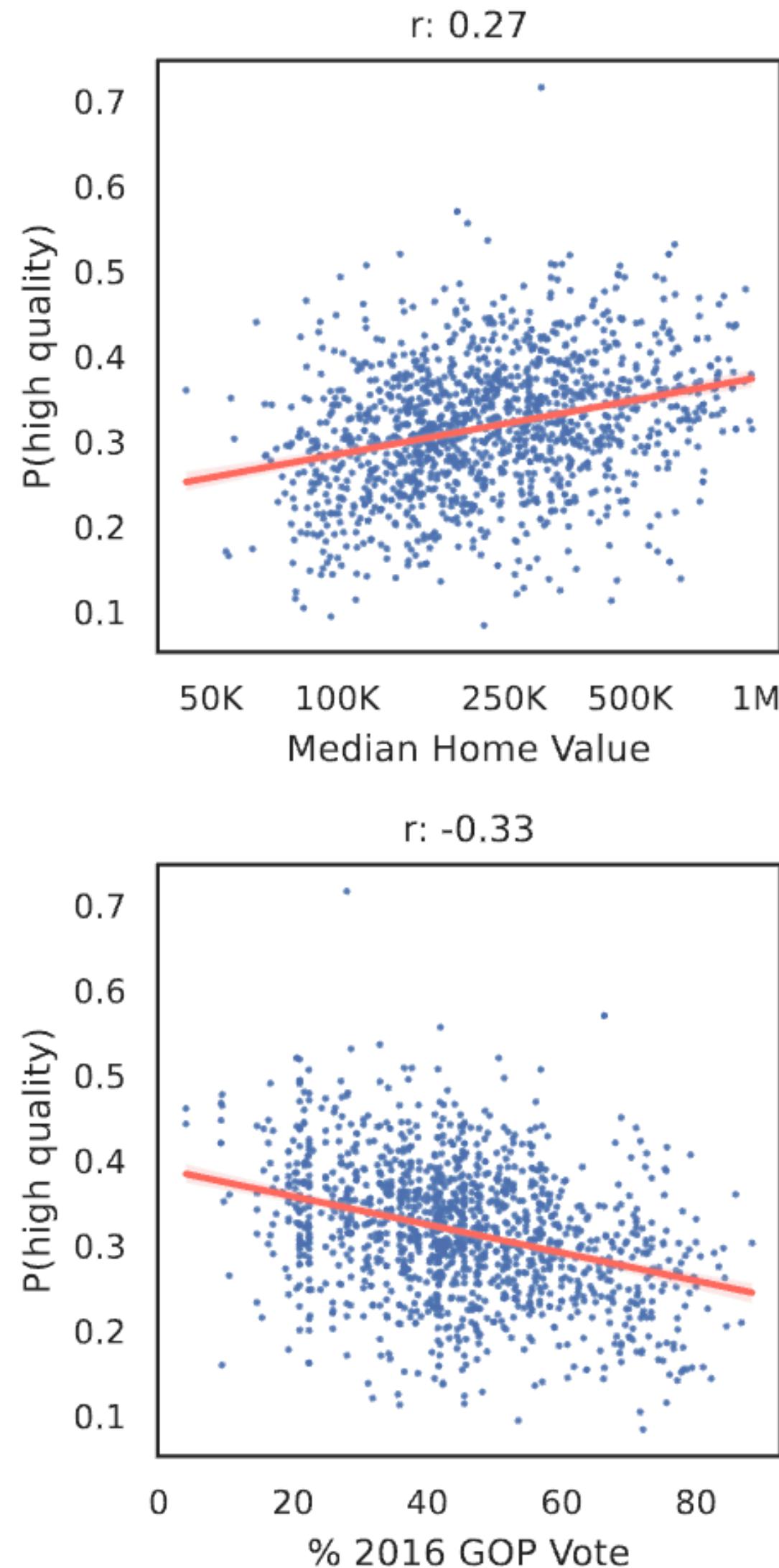
- A comparison of articles from high-school newspapers with mainstream newspapers
- Articles from school newspapers are more likely to be marked as low-quality



- Types of articles written in high-school newspapers can determine if the article passes the quality filter

Feature	Coefficient
<i>Intercept</i>	0.471***
Topic 5 (<i>christmas, dress, holiday</i>)	-0.056***
Topic 2 (<i>school, college, year</i>)	-0.037***
Topic 6 (<i>student, school, class</i>)	-0.004
Topic 1 (<i>people, just, like</i>)	0.003
Topic 7 (<i>movie, film, movies</i>)	0.062***
Topic 3 (<i>music, album, song</i>)	0.113***
Topic 4 (<i>people, women, media</i>)	0.197***
Topic 9 (<i>game, team, players</i>)	0.246***
Topic 8 (<i>Trump, president, election</i>)	0.346***
Presence of first/second person pronoun	-0.054***
Presence of third person pronoun	0.024
$\log_2(\text{Number of tokens})$	0.088***
R^2	0.336
adj. R^2	0.336

- Daily life topics tend to be negatively associated with quality



Dependent variable: $P(\text{high quality})$
 Observations: 968 schools

Feature	Coefficient
<i>Intercept</i>	0.076
% Rural	-0.069***
% Adults \geq Bachelor Deg.	0.059**
$\log_2(\text{Median Home Value})$	0.010*
$\log_2(\text{Number of students})$	0.006*
$\log_2(\text{Student:Teacher ratio})$	-0.007
Is Public	0.015*
Is Magnet	0.013
Is Charter	0.033
R^2	0.140
adj. R^2	0.133

School level prediction of quality articles

CONSEQUENCES

- Use of Wikipedia, newswire, etc as exemplars of high quality text implicitly favors content written by authors from larger schools in wealthier, educated, urban areas of the United States.
- The training input of LLMs can significantly determine the output of these models
- Selection bias in articles can perpetuate inequities

LEARNING METHODS

		Unsupervised learning	Supervised learning
Input	$\{x\}_{i=1}^N; x_i \in \mathcal{X}$	$\{(x, y)\}_{i=1}^N; x_i \in \mathcal{X}, y_i \in \mathcal{Y}$	
Example	Topic modeling, clustering	Classification, Regression	

“How do we get the labels for our data?”

LABELING TECHNIQUES

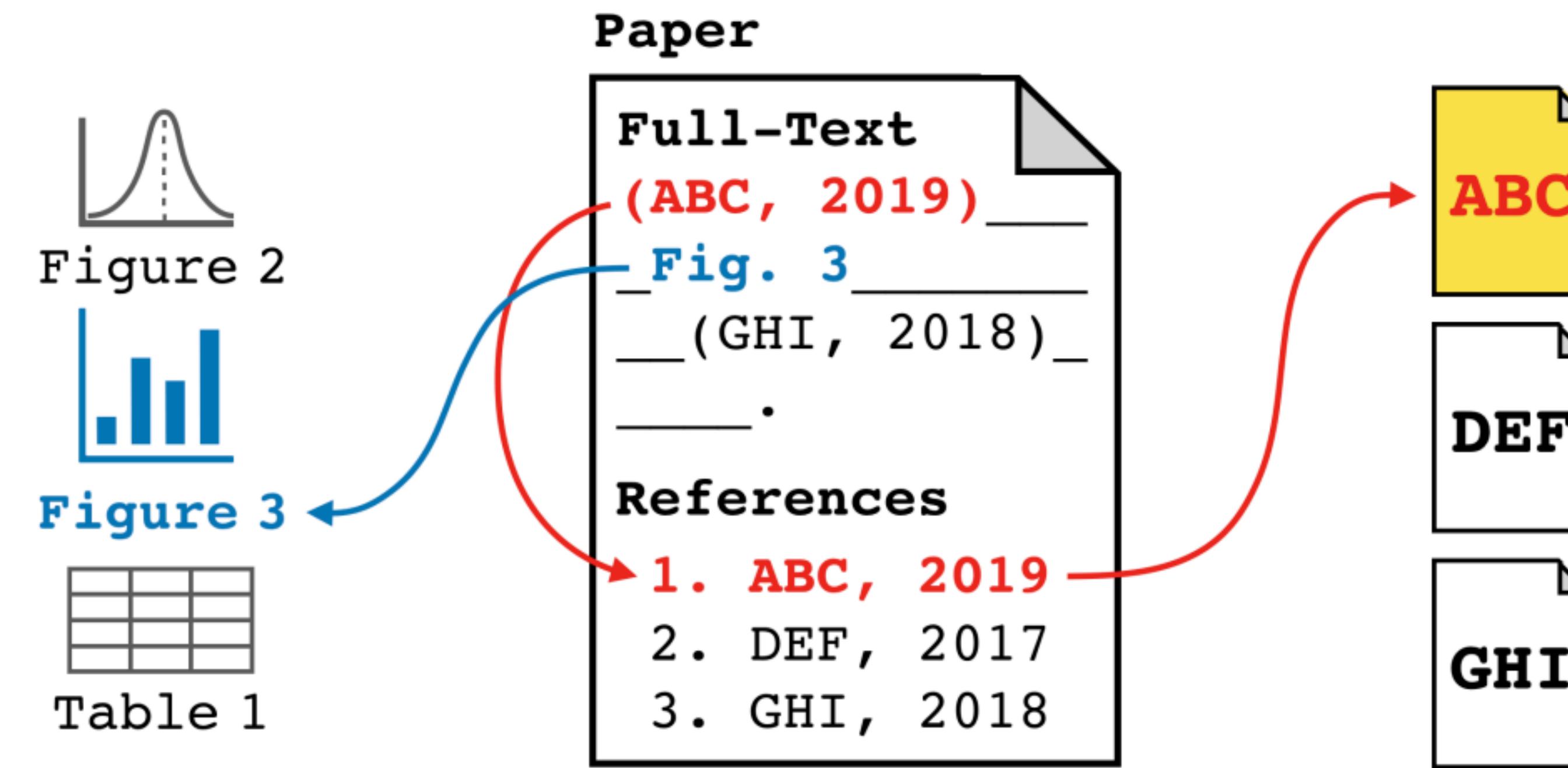
LABELING TECHNIQUES

- Metadata
- Human annotations
- Distant labels

METADATA

METADATA

- Often we have the data we're interested in but supplementary data that gives us more context



Figure/table references linked to figure/table objects

Inline citations linked to bibliography entries

Bibliography entries linked to papers in corpus

```
{  
    "paperId": "649def34f8be52c8b66281af98ae8",  
    "corpusId": 2314124,  
    + "externalIds": { ... },  
    "url": "https://www.semanticscholar.org/p.../  
    "title": "Construction of the Literature  
    "abstract": "We describe a deployed scala  
    "venue": "International Conference on Sof  
    + "publicationVenue": { ... },  
    "year": 2018,  
    "referenceCount": 321,  
    "citationCount": 987,  
    "influentialCitationCount": 654,  
    "isOpenAccess": true,  
    + "openAccessPdf": { ... },  
    + "fieldsOfStudy": [ ... ],  
    + "s2FieldsOfStudy": [ ... ],  
    + "publicationTypes": [ ... ],  
    "publicationDate": "2015-01-17",  
    + "journal": { ... },  
    + "citationStyles": { ... },  
    + "authors": [ ... ],  
    + "citations": [ ... ],  
    + "references": [ ... ],  
    + "embedding": { ... },  
    + "tldr": { ... }  
}
```

In addition to the text of the paper, we also have information about authors, publication date, venues, etc from which you can induce the labels

```
{  
    "paperId": "649def34f8be52c8b66281af98ae8",  
    "corpusId": 2314124,  
    + "externalIds": { ... },  
    "url": "https://www.semanticscholar.org/p.../  
    "title": "Construction of the Literature  
    "abstract": "We describe a deployed scala  
    "venue": "International Conference on Sof  
    + "publicationVenue": { ... },  
    "year": 2018,  
    "referenceCount": 321,  
    "citationCount": 987,  
    "influentialCitationCount": 654,  
    "isOpenAccess": true,  
    + "openAccessPdf": { ... },  
    + "fieldsOfStudy": [ ... ],  
    + "s2FieldsOfStudy": [ ... ],  
    + "publicationTypes": [ ... ],  
    "publicationDate": "2015-01-17",  
    + "journal": { ... },  
    + "citationStyles": { ... },  
    + "authors": [ ... ],  
    + "citations": [ ... ],  
    + "references": [ ... ],  
    + "embedding": { ... },  
    + "tldr": { ... }  
}
```

Field	Description
id	The comment's identifier, e.g., "dbumnq8" (String).
author	The account name of the poster, e.g., "example_username" (String).
link_id	Identifier of the submission that this comment is in, e.g., "t3_51954r" (String).
parent_id	Identifier of the parent of this comment, might be the identifier of the submission if it is top-level comment or the identifier of another comment, e.g., "t1_dbu5bpp" (String).
created_utc	UNIX timestamp that refers to the time of the submission's creation, e.g., 1483228803 (Integer).
subreddit	Name of the subreddit that the comment is posted. Note that it excludes the prefix /r/. E.g., 'AskReddit' (String).
subreddit_id	The identifier of the subreddit where the comment is posted, e.g., "t5_2qh1i" (String).
body	The comment's text, e.g., "This is an example comment" (String).
score	The score of the comment. The score is the number of upvotes minus the number of downvotes. Note that Reddit fuzzes the real score to prevent spam bots. E.g., 5 (Integer).
distinguished	Flag to determine whether the comment is made by the moderators or admins. "null" means not distinguished (String).
edited	Flag indicating if the comment has been edited. Either the UNIX timestamp that the comment was edited at, or "false".
stickied	Flag indicating whether the submission is set as sticky in the subreddit, e.g., false (Boolean).
retrieved_on	UNIX timestamp that refers to the time that we crawled the comment, e.g., 1483228803 (Integer).
gilded	The number of times this comment received Reddit gold, e.g., 0 (Integer).
controversiality	Number that indicates whether the comment is controversial, e.g., 0 (Integer).
author_flair_css_class	The CSS class of the author's flair. This field is specific to subreddit (String).
author_flair_text	The text of the author's flair. This field is specific to subreddit (String).

/r/LifeProTips (LT)

LPT: Check the Facebook app to find the owner of a lost smartphone

or simply call her 'mum'? Also slightly less intrusive IMO.

63 comments, 72% upvoted

LPT: get your pets to take their medicine with butter.

This is much better! I have been trying ice cream but my dog is too smart.

62 comments, 72% upvoted

LPT: For a cleaner home with little effort, never leave a room empty-handed. There is almost always something you can put back in its place on your way.

Woah.

115 comments, 93% upvoted

/r/Fitness (FT)

tl;dr quit whining cuz r/fitness didn't respond they way you wanted...

Unfortunately, I doubt this kind of post is going to change anything...

237 comments, 71% upvoted

Interesting New Study: Red Meat Linked With Increased Mortality Risk. Thought this study is worth a discussion...

Man, it seems like everything these days will lower your life span.

66 comments, 63% upvoted

What type of snack should I have preworkout to avoid lethargy at the gym? I don't wanna be sluggish at the gym...

Apples slices with peanut butter.

394 comments, 90% upvoted

/r/personalfinance (PF)

Tipping as legal discrimination: Black servers get tipped 3.25% less... [LINK]...

Tipping should be abandoned anyway, it's ridiculous....

61 comments, 57% upvoted

Am I crazy for wanting this car/payment? Short of it .. car is \$45,000...

Needing a car for work and purchasing \$45k car are two entirely different things.

125 comments, 62% upvoted

Accumulating wealth via homeownership vs accumulating wealth as a renter. One of the often cited benefits of homeownership ...

Use this handy calculator from the NY Times. If you're diligent...

110 comments, 97% upvoted

	AM	AW	FT	LT	PF	RL
HAND	55.4	52.2	61.9	59.7	54.5	60.8
TFIDF	57.4	60.1	63.3	59.1	58.7	65.4
ARORA	58.6	62.0	60.5	59.4	57.2	62.1
W2V	60.7	62.1	63.1	61.4	59.9	64.3
LSTM	58.9	58.2	63.6	61.5	60.0	63.1
BERT-LSTM	64.5	65.1	66.2	65.0	65.1	67.8
BERT-MP	63.4	64.0	64.4	65.7	64.1	67.0
BERT-MP-512	63.9	64.0	64.7	65.8	65.6	67.7
HAND+W2V	61.3	62.3	64.9	63.2	60.0	66.3
HAND+BERTMP512	63.6	63.5	64.9	64.1	64.4	68.0

METADATA

METADATA

- Not everything is present in the metadata
- For example, given a reddit comment, we don't know the degree to which the comment can be considered as hate speech or counter speech.

ANNOTATING BY HAND

ANNOTATING BY HAND

- Some key issues:
 - Who should annotate?
 - How many annotations are required?
 - How to evaluate that the annotations are consistent?

EXPERTS VS NON-EXPERTS

EXPERTS VS NON-EXPERTS

- The choice depends on the task
- Some tasks require domain knowledge, cultural context, etc – which is something an expert will generally have
- Other tasks are difficult for computational methods but not so much for humans

CC	Coordinating conj.
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PP\$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	infinitival <i>to</i>
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/present pple
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sg. present
VBZ	Verb, 3rd ps. sg. present
WDT	Wh-determiner
WP	Wh-pronoun
WP\$	Possessive wh-pronoun
WRB	Wh-adverb
#	Pound sign
\$	Dollar sign
.	Sentence-final punctuation
,	Comma
:	Colon, semi-colon
(Left bracket character
)	Right bracket character
"	Straight double quote
'	Left open single quote
"	Left open double quote
,	Right close single quote
"	Right close double quote

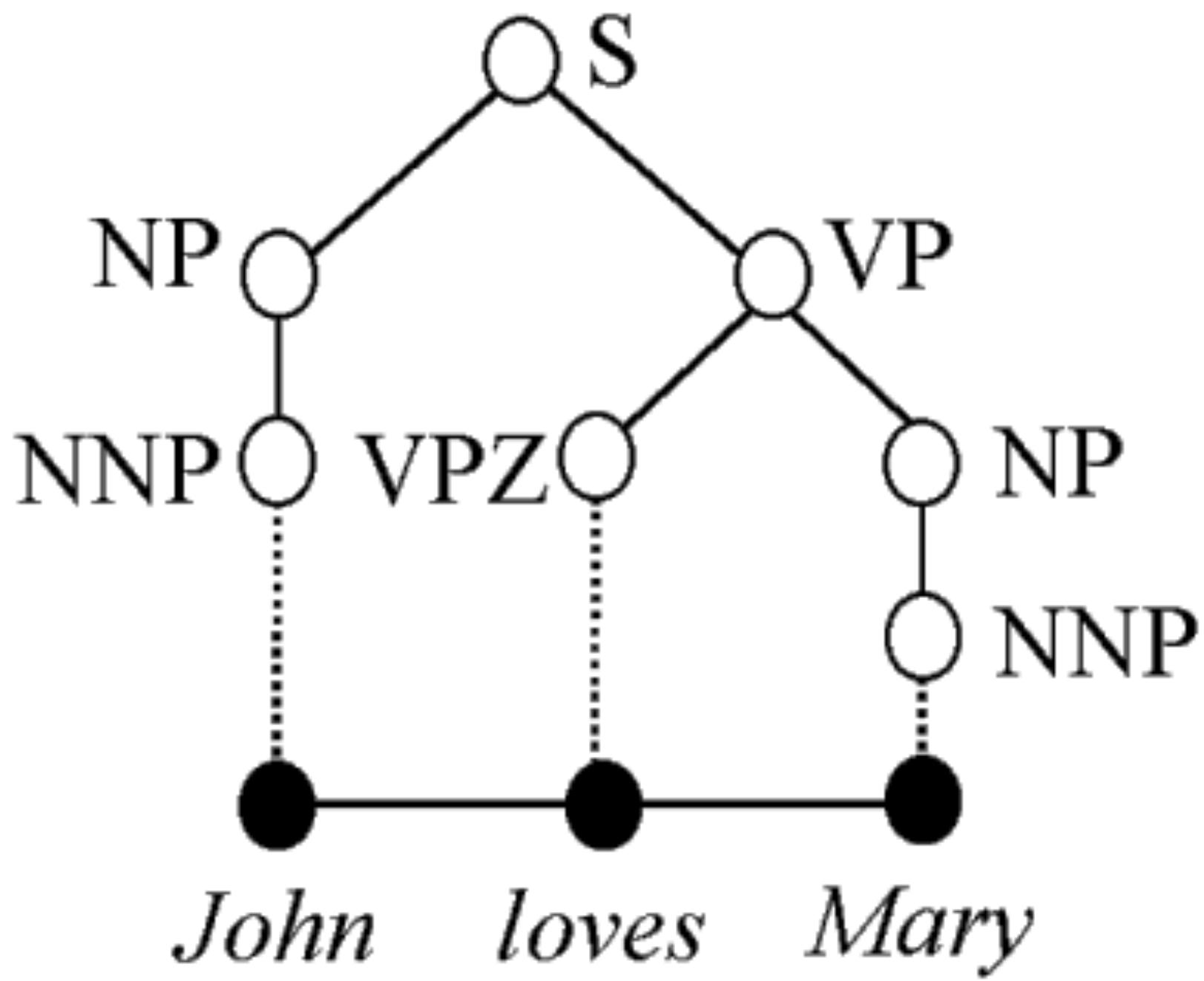


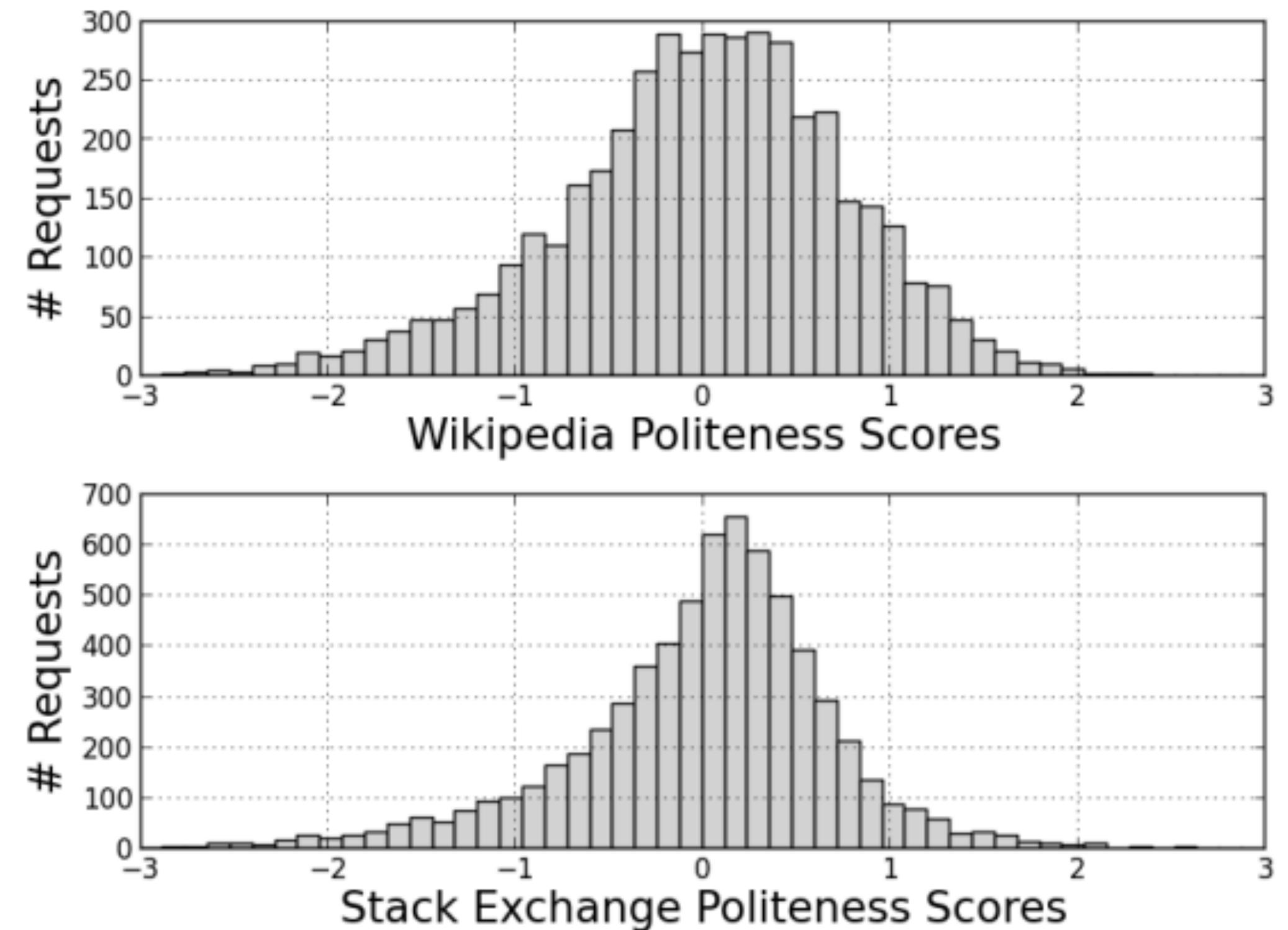
Image from Wikipedia

A computational approach to politeness with application to social factors

Cristian Danescu-Niculescu-Mizil^{*‡}, Moritz Sudhof[†], Dan Jurafsky[†],
Jure Leskovec^{*}, and Christopher Potts[†]

^{*}Computer Science Department, [†]Linguistics Department

^{*†}Stanford University, [‡]Max Planck Institute SWS



Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Text Categories

Anti-Black

Going to Africa. Hope I don't get AIDS.
Just kidding. I'm white!



AAE

If u grown & still get thirsty for Jordans
knowin erbody else gon havem & u
still feel like u accomplished
something that say alot about u



Vulgar

I got mosquito bites on my foot and
they fucking hurt



Key Findings: Identity and Attitude Biases in Toxicity Detection

Breadth-of-Workers

Breadth-of-Posts

Less offensive/racist for annotators who...

score higher in FREEOFFSPEECH, RACISTBELIEFS, LINGPURISM,
TRADITIONALISM; are more conservative, men, White

More offensive/racist for annotators who...

score higher in EMPATHY, ALTRUISM, HARMOFHATESPEECH;
are more liberal, women, Black

Less offensive/racist for annotators who...

score higher in RACISTBELIEFS

More offensive/racist for annotators who...

score higher in HARMOFHATESPEECH

More racist for annotators who...

score higher in FREEOFFSPEECH,
TRADITIONALISM; are more conservative

More offensive for annotators who...

score higher in RACISTBELIEFS,
TRADITIONALISM; are more conservative

More racist for annotators who...

score higher in RACISTBELIEFS;
are more conservative

More offensive for annotators who...

score higher in LINGPURISM, TRADITIONALISM;
are more conservative

HOW MANY ANNOTATORS?

- Usually more than 1; higher bound usually depends on cost
- More annotators is usually better if the annotations are independent
- Key question: how to quantify the disagreement between annotators?

QUANTIFYING DISAGREEMENT

- Let's say we have two annotators and each annotator is given an example and asked to provide a binary label

QUANTIFYING DISAGREEMENT

	1	2	3	4	5	6	7	8	9	10
Annotator 1	spam	spam	spam	spam	spam	ham	ham	ham	ham	ham
Annotator 2	spam	spam	ham	spam	spam	spam	ham	spam	ham	spam

QUANTIFYING DISAGREEMENT

Annotator2 \ Annotator1	A1=spam	A1=ham
A2 = spam	4	3
A2=ham	1	2

QUANTIFYING DISAGREEMENT

Annotator2 \ Annotator1		A1=spam	A1=ham
A2 = spam	True positives	False positives	
A2=ham	False negatives	True negatives	

QUANTIFYING DISAGREEMENT

- One can simply quantify disagreement or agreement by calculating a measure similar to accuracy

- $$\frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N}$$

QUANTIFYING DISAGREEMENT

- Why is this not good?
- The metric does not quantify whether two annotators will disagree or agree on examples just by chance
- It does not account for one annotator being more strict than another, or one label more likely than another

COHEN'S κ

- What is the probability that A1 and A2 agree on the observed data?

COHEN'S κ

		A1=spam	A1=ham	
		4	3	7
Annotator2 \ Anno tator1	A2 = spam			
	A2=ham	1	2	3
		5	5	

- What is the probability that A1 and A2 agree on the observed data?

COHEN'S K

Annotator2 \ Anno tator1		A1=spam	A1=ham	
A2 = spam	4	3	7	
A2=ham	1	2	3	
	5	5		

- $P_o = P((A1=\text{spam} \cap A2=\text{spam}) \cup (A1=\text{ham} \cap A2=\text{ham}))?$
- $P_o = 0.4 + 0.2$

COHEN'S κ

- What is the probability that A1 and A2 agree by chance?

COHEN'S κ

		A1=spam	A1=ham	
		4	3	7
A2 = spam	A2 = spam	4	3	7
	A2=ham	1	2	3
		5	5	

- What is the probability that A1 and A2 agree by chance?

COHEN'S K

		A1=spam	A1=ham	
Annotator2 \ Anno tator1	A2 = spam	4	3	7
	A2=ham	1	2	3
		5	5	

- $P(A1=\text{spam}) = 0.5; P(A1=\text{ham}) = 0.5$
- $P(A2=\text{spam}) = 0.7; P(A2=\text{ham}) = 0.3$

COHEN'S K

		A1=spam	A1=ham	
		4	3	7
A2 = spam	A2 = spam	4	3	7
	A2=ham	1	2	3
		5	5	

- What is $P((A1=\text{spam} \cap A2=\text{spam}) \cup (A1=\text{ham} \cap A2=\text{ham}))$?

COHEN'S K

		A1=spam	A1=ham	
		4	3	7
Annotator2 \ Anno tator1	A2 = spam			
	A2=ham	1	2	3
		5	5	

- $P(A1=\text{spam} \cap A2=\text{spam}) = P(A1=\text{spam}) * P(A2=\text{spam})$
- $P(A1=\text{spam} \cap A2=\text{spam}) = 0.5 * 0.7 = 0.35$

COHEN'S κ

		A1=spam	A1=ham	
		4	3	7
Annotator2 \ Anno tator1	A2 = spam			
	A2=ham	1	2	3
		5	5	

- $P(A1=\text{ham} \cap A2=\text{ham}) = P(A1=\text{ham}) * P(A2=\text{ham})$
- $P(A1=\text{ham} \cap A2=\text{ham}) = 0.5 * 0.3 = 0.15$

COHEN'S K

		A1=spam	A1=ham	
Annotator2 \ Anno tator1	A2 = spam	4	3	7
	A2=ham	1	2	3
		5	5	

- $P_e = P((A1=\text{spam} \cap A2=\text{spam}) \cup (A1=\text{ham} \cap A2=\text{ham}))$
- $P_e = 0.35 + 0.15 = 0.5$

COHEN'S κ

A2 \ A1	A1=spam	A1=ham
A2 = spam	4	3
A2=ham	1	2

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

$$\kappa = \frac{0.6 - 0.5}{1 - 0.5}$$

$$\kappa = 0.2$$

INTERANNOTATOR AGREEMENT

κ	Level of agreement
< 0.20	Poor
$0.21 - 0.40$	Fair
$0.41 - 0.60$	Moderate
$0.61 - 0.80$	Good
$0.81 - 1.00$	Perfect

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

- Agreement levels for the kappa statistic

INTERANNOTATOR AGREEMENT

- Cohen's κ corrects for chance agreement
- Scott's π calculates chance agreement in a different way
- Fleiss' κ extends agreement measurement to more than two raters

Language from police body camera footage shows racial disparities in officer respect

Rob Voigt^{a,1}, Nicholas P. Camp^b, Vinodkumar Prabhakaran^c, William L. Hamilton^c, Rebecca C. Hetey^b, Camilla M. Griffiths^b, David Jurgens^c, Dan Jurafsky^{a,c}, and Jennifer L. Eberhardt^{b,1}

^aDepartment of Linguistics, Stanford University, Stanford, CA 94305; ^bDepartment of Psychology, Stanford University, Stanford, CA 94305; and ^cDepartment of Computer Science, Stanford University, Stanford, CA 94305

Contributed by Jennifer L. Eberhardt, March 26, 2017 (sent for review February 14, 2017; reviewed by James Pennebaker and Tom Tyler)

INFORMAL TITLE	ASK FOR AGENCY	ADVERBIAL "JUST"	
All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick.		-0.51	
"HANDS ON THE WHEEL"			

APOLOGY	INTRODUCTION	LAST NAME	
Sorry to stop you. My name's Officer [name]			0.84
with the Police Department.			

Batch	Formal	Friendly	Impartial	Polite	Respectful
1	0.82	0.86	0.84	0.86	0.83
2	0.88	0.89	0.86	0.86	0.87
3	0.80	0.87	0.73	0.84	0.78
4	0.85	0.91	0.79	0.88	0.87
5	0.77	0.89	0.81	0.87	0.87
6	0.91	0.82	0.81	0.87	0.86
7	0.85	0.86	0.84	0.84	0.84

Interannotator agreement chart

AGREEMENT ON CONTINUOUS VALUES?

- IAA metrics such as Cohen's κ works well for categorical variables
- For continuous variables, we usually calculate some correlation between the annotators

NOISY LABELS

I'm going to QTM 340 today 😊



Text = "I'm going to QTM 340
today"

Label= Positive

Twitter Sentiment Classification using Distant Supervision

Alec Go
Stanford University
Stanford, CA 94305
alecmgo@stanford.edu

Richa Bhayani
Stanford University
Stanford, CA 94305
rbhayani@stanford.edu

Lei Huang
Stanford University
Stanford, CA 94305
leirocky@stanford.edu

IN CLASS

- Data annotation