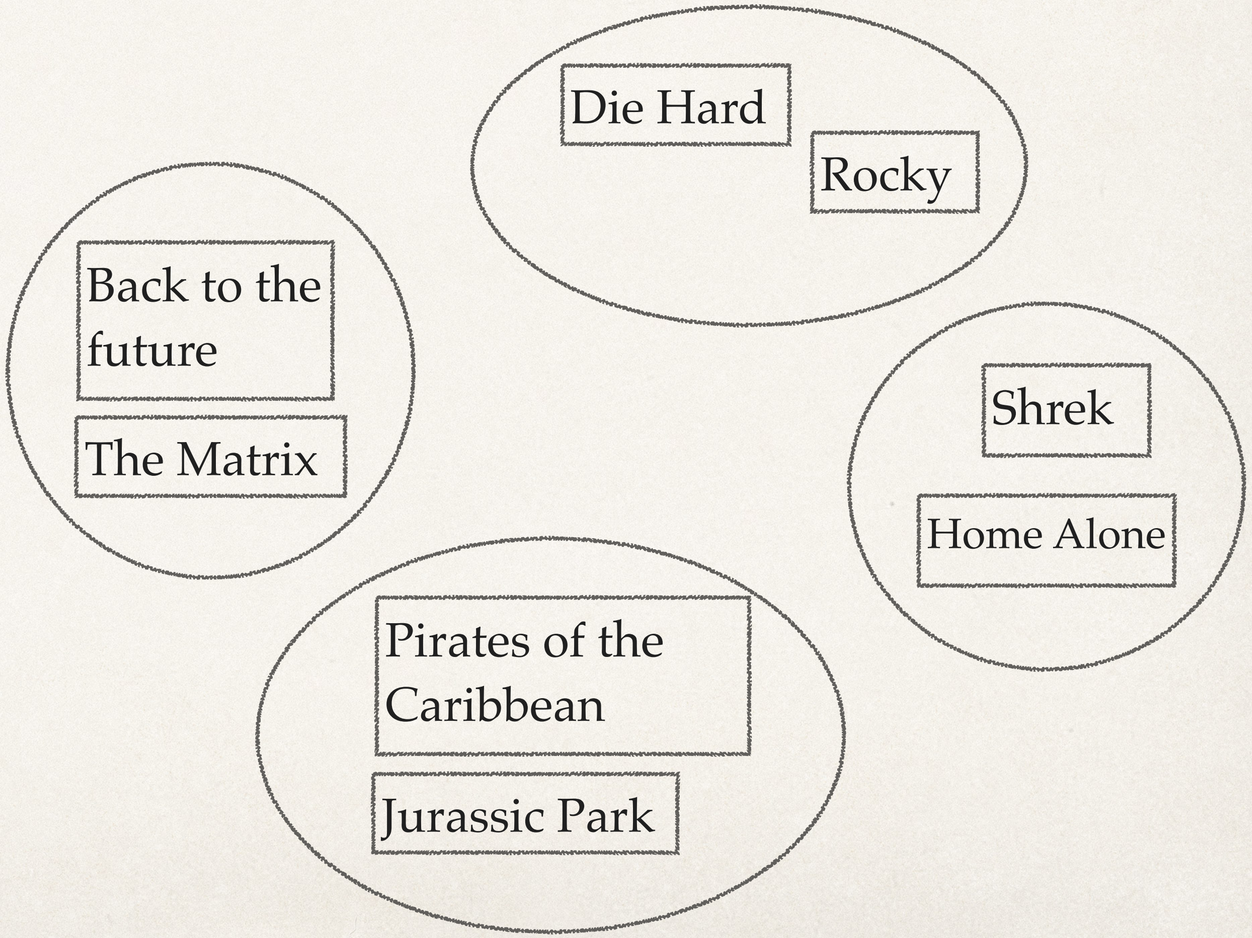
Data Labeling and Annotations

Sandeep Soni

10/11/2023

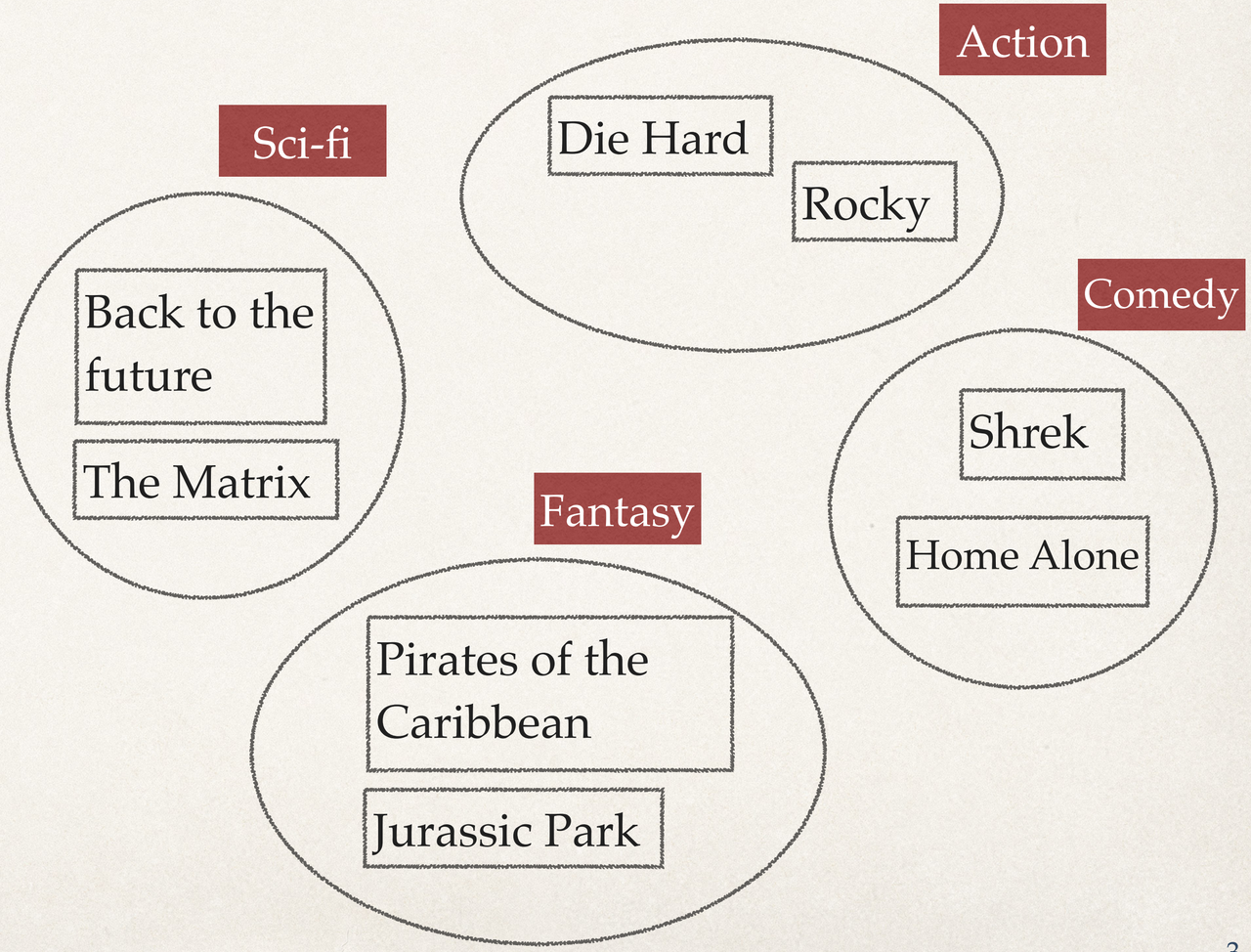
Learning methods

Pirates of the Caribbean	Shrek
Back to the future	Home Alone
The Matrix	Jurassic Park
Die Hard	Rocky



Learning methods

Pirates of the Caribbean	Shrek
Back to the future	Home Alone
The Matrix	Jurassic Park
Die Hard	Rocky



Learning methods

Fantasy

Pirates of the Caribbean

Comedy

Shrek

Sci-fi

Back to the future

Comedy

Home Alone

Sci-fi

The Matrix

Fantasy

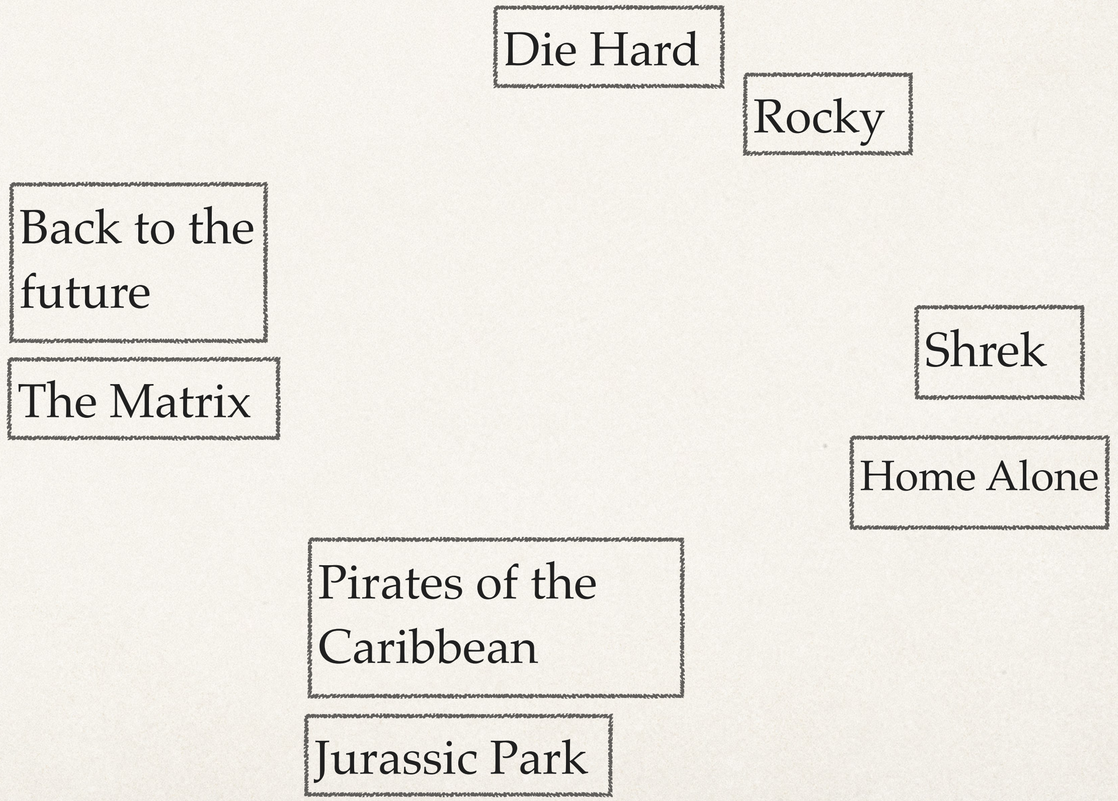
Jurassic Park

Action

Die Hard

Action

Rocky



Learning methods

Fantasy

Pirates of the Caribbean

Comedy

Shrek

Sci-fi

Back to the future

Comedy

Home Alone

Sci-fi

The Matrix

Fantasy

Jurassic Park

Action

Die Hard

Action

Rocky

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

Question for the day

"How do we get the labels for our data?"

Labeling techniques

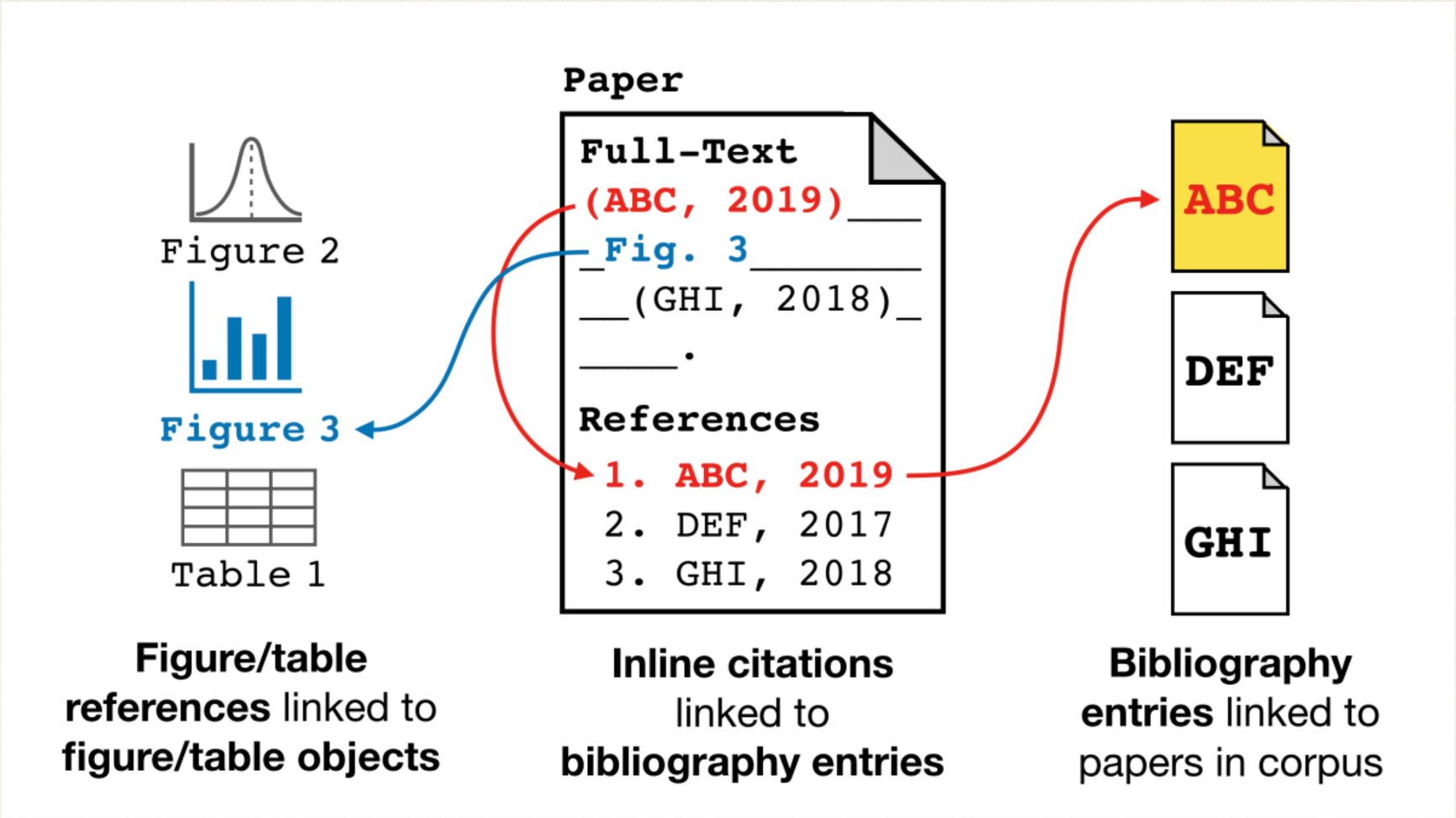
Labeling techniques

- Metadata
- Human annotations
- Distant labels

Metadata

Metadata

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



Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

```
{
   "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": { … }
}
```


Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

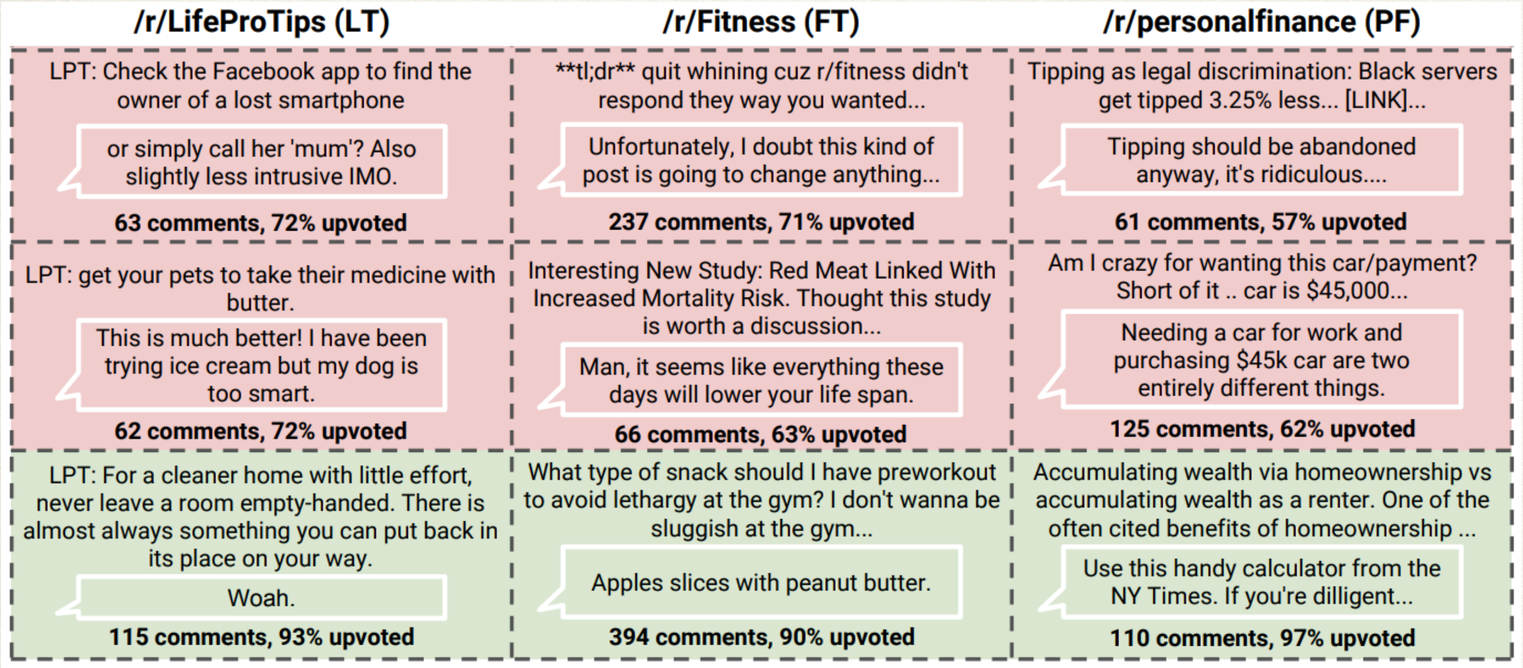
In addition to the text of the paper, we also have information about authors, publication date, venues, etc from which you can induce the labels


```
{
   "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": { … }
}
```


Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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_5l954r" (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).

Baumgartner, Jason, et al. "The pushshift reddit dataset." Proceedings of the international AAAI conference on web and social media. Vol. 14. 2020.



	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**

Hessel, Jack, and Lillian Lee. "Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features." Proceedings of NAACL-HLT. 2019.

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

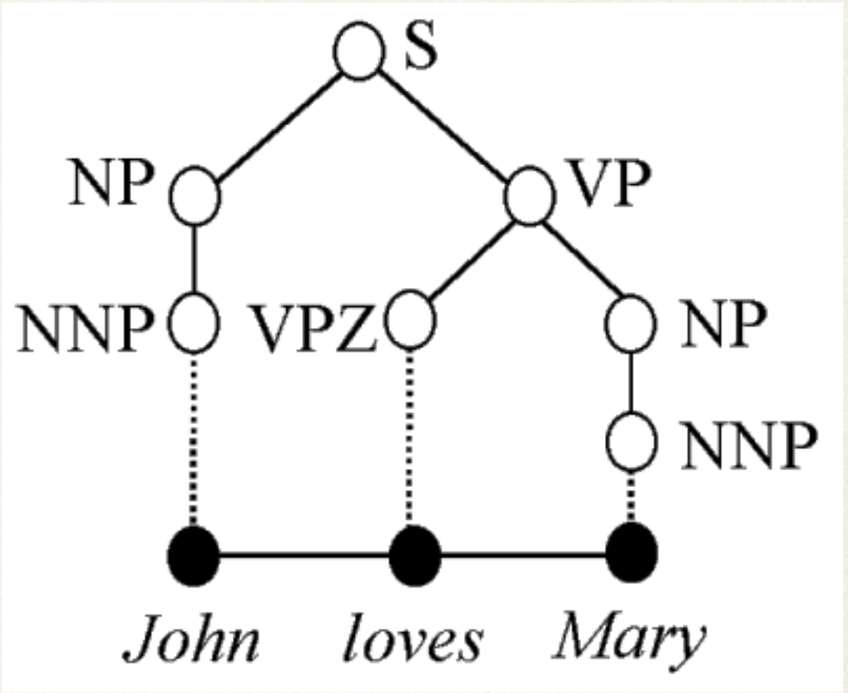
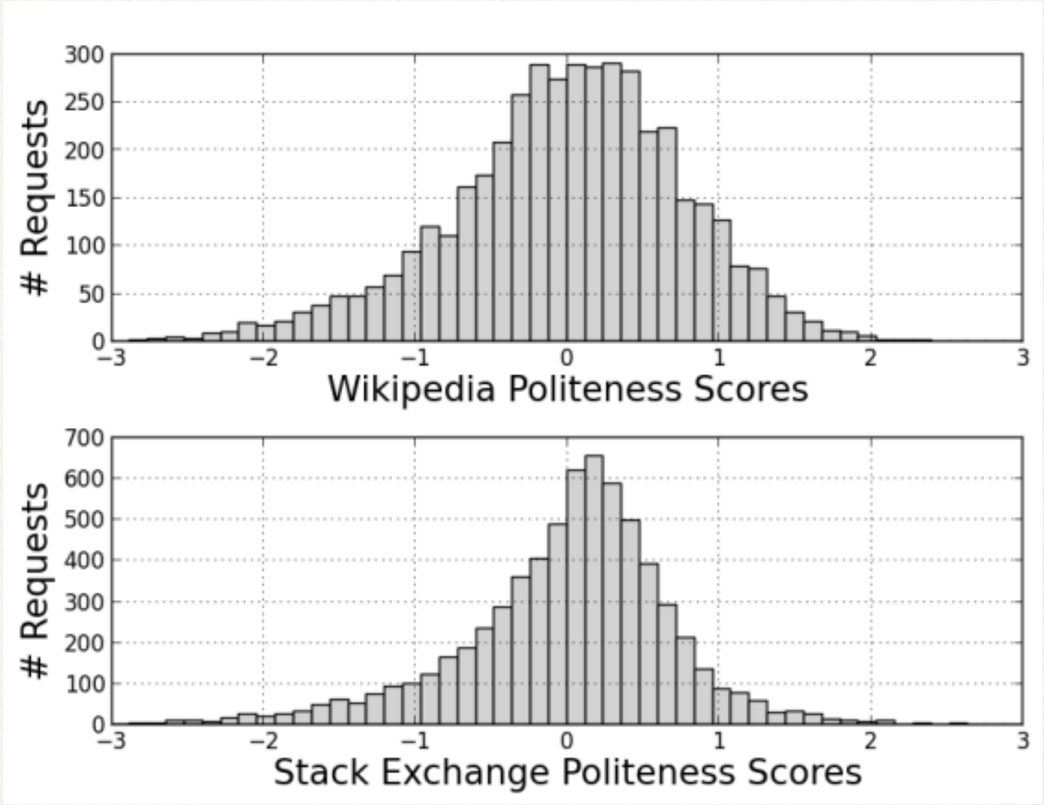


Image from Wikipedia

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

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		**Key Findings**: *Identity and Attitude Biases in Toxicity Detection*	
		Breadth-of-Workers	*Breadth-of-Posts*
Anti-Black	*Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!*	***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
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*	***More racist** for annotators who...* score higher in RACISTBELIEFS; are more conservative	***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
Vulgar	*I got mosquito bites on my foot and they fucking hurt*	***More offensive** for annotators who...* score higher in LINGPURISM, TRADITIONALISM; are more conservative	——

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

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 case
- 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

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

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

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- P_o = P((A1=spam ∩ A2=spam) ∪ (A1=ham ∩ A2=ham))?
- P_o = 0.4 + 0.2

Cohen's κ

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

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
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 κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- P(A1=spam) = 0.5; P(A1=ham) = 0.5
- P(A2=spam) = 0.7; P(A2=ham) = 0.3

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- What is P((A1=spam ∩ A2=spam) ∪ (A1=ham ∩ A2=ham))?

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- P(A1=spam ∩ A2=spam) = P(A1=spam) * P(A2=spam)
- P(A1=spam ∩ A2=spam) = 0.5 * 0.7 = 0.35

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- P(A1=ham ∩ A2=ham) = P(A1=ham) * P(A2=ham)
- P(A1=ham ∩ A2=ham) = 0.5 * 0.3 = 0.15

Cohen's κ

Annotator2\Annotator1	A1=spam	A1=ham	
A2 = spam	4	3	7
A2=ham	1	2	3
	5	5	

- $P_e =$ P((A1=spam ∩ A2=spam) ∪ (A1=ham ∩ A2=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

- 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

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

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

Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N project report, Stanford* 1.12 (2009): 2009.

In class

- Data annotation