Weakly Supervised Labeling Strategies for Classifying User-generated Content

by

Matti Wiegmann

Disputation to obtain the degree
Dr. rer. nat.

Part 1

Supervised Learning

Motivation

Example: Chihuahua or Muffin?

Supervised Learning

Motivation

Example: Chihuahua or Muffin?

Supervised Learning

Motivation

Classification Problems

Determine the label $c \in C$ of a data point $x \in X$.

Supervised Learning

Find an optimal model $y : X \to C$ over a set D of examples.
$\rightsquigarrow$ The classifier learns from labeled data.

$$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times C$$

Example: Chihuahua or Muffin?

Supervised Learning

Classification Problems

Determine the label $c \in C$ of a data point $x \in X$.

Supervised Learning

Find an optimal model $y : X \to C$ over a set D of examples.

$\rightsquigarrow$ The classifier learns from labeled data.

$$D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times C$$

Assumption: The labels stem from an ideal label function.

- The labels are correct and complete.
- Human annotation is considered an ideal label function. In NLP, IR, CSS, ...

Example: Chihuahua or Muffin?

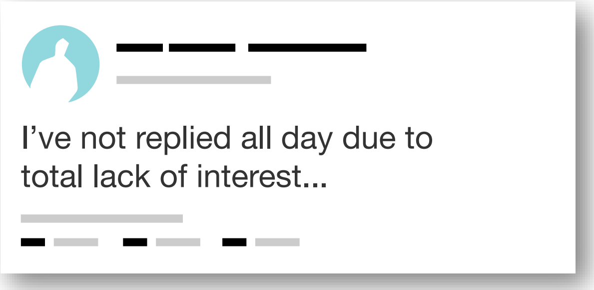
Labeling Functions

Motivation

Problems of human annotation:

- Limited human ability
  Subjectivity, limited domain expertise, complex labels
- Scaling and cost

Is the user in a depression or not?



Labeling Functions

Motivation

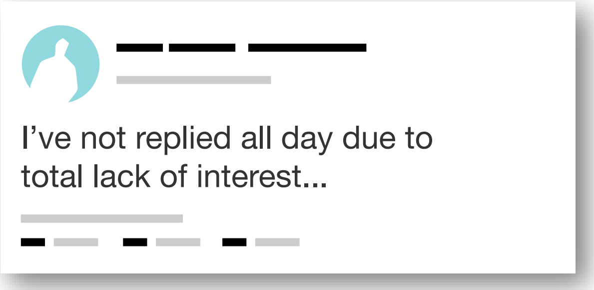
Problems of human annotation:

- Limited human ability
  Subjectivity, limited domain expertise, complex labels
- Scaling and cost

⤳ Automatic labeling functions:

- Semi-supervised learning
- Self-supervised learning
- Weak supervision

Is the user in a depression or not?



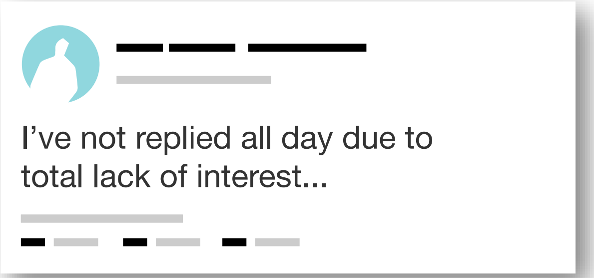
Weak Supervision

Motivation

Use a distant source of knowledge to derive the label.

- Use a heuristic labeling function to link data and distant knowledge.

Is the user in a depression or not?

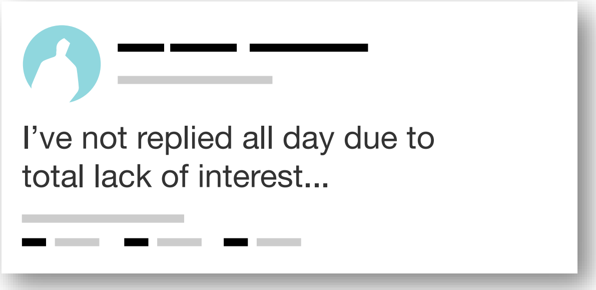


Weak Supervision

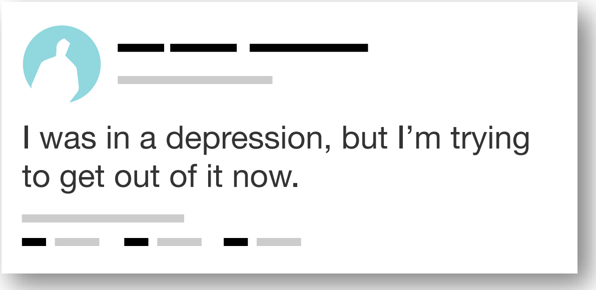
Use a distant source of knowledge to derive the label.

- Use a heuristic labeling function to link data and distant knowledge.

Is the user in a depression or not?



Use knowledge from later posts

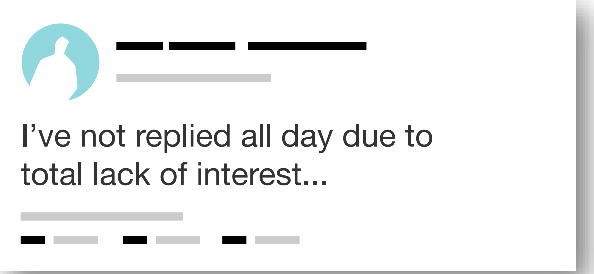


Weak Supervision

Use a distant source of knowledge to derive the label.

- Use a heuristic labeling function to link data and distant knowledge.

Is the user in a depression or not?

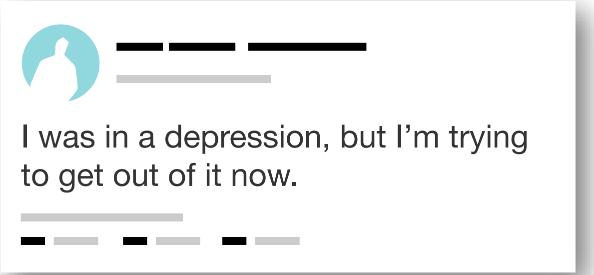


Problems

There is no general theory on weak supervision.

- What sources of data and knowledge are available?
- What are pitfalls of common labeling functions?
- How to evaluate the labeling functions?
- . . .

Use knowledge from later posts



Contributions

1. **Surveying successful applications** to establish a theoretic foundation.

2. **Constructing novel datasets** via new, complex labeling functions.

3. **Answering research questions** based on the new datasets.

Contributions

1. **Surveying successful applications** to establish a theoretic foundation.

2. **Constructing novel datasets** via new, complex labeling functions.

3. **Answering research questions** based on the new datasets.

Profiling Influencers on Twitter
[Wiegmann et al., ACL 2019] [Wiegmann et al., PAN@CLEF 2019] [Wiegmann et al., PAN@CLEF 2020]

Analyzing the Persuasiveness of Debaters
[Wiegmann et al., COLING 2022]

Trigger Warning Assignment
[Wiegmann et al., ACl 2023] [Wiegmann et al., PAN@CLEF 2023] [Wolska and Wiegmann et al., EMNLP 2023]
[Wiegmann et al., CLEF 2024]

Part 2

Surveying Successful Applications

Survey Method

What are eligible sources of data?

What are sources of distant knowledge?

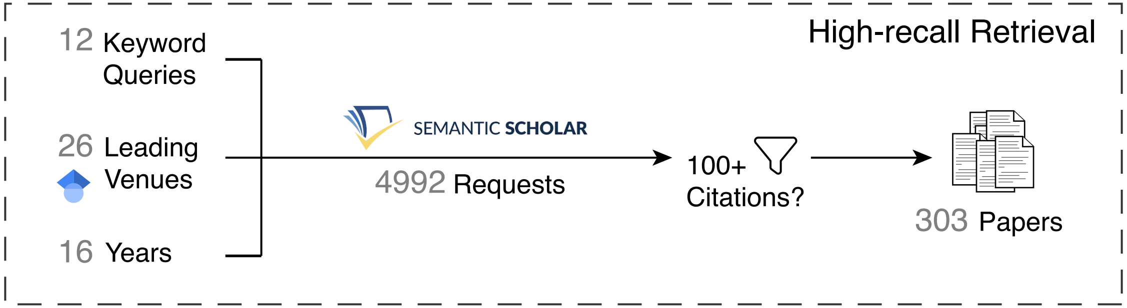
What are common labeling functions?

How can we evaluate labeling functions?

Surveying Successful Applications

Survey Method

Identify successful papers in NLP, IR, ML, and WSM research.



What are eligible sources of data?

What are sources of distant knowledge?

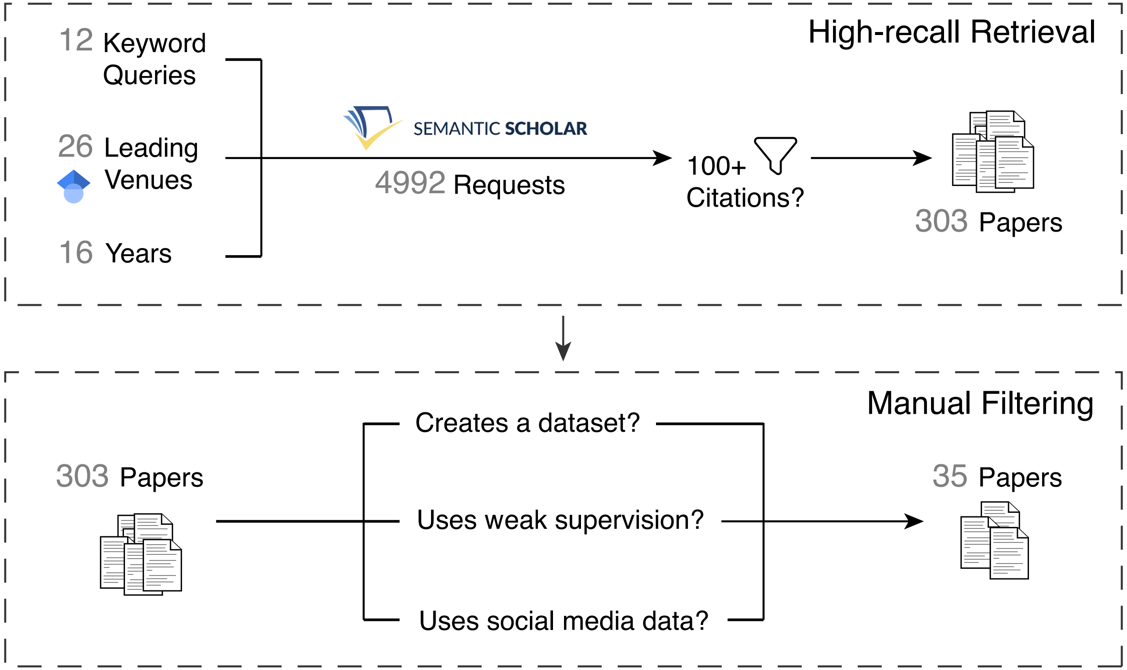
What are common labeling functions?

How can we evaluate labeling functions?

Surveying Successful Applications

Survey Method

Identify successful papers in NLP, IR, ML, and WSM research.



What are eligible sources of data?

What are sources of distant knowledge?

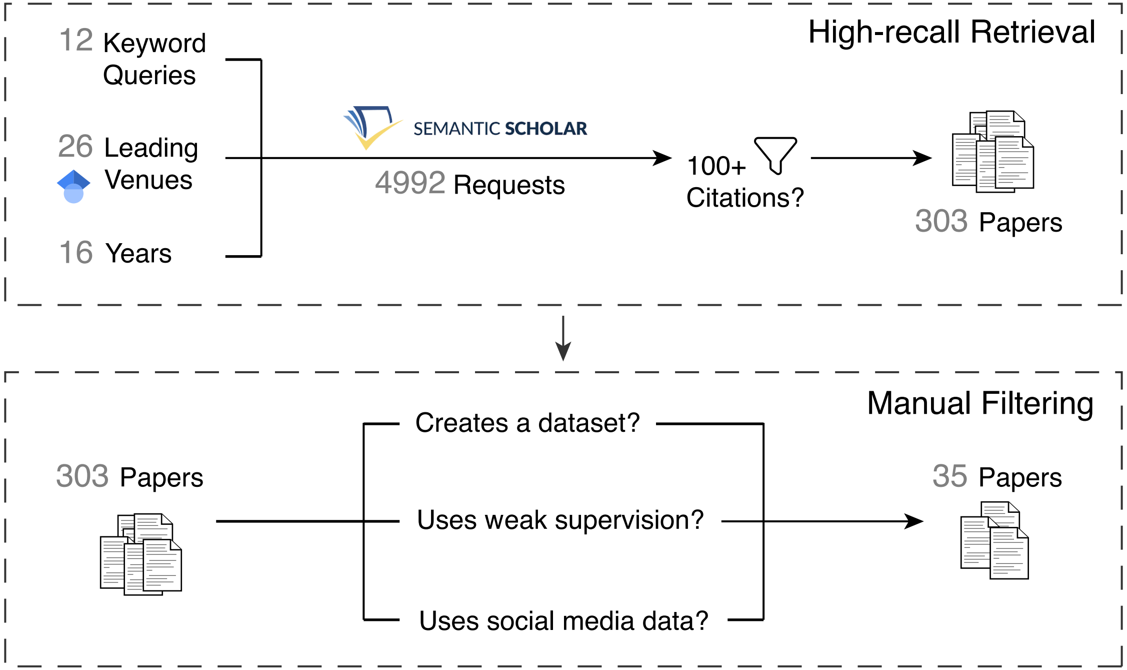
What are common labeling functions?

How can we evaluate labeling functions?

Surveying Successful Applications

Survey Method

Identify successful papers in NLP, IR, ML, and WSM research.



What are eligible sources of data?

What are sources of distant knowledge?

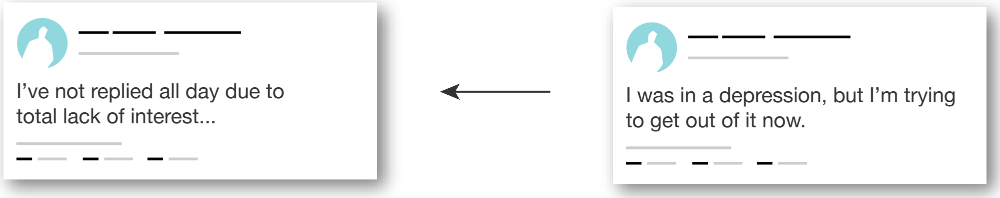
What are common labeling functions?

How can we evaluate labeling functions?

Heuristic Distance

Theory

Use knowledge from later posts *(short distance)*



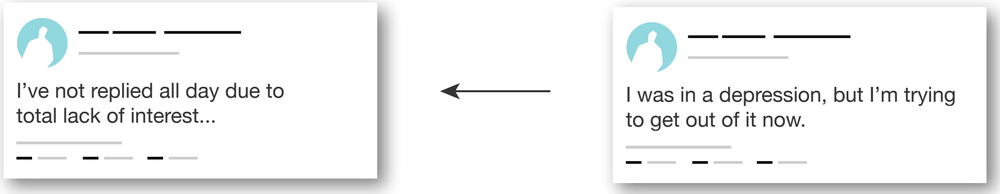
Heuristics:

- ❑ Time of the depression is between both posts.

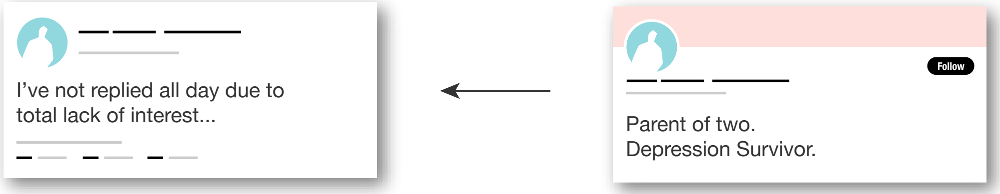
Heuristic Distance

Theory

Use knowledge from later posts *(short distance)*



Use knowledge from user bio



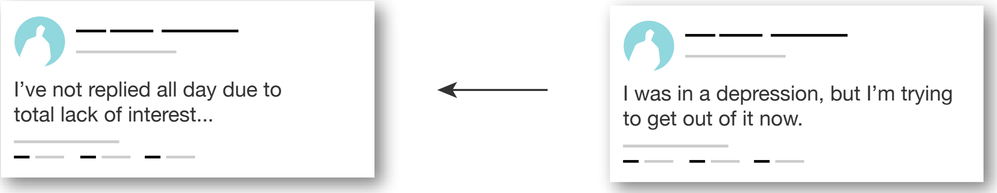
Heuristics:

- ❑ Time of the depression is between both posts.
- ❑ Time of the depression is the complete post history.

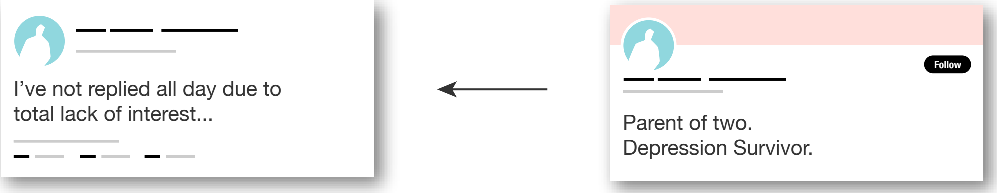
Heuristic Distance

Theory

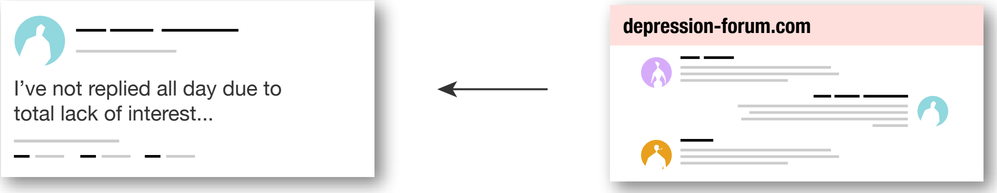
Use knowledge from later posts *(short distance)*



Use knowledge from user bio



Use knowledge from external site *(long distance)*



Heuristics:

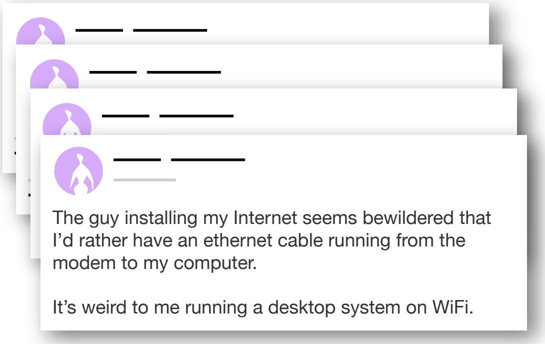
- ❑ Time of the depression is between both posts.
- ❑ Time of the depression is the complete post history.
- ❑ Identical account names mean it is same person.
- ❑ Forum users are in a depression.

Part 3

Profiling Influencers on Twitter

Task: Author Profiling

Given a Twitter timeline, determine the user's personal attributes.



Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

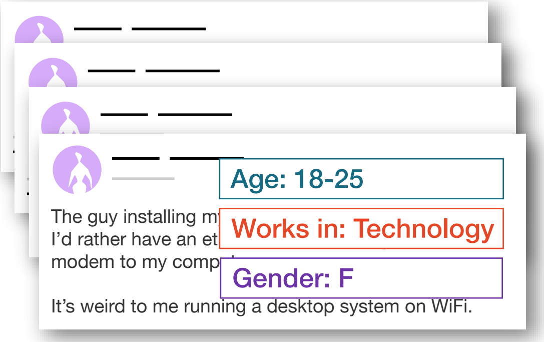
Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Task: Author Profiling

Given a Twitter timeline, determine the user's personal attributes.



Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Problems for human annotation

- Many labels are rare.
- Humans can not assign the labels.

Profiling Influencers on Twitter

Heuristic labeling function

Link Twitter accounts to Wikidata pages.

Verified Users
297 878

WIKIDATA

Platform
Twitter

Data
Timeline of a user’s tweets

Size
71K timelines
239 attributes

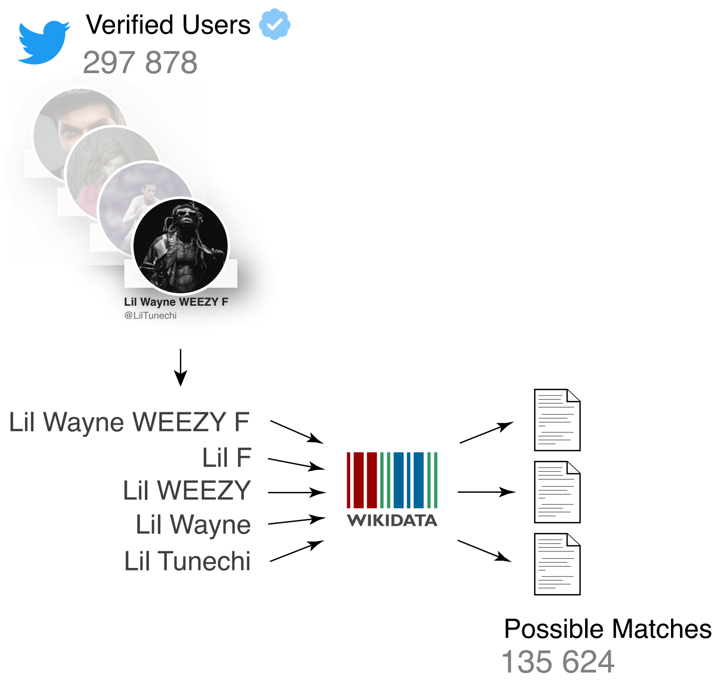
Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Heuristic labeling function

Link Twitter accounts to Wikidata pages.



Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

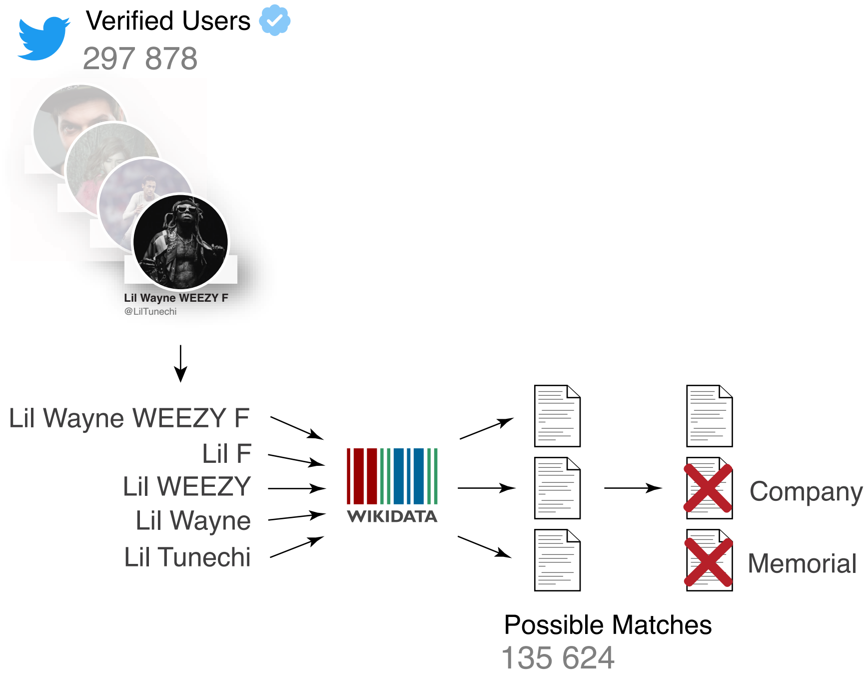
Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Heuristic labeling function

Link Twitter accounts to Wikidata pages.



Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

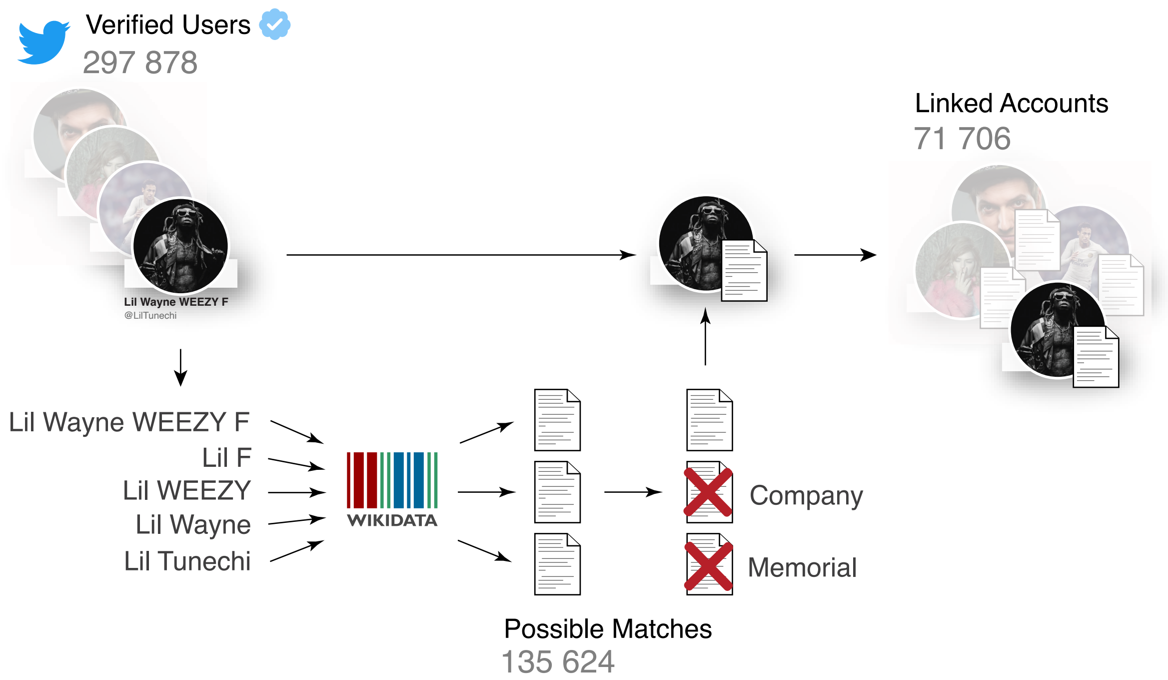
Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Heuristic labeling function

Link Twitter accounts to Wikidata pages.



Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Case Studies

Evaluation of the labeling function

Weak labels

- 28K Wikidata entities contain a Twitter handle.

$\rightsquigarrow$ 7,751 are not in our dataset (0.72 recall)

$\rightsquigarrow$ 124 are incorrectly linked (0.99 precision)

- Errors can be attributed to the individual name candidate rules.

Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Profiling Influencers on Twitter

Answering research questions

RQ 1. Can we transfer profilers between populations?

- Transfer learning [ACL 2019]
  Train and test on different datasets
- Shared task evaluation [CLEF 2019]
  Finding the best classifiers; 8 submissions

RQ 2. Are fan posts indicative of influencer attributes?

- Profiling via follower tweets [CLEF 2020]
  Shared task evaluation; 3 submissions

Platform
Twitter

Data
Timeline of a user's tweets

Size
71K timelines
239 attributes

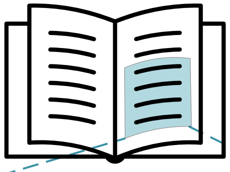
Knowledge
Database (Wikidata)

Evaluation
Weak Labels

Trigger Warning Assignment

Task: Trigger Warning Assignment

Given a document, assign it a warning label if needed.



The disfigurement of each hapless undead body, some missing limbs, covered in blood and ooze, ...

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

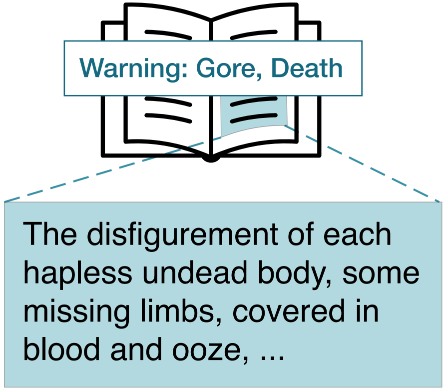
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Task: Trigger Warning Assignment

Given a document, assign it a warning label if needed.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

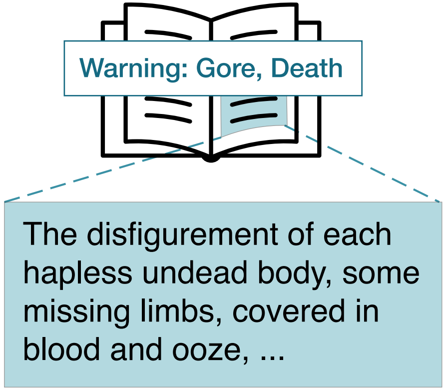
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

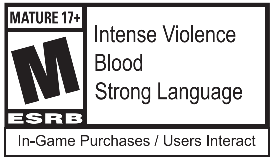
Trigger Warning Assignment

Task: Trigger Warning Assignment

Given a document, assign it a warning label if needed.



ESRB Game Ratings



MPAA Movie Ratings

Problems for human annotation

- Documents are too long for annotation.
- Some objectionable topics are very rare.

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Heuristic labeling function

Link freeform text descriptors to a label taxonomy.

Fiction Documents
7.9 Million

Rating:	Teen And Up Audiences
Archive Warning:	Graphic Depictions Of Violence
Fandom:	僕のヒーローアカデミア | Boku no Hero Academia | My Hero Academia
Relationships:	Midoriya Izuku & Yagi Toshinori | All Might, Midoriya Izuku & Todoroki Shouto, Midoriya Izuku & Uraraka Ochako, Iida Tenya & Midoriya Izuku
Characters:	Midoriya Izuku, Yagi Toshinori | All Might, Midoriya Inko, Shimura Nana, Bakugou Katsuki, Todoroki Shouto, Uraraka Ochako, Iida Tenya, Iida Tensei | Ingenium
Additional Tags:	Alternate Universe - Canon Divergence, BAMF Midoriya Izuku, The Sixth Sense AU, Bakugou Katsuki Swears A Lot, Izuku Sees Dead People, Queerplatonic Relationships, Midoriya Izuku Has a Quirk, Platonic Slow Burn, Panic Attacks, past trauma, Body Horror, Character Death, Implied/Referenced Child Abuse, CONTENT WARNINGS CAN BE FOUND IN CHAPTER ENDNOTES
Language:	English
Stats:	Published: 2016-10-21 Completed: 2019-10-12 Words: 424,070 Chapters: 60/60 Comments: 24,894 Kudos: 95,593 Bookmarks: 23,262 Hits: 3,501,502

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

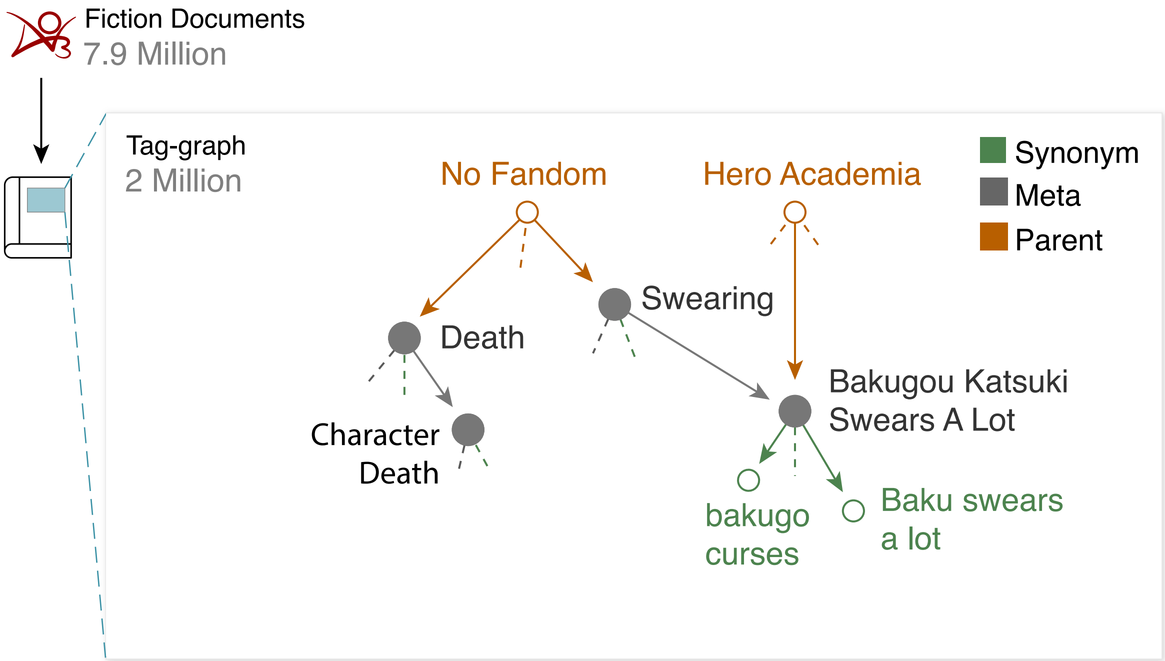
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Heuristic labeling function

Link freeform text descriptors to a label taxonomy.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

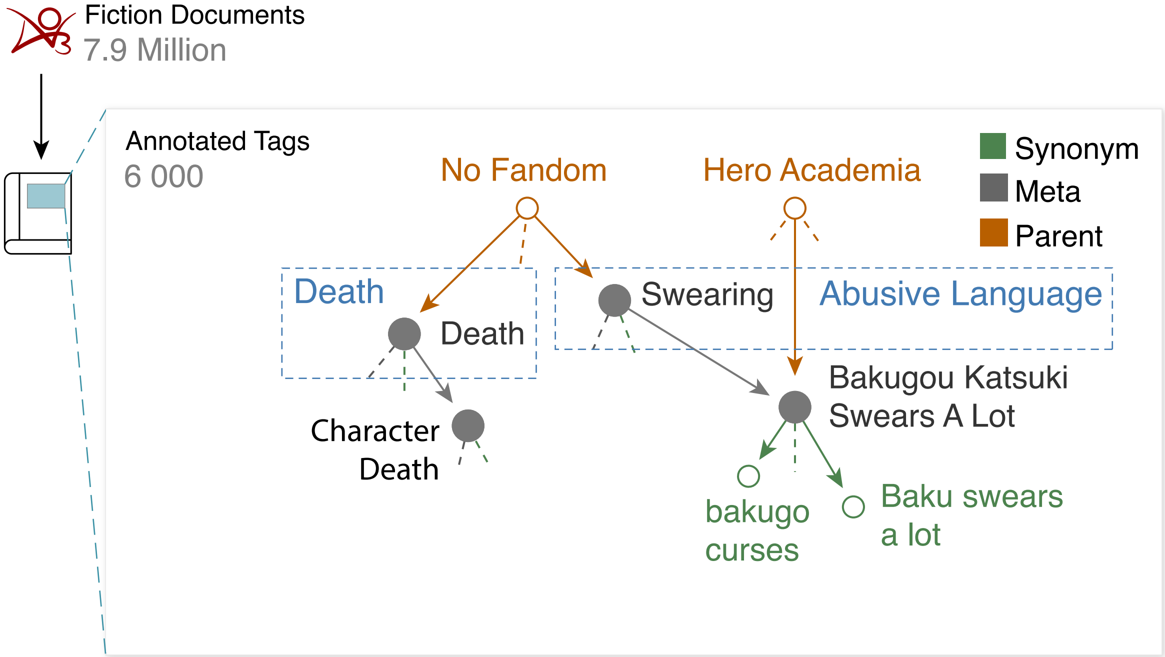
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Heuristic labeling function

Link freeform text descriptors to a label taxonomy.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

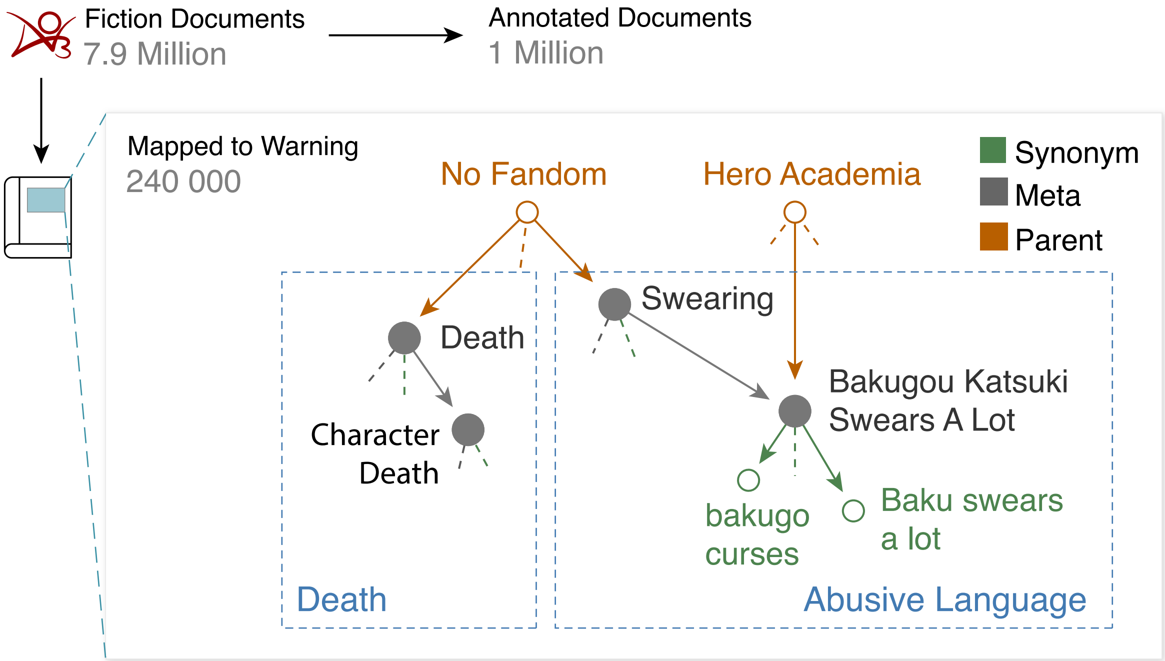
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Heuristic labeling function

Link freeform text descriptors to a label taxonomy.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Case Studies

Evaluation of the labeling function

Spot checks

- Manually annotated test sets.
- 0.94 F_1 on 2,000 most common tags.
- 0.96 F_1 on 10-11k most common tags.

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Case Studies

Evaluation of the labeling function

Spot checks

- Manually annotated test sets.
- 0.94 F_1 on 2,000 most common tags.
- 0.96 F_1 on 10-11k most common tags.

But
Tag-graph covers only ~80% of tag occurrences and ~20% of all unique tags.

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Case Studies

Answering research questions

RQ 1. Can we assign trigger warnings to documents?

- ❑ Violence Classification [EMNLP 2023]
  Input vs document length, popularity, confounder analysis
- ❑ Multi-label Classification [ACI 2023]
  Role of support for each tag, granularity of the taxonomy
- ❑ Shared Task Evaluation [PAN@CLEF 2023]
  Finding the best classifiers; 6 submissions

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

Trigger Warning Assignment

Case Studies

Answering research questions

RQ 2. Does label noise influence model evaluation?

- ❑ Noise Reduction [CLEF 2024]
  LLM-based pruning to remove noisy labels from test data

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Labels

1. **Surveying successful applications** to establish a theoretic foundation.

2. **Constructing novel datasets** via new, complex labeling functions.

3. **Answering research questions** based on the new datasets.

Profiling Influencers on Twitter
[Wiegmann et al., ACL 2019] [Wiegmann et al., PAN@CLEF 2019] [Wiegmann et al., PAN@CLEF 2020]

Analyzing the Persuasiveness of Debaters
[Wiegmann et al., COLING 2022]

Trigger Warning Assignment
[Wiegmann et al., ACl 2023] [Wiegmann et al., PAN@CLEF 2023] [Wolska and Wiegmann et al., EMNLP 2023]
[Wiegmann et al., CLEF 2024]

Appendix

Distant Knowledge

- Curated list
  Emoticons to emotion label
  Phrases (*“I’m 35 as of today”*)to demographic group
- Database (structured, unstructured)
  Wikidata, a database of known bots, Google . . .
- Metadata (direct, distant, computed)
  Geo-tags as user home location . . .
- Classifiers

Evaluation strategies

- ❑ Spot checks
- ❑ Weak labels
- ❑ Annotated data
- ❑ Models

Profiling Influencers on Twitter

Evaluation of the labeling function

- ❑ 28K Wikidata entities contain a Twitter handle.
- $\rightsquigarrow$ 7,751 are not in our dataset (0.72 recall)
- $\rightsquigarrow$ 124 are incorrectly linked (0.99 precision)

Error rates and matches by name candidate:

	Name Candidate Rule						
	I	II	III	IV	V	VI	all
Matches	91.8%	2.8%	0.1%	1.8%	2.9%	0.3%	71,706
Errors	50.0%	3.2%	0.0%	23.3%	21.8%	1.6%	124

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Name Candidate Rules

(1) Remove non-alphanumeric characters from *display name*.

(2) Split *handle* at capitalized characters. (`@FirstLast`)

(3) Split off the *display name* from the *handle*.

(4) Split (1) on whitespace, use first and last parts.

(5) Split (1) on whitespace, use all but the last part.

(6) Split (1) on whitespace, use all but the last two parts.

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Labels

Label	Occurrences		Most frequent value	
Sex	65,035	90.1%	Male	71.7%
Occupation	63,017	87.9%	Actor	15.3%
Date of birth	60,493	84.4%	-	-
Educated at	28,134	39.2%	Harvard	2.1%
Sport	18,688	26.1%	Football	30.8%
Languages spoken	12,094	16.9%	English	54.9%
Political party	6,703	9.4%	Republican	16.4%
Genre	6,699	9.3%	Pop Music	21.6%
Race	3,531	0.5%	African Am.	66.5%
Religion	2,960	0.4%	Islam	23.5%

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Classifier transfer

Model	Test Dataset				
	PAN15	PAN16	PAN17	PAN18	Celeb
alvarezcamona15	**0.859**	–	–	–	0.723
nissim16	–	**0.641**	–	–	0.740
nissim17	–	–	**0.823**	–	0.855
danehsvar18	–	–	–	**0.822**	0.817
CNN (Celeb)	0.747	0.590	0.747	0.756	**0.861**

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Shared task evaluation campaign.

Classification across four personal attributes.

Participant	Gender (3)	Age (5)	Renown (3)	Occupation (8)
Radivchev	**0.609**	**0.657**	**0.548**	0.461
Pelzer	0.547	<u>0.518</u>	0.460	<u>0.481</u>
Moreno-Sandoval	0.561	0.516	0.518	0.418
Martinc	<u>0.594</u>	0.347	0.507	**0.486**
Petrik	0.555	0.360	<u>0.526</u>	0.385
Fernquist	0.465	0.467	0.482	0.300
Asif	0.588	0.254	0.504	0.427
Bryan	0.335	0.207	0.289	0.165

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Shared task evaluation campaign.

Class-wise scores of the most effective submitted system.

Gender	F_1
Male	0.951
Female	0.881
Diverse	0.307

Renown	F_1
High	0.874
Medium	0.469
Low	0.261

Occupation	F_1
Sports	0.90
Entertainer	0.79
Politician	0.74
Creator	0.57
Scientist	0.32
Clergy	0.27
Manager	0.23
Professional	0.21

Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

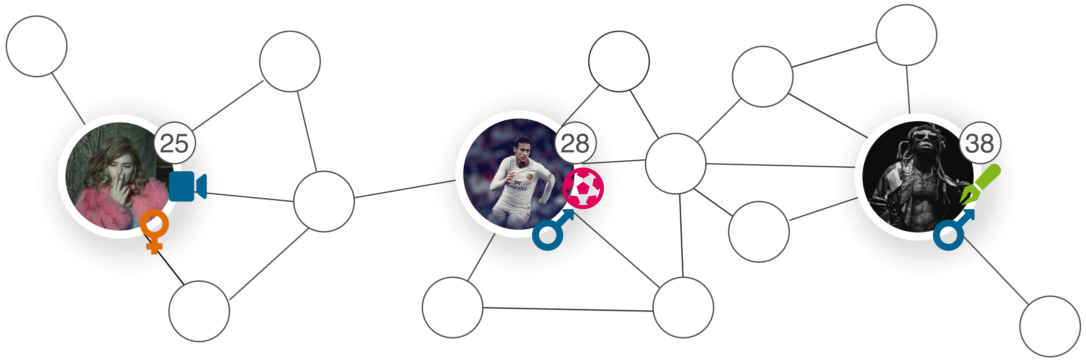
Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Profiling via follower tweets. [CLEF 2020]

Dataset extension method



Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

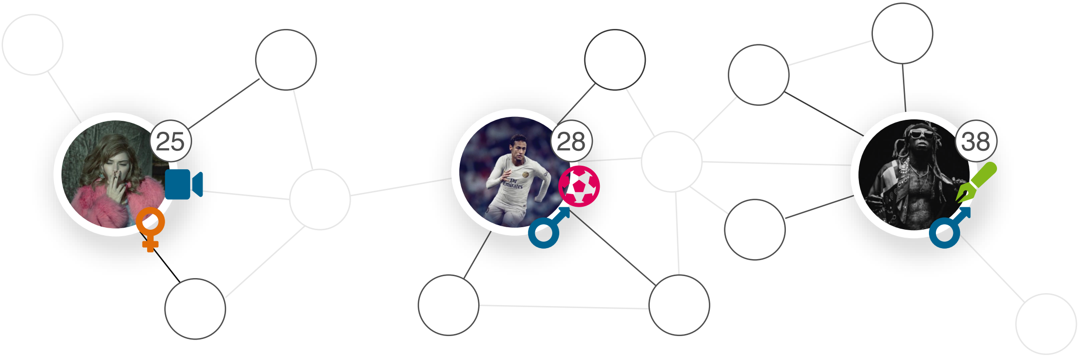
Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Profiling via follower tweets. [CLEF 2020]

Dataset extension method



Platform
Twitter.

Data
Timeline of a users tweets.

Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Profiling Influencers on Twitter

Profiling via follower tweets. [CLEF 2020]

Results of the shared task evaluation

Participant	**Age** (5)	**Gender** (2)	**Occupation** (4)
baseline-oracle	0.50	0.75	0.70
Hodge	0.43	0.68	0.71
Koloski	0.41	0.62	0.60
Alroobaea	0.32	0.70	0.60
baseline	0.36	0.58	0.52

Appendix

Platform
Twitter.

Data
Timeline of a users tweets.

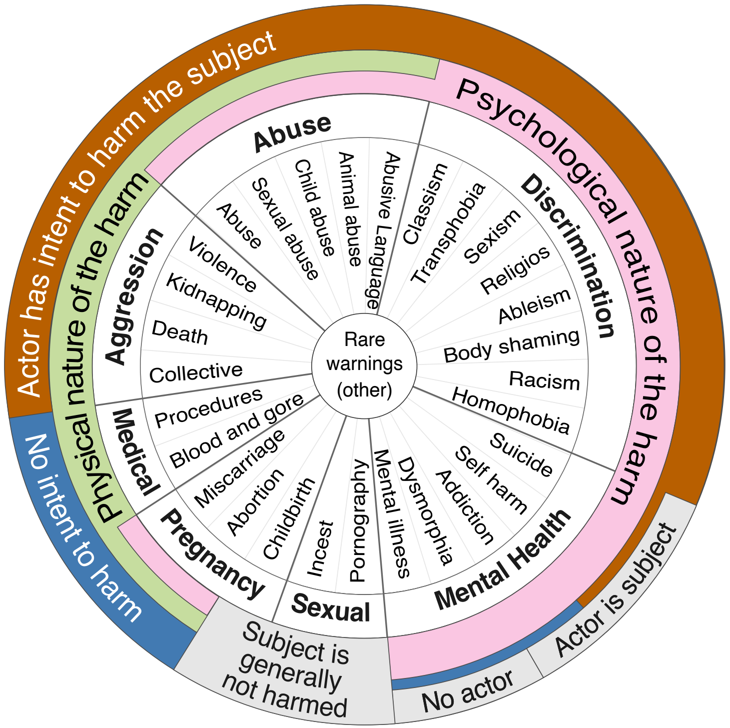
Size
71K timelines.
239 different attributes.

Knowledge
Database (Wikidata properties).

Evaluation
Weak Labels.

Trigger Warning Assignment

Warning Taxonomy



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

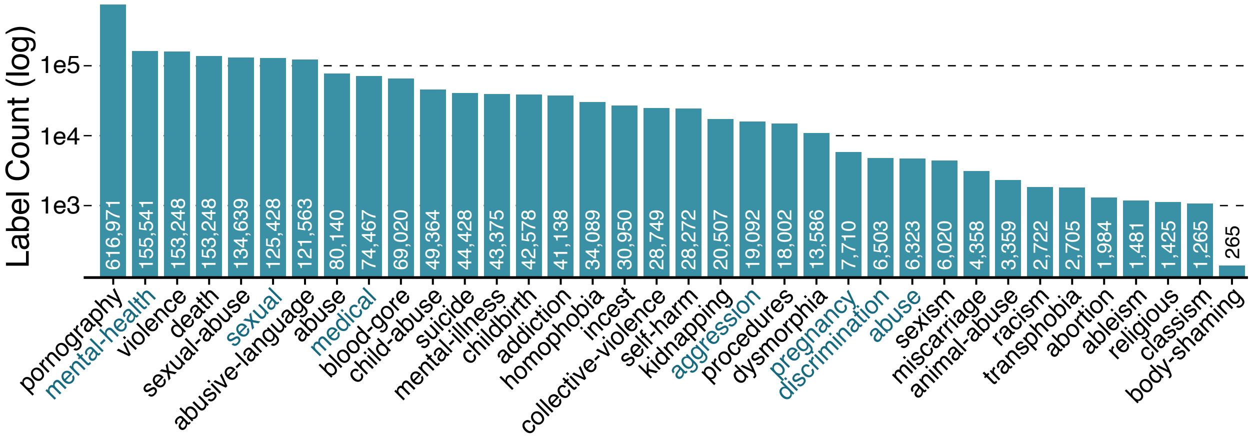
Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Dataset Statistics



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

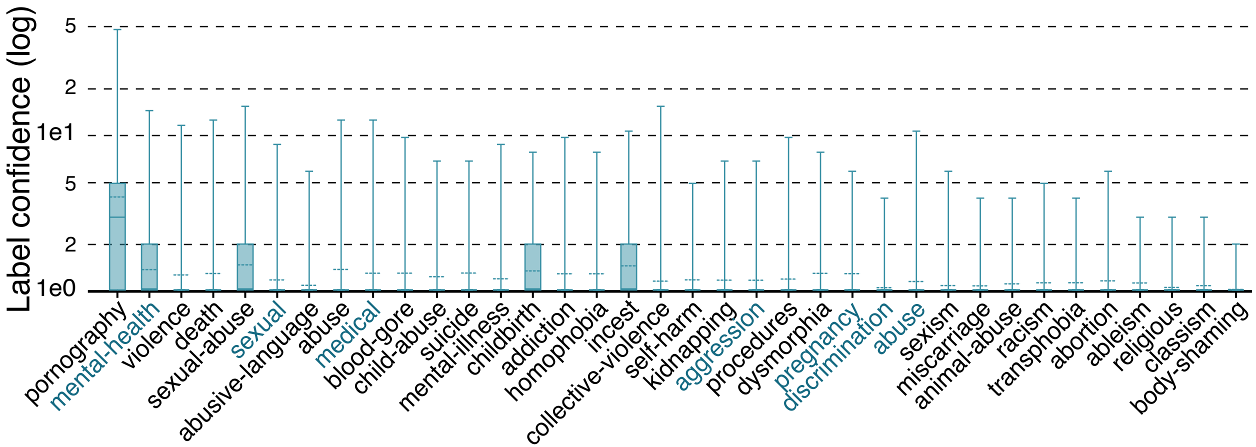
Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Dataset Statistics



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

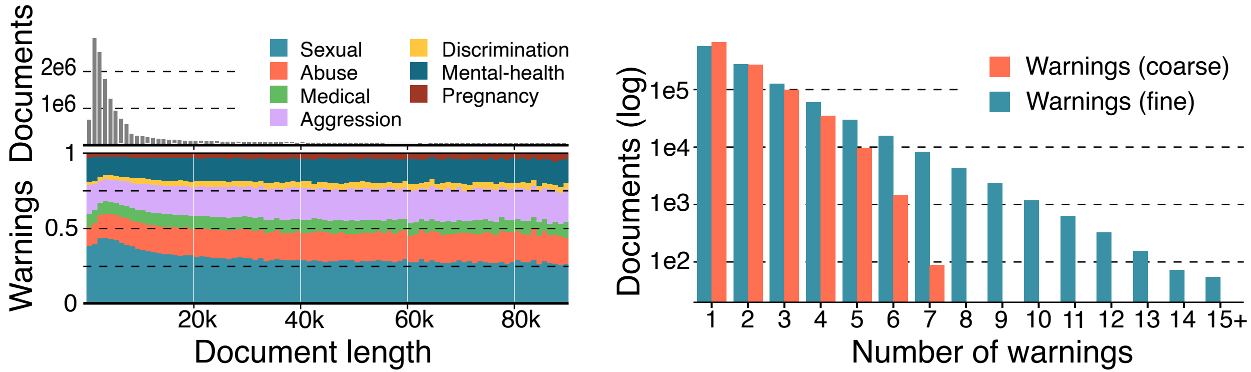
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Dataset Statistics

Left: Warning distribution by document length.
Right: Number of warnings per document.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Evaluation of the labeling function

Manually annotated test sets:

- 0.94 F_1 on 2,000 most common tags.
- 0.96 F_1 on 10-11k most common tags.

Via verbatim warnings. (‘warning: abuse’, ‘tw: needles’, . . .)

	Occurrences	**Unique Tags**
Total	62,316	27,694
Classified as warning	34,806	9,595
- of all wrangled	0.86	0.79
- of all free-form	0.56	0.35

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

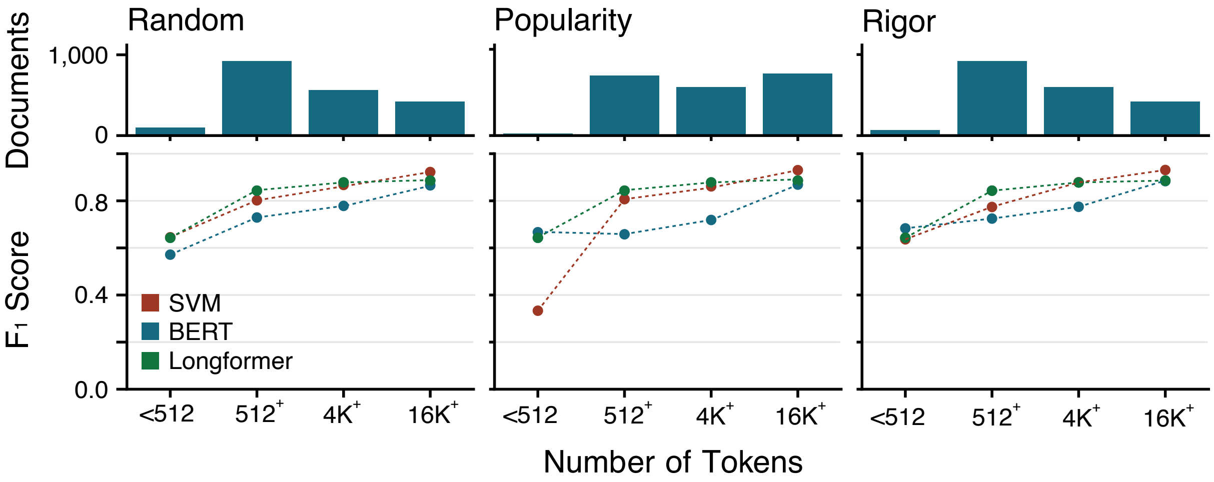
Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Violence Classification.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Appendix

Violence Classification.

Features indicating violence

Random		Popularity		Rigor	
4.65	blood	3.82	blood	4.54	blood
2.40	dead	2.32	screams	2.62	dead
2.37	kill	2.02	scream	2.23	screams
2.33	screams	1.94	dead	2.13	pain
1.99	screamed	1.91	kill	2.03	bloody
1.95	flesh	1.89	pain	1.96	scream
1.89	screaming	1.89	killed	1.93	bleeding
1.86	scream	1.84	bloody	1.93	blade
1.79	pain	1.81	bleeding	1.91	kill
1.77	killed	1.75	blade	1.87	killed
⋮		⋮		⋮	
0.91	hannibal (84)	0.55	sith (341)	0.97	hannibal (67)

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Violence Classification.

Features indicating non-violence

Random		Popularity		Rigor	
-1.67	kiss	-1.16	kiss	-1.86	kiss
-1.07	managed	-0.96	embarrassing	-1.00	teasing
-1.01	ridiculous	-0.91	halfway	-0.93	spent
-0.92	admit	-0.90	experience	-0.92	demanded
-0.91	teasing	-0.90	surprised	-0.90	hadn
-0.91	shoulders	-0.87	close	-0.89	fin
-0.89	snorted	-0.82	dance	-0.89	flushed
-0.89	curled	-0.81	teasing	-0.87	imagined
-0.88	weekend	-0.80	ridiculous	-0.85	ridiculou
-0.88	surprised	-0.80	kissing	-0.84	carefully

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

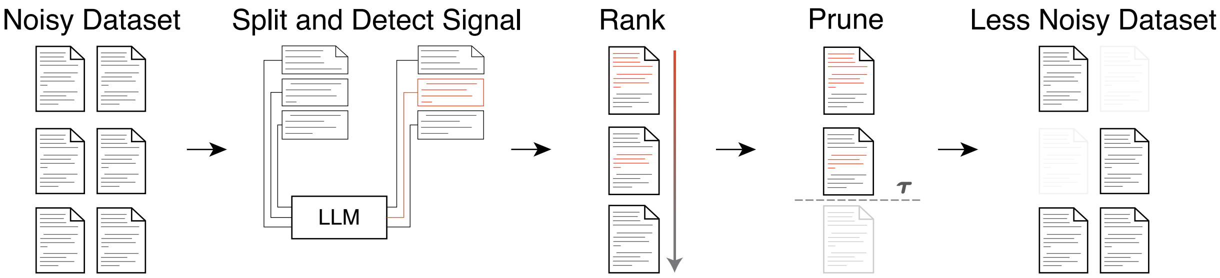
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Noise Reduction.

Estimate the aggregated “signal strengt” for each label.



Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Noise Reduction Evaluation.

1. Find reliable labels $\rightsquigarrow$ should not be removed.

Some authors add detailed warnings to individual chapters.

Chapter 3

Notes:

Edit 12/26/17: By popular demand and my own personal desire, I have made a minor aesthetic modification to Izuku in this story; this chapter has been edited to include it.

CW: Gore, discussions of past domestic abuse, car accidents, and murder.

[PitViperOfDoom, 2016]

Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

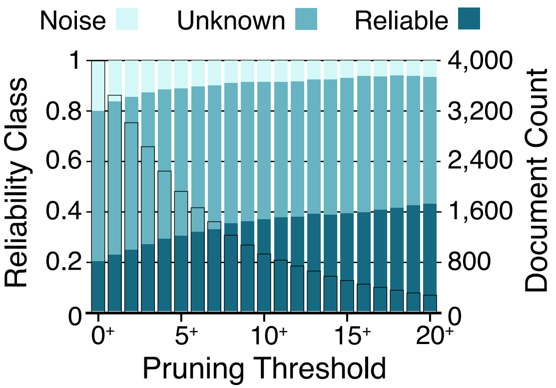
Knowledge
Curated List, Document Metadata

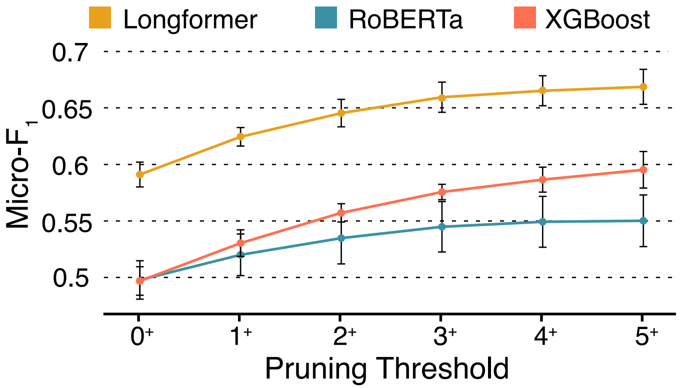
Evaluation
Spot Checks, Weak Label

Trigger Warning Assignment

Noise Reduction Evaluation.

1. Find reliable labels ⇝ should not be removed.
2. Add artificial label noise ⇝ should be removed.
3. Model F_1 and model differences should increase.





Platform
Archive of Our Own (AO3)

Data
Fanfiction documents

Size
1M documents
36 labels

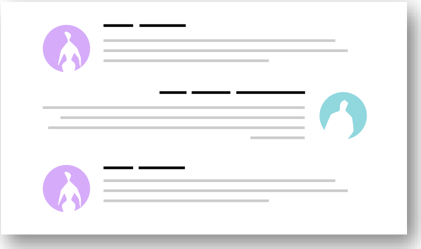
Knowledge
Curated List, Document Metadata

Evaluation
Spot Checks, Weak Label

Persuasiveness of Debaters

Task: Debater analysis

Given a debaters post histroy, is the debater persuasive or not?



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Task: Debater analysis

Given a debaters post histroy, is the debater persuasive or not?

Platform

Reddit (`/r/ChangeMyView`)

Data

Debater histories

Size

3.8K histories

3 labels

Knowledge

Metadata (Delta)

Evaluation

–

Persuasiveness of Debaters

Task: Debater analysis

Given a debaters post histroy, is the debater persuasive or not?

Problem for human annotation

- ❑ Persuasiveness is subjective.
- ❑ Need many debates for each of many debaters.

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

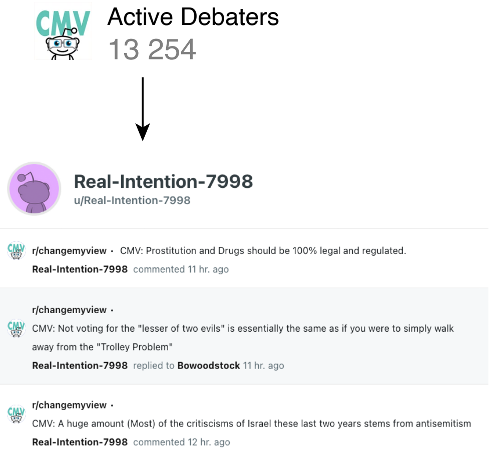
Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Heuristic labeling Function

Aggregate debate delta across debate histories.



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

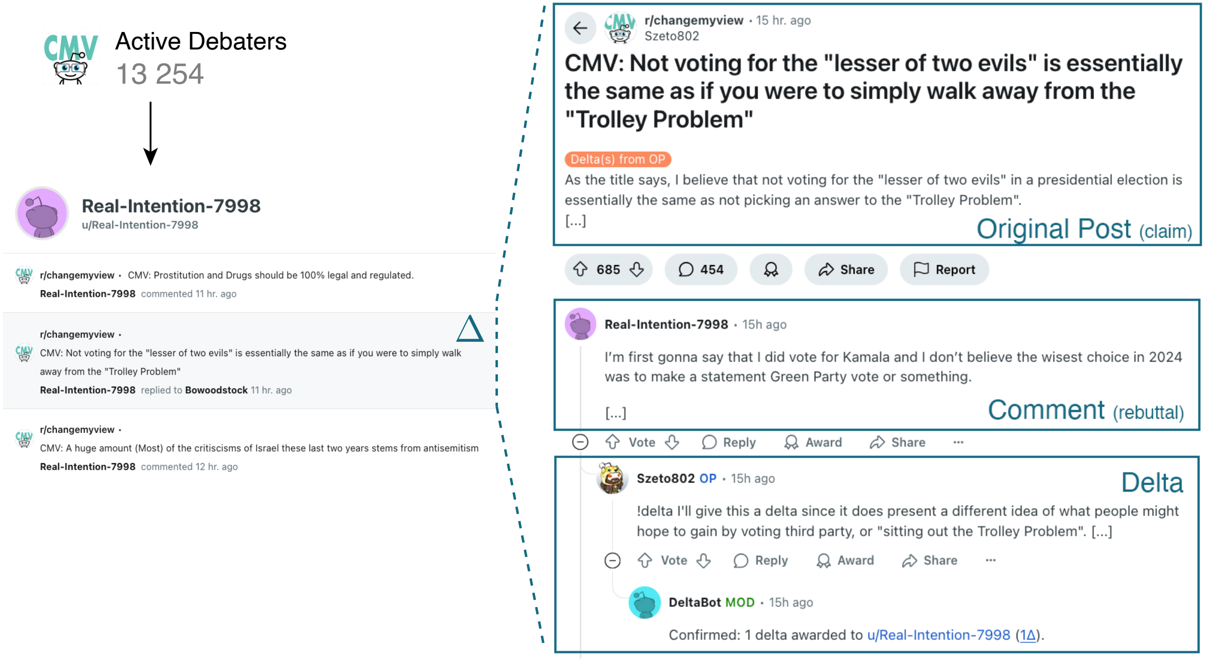
Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Heuristic labeling Function

Aggregate debate delta across debate histories.



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

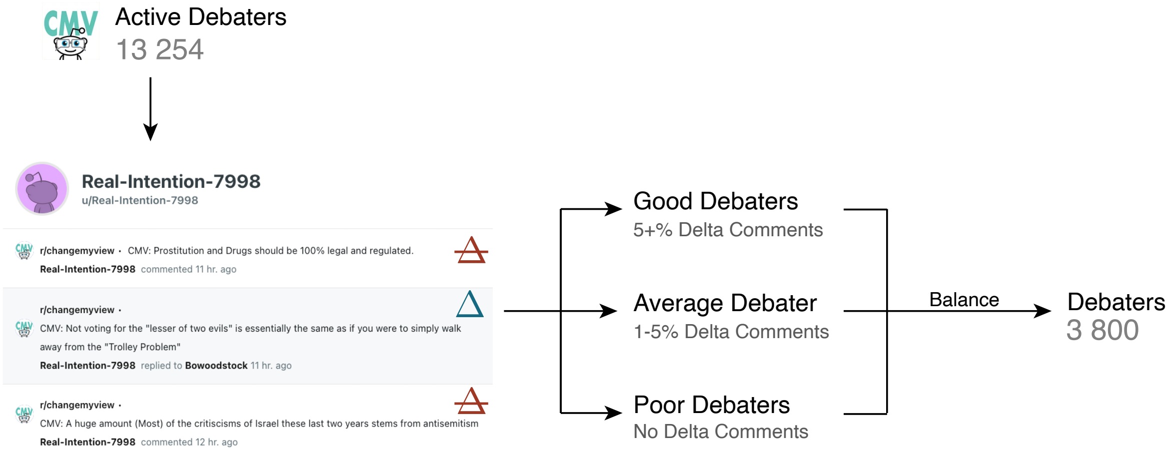
@Wiegmann, 2025

Persuasiveness of Debaters

Appendix

Heuristic labeling Function

Aggregate debate delta across debate histories.



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Appendix

Answering research questions

RQ 1. Why are some debaters more persuasive than others?

- Diachronic analysis. [COLING 2022]
  Role of engagement and experience in persuasiveness
- Feature analysis.
  Which features predict persuasiveness in a classifier?
- Style analysis.
  Which lexical, syntactic, and semantic features explain persuasiveness?

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

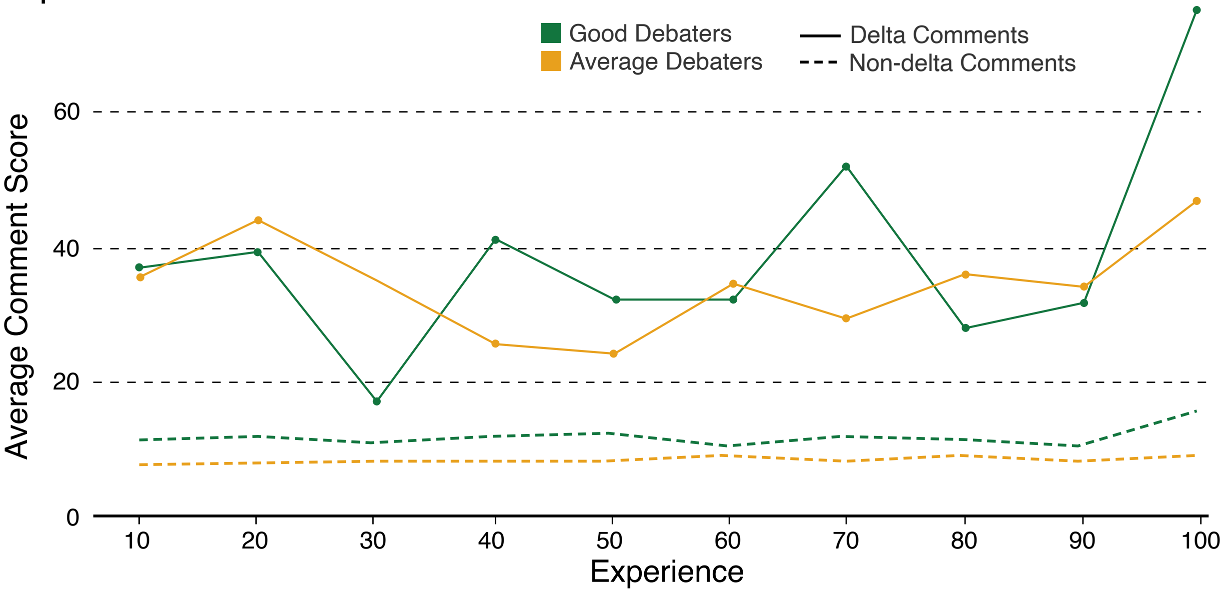
Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Diachronic analysis

Comment score of delta/non-delta comments with increasing debater experience.



Appendix

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

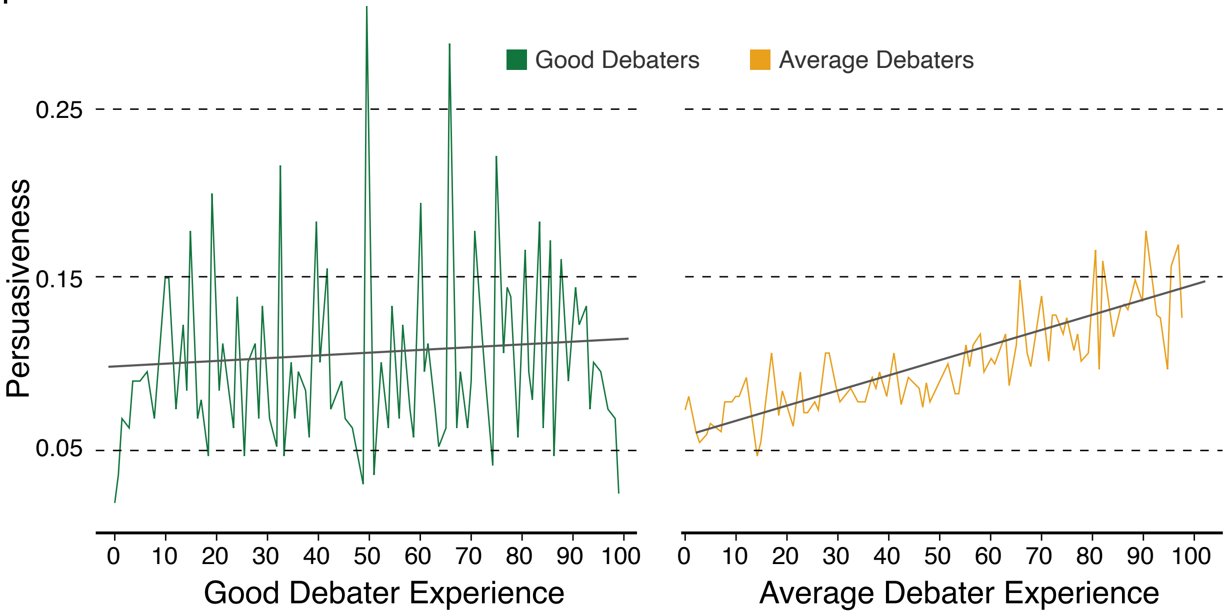
Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Diachronic analysis

Persuasiveness with increasing experience for debaters in different persuasiveness classes.



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Feature analysis

Features	Good vs	
	Average	Poor
Baseline Features		
Bag of Words	0.60	0.68
Stylometry	0.62	0.67
Vocabulary Interplay	0.58	0.67
Syntactic Features		
Word class n-grams	0.57	0.51
Text Complexity	0.65	0.61
Semantic Features		
Word Mover's Distance	0.59	0.63

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Feature analysis

Features	Good vs	
	Average	Poor
Pragmatic Features		
Elementary Units	0.51	0.59
Claim or Premise	0.47	0.55
Claim Type	0.48	0.58
Premise Type	0.48	0.58
Claim and Premise Types	0.48	0.58
Frames	**0.70**	**0.72**

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Style analysis

Persuasive debaters

- write long comments,
- have lower lexical diversity and syntactic complexity,
- have a higher semantic diversity,
- more often use rhetorical statements, and
- more often use political and cultural identity frames.

Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

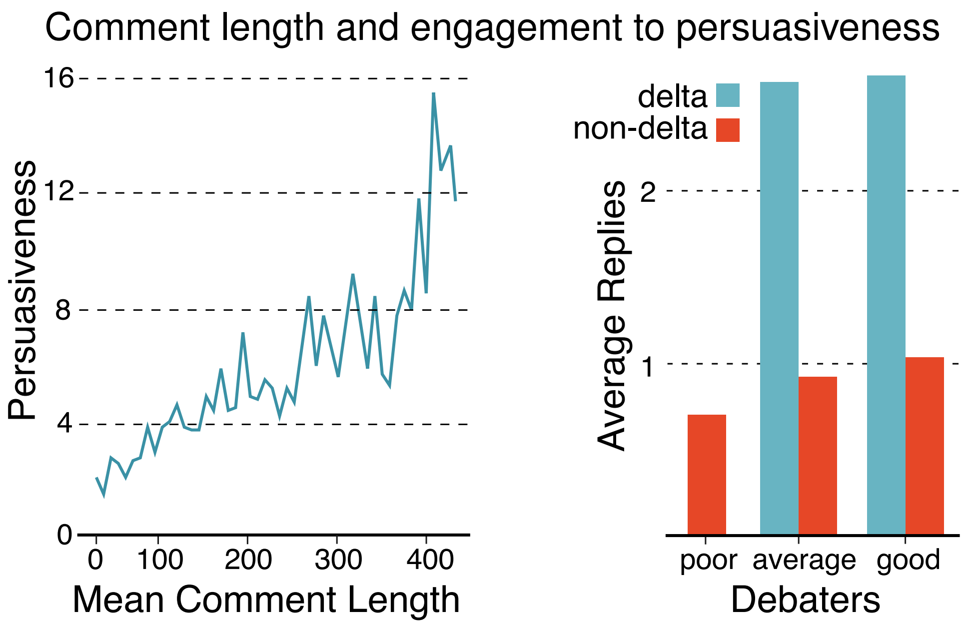
Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–

Persuasiveness of Debaters

Style analysis



Platform
Reddit (`/r/ChangeMyView`)

Data
Debater histories

Size
3.8K histories
3 labels

Knowledge
Metadata (Delta)

Evaluation
–