Computational Approaches to Unveiling Biases in Stories and Models

Yulia Tsvetkov

How do we make decisions

System 1

automatic

fast
parallel
automatic
effortless
associative
slow-learning

System 2

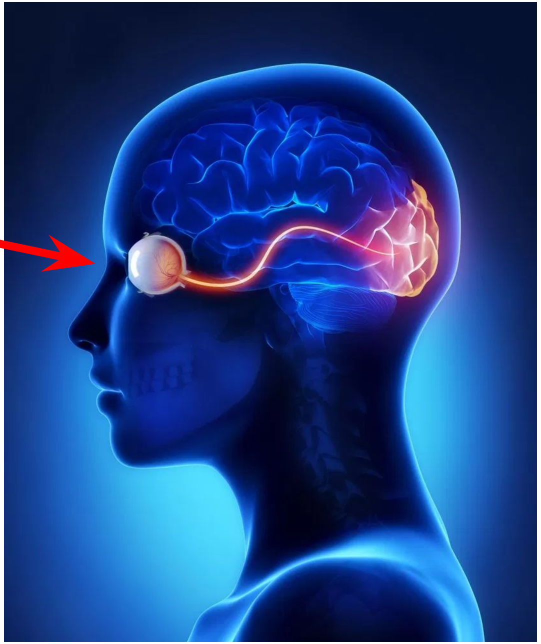
effortful

slow
serial
controlled
effort-filled
rule-governed
flexible

Kahneman & Tversky 1973, 1974, 2002

The brain needs to delegate as much as possible to System 1

~10MP



System 1

automatic

System 2

effortful

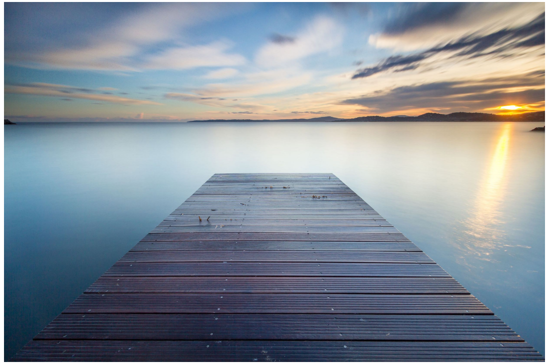
Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments.

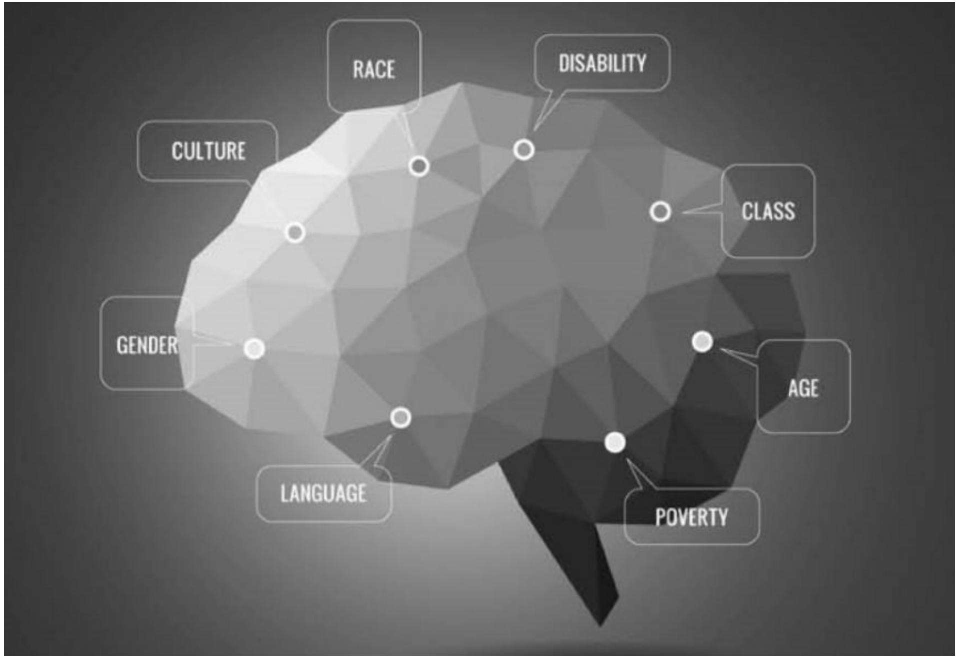
Although we identify with System 2, over 95% of cognition is relegated to the System 1’s “auto-pilot.”

Psychological perspective on implicit bias

Stereotypes inevitably form because of the innate tendency of the human mind to:

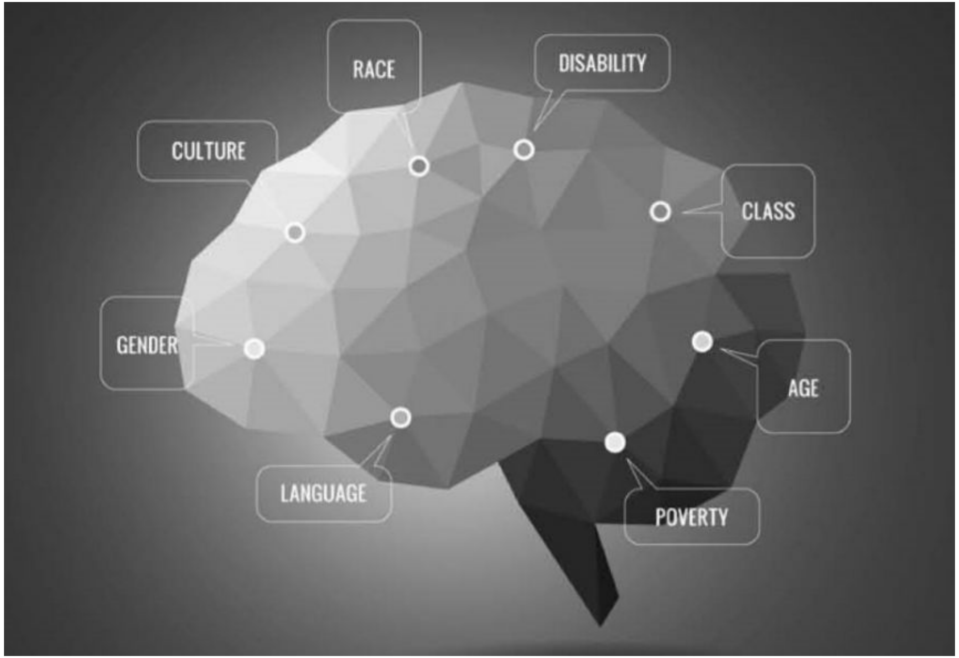
- Categorize the world to simplify processing
- Store learned information in mental representations (called schemas)
- Automatically and unconsciously activate stored information whenever one encounters a category member





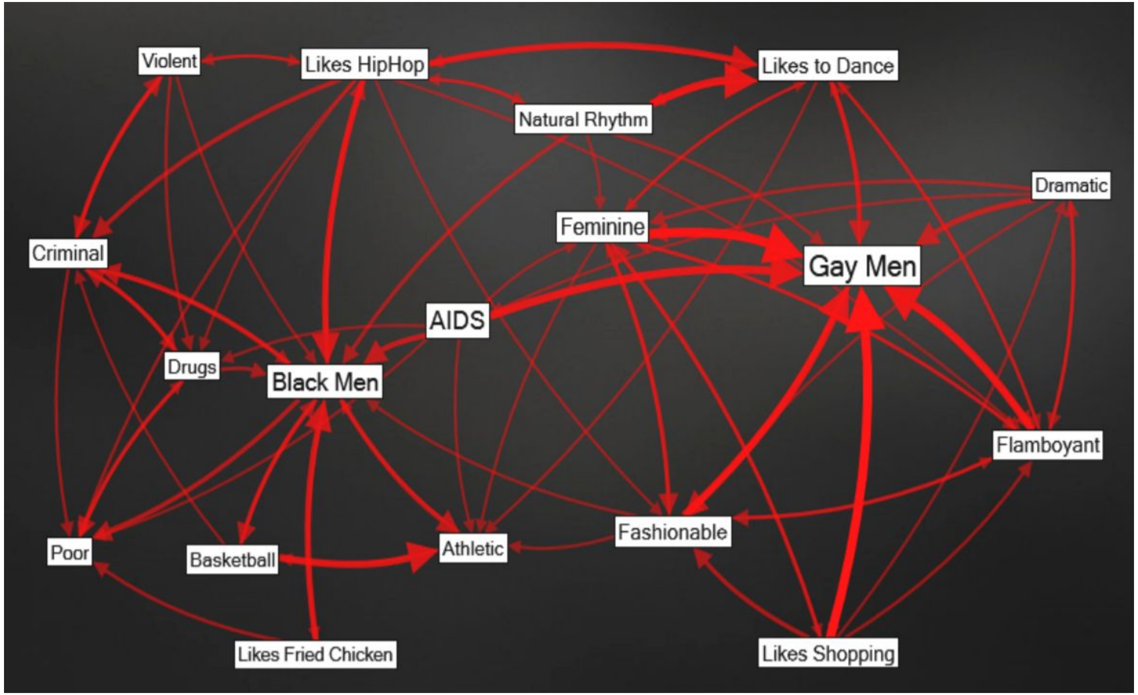
[Image credit: Geoff Kaufman]

Stereotypes are internalized as associations through natural processes of learning and categorization



[Image credit: Geoff Kaufman]

Social stereotypes are not necessarily negative, but still have negative effect



Implicit biases are distressingly pervasive, operate largely unconsciously, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

AI is only System 1

- Conversational agents
- Personal assistants
- Search engines
- Translation engines
- Medical research assistants

Online data is riddled with **SOCIAL STEREOTYPES**

Biased NLP technologies

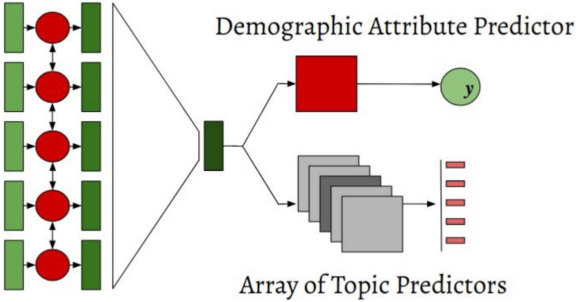
- Bias in word embeddings (Bolukbasi et al. 2017; Caliskan et al. 2017; Garg et al. 2018)
- Bias in Language ID (Blodgett & O'Connor. 2017; Jurgens et al. 2017)
- Bias in Visual Semantic Role Labeling (Zhao et al. 2017)
- Bias in Natural Language Inference (Rudinger et al. 2017)
- Bias in Coreference Resolution (Rudinger et al. 2018; Zhao et al. 2018)
- Bias in Automated Essay Scoring (Amorim et al. 2018)
- Bias in Sentiment Analysis (Kiritchenko & Mohammad et al. 2018)
- Bias in Text Classification (De-Arteaga et al. 2019)
- Bias in Machine Translation (Prates et al. 2018)

This talk: veiled implicit biases in data & text classification

Part 1

[Field et al., ICWSM'19]

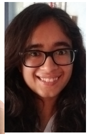
Part 2



[Kumar et al., ongoing]

Contextual Affective Analysis: A Case Study of People Portrayals in Online #MeToo Stories

[Field et al., ICWSM'19]



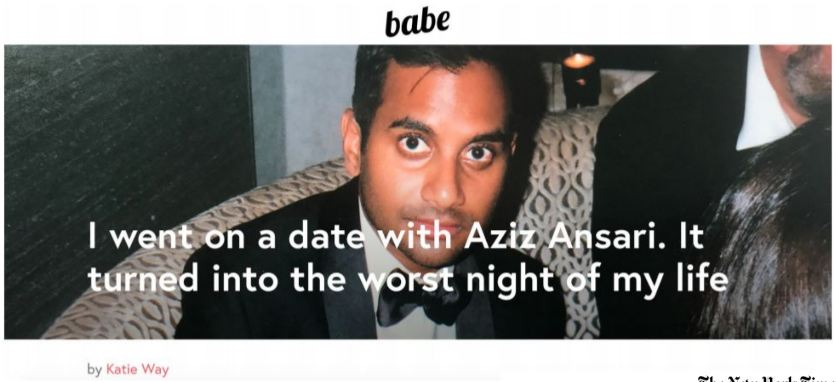
Background: the #MeToo movement

- 2006: Tarana Burke coins phrase "Me Too." Burke is a survivor of sexual assault and wanted to do something to help women and girls of color who had also survived sexual violence
- Oct. 5 2017: Actress Ashley Judd accuses media mogul Harvey Weinstein in a breaking story by The New York Times.
- Oct 15 2017: Actress Alyssa Milano reignites "Me Too" with the tweet "If you've been sexually harassed or assaulted write 'me too' as a reply to this tweet," and it quickly turned into a movement.
- Oct 18 2017: Olympic gymnast McKayla Maroney tweets that she was sexually assaulted by former team doctor Lawrence G. Nassar

  ...
- Jan 23 2019: An article published Wednesday online in the Atlantic contains new allegations against "X-Men" Director Bryan Singer,

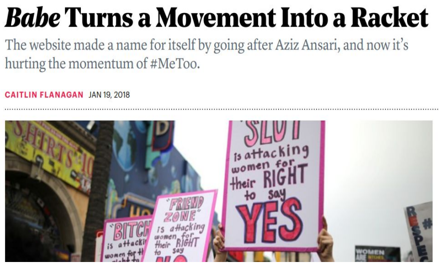
https://www.chicagotribune.com/lifestyles/ct-me-too-timeline-20171208-htmlstory.html

Jan 13, 2018


babe

I went on a date with Aziz Ansari. It turned into the worst night of my life

by Katie Way


***Babe* Turns a Movement Into a Racket**

The website made a name for itself by going after Aziz Ansari, and now it's hurting the momentum of #MeToo.

CAITLIN FLANAGAN JAN 19, 2018


The New York Times

OPINION

Aziz Ansari Is Guilty. Of Not Being a Mind Reader.

By Bari Weiss

Jan. 15, 2018

https://babe.net/2018/01/13/aziz-ansari-28355

Importance of power and agency in narratives of sexual harassment

“The single most distressing thing to me about this story is that the only person with any **agency** in the story seems to be Aziz Ansari. The woman is merely acted upon.”, Bari Weiss, New York Times

Tarana Burke described her goal in founding the #MeToo movement as: “**empowerment** through **empathy**”

Importance of power and agency in narratives of sexual harassment

Something inherently important about portrayals of power and agency: the type of response they elicit from readers

"Victim" vs "Survivor"

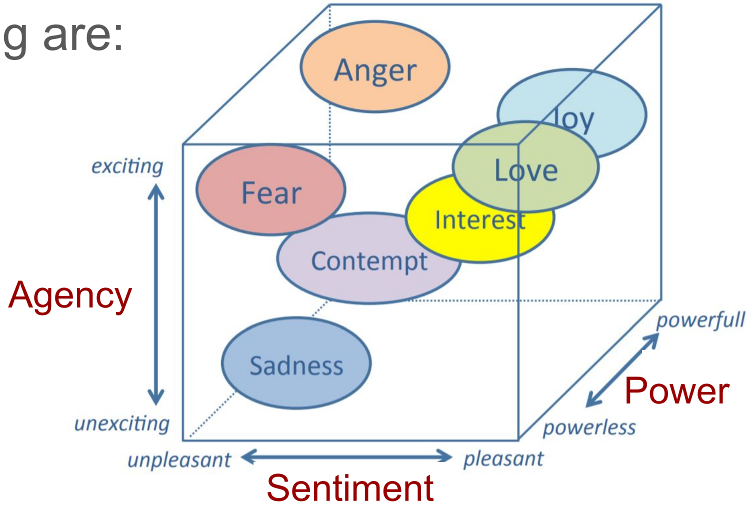
- Someone who underwent trauma; evokes pity
- Survivor: Someone who fought through trauma; evokes admiration

Spry, Tami. "In the absence of word and body: Hegemonic implications of" victim" and" survivor" in women's narratives of sexual violence." Women and Language 13.2 (1995): 27.

Affect Control Theory

Besides a denotative meaning, three most important, largely independent, dimensions of word meaning are:

- Valence / Sentiment
  - positive–negative
  - pleasant–unpleasant
- Arousal / Agency
  - active–passive
- Dominance / Power
  - dominant–submissive



[Image credit: Tobias Schröder]

Osgood, C.; Suci, G.; and Tannenbaum, P. 1957. The Measurement of Meaning. Illini Books, IB47. University of Illinois Press
Mohammad, Saif. "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words." Proc. ACL'18

Research Questions

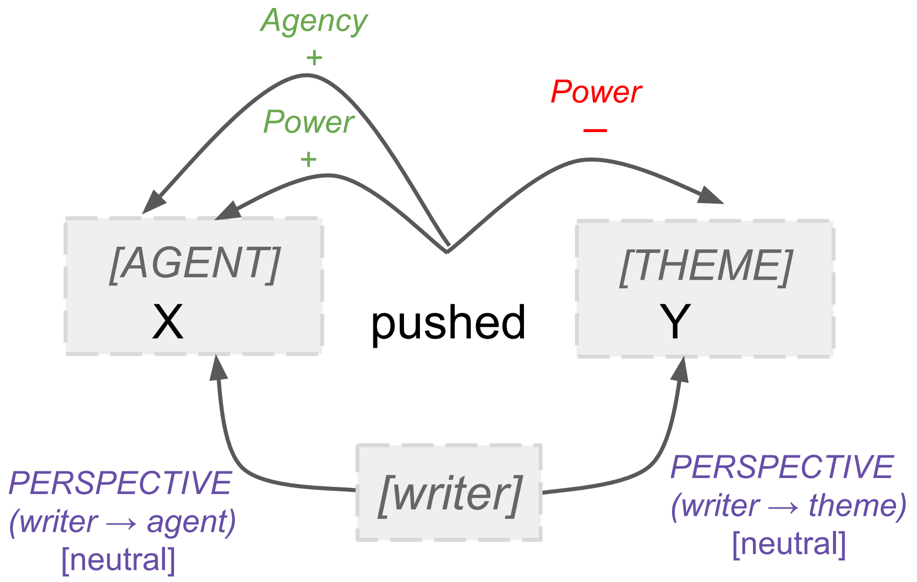
The #MeToo movement has largely been viewed as “empowering” but journalists have a choice in how they portray people (victim vs. survivor)

In news articles about the #MeToo movement:

- Who is portrayed as powerful?
  - Women? Men? Accusers? Accused? Someone else?
- Who is portrayed as sympathetic?
- Who is portrayed as having high agency?
- How do these portrayals differ across narratives and news outlets?

How do we measure power, agency, and sentiment?

Connotation frames (Rashkin et al. 2016)



Rashkin, H., Singh, S, and Choi, Y. "Connotation Frames: A Data-Driven Investigation." Proc. ACL'16
Sap, M., Prasetio, M. C., Holtzman, A., Rashkin, H., Choi, Y., "Connotation Frames of Power and Agency in Modern Films." Proc. EMNLP'17

Connotation frames (Rashkin et al. 2016)

She pushed him away

How do you think she feels about the outcome of this event?

Positive | Either Positive or Neutral | Neutral | Either Negative or Neutral | Negative | Can't have feelings

How do you think he feels about the outcome of this event?

Positive | Either Positive or Neutral | Neutral | Either Negative or Neutral | Negative | Can't have feelings

How the **writer** feels about she:

Positive | Either Positive or Neutral | Neutral | Either Negative or Neutral | Negative

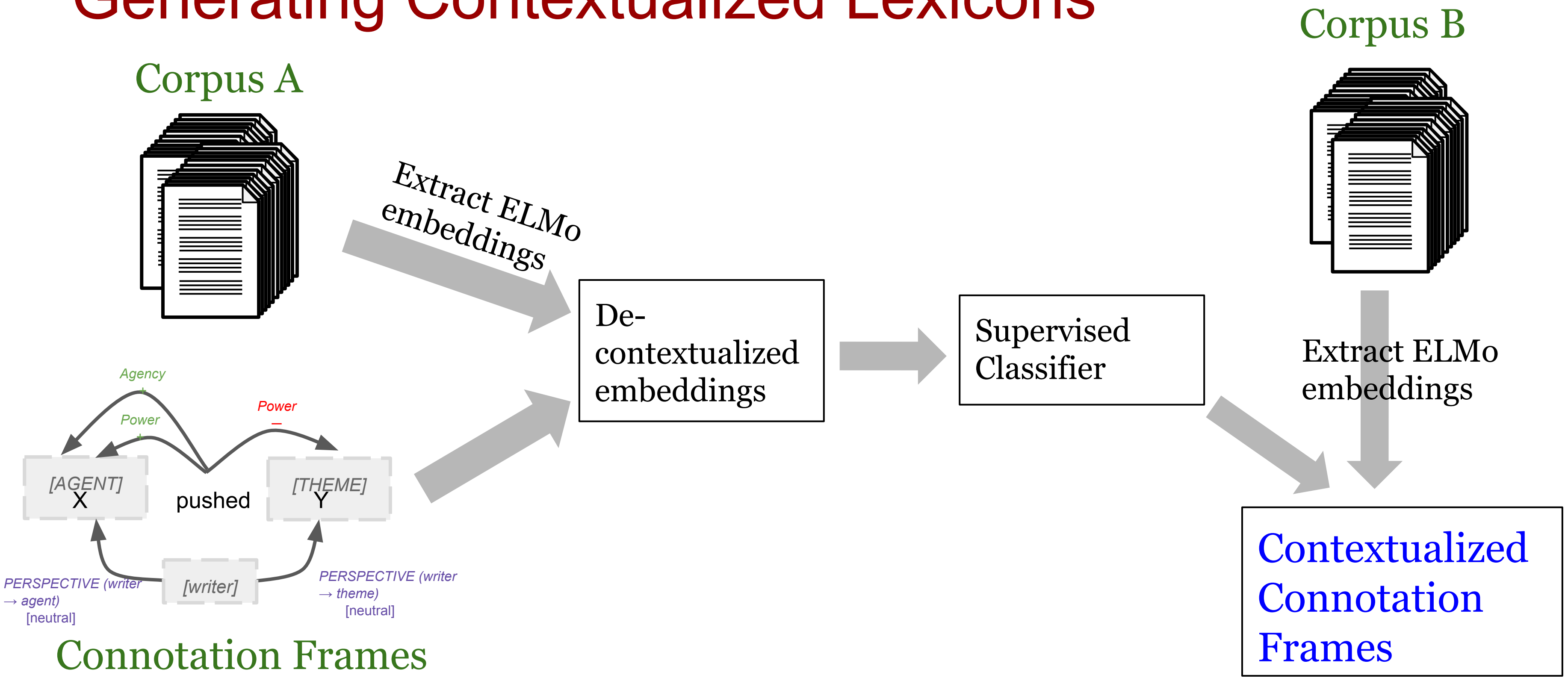
Annotations on **verbs** for various traits from various perspectives

What extensions do we need beyond existing annotations?

- Annotations are over *verbs* but we are interested in *entities*
- How do we handle verbs without annotations?
- Each verb has a single annotation for each dimension, but verbs have different connotations in different contexts

The hero deserves appellation
The boy deserves punishment

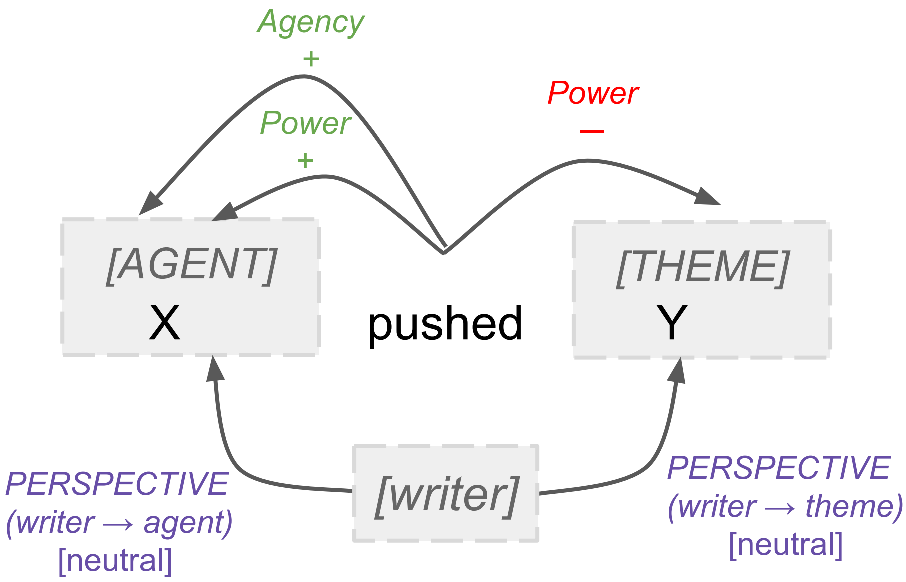
Generating Contextualized Lexicons

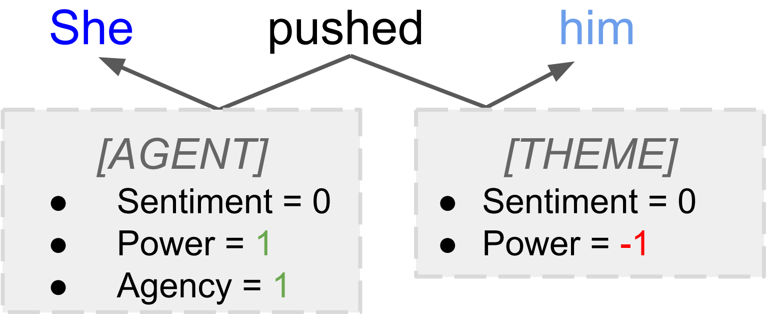


What extensions do we need beyond existing annotations?

- ❑ Annotations are over *verbs* but we are interested in *entities*
- ✓ How do we handle verbs without annotations?
- ✓ Each verb has a single annotation for each dimension, but verbs have different connotations in different contexts

Connotation frames vs Contextual affective analysis





She said that...She pushed him...He pushed back... She tried to...

She

- Sentiment = 0.7
- Power = 0.6
- Agency = 0.2

He

- Sentiment = -0.5
- Power = 0.8
- Agency = 0.7

Evaluation of contextualization

	Verb-level	**Sent.-level**
Sentiment(theme)	41.05	44.35
Sentiment(agent)	51.37	52.80

F1 scores over sentence-level annotations

Evaluation of entity scoring

Off-the-shelf	Frequency	Ours
57.1	59.1	71.4

- Task: Given a pair of entities mentioned in the same set of newspaper articles, choose which entity is more powerful

- Accuracy compared to manual annotations

Analysis of #MeToo Data

Data

A corpus of newspaper articles and blog posts containing the keyword ***#metoo*** using NewsApi

- November 2, 2017 -- May 29, 2018
- Discarded 404 errors, videos, non-English articles and removed duplicates
- **27,602 articles** across **1,576 outlets**
- **3,132,389 entity-verb tuples**

Caveats

- Sample of articles may not be representative
- Bias of researchers may have influenced results
- Our analysis is not intended to have any impact on the individuals described or how they are perceived

Analyses

1. **Corpus-level**
   broad trends in coverage of all common entities across the entire corpus

2. **Role-level**
   how people in similar roles across separate incidents are portrayed

3. **Incident-level**
   analysis of people portrayals involved in a specific incident

Corpus-level analysis: Who are the most powerful, sympathetic, and high agency people?

Most Positive: Kara Swisher, Tarana Burke, Meghan Markle, Frances McDormand, Oprah Winfrey
Most Negative: Bill Cosby, Harvey Weinstein, Eric Schneiderman, Kevin Spacey, Ryan Seacrest, Woody Allen

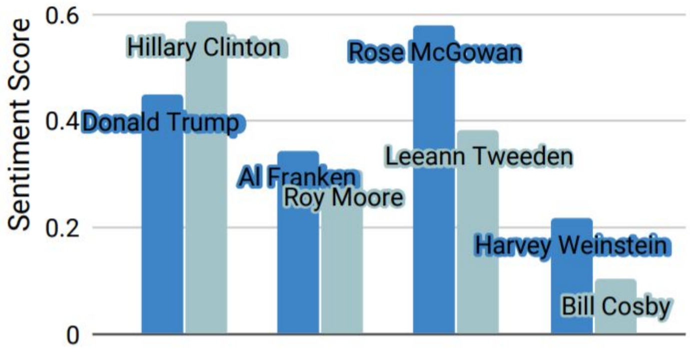
Highest Power: the #MeToo movement, Judge Steven O'Neill, The New York Times, Congress, Facebook, Twitter, Eric Schneiderman, Donald Trump
Lowest Power: Kevin Spacey, Andrea Constand, Leeann Tweeden, Dylan Farrow, Uma Thurman

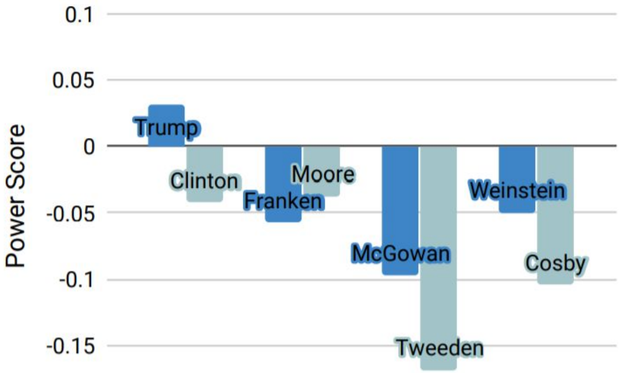
Highest Agency: Judge Steven O'Neill, Eric Schneiderman, Russell Simmons, The New York Times, Frances McDormand, CNN, Donald Trump, Hillary Clinton
Lowest Agency: Kara Swisher, the United States, Hollywood, Meryl Streep

Corpus-level analysis: Who are the most powerful, sympathetic, and high agency people?

- Male accused are portrayed with negative sentiment but with high power
- Female accusers are portrayed among the least powerful entities
- Prominence of 3rd party commenters:
  - Lots of positive sentiment and often high-powered
- Prominence of abstract entities: the #MeToo movement, Congress, Twitter
  - High powered, sometimes high agency

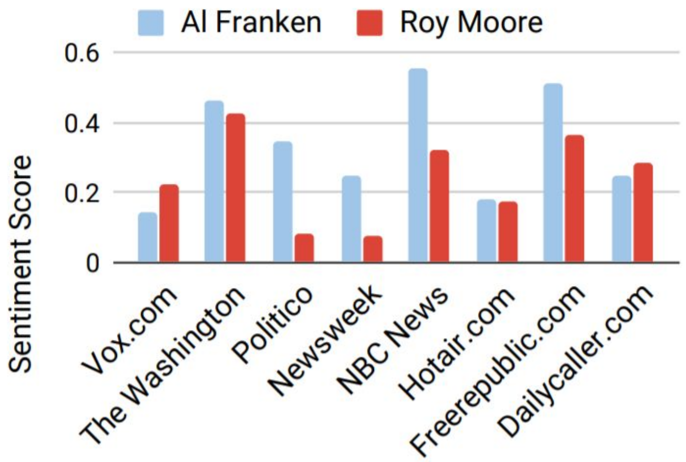
Role-level analysis: how do people in similar roles compare?





- Rose McGowan and Leeann Tweeden are both portrayed with positive sentiment but Rose McGowan has much higher power
- Clinton has more positive sentiment but Trump has higher power
- Politicians Al Franken and Roy Moore have more positive sentiment than Weinstein and Cosby

Cross-outlet comparison: journalistic bias

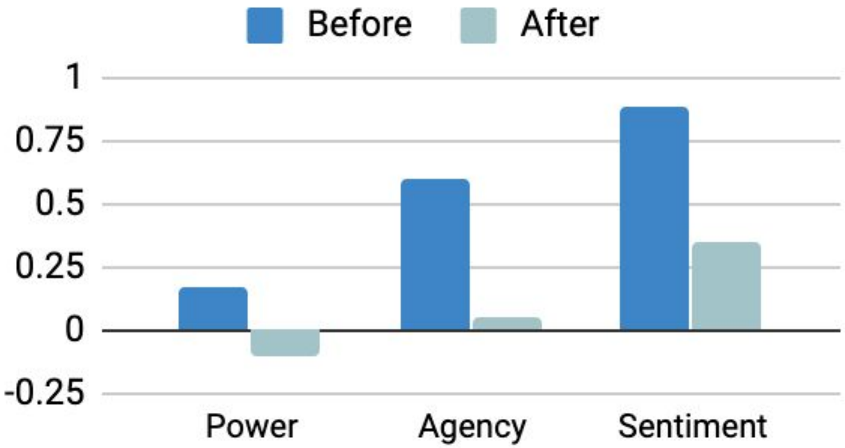


Left-leaning (Democratic): Vox.com, The Washington Post, Newsweek, NBC.
Right-leaning (Republican) outlets: Hotair.com, Freerepublic.com, Dailycaller.com.
Centrist: Politico

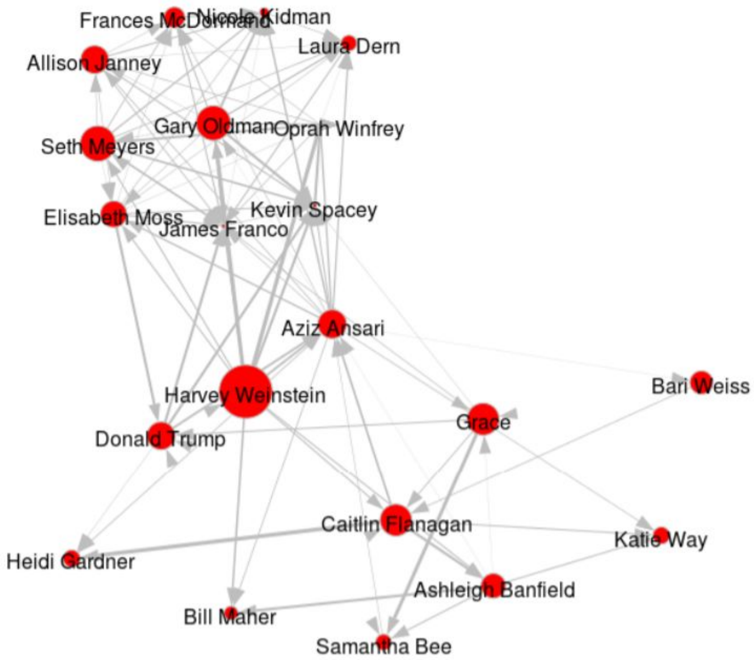
- Al Franken (Democrat) and Roy Moore (Republican) were both politicians accused of sexual misconduct
- Sentiment portrayals does not fall along party lines
- Right-leaning articles present Al Franken as a scapegoat, forced out of office by other Democrats without a fair ethics hearing.

https://dailycaller.com/2018/01/01/railroaded-the-real-reasons-al-franken-is-no-longer-a-senator/

Returning to our motivating example: Ansari

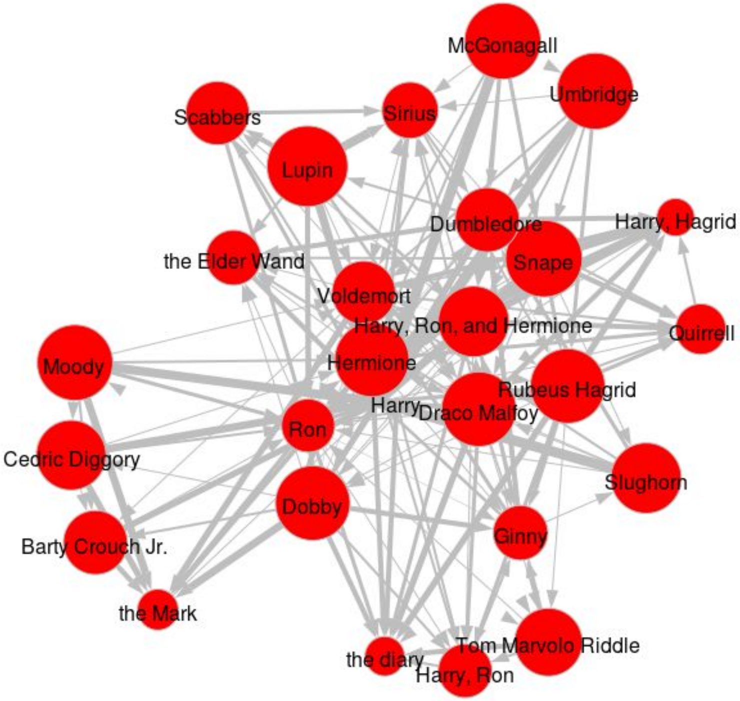


Incident-level: the power landscape surrounding Ansari



- Top left: focused on Golden globes
- Bottom Right: focused on Babe.net articles
- Journalists become powerful entities in the narrative: Caitlin Flanagan, Ashleigh Banfield, Bari Weiss, etc.
- Grace is generally less powerful than Aziz Ansari

Power graph visualization of Wikipedia summaries of Harry Potter



- Voldemort was unable to **kill** him
- working behind the scenes to **kill** Harry
- attempts to seize the stone and **kill** Harry
- attempt to **murder** Harry
- tried to **murder** Harry
- Before Moody can **kill** Harry
- arrives to **kill** Harry
- Horcrux tries to **kill** him
- allow Voldemort to **kill** him

Conclusions

We combine psychology literature and affective control theory with NLP connotation frames to develop contextualized affective analysis

We examine dimensions of power, agency, and sentiment media coverage of the #MeToo movement

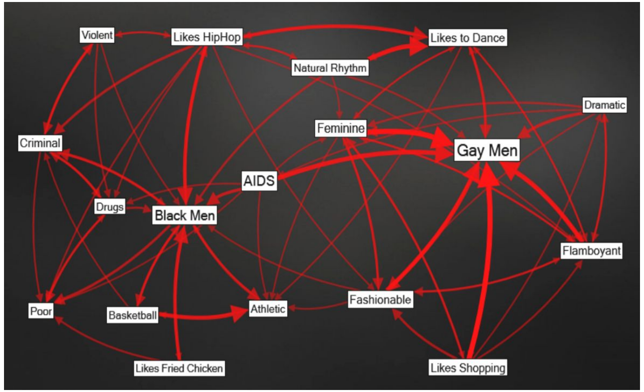
Female accusers are highly sympathetic entities but accused men are portrayed as more powerful

Journalists / other 3rd parties commenting on events become powerful entities in the narrative

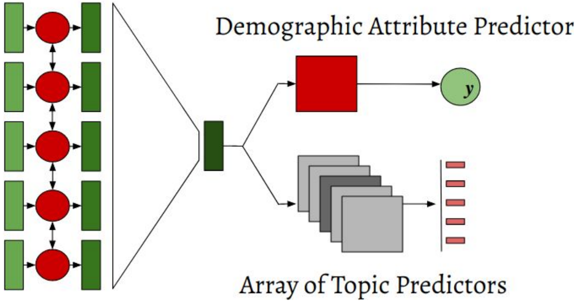
Limitations and future work

- Our analysis is restricted to verbs:
  - What about other parts of speech? Adjectives? Apposition nouns?
  - Syntactic features, quoting patterns, location of mention in the article, etc.
- Power, agency, and sentiment are not binary attributes
- Random sampling of articles may not be entirely representative
- Can we measure impact of articles? How do readers respond to them?
- How can we incorporate the role of social media?
- Cross-lingual analysis and evaluation

How can this research be used practically?

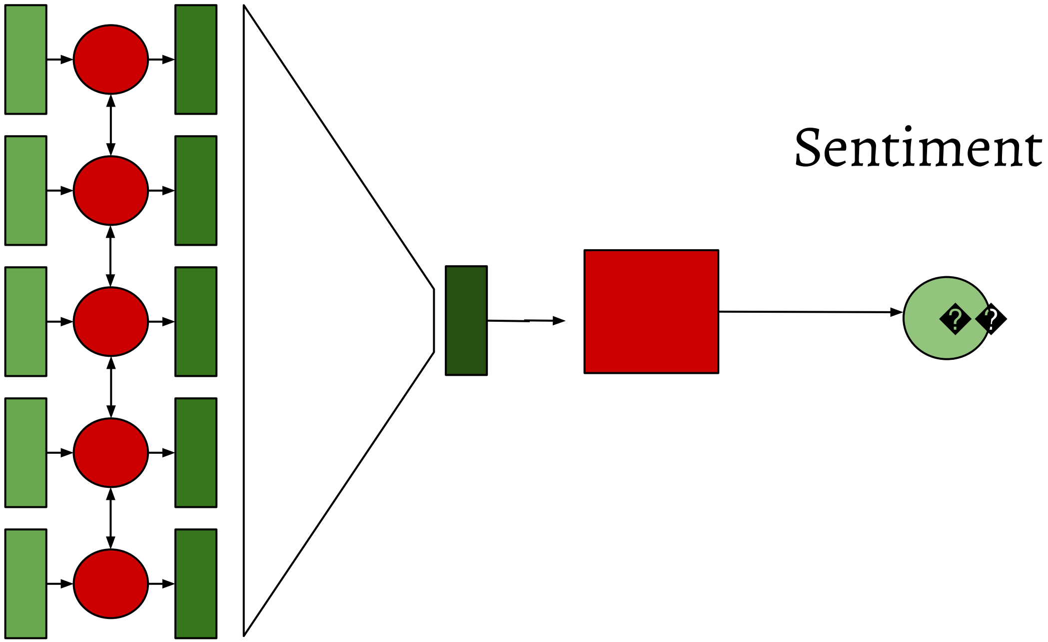


Part 2: implicit biases text classification (ongoing)

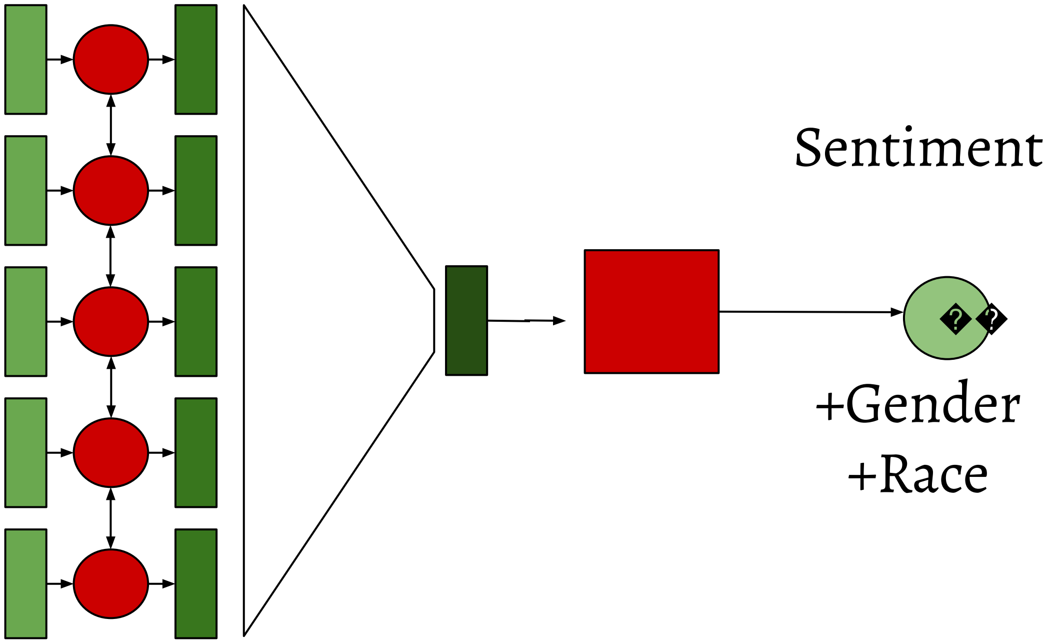


[Kumar et al., ongoing]

Text classification



Bias in text classification

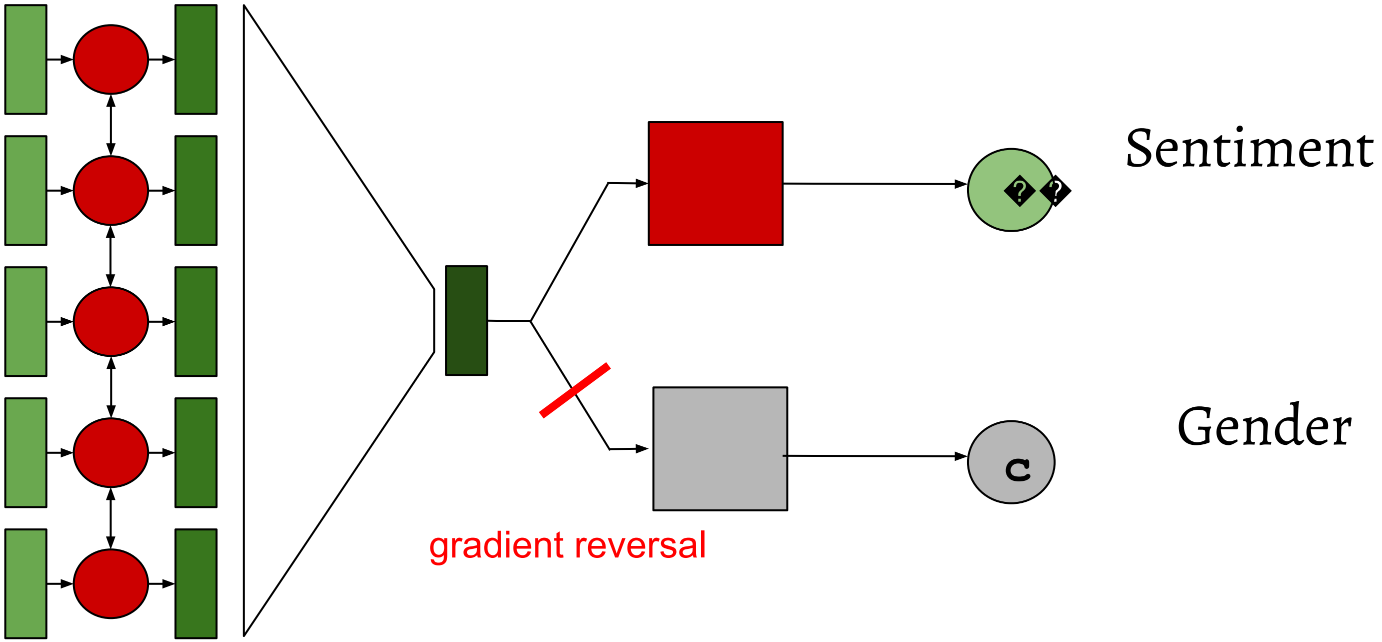


- Bias in Sentiment Analysis (Kiritchenko & Mohammad et al. 2018)

‘The conversation with Amanda was heartbreaking’
‘The conversation with Alonzo was heartbreaking’
‘The conversation with Lakisha was heartbreaking’

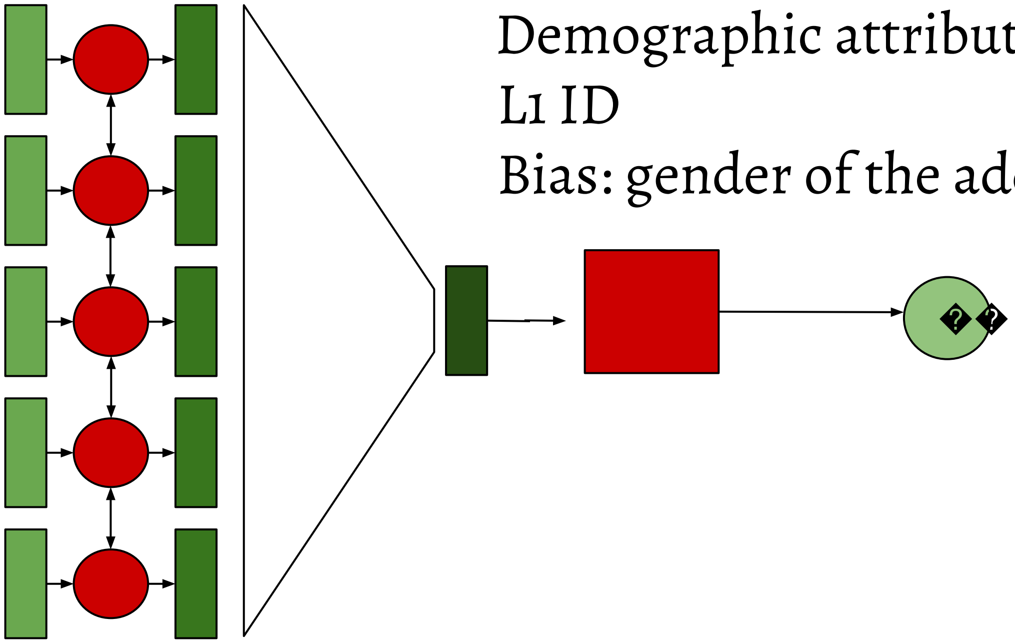
- Bias in Sentiment Analysis (Kiritchenko & Mohammad et al. 2018)

Possible debiasing approach: adversarial multi-task learning



Beutel et al. (2017), Zhang et al. (2018), Pryzant et al. (2018), Elazar & Goldberg (2018)

Making predictions about people or their writing style

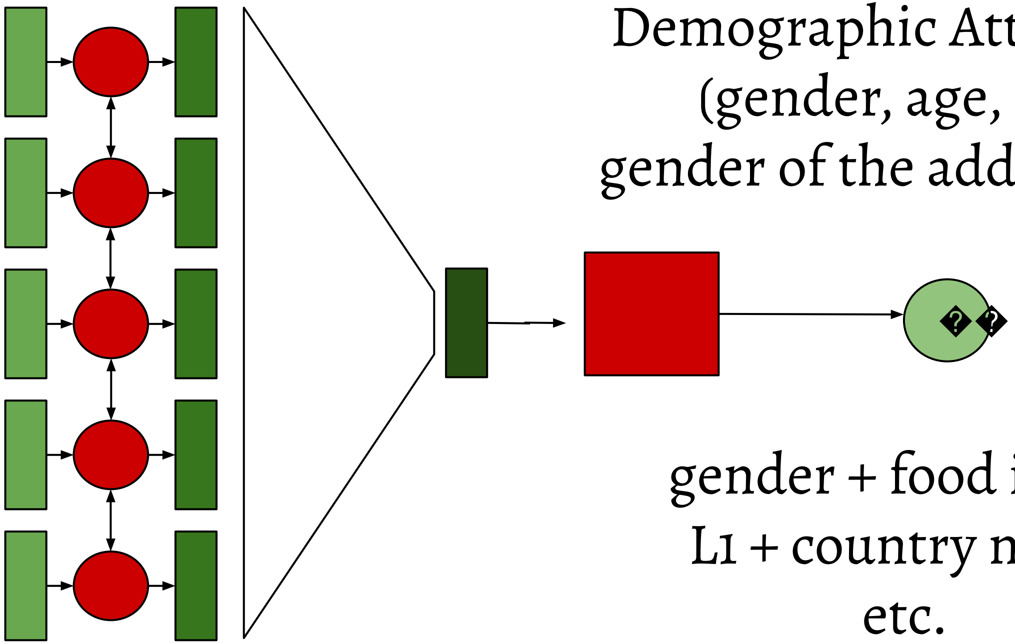


Demographic attributes (gender, age)

L1 ID

Bias: gender of the addressee

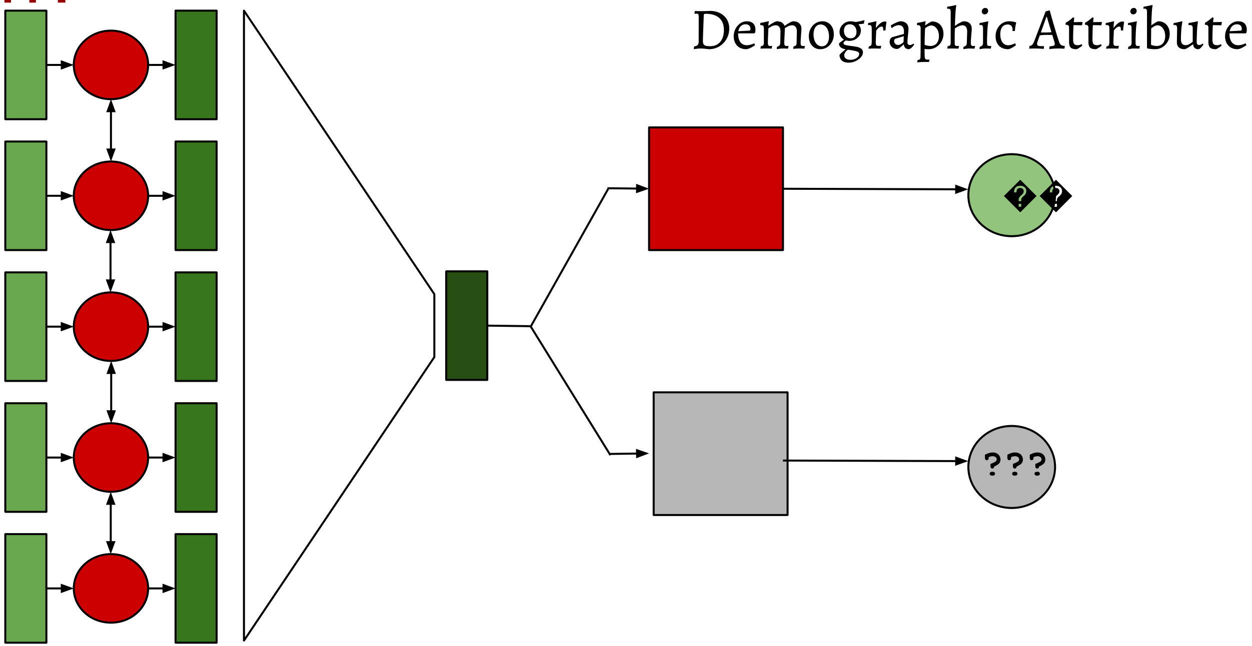
Making predictions about people: multiple confounds



Demographic Attribute
(gender, age, L1,
gender of the addressee)

gender + food items
L1 + country name
etc.

Research question: how to train deconfounded attribute classification?



Salient words in L2 corpora

English	ireland irish british britain russia scotland england states american london brexit
Finnish	finland finnish finns helsinki swedish finn nordic sweden sauna nokia estonian
French	french france paris sarkozy macron fillon hollande gaulle hamon marine valls breton
German	german germany austria merkel refugees asylum germans bavaria austrian berlin also
Greece	greek greece greeks syriza macedonia athens turkey macedonians fyrom turkish ancient
Dutch	dutch netherlands amsterdam wilders rotterdam holland rutte belgium bike hague
Polish	poland polish poles warsaw lithuanian lithuania judges jews ukranians imho tusk
Romanian	romania romanian romanians moldova bucharest hungarian hungarians transistria
Spanish	spain catalan spanish catalonia catalans madrid barcelona independence spaniards
Swedish	sweden swedish swedes stockholm swede malmo danish nordic denmark finland

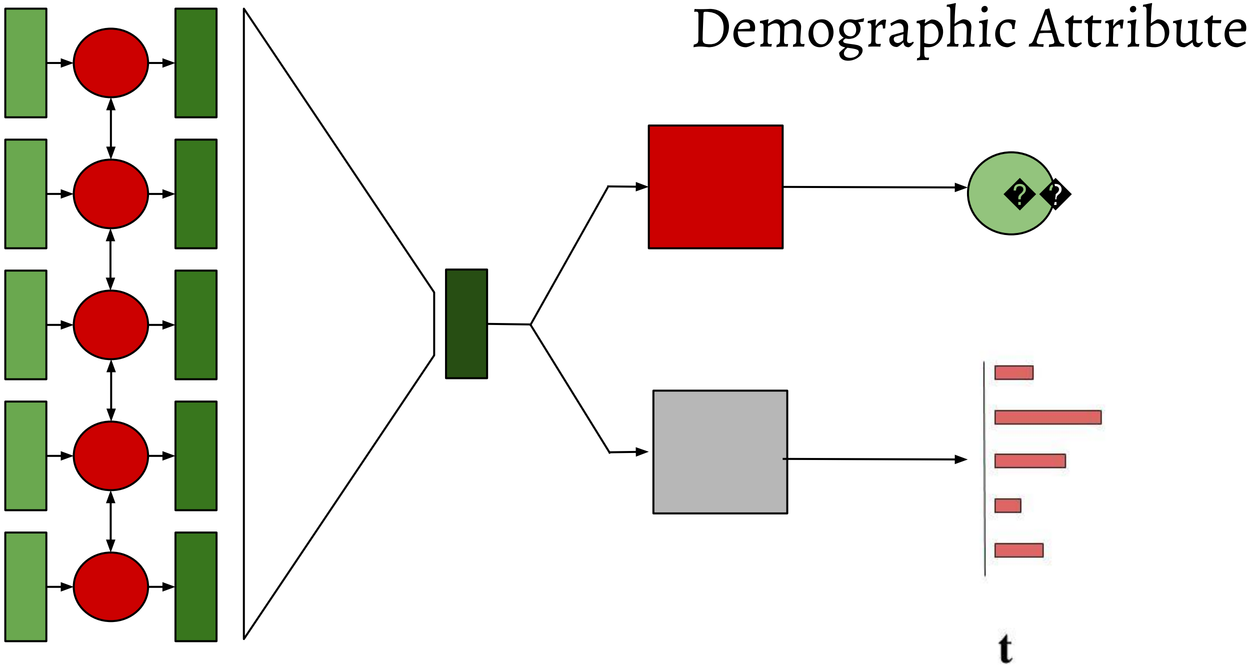
Table 3: Top words based on log-odds scores for each label in the Reddit dataset

Motivation for demoting confounds

	In-Domain	Out-of-Domain
NO-ADV	52.5	25.7
+MASK TOP-20	32.8	21.0
+MASK TOP-50	31.6	20.4
+MASK TOP-100	30.1	19.7
+MASK TOP-200	28.5	18.7

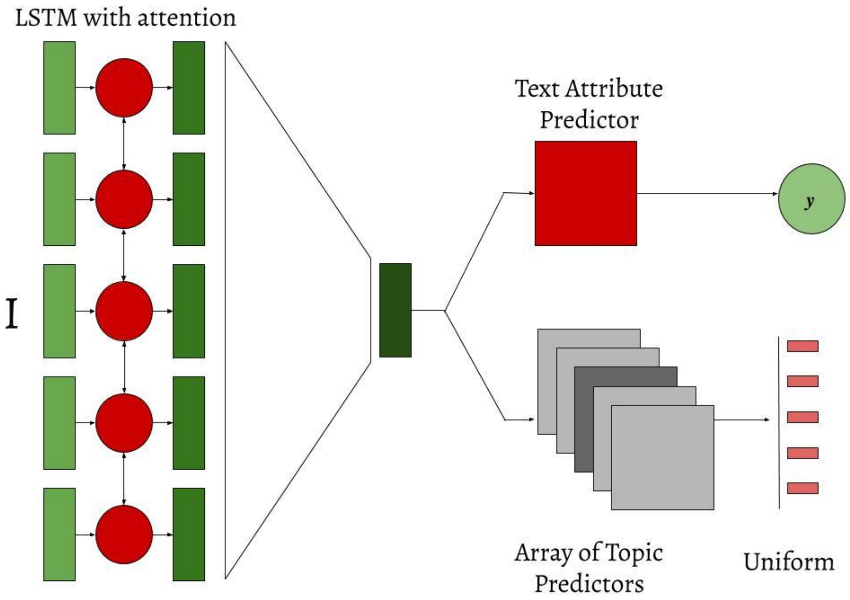
- 10 most frequent L1s in L2-Reddit corpus (Rabinovich et al. 2018)
- In-domain: Europe-related Reddit forums (r/Europe, r/AskEurope, r/EuropeanCulture); Out-of-domain: other forums

Demoting latent confounds

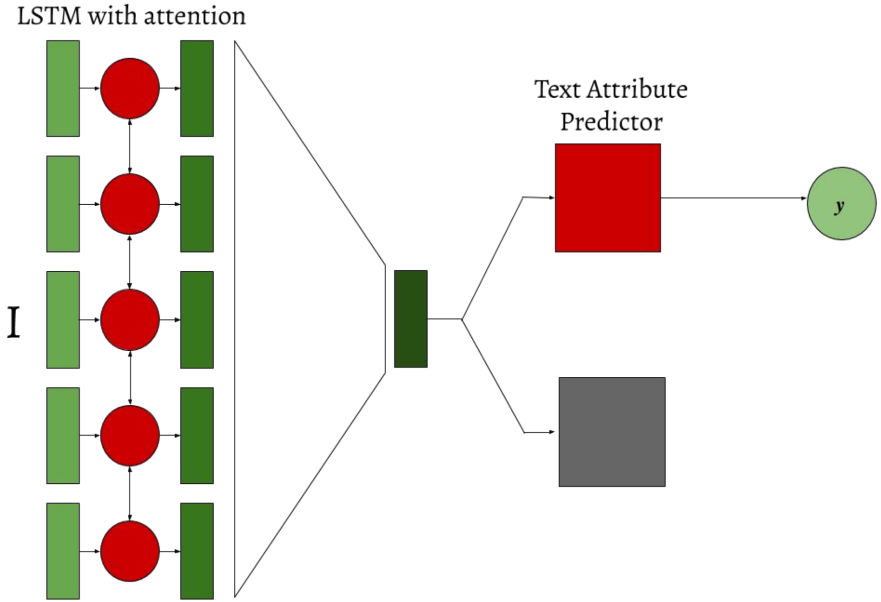


Topics to avoid: Demoting Latent Confounds in Text Classification

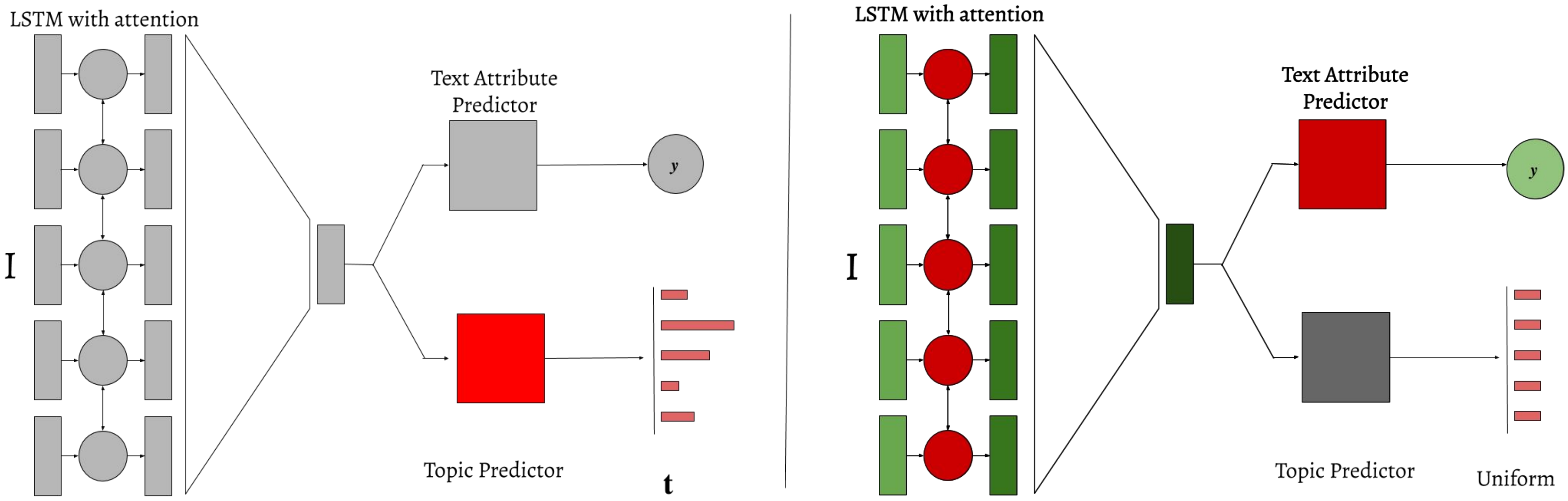
- Instead of gradient reversal -- learning schedule: alternating optimization of classifier and adversary
- adversarial training with multiple adversaries, to alleviate the problem of drifting parameters
- new method of representing and extracting variables which are confounds in text classification



Alternating optimization of classifier and adversary

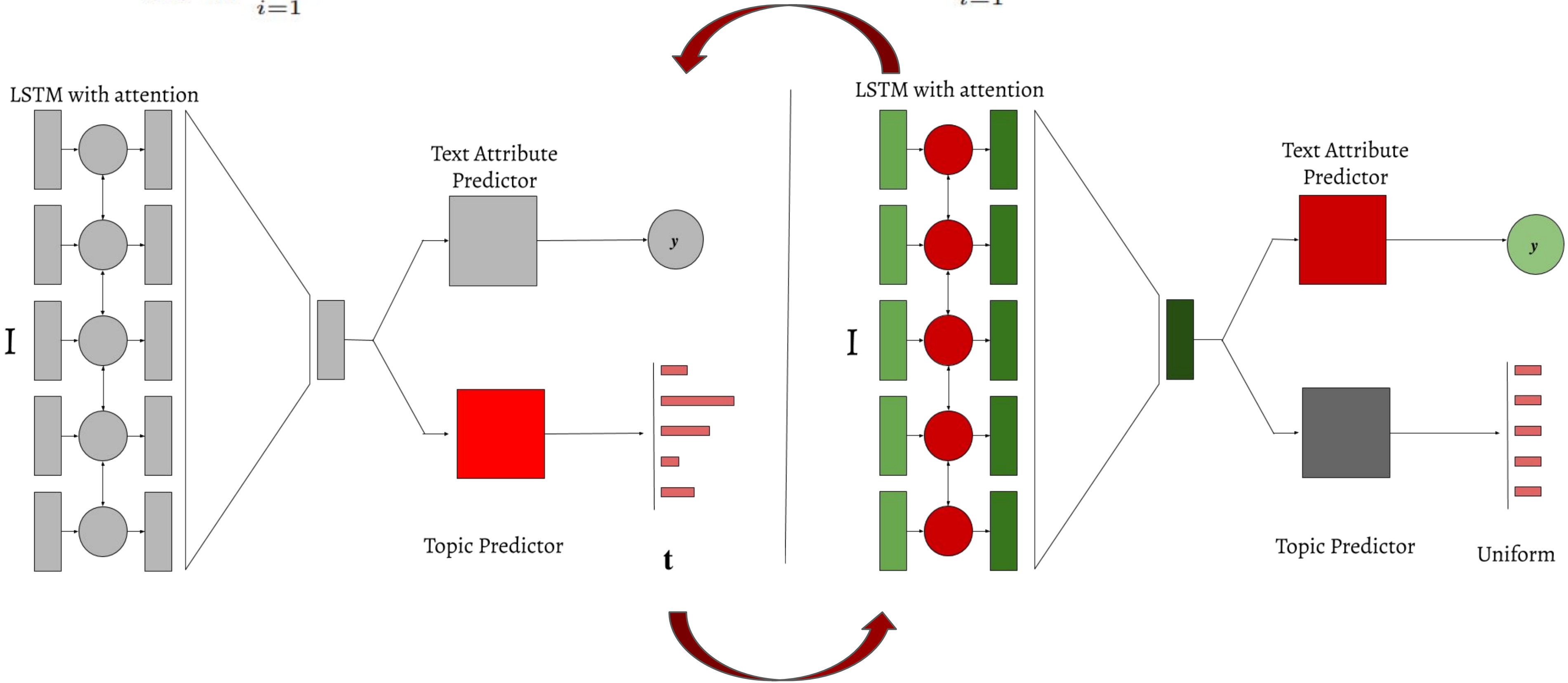


Alternating optimization of classifier and adversary

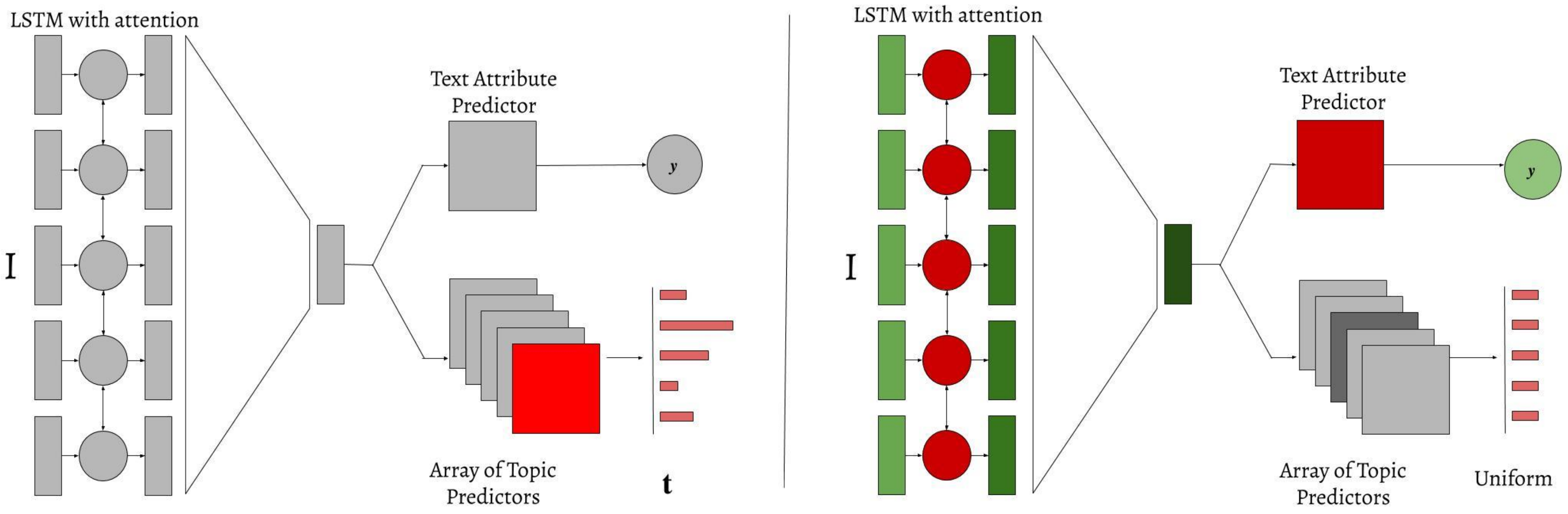


$$\arg\min_{\text{adv}} \frac{1}{N}\sum_{i=1}^{N} \text{CE}(\text{adv}(h(x_i)), t_i);$$

$$\arg\min_{c,h} \frac{1}{N}\sum_{i=1}^{N} \text{CE}(c(h(x)), y) - \text{CE}(\text{adv}(h(x)), t)$$



Training with multiple adversaries



Latent confounds

$$p(y \mid x) \propto p(y)p(x \mid y)$$

$$\propto p(y) \prod_{i=1}^{n} p(w_i \mid y)$$

$$p(w_i \mid y) \propto \sigma(lo(w_i, y))$$

Log-odds ratio with Dirichlet prior

TOEFL

	-P0	-P1	-P2	-P3	-P4	-P5	-P6	-P7
LR	52.8/44.3	54.5/39.4	56.9/53.5	54.4/58.1	57.0/53.7	52.6/49.1	63.9/54.8	50.4/54.5
NO-ADV	62.0/57.0	62.1/**58.0**	**62.1**/54.8	60.4/58.1	62.0/64.4	61.7/50.0	62.4/**63.9**	**63.1**/60.1
ALT-LO	**63.1/63.0**	**62.2**/55.0	59.9/**56.7**	**61.3**/58.1	**62.5/65.2**	**62.0/50.9**	**63.2**/62.0	62.8/**68.5**

Table 7: Accuracy results on the TOEFL dataset

- Baseline features: function words, POS trigrams and sentence length, all of which are reflective of the style of writing (Goldin et al. 2018)
- in-domain/out-of-domain

L2-Reddit (Rabinovich et al. 2018)

	In-Domain	Out-of-Domain
LR	21.2	18.5
LO-TOP-20	38.7	21.9
LO-TOP-50	36.4	21.4
LO-TOP-100	35.8	21.2
LO-TOP-200	34.7	20.8

Table 2: Baseline classification accuracy

	In-Domain	Out-of-Domain
GR-LO	22.5	15.7
ALT-LDA	46.2	21.9
ALT-LO	**48.8**	**22.9**

- Baseline features: function words, POS trigrams and sentence length, all of which are reflective of the style of writing (Goldin et al. 2018)
- Predicting L1 on Reddit is a much harder task due to the high proficiency level of the authors

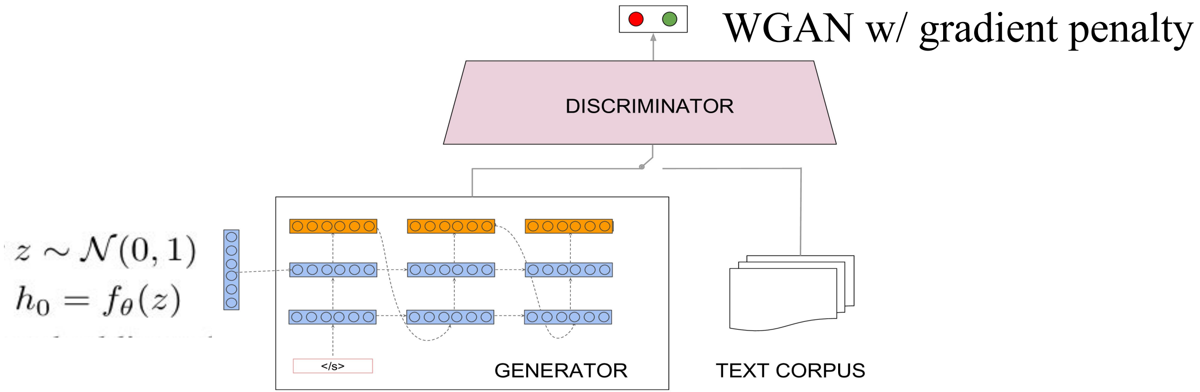
Attentions

NO-ADV	sweden france greece finland poland spain greek germany french eu romania polish dutch german spanish swedish netherlands finnish
LO-TOP-50	eu ’s ’re ’m ’ & uk us because ’ve am its nt english these usa nt here ’ll especially correct pis de within
ALT-LO	the in to of that a i is and ’t as from with by ? on but & they are about at because like was would have you

Table 8: Few highest scoring words in lexicons generated using attention scores

Follow-up work

- Interpreting predictions
- Detection of gender bias
- Controllable generation, style transfer



THANK YOU!

anjalief@cs.cmu.edu sachink@cs.cmu.edu