Social Applications of Pre-trained Language Models

Anjalie Field

Overview

- Defining computational social science
- Methodology:
  - Supervised
    - Classification
  - Unsupervised
    - Topic modeling
  - Pretrained representations
    - Entity representations
    - Metaphor detection
- Recap:
  - Open Challenges

“The study of social phenomena using digitized information and computational and statistical methods” [Wallach 2018]

Social Science

- When and why do senators deviate from party ideologies?
- Analyze the impact of gender and race on the U.S. hiring system
- Examine to what extent recommendations affect shopping patterns vs. other factors

Explanation

NLP

- How many senators will vote for a proposed bill?
- Predict which candidates will be hired based on their resumes
- Recommend related products to Amazon shoppers

Prediction

Hanna Wallach (2018) “Computational Social Science ≠ Computer Science + Social Data”. *Communications of the ACM*

Grimmer and Stewart (2013) Survey of Text as Data

- Classification
  - Hand-coding + supervised methods
  - Dictionary Methods
- Time series / frequency analysis
- Clustering (when classes are unknown)
  - Single-membership (ex. K-means)
  - Mixed membership models (ex. LDA)
- Scaling (Map actors to ideological space)
  - Word scores
  - Word fish (generative approach)

Justin Grimmer and Brandon M. Stewart (2013) “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts” *Political Analysis*

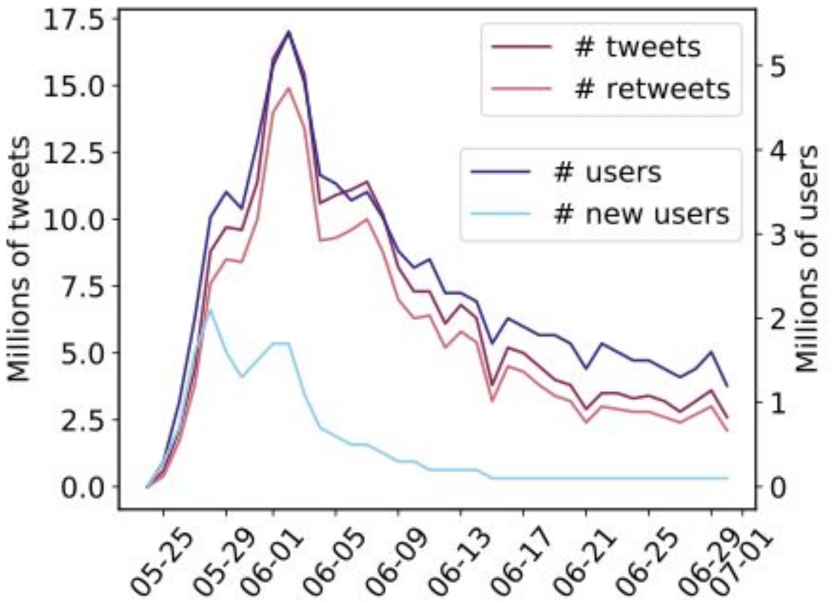
Supervised Classification

- “An Analysis of Emotions and the Prominence of Positivity in #BlackLivesMatter Tweets” Anjalie Field, Chan Young, Park, Antonio Theophilo, Jamelle Watson-Daniels, and Yulia Tsvetkov (PNAS, 2022)

Background: Black Lives Matter movement

The term #BlackLivesMatter originated in posts made by activists Alicia Garza and Patrisse Cullors in 2013

#BlackLivesMatter
#JusticeForGeorgeFloyd
#ICantBreathe



NLP models can facilitate analysis of *emotions*

- "Moral shocks" can cause people to join social movements, but sense of camaraderie, optimism, and hope for change are necessary for sustained involvement
- "Angry Black" stereotypes have lead to tangible harms
    - Media portrayals of protestors as "thugs"

J.M. Jasper (2011) "Emotions and Social Movements: Twenty Years of Theory and Research" *Annual Review of Sociology*
Jeff Goodwin, J.M. Jasper, and Francesca Polletta (2007) "Emotional Dimensions of Social Movements" *The Blackwell Companion to Social Movements*
P.H. Collins (1990) "Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment" *Perspectives on Gender*

Challenges in NLP model development

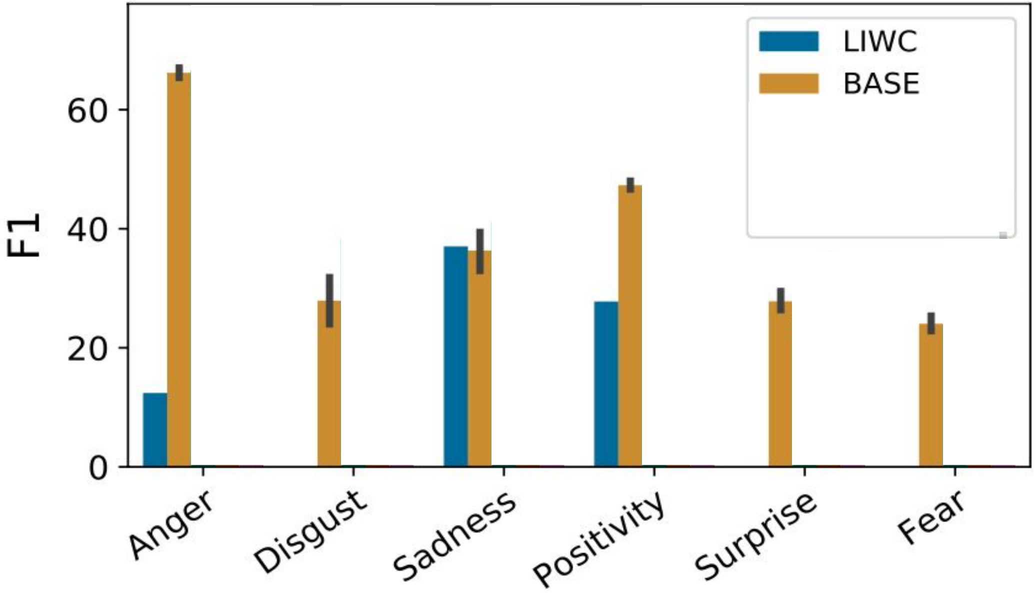
- Emotion taxonomy
  - Ekman's 6 core emotions: anger, disgust, fear, positivity, surprise, sadness
- Annotated Data
  - Existing data sets: GoEmotions and HurricaneEmo
  - New data: 700 BLM tweets annotated according to Ekman's taxonomy
- Domain adaptation
  - Protest movements often raise new ideas in short time spans, e.g. NRC lexicons associate *police* with *trust*

Paul Ekman (1992) "Are there basic emotions?" *Psychological Review*
Demszky et al. (2020) "GoEmotions: A Dataset of Fine-Grained Emotions" *ACL*
Desai et al. (2020) "Detecting Perceived Emotions in Hurricane Disasters" *ACL*

BERT-based classifier

BASE
Large general data set training

Evaluate over annotated BLM tweets

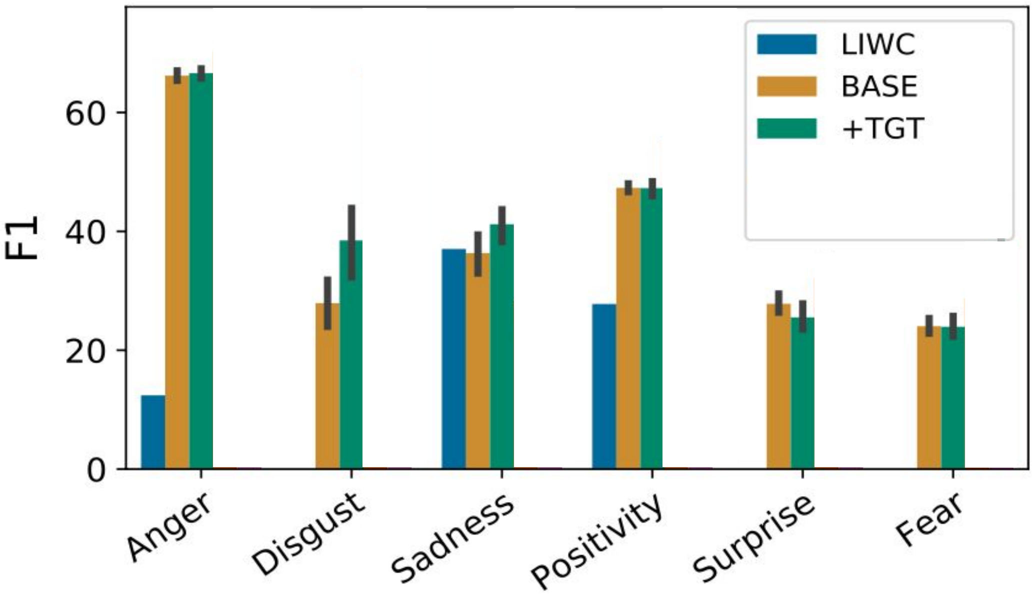


BERT-based classifier

BASE
Large general data set training

Evaluate over annotated BLM tweets

TGT
Masked-language model pre-training



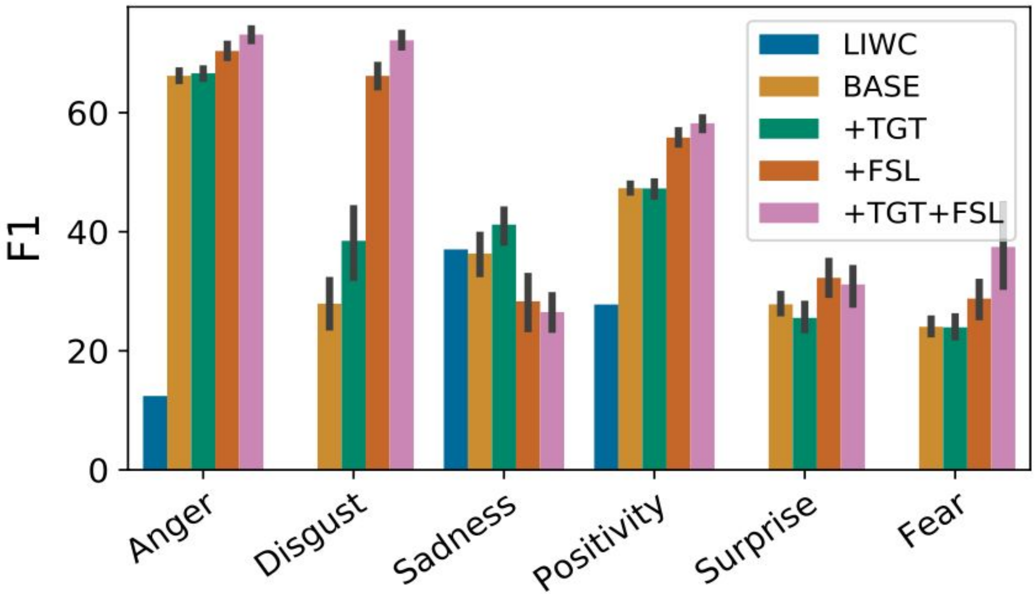
BERT-based classifier

BASE
Large general data set training

Evaluate over annotated BLM tweets

TGT
Masked-language model pre-training

+FSL
Fine-tuning with small newly annotated data

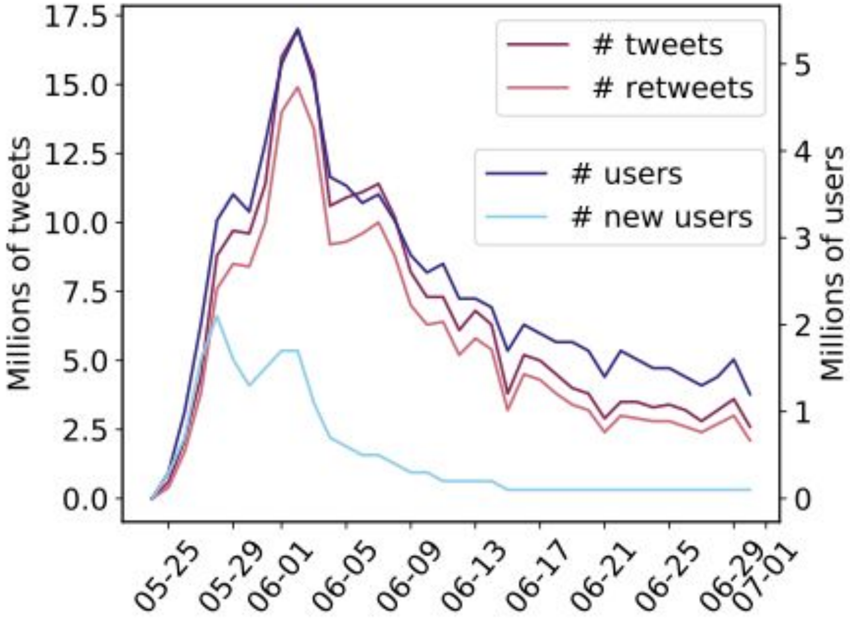


Data: 34M tweets about Black Lives Matter Protests from June 2020

Pro-BLM Hashtags: #BlackLivesMatter, #GeorgeFloyd, #ICantBreathe, #BLM...

Anti-BLM Hashtags: #BlueLivesMatter, #AllLivesMatter...

Police: cops, police,
Protests: protests, protesters, protestors,
Other: george floyd, derek chauvin, protest riot, riots, rioters, looting, looters,



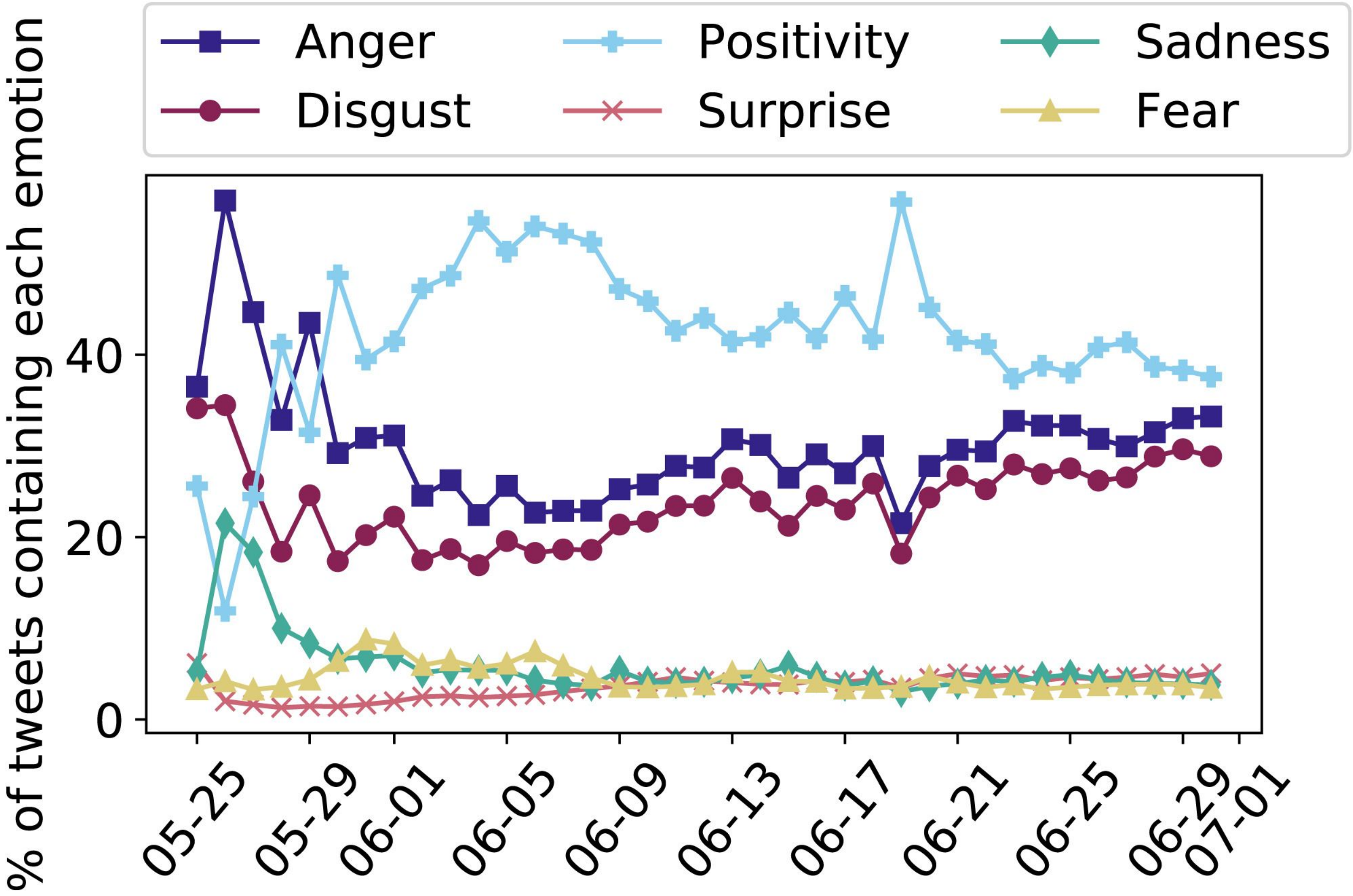
Ethical Considerations and Limitations

- Sample of tweets may not be representative
- Measuring emotions ***perceived in tweets***
  - Cannot draw conclusions about what emotions people actually experienced
- Privacy and consent
  - Not showing any specific examples or usernames from the data

Anger

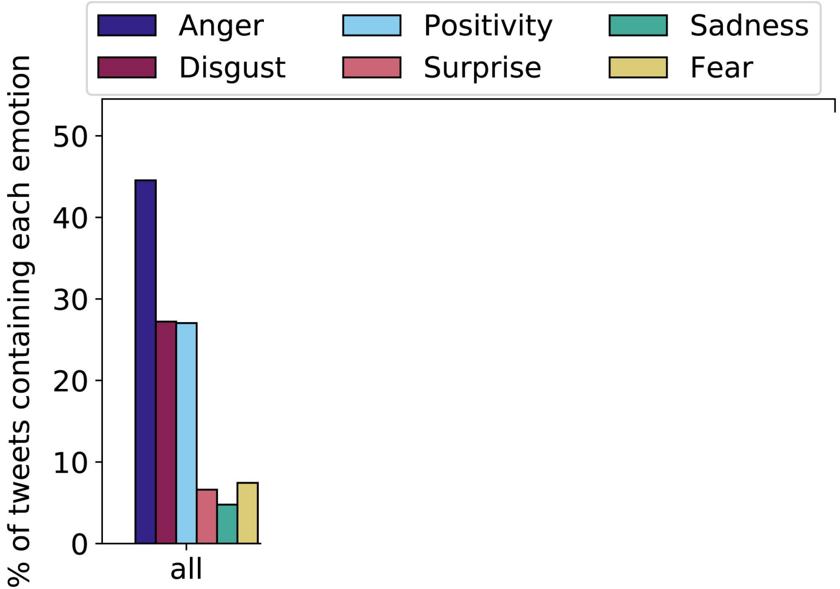
BREONNATAYLOR
BreonnaTaylor
Trump
GeorgeFloyd
DefundThePolice
PoliceBrutality
GeordgeFloydWasMurdered
AntifaTerrorists
Antifa
ACAB
MAGA
FoxNews

Anger	Disgust	Positivity	Surprise	Sadness
BREONNATAYLOR	AllLivesMatter	BlackLivesMatter	BLM	GeorgeFloyd
BreonnaTaylor	Racist	blacklivesmatter	GeorgeFloyd	RIPGeorgeFloyd
Trump	BunkerBoy	Blackouttuesday	AllLivesMatter	JusticeForGeorgeFloyd
GeorgeFloyd	RacistInChief	RAISETHEDEGREE	askingforafriend	RIP
DefundThePolice	BLM	VidasNegrasImportam	DavidDorn	sad
PoliceBrutality	DefundThePolice	love	confused	BlackLivesMatter
GeordgeFloydWasMurdered	FakeNews	BLACK_LIVES_MATTERS	WhiteLivesMatter	JusticeForFloyd
AntifaTerrorists	TrumpResignNow	BlackOutTuesday	AskingForAFriend	ICantBreathe
Antifa	Trump	MatchAMillion	Antifa	RestInPower
ACAB	ACAB	Juneteenth	JustAsking	RIPHumanity
MAGA	ScumMedia	PrideMonth	Blm	rip
FoxNews	MAGA	art	TrumpSupremacist	

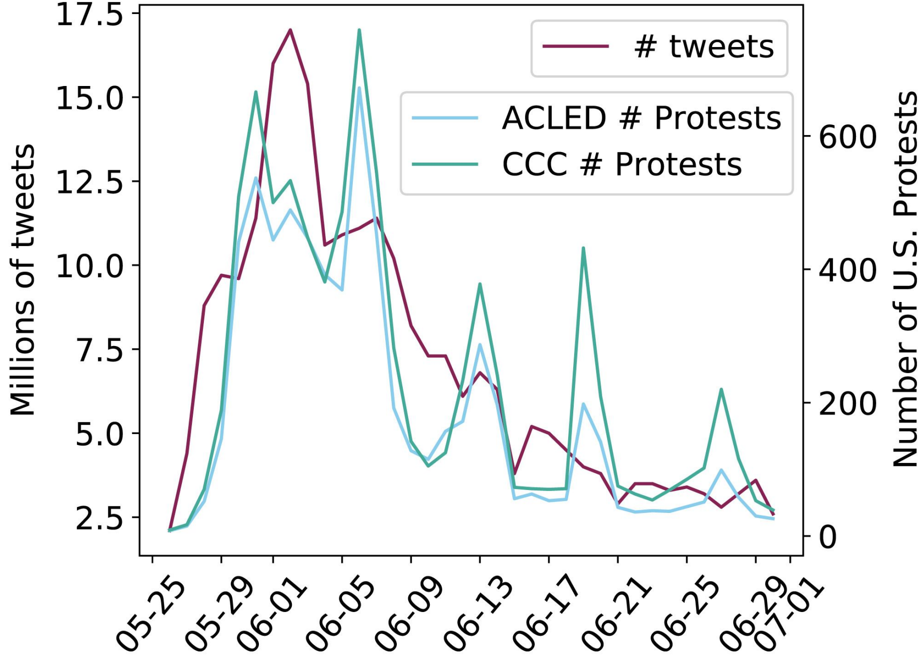


Anger
Positivity
Sadness
Disgust
Surprise
Fear
% of tweets containing each emotion
0
20
40
05-25
05-29
06-01
06-05
06-09
06-13
06-17
06-21
06-25
06-29
07-01

Positivity is more prevalent in tweets with pro-BLM hashtags



Positivity is correlated with in-person protests



	Correlation with protest across states	Correlation with protests across cities
Anger	-0.43*	-0.16*
Disgust	-0.24	-0.21*
Positivity	0.48*	0.12*
Sadness	-0.38*	0.06
Surprise	-0.25	0.09

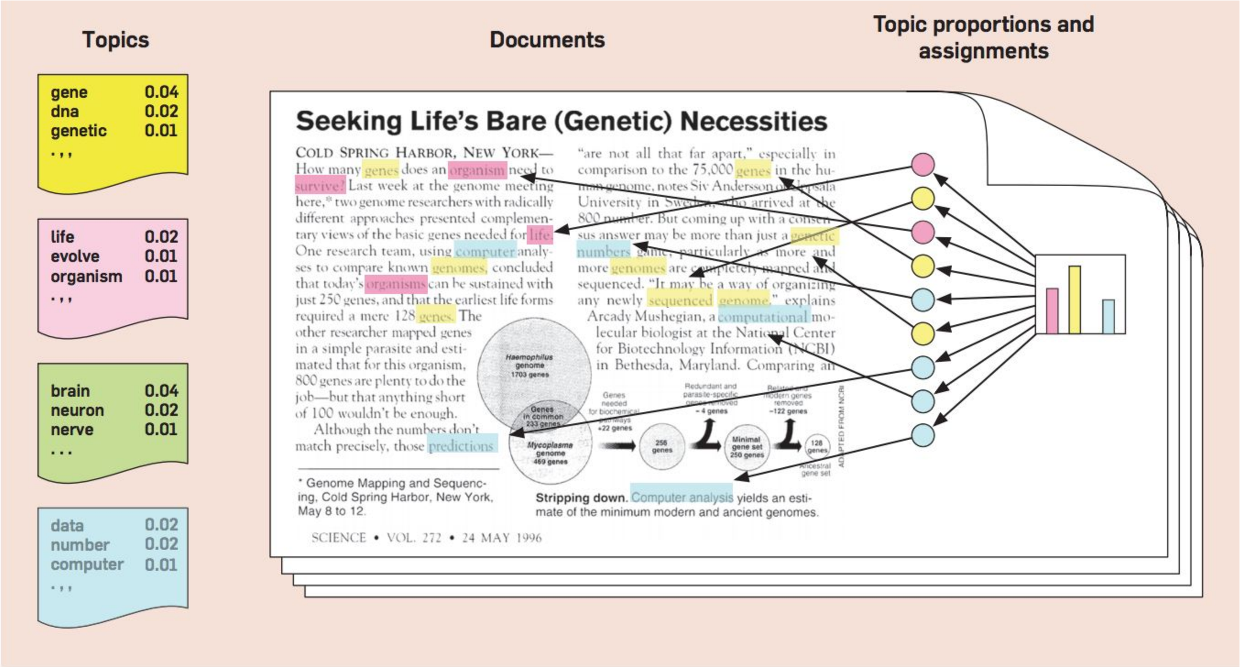
Recap:

- Findings in this example:
    - While stereotypical portrayals of protestors emphasize anger and outrage, our analysis demonstrates that positive emotions like hope and optimism are also prevalent on Twitter
    - Refutes overly-simplified portrayals of people involved in social movements and discourage stereotyping
- Can use pre-trained language model as improved classifier
    - Pre-training objective facilitates domain adaptation
    - Still need in-domain annotations to improve performance and evaluate

Unsupervised Clustering

- “Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence” Federico Bianchi, Silvia Terragni, Dirk Hovy (ACL, 2021)
- “Challenges in Opinion Manipulation Detection: An Examination of Wartime Russian Media” Chan Young Park, Julia Mendelsohn, Anjalie Field, Yulia Tsvetkov (Findings of EMNLP, 2022)

Quick Overview of “Topic Modeling”



- Assume each document contains a mixture of “topics”
- Each topic uses mixtures of vocabulary words
- Goal: recover topic and vocabulary distributions

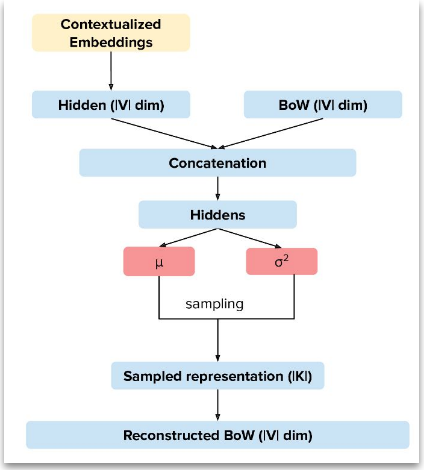
Clustering: Contextualized Topic Models

for *each document* **w** **do**
  Draw topic distribution $\theta \sim \text{Dirichlet}(\alpha)$;
  for *each word at position* n **do**
    Sample topic $z_n \sim \text{Multinomial}(1, \theta)$;
    Sample word $w_n \sim \text{Multinomial}(1, \beta_{z_n})$;
  end
end

Latent Dirichlet Allocation (LDA)
2003

AEVB inference algorithm to slightly generalized/ modified LDA

ProdLDA
2017



Contextualized Topic Models
2021

David M. Blei, Andrew Y. Ng, and Michael I. Jordan (2003) "Latent dirichlet allocation" *JMLR*
Akash Srivastava and Charles Sutton (2017) "Autoencoding variational inference for topic models" *ICLR*
Federico Bianchi, Silvia Terragni, Dirk Hovy (2021) "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence" *ACL*

Control areas as of July 22

Sources: Institute for the Study of War, AEI's Critical Threats Project, Post reporting


DON'T
LISTEN TO
RUSSIAN
PROPAGANDA

Example: Contextualized Topic Models in Social Media Posts about Russia-Ukraine War

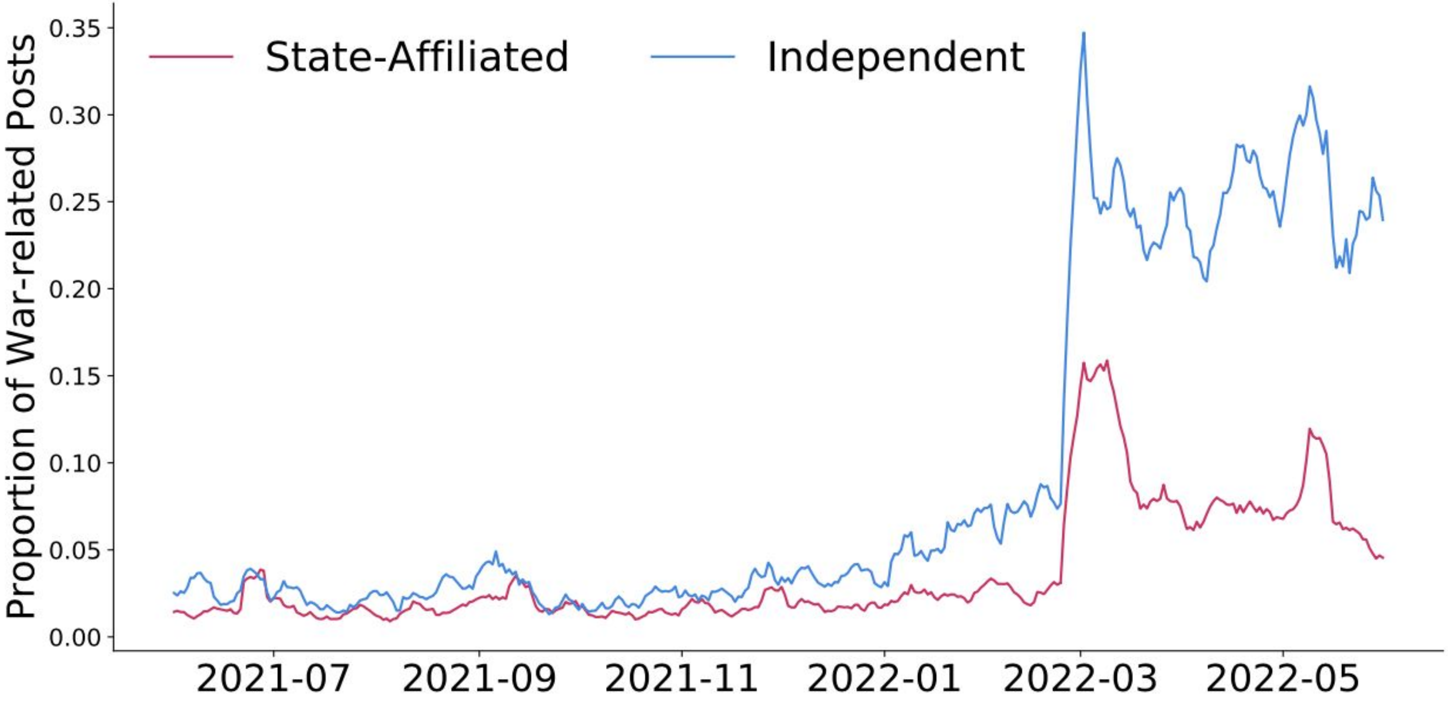
- Emerging social media data set
  - Don’t have in-domain annotated data
  - Open-ended exploratory questions

Dataset Collection

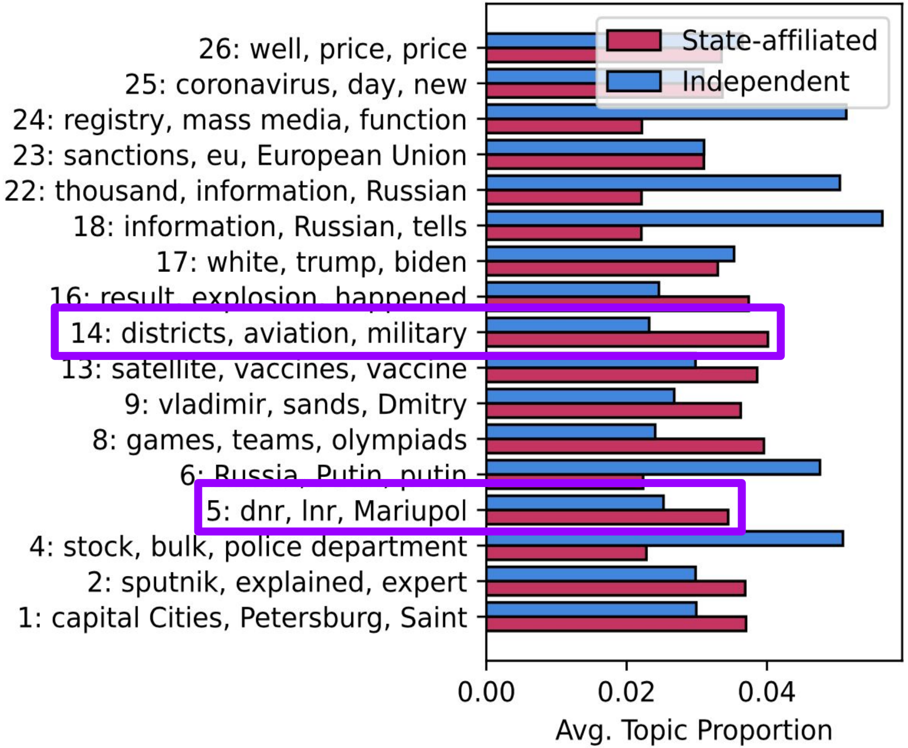
- Jan 01 2021 ~ Present (analysis ends May 31 2022)
- Three dimensions
  - **Time**: pre-war, during-war
  - **Platform**: Twitter, VKontakte (VK)
  - **Media ownership**: state-affiliated, independent

23 State-affiliated outlets		**20 Independent outlets**	
RT_com	rbc	tvrain	snob_project
life	ria	Forbes	golosameriki
tassagency	gazeta	novgaz	svobodaradio
tv5	vesti	meduzaproject	BBC
rgru	Ukraina RU	rtvi	The insiders

Example: Contextualized Topic Models in Tweets about Russo-Ukraine War



Example: Contextualized Topic Models in Tweets about Russo-Ukraine War



- CTM suggests war-related topics are more common in state-affiliated outlets
- Pro: didn't need to do any data annotations, able to run quickly
- Con: Not sure if we're measuring the right thing

Embedding Projections / Ideology Mapping

- “Entity-Centric Contextual Affective Analysis” Anjalie Field and Yulia Tsvetkov (ACL, 2019)
- “Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration” Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky (PNAS, 2022)

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

nlp debiasing word embeddings

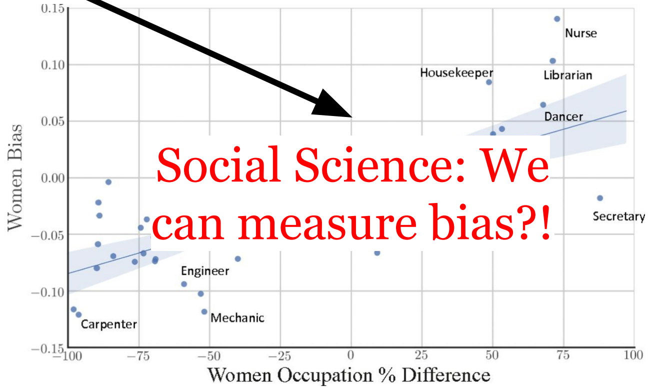
About 3,280 results (**0.11** sec)

Man is to computer programmer as woman is to homemaker? **debiasing word embeddings**
T Bolukbasi, KW Chang, JY Zou… - Advances in neural … 2016 - proceedings.neurips.cc
… and **natural language processing** tasks. We sh… at even **word embeddings** trained on …
is first shown to be captured by a direction in the **word embedding**. Second, gender neutral …
☆ Save 99 Cite

Gender-pre…
M Kaneko, D …
… **word embe**… **rd embeddings**
… information for downstream **NLP** tasks that use those **debiased word embeddings**. To …
☆ Save 99 Cite Cited by 82 Related articles All 5 versions

Lipstick on a pig: **Debiasing** methods cover up systematic gender biases in **word embeddings** but do not remove them
H Gonen, Y Goldberg - arXiv preprint arXiv:1903.03862, 2019 - arxiv.org
… **Word embeddings** are widely used in **NLP** for a vast range of … For each **debiased word** embedding we quantify the hidden bias … For HARD-**DEBIASED** we compare to the **embeddings** …
☆ Save 99 Cite Cited by 400 Related articles All 10 versions

NLP: Oh no! My models are biased!



Social Science: We can measure bias?!

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou (2018) “Word embeddings quantify 100 years of gender and ethnic stereotypes”

Entity Representations: Power, Agency, and Sentiment in News

"Entity-Centric Contextual Affective Analysis" Anjalie Field and Yulia Tsvetkov (ACL, 2019)

Goal: Examine how people are described in terms of power, agency, and sentiment in narrative text

Example: Do news articles portray women as less powerful than men?

Annotated Lexicons

- “Computer Programmer” and “homemaker” come from lists of occupational stereotypes

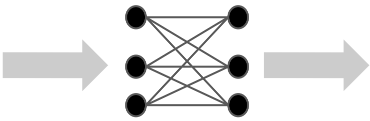
- Lexicons annotated for power, agency and sentiment

	Low	High
Power	timid	resourceful
	weakly	powerfully
	cowardly	courageous
	inferior	superior
	clumsy	skillful
Sentiment	negative	positive
	pessimistic	optimistic
	annoyed	amused
	pessimism	optimism
	disappointed	pleased
Agency	silently	furiously
	meek	lusty
	homely	sexy
	bored	flustered
	quietly	frantically

Methodology

Extract embeddings for words in the lexicon:

"The **king** won the war"

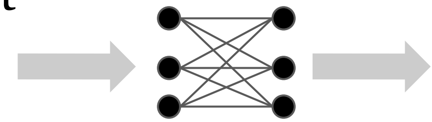


0.9
5.6
9.3
…

Extract embeddings for entities we want to measure:

"**Hillary Clinton** lost the 2016 election"

"**Donald Trump** won the 2016 election"



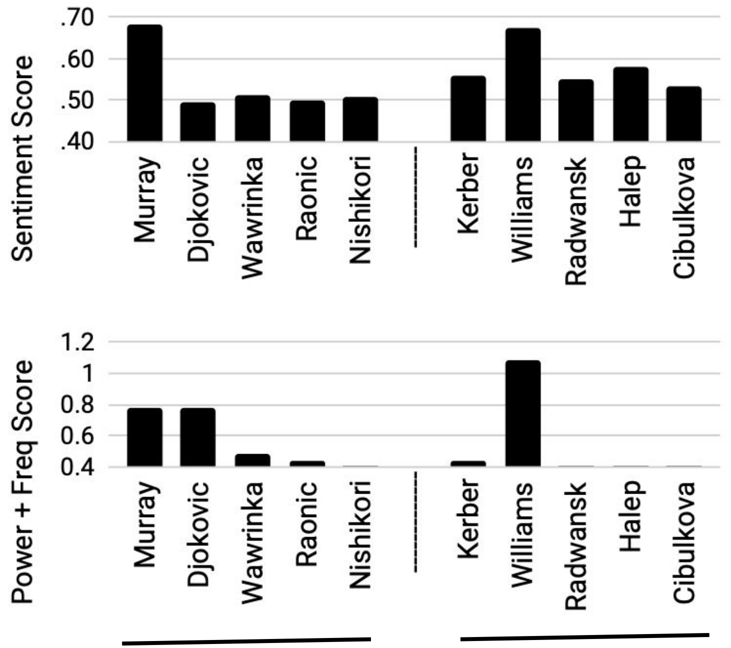
0.9
5.6
9.3
…

0.4
5.4
3.8
…

"Regression"
Directly train supervised classifier, using embeddings as features and lexicon annotations as labels

"ASP" Use lexicons to identify "power", "agency", and "sentiment" subspaces and project entity embeddings

Sample Results



Men Tennis Players

Women Tennis Players

Women tennis players tend to be portrayed with more **positive sentiment**

Men tennis players tend to be portrayed with more **power**

Problem: Can't distinguish model training data from corpora

Full annotation set (383 pairs)

	Regression	ASP
ELMo	44.9	43.6
BERT	41.8	49.3
BERT-masked	49.6	**59.0**
Frequency Baseline	58.0	

Reduced annotation set (49 pairs)

	Regression	ASP
ELMo	36.7	42.8
BERT	42.9	49.0
BERT-masked	53.1	55.1
Frequency Baseline	57.1	
Field et al. (2019)	**71.4**	

Over evaluation data set that inverts traditional power roles (#MeToo movement), method performs poorly

Pre-trained models have strong token signal from pre-training data:
"**Hillary Clinton** lost the 2016 election"
--captures how "Hillary Clinton" was depicted in pre-training data, not just this sentence

Problem: Can't distinguish model training data from corpora

Full annotation set (383 pairs)

	Regression	ASP
ELMo	44.9	43.6
BERT	41.8	49.3
BERT-masked	49.6	**59.0**
Frequency Baseline	58.0	

Reduced annotation set (49 pairs)

	Regression	ASP
ELMo	36.7	42.8
BERT	42.9	49.0
BERT-masked	53.1	55.1
Frequency Baseline	57.1	
Field et al. (2019)	**71.4**	

"[MASK] lost the 2016 election"
Masking helps a little

also degrades embedding quality

Regression

	Power	Sentiment	Agency
ELMo	0.78	**0.84**	0.76
BERT	**0.79**	0.83	**0.78**
BERT-masked	0.64	0.70	0.62

ASP

	Power	Sentiment	Agency
ELMo	0.65	0.76	0.63
BERT	0.65	0.71	0.66
BERT-masked	0.41	0.47	0.41

Problem: Can't distinguish model training data from corpora

Still maybe useful when:

- We don't care about results specific to a domain (how are people depicted in model representations / general large corpora?)
- We are looking at comparative questions: is X portrayed different over time?

Metaphor Detection

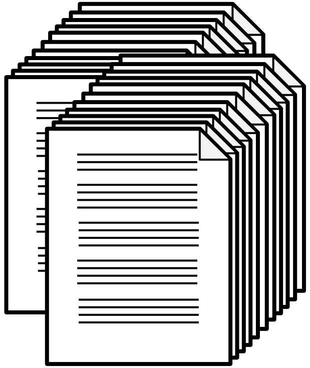
“Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration” Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky (PNAS, 2022)

Goal: measure implicit dehumanizing metaphors long associated with immigration (animals, cargo, etc.)

- “Animal” metaphor: “the herding of these [. . .] into stockades is pictured.”
- “Cargo” metaphor: “I voted last week for an antidumping bill to prevent the dumping of manufactured products into this country, and I will vote for any bill to prevent the dumping of undesirable [...] into this country.”

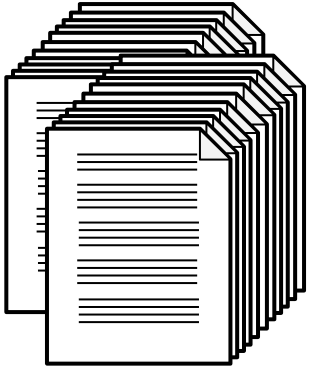
Metaphor Detection: Methodology

Analysis corpora: US political speeches



"The tendency of ***displaced persons*** to flock together"

BERT training corpora



"The tendency of ***birds*** to flock together"

Key idea: If "displaced persons", "immigrants", "Germans", etc. are used in similar sentences in the analysis corpora as words like "animals" and "cargo" in pretraining corpora, this implies dehumanizing metaphors

Metaphor Detection: Methodology

Identify sentences in analysis corpora where immigrants are mentioned (“immigrants”, “displaced persons”, “Germans”, etc.)

“The tendency of *displaced persons* to flock together”

Mask out mentions, and compute BERT output scores for words from common dehumanization metaphors

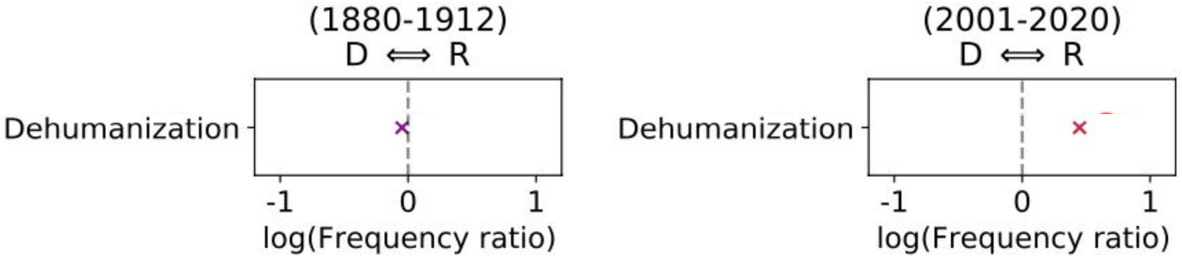
“The tendency of [MASK] to flock together”

“Animals” (0.6)
“birds” (0.9)
“Sheep” (0.6)
“Dogs” (0.3)
…

“cargo” (0.2)
“shipment” (0.2)
…

Metaphor Detection: Findings

In modern political speeches, Republicans use more dehumanizing language about immigration than Democrats



Recap

- Supervised classification
  - Emotions in tweets about Black Lives Matter
- Unsupervised topic modeling
  - “Operation” vs. “War” in state-affiliated vs. independent Russian media outlets
- Embedding projections / ideology mapping
  - Entity representations for analyzing corpora of narrative text
    - Measureing power, agency, and sentiment in news
  - Metaphor detection
    - Dehumanization about immigrants in political speeches

Open Challenges

- Supervised classification
- Unsupervised topic modeling
- Embedding projections / ideology mapping

- Annotation data is slow and difficult
- Need to disentangle pre-training data from analysis data
- Difficult to interpret / determine what exactly we're measuring
- Models are not user-friendly and require lots of compute

End