SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

IC²S²

A framework for the computational linguistic analysis of dehumanization

Julia Mendelsohn
juliamendelsohn.github.io
@jmendelsohn2
juliame@umich.edu

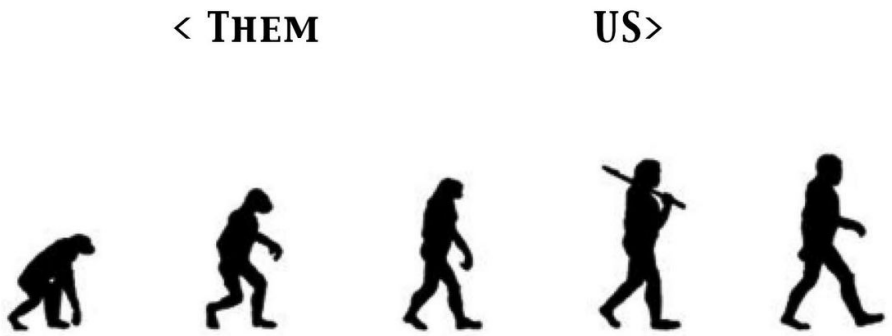
Dan Jurafsky
stanford.edu/~jurafsky
@jurafsky
jurafsky@stanford.edu

Yulia Tsvetkov
cs.cmu.edu/~ytsvetko
ytsvetko@cs.cmu.edu

(https://cdn.psychologytoday.com)

Dehumanization

The act of perceiving or treating people as less than human [Haslam & Stratemeyer, 2016]



Leads to extreme intergroup bias, hate speech, violence

This talk

- We identify **linguistic analogs** for several dimensions of dehumanization and propose **computational techniques** to measure these linguistic correlates.
- **Case Study**: changing representations of LGBTQ groups in the New York Times over three decades.
- Through this lens, we investigate differences in social meaning between seemingly similar group labels.

Components of dehumanization

1. Negative evaluations of the target group
2. Moral disgust
3. Associations with non-humans (especially vermin)
4. Denial of agency
5. Psychological distance
6. Essentialism
7. Denial of subjectivity

Components of dehumanization

1. Negative evaluations of the target group
2. Moral disgust
3. Associations with non-humans (especially vermin)

We operationalize these three components by identifying and measuring **lexical semantic** analogs.

Components of dehumanization

1. Negative evaluations of the target group

Attribution of negative characteristics to target group categorizes groups that are “excluded from the realm of acceptable norms and values” [Bar-Tal, 1990]

Components of dehumanization

2. Moral Disgust

Disgust ➔ perception of target group's negative social value [Sherman & Haidt, 2011]

Moral disgust "facilitates moral exclusion of out-groups" [Buckels & Trapnell, 2013]

Components of dehumanization

3. Associations with non-humans (especially vermin)

   Vermin metaphor conceptualizes the target group as “engaged in threatening behavior, but devoid of thought or emotional desire” [Tipler & Ruscher, 2014]

Quantifying *negative evaluations* (1)

Valence: aspect of meaning ranging from negative emotion (unpleasant) to positive (pleasant)

NRC VAD lexicon: valence scores from 0 to 1 for 20k English words

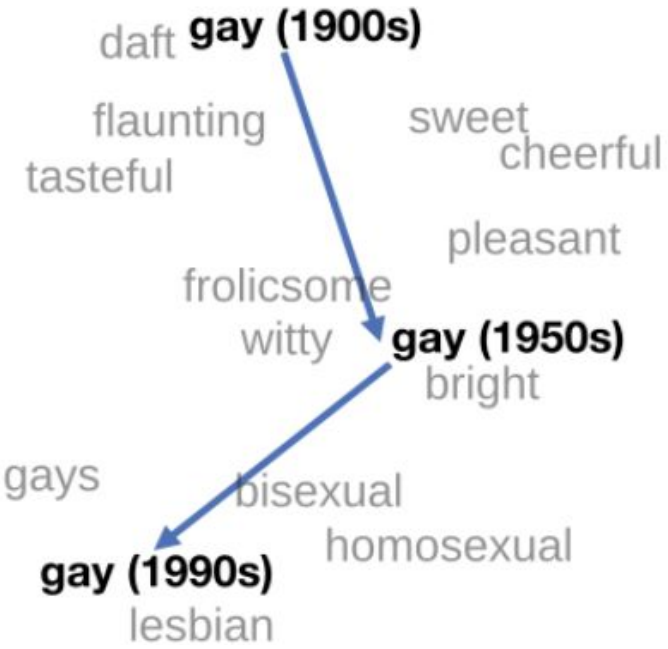
Word	Score
love	1.000
happy	1.000
happily	1.000
toxic	0.008
nightmare	0.005
shit	0.000

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Mohammad,S. (2018). ACL.

Quantifying *negative evaluations* (2)

The cosine similarity between words in vector space models reflects similarity in meaning

We estimate a group label's valence by training word vectors, measuring average valence over the label's nearest K neighbors



Hamilton, WL, et al. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. ACL.

Quantifying *moral disgust*

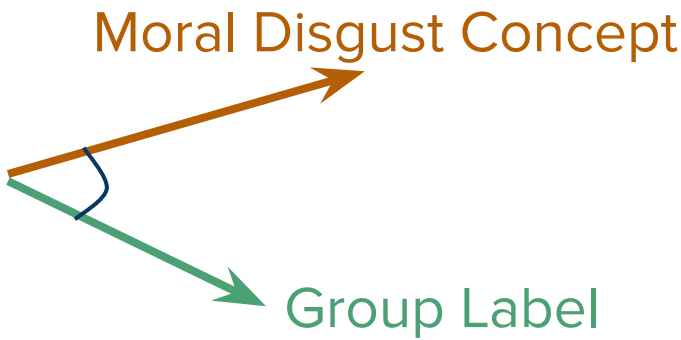
Create vector representation for Moral Disgust Concept

Weighted average of word vectors from Moral Foundations Dict (46 words/stems)

Cosine similarity between Moral Disgust Concept and group label

*disgust**	*sin*
*filth**	*gross*
*repuls**	*pervert*
*profan**	*obscen**

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations..



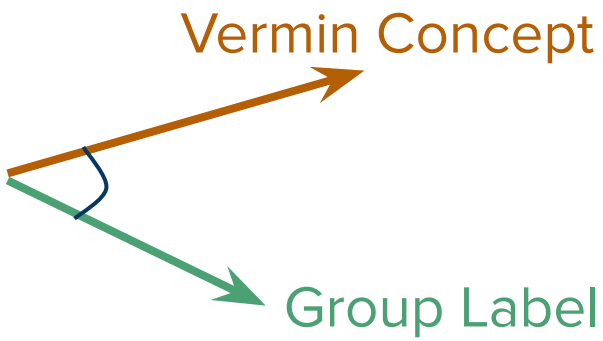
Quantifying *vermin metaphors*

Create vector representation for Vermin Concept

Weighted average of verminy word vectors

Cosine similarity between Vermin Concept and group label

vermin	*rodent(s)*
rat(s)	*cockroach(es)*
mice	*termite(s)*
fleas	*bedbug(s)*



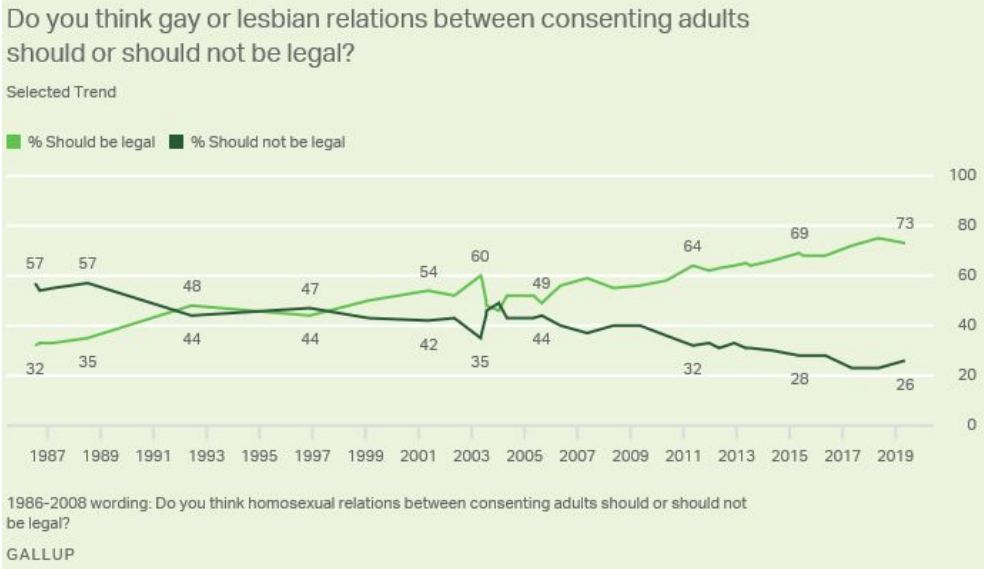
Methods Summary

Dehumanization Element	Operationalization
Negative evaluation of target group	Paragraph-level valence analysis Connotation frames of perspective **Word embedding neighbor valence**
Denial of agency	Connotation frames of agency Word embedding neighbor agency
Moral disgust	**Vector similarity to *disgust***
Vermin metaphor	**Vector similarity to *vermin***

LGBTQ representation in the *New York Times*

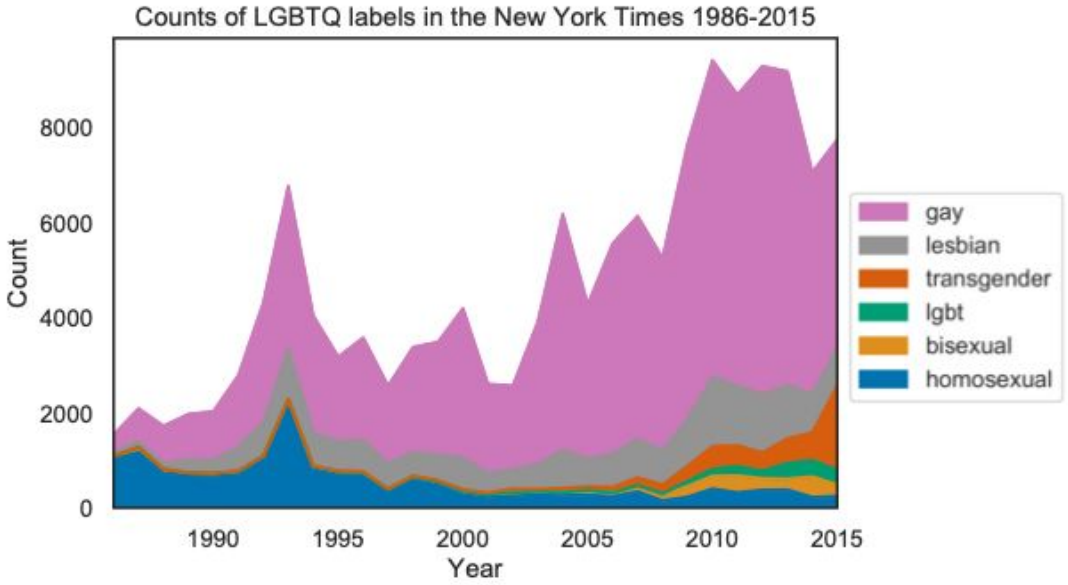
Americans have become more supportive of LGBTQ rights, but LGBTQ people still face significant discrimination

Homosexual: outdated label with clinical and sexual associations



Data

Word embeddings trained per year on full NYT 1986-2015



Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	bestiality
feminist	promiscuity	marriage	pedophilia
suffrage	polygamy	closeted	adultery
sexism	anal	equality	infanticide
a.c.l.u.	intercourse	abortion	abhorrent
amen	consenting	unmarried	feticide
queer	consensual	openly	fornication

Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	bestiality
feminist	promiscuity	marriage	pedophilia
suffrage	polygamy	closeted	adultery
sexism	anal	equality	infanticide
a.c.l.u.	intercourse	abortion	abhorrent
amen	consenting	unmarried	feticide
queer	consensual	openly	fornication

Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	bestiality
feminist	promiscuity	marriage	pedophilia
suffrage	polygamy	closeted	adultery
sexism	anal	**equality**	infanticide
a.c.l.u.	intercourse	abortion	abhorrent
amen	consenting	unmarried	feticide
queer	consensual	openly	fornication

Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	**premarital**	interracial	**premarital**
women	**sexual**	couples	bestiality
feminist	**promiscuity**	marriage	pedophilia
suffrage	polygamy	closeted	adultery
sexism	**anal**	equality	infanticide
a.c.l.u.	**intercourse**	abortion	abhorrent
amen	**consenting**	unmarried	feticide
queer	**consensual**	openly	fornication

Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	**interracial**	premarital
women	sexual	**couples**	bestiality
feminist	promiscuity	**marriage**	pedophilia
suffrage	**polygamy**	closeted	adultery
sexism	anal	equality	infanticide
a.c.l.u.	intercourse	abortion	abhorrent
amen	consenting	**unmarried**	feticide
queer	consensual	openly	fornication

Word embedding top nearest neighbors

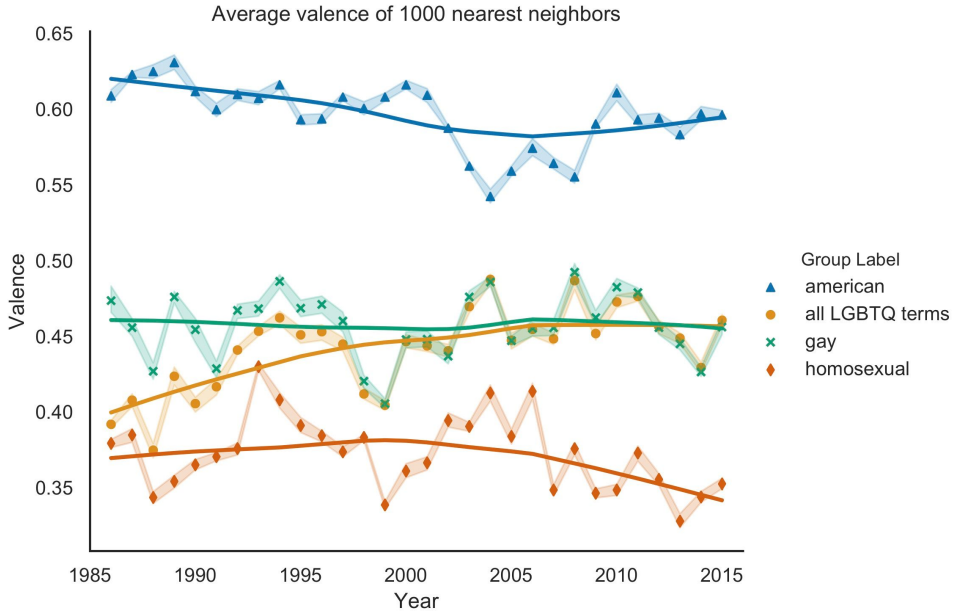
1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	bestiality
feminist	promiscuity	marriage	pedophilia
suffrage	polygamy	**closeted**	adultery
sexism	anal	equality	infanticide
a.c.l.u.	intercourse	abortion	abhorrent
amen	consenting	unmarried	feticide
queer	consensual	**openly**	fornication

Word embedding top nearest neighbors

1986		2015	
gay	***homosexual***	***gay***	***homosexual***
homophobia	premarital	interracial	premarital
women	sexual	couples	**bestiality**
feminist	**promiscuity**	marriage	**pedophilia**
suffrage	polygamy	closeted	**adultery**
sexism	anal	equality	**infanticide**
a.c.l.u.	intercourse	abortion	**abhorrent**
amen	consenting	unmarried	**feticide**
queer	consensual	openly	**fornication**

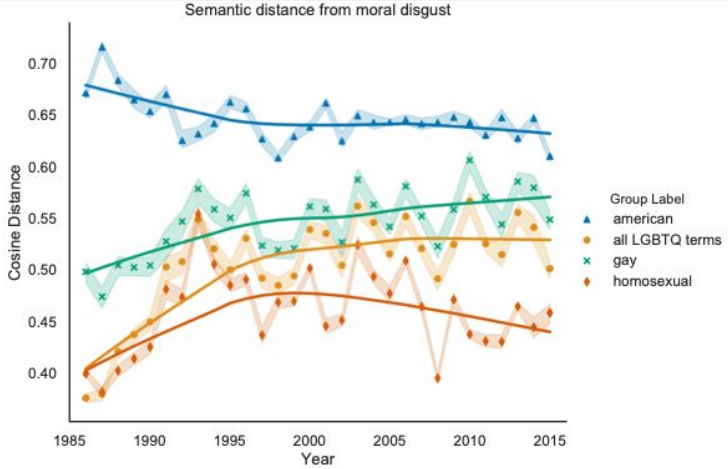
Results: *negative evaluations*

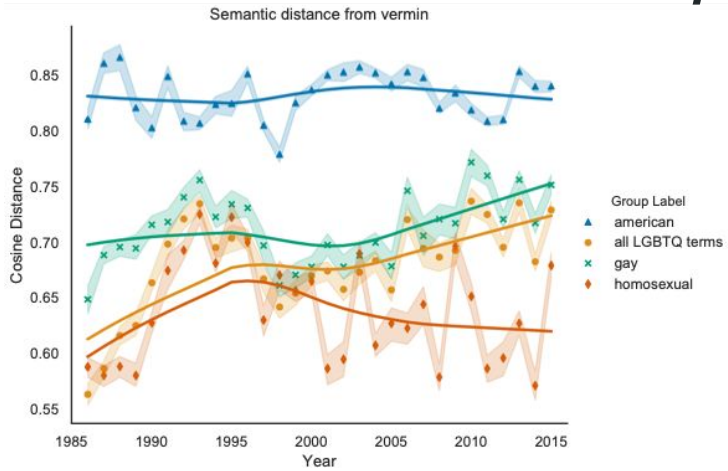
- Evaluations of LGBTQ people have improved over time
- *Homosexual* associated with more negative words than gay



- *Homosexual's* neighboring words become <u>more</u> negative, suggesting that this term is used in more negative (and potentially dehumanizing) contexts

Results: *moral disgust* & *vermin metaphor*





- LGBTQ terms have become less associated with **moral disgust** and **vermin** over time
- *Homosexual* is more associated with **moral disgust** and **vermin** than *gay*, especially after 2000

Summary

Our framework involves:

1. Identifying aspects of dehumanization from literature
2. Measuring lexical semantic correlates with computational methods
3. Qualitative & quantitative evaluation (in paper)

Our study of LGBTQ representation in the *NYT* revealed:

- Increasingly humanizing descriptions of LGBTQ people
- *Homosexual* emerged as an index of more dehumanizing attitudes than other terms (esp. *gay*)

Interdisciplinary Contributions

Framework for large-scale study of dehumanization

Linguistics: language variation and change in discourses surrounding marginalized social groups

Psych: complement small-scale dehumanization studies

CS: Detection of media bias and abusive language

Thank you!

A preprint of our paper is available here.

Julia Mendelsohn
juliamendelsohn.github.io
@jmendelsohn2
juliame@umich.edu

Dan Jurafsky
stanford.edu/~jurafsky
@jurafsky
jurafsky@stanford.edu

Yulia Tsvetkov
cs.cmu.edu/~ytsvetko
ytsvetko@cs.cmu.edu

References

- Bar-Tal, D. (1990). Causes and consequences of delegitimization: Models of conflict and ethnocentrism. Journal of Social issues 46, 65–81
- Buckels, E. E. and Trapnell, P. D. (2013). Disgust facilitates outgroup dehumanization. Group Processes & Intergroup Relations 16, 771–780
- Gallup (2019). Gay and lesbian rights. http://news.gallup.com/poll/1651/gay-lesbian-rights.aspx
- Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology 96-1029
- Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)
- Haslam, N. and Stratemeyer, M. (2016). Recent research on dehumanization. Current Opinion in Psychology 11, 25–29
- Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In Proceedings of The 56th Annual Meeting of the Association for Computational Linguistics (ACL)
- Opotow, S. (1990). Moral exclusion and injustice: An introduction. Journal of social issues 46, 1–20
- Rashkin, H., Singh, S., and Choi, Y. (2016). Connotation frames: A data-driven investigation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics
- Sap, M., Prasettio, M. C., Holtzman, A., Rashkin, H., and Choi, Y. (2017). Connotation frames of power and agency in modern films. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2329–2334
- Sherman, G. D. and Haidt, J. (2011). Cuteness and disgust: the humanizing and dehumanizing effects of emotion. Emotion Review 3, 245–251
- Tipler, C. and Ruscher, J. B. (2014). Agency's role in dehumanization: Non-human metaphors of out-groups. Social and Personality Psychology Compass 8, 214–228

Additional slides

Bias in human-annotated VAD lexicon

We filtered LGBTQ labels before calculating valence

LGBTQ term	Valence	Other term	Valence
transsexual	0.264	*woman*	0.865
homosexual	0.333	*human*	0.767
lesbian	0.385	*man*	0.688
gay	0.388	*person*	0.646
bisexual	0.438	*heterosexual*	0.561

Quantifying *negative evaluations* (1)

Valence: evaluation from negative (unpleasant) to positive (pleasant)

NRC VAD lexicon: valence scores from 0 to 1 for 20k English words

Calculate average valence score over all words in the text

Word	Score
love	1.000
happy	1.000
happily	1.000
toxic	0.008
nightmare	0.005
shit	0.000

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Mohammad,S. (2018). ACL.

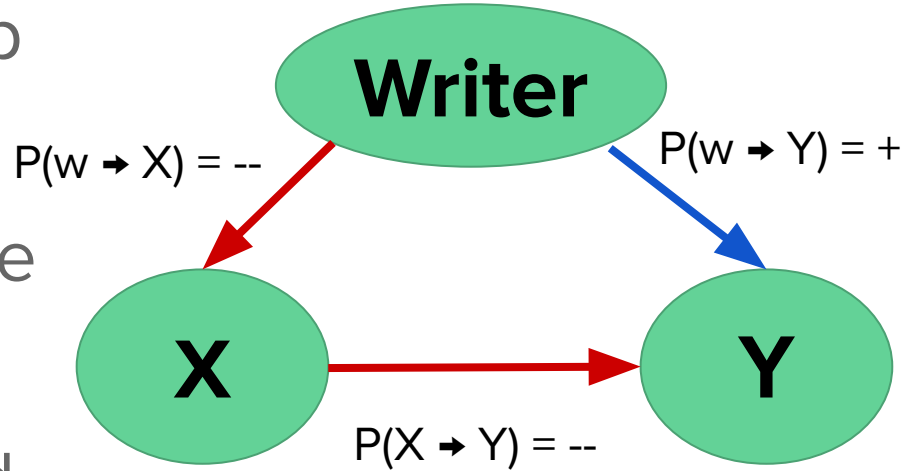
Quantifying *negative evaluations* (2)

We want to measure valence **directed towards** target group

Connotation Frames Lexicon: 900 verbs, writer's perspective towards subj and obj

Extracted SVO tuples for head verbs where group label was in subj or obj NP

X violates Y



Rashkin, H., Singh, S., & Choi, Y. (2016). Connotation Frames: A Data-Driven Investigation. ACL.

Components of dehumanization

4. Denial of agency

Agency: The ability to:
(1) experience emotion & feel pain (affective mental states)
(2) act & produce effect on environment (behavioral potential)
(3) think & hold beliefs (cognitive mental states)
[Tipler & Ruscher, 2014]

Quantifying *denial of agency*

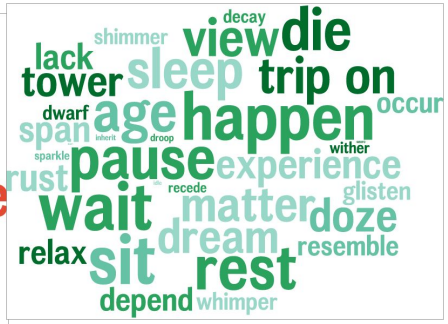
Agency Connotation Frames:

2k verbs labeled for agency

High agency: high control, active decision-makers

Low agency: more passive

Fraction of high-agency subjects in SV pairs containing group label



Sap, M. et al. (2017). Connotation frames of power and agency in modern films. EMNLP.

Quantifying *denial of agency* (2)

NRC VAD lexicon: dominance scores from 0 to 1 for 20k words

Calculate dominance score over nearest K word2vec neighbors

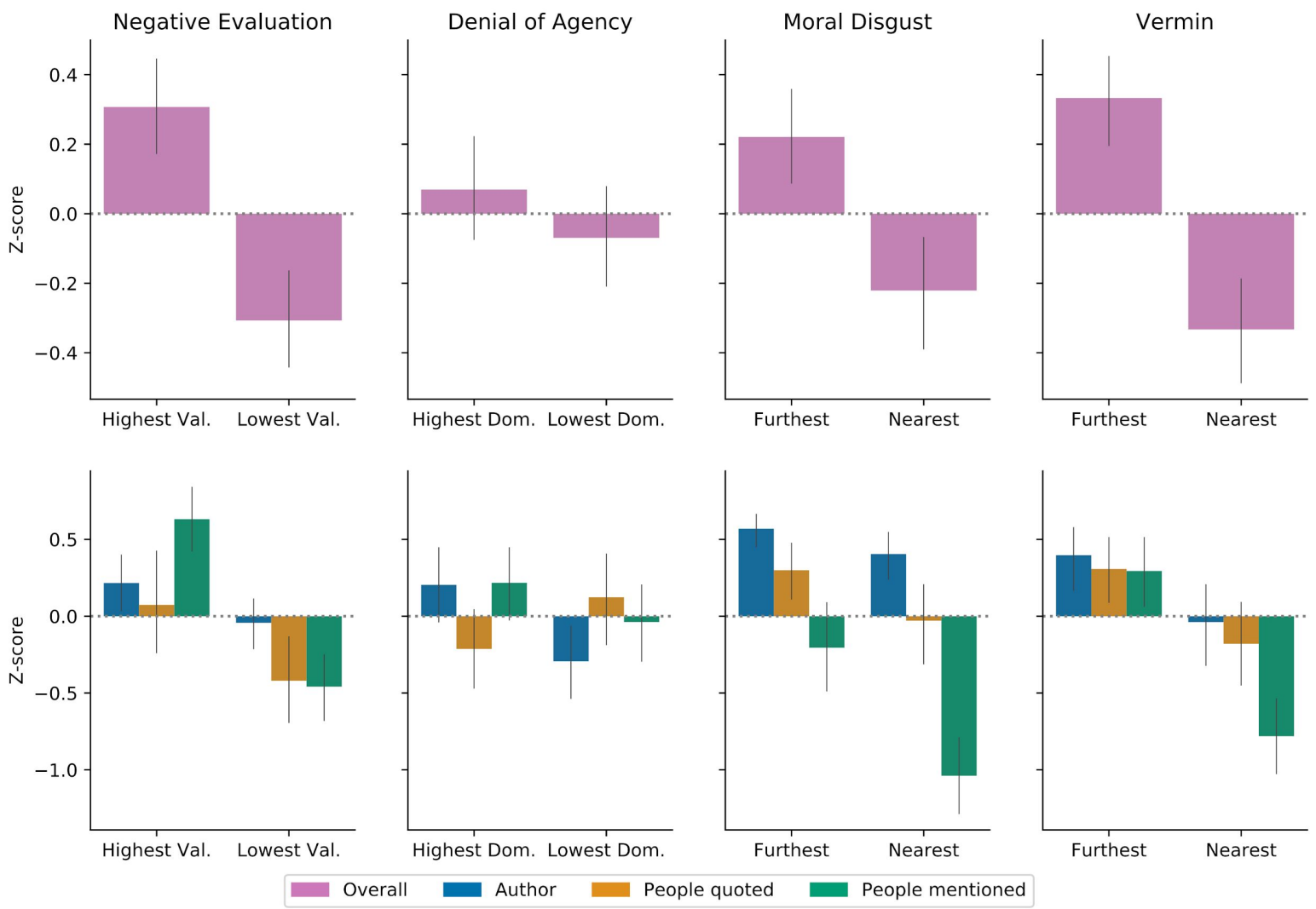
Limitation: power != agency

Word	Score
powerful	0.991
leadership	0.983
success	0.981
empty	0.081
frail	0.069
weak	0.045

Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. Mohammad,S. (2018). ACL.

Tradeoffs: *negative evaluation* methods

Paragraph	**Connotation frames**	**Vector neighbors**
interpretable	interpretable	less interpretable
broader context	limited scope	broader context
not directed	directed	directed
topical effects	syntax is hard	major events
Disentangling perspectives within text		


Negative Evaluation
Denial of Agency
Moral Disgust
Vermin
Z-score
0.4
0.2
0.0
−0.2
−0.4
Highest Val.
Lowest Val.
Highest Dom.
Lowest Dom.
Furthest
Nearest
Furthest
Nearest
Z-score
0.5
0.0
−0.5
−1.0
Highest Val.
Lowest Val.
Highest Dom.
Lowest Dom.
Furthest
Nearest
Furthest
Nearest
Overall
Author
People quoted
People mentioned

Future directions

- More sophisticated methods (contextual embeddings)
- Measure other dimensions of dehumanization and non-lexical semantic cues
  - Denial of subjectivity (quote attribution)
  - Psychological distance (definite plurals)
  - Essentialism (noun v. adjective forms)
- Other LGBTQ terms, groups, data sources, languages

Ethical concerns

- Biases in lexicons and methods
- Vectors are dehumanizing
- Case Study: Aggregated LGBTQ representations suppress diversity of identities within this umbrella
- Emphasis on *gay* and *homosexual* and erasure of marginalized people within LGBTQ communities