Natural Language Processing for Law and Social Science

7. Document Embeddings

Outline

Outline

Research Objectives

1. **What is the research question?**
2. Corpus and Data.

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
   - **What are we trying to measure?**

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
    - **What are we trying to measure?**
    - Select a model and train it.
    - **Validate that the model is measuring what we want and that there are no clear confounders.**

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
   - **What are we trying to measure?**
   - Select a model and train it.
   - **Validate that the model is measuring what we want and that there are no clear confounders.**
4. Empirical analysis
   - Produce statistics or predictions with the trained model.
   - **Answer the research question.**

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

Implicit attitudes

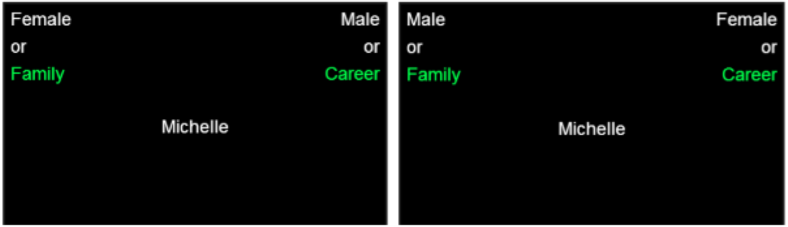
"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

- Generally measured using Implicit Association Tests (IATs)

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)

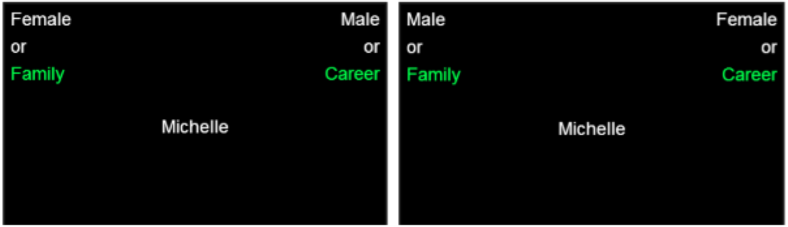


- Comparing reaction times across trials with different word pairs:
  - subjects tend to be slower and more error-prone in assignments against stereotype (e.g. “Michelle” goes to “Female or Career”).

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)



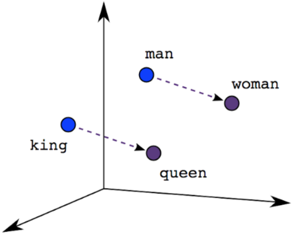
- Comparing reaction times across trials with different word pairs:
  - subjects tend to be slower and more error-prone in assignments against stereotype (e.g. "Michelle" goes to "Female or Career").
  - IAT score = difference in reaction time between stereotype-consistent and stereotype-inconsistent rounds.

Caliskan, Bryson, and Narayanan (*Science* 2017)

- "We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. . . "

Caliskan, Bryson, and Narayanan (*Science* 2017)

- "We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. . . "

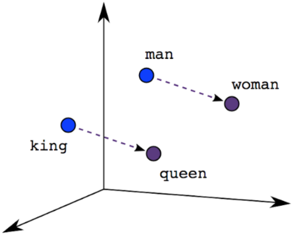


Analogies

- king : queen :: man : woman
- walked : walking :: swam : swimming

Caliskan, Bryson, and Narayanan (*Science* 2017)

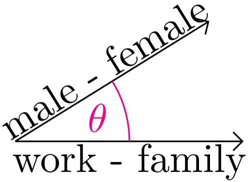
- “We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. . . “



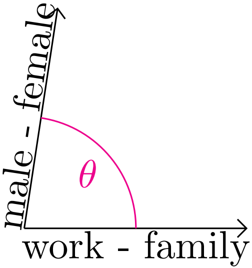
Analogies

- king : queen :: man : woman
- walked : walking :: swam : swimming
- **man : programmer :: woman : homemaker**
- **he : physician :: she : nurse**

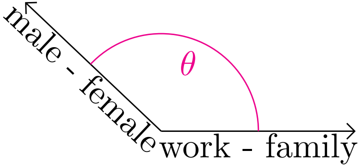
Measuring Gender Stereotypes using Cosine Similarity



(a)



(b)



(c)

Example Stimuli

- Targets:
  - **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
  - **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.

Example Stimuli

- Targets:
  - **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
  - **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- Attributes:
  - **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
  - **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names

Results

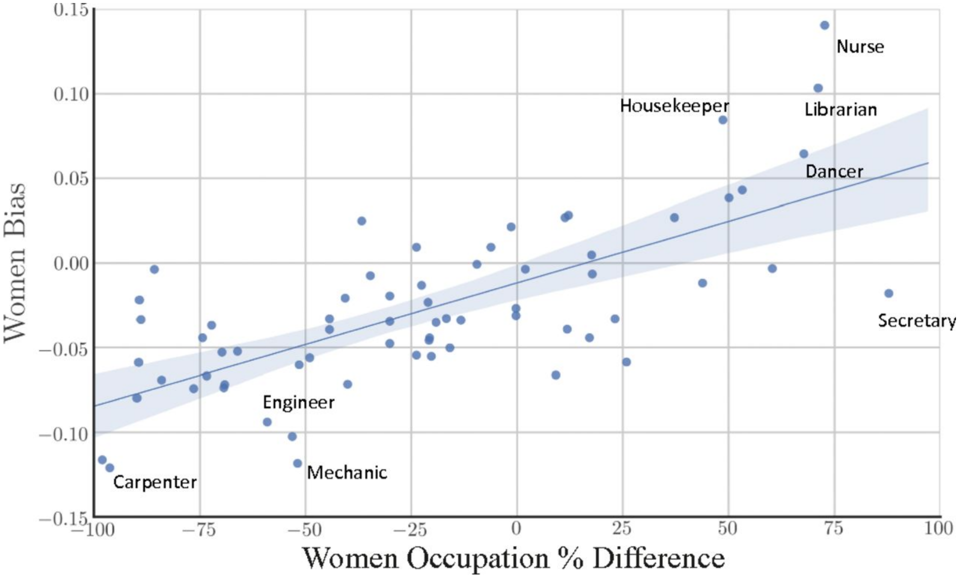
- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names
- Male names vs. Female names:

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names
- Male names vs. Female names:
  - Career words (e.g. professional, corporation, ...) vs. family words (e.g. home, children, ...)
  - Math/science words vs arts words

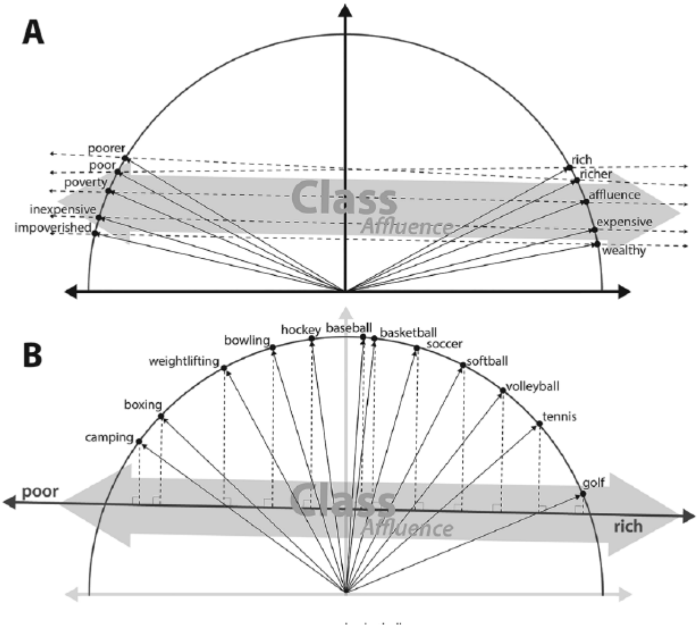
What do we learn from this?

Garg, Schiebinger, Jurafsky, and Zou (PNAS 2018)



Women's occupation relative percentage vs. embedding bias in Google News vectors.

Kozlowski, Evans, and Taddy (ASR 2019)



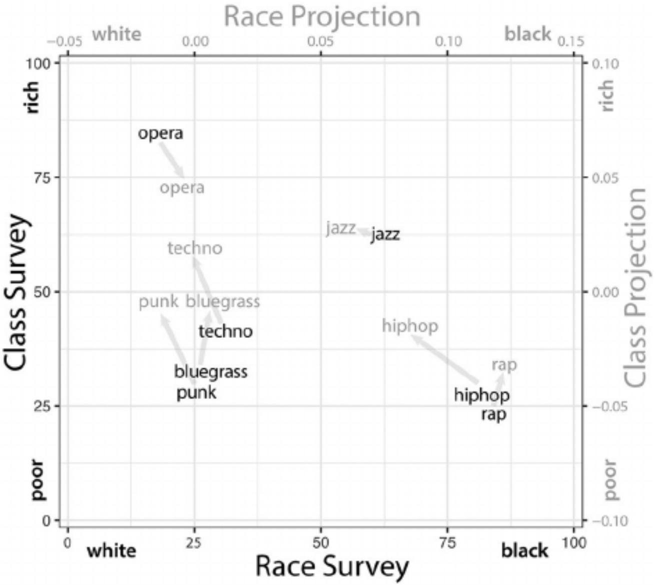


Figure 3. Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)

Time Series Analysis of Affluence

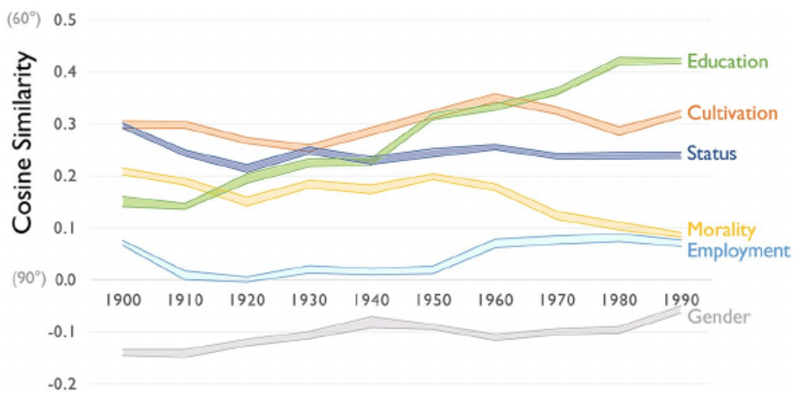


Figure 5. Cosine Similarity between the Affluence Dimension and Six Other Cultural Dimensions of Class by Decade; 1900 to 1999 Google Ngrams Corpus
Note: Bands represent 90 percent bootstrapped confidence intervals produced by subsampling.

"Among the 10 nouns most highly projecting on the affluence dimension in the first decade of the twentieth century are "fragrance," "perfume," "jewels," and "gems," ..."

Discussion

Discussion

- What are we measuring?
  - e.g., how do measurements by person or by group correlate with actual attitudes, for example as measured by IAT scores?

Discussion

- What are we measuring?
  - e.g., how do measurements by person or by group correlate with actual attitudes, for example as measured by IAT scores?
- What attitudes can be reliably measured?
  - e.g. racial sentiment, especially in light of black sheep problem.

Discussion

- What are we measuring?
  - e.g., how do measurements by person or by group correlate with actual attitudes, for example as measured by IAT scores?
- What attitudes can be reliably measured?
  - e.g. racial sentiment, especially in light of black sheep problem.
- In what domains is this relevant?
  - social media, news media, politics, legal, scientific, ...

Discussion

- What are we measuring?
  - e.g., how do measurements by person or by group correlate with actual attitudes, for example as measured by IAT scores?
- What attitudes can be reliably measured?
  - e.g. racial sentiment, especially in light of black sheep problem.
- In what domains is this relevant?
  - social media, news media, politics, legal, scientific, ...
- Does language matter?
  - Djourelova (2020): style change from “illegal” to “undocumented” immigrant softened attitudes toward immigration.

Outline

Bias in NLP Systems

Sentiment Analysis

```
text_to_sentiment("Let's go get Italian food")
2.0429166109
text_to_sentiment("Let's go get Chinese food")
1.4094033658
text_to_sentiment("Let's go get Mexican food")
0.3880198556
```

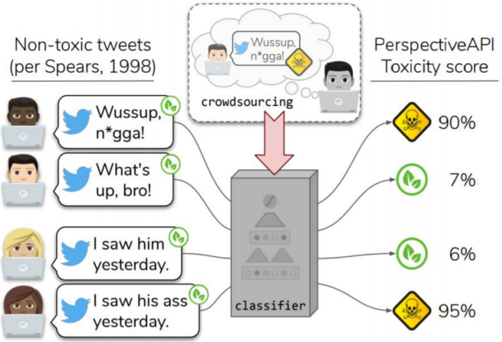
```
text_to_sentiment("My name is Emily")
2.2286179365
text_to_sentiment("My name is Heather")
1.3976291151
text_to_sentiment("My name is Yvette")
0.9846380213
text_to_sentiment("My name is Shaniqua")
-0.4704813178
```

Is this sentiment model racist?

Source: Kareem Carr slides.

Bias in NLP Systems

Toxicity Detection



	Within dataset proportions				
DWMW17			% false identification		
Group	Acc.	None	Offensive	Hate	
AAE	94.3	1.1	**46.3**	0.8	
White	87.5	**7.9**	9.0	**3.8**	
Overall	91.4	2.9	17.9	2.3	
FDCL18			% false identification		
Group	Acc.	None	Abusive	Hateful	
AAE	81.4	4.2	**26.0**	**1.7**	
White	82.7	**30.5**	4.5	0.8	
Overall	81.4	20.9	6.6	0.8	

Is this toxicity detection model racist?

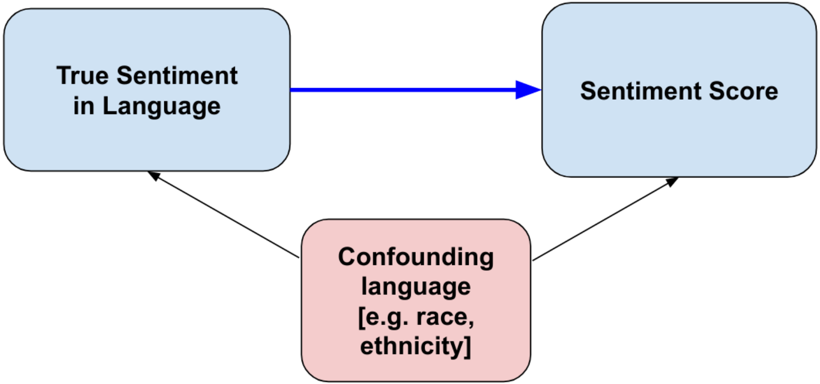
Source: Jacobs and Wallach slides.

NLP “Bias” is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.

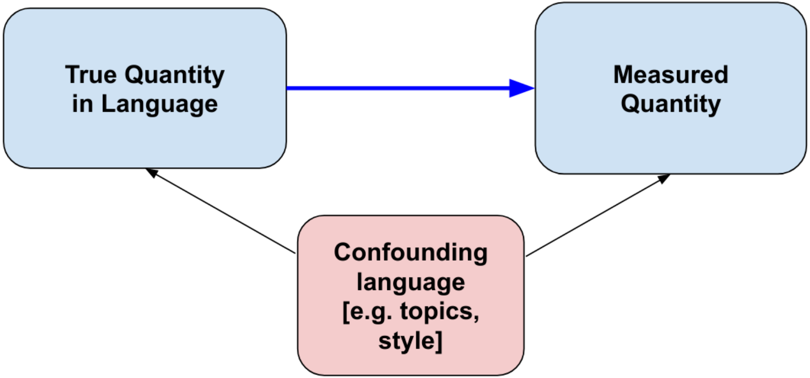
NLP "Bias" is statistical bias

- Sentiment scores that are trained on annotated datasets also learn from the correlated non-sentiment information.



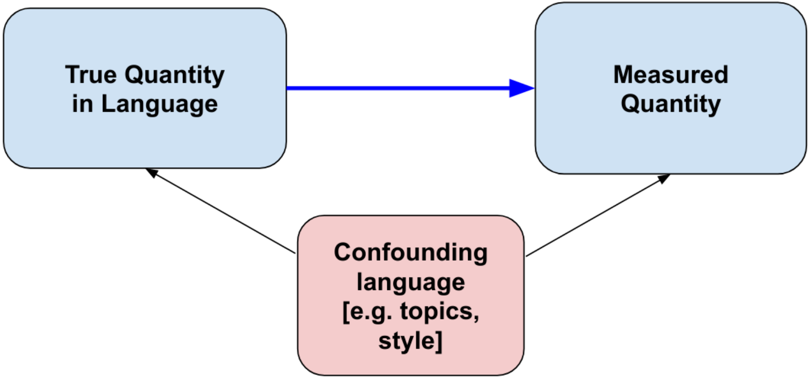
- Supervised sentiment models are confounded by correlated language factors.
  - e.g., in the training set maybe people complain about Mexican food more often than Italian food.

This is a universal problem



- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

This is a universal problem



- supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- An important exception: dictionary methods (perhaps explaining why they are often used by economists). But they have other serious limitations.

Examples: Confounders in Measurement from Text

What quantity do we care about? vs. What do we measure?

Examples: Confounders in Measurement from Text

What quantity do we care about? vs. What do we measure?

- Positive/negative sentiment → Count positive/negative words, or predict text annotations.
- Toxicity → Count toxic words, or predict text annotations.

confounders?

Examples: Confounders in Measurement from Text

What quantity do we care about? vs. What do we measure?

- Positive/negative sentiment $\rightarrow$ Count positive/negative words, or predict text annotations.
- Toxicity $\rightarrow$ Count toxic words, or predict text annotations.

confounders?

- Student performance $\rightarrow$ predicted essay grade based on labeled essay documents.
- Credit worthiness $\rightarrow$ predicted probability of default based on loan application documents.

confounders?

Examples: Confounders in Measurement from Text

What quantity do we care about? vs. What do we measure?

- Positive/negative sentiment $\rightarrow$ Count positive/negative words, or predict text annotations.
- Toxicity $\rightarrow$ Count toxic words, or predict text annotations.

confounders?

- Student performance $\rightarrow$ predicted essay grade based on labeled essay documents.
- Credit worthiness $\rightarrow$ predicted probability of default based on loan application documents.

confounders?

- Political partisanship $\rightarrow$ predicted probability being Democrat/Republican based on speeches.

confounders?

Examples: Confounders in Measurement from Text

What quantity do we care about? vs. What do we measure?

- Positive/negative sentiment $\rightarrow$ Count positive/negative words, or predict text annotations.
- Toxicity $\rightarrow$ Count toxic words, or predict text annotations.

confounders?

- Student performance $\rightarrow$ predicted essay grade based on labeled essay documents.
- Credit worthiness $\rightarrow$ predicted probability of default based on loan application documents.

confounders?

- Political partisanship $\rightarrow$ predicted probability being Democrat/Republican based on speeches.

confounders?

- Policy priorities $\rightarrow$ predicted probability of speeches/laws being about a particular policy topic.

confounders?

When is measurement confounding important?

- By itself, producing measurements that are biased by confounders might not be a problem.
- e.g.:
  - an NLP-based credit score that learns confounders $\rightarrow$ not a problem unless debtors learn about it and strategically alter their documents.
    - similarly with automated essay grading

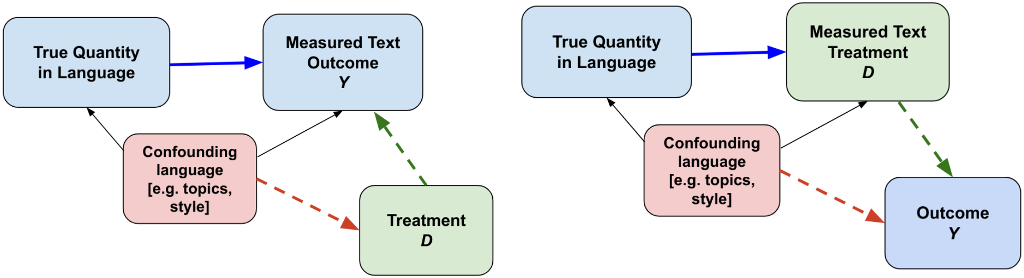
When is measurement confounding important?

- By itself, producing measurements that are biased by confounders might not be a problem.
- e.g.:
  - an NLP-based credit score that learns confounders $\rightarrow$ not a problem unless debtors learn about it and strategically alter their documents.
    - similarly with automated essay grading
- for measuring political divisiveness or policy priorities
  - probably won't matter for in-domain summary statistics
  - but would matter a lot for summary statistics in a new domain

When is measurement confounding important?

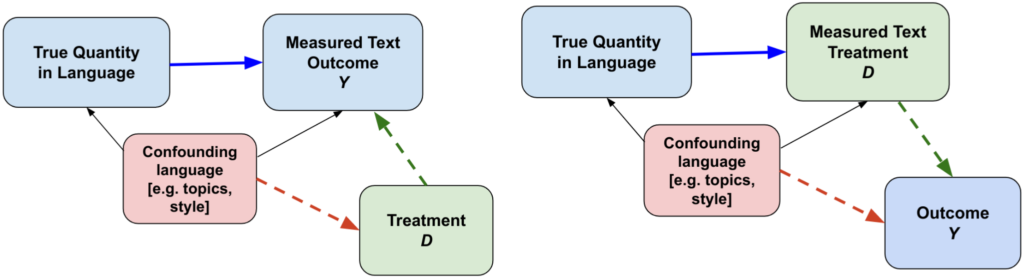
- By itself, producing measurements that are biased by confounders might not be a problem.
- e.g.:
  - an NLP-based credit score that learns confounders $\rightarrow$ not a problem unless debtors learn about it and strategically alter their documents.
    - similarly with automated essay grading
- for measuring political divisiveness or policy priorities
  - probably won't matter for in-domain summary statistics
  - but would matter a lot for summary statistics in a new domain
- even in domain, will matter for assessing the causal effect of a treatment, e.g. the electoral cycle:
  - elections might cause politicians to focus on social issues rather than economic issues,
  - if social/economic issues are confounded with partisanship, the resulting estimates are biased.

When is measurement confounding important?



- When text is outcome, the confounders cannot be correlated with the treatment.
- When text is treatment, the confounders cannot be correlated with the outcome.

When is measurement confounding important?



- When text is outcome, the confounders cannot be correlated with the treatment.
- When text is treatment, the confounders cannot be correlated with the outcome.
  - e.g.: estimating the effect of politician speech sentiment on his/her reelection chances?

Steps for de-biasing

- Language features that are often confounded with the quantity of interest:
  - stopwords
  - named entities: person/organization/place names
- These can be dropped during pre-processing to reduce the influence of confounders in subsequent measurements.
- Can control for topic or style features or other potential confounders in regressions.

De-Biasing Word Embeddings

De-Biasing Word Embeddings

- Bolukbasi et al (NIPS 2016):
  - “Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**.“

De-Biasing Word Embeddings

- Bolukbasi et al (NIPS 2016):
  - "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**.“
  - "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."

De-Biasing Word Embeddings

- Bolukbasi et al (NIPS 2016):
  - "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."
  - "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."
  - "Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female."

De-Biasing Word Embeddings

- Bolukbasi et al (NIPS 2016):
  - "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."
  - "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."
  - "Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female."
- But: Gonen and Goldberg (2019):
  - ***"... we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between 'gender-neutralized' words in the debiased embeddings, and can be recovered from them..."***
- Project idea: use double machine learning to de-bias word embeddings.

Outline

What is (Document) Embedding?

“Embedding”: a lower-dimensional dense vector representation of a higher-dimensional object

- also refers to algorithm for making such vectors

What is (Document) Embedding?

“Embedding”: a lower-dimensional dense vector representation of a higher-dimensional object

- also refers to algorithm for making such vectors

Document vectors:

- quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.

What is (Document) Embedding?

“Embedding”: a lower-dimensional dense vector representation of a higher-dimensional object

- also refers to algorithm for making such vectors

Document vectors:

- quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- standard approach:
  - represent documents as sparse vectors of token counts/frequencies.
  - e.g., to compute document similarity measures, to learn topics, or to use as features in supervised learning

What is (Document) Embedding?

"Embedding": a lower-dimensional dense vector representation of a higher-dimensional object

- also refers to algorithm for making such vectors

Document vectors:

- quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- standard approach:
  - represent documents as sparse vectors of token counts/frequencies.
  - e.g., to compute document similarity measures, to learn topics, or to use as features in supervised learning
- **Embedding approach**:
  - low-dimensional dense vectors rather than high-dimensional sparse vectors
  - Embedding without neural nets:
    - PCA reductions of the document-term matrix
    - LDA topic shares

What is (Document) Embedding?

"Embedding": a lower-dimensional dense vector representation of a higher-dimensional object

- also refers to algorithm for making such vectors

Document vectors:

- quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- standard approach:
  - represent documents as sparse vectors of token counts/frequencies.
  - e.g., to compute document similarity measures, to learn topics, or to use as features in supervised learning
- **Embedding approach**:
  - low-dimensional dense vectors rather than high-dimensional sparse vectors
  - Embedding without neural nets:
    - PCA reductions of the document-term matrix
    - LDA topic shares
  - Embedding with neural nets (today):
    - many useful ways to do this.

Embedding layers can produce document embeddings

- **Embedding layers** take a categorical variable as input and produce a low-dimensional dense representation.

Embedding layers can produce document embeddings

- **Embedding layers** take a categorical variable as input and produce a low-dimensional dense representation.

Can be used to produce document embeddings:

- Tokenize document to fixed length n_L
- Inputs are each word position, input categorical (word) to n_E-dimensional embedding layer:
$$\boldsymbol{x}_{1:n_L} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_t & \dots & \boldsymbol{x}_{n_L} \end{bmatrix}$$
- pipe to further hidden layers of network.

Embedding layers can produce document embeddings

- **Embedding layers** take a categorical variable as input and produce a low-dimensional dense representation.

Can be used to produce document embeddings:

- Tokenize document to fixed length n_L
- Inputs are each word position, input categorical (word) to n_E-dimensional embedding layer:
$$\boldsymbol{x}_{1:n_L} = [\ \boldsymbol{x}_1 \quad \ldots \quad \boldsymbol{x}_t \quad \ldots \quad \boldsymbol{x}_{n_L}\ ]$$
- pipe to further hidden layers of network.
- document embedding $=$ $n_L n_E$-dimensional vector of concatenated word embeddings.
    - computationally demanding and only works with short documents.

Autoencoder Encodings

- Autoencoder compresses a document (e.g. a sentence) into a vector to be reconstructed.
  - Can use the compressed representation as a document embedding.
- Standard (that is, non-transformer) autoencoder embeddings don't tend to work well for sentence similarity tasks because autoencoders try to reproduce the specific wording (reconstruction objective), rather than the semantic meaning.
  - transformer-based autoencoders, e.g. BART, address this issue (Week 9)

Outline

Word Vectors can produce Document Vectors

$$\vec{D} = \sum_{w \in D} a_w \vec{w}$$

- The "continuous bag of words" representation for document D is the sum, or the average (potentially weighted by a_w), of the vectors $\vec{w}$ for each word w in the document.
  - word vectors $\vec{w}$ constructed using Word2Vec or GloVe (pre-trained or trained on the corpus).
  - "Document" could be sentence, paragraph, section, etc.

Word Vectors can produce Document Vectors

$$\vec{D} = \sum_{w \in D} a_w \vec{w}$$

- The "continuous bag of words" representation for document D is the sum, or the average (potentially weighted by a_w), of the vectors $\vec{w}$ for each word w in the document.
  - word vectors $\vec{w}$ constructed using Word2Vec or GloVe (pre-trained or trained on the corpus).
  - "Document" could be sentence, paragraph, section, etc.
- Arora, Liang, and Ma (2017) provide a "tough to beat baseline", the SIF-weighted ("smoothed inverse frequency") average of the vectors:

$$a_w = \frac{\alpha}{\alpha + p_w}$$

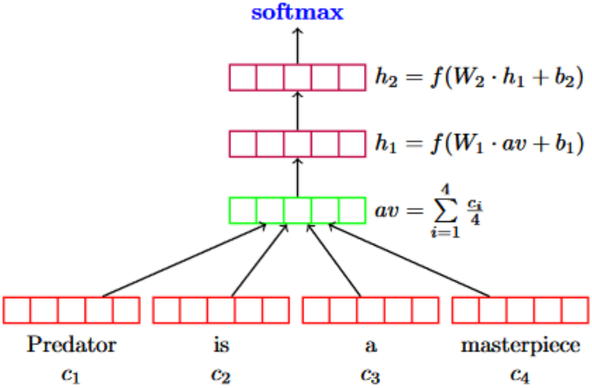
where p_w is the probability (frequency) of the word and $\alpha = .001$ is a smoothing parameter.

Deep Averaging Network (Iyyer et al 2015)

- Similar to the previous aggregated word embedding methods, but embeddings are learned during training:

Deep Averaging Network (Iyyer et al 2015)

- Similar to the previous aggregated word embedding methods, but embeddings are learned during training:



1. Trainable embedding layer for words, initialized with pre-trained embeddings
2. Average the embeddings, with dropout (sometimes words left out of average)
3. Average embedding fed into MLP with multiple hidden layers
4. MLP outputs used for classification or regression

fastText: Hashed N-Gram Embeddings (Joulin et al 2016)

Combines the Iyyer et al (2015) approach with the hashing n-gram vectorizer.

fastText: Hashed N-Gram Embeddings (Joulin et al 2016)

Combines the Iyyer et al (2015) approach with the hashing n-gram vectorizer.

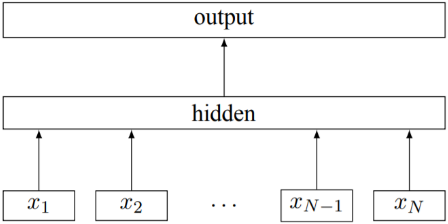


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

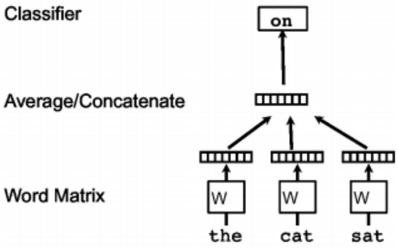
1. Allocate $n_w \approx 10$ million rows to embedding matrix.
2. Assign n-grams to embedding indexes with hashing function.
3. sentence embedding = average of n-gram embeddings
4. send to dense hidden layer(s)
5. send to output (e.g. classifier / regressor).

- Captures the local predictive power of n-grams without building vocabulary or costly training of CNN.

Outline

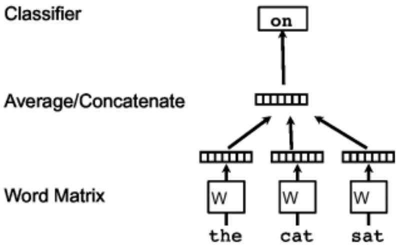
Doc2Vec (Le and Mikolov)

- Recall that Word2Vec trains word embeddings to predict a word given neighboring context words:

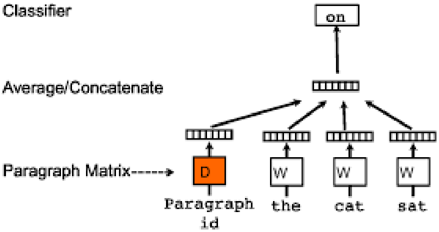


Doc2Vec (Le and Mikolov)

- Recall that Word2Vec trains word embeddings to predict a word given neighboring context words:



- Doc2Vec augments Word2Vec with a categorical embedding for the document (e.g. paragraph):



Doc2Vec on Wikipedia (Dai, Olah, and Le 2015)

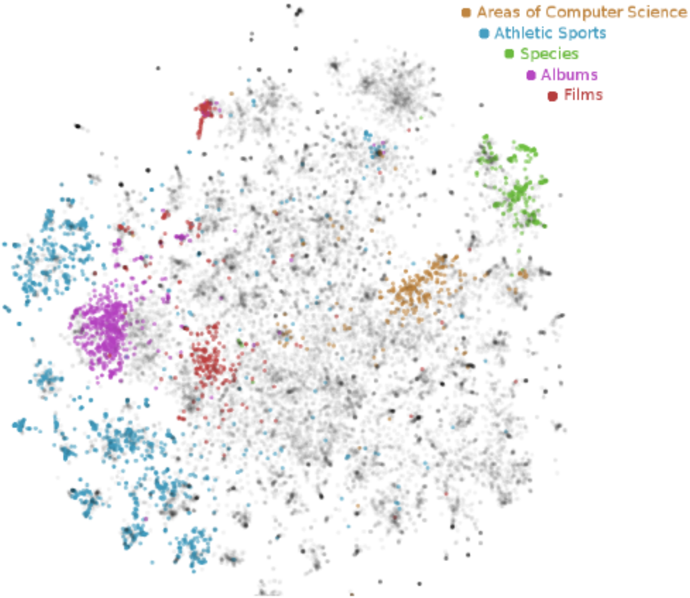
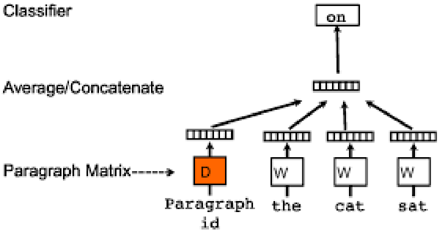


Figure 3: Visualization of Wikipedia paragraph vectors using t-SNE.

Vectorizing New Documents



- A new document that wasn't in training does not have a vector.
- Document inference step:
  - freeze word embeddings in input layer and in output layer.
  - learn embedding for new document to predict sampled words in new document.

Document Embeddings Geometry

- With topic models, each dimension has a topical interpretation.
- With document embeddings, a direction (might) have a topical interpretation.

Document Embeddings Geometry

- With topic models, each dimension has a topical interpretation.
- With document embeddings, a direction (might) have a topical interpretation.
- Analogous with word embeddings, directions in document embedding capture analogous dimensions of documents:

Table 2: Wikipedia nearest neighbours

(a) Wikipedia nearest neighbours to "Lady Gaga" using Paragraph Vectors. All articles are relevant.

Article	Cosine Similarity
Christina Aguilera	0.674
Beyonce	0.645
Madonna (entertainer)	0.643
Artpop	0.640
Britney Spears	0.640
Cyndi Lauper	0.632
Rihanna	0.631
Pink (singer)	0.628
Born This Way	0.627
The Monster Ball Tour	0.620

(b) Wikipedia nearest neighbours to "Lady Gaga" - "American" + "Japanese" using Paragraph Vectors. Note that Ayumi Hamasaki is one of the most famous singers, and one of the best selling artists in Japan. She also has an album called "Poker Face" in 1998.

Article	Cosine Similarity
Ayumi Hamasaki	0.539
Shoko Nakagawa	0.531
Izumi Sakai	0.512
Urbangarde	0.505
Ringo Sheena	0.503
Toshiaki Kasuga	0.492
Chihiro Onitsuka	0.487
Namie Amuro	0.485
Yakuza (video game)	0.485
Nozomi Sasaki (model)	0.485

Doc2Vec for Judicial Opinions (Ash and Chen 2018)

- Corpus: 300,000 cases from U.S. Circuit Courts, 1870-2010.
- Produce document vectors for each case to understand differences between judges and courts.

Doc2Vec for Judicial Opinions (Ash and Chen 2018)

- Corpus: 300,000 cases from U.S. Circuit Courts, 1870-2010.
- Produce document vectors for each case to understand differences between judges and courts.
- De-mean vectors by group (court, topic, or year) to extract relevant information:
  - de-mean by topic-year to distinguish courts.
  - de-mean by court-topic to distingush years.
  - de-mean by court-year to distinguish topics.

Figure 1: Centered by Topic-Year, Averaged by Judge, Labeled by Court

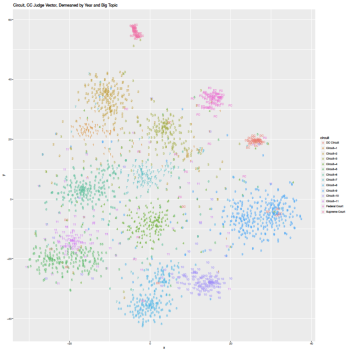


Figure 2: Centered by Court-Topic, Averaged by Court-Year, Labeled by Decade

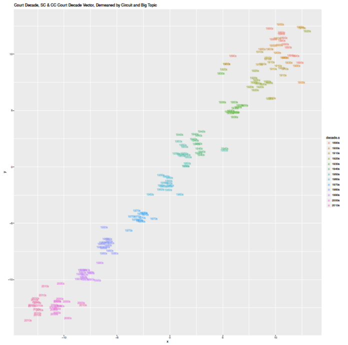
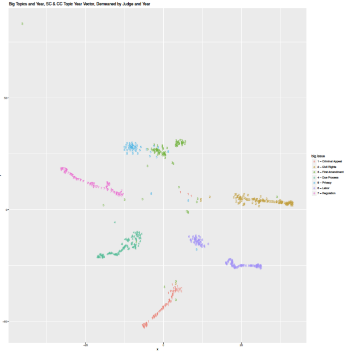


Figure 3: Centered by Judge-Year, Averaged by Topic-Year, Labeled by Topic



Outline

StarSpace: Embed Anything (Wu et al 2018)

Generalize two key embedding ingredients from NLP to much broader set of tasks:

StarSpace: Embed Anything (Wu et al 2018)

Generalize two key embedding ingredients from NLP to much broader set of tasks:

1. aggregate embeddings across words or phrases by document $\rightarrow$ aggregate embeddings across features by entity
2. negative sampling of co-locating words vs random words $\rightarrow$ negative sampling of related entities vs unrelated entities

Entities and Features

- **features** are categorical variables.
  - learn $n_F \times n_E$ embedding matrix F with n_F features and embedding dimension n_E.
- **entities** are bags of features:
  - for entity consisting of features $a = \{1, 2, ..., i, ...\}$, sum over feature embeddings:

$$\vec{a} = \sum_{i \in a} F_i$$

where F_i indicates the associated row of F.

Entities and Features

- **features** are categorical variables.
  - learn $n_F \times n_E$ embedding matrix F with n_F features and embedding dimension n_E.
- **entities** are bags of features:
  - for entity consisting of features $a = \{1, 2, ..., i, ...\}$, sum over feature embeddings:

    $$\vec{a} = \sum_{i \in a} F_i$$

    where F_i indicates the associated row of F.
- Then by construction, entities and features are in the same space.

StarSpace Negative Sampling Objective

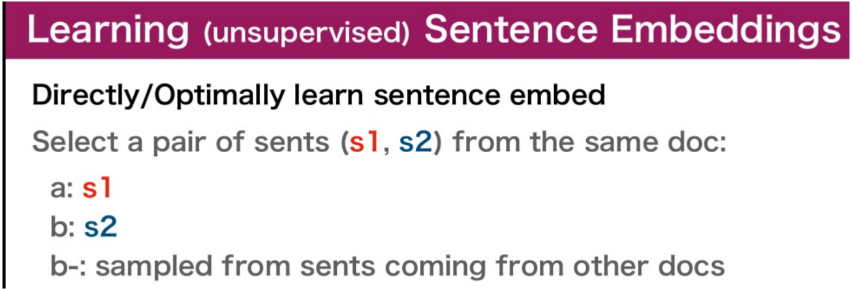
- For entity a selected at current training batch:
  - positive sample: related entity b (e.g. two sentences from the same document).
  - negative samples: k unrelated entities $b_1^-, \ldots b_k^-$ (e.g. sentences in other documents).

StarSpace Negative Sampling Objective

- For entity a selected at current training batch:
  - positive sample: related entity b (e.g. two sentences from the same document).
  - negative samples: k unrelated entities $b_1^-, ...b_k^-$ (e.g. sentences in other documents).
- Compute vectors $\vec{a} = \sum_{i \in a} F_i$, $\vec{b}$, $\vec{b}_1^-, ..., \vec{b}_k^-$
- Compute cosine similiarities sim$(\vec{a}, \vec{b})$,sim$(\vec{a}, \vec{b}_1^-)$, ..., sim$(\vec{a}, \vec{b}_k^-)$,

StarSpace Negative Sampling Objective

- For entity a selected at current training batch:
  - positive sample: related entity b (e.g. two sentences from the same document).
  - negative samples: k unrelated entities $b_1^-, \ldots b_k^-$ (e.g. sentences in other documents).
- Compute vectors $\vec{a} = \sum_{i \in a} F_i$, $\vec{b}$, $\vec{b}_1^-, \ldots, \vec{b}_k^-$
- Compute cosine similiarities sim$(\vec{a}, \vec{b})$,sim$(\vec{a}, \vec{b}_1^-)$, ..., sim$(\vec{a}, \vec{b}_k^-)$,
- Ranking loss objective gives a reward if sim$(\vec{a}, \vec{b})$ gets a higher rank relative to the negative samples, and gives a penalty if it is lower rank.



- but StarSpace can be used for anything.
- the trained model can provide similarities between entities, between features, and between entities and features.

No social science papers with StarSpace

But many opportunities:

- embed judicial opinions as bundles of citations
- embed academic articles as bundles of citations
- embed politicians as bundles of roll call votes

Check for Understanding: True/False

1. A limitation of the Arora et al (2017) "tough-to-beat" sentence embeddings is that the vectors do not contain any information about word order.
2. Doc2Vec addresses the limits of the Arora et al (2017) embeddings by adding information on word order.
3. Unlike the other document embeddings, FastText embeddings (averaged hashed n-gram embeddings) do not have a geometric interpretation.
4. StarSpace embeddings could put judges, cases, and words into a single space.

Outline

Analyzing polarization in social media: Method and application to tweets on 21 mass shootings

Demszky, Garg, Voigt, Zou, Gentzkow, Shapiro, and Jurafsky (2019)

- Dataset:
  - tweets about 21 mass shooting events in USA, 2015-2018.
  - $N = 10,000$ (out of 4.4 million tweets from the firehose archive).
  - Party affiliation identified off of whether account follows more Democrats or Republicans
- Text partisanship:
  - measure from Gentzkow, Shapiro, and Taddy (2019) – roughly, text distance between Democrat and Republican twitter accounts.

Sentence Embeddings for Topic Assignment

- Train GloVe embeddings on tweets and create Create Arora et al (2017) embeddings:
- Cluster the embeddings using k-means
- Identify and drop hard-to-classify tweets:
    1. compute ratio of distance to closest topic and distance to second-closest topic.
    2. drop tweets above the 75th percentile.
- Validation using Amazon Mechanical Turk to choose number of clusters:
    - Identify word intruder: five from one cluster, one from another cluster.
    - Identify tweet intruder: three from one cluster, and one from another cluster.

Topic Content

Topic	10 Nearest Stems
news (19%)	break, custodi, #breakingnew, #updat, confirm, fatal, multipl, updat, unconfirm, sever
investigation (9%)	suspect, arrest, alleg, apprehend, custodi, charg, accus, prosecutor, #break, ap
shooter's identity & ideology (11%)	extremist, radic, racist, ideolog, label, rhetor, wing, blm, islamist, christian
victims & location (4%)	bar, thousand, california, calif, among, los, southern, veteran, angel, via
laws & policy (14%)	sensibl, regul, requir, access, abid, #gunreformnow, legisl, argument, allow, #guncontolnow
solidarity (13%)	affect, senseless, ach, heart, heartbroken, sadden, faculti, pray, #prayer, deepest
remembrance (6%)	honor, memori, tuesday, candlelight, flown, vigil, gather, observ, honour, capitol
other (23%)	dude, yeah, eat, huh, gonna, ain, shit, ass, damn, guess

- The embedding method resulted in more coherent topics (better MTurk validation for words and tweets) than a topic model. $k = 8$ got best coherence.
  - Appendix reports samples of tweets for each topic (but does not say how samples were selected).

Between-topic vs within-topic polarization

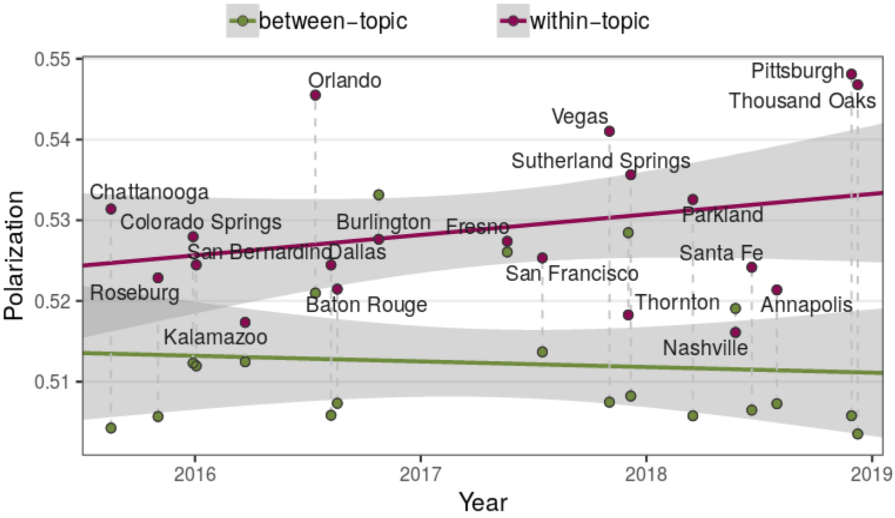
- Within-topic polarization: compute partisan text distance separately by the tweet clusters.

Between-topic vs within-topic polarization

- Within-topic polarization: compute partisan text distance separately by the tweet clusters.
- Between-topic polarization: Compute partisan text distance using cluster counts, rather than token counts.

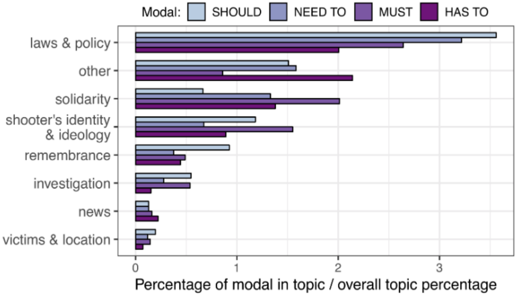
Between-topic vs within-topic polarization

- Within-topic polarization: compute partisan text distance separately by the tweet clusters.
- Between-topic polarization: Compute partisan text distance using cluster counts, rather than token counts.



Modality

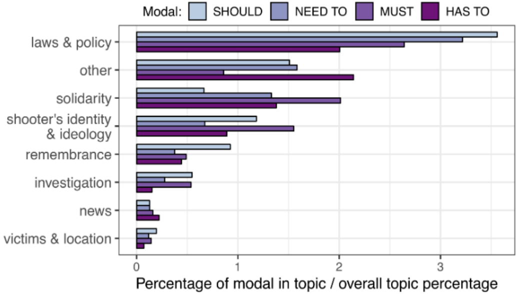
This roller coaster debate **MUST** STOP! Sensible gun ownership is one thing but assault weapons massacre innocent lives. The savagery of gore at #Parkland was beyond belief & **must** be the last.
In times of tragedy **shouldn't** we all come together?! Prayers for those harmed in the #PlannedParenthood shooting.
Communities **need to** step up and address white on white crime like the Las Vegas massacre. White men are out of control.
he BLM protest shooting, planned parenthood, now cali... domestic terrorism will crumble this country, SANE PPL **HAVE TO** FIGHT BACK
Shooting cops is horrible, cannot be condoned. But **must be** understood these incidents are outgrowth of decades of police abuses. #BatonRouge
1. Islamic terrorists are at war with us 2. Gun free zones = kill zones 3. Americans **should be** allowed to defend themselves #Chattanooga
Las Vegas shooting Walmart shooting and now 25 people killed in Texas over 90 people killed Mexico **should** build that wall to keep the US out
CNN reporting 20 dead, 42 injured in Orlando night club shooting. Just awful. The US **must** act to control guns or this carnage will continue.



- Count the four most frequent necessity modals in the data: should, must, have to, need to.
  - in this context, they are used as calls to action.

Modality

This roller coaster debate **MUST** STOP! Sensible gun ownership is one thing but assault weapons massacre innocent lives. The savagery of gore at #Parkland was beyond belief & **must** be the last.
In times of tragedy **shouldn't** we all come together?! Prayers for those harmed in the #PlannedParenthood shooting.
Communities **need to** step up and address white on white crime like the Las Vegas massacre. White men are out of control.
he BLM protest shooting, planned parenthood, now cali... domestic terrorism will crumble this country, SANE PPL **HAVE TO** FIGHT BACK
Shooting cops is horrible, cannot be condoned. But **must be** understood these incidents are outgrowth of decades of police abuses. #BatonRouge
1. Islamic terrorists are at war with us 2. Gun free zones = kill zones 3. Americans **should be** allowed to defend themselves #Chattanooga
Las Vegas shooting Walmart shooting and now 25 people killed in Texas over 90 people killed Mexico **should** build that wall to keep the US out
CNN reporting 20 dead, 42 injured in Orlando night club shooting. Just awful. The US **must** act to control guns or this carnage will continue.



- Count the four most frequent necessity modals in the data: should, must, have to, need to.
  - in this context, they are used as calls to action.
- Democrats use modals more than Republicans; Republicans seem more fatalistic.

Partisanship of Topics, by Race of Shooter

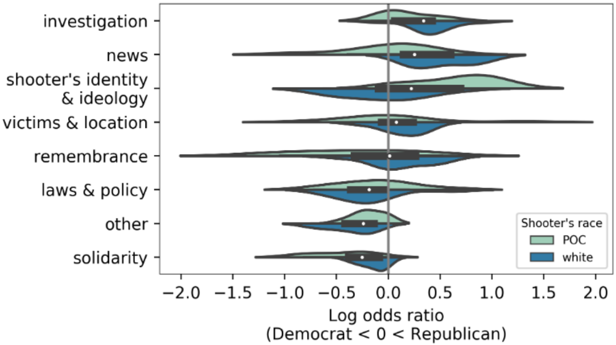


Figure 7: The plot shows the kernel density of the partisan log odds ratios of each topic (one observation per event). The white points show the median and the black rectangles the interquartile range across events.

Partisan Framing Devices: Words

- Partisanship of phrases from supervised model:

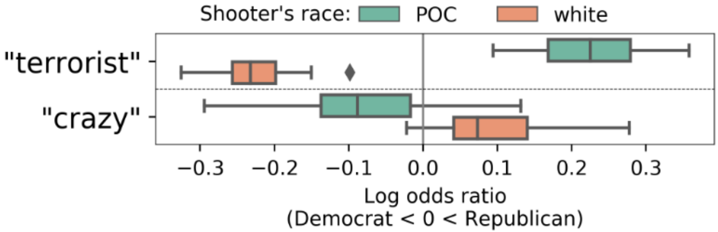


Figure 8: The log odds ratios of "terrorist" and "crazy" across events, grouped by the shooter's race. The boxes show the interquartile range and the diamond an outlier.

- Partisan valence of "terrorist" and "crazy" flip depending on race of shooter (these words have the largest racial difference in the joint vocabulary).

Affect (Emotions)

- Starting point: Emotion lexicon from Mohammad and Turney (2013), available at `saifmohammad.com`.
  - 14,182 words assigned to sentiment (positive/negative) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust).

Affect (Emotions)

- Starting point: Emotion lexicon from Mohammad and Turney (2013), available at `saifmohammad.com`.
  - 14,182 words assigned to sentiment (positive/negative) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust).
- Domain propagation (Hamilton et al 2018):
  - pick 5-11 representative words per emotion category (Appendix E)
  - for each word in vocabulary, compute average distance to each member of each category. take 30 closest words as lexicon.

Affect (Emotions)

- Starting point: Emotion lexicon from Mohammad and Turney (2013), available at `saifmohammad.com`.
  - 14,182 words assigned to sentiment (positive/negative) and emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust).
- Domain propagation (Hamilton et al 2018):
  - pick 5-11 representative words per emotion category (Appendix E)
  - for each word in vocabulary, compute average distance to each member of each category. take 30 closest words as lexicon.

sadness senseless, loss, tragedi, lost, devast, sad, love, griev, horrif, terribl, pain, violenc, condol, broken, hurt, feel, victim, mourn, horrifi, will, grief, ach, suffer, sick, kill, aw, sicken, evil, massacr, mad

disgust disgust, sick, shame, ignor, wrong, blame, hell, ridicul, idiot, murder, evil, coward, sicken, feel, disgrac, slaughter, action, bad, insan, attack, pathet, outrag, polit, terrorist, mad, damn, lose, shit, lie, asshol

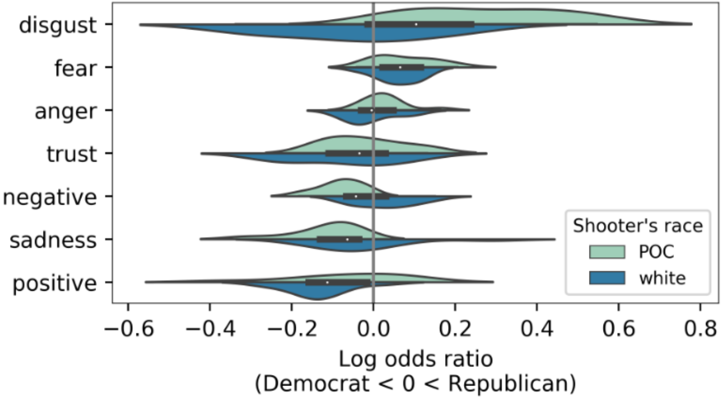
anger gun, will, murder, kill, violenc, wrong, shoot, bad, death, attack, feel, shot, action, arm, idiot, crazi, crimin, terrorist, mad, hell, crime, blame, fight, ridicul, insan, shit, die, threat, terror, hate

fear danger, threat, fear, arm, gun, still, shooter, attack, feel, fight, hide, murder, shot, shoot, bad, kill, chang, serious, violenc, forc, risk, defend, warn, govern, concern, fail, polic, wrong, case, terrorist

trust school, like, good, real, secur, show, nation, don, protect, call, teacher, help, law, great, save, true, wonder, respons, sad, answer, person, feel, safe, thought, continu, love, guard, church, fact, support

Partisanship of Affect Categories

- Compute partisanship scores using affect-category counts:



- Disgust affect flips along partisan lines depending on race of shooter.

We provide an NLP framework to uncover four linguistic dimensions of political polarization in social media: topic choice, framing, affect and illocutionary force. We quantify these aspects with existing lexical methods, and propose clustering of tweet embeddings as a means to identify salient topics for analysis across events; human evaluations show that our approach generates more cohesive topics than traditional LDA-based models. We apply our methods to study 4.4M tweets on 21 mass shootings. We provide evidence that the discussion of these events is highly polarized politically and that this polarization is primarily driven by partisan differences in framing rather than topic choice. We identify framing devices, such as grounding and the contrasting use of the terms "terrorist" and "crazy", that contribute to polarization. Results pertaining to topic choice, affect and illocutionary force suggest that Republicans focus more on the shooter and event-specific facts (news) while Democrats focus more on the victims and call for policy changes. Our work contributes to a deeper understanding of the way group divisions manifest in language and to computational methods for studying them.

1. What is the research question?
2. What is the problem solved?
3. What is being measured?
4. How does the measurement help answer the research question?