Sequencing Legal DNA

NLP for Law and Political Economy

9. Document Embeddings

Objectives

1. **What is the research question?**

Objectives

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata

Objectives

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Embedding phrases and documents:
   - **What are we trying to measure?**

Objectives

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Embedding phrases and documents:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model is measuring what we want.

Objectives

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Embedding phrases and documents:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model is measuring what we want.
4. Empirical analysis
   - Produce statistics or predictions with the trained model.
   - **Answer the research question.**

Vectorizing Documents

- Quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- We started with the baseline approach: documents become sparse vectors of token counts/frequencies.

Vectorizing Documents

- Quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- We started with the baseline approach: documents become sparse vectors of token counts/frequencies.
  - high-dimensionality can cause issues, but sparsity mitigates.
  - can use documents of arbitrary length
  - can capture local word order with n-grams, but long-run word order is lost.

Embedding layers

- Previously, we introduced embedding layers:
  - take the whole document as input, pad documents to the same length, and represent the document as a flattened series of embedding vectors.

Embedding layers

- Previously, we introduced embedding layers:
  - take the whole document as input, pad documents to the same length, and represent the document as a flattened series of embedding vectors.
  - potentially captures information on long-range ordering of features in documents
  - DNNs work better with dense vectors
  - computationally demanding
  - only works with short documents

Outline

Outline

From Word Vectors to 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.
- Wieting et al (2016) find that this simple representation often out-performs complex recurrent architectures, for example on sentence entailment.

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

- Can filter tokens:
  - drop stopwords
  - filter on parts of speech (e.g., keep only nouns, adjectives, and verbs)

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

- Can filter tokens:
  - drop stopwords
  - filter on parts of speech (e.g., keep only nouns, adjectives, and verbs)
- Token weighting:
  - set a_w to weight words by inverse term frequency or inverse document frequency (that is, up-weight rare/informative words)

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

- Can filter tokens:
  - drop stopwords
  - filter on parts of speech (e.g., keep only nouns, adjectives, and verbs)
- Token weighting:
  - set a_w to weight words by inverse term frequency or inverse document frequency (that is, up-weight rare/informative words)
  - **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.

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

- Can filter tokens:
  - drop stopwords
  - filter on parts of speech (e.g., keep only nouns, adjectives, and verbs)
- Token weighting:
  - set a_w to weight words by inverse term frequency or inverse document frequency (that is, up-weight rare/informative words)
  - **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.
- Can also normalize weights to sum to one:

$$\vec{D} = \frac{1}{\sum_{w \in D} a_w} \sum_{w \in D} a_w \vec{w}$$

Outline

- An interesting factor in political psychology is the role of **cogition and emotion** in political messaging.
  - What works better: a **logical argument** or an **emotional appeal**?

- An interesting factor in political psychology is the role of **cogition and emotion** in political messaging.
  - What works better: a **logical argument** or an **emotional appeal**?
- In this paper, we build a new measure of cognitive/emotive valence in language and apply it to speeches of U.S. Congress members.

Cognitive/Affective Dictionary

- Dictionary: a new domain-appropriate list of words for:
  - **Cognitive Processing (“thinking”)**: insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusion, and exclusion
  - **Affective Processing (“feeling”):** positive and negative emotions, pleasure, pain, happiness, anxiety, anger, and sadness.

Cognitive/Affective Dictionary

- Dictionary: a new domain-appropriate list of words for:
  - **Cognitive Processing ("thinking")**: insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusion, and exclusion
  - **Affective Processing ("feeling"):** positive and negative emotions, pleasure, pain, happiness, anxiety, anger, and sadness.
- Drawn from LIWC, but many false positives removed (e.g., "admir*" matches to "admiral", so that's dropped).

Word Embeddings and Metric

- We train word2vec on all speeches in U.S. House of Representatives and Senate for the years 1994-2014.

Word Embeddings and Metric

- We train word2vec on all speeches in U.S. House of Representatives and Senate for the years 1994-2014.
- We get the embeddings for each word in our dictionaries, and take the centroid (average) for affective $\vec{A}$ and cognitive $\vec{C}$.

Word Embeddings and Metric

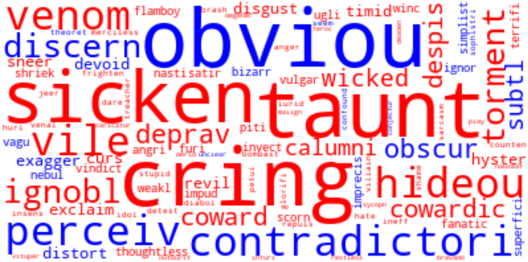
- We train word2vec on all speeches in U.S. House of Representatives and Senate for the years 1994-2014.
- We get the embeddings for each word in our dictionaries, and take the centroid (average) for affective $\vec{A}$ and cognitive $\vec{C}$.
- Next: let $\vec{d}_i$ be the SIF-weighted average of the embeddings for the words in document i

Word Embeddings and Metric

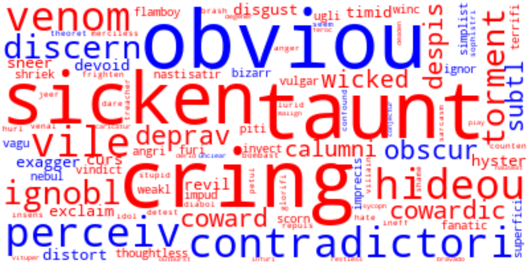
- We train word2vec on all speeches in U.S. House of Representatives and Senate for the years 1994-2014.
- We get the embeddings for each word in our dictionaries, and take the centroid (average) for affective $\vec{A}$ and cognitive $\vec{C}$.
- Next: let $\vec{d}_i$ be the SIF-weighted average of the embeddings for the words in document i
- Then: the measure of relative emotionality in i is

$$Y_i = \frac{\text{sim}(\vec{d}_i, \vec{A}) + 1}{\text{sim}(\vec{d}_i, \vec{C}) + 1}$$

where sim() is cosine similarity.

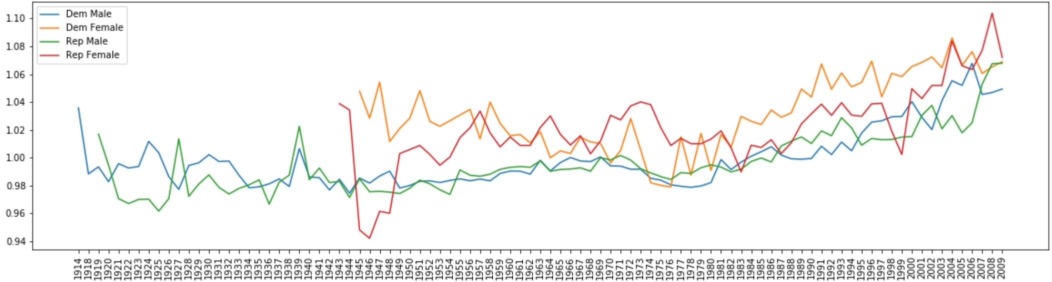


- Top cognitive sentences:
  - "In my judgment, neither is true in the case of this amendment."
  - “Is that correct?”
  - “R. 15 contains a provision that is similar but, in fact, broader in scope.”



- Top cognitive sentences:
  - "In my judgment, neither is true in the case of this amendment."
  - "Is that correct?"
  - "R. 15 contains a provision that is similar but, in fact, broader in scope."
- Top emotional sentences:
  - "There is nothing to trouble any heart, nothing to hurt at all; death is only a quiet door, in an old garden wall."
  - "With joy in his heart and a smile on his face he graced practically every social occasion with a song."
  - "We Democrats may disagree, but we love our fellow men and we never hate them."

Emotional language in the very long run



Outline

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

Demszky, Garg, Voigt, Zou, Gentzkow, Shapiro, and Jurafsky

- Research Object:
  - use NLP to understand four "dimensions" of social media polarization: topic choice, framing, affect, modality.

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

Demszky, Garg, Voigt, Zou, Gentzkow, Shapiro, and Jurafsky

- Research Object:
  - use NLP to understand four "dimensions" of social media polarization: topic choice, framing, affect, modality.
- Context:
  - tweets in response to mass shooting events.

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

Demszky, Garg, Voigt, Zou, Gentzkow, Shapiro, and Jurafsky

- Research Object:
  - use NLP to understand four "dimensions" of social media polarization: topic choice, framing, affect, modality.
- Context:
  - tweets in response to mass shooting events.
- Research question:
  - does political partisanship manifest in polarized responses to violent/polarizing events.

Dataset

- 21 mass shooting events, 2015-2018, from Gun Violence Archive

Dataset

- 21 mass shooting events, 2015-2018, from Gun Violence Archive
- tweets about those events, identified by:
  - location keywords (e.g. chattanooga, roseburg, san bernardino, fresno, etc.)
  - event keywords (lemmas): shoot, gun, kill, attack, massacre, victim

Dataset

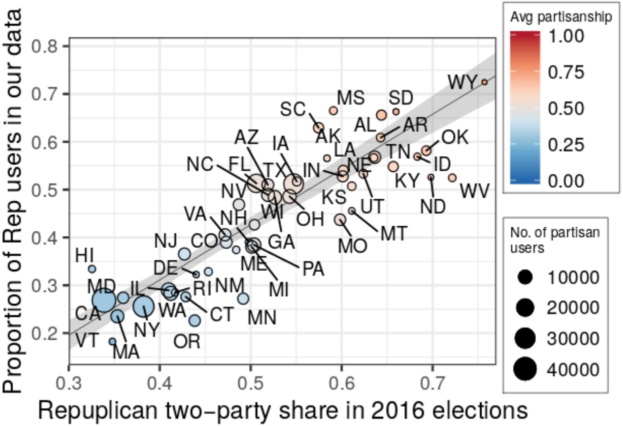
- 21 mass shooting events, 2015-2018, from Gun Violence Archive
- tweets about those events, identified by:
  - location keywords (e.g. chattanooga, roseburg, san bernardino, fresno, etc.)
  - event keywords (lemmas): shoot, gun, kill, attack, massacre, victim
  - filter out retweets and tweets from deactivated accounts
  - $N = 10,000$ (out of 4.4 million tweets from the firehose archive).

Identifying party affiliation of Twitter users

- Party affiliation identified off of whether you follow more Democrats or Republicans, from a list of Twitter accounts associated with legislators, presidential candidates, and party organizations (Volkova et al 2014).
  - at least 51% of tweets for each event can be assigned partisanship this way.

Identifying party affiliation of Twitter users

- Party affiliation identified off of whether you follow more Democrats or Republicans, from a list of Twitter accounts associated with legislators, presidential candidates, and party organizations (Volkova et al 2014).
  - at least 51% of tweets for each event can be assigned partisanship this way.
- For geolocated users this matches up pretty well with party vote shares by state ($R^2 = .82$):



Pre-processing for partisanship

- Stemming and stopword removal.
- Event-specific vocabulary:
  - unigrams and bigrams
  - occur in event's tweets at least 50 times
  - must be used by at least two tweeters.

Partisanship

- Leave-one-out estimator from Gentzkow et al (2019), applied to each shooting event:

$$\pi = \frac{1}{2}\left(\frac{1}{|D|}\sum_{i \in D} \hat{\boldsymbol{q}}_i \cdot \hat{\boldsymbol{\rho}}_{-i} + \frac{1}{|R|}\sum_{i \in R} \hat{\boldsymbol{q}}_i \cdot (1 - \hat{\boldsymbol{\rho}}_{-i})\right)$$

Partisanship

- Leave-one-out estimator from Gentzkow et al (2019), applied to each shooting event:

$$\pi = \frac{1}{2}\left(\frac{1}{|D|}\sum_{i\in D}\hat{\boldsymbol{q}}_i \cdot \hat{\boldsymbol{\rho}}_{-i} + \frac{1}{|R|}\sum_{i\in R}\hat{\boldsymbol{q}}_i \cdot (1-\hat{\boldsymbol{\rho}}_{-i})\right)$$

  - $\hat{\boldsymbol{q}}_i$ = token frequencies for user i, drawn from set of democrats D and set of republicans R

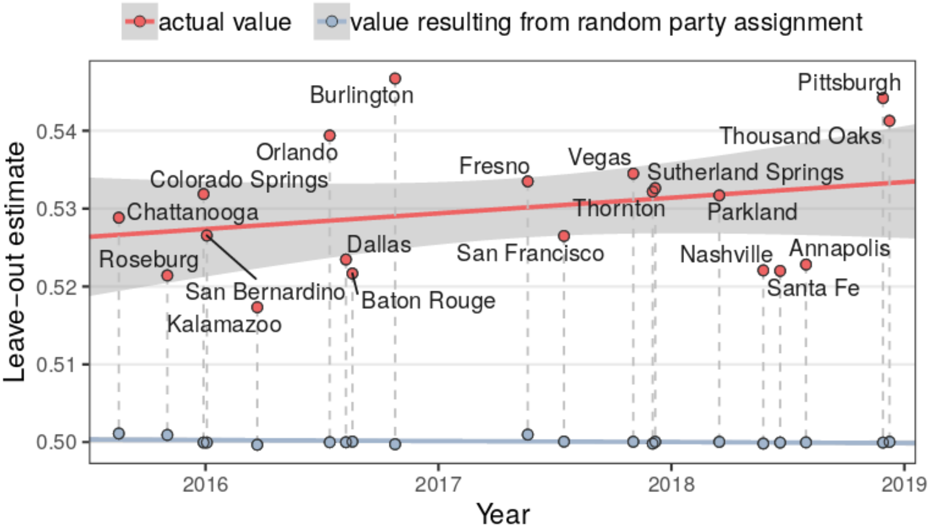
Partisanship

- Leave-one-out estimator from Gentzkow et al (2019), applied to each shooting event:
$$\pi = \frac{1}{2}\left(\frac{1}{|D|}\sum_{i\in D}\hat{\boldsymbol{q}}_i \cdot \hat{\boldsymbol{\rho}}_{-i} + \frac{1}{|R|}\sum_{i\in R}\hat{\boldsymbol{q}}_i \cdot (1-\hat{\boldsymbol{\rho}}_{-i})\right)$$
  - $\hat{\boldsymbol{q}}_i$ = token frequencies for user i, drawn from set of democrats D and set of republicans R
  - $\hat{\boldsymbol{\rho}}_{-i}$ has elements
$$\rho_{-i} = \frac{q_i^D}{q_i^D + q_i^R}$$
    empirical poster probabilities computed from all other users.

Partisanship

- Leave-one-out estimator from Gentzkow et al (2019), applied to each shooting event:
$$\pi = \frac{1}{2}\Big(\frac{1}{|D|}\sum_{i\in D}\hat{\boldsymbol{q}}_i \cdot \hat{\boldsymbol{\rho}}_{-i} + \frac{1}{|R|}\sum_{i\in R}\hat{\boldsymbol{q}}_i \cdot (1-\hat{\boldsymbol{\rho}}_{-i})\Big)$$
  - $\hat{\boldsymbol{q}}_i$ = token frequencies for user i, drawn from set of democrats D and set of republicans R
  - $\hat{\boldsymbol{\rho}}_{-i}$ has elements
  $$\rho_{-i} = \frac{q_i^D}{q_i^D + q_i^R}$$
  empirical poster probabilities computed from all other users.
- π is an estimate for expected posterior probability that a Bayesian observer would correctly predict party after observing one randomly sampled token.
  - consistency assumes tokens are drawn from multinomial logit.

Tweets about mass shootings are polarized



- comparable to $\pi = .53$ in Congressional speeches (GST 2019).
- The increase in polarization over time is not statistically significant.

Questions

- How polarized are tweets about other topics (not mass shootings)?
  - why not use a tweeter fixed effect and compare to their other tweets?
  - why not show pre-trends in polarization?

Questions

- How polarized are tweets about other topics (not mass shootings)?
  - why not use a tweeter fixed effect and compare to their other tweets?
  - why not show pre-trends in polarization?
- Can show polarization separately by party?

Questions

- How polarized are tweets about other topics (not mass shootings)?
  - why not use a tweeter fixed effect and compare to their other tweets?
  - why not show pre-trends in polarization?
- Can show polarization separately by party?
- Validating π:
  - How accurate is π at the individual level?
  - Where is the binscatter of π versus actual party affiliation?

Sentence Embeddings for Topic Assignment

1. Make a new vocabulary:
   1.1 Sample 10,000 tweets from each event
   1.2 vocabulary of stemmed words occuring at least ten times in at least three events ($N = 2000$)

Sentence Embeddings for Topic Assignment

1. Make a new vocabulary:
    1.1 Sample 10,000 tweets from each event
    1.2 vocabulary of stemmed words occuring at least ten times in at least three events ($N = 2000$)
2. Train GloVe embeddings on random samples of tweets from each event (samples were different sizes, this is not explained)

Sentence Embeddings for Topic Assignment

1. Make a new vocabulary:
   1.1 Sample 10,000 tweets from each event
   1.2 vocabulary of stemmed words occuring at least ten times in at least three events ($N = 2000$)
2. Train GloVe embeddings on random samples of tweets from each event (samples were different sizes, this is not explained)
3. Create Arora et al (2017) embeddings:
   3.1 for each tweet t, compute weighted average vectors v_t for each word, weighted by inverse frequency.
   3.2 take out first principal component of matrix whose rows are v_t

Topics = Embedding Clusters

1. Cluster the embeddings using k-means

Topics = Embedding Clusters

1. Cluster the embeddings using k-means
2. Identify and drop hard-to-classify tweets:
   2.1 compute ratio of distance to closest topic and distance to second-closest topic.
   2.2 drop tweets above the 75th percentile.

Topics = Embedding Clusters

1. Cluster the embeddings using k-means
2. Identify and drop hard-to-classify tweets:
   2.1 compute ratio of distance to closest topic and distance to second-closest topic.
   2.2 drop tweets above the 75th percentile.

- Validation using Amazon Mechanical Turk:
  - 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

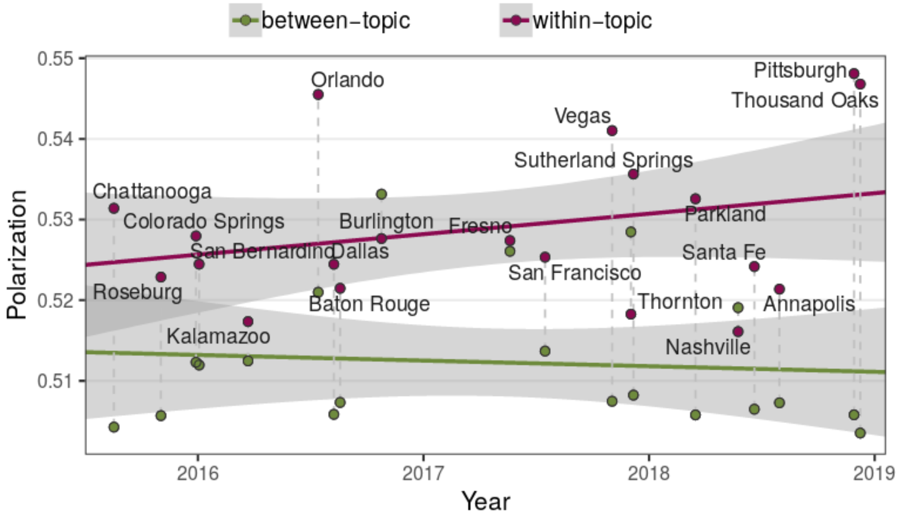
- Within-topic polarization: compute π separately by the tweet clusters.

Between-topic vs within-topic polarization

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

Between-topic vs within-topic polarization

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



Trends in within-topic polarization

- Most polarized topics: shooter's identity & ideology (.55), laws & policy (.54)

Figure 6: Las Vegas within-topic polarization in the days after the event. The bar charts show the proportion of each topic in the data at a given time.

- "measuring polarization of topics for other events over time is noisy".

Partisanship of Topics, by Race of Shooter

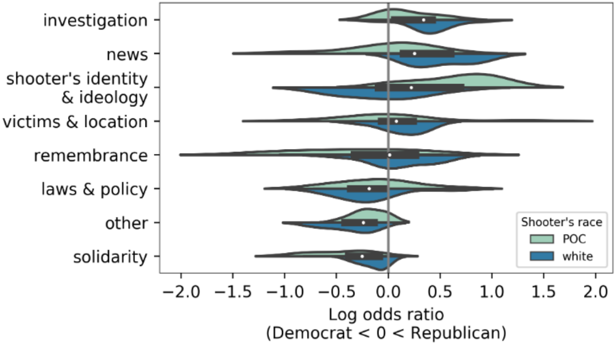


Figure 7: The plot shows the kernel density of the partisan log odds ratios of each topic (one observation per event). The white points show the median and the black rectangles the interquartile range across events.

Partisan Framing Devices: Words

- Partisanship of phrases from GST model:

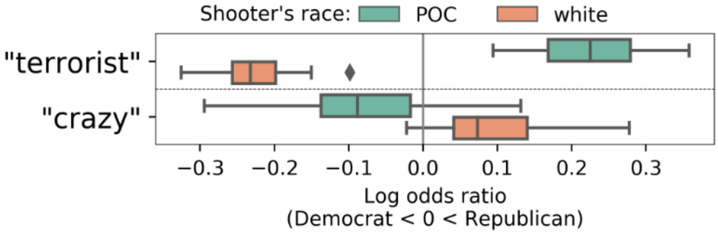
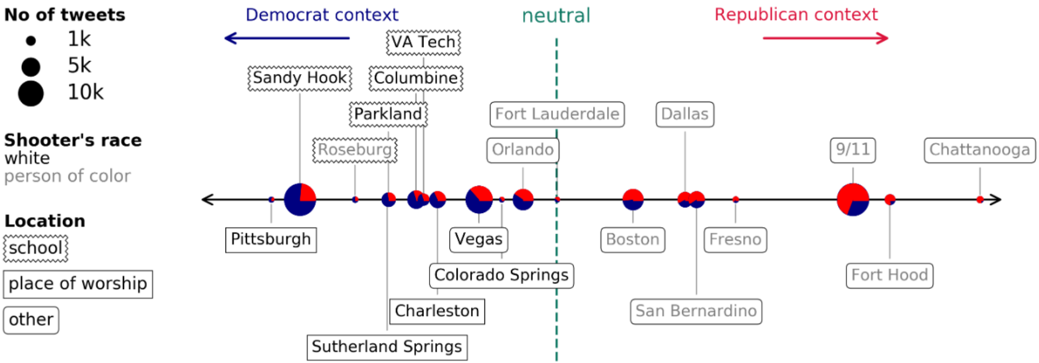


Figure 8: The log odds ratios of "terrorist" and "crazy" across events, grouped by the shooter's race. The boxes show the interquartile range and the diamond an outlier.

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

Partisan Framing Devices: Events

- Partisanship of keywords for previous events from GST model:



- Democrats invoke white shooters, Republicans invoke POC shooters.

Affect

- 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

- 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:
  - 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

- 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:
  - 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

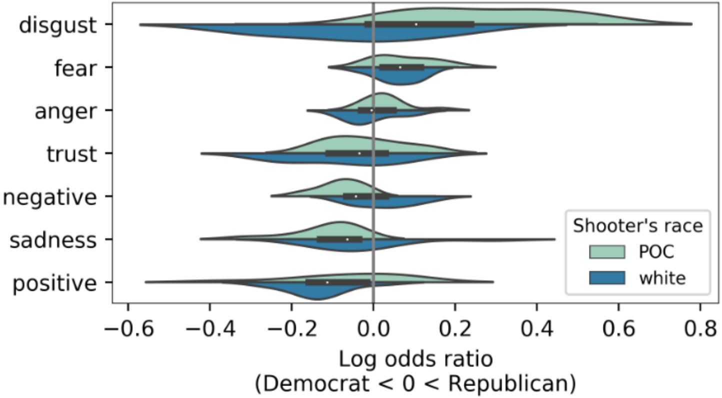
anger gun, will, murder, kill, violenc, wrong, shoot, bad, death, attack, feel, shot, action, arm, idiot, crazi, crimin, terrorist, mad, hell, crime, blame, fight, ridicul, insan, shit, die, threat, terror, hate

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

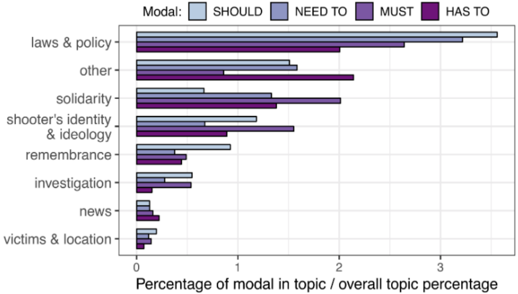
- Compute GST partisanship scores using affect-category counts:



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

Modality

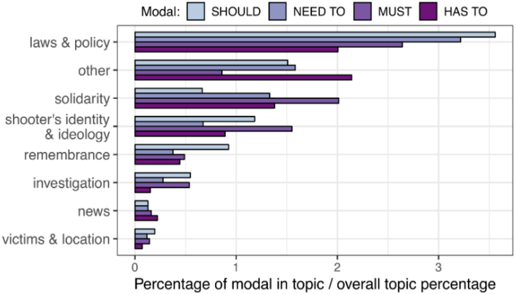
This roller coaster debate **MUST** STOP! Sensible gun ownership is one thing but assault weapons massacre innocent lives. The savagery of gore at #Parkland was beyond belief & **must** be the last.
In times of tragedy **shouldn't** we all come together?! Prayers for those harmed in the #PlannedParenthood 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 are more fatalistic.

Comments

- This is an impressive array of NLP tools aimed at the same research question.
  - could be moving toward a standard for analyzing interpretable dimension in language.

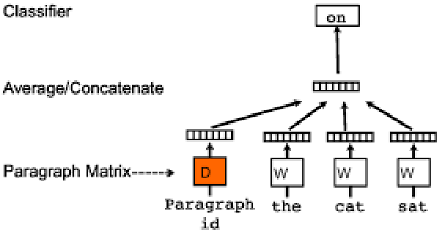
Comments

- This is an impressive array of NLP tools aimed at the same research question.
  - could be moving toward a standard for analyzing interpretable dimension in language.
- For all outcomes, would help to compare to other types of events, and to show pre-trends.
  - there is no baseline for polarization for comparison.
  - they do not distinguish whether outcomes are driven by different people selecting into tweeting, vs within-user changes.

Outline

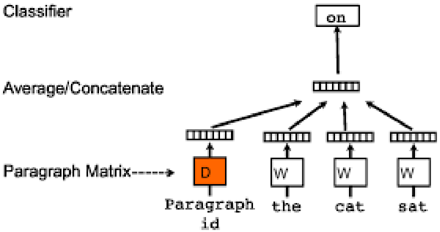
Outline

Doc2Vec (Le and Mikolov)



- Doc2Vec generalizes Word2Vec to documents:
  - predict a word using both the immediate neighbors, as well as **a bag-of-words representation of the whole document**.

Doc2Vec (Le and Mikolov)



- Doc2Vec generalizes Word2Vec to documents:
  - predict a word using both the immediate neighbors, as well as **a bag-of-words representation of the whole document**.
- In Doc2Vec, both words **and documents** are assigned a learned vector representation through an embedding layer.

Document Embeddings Geometry

- Just as directions in word space encode semantic information about the words, directions in document space encode topical information about the documents.

Document Embeddings Geometry

- Just as directions in word space encode semantic information about the words, directions in document space encode topical information about the documents.
- In topic models, each dimension has a topical interpretation; in document embeddings, a direction (might) have a topical interpretation.

Doc2Vec in gensim

- syntax is the same as Word2Vec.
- can train both document vectors and word vectors.
- can get similarity between documents, and use clustering to get groups of related documents.

Outline

Doc2Vec on Wikipedia

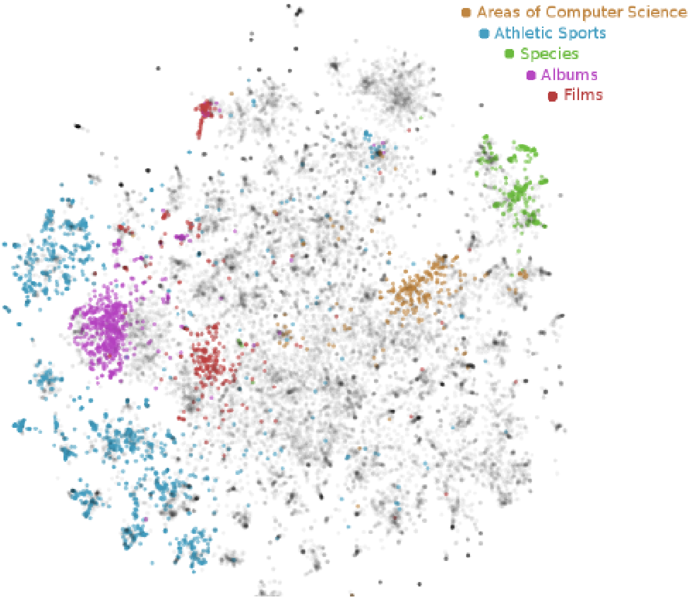


Figure 3: Visualization of Wikipedia paragraph vectors using t-SNE.

Table 1: Nearest neighbours to "Machine learning." Bold face texts are articles we found unrelated to "Machine learning." We use Hellinger distance for LDA and cosine distance for Paragraph Vectors as they work the best for each model.

LDA	**Paragraph Vectors**
Artificial neural network	Artificial neural network
Predictive analytics	Types of artificial neural networks
Structured prediction	Unsupervised learning
Mathematical geophysics	Feature learning
Supervised learning	Predictive analytics
Constrained conditional model	Pattern recognition
Sensitivity analysis	Statistical classification
SXML	Structured prediction
Feature scaling	Training set
Boosting (machine learning)	Meta learning (computer science)
Prior probability	Kernel method
Curse of dimensionality	Supervised learning
Scientific evidence	Generalization error
Online machine learning	Overfitting
N-gram	Multi-task learning
Cluster analysis	Generative model
Dimensionality reduction	Computational learning theory
Functional decomposition	Inductive bias
Bayesian network	Semi-supervised learning

Table 5: arXiv nearest neighbours to "Distributed Representations of Sentences and Documents" using Paragraph Vectors.

Title	Cosine Similarity
Evaluating Neural Word Representations in Tensor-Based Compositional Settings	0.771
Polyglot: Distributed Word Representations for Multilingual NLP	0.764
Lexicon Infused Phrase Embeddings for Named Entity Resolution	0.757
A Convolutional Neural Network for Modelling Sentences	0.747
Distributed Representations of Words and Phrases and their Compositionality	0.740
Convolutional Neural Networks for Sentence Classification	0.735
SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation	0.735
Exploiting Similarities among Languages for Machine Translation	0.731
Efficient Estimation of Word Representations in Vector Space	0.727
Multilingual Distributed Representations without Word Alignment	0.721

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

Table 7: arXiv nearest neighbours to "Distributed Representations of Sentences and Documents" - "neural" + "Bayesian". I.e., the Bayesian equivalence of the Paragraph Vector paper.

Title	**Cosine Similarity**
Content Modeling Using Latent Permutations	0.629
SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation	0.611
Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA	0.579
Evaluating Neural Word Representations in Tensor-Based Compositional Settings	0.572
Syntactic Topic Models	0.548
Training Restricted Boltzmann Machines on Word Observations	0.548
Discrete Component Analysis	0.547
Resolving Lexical Ambiguity in Tensor Regression Models of Meaning	0.546
Measuring political sentiment on Twitter: factor-optimal design for multinomial inverse regression	0.544
Scalable Probabilistic Entity-Topic Modeling	0.541

Outline

Document Vectors for Judicial Opinions

- Ash and Chen (2018) produce document vectors for each case to understand differences between judges and courts.
  - Corpus: 300,000 cases from U.S. Circuit Courts, 1870-2010.

Document Vectors for Judicial Opinions

- Ash and Chen (2018) produce document vectors for each case to understand differences between judges and courts.
  - Corpus: 300,000 cases from U.S. Circuit Courts, 1870-2010.
- We 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

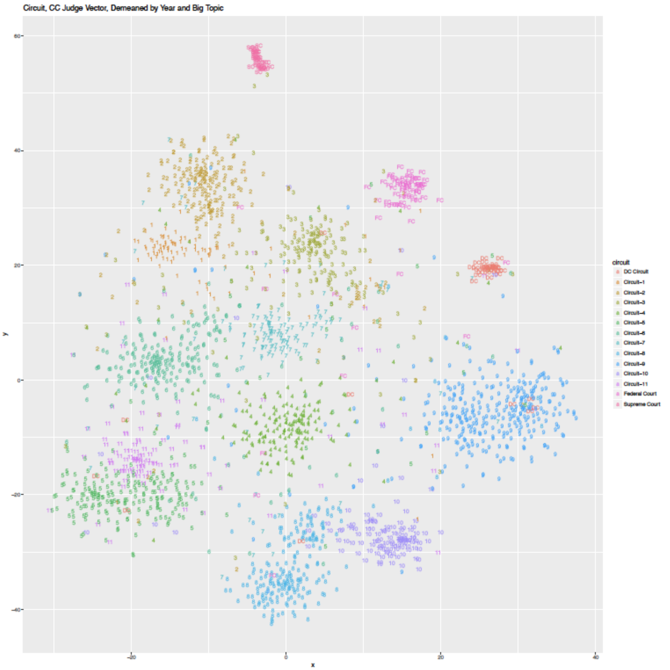


Figure 2: Centered by Court-Topic, Averaged by Court-Year, Labeled by Decade

tsne plot

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

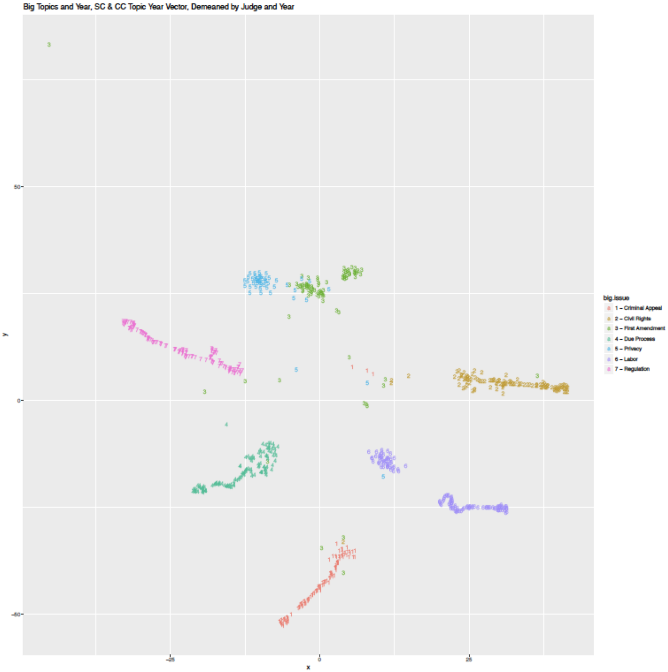
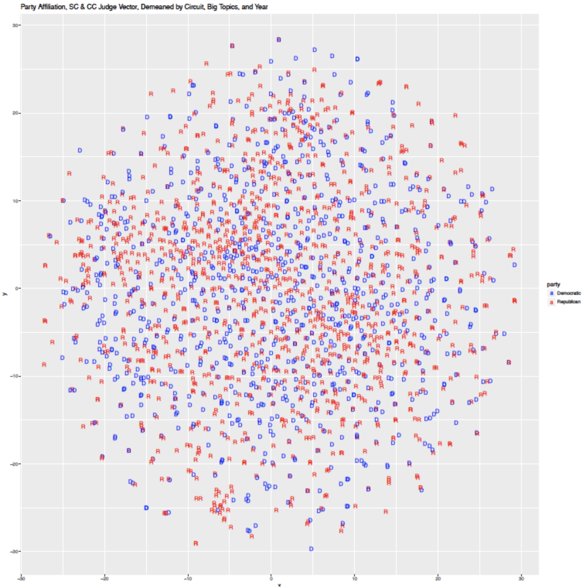


Figure 5: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Judge Birth Cohort

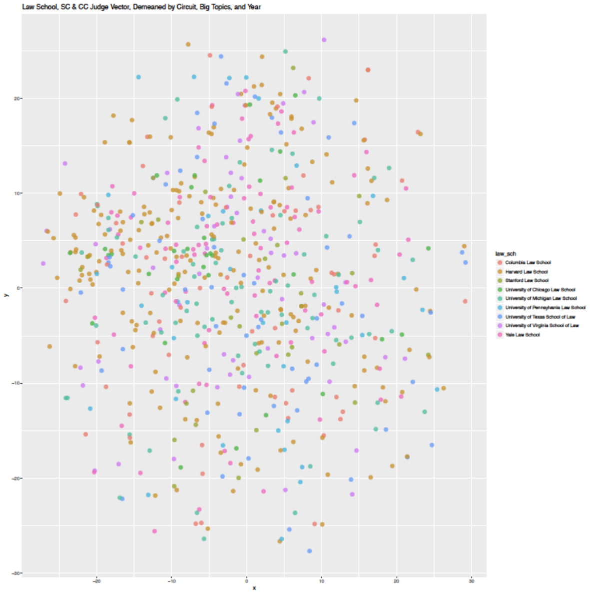
tsne plot

Figure 4: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Political Party



tsne plot

Figure 6: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Law School Attended



Relatedness between judges (e.g. Richard Posner)

Circuit Judge Name	Similarity	Rank	Circuit Judge Name	Similarity	Rank
POSNER, RICHARD A.	1.000	1	TONE, PHILIP W.	0.459	16
EASTERBROOK, FRANK H.	0.663	2	SIBLEY, SAMUEL	0.459	17
SUTTON, JEFFREY S.	0.620	3	SCALIA, ANTONIN	0.456	18
NOONAN, JOHN T.	0.596	4	COLLOTON, STEVEN M.	0.445	19
NELSON, DAVID A.	0.592	5	DUNIWAY, BENJAMIN	0.438	20
CARNES, EDWARD E.	0.567	6	GIBBONS, JOHN J.	0.422	21
FRIENDLY, HENRY	0.566	7	BOGGS, DANNY J.	0.420	22
KOZINSKI, ALEX	0.563	8	BREYER, STEPHEN G.	0.414	23
GORSUCH, NEIL M.	0.559	9	GOODRICH, HERBERT	0.412	24
CHAMBERS, RICHARD H.	0.546	10	LOKEN, JAMES B.	0.410	25
FERNANDEZ, FERDINAND F.	0.503	11	WEIS, JOSEPH F.	0.408	26
EDMONDSON, JAMES L.	0.501	12	SCALIA, ANTONIN (SCOTUS)	0.406	27
KLEINFELD, ANDREW J.	0.491	13	BOUDIN, MICHAEL	0.403	28
WILLIAMS, STEPHEN F.	0.481	14	RANDOLPH, A. RAYMOND	0.397	29
KETHLEDGE, RAYMOND M.	0.459	15	MCCONNELL, MICHAEL W.	0.390	30

Document vectors demeaned by court, year, and topic, then aggregated by judge.

Outline

Doc2Vec for Judge Sentiment

- ...

Outline

ELMO

- ...

Outline

Autoencoder Encodings

- A recurrent autoencoder compressed a document (e.g. a sentence) into a vector to be reconstructed.
  - Can use the compressed representation as a document embedding.
- But these embeddings don't tend to work well for sentence similarity tasks because autoencoders try to reproduce the specific wording (reconstruction objective), rather than the conceptual meaning.

Machine Translation

- Machine translators produce a sentence vector that must be decoded into another language.
  - if the vector produces a good translation, it must contain the important information in the sentence.
- The Multilingual Universal Sentence Encoder produces precisely this type of embedding.

Skip-Thought Embeddings

- Kiros et al (2015), drawing on the intuition of skip-gram embeddings in word2vec, produce sentence embeddings for a sentence prediction task.
  - the encoder vectorizes a sentence, and the decoder tries to reproduce the next sentence.
  - could also do negative sampling: produce embeddings to guess whether two sentences are in the same paragraph.

Skip-Thought Embeddings

- Kiros et al (2015), drawing on the intuition of skip-gram embeddings in word2vec, produce sentence embeddings for a sentence prediction task.
  - the encoder vectorizes a sentence, and the decoder tries to reproduce the next sentence.
  - could also do negative sampling: produce embeddings to guess whether two sentences are in the same paragraph.
- Works really well in practice (Goldberg 2017, p 204), e.g. assigning similar vectors to
  - he ran his hand inside his coat, double-checking that the unopened letter was still there.
  - he slipped his hand between his coat and his shirt, where the folded copies lay in a brown envelope.

StarSpace

- ...

FastText Embeddings

- ...

Universal Sentence Encoder

- ...

BERT

- …