Text as Data for the Social Sciences (II)

Germain Gauthier, Felix Lennert, Etienne Ollion

germain.gauthier@polytechnique.edu

SICSS Paris

June 2022

Recap of the Last Lecture

- General pipeline for text-as-data

  Raw corpus $\rightarrow$ Featurization $\rightarrow$ Measurement $\rightarrow$ Insights

Recap of the Last Lecture

- General pipeline for text-as-data

  Raw corpus $\rightarrow$ Featurization $\rightarrow$ Measurement $\rightarrow$ Insights

- A common featurization is the bag-of-words representation, that boils down to counting words.
- Easy, transparent, it has led to tens of thousands of publications in the social sciences.

Recap of the Last Lecture

- General pipeline for text-as-data

  Raw corpus $\rightarrow$ Featurization $\rightarrow$ Measurement $\rightarrow$ Insights

- A common featurization is the bag-of-words representation, that boils down to counting words.
- Easy, transparent, it has led to tens of thousands of publications in the social sciences.
- Nonetheless, it also comes with some limitations.

Limitations of Bag-of-words

- Bag-of-words basically disregards **context and syntax**.

Limitations of Bag-of-words

- Bag-of-words basically disregards **context and syntax**.

  Example 1

  "The terrorists killed American soldiers."

  "American soldiers killed the terrorists."

  $\rightarrow$ These two sentences have the same representation!

Limitations of Bag-of-words

- Bag-of-words basically disregards **context and syntax**.

  Example 1

  "The terrorists killed American soldiers."

  "American soldiers killed the terrorists."

  → These two sentences have the same representation!

  Example 2

  "This is a sentient being."

  "I love being the way I am."

  → "being" has different meanings and grammatical functions, but the same representation.

Limitations of Bag-of-words

- No notion of **distance** between tokens.

  For example, *"hi"* and *"hello"* are not considered more similar to one another than to *"sociology"*.

  In fact, tokens are all at the same distance of one another.

Limitations of Bag-of-words

- No notion of **distance** between tokens.

  For example, *"hi"* and *"hello"* are not considered more similar to one another than to *"sociology"*.

  In fact, tokens are all at the same distance of one another.

- In other words, tokens are **discrete** features.

  Once again, *"hi"* and *"hello"* are completely distinct features for predicting whether a message is greeting somebody.

  "sociology" and *"economics"* are completely distinct features for predicting whether a message is about the social sciences.

Today, we will focus on text features that capture context...

Today, we will focus on text features that capture context...

But what do we mean by context?

Today, we will focus on text features that capture context...

But what do we mean by context?

And why should we care?

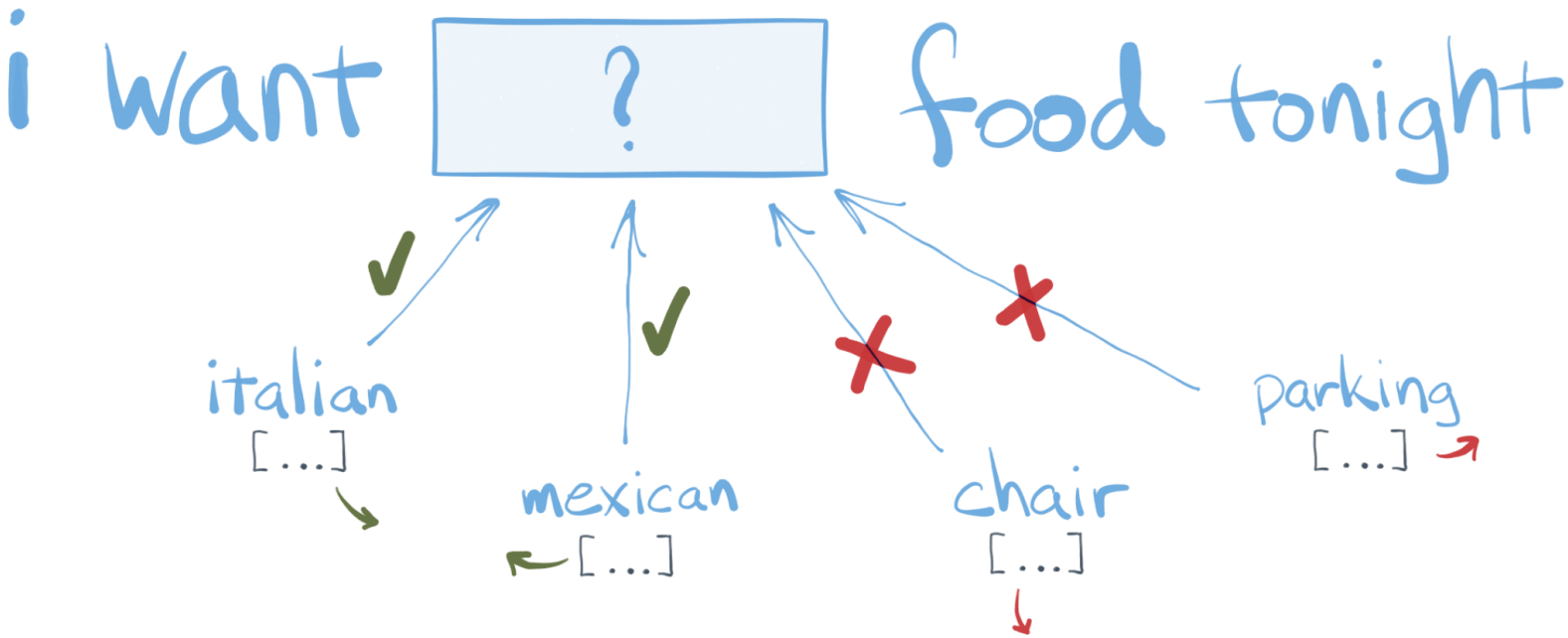
An Example to Build Some Intuition

Figure: Can you complete this text snippet?

Source: Patrick Harrison, S&P Global Market Intelligence

An Example to Build Some Intuition

Figure: Can you complete this text snippet?



Source: Patrick Harrison, S&P Global Market Intelligence

Language in Context (and vice-versa)

"you shall know a word by the company it keeps"

(J. R. Firth, 1957)

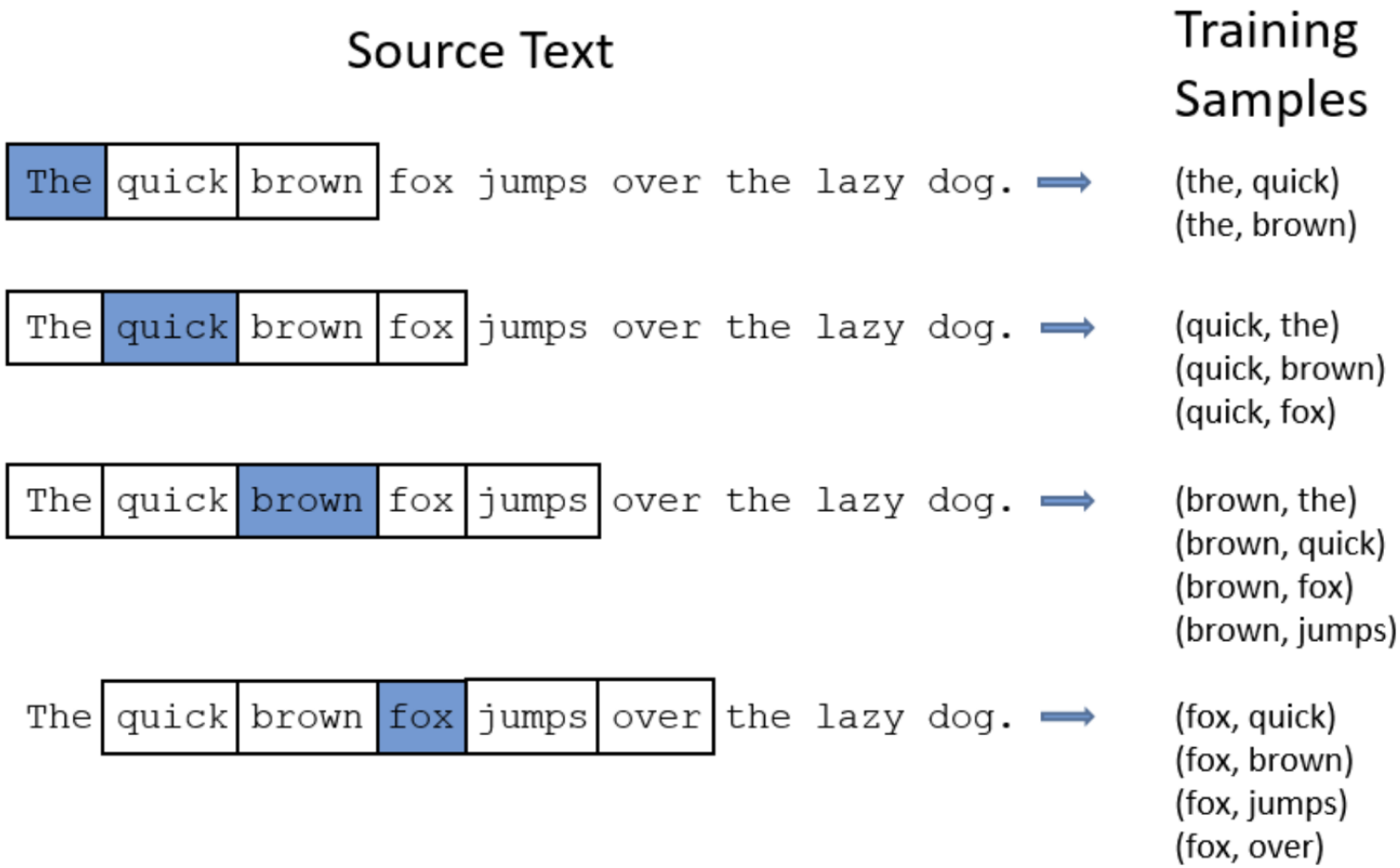
- Neighboring words provide us with additional information to interpret a word's meaning.
- In other words, **word co-occurrences capture context.**
- This information is useful for machine learning applications.

  e.g., document classification, machine translation, syntax prediction, machine comprehension, etc.

The Brute Force Approach

- **Build a large word co-occurrence matrix C.**
- Notations:
  - V is a vocabulary of $|V|$ words.
  - M is an integer called the **window**.
  - The M words preceding and the M words following a word consitute its **context**.
- The cell (i, j) of C represents how many times the word i co-occurs with word j in the window.
- Each of the lines of C is a vector representation of a word that contains more information than one-hot vectors (i.e., bag-of-words).

Concrete Example for the Window Size



Source: Julian Gilyadov. Window size $M = 2$.

The Limits to the Brute Force Approach

- However, the resulting co-occurrence matrix C is **high-dimensional and sparse**.
- As the vocabulary size increases, working with this matrix becomes intractable.

The Limits to the Brute Force Approach

- However, the resulting co-occurrence matrix C is **high-dimensional and sparse**.
- As the vocabulary size increases, working with this matrix becomes intractable.
- **Can we approximate C in a low-dimensional, dense vector space? (i.e., such that $p << |V|$)**

The Limits to the Brute Force Approach

- However, the resulting co-occurrence matrix C is **high-dimensional and sparse**.
- As the vocabulary size increases, working with this matrix becomes intractable.
- **Can we approximate C in a low-dimensional, dense vector space? (i.e., such that $p << |V|$)**
- Yes. This is precisely what text embeddings are all about.

Plan

The First Generation of Embeddings

- The three most famous models are:
  - `Word2Vec` (Mikolov et al. 2013)
  - `GloVe` (Pennington et al., 2014)
  - `FastText` (Mikolov et al., 2018)

The First Generation of Embeddings

- The three most famous models are:
  - `Word2Vec` (Mikolov et al. 2013)
  - `GloVe` (Pennington et al., 2014)
  - `FastText` (Mikolov et al., 2018)
- Let's dig deeper into `Word2Vec`.

Tomas Mikolov

FOLLOW

Senior Researcher, CIIRC CTU
Verified email at cvut.cz

Artificial Intelligence Machine Learning Language Modeling Natural Language Processing

TITLE	CITED BY	YEAR
Distributed representations of words and phrases and their compositionality T Mikolov, I Sutskever, K Chen, GS Corrado, J Dean Neural information processing systems	34060	2013

A Supervised Learning Problem

- `Word2Vec` reformulates learning word co-occurrences as two prediction tasks:

A Supervised Learning Problem

- `Word2Vec` reformulates learning word co-occurrences as two prediction tasks:
  - **Continuous Bag of Words (CBOW):** Given its context words, predict a focus word.

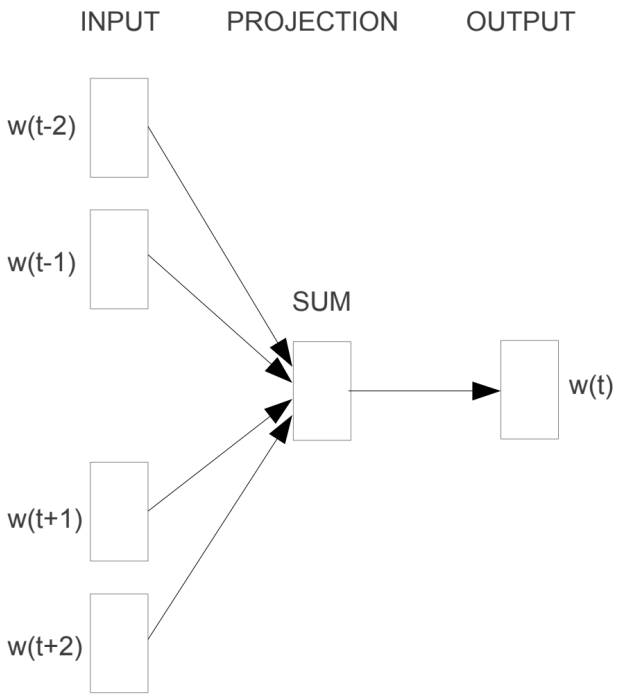
A Supervised Learning Problem

- `Word2Vec` reformulates learning word co-occurrences as two prediction tasks:
  - **Continuous Bag of Words (CBOW):** Given its context words, predict a focus word.
  - **Skipgram:** Given a focus word, predict all its context words.

A Supervised Learning Problem

- `Word2Vec` reformulates learning word co-occurrences as two prediction tasks:
  - **Continuous Bag of Words (CBOW):** Given its context words, predict a focus word.
  - **Skipgram:** Given a focus word, predict all its context words.
- In both cases, the model results in a low-dimensional, dense vector space representation of C.

CBOW – Intuition



Source: Efficient Estimation of Word Representations in Vector Space, Mikolov et al. (2013)

CBOW – Likelihood

- Recall M the size of the context window (often between 5 and 10).
- Given a sequence of T words, the log-likelihood is

$$\frac{1}{T}\sum_{t=1}^{T}\log\left(P\left(w_t|\{w_{t+j}\}_{-M\leq j\leq M,j\neq 0}\right)\right)$$

CBOW – Likelihood

- Recall M the size of the context window (often between 5 and 10).
- Given a sequence of T words, the log-likelihood is

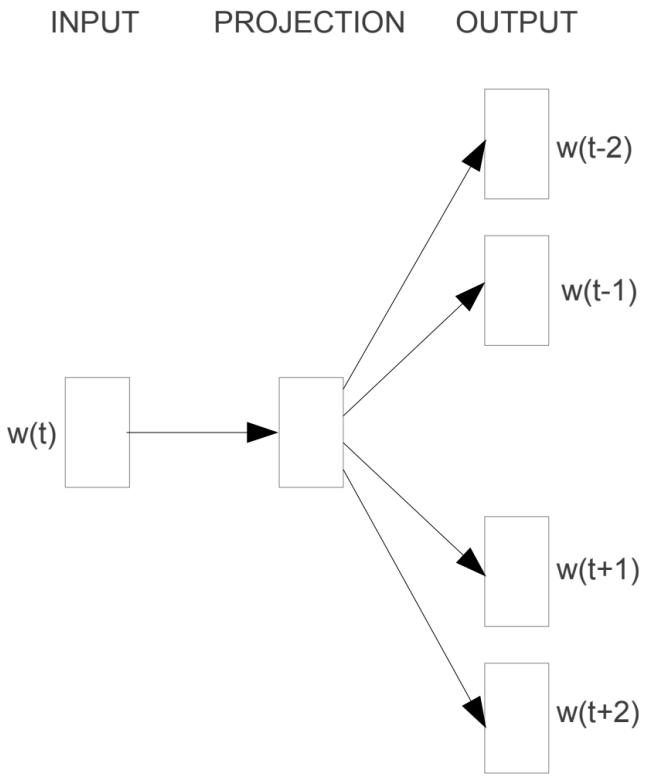
$$\frac{1}{T}\sum_{t=1}^{T}\log\left(P(w_t|\{w_{t+j}\}_{-M\leq j\leq M, j\neq 0})\right)$$

- The probability is given by the softmax function:

$$P(w_t|\{w_{t+j}\}_{-M\leq j\leq M, j\neq 0}) = \frac{\exp(w_t'\cdot\bar{u}_t)}{\sum_{k=1}^{|V|}\exp(w_k'\cdot\bar{u}_t)}$$

($\bar{u}_t$ is the average of the vectors for words in the context window)

Skipgram – Intuition



Source: Distributed Representations of Words and Phrases and their Compositionality, Mikolov et al. (2013)

Skipgram – Likelihood

- Recall M the size of the context window (often between 5 and 10).
- Given a sequence of T words, the log-likelihood is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-M\leq j\leq M, j\neq 0}\log\left(P(w_{t+j}|w_t)\right)$$

Skipgram – Likelihood

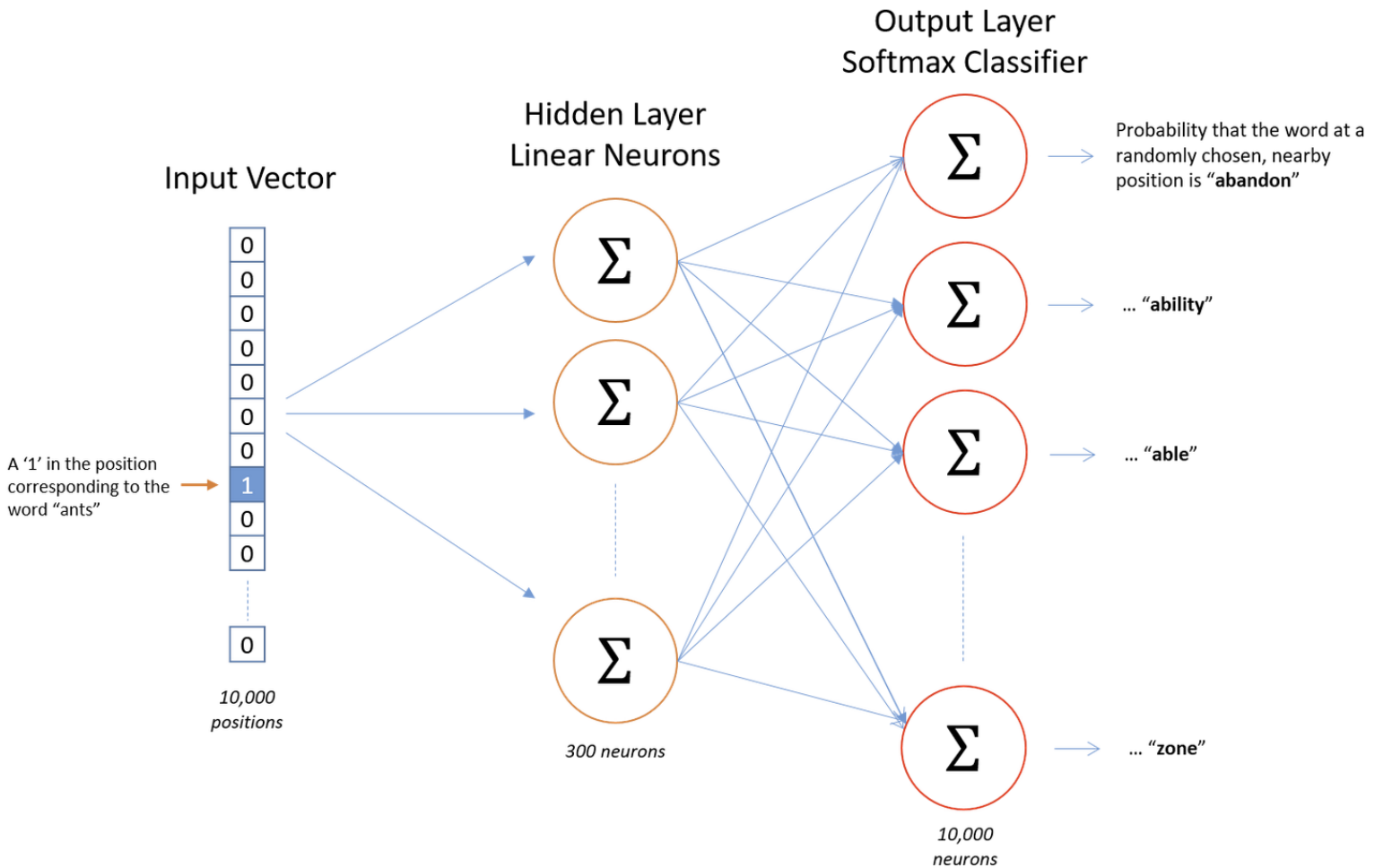
- Recall M the size of the context window (often between 5 and 10).
- Given a sequence of T words, the log-likelihood is

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-M\leq j\leq M, j\neq 0}\log\left(P(w_{t+j}|w_t)\right)$$

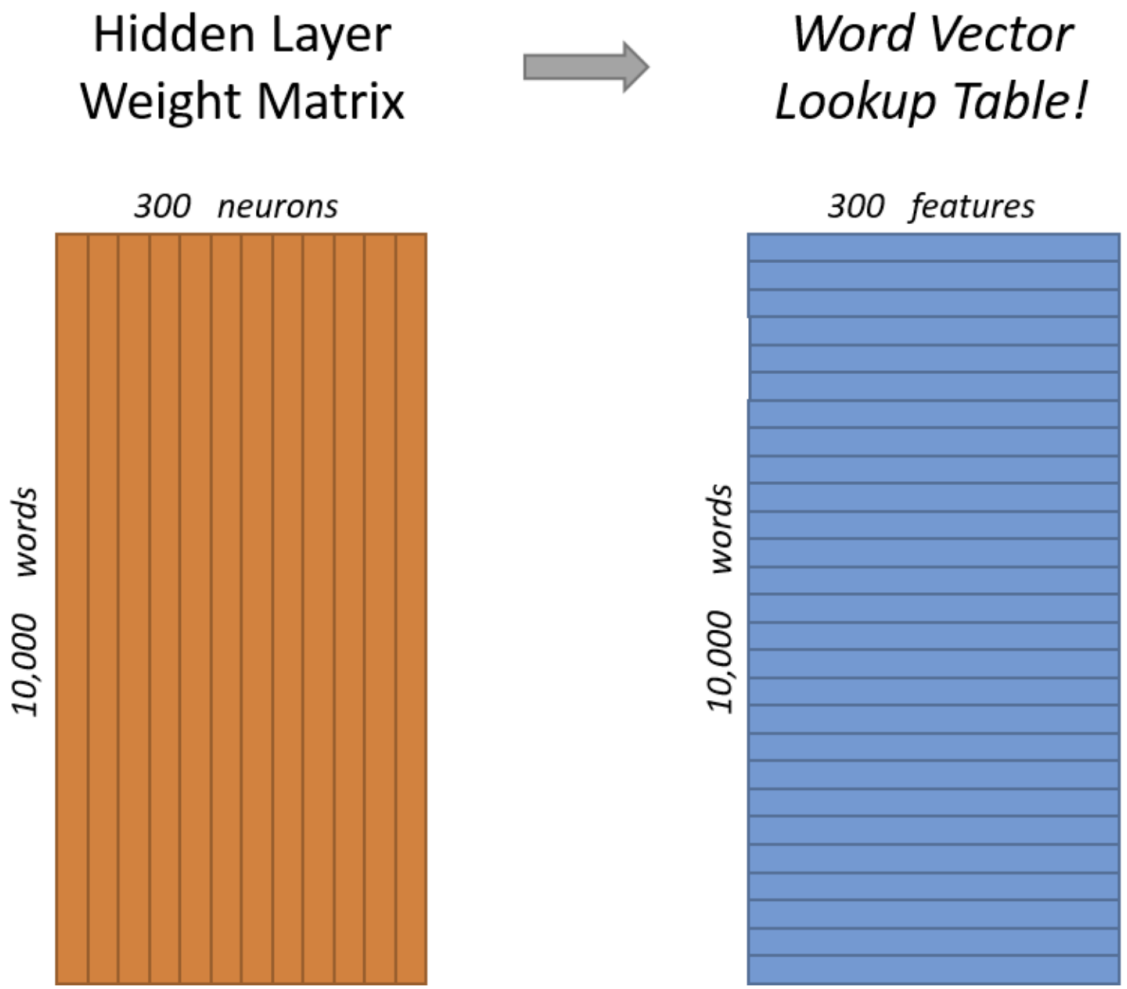
- The probability is given by the softmax function:

$$P(w_{t+j}|w_t) = \frac{\exp(w'_{t+j}\cdot w_t)}{\sum_{k=1}^{|V|}\exp(w'_k\cdot w_t)}$$

Neural Network Representation



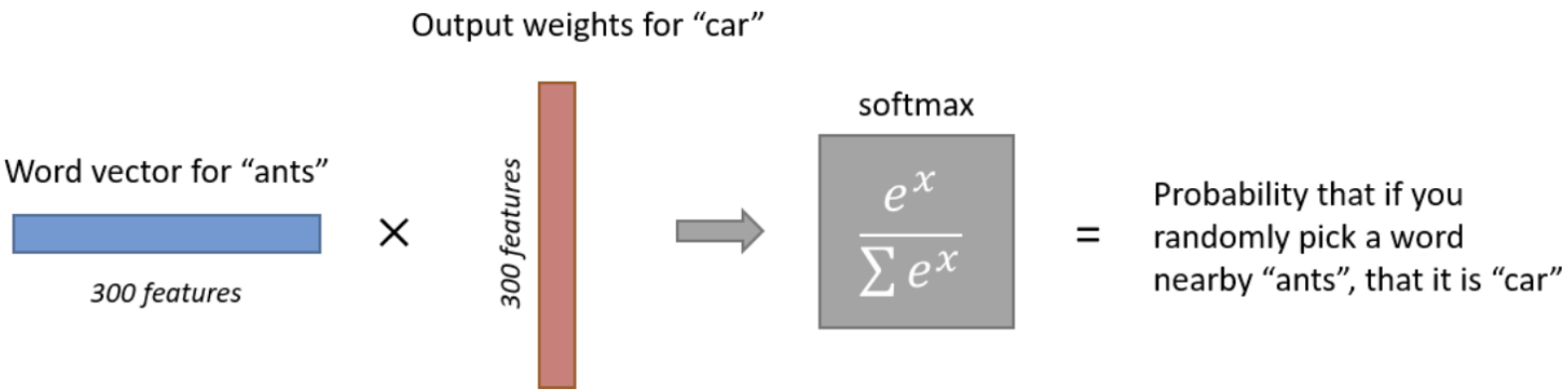
Source: Julian Gilyadov. Contrary to most supervised learning tasks, the hidden layer is what we actually care about here. It represents the word vectors!



Source: Julian Gilyadov

$$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = \begin{bmatrix} 10 & 12 & 19 \end{bmatrix}$$

Source: Julian Gilyadov



Source: Julian Gilyadov

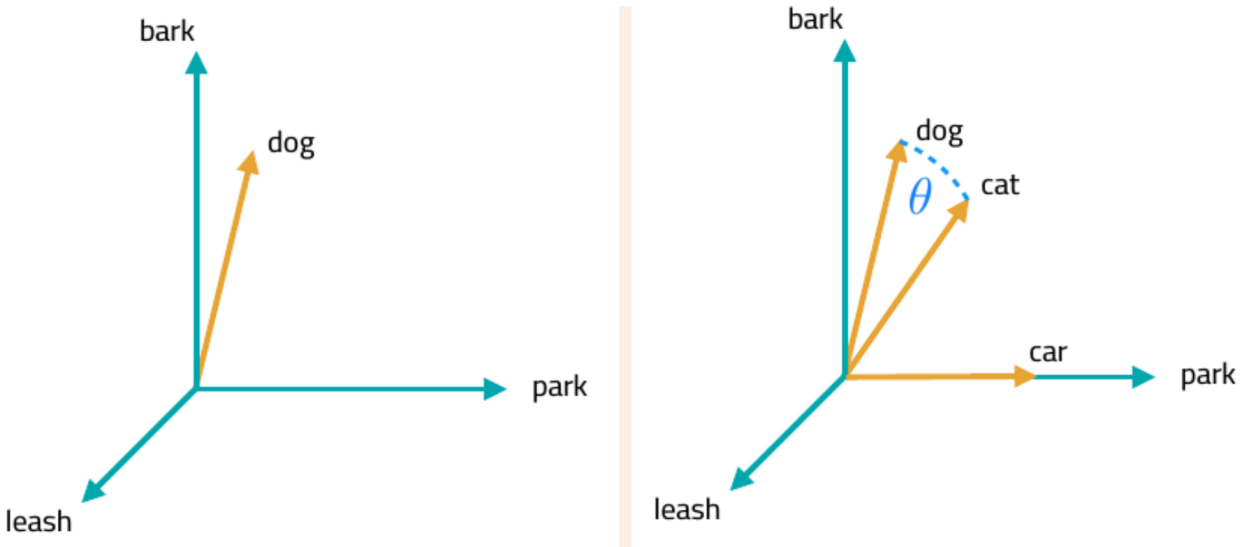
Distance Between Texts

- With embeddings, we can use linear algebra to understand **relationships between words**.
- In particular, words that are geometrically close to each other are **similar**.
- The standard metric for comparic vectors is **cosine similarity**:

$$\cos\theta = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

- When vectors are normalized, cosine similarity is:
  - Simply the dot product of both vectors
  - Proportional to the Euclidean distance (so you can use it, too)

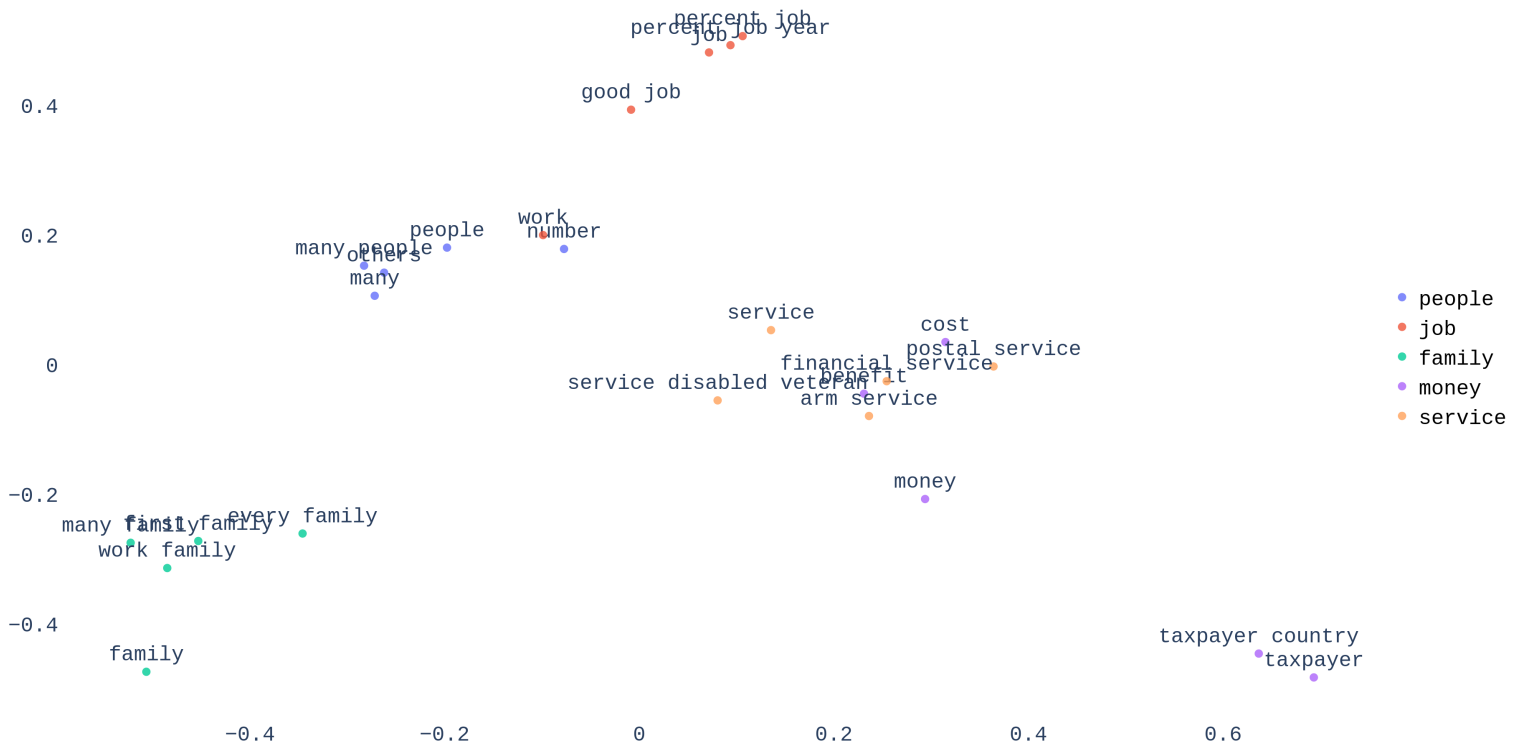
Distance Between Texts



Visualizing Embeddings

- One can also visualize the resulting embedding space by **projecting it on a two-dimensional space**.
- Two commonly used techniques are:
  - Principal Component Analysis (PCA)
  - t-distributed stochastic neighbor embedding (t-SNE)

Visualizing Embeddings

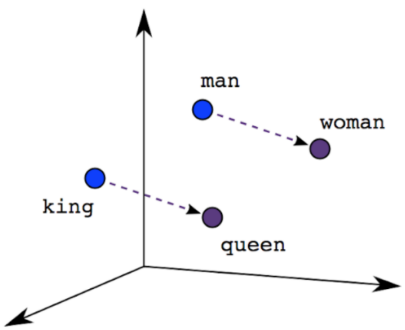


Source: RELATIO: Text Semantics Capture Political and Economic Narratives, Ash et al. (2022)

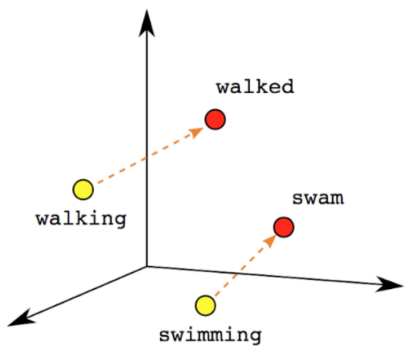
Basic arithmetic often carries meaning.

- `Word2vec` algebra can depict conceptual, analogical relationships between words.

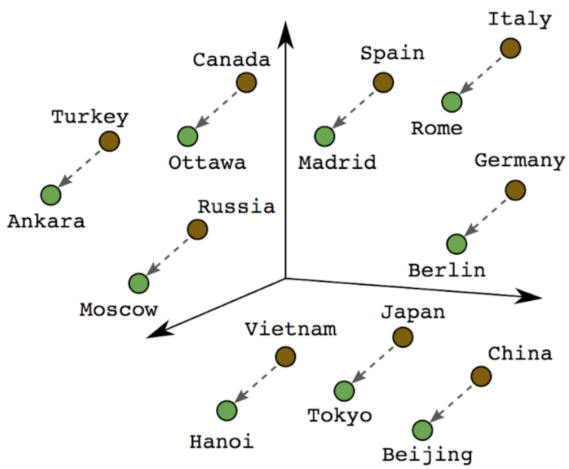
  e.g., $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\textit{queen}}$



Male-Female



Verb Tense



Country-Capital

Some Refinements

- The main assumption behind `word2vec` is that **context words are exchangeable**.
- In other words, the ordering of words is not accounted for.
- Recent models relax this assumption. They are called **transformers**.
  - `ElMo` (Peters et al., 2018)
  - `BERT` (Devlin et al., 2019)
- And consistently outperform previous language models in a large variety of tasks.
- They are particularly good at **transfer learning**.

Pros and Cons

- **Pros**
  - Many pre-trained models for different languages are freely available online.
  - Many packages to train models from scratch or fine-tune existing models to a specific corpus.
  - More often than not, they provide sizable gains in prediction accuracy.
- **Cons**
  - Clear loss of interpretability relative to bag-of-words
  - Neighbouring words are not the only forms of context.
  - Often critiqued as "stochastic parrots" (Bender et al., 2021).

Plan

Identifying Political Frames

- Politically-contested issues are often discussed with different emphases by different people. This emphasis is called a **frame**.
- Embeddings can help us to identify these frames.

Identifying Political Frames

- Politically-contested issues are often discussed with different emphases by different people. This emphasis is called a **frame**.
- Embeddings can help us to identify these frames.

Figure: Frames Related to Immigration

Immigration Specific	Victim: Global Economy	Immigrants are victims of global poverty, underdevelopment and inequality
	Victim: Humanitarian	Immigrants experience economic, social, and political suffering and hardships
	Victim: War	Focus on war and violent conflict as reason for immigration
	Victim: Discrimination	Immigrants are victims of racism, xenophobia, and religion-based discrimination
	Hero: Cultural Diversity	Highlights positive aspects of differences that immigrants bring to society
	Hero: Integration	Immigrants successfully adapt and fit into their host society
	Hero: Worker	Immigrants contribute to economic prosperity and are an important source of labor
	Threat: Jobs	Immigrants take nonimmigrants' jobs or lower their wages
	Threat: Public Order	Immigrants threaten public safety by being breaking the law or spreading disease
	Threat: Fiscal	Immigrants abuse social service programs and are a burden on resources
	Threat: National Cohesion	Immigrants' cultural differences are a threat to national unity and social harmony

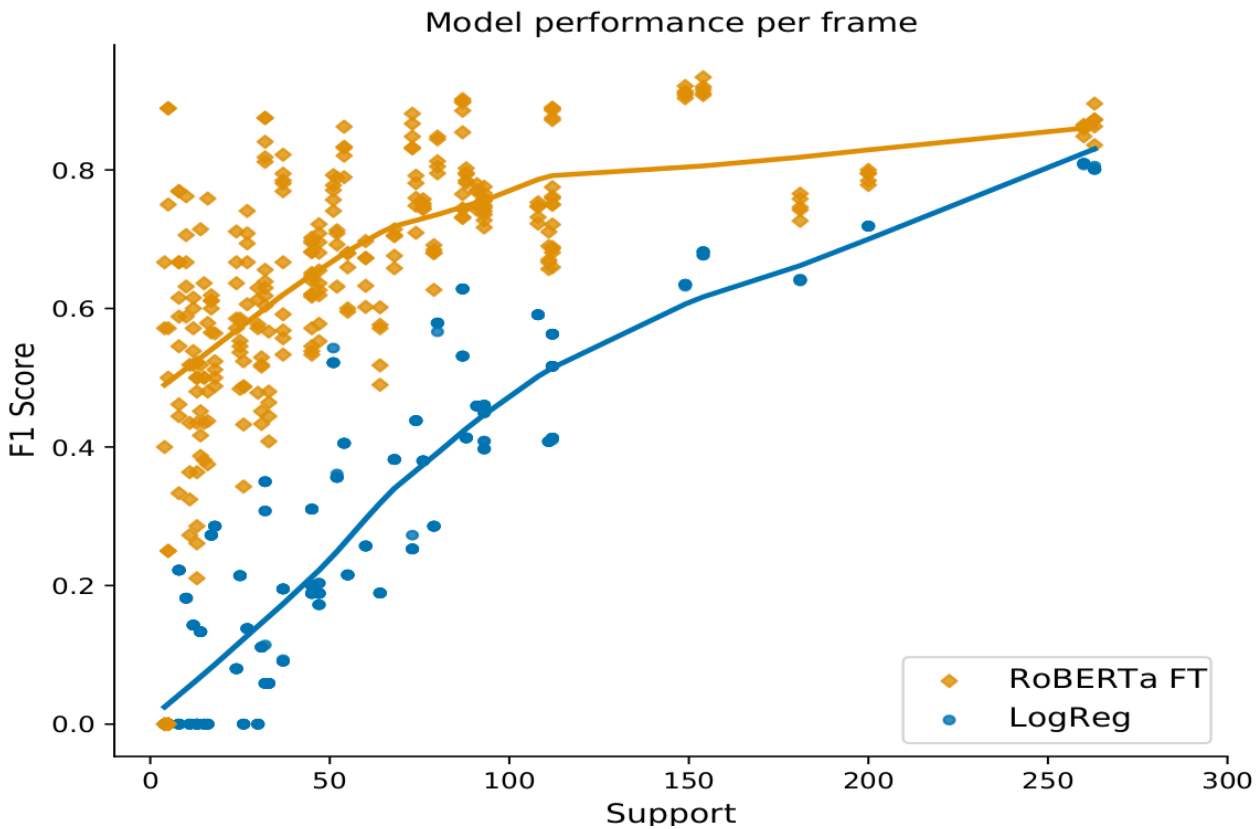
Source: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021)

Identifying Political Frames

[WHERE THE JOBS ARE]$_{\text{Economic}}$
[Critics of illegal immigration can make many cogent arguments to support the position that the U.S. Congress and the Colorado legislature must develop effective and well-enforced immigration policies that will restrict the number of people who migrate here legally and illegally.]$_{\text{Public opinion}}$
[It's true that all forms of immigration exert influence over our economic and [cultural make-up.]$_{\text{Cultural identity}}$ **In some ways, immigration improves our economy by adding laborers, taxpayers and consumers, and in other ways [immigration detracts from our economy by increasing the number of students, health care recipients and other beneficiaries of public services.]**$_{\text{Capacity}}$**]**$_{\text{Economic}}$
[Some economists say that immigrants, legal and illegal, produce a net economic gain, while others say that they create a net loss.]$_{\text{Economic}}$ There are rational arguments to support both sides of this debate, and it's useful and educational to hear the varying positions.

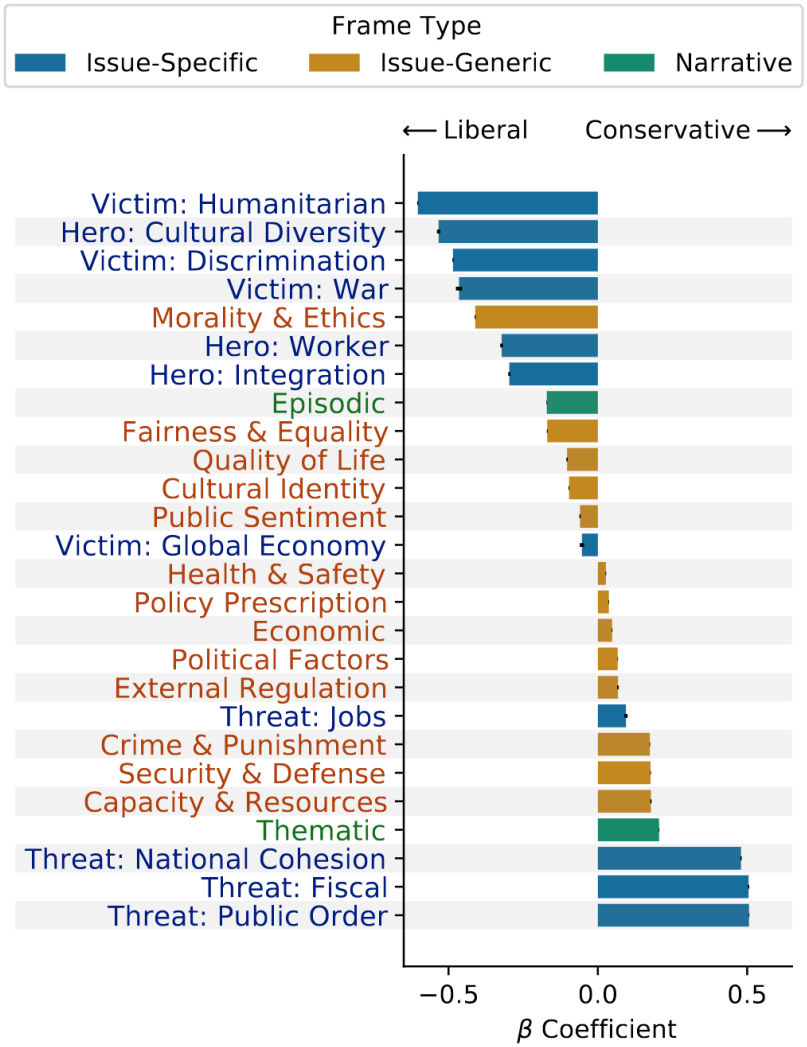
Source: The Media Frames Corpus: Annotations of Frames Across Issues, Card et al. (2015)

Identifying Political Frames



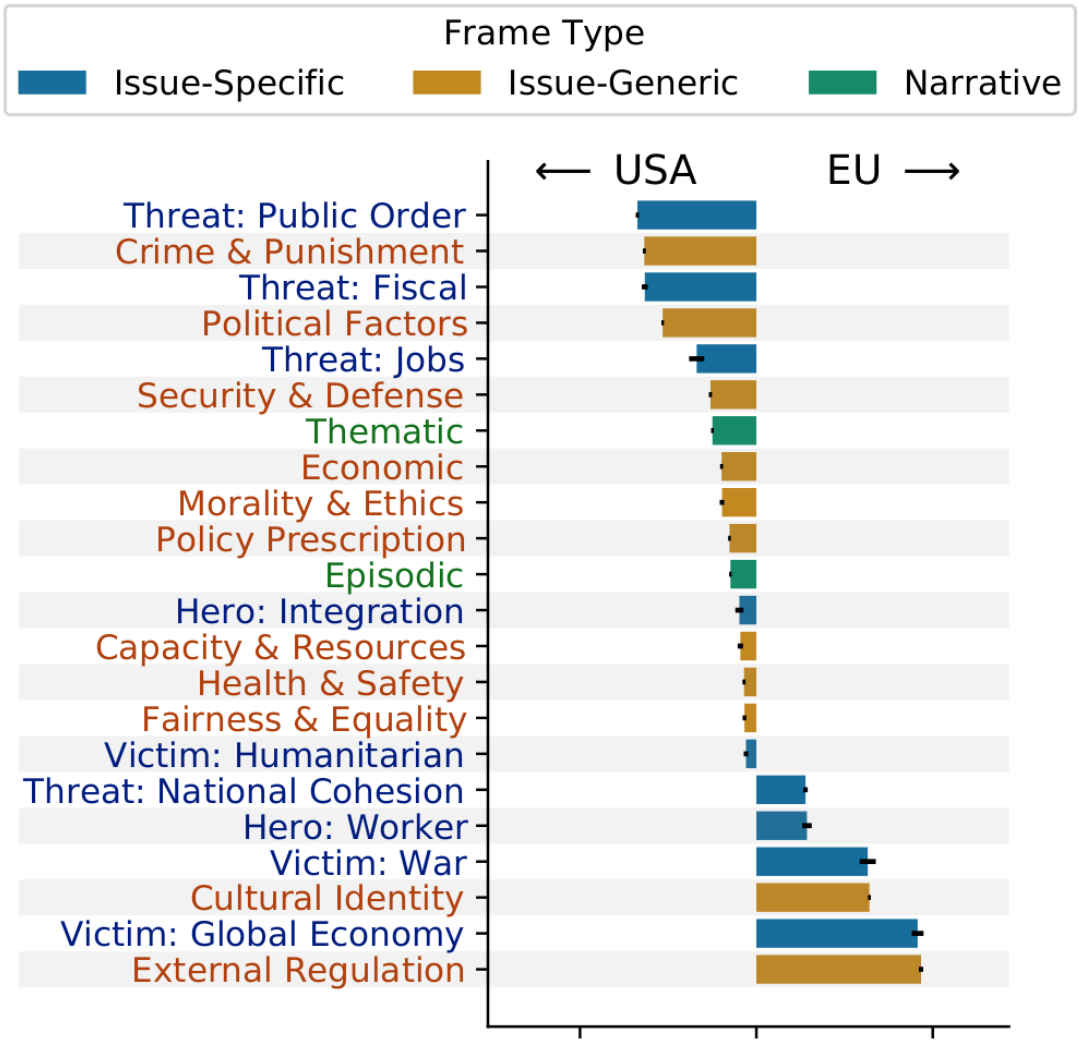
Source/Notes: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021). Modern transformers have high accuracy ($\approx$ 3-4 times better than a random guess).

Identifying Political Frames



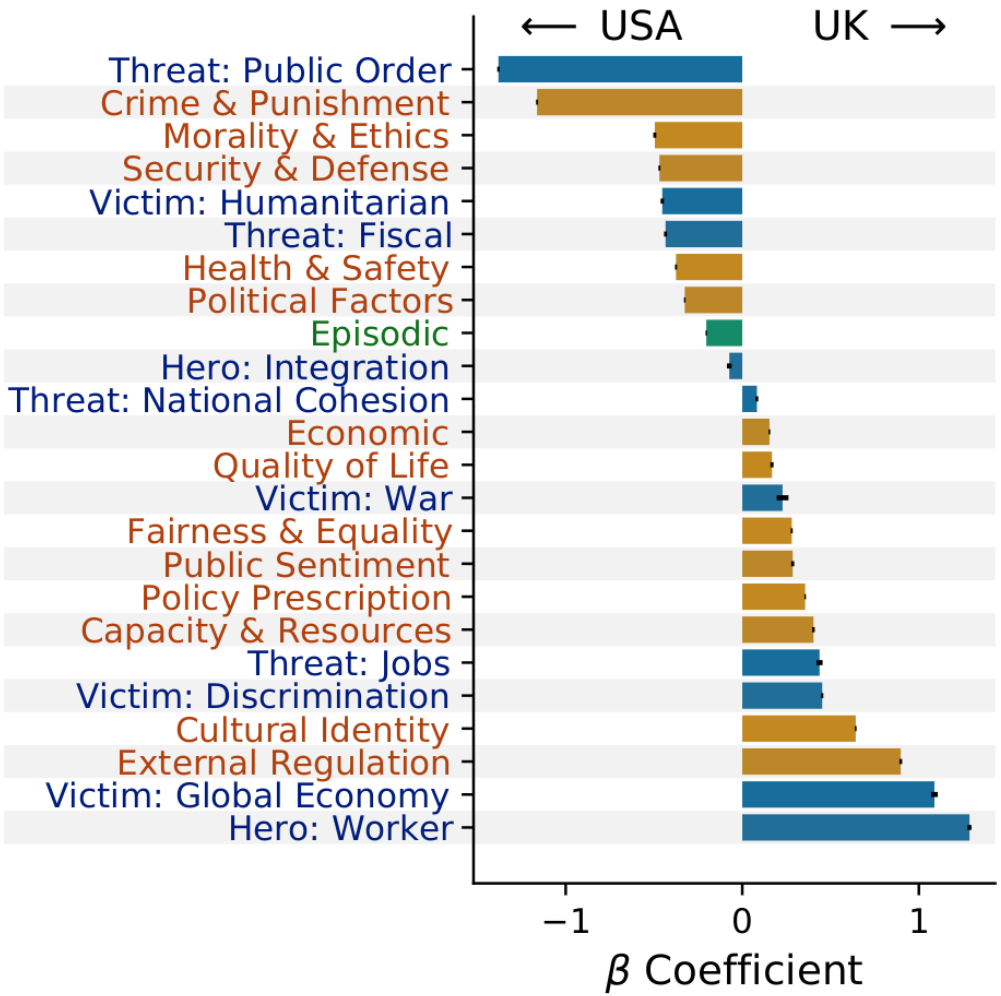
Source: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021)

Identifying Political Frames



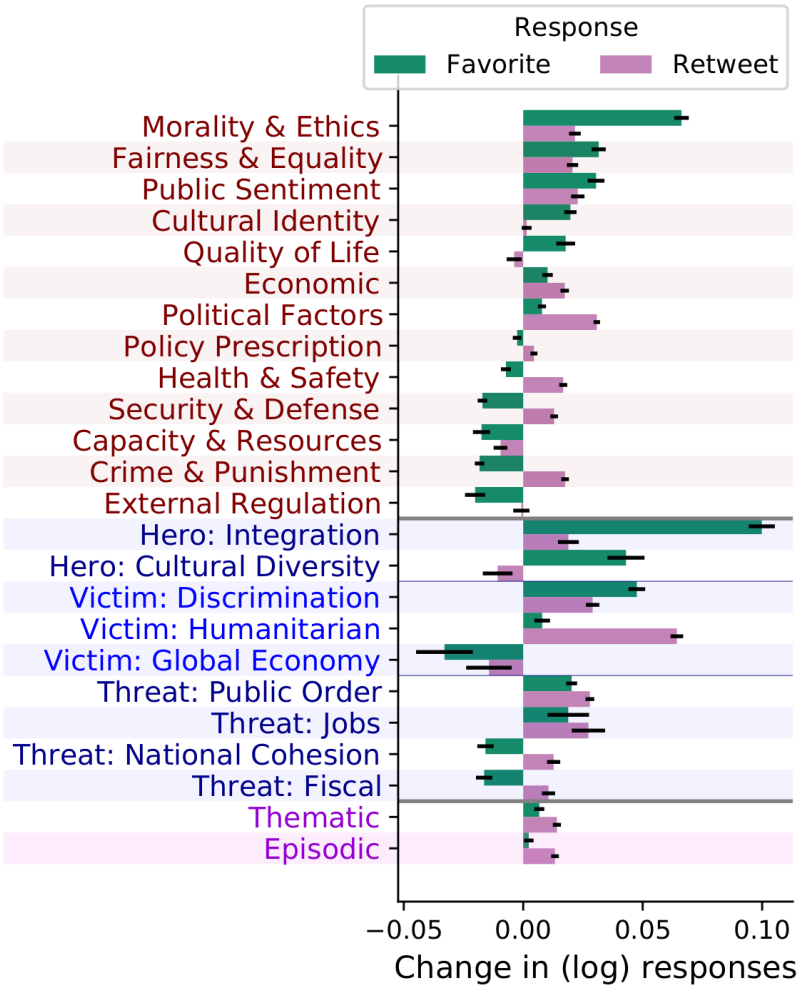
Source: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021)

Identifying Political Frames



Source: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021)

Identifying Political Frames



Source: Modeling Framing in Immigration Discourse on Social Media, Mendelsohn et al. (2021)

Other Applications

- Modern embeddings models are very efficient for supervised learning tasks.
- Some additional examples:
  - Sentiment analysis
  - Ideology prediction
  - Toxic language detection
  - Bot detection
- Come up with your own!

Plan

Method 1: Topic Modeling for Short Text Snippets

Method:

1. Represent all your documents as vectors in embeddings space.
2. Apply a clustering algorithm (e.g., K-Means).

Method 1: Topic Modeling for Short Text Snippets

Method:

1. Represent all your documents as vectors in embeddings space.
2. Apply a clustering algorithm (e.g., K-Means).

- Each cluster can be interpreted as a topic (Demszky et al., 2019).
- For very short text snippets (e.g., phrases of two or three words), the resulting clusters can also be interpreted as entities (Ash et al., 2022).

The Return of the Yellow Vests!

- Boyer et al. (2022) study the role of social media in the Yellow Vests protests that took place in 2018 in France.
- 600+ active Yellow Vests Facebook pages between November 2018 and April 2019
- They cluster embedded sentences via the K-Means algorithm (15 topics).
- The share of antagonistic topics increases over time in the corpus. Moderate users progressively fell into inactivity, but those who remained radicalized.

Figure: Topic Model on Yellow Vests Facebook Pages

A. Economic concerns

B. Support

C. Politics

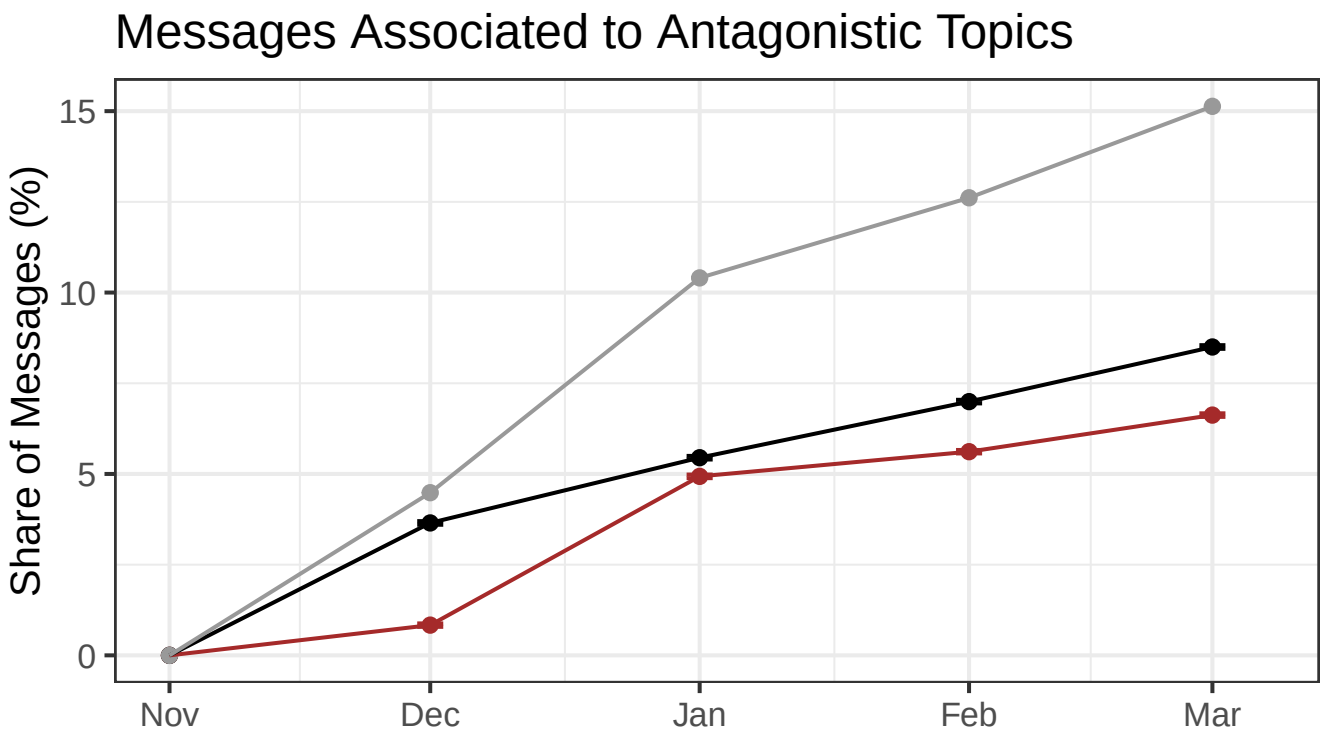
D. Insults

E. Violence

F. Conspiracy Theories

Source: Mobilization Without Consolidation: Social Media and the Yellow Vests Protests, Boyer at al. (2022)

Figure: Intensive and Extensive Margins of Radicalization



Source/Notes: Mobilization without Consolidation: Social Media and the Yellow Vests Protests, Boyer et al. (2022). The regression equation is $Y_{i,t} = \delta_i + \gamma_t + \varepsilon_{i,t}$. This simple linear decomposition suggests moderate users left the movement (red line), and those who stayed radicalized (black line).

Method 2: Measuring Dimensions of Language

Method:

1. Represent two dimensions of language as vectors.
2. Compute their cosine similarity.

Method 2: Measuring Dimensions of Language

Method:

1. Represent two dimensions of language as vectors.
2. Compute their cosine similarity.

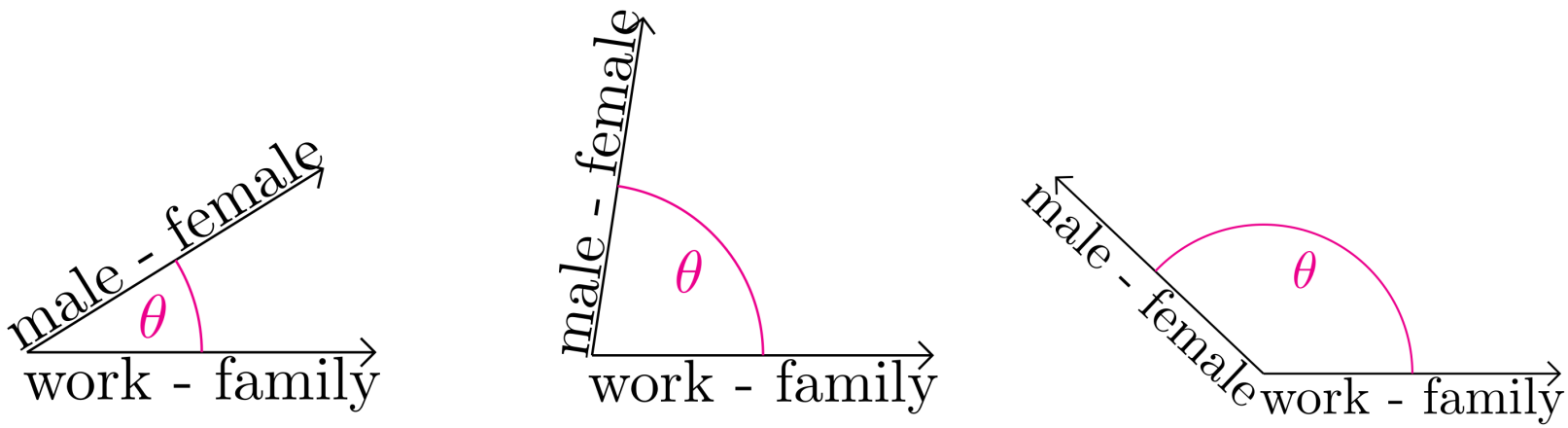
- For example:

  $\vec{\text{male}} - \vec{\text{female}} \rightarrow$ Gender dimension

  $\vec{\text{work}} - \vec{\text{family}} \rightarrow$ Work dimension

- Often, the dimensions are based on dictionaries.
- For example, the vector $\vec{\text{male}}$ can be computed as the average of a list of words related to men (*e.g., Marc, Thomas, John, he, etc.*).

A Graphical Illustration



(a) (b) (c)

Notes: The larger θ, the more dissimilar are the two dimensions of language. In Panel (a), men are strongly associated to work. In Panel (c), it is the opposite.

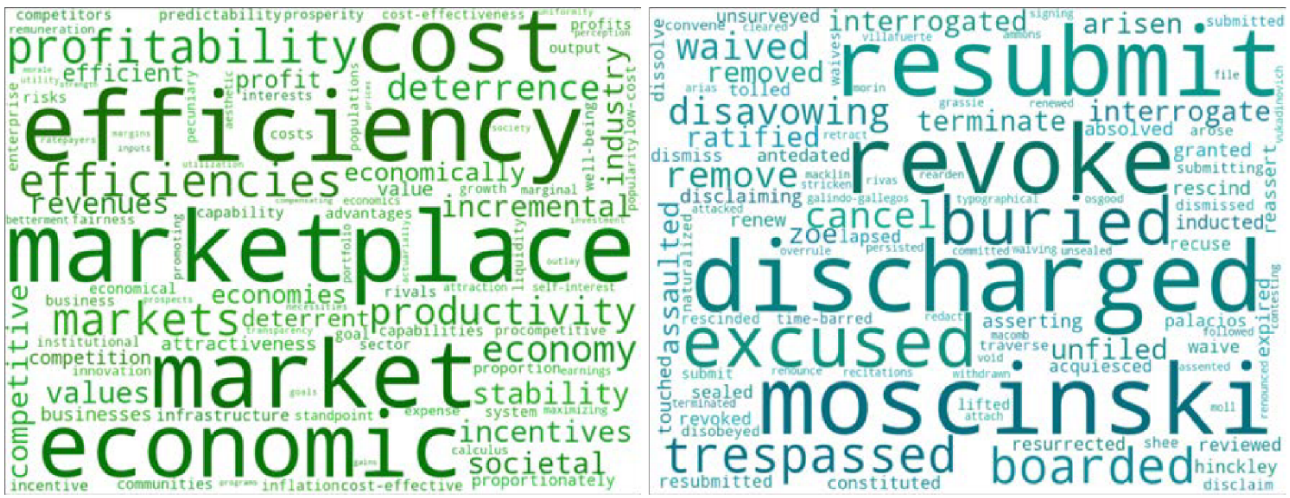
Application to Economic Jargon

- Elliott Ash, Daniel Chen & Suresh Naidu (2022) study the impact of the early law-and-economics movement on the U.S. judiciary.
- Focus on the Manne Economics Institute for Federal Judges, an intensive economics course that trained almost half of federal judges between 1976 and 1999.
- After attending the economics training program, were judges more likely to use economic jargon in their rulings?
- **The dimensions of economic jargon are based on a dictionary of legal and economic terms.**

Application to Economic Jargon

Figure A.4: Words Correlated with Law-and-Economics Lexicon Dimension

(a) Positively Associated Words (a) Negatively Associated Words

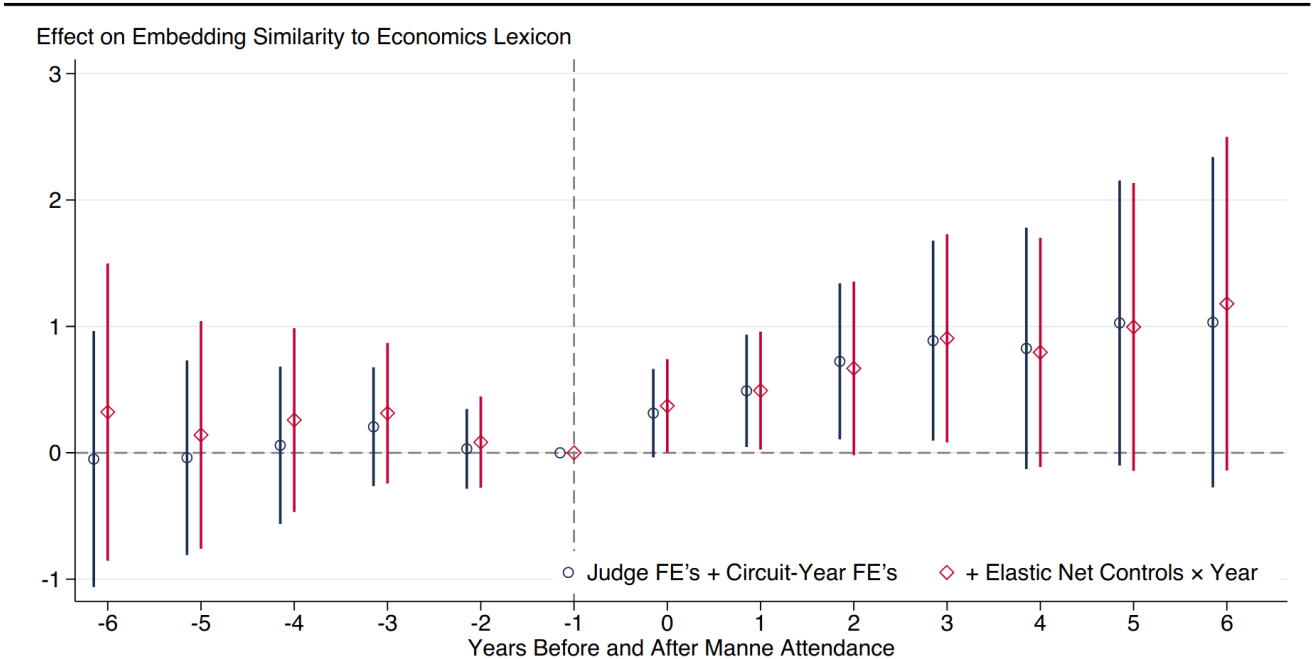


Notes. The left word cloud lists the set of words that have the highest cosine similarity to the average word vector for Ellickson phrases in the word embedding space. The right word cloud gives the words that have the lowest (most negative) cosine similarity to this vector.

Source: Ideas Have Consequences: The Impact of Law and Economics on American Justice, Ash et al. (2022)

Application to Economic Jargon

Figure 3: Effect of Manne Program on Economics Language



Notes. Event study effect of Manne attendance on Word Embedding Similarity to Law-and-Economics Lexicon (from Ellickson, 2000). Sample is limited to case authors. Regressions include judge and circuit-year fixed effects (blue circles), with additional specification adding elastic-net-selected controls interacted with year fixed effects (red diamonds). Observations are weighted to treat judge-years equally. Error spikes give 95% confidence intervals, with standard errors clustered by judge.

Source: Ideas Have Consequences: The Impact of Law and Economics on American Justice, Ash et al. (2022)

Application to Emotional Content

- Gloria Gennaro and Elliott Ash (2022) study appeals to emotions in American political discourse.
- They rely on the U.S. Congressional Record (1858–2013).
- Emotionality has been increasing over time (in particular since the 1980s).
- **The dimensions of emotion/reason are based on the Linguistic Inquiry and Word Count (LIWC) dictionary.**

Application to Emotional Content

(a) Cognitive/Rational Language

(b) Affective/Emotional Language

Fig. 1. *Semantic Poles for Rationality and Emotion.*
Notes: The wordclouds show the dictionary words that are closest to the respective 'poles' of the dimension in the embedding space corresponding to rationality/cognition (a) and affect/emotion (b). Size denotes closeness to the respective word-vector centroid.

Source/Notes: Emotion and Reason in Political Language, Gennaro and Ash (2021).

Application to Emotional Content

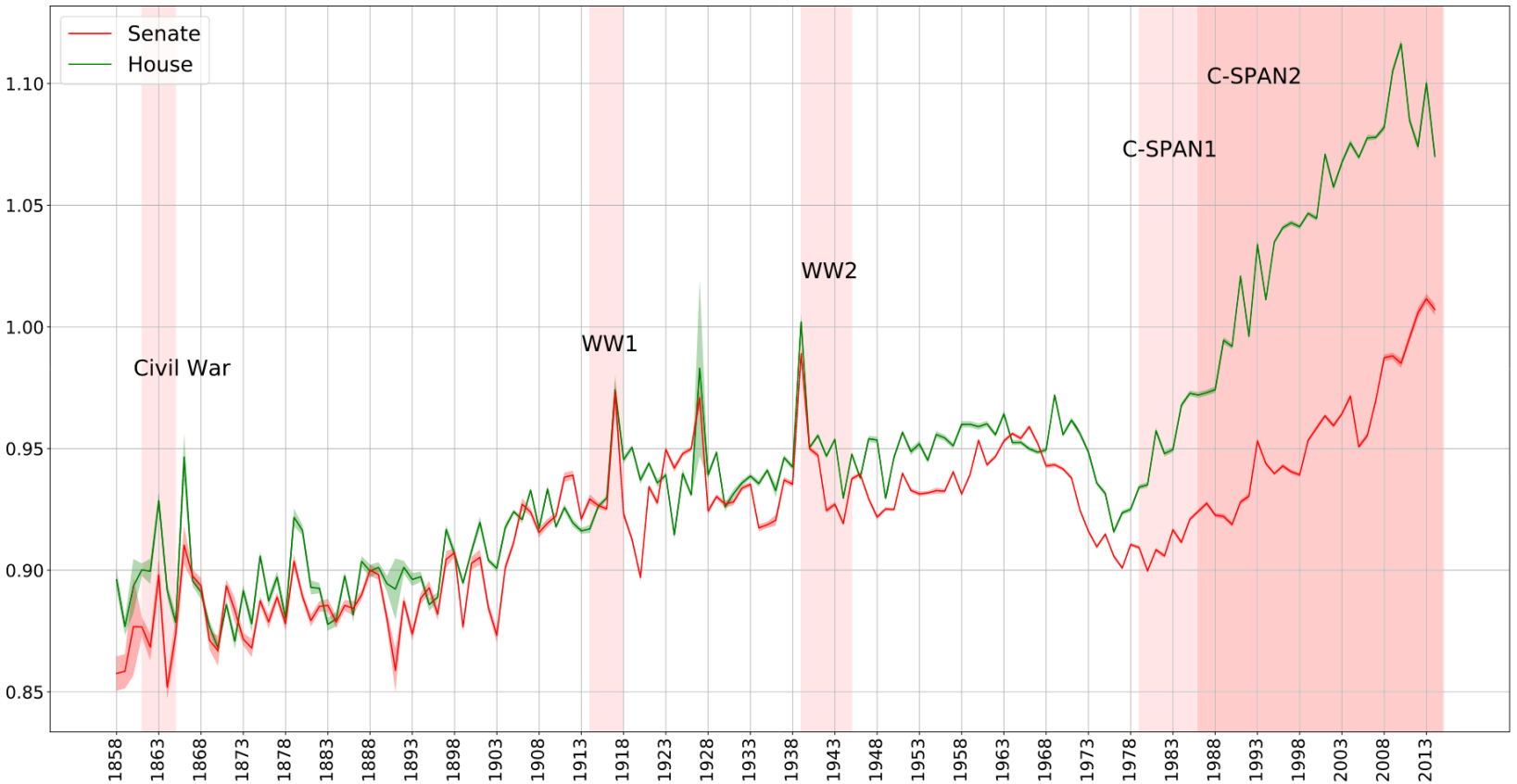


Fig. 2. *Emotionality in U.S. Congress by Chamber, 1858–2014.*
Notes: Time series of emotionality in the Senate (red) and the House of Representatives (green).

Source: Emotion and Reason in Political Language, Gennaro and Ash (2021)

Plan

The Current Research Frontier

- GPT-3 and other monsters
- Other related fields build on similar concepts and methods:
  - Speech recognition
  - Image recognition
- Many media take the form or spoken text or images...
- A new treasure trove for social scientists?

A Few Books To Go Further

- Natural Language Processing in Python, Third Edition ("NLTK Book")
- Yoav Goldberg, Neural Network Methods for Natural Language Processing (2017)
- Jurafsky and Martin, Speech and Language Processing (3d Ed. 2019).

Thanks for listening!

Text as Data for the Social Sciences (II)

Germain Gauthier, Felix Lennert, Etienne Ollion

germain.gauthier@polytechnique.edu

SICSS Paris

June 2022