Unstructured Data for Economics

Lecture 1: Word Embeddings

Stephen Hansen
University College London

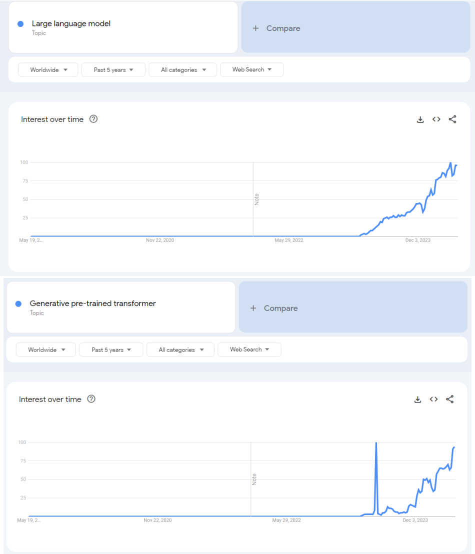
Reading

Background material for the lecture:

1. Ash and Hansen, *Text Algorithms in Economics*, `https://bityl.co/QQRC`
2. Jurafsky and Martin, *Speech and Language Processing*, Ch. 6 `https://bityl.co/QQRS`

All lecture materials and example notebooks available on Dropbox.

Surge in Interest in AI for Natural Language



What is New Here?

Text data has a long(ish) history in accounting, economics, and finance.

Primary motivation is to measure important phenomena difficult to capture with traditional data.

Well-known examples include competition, policy uncertainty, sentiment, polarization, etc.

Automated methods have largely counted words or their co-occurrence.

Main limitation is limited understanding of context: "economic growth is weak but long-term productivity trends are strong" vs. "economic growth is strong but long-term productivity trends are weak."

Deep Learning

The key breakthrough in modeling context came with deep learning models with Transformer architecture [Vaswani et al., 2017].

Contrast in approach from previous approaches to language:

- Expert systems
- Raw feature counts
- Dimensionality reduction of feature counts

Compared to previous models, LLMs (i) are easier to use; (ii) better incorporate contextual data; (iii) produce more human-like output; (iv) have an increasingly impenetrable internal structure.

Outline

1. Background
2. Word2vec
3. Embedding Non-Textual Data

Background

Notation

The corpus is composed of D documents indexed by d.

After pre-processing, each document is a finite, length-N_d list of terms $\mathbf{w}_d = (w_{d,1}, \ldots, w_{d,N_d})$ with generic element $w_{d,n}$.

Let $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_D)$ be a list of all terms in the corpus, and let $N \equiv \sum_d N_d$ be the total number of terms in the corpus.

Suppose there are V **unique** terms in $\mathbf{w}$, where $1 \leq V \leq N$, each indexed by v.

We can then map each term in the corpus into this index, so that $w_{d,n} \in \{1, \ldots, V\}$.

Example

Consider three documents:

1. 'stephen is nice'
2. 'john is also nice'
3. 'george is mean'

We can consider the set of unique terms as $\{\text{stephen}, \text{is}, \text{nice}, \text{john}, \text{also}, \text{george}, \text{mean}\}$ so that $V = 7$.

Construct the following index:

stephen	is	nice	john	also	george	mean
1	2	3	4	5	6	7

We then have $\mathbf{w}_1 = (1, 2, 3)$; $\mathbf{w}_2 = (4, 2, 5, 3)$; $\mathbf{w}_3 = (6, 2, 7)$.

Which Corpus?

Much of traditional text-as-data analysis fits models on corpora drawn from domain of interest.

Large language models were first fit on generic corpora like Common Crawl, Wikipedia, or Google Books.

More recent iterations expand the training data (but details becoming more obscure).

Important to realize that the training data contains the knowledge that a model can encode.

Any biases in the training data can also be inherited by the model.

How to Preprocess?

Preprocessing in traditional text-as-data analysis follows a standard sequence of steps:

1. Break strings into individual tokens.
2. Remove common and rare words.
3. Fold into common linguistic roots (lower case, stem/lemmatize)

Detail outside scope of course but see `https://bityl.co/QIKa` for code demonstration.

How to Preprocess?

Preprocessing in traditional text-as-data analysis follows a standard sequence of steps:

1. Break strings into individual tokens.
2. Remove common and rare words.
3. Fold into common linguistic roots (lower case, stem/lemmatize)

Detail outside scope of course but see `https://bityl.co/QIKa` for code demonstration.

In LLMs there is much more limited pre-preprocessing in order to preserve meaning of language; see notebook for more details.

Document-Term Matrix

Let $x_{d,v}$ be the count of term v in document d.

A popular quantitative representation of text is the *document-term matrix* $\mathbf{X}$, which collects the counts $x_{d,v}$ into a $D \times V$ matrix.

In the previous example, we have

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Text analysis that relies on the document-term matrix is called the bag-of-words approach.

Documents as Vectors

We can view the documents that make up the rows of $\mathbf{X}$ as vectors.

Let each vocabulary term v have its own vector $\mathbf{e}_v \in \mathbb{R}^V$ where

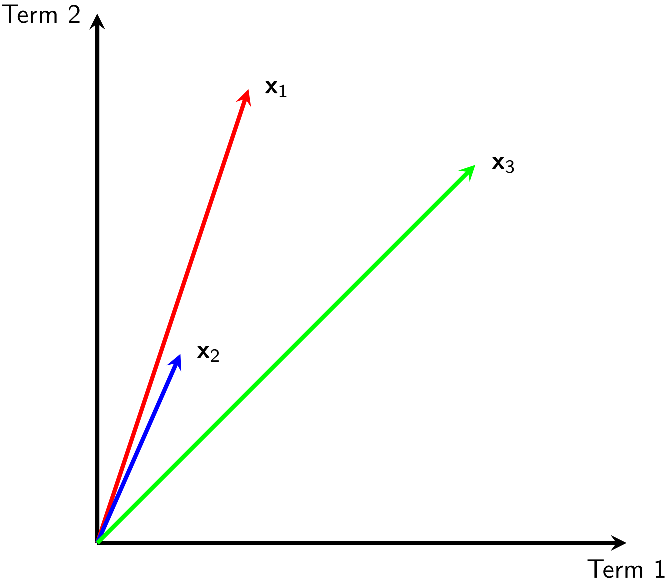
$$e_{v,v'} = \begin{cases} 1 & \text{if } v = v' \\ 0 & \text{otherwise} \end{cases}$$

Note that each term's vector is orthogonal to every other term's vector.

We can express document d as

$$\mathbf{x}_d = x_{d,1}\mathbf{e}_1 + x_{d,2}\mathbf{e}_2 + \ldots + x_{d,V}\mathbf{e}_V$$

Three Documents



Cosine Similarity

Define the cosine similarity between documents i and j as

$$CS(i,j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}$$

1. Since document vectors have no negative elements $CS(i,j) \in [0,1]$.
2. $\mathbf{x}_i / \|\mathbf{x}_i\|$ is unit-length, correction for different distances.

Limitations of Bag-of-Words + Cosine Similarity

Synonomy

economic growth is weak but long-term productivity trends are strong
economic growth is tepid but long-term productivity trends are strong

Limitations of Bag-of-Words + Cosine Similarity

Synonomy

economic growth is weak but long-term productivity trends are strong
economic growth is tepid but long-term productivity trends are strong

Polysemy

economic statistics lie about current well-being
my cat's favorite activity is to lie on our bed

Limitations of Bag-of-Words + Cosine Similarity

Synonomy

economic growth is weak but long-term productivity trends are strong
economic growth is tepid but long-term productivity trends are strong

Polysemy

economic statistics lie about current well-being
my cat's favorite activity is to lie on our bed

Sequence

economic growth is weak but long-term productivity trends are strong
economic growth is strong but long-term productivity trends are weak

Synonomy and Factor Models

The synonomy problem is similar to the factor model problem.

A large set of correlated features (words) across observations (documents) motivate dimensionality reduction.

Well-known factor models for text:

1. **Latent Semantic Analysis** [Deerwester et al., 1990]

   Application: [Iaria et al., 2018]
2. **Probabilistic LSI / NMF** [Ding et al., 2006]

   Application: [Ke et al., 2021]
3. **Latent Dirichlet Allocation** [Blei et al., 2003]

   Application: [Hansen et al., 2018]

Richer Feature Count Tabulation?

These factor models still operate on the document-term matrix and so are not aware of context.

One possible idea is to tabulate richer features, e.g. sequences of words rather than individual words.

These are called *n*-gram models ($n = 2$ bigram, $n = 3$ trigram).

But the number of unique length-N sequences composed of V total features is V^N.

We need to work in lower-dimensional space to compare token sequences!

Word2vec

Word Embeddings

A word embedding is a low-dimensional vector representation of a word.

Ideally in this low-dimensional vector space words with similar meanings will lie close together.

The construction of word embeddings via neural networks was a major step that preceded the development of large language models.

Word2vec [Mikolov et al., 2013a, Mikolov et al., 2013b] is a particularly well-known algorithm for the construction of word embeddings.[1]

[1] See also [Bengio et al., 2003].

Distributional Hypothesis

The distributional hypothesis states that words that share similar contexts share similar meanings.

Example from JM:

(6.1) Ongchoi is delicious sauteed with garlic.

(6.2) Ongchoi is superb over rice.

(6.3) ...ongchoi leaves with salty sauces...

And suppose that you had seen many of these context words in other contexts:

(6.4) ...spinach sauteed with garlic over rice...

(6.5) ...chard stems and leaves are delicious...

(6.6) ...collard greens and other salty leafy greens

Formalizing Local Context

Recall that $w_{d,n}$ is the nth word in document d.

The *context* of $w_{d,n}$ is a length-$2L$ window of words around $w_{d,n}$:

$$\mathsf{C}(w_{d,n}) = [w_{d,n-L}, w_{d,n-L+1}, \ldots, w_{d,n+L-1}, w_{d,n+L}]$$

Can truncate context appropriately if window stretches past beginning or end of text.

In line with distributional hypothesis, word embedding models seek to generate similar embeddings for words that share similar contexts.

Self-Supervised Learning

The 'meaning' of a word is an unobserved and subjective concept.

Difficult to directly formulate an objective function.

Important conceptual idea is to formulate word prediction tasks that are solved using word embeddings.

Although word embeddings are formulated to solve prediction problems, they are nevertheless useful for the primary task of revealing meaning.

The approach of using auxiliary word prediction tasks to build high-quality embeddings is called self-supervised learning.

Prediction Tasks

There are two variants of word2vec which correspond to differing prediction tasks.

Skipgram model

1. Predict presence of each $w_{d,n-l} \in \mathrm{C}(w_{d,n})$ given $w_{d,n}$.
2. Predict absence of randomly sampled words from the corpus given $w_{d,n}$.

Continuous Bag-of-Words model

1. Predict presence of $w_{d,n}$ given $\mathrm{C}(w_{d,n})$.
2. Predict absence of randomly sampled words from the corpus given $\mathrm{C}(w_{d,n})$.

Words and Context in Skipgram Model

"economic growth is weak but long-term productivity trends are strong"

Suppose $L = 2$.

Positive Examples		**Negative Examples**	
Word	Context	Word	Context
economic	growth	economic	down
economic	is	economic	towards
growth	economic	growth	inflation
growth	is	growth	mild
growth	weak	growth	very
is	economic	is	not
is	growth	is	can
is	weak	is	rate
is	but	is	how
.	.	.	.
strong	are	strong	many

The number of negative examples to sample per positive example is a modeling choice.

Parametrization of the Prediction Problems

Endow each word v in the vocabulary with an embedding vector $\boldsymbol{\rho}_v \in \mathbb{R}^K$ and a context vector $\boldsymbol{\alpha}_v \in \mathbb{R}^K$ where $K \ll V$.

The positive examples are modeled as

$$\Pr\left[\, w_{d,n-l} \in \mathrm{C}(w_{d,n}) \mid w_{d,n} \,\right] = \frac{\exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}{1 + \exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}$$

and the negative examples are modeled as

$$\Pr\left[\, w_{d,n-l} \notin \mathrm{C}(w_{d,n}) \mid w_{d,n} \,\right] = 1 - \frac{\exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}{1 + \exp\left(\boldsymbol{\rho}_{w_{d,n}}^T \boldsymbol{\alpha}_{w_{d,n-l}}\right)}$$

Example

The first row of the table above would contribute the following elements to the loss function:

$$\Pr\left[\,\mathrm{growth} \in \mathsf{C}(w_{d,n}) \mid w_{d,n} = \mathrm{economic}\,\right] = \frac{\exp\left(\boldsymbol{\rho}_{\mathrm{economic}}^{T}\boldsymbol{\alpha}_{\mathrm{growth}}\right)}{1+\exp\left(\boldsymbol{\rho}_{\mathrm{economic}}^{T}\boldsymbol{\alpha}_{\mathrm{growth}}\right)}$$

$$\Pr\left[\,\mathrm{down} \notin \mathsf{C}(w_{d,n}) \mid w_{d,n} = \mathrm{economic}\,\right] = \frac{1}{1+\exp\left(\boldsymbol{\rho}_{\mathrm{economic}}^{T}\boldsymbol{\alpha}_{\mathrm{down}}\right)}$$

Loss function multiplies all such probabilities together and optimizes using gradient methods.

Terms Close to Uncertainty in FOMC Transcripts

term	sim
uncertainties	0.741
anxiety	0.48
pessimism	0.479
skepticism	0.465
optimism	0.445
caution	0.442
gloom	0.437
uncertain	0.433
sensitivity	0.427
angst	0.426

term	sim
challenges	0.415
fragility	0.405
clarity	0.401
concerns	0.4
risks	0.397
disagreement	0.387
volatility	0.384
tension	0.383
certainty	0.382
skepticism	0.38

Terms Close to Risk

term	sim
risks	0.737
threat	0.609
danger	0.541
dangers	0.463
vulnerability	0.457
chances	0.451
breakout	0.433
probability	0.426
possibility	0.409
likelihood	0.406

term	sim
misdirected	0.385
odds	0.379
uncertainty	0.375
concern	0.371
prospect	0.37
instability	0.363
potentially	0.352
concerns	0.352
challenges	0.346
risking	0.342

Importance of Training Corpus

Relationships among words can vary depending on the training corpus.

Example of training word embeddings on Wiki/Newswire text and on Harvard Business Review.

team		leader	
HBR	Generic	HBR	Generic
teams	teams	leadership	leaders
project_team	squad	leaders	leadership
management_team	players	manager	party
executive_team	football	person	opposition
group	coach	strong_leader	led
staff	league	chief_executive	rebel

Embeddings and Cultural Attitudes [Garg et al., 2018]

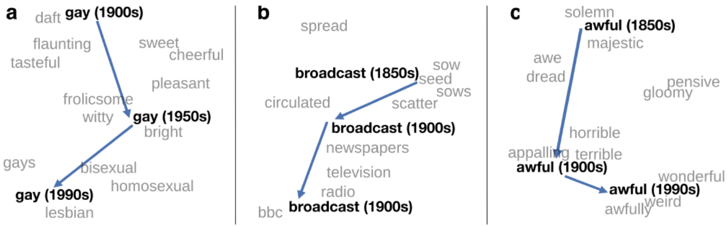
Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Table 3. Top Asian (vs. White) adjectives in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Irresponsible	Disorganized	Inhibited
Envious	Outrageous	Passive
Barbaric	Pompous	Dissolute
Aggressive	Unstable	Haughty
Transparent	Effeminate	Complacent
Monstrous	Unprincipled	Forceful
Hateful	Venomous	Fixed
Cruel	Disobedient	Active
Greedy	Predatory	Sensitive
Bizarre	Boisterous	Hearty

Evolution of Word Meanings [Hamilton et al., 2016]



Concept Detection

Expanding Dictionaries

One application of word embeddings is to augment human judgment in the construction of dictionaries.

Motivation is that economists are experts in which concept might be most important in a particular setting, but not in which words relate to that concept.

One can specify a set of ‘seed’ words and then find nearest neighbors of those words to populate a dictionary.

Strategy adopted by several recent papers:

1. [Hanley and Hoberg, 2019]
2. [Li et al., 2021]
3. [Bloom et al., 2021]
4. [Davis et al., 2020]

Embedding Dictionaries

Dictionaries provide a coarse representation of concepts in that some relevant terms might be missing altogether, and strength of association with concept isn't accounted for.

One strategy is to measure the association between documents and word lists in an embedding space rather than the bag-of-words space.

Recent example is [Gennaro and Ash, 2022] which studies emotional language in politics using the Congressional Record corpus.

Set A of words represents emotion, and set C of words represents cognition (both from LIWC).

Emotionality of speech i is

$$Y_i = \frac{\mathrm{sim}(\boldsymbol{d}_i, \boldsymbol{A}) + b}{\mathrm{sim}(\boldsymbol{d}_i, \boldsymbol{C}) + b}$$

Results

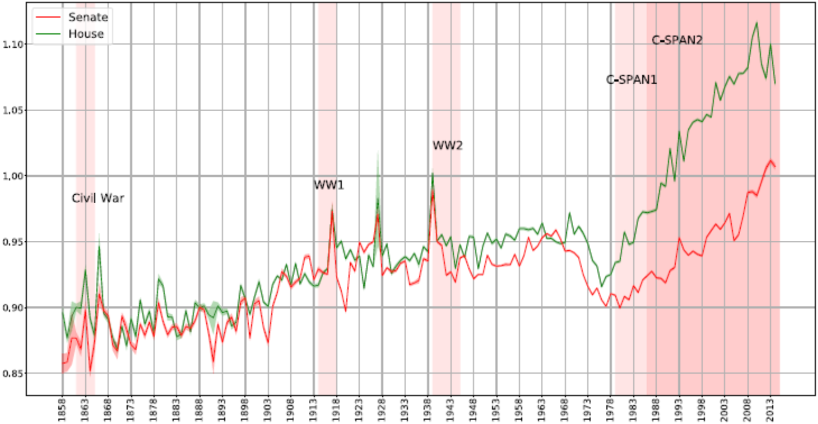
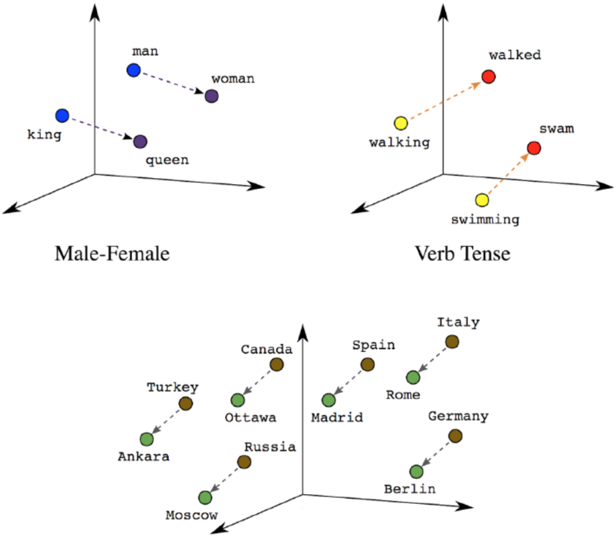


Fig. 2. *Emotionality in U.S. Congress by Chamber, 1858–2014.*
Notes: Time series of emotionality in the Senate (red) and the House of Representatives (green).

Relationship Among Concepts

Directions Encode Meaning



Male-Female

Verb Tense

Country-Capital

Word Embeddings and Cultural Attitudes

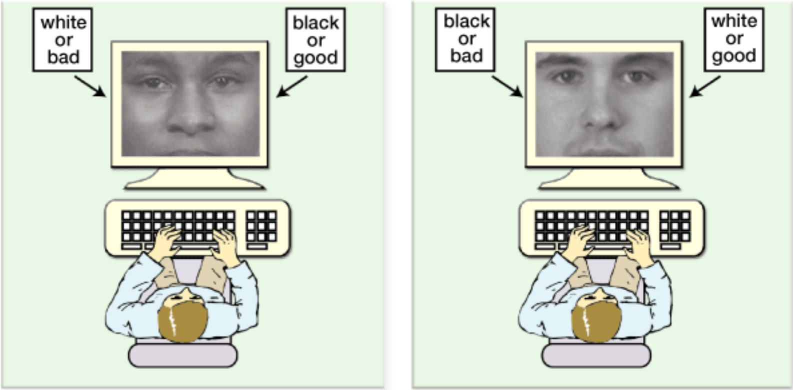
Because word embeddings appear to capture semantically meaningful relationships among words, there is interest in using them to measure cultural attitudes.

In psychology there is a long-standing Implicit Association Test that measures participants' time to correctly classify images depending on word combinations.

The hypothesis is that reaction times are shorter when word combinations more naturally belong together, which allows a measure of bias.

[Caliskan et al., 2017] have use word embeddings to ask whether similar biases exist in natural language.

Implicit Association Test



Word-Embedding Association Test

The Word-Embedding Association Test (WEAT) measures whether two sets of target words X, Y (e.g. male, female words) differ in their relative similarity to two sets of attribute words A, B (e.g. career, family words).

Let $\cos(\mathbf{x}, \mathbf{y})$ be cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$.

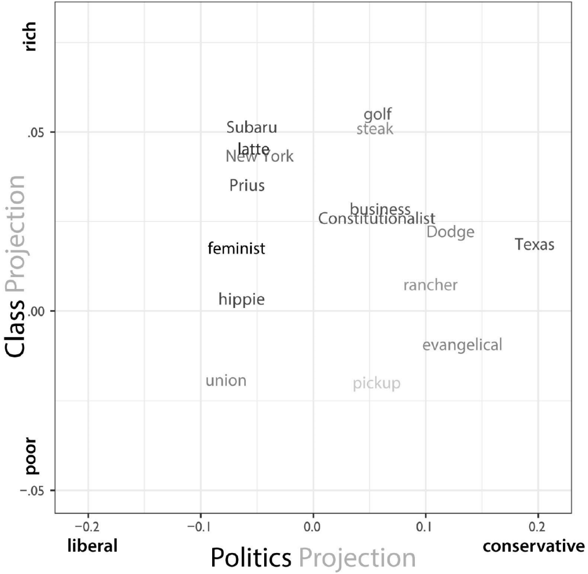
Let $s(\mathbf{w}, A, B) = \text{mean}_{\mathbf{a} \in A} \cos(\mathbf{w}, \mathbf{a}) - \text{mean}_{\mathbf{b} \in B} \cos(\mathbf{w}, \mathbf{b})$.

$$\text{WEAT} = \frac{\sum_{\mathbf{x} \in X} s(\mathbf{x}, A, B) - \sum_{\mathbf{y} \in Y} s(\mathbf{y}, A, B)}{\text{std}_{\mathbf{x} \in X \cup Y}\, s(\mathbf{x}, A, B)}$$

IAT vs WEAT

Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(5)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(5)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(5)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (5)	(7)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (9)	(7)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(9)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(9)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(9)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Language and Culture [Kozlowski et al., 2019]



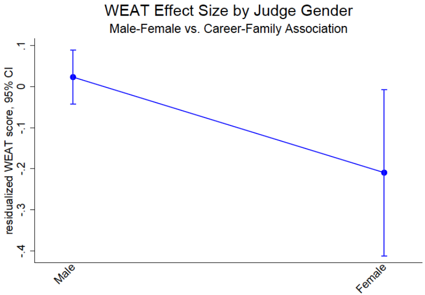
Does Language affect Decisions?

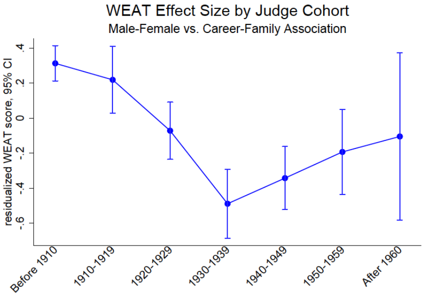
[Ash et al., 2024] use a measure similar to WEAT to measure linguistic gender bias among judges using written opinions.

They then match judge-specific bias scores with individual judge decisions to see whether the two are related.

Data is the universe of US appellate court decisions from 1890-2013.

WEAT and Judge Characteristics





Effects of WEAT

Judges with higher lexical bias are:

- Less likely to cast vote in favor of women's interests
- More likely to vote more conservatively across all issues
- Less likely to cite women in their opinions
- More likely to reverse female district judges

Document Similarity

Embedding-Based Similarity

Several papers use the distance between documents as captured by average embedding vectors.

[Kogan et al., 2019] measures distance between patents and occupation descriptions to proxy exposure of jobs to technical change.

[Hansen et al., 2021] measures distance between O*NET occupation descriptions and job postings to proxy skill demand.

Word2Vec Summary

Word2Vec introduces several ideas that remain influential:

1. Words as low-dimensional embedding vectors.
2. Self-supervised learning using auxiliary word-prediction tasks to build informative representations of language.
3. Neural network estimation in place of statistical models.
4. Surprising behavior of estimated latent meaning space.

One important limitation: vectors are built using local context but they do not vary with local context.

References I

Ash, E., Chen, D. L., and Ornaghi, A. (2024).
Gender Attitudes in the Judiciary: Evidence from US Circuit Courts.
American Economic Journal: Applied Economics, 16(1):314–350.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003).
A neural probabilistic language model.
The Journal of Machine Learning Research, 3(null):1137–1155.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
The Journal of Machine Learning Research, 3(null):993–1022.

Bloom, N., Hassan, T. A., Kalyani, A., Lerner, J., and Tahoun, A. (2021).
The Diffusion of Disruptive Technologies.
Working Paper 28999, National Bureau of Economic Research.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017).
Semantics derived automatically from language corpora contain human-like biases.
Science, 356(6334):183–186.

References II

Davis, S. J., Hansen, S., and Seminario-Amez, C. (2020).
Firm-Level Risk Exposures and Stock Returns in the Wake of COVID-19.
Working Paper 27867, National Bureau of Economic Research.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990).
Indexing by latent semantic analysis.
Journal of the American Society for Information Science, 41(6):391–407.

Ding, C., Li, T., and Peng, W. (2006).
Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method.
In Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1, AAAI'06, pages 342–347, Boston, Massachusetts. AAAI Press.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018).
Word embeddings quantify 100 years of gender and ethnic stereotypes.
Proceedings of the National Academy of Sciences, 115(16):E3635–E3644.

References III

Gennaro, G. and Ash, E. (2022).
Emotion and Reason in Political Language.
The Economic Journal, 132(643):1037–1059.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016).
Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.
In Erk, K. and Smith, N. A., editors, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hanley, K. W. and Hoberg, G. (2019).
Dynamic Interpretation of Emerging Risks in the Financial Sector.
The Review of Financial Studies, 32(12):4543–4603.

Hansen, S., McMahon, M., and Prat, A. (2018).
Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach.
The Quarterly Journal of Economics, 133(2):801–870.

References IV

Hansen, S., Ramdas, T., Sadun, R., and Fuller, J. (2021).
The Demand for Executive Skills.
Technical Report 28959, National Bureau of Economic Research, Inc.

Iaria, A., Schwarz, C., and Waldinger, F. (2018).
Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science.
The Quarterly Journal of Economics, 133(2):927–991.

Ke, S., Olea, J. L. M., and Nesbit, J. (2021).
Robust Machine Learning Algorithms for Text Analysis.
Unpublished manuscript.

Kogan, L., Papanikolaou, D., Schmidt, L., and Seegmiller, B. (2019).
Technology, Vintage-Specific Human Capital, and Labor Displacement: Evidence from Linking Patents with Occupations.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019).
The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings.
American Sociological Review, 84(5):905–949.

References V

Li, K., Mai, F., Shen, R., and Yan, X. (2021).
Measuring Corporate Culture Using Machine Learning.
The Review of Financial Studies, 34(7):3265–3315.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a).
Efficient Estimation of Word Representations in Vector Space.
arXiv:1301.3781 [cs].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b).
Distributed Representations of Words and Phrases and their Compositionality.
arXiv:1310.4546 [cs, stat].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is All you Need.
In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.