Machine Learning for Textual and Unstructured Data

Lecture 3:
Word Embeddings

Stephen Hansen
University College London

Introduction

The dimensionality reduction methods in the previous slides rely on global co-occurrence patterns.

In natural language, local context is arguably a more natural guide to meaning: "You shall know a word by the company it keeps" (Firth).

Suppose we wish to understand the meaning of the 'alien word' $%& which has been inserted in place of an English word in the following sentence:

> *Every morning last summer in Greece, I visited the $%& where I would swim, play in the sand, and sunbathe.*

The words present in this text snippet give strong clues about the meaning of $%&.

Words present many sentences away would generally be less informative.

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 has been a major topic in NLP in the past decade.

Embedding vectors can be of interest in their own right, or else form part of the representation of a document for other tasks.

Local Contexts and Word Embeddings

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 $w_{d,n}$:

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

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

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

GloVe

The GloVe model [?] begins with a $V \times V$ matrix $\mathbf{W}$ of local word co-occurrences.

W_{ij} is the number of times term j appears within the context of i.

Assign to each term v an embedding vector $\boldsymbol{\rho}_v \in \mathbb{R}^V$.

$$\min \sum_{i,j} f(W_{i,j}) \left(\boldsymbol{\rho}_i^T \boldsymbol{\rho}_j - \log\left(W_{i,j}\right)\right)^2$$

Terms that co-occur frequently will have more highly correlated embedding vectors.

Word2Vec

Word2vec [Mikolov et al., 2013a, Mikolov et al., 2013b] is another well-known model for word embeddings that incorporates local context.

In addition to an embedding vector, each term is assigned a context vector $\boldsymbol{\alpha}_v \in \mathbb{R}^V$.

Word vectors are chosen to solve word-prediction tasks:

$$\Pr[\, w_{d,n} = v \mid C(w_{d,n}) \,] = \frac{\exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_v)}{\sum_{v'} \exp(\overline{\boldsymbol{\alpha}}_{d,n}^T \boldsymbol{\rho}_{v'})} \text{ where } \overline{\boldsymbol{\alpha}}_{d,n} = \frac{1}{2L} \sum_{w \in C(w_{d,n})} \boldsymbol{\alpha}_w$$

Example of self-supervised learning.

The version of Word2Vec is the continuous-bag-of-words model; the skip-gram model instead predicts context words given a center word.

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

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

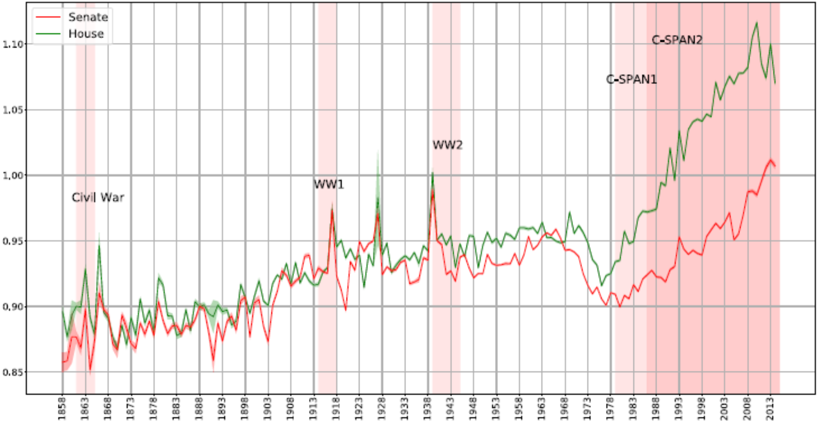


Fig. 2. *Emotionality in U.S. Congress by Chamber, 1858–2014.*
Notes: Time series of emotionality in the Senate (red) and the House of Representatives (green).

Transfer Learning

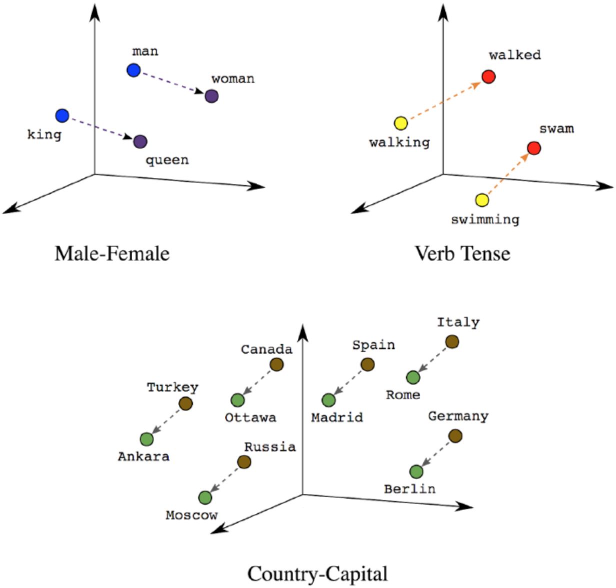
In the above examples, Word2vec is fit directly to the data of interest.

In many use cases, one instead uses an estimated model from an auxiliary corpus for word embeddings, or as starting values in embedding estimation.

This is known as transfer learning and becomes particularly important for large-scale language models.

Relationship Among Concepts

Directions Encode Meaning



Word Embeddings and Cultural Attitudes

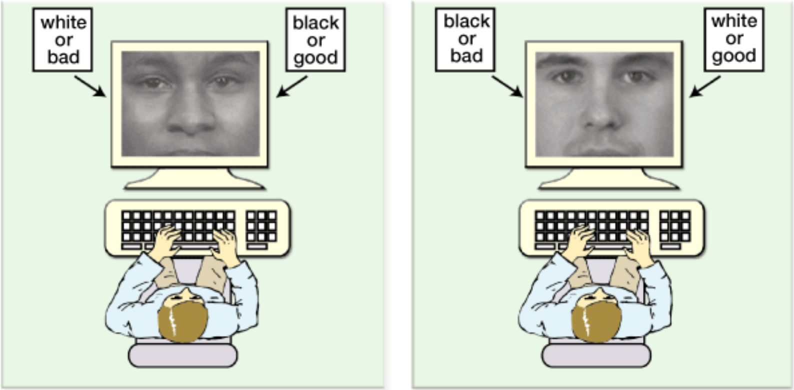
Because word embeddings appear to capture semantically meaningful relationships among words, there is interest in using them to measure cultural attitudes.

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 $\mathbf{y}$.

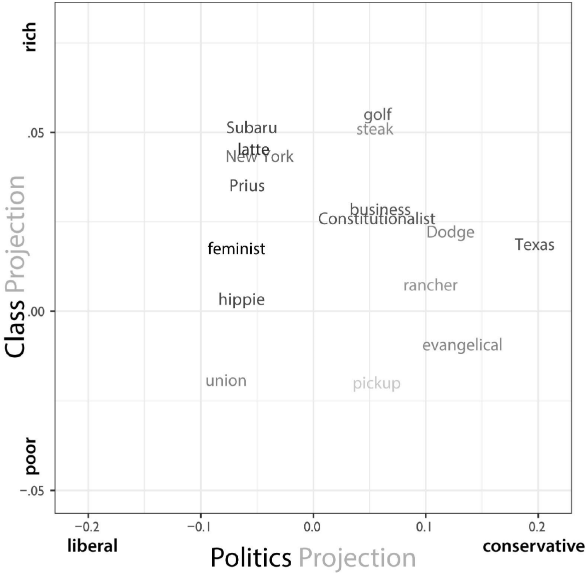
Let $s(\mathbf{w}, A, B) = \mathrm{mean}_{\mathbf{a} \in A} \cos(\mathbf{w}, \mathbf{a}) - \mathrm{mean}_{\mathbf{b} \in B} \cos(\mathbf{w}, \mathbf{b})$.

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

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]



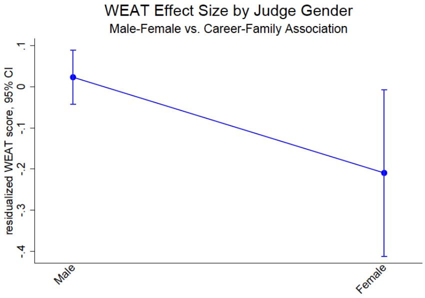
Does Language affect Decisions?

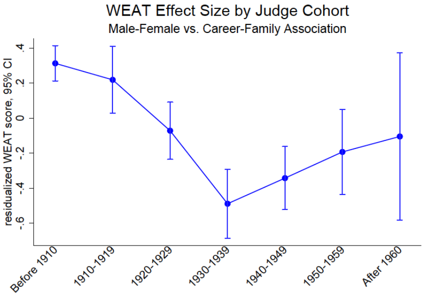
[Ash et al., 2020] use a measure similar to WEAT to measure linguistic gender bias among judges using written opinions.

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.

Choosing Among Algorithms

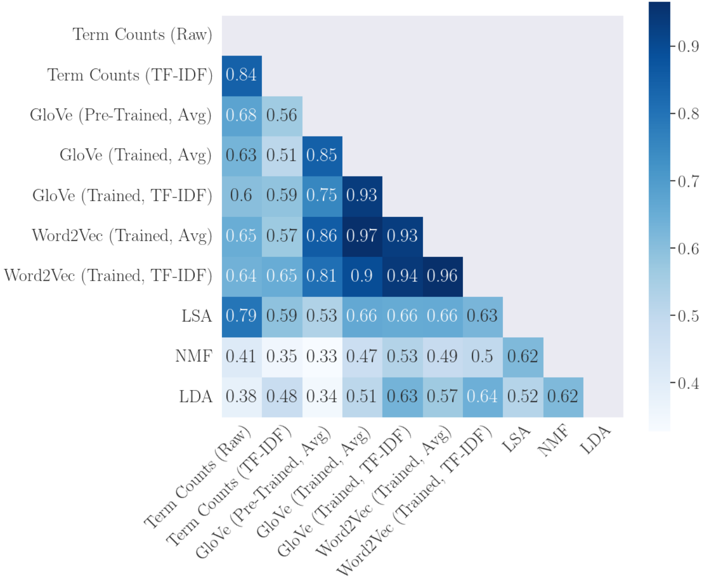
By now we have seen multiple algorithms for document similarity, but provided no means to assess which one to choose.

Does the choice of algorithm matter?

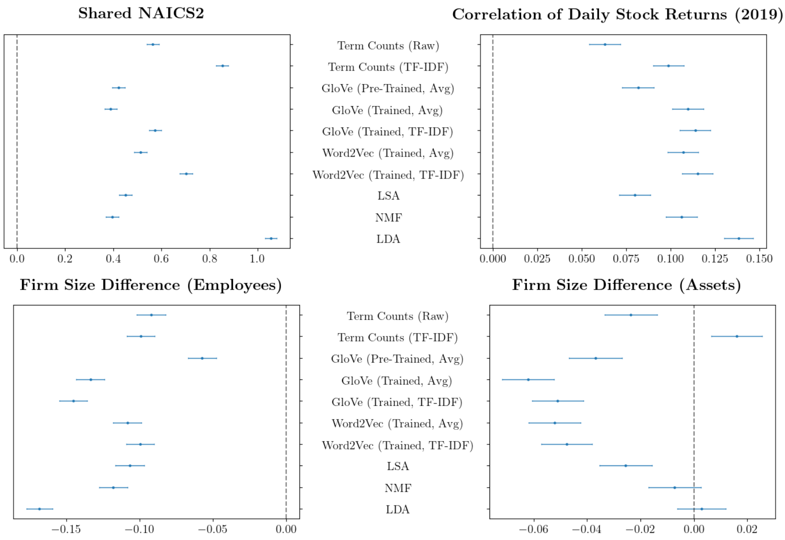
We evaluate document similarity in the context of 10-K risk factors using randomly sampled pairs from the universe of 2019 filing firms.

Keep data constant, and vary the algorithm used for similarity comparison.

Pearson Correlation Between Similarity Scores Across Pairs



Downstream Regression Results



How to Proceed?

Clear need for some validation task against which to compare these algorithms.

Unclear that existing ideas from computer science are relevant in economics.

Inevitable need for some human input to judge some baseline truth.

Ideally one would find validation tasks that were relevant across research questions so that each author doesn't have to start from scratch.

Topic Modeling for Embeddings

Incorporating Document Heterogeneity

Topic models express heterogeneity across documents within the context of the bag-of-words model.

Word embeddings exploit local context to construct a semantically meaningful vector space.

Can we combine the strengths of the two approaches?

[Dieng et al., 2020] present the embedded topic model.

Statistical Model

1. Draw topic proportions $\boldsymbol{\theta}_d \sim \mathcal{LN}(0, 1)$.
2. For each word $w_{d,n}$:
    2.1 Draw a topic assignment $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d, 1)$.
    2.2 $w_{d,n} \sim \text{Mult}\left(\text{softmax}(\mathbf{P}^T \boldsymbol{\alpha}_{z_{dn}}), 1\right)$.

Where $\mathbf{P}$ is a $K \times V$ matrix whose vth column is the word embedding $\boldsymbol{\rho}_v$.

Topics are now represented by vectors in the embedding space.

Topic in Embedding Space

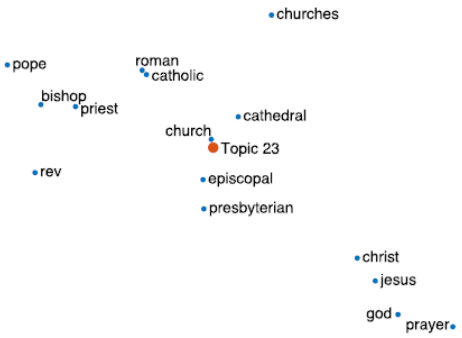


Figure 2: A topic about Christianity found by the ETM on *The New York Times*. The topic is a point in the word embedding space.

Most Common Topics

ETM

game	music	united	wine	company	yankees	art
team	mr	israel	food	stock	game	museum
season	dance	government	sauce	million	baseball	show
coach	opera	israeli	minutes	companies	mets	work
play	band	mr	restaurant	billion	season	artist

Table 3: Top five words of seven most used topics from different document models on 1.8M documents of the *New York Times* corpus with vocabulary size 212,237 and $K = 300$ topics.

Performance vs LDA

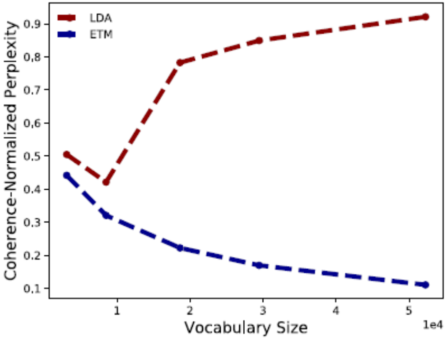


Figure 1: Ratio of the held-out perplexity on a document completion task and the topic coherence as a function of the vocabulary size for the ETM and LDA on the *20NewsGroup* corpus. The perplexity is normalized by the size of the vocabulary. While the performance of LDA deteriorates for large vocabularies, the ETM maintains good performance.

Embedding Items from Consumer Choice

Contextual Data

More generally, embeddings model data given its context.

This idea extends well beyond text.

One recent area of application is to consumer choice data.

'word' is replaced by an item that a consumer is observed to purchase.

'context' is replaced by the other items a consumer has bought.

See auxiliary slides for illustration to movie data.

SHOPPER Model

[Ruiz et al., 2020] builds a probabilistic model of consumer choice that incorporates context.

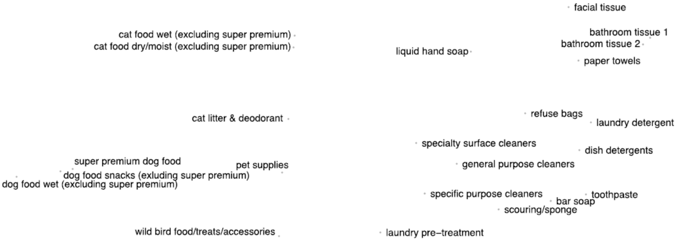
Basic choice probability is

$$\Pr\left[\,\text{item } c \mid \text{items in basket}\,\right] \propto \exp\left\{\boldsymbol{\theta}_u^T \boldsymbol{\alpha}_c + \boldsymbol{\rho}_c\left(\frac{1}{i-1}\sum_{j=1}^{i-1}\boldsymbol{\alpha}_{y_j}\right)\right\}$$

Extended choice probability incorporates prices, seasonal effects.

Likelihood formed by summing over unobserved ordering of product choices.

Model competitive for predicting how consumer choice reacts to price changes.



(a) Pet food and supplies. (b) Cleaning and hygiene.

FIG. 3. *Two regions of the two-dimensional* T-SNE *projection of the features vectors* α_c *for the category-level experiment.*

TABLE 9

Items with the highest complementarity and lowest exchangeability metrics for some query items

Query items	Complementarity score		Exchangeability score	
Mission tortilla soft taco 1	2.40	taco bell taco seasoning mix	0.05	mission fajita size
	2.26	mcrmck seasoning mix taco	0.07	mission tortilla soft taco 2
	2.24	lawrys taco seasoning mix	0.13	mission tortilla fluffy gordita
Private brand hot dog buns	2.99	bp franks meat	0.11	ball park buns hot dog
	2.63	bp franks bun size	0.13	private brand hotdog buns potato 1
	2.37	bp franks beed bun length	0.15	private brand hotdog buns potato 2
Private brand mustard squeeze bottle	0.50	private brand hot dog buns	0.15	frenchs mustard classic yellow squeeze
	0.41	private brand cutlery full size forks	0.16	frenchs mustard classic yellow squeezed
	0.24	best foods mayonnaise squeeze	0.21	heinz ketchup squeeze bottle
Private brand napkins all occasion	0.78	private brand selection plates 6 7/8 in	0.09	vnty fair napkins all occasion 1
	0.50	private brand selection plates 8 3/4 in	0.11	vnty fair napkins all occasion 2
	0.49	private brand cutlery full size forks	0.12	private brand selection premium napkins

Conclusion

Embedding models for language learn vector representations for words that depend on local context.

Word2vec was an important milestone for demonstrating how to estimate a neural language model that produced semantically coherent embeddings.

Natural next step: embedding text sequences.

References I

Ash, E., Chen, D. L., and Ornaghi, A. (2020).
Gender attitudes in the judiciary : Evidence from U.S. circuit courts.
https://warwick.ac.uk/fac/soc/economics/research/workingpapers/2020/twerp_1256_-_ornaghi.pdf.

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.

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.

Dieng, A. B., Ruiz, F. J. R., and Blei, D. M. (2020).
Topic Modeling in Embedding Spaces.
Transactions of the Association for Computational Linguistics, 8:439–453.

References II

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

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., Ramdas, T., Sadun, R., and Fuller, J. (2021).
The Demand for Executive Skills.

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 III

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].

Ruiz, F. J. R., Athey, S., and Blei, D. M. (2020).
SHOPPER: A probabilistic model of consumer choice with substitutes and complements.
The Annals of Applied Statistics, 14(1):1–27.