Text as Data for the Social Sciences

Germain Gauthier

ETH Zürich
Ecole Polytechnique, CREST

IOEA – May 16, 2023

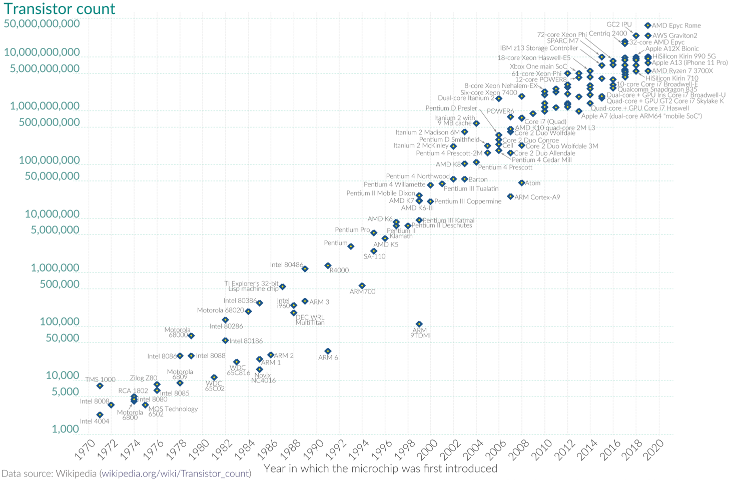
The Rise of Text Data

- The digital era generates considerable amounts of text.
  - Social media and internet queries
  - Wikipedia, online newspapers, TV transcripts
  - Digitized books, speeches, laws
- It is matched with a similar increase in computational resources.
  - Moore’s law = processing power of computers doubles every two years (since the 70s!)

Moore's Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Our World in Data



Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.

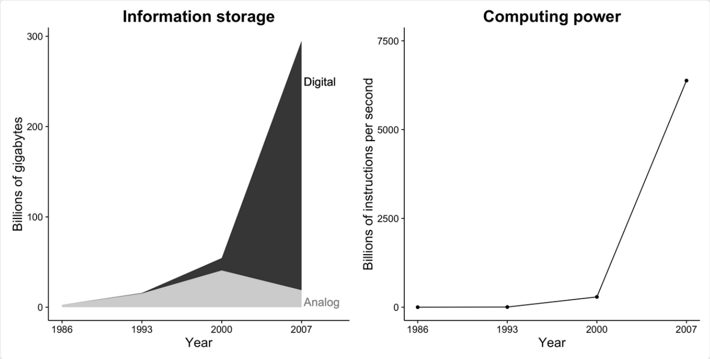


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

Source: Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

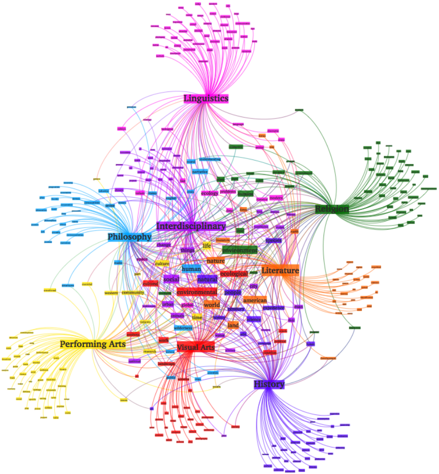
Natural Language Processing

- Natural language processing is a *data-driven* approach to the analysis of text documents.
- Applications in your everyday life:
  - Search engines, translation services, spam detection
- Applications in the social sciences:
  - Measuring economic policy uncertainty, news sentiment, racial and misogynistic bias, political and economic narratives, speech polarization
  - Predicting protests, GDP growth, financial market fluctuations

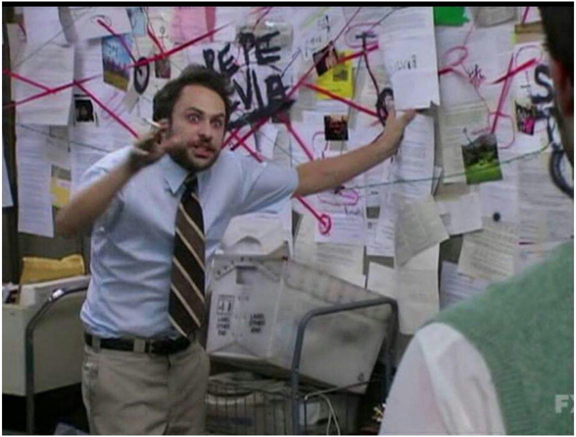
Today's Roadmap

- Overview of text as data **in theory** and **in practice**.
- We will look at text as data through the history of its **methods**:
  1. Bag-of-words
  2. Static embeddings
  3. Sequence embeddings
- For each method, we will focus on **applications for social scientists**.
- We will end with more general considerations on the **research frontier**.

We want to do cool graphs like the one below...



But for this you have to bear with me...



Plan

Bag-of-words

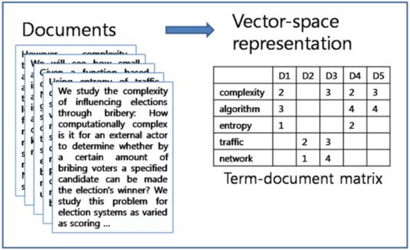
Static Embeddings

Sequence Embeddings

Issues and Challenges

Counting Words

- The simplest way to represent text documents is word frequencies.
- This is referred to as the **bag-of-words** approach.
- The corpus is featurized as a term-document matrix **W**:



Documents

However, complexity

We will see how small

Given a function based

Using entropy of traffic

We study the complexity of influencing elections through bribery: How computationally complex is it for an external actor to determine whether by a certain amount of bribing voters a specified candidate can be made the election's winner? We study this problem for election systems as varied as scoring ...

Vector-space representation

	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

Some Refinements

- The general unit for counts is called a **token**.
- The final set of tokens considered is the **vocabulary**.
- Tokens can also include symbols and digits.

  e.g., #, !, ?, haha, 2008, etc.
- Depending on the application, some tokens are uninformative and can be removed (i.e., stopwords).

  e.g., she, he, the, a, etc.
- Some tokens mean the same thing and can be grouped together (via stemming or lemmatization).

  e.g., animal and animals, eating and eat, etc.

So what can we do with W?

Dictionaries

- The dictionary approach consists of:
    - A *narrow* vocabulary based on a set of pre-defined tokens.
    - A *deterministic* mapping f from the features W to the outcomes Y.
- Extensively used for sentiment analysis:
    - Let (w_i, s_i) be pairs of words w_i and their associated sentiment score $s_i \in [-1, 1]$.

      e.g., ("perfect", 0.8), ("awful", -0.9)
    - The sentiment score for any phrase j of k tokens is a weighted average:

$$s_j = \frac{1}{K} \sum_{i=1}^{k} s_i.$$

Application – “Gilets Jaunes” Facebook Interactions

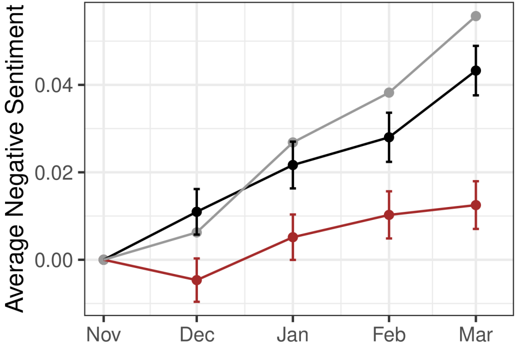
Examples of the most positive sentences:

honneur gilet jaune
mdr
bravo
mercii jeune meilleur facon aider progres meilleur monde
bravo gabin media honnete souhaite reussite merite equipe bravo gj

Examples of the most negative sentences:

macron demission
macron cabanon castananer enfer
florence menteur
bande pourriture batard
castaner assassin degage voleur menteur

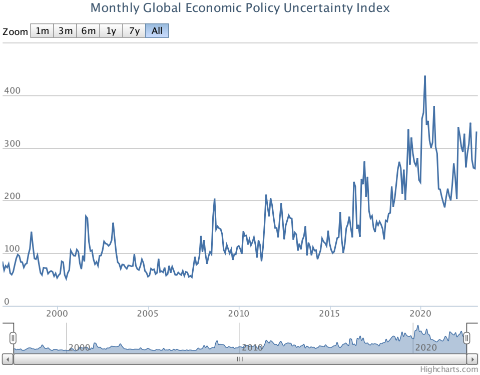
Figure: Moderate users left, and those who remained radicalized.



Notes: The regression equation is $Y_{i,t} = \delta_i + \gamma_t + \varepsilon_{i,t}$. The red line is the composition effect. The black line is individual-level radicalization effect. The grey line is the total observed trend in the data.

Source: Social Media and the Dynamics of Protests, Boyer et al. (2023).

Application – Economic Policy Uncertainty Index



Notes: Normalized counts of articles containing “uncertainty” or “uncertain”; “economic” or “economy”; “congress” or “deficit” or “Federal Reserve” or “legislation” or “regulation” or “white house” in ten major US newspaper outlets.

Source: Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics, 131(4)*, 1593-1636.

Text Regressions

- We can generalize dictionaries with a text regression:

$$Y_i = W_i^T \beta + \varepsilon_i.$$

- But W can be very large as the vocabulary size grows.
- A penalized regression is often required to estimate β:

$$\min_{\beta} \Big(\sum_{i=1}^{n} Y_i - W_i^T \beta \Big)^2 \text{ such that } \sum_{j=1}^{V} \beta_j^2 \leq \lambda.$$

- Other common approches: random forests and support vector machines.
- Other common transforms: logistic and multinomial logistic.

Application – Speech Polarization in the U.K.

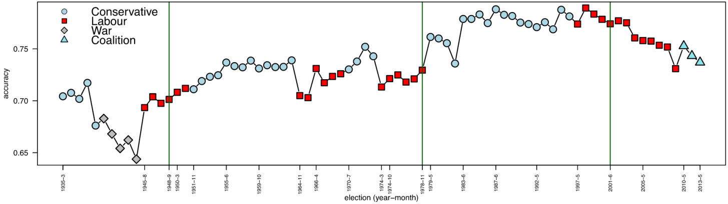


Figure 3. Estimates of parliamentary polarization, by session. Election dates mark *x*-axis. Estimated change points are [green] vertical lines.

Notes: Predictive accuracy of Ridge regressions by parliamentary session.

Source: Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *Political Analysis, 26(1)*, 120-128.

Topic Models

- In some cases, we do not have the labels and would like to regroup similar documents together.
- Topics models infer *latent* topics in the corpus:
  - Documents as distributions over topics
  - Topics as distributions over words
- Formally, $\mathbf{W}$ is decomposed into two matrices:

$$\mathbf{W} = \mathbf{\Theta} \times \mathbf{B}^T$$

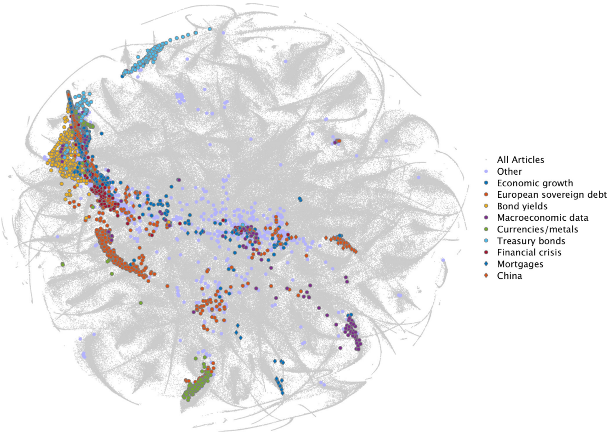
  where $\mathbf{W} \in D \times V$, $\mathbf{\Theta} \in D \times K$, and $\mathbf{B} \in V \times K$.
- Often, priors are used to induce sparsity in $\mathbf{\Theta}$ and $\mathbf{B}$.
  - For instance, as its name suggests, Latent Dirichlet Allocation (LDA) assumes Dirichlet priors.

"We refer to the latent multinomial variables in the LDA model as topics so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words."

Source: Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993-1022.

Application – Analyzing business news...

Figure 14: Articles Featuring the Federal Reserve



Source: Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021). *Business news and business cycles* (No. w29344). National Bureau of Economic Research.

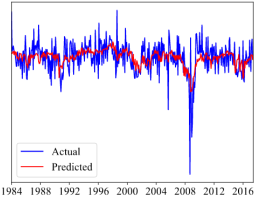
To predict macroeconomic variables.

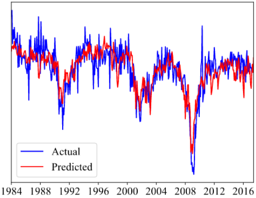
Industrial Production Growth

Topic	Coeff.	p-val.
Recession	-0.38	0.00
Oil market	-0.17	0.00
Southeast Asia	0.11	0.10
Health insurance	0.06	0.93
Clintons	0.03	0.40
In-Sample R^2	0.21	
Out-of-Sample R^2	0.06	

Employment Growth

Topic	Coeff.	p-val.
Recession	-0.61	0.00
Rail/trucking/shipping	0.22	0.01
Bush/Obama/Trump	-0.15	0.09
Iraq	-0.14	0.01
Clintons	0.12	0.01
In-Sample R^2	0.59	
Out-of-Sample R^2	0.48	





Source: Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021). *Business news and business cycles* (No. w29344). National Bureau of Economic Research.

Limitations of Bag-of-words

- Easy, transparent, the bag-of-words approach has led to tens of thousands of publications... But it comes with clear limitations.
- No notion of **semantic proximity** between tokens. Tokens are all at the same distance of one another.
  - For example, *"sociology"* and *"economics"* are not considered more similar to one another than to *"hello"*.
- In other words, tokens are discrete features.
  - *"sociology"* and *"economics"* are completely distinct features for predicting whether a paper is about the social sciences.

What's the solution? *Static embeddings*!

Plan

Bag-of-words

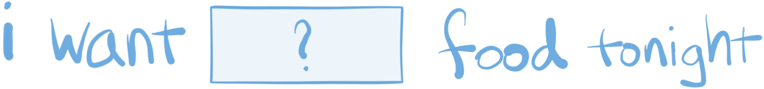
Static Embeddings

Sequence Embeddings

Issues and Challenges

Building Some Intuition

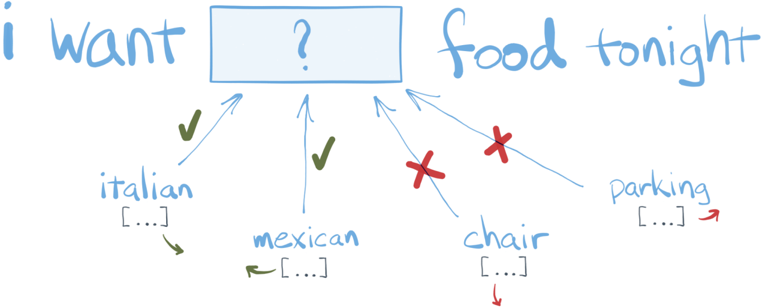
Figure: Can you complete this text snippet?



Source: Patrick Harrison, *S&P Global Market Intelligence*

Building Some Intuition

Figure: Can you complete this text snippet?



Source: Patrick Harrison, *S&P Global Market Intelligence*

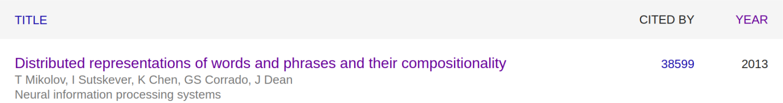
Language in Context (and vice-versa)

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

(J. R. Firth, 1957)

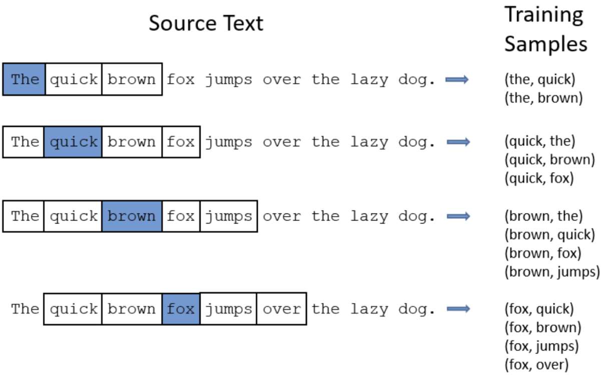
- Neighboring words provide us with additional information to interpret a word's meaning.
- In other words, **word co-occurrences capture context.**
- This information is useful for downstream applications.
- Let's look into one algorithm that learns word embeddings: `word2vec`.

word2vec



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	38599	2013

Training Samples

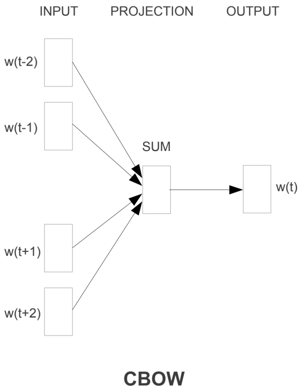


Source Text	Training Samples
The quick brown fox jumps over the lazy dog.	(the, quick) (the, brown)
The quick brown fox jumps over the lazy dog.	(quick, the) (quick, brown) (quick, fox)
The quick brown fox jumps over the lazy dog.	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown fox jumps over the lazy dog.	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

Notes: Window size $M = 2$.

Source: Julian Gilyadov.

`word2vec` – Intuition



Source: Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

word2vec – Formal Parametrization

- Let M be the size of the context window (e.g., $M = 10$).
- Assume that each word can be represented as a K-dimensional vector (e.g., $K = 300$).
- The probability of the focus word given its context words is given by the softmax function:

$$P(w_t|\{w_{t+j}\}_{-M\leq j\leq M, j\neq 0}) = \frac{\exp(w_t^T \bar{u}_t)}{\sum_{k=1}^{V} \exp(w_k^T \bar{u}_t)},$$

where $\bar{u}_t$ is the average of the context vectors for words in the context window and w vectors are word vectors.

- This parametrization forces the vectors of words that co-occur together to be close in the embeddings space.

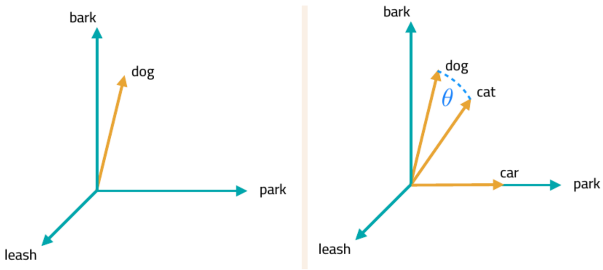
Distance Between Texts

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

$$\cos\theta = \frac{w_1^T w_2}{||w_1||||w_2||}$$

- 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)

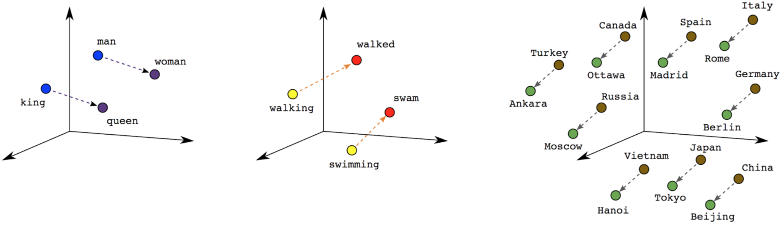
Distance Between Texts



Basic arithmetic often carries meaning.

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

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

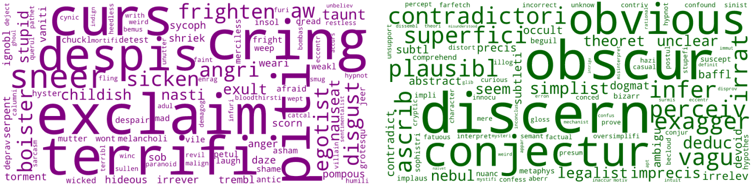


Male-Female

Verb Tense

Country-Capital

Application – Emotion and Reason in Politics


(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: Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. *The Economic Journal*, 132(643), 1037-1059.

Application – Emotion and Reason in Politics

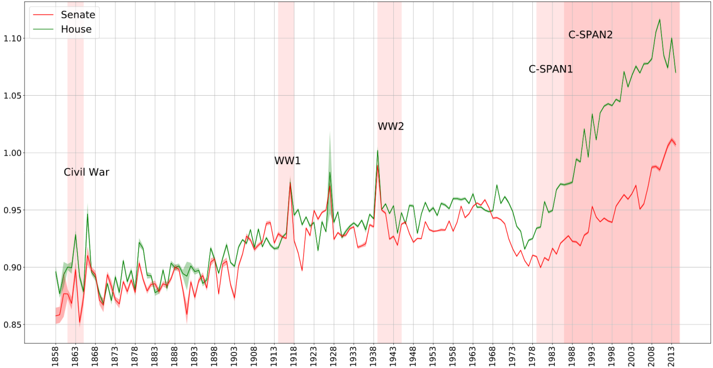


Fig. 2. *Emotionality in U.S. Congress by Chamber, 1858–2014.*
Notes: Time series of emotionality in the Senate (red) and the House of Representatives (green).

Source: Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. *The Economic Journal*, 132(643), 1037-1059.

Limitations of Static Embeddings

- `word2vec` assumes that **context words are exchangeable**.
  - The ordering of words is *not* accounted for.
  - All context word vectors are weighted *equally* to predict the focus word.
- Intuitive example:
  - "As a leading firm in the [MASK] sector, we hire highly skilled software engineers."
  - "As a leading firm in the [MASK] sector, we hire highly skilled petroleum engineers."

Limitations of Static Embeddings

- `word2vec` assumes that **context words are exchangeable**.
  - The ordering of words is *not* accounted for.
  - All context word vectors are weighted *equally* to predict the focus word.
- Intuitive example:
  - "As a leading firm in the **tech** sector, we hire highly skilled software engineers."
  - "As a leading firm in the **energy** sector, we hire highly skilled petroleum engineers."

Limitations of Static Embeddings

- `word2vec` assumes that **context words are exchangeable**.
  - The ordering of words is *not* accounted for.
  - All context word vectors are weighted *equally* to predict the focus word.
- Intuitive example:
  - "As a leading firm in the **tech** sector, we hire highly skilled software engineers."
  - "As a leading firm in the **energy** sector, we hire highly skilled petroleum engineers."
- Clearly, some words matter more than others to predict the focus word.

Solution? *Get the models to pay attention!*

Plan

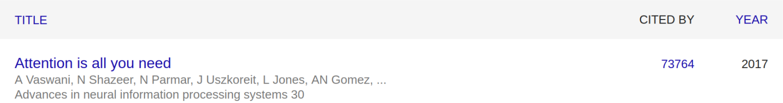
Bag-of-words

Static Embeddings

Sequence Embeddings

Issues and Challenges

“Attention is all you need.”



TITLE	CITED BY	YEAR
Attention is all you need A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, ... Advances in neural information processing systems 30	73764	2017

The Attention Mechanism

- **Self-attention functions** give some context words more weight than others in the focus word's vector representation.
- Depending on its surrounding words, the same word will not have the same vector representation anymore.
- Intuitive example:
  - "she filed suit under *class* action"
  - "she graduated top of *class*"

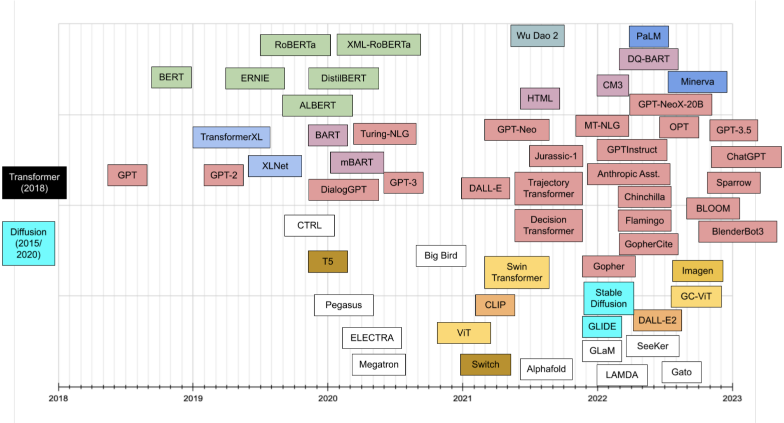
The Attention Mechanism

- Formally, let $(\rho^0_{d,1}, ..., \rho^0_{d,Nd})$ be the initial embeddings that make up a document.
- The new embedding at each position n is given by:

$$\rho^1_{d,n} = \sum_{n'=1}^{N_d} w_{(d,n),n'} \rho^0_{d,n'} \quad \text{such that} \quad \sum_{n'=1}^{N_d} w_{(d,n),n'} = 1.$$

  where $w_{(d,n),n'}$ are non-negative self-attention weights that determine the importance of a context word for a given focus word.
- Since 2017, a new class of models – **transformers** – learn attention parameters rather fast and at scale (Vaswani et al., 2017).

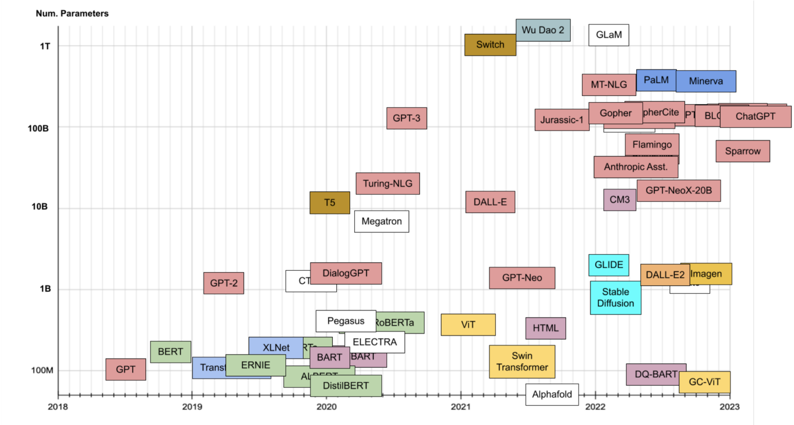
Different transformer families



Notes: Transformer timeline, with colors describing the Transformer family.

Source: Amatriain (2023).

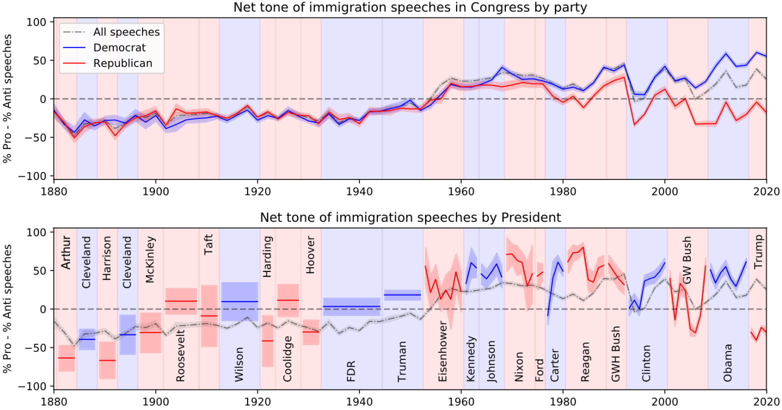
Different transformer families



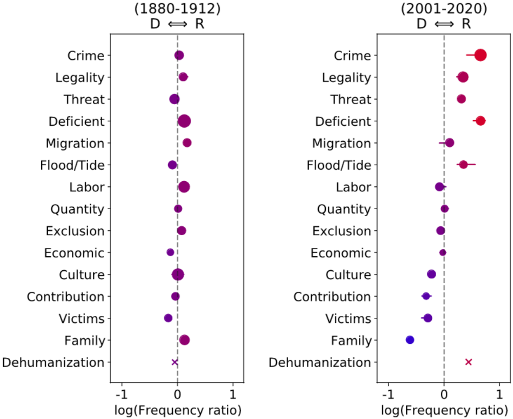
Notes: Transformer timeline. On the vertical axis: the number of parameters.

Source: Amatriain (2023).

Application – Framing Immigration in Political Discourse



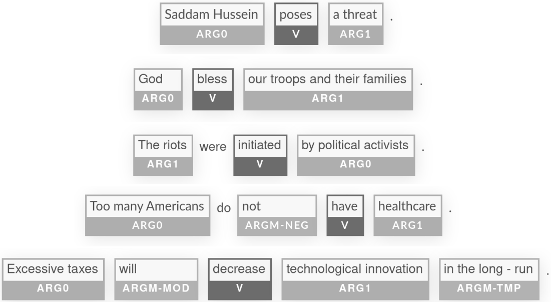
Source: Source: Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., ... & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences, 119(31)*, e2120510119.



Source: Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., ... & Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences, 119(31)*, e2120510119.

Application – Mining Narratives from Large Text Corpora

Figure: Examples of Semantic Role Labeling Annotations



Saddam Hussein ARG0 | poses V | a threat ARG1 | .

God ARG0 | bless V | our troops and their families ARG1 | .

The riots ARG1 | were | initiated V | by political activists ARG0 | .

Too many Americans ARG0 | do | not ARGM-NEG | have V | healthcare ARG1 | .

Excessive taxes ARG0 | will ARGM-MOD | decrease V | technological innovation ARG1 | in the long - run ARGM-TMP | .

Notes: See `https://demo.allennlp.org/semantic-role-labeling` for additional examples.

Source: Ash, E., Gauthier, G., & Widmer, P. (2023). Text semantics capture political and economic narratives. *Political Analysis*

- Consider these two sentences:
  - "Millions of Americans lost their unemployment benefits."
  - "Many Americans lost their much-needed unemployment benefits."
- A human recognizes that both sentences refer to the same underlying narrative:

$$\texttt{Americans} \xrightarrow{\texttt{lost}} \texttt{unemployment benefits}$$

→ We need some **dimension-reduction** strategy.

We consider two complementary approaches for dimension reduction.

- **Named Entity Recognition** (NER)
  - Identify named entities such as events, organizations, places, and people – i.e., proper nouns.
- **Clustering** of common entities with sequence embeddings
  - For the remaining entity strings, we first produce phrase embeddings.
  - We then cluster the embeddings, e.g. with K-means/DBSCAN, perhaps after dimension reduction.

Figure: Examples of Entities Based on Trump Tweets

taliban | people middle east mess | energy middle east | story respect policy middle east | bowe bergdahl | billion afghanistan | afghanis | karzai afghanistan | freed taliban hostage | taliban leader deserter | afghan casualties | iranians | policeman middle east

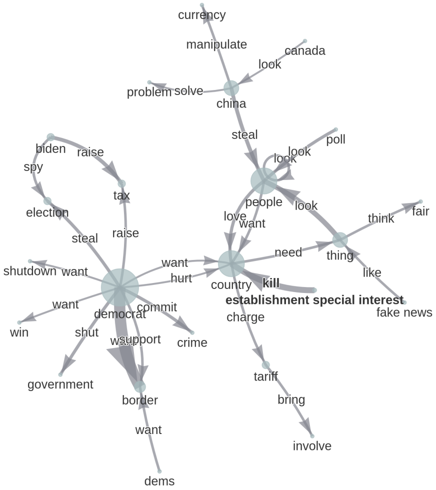
golf tournament | world golf championship cadillac | golf odyssey | golf | job architect golf course ferry point | golf baseball | finishing hole golf | golf major | ad golf turnberry | luxury villa golf course | object golf | time the open golf course

w h decision | testing | cross examination | testing thing world | watchdog | testing site | testing problem | drug test result | probe | fight individual mandate | checker | minute testing apparatus | marching order | observer | drug test | inspection grade | audit |

snowden amp secret | whistleblowers complaint | acting director national intelligence | allegation opponent spy | classified information nda | ex nsa contractor | snowden russiahe | cia director | accuser whistleblower | cia chiefs | standing whistleblower rules

hurricane harvey | storm hurricane | hurricane effect | katrina | storm path | storm flooding oklahoma | story weapon hurricane shore | damage hurricane laura | fema relief funds victim hurricane harvey | tornado tennessee | preparation home amp flooding amp storm

Figure: Trump Narratives on Twitter



Source: Ash, E., Gauthier, G., & Widmer, P. (2023). Text semantics capture political and economic narratives. *Political Analysis*

Plan

Bag-of-words

Static Embeddings

Sequence Embeddings

Issues and Challenges

Incorporating Document Metadata

- Most documents have associated *metadata*, but this metadata is rarely explicitly used in modern language models.
- Social scientists often use the output of text as data methods in *downstream* regressions.
- Two obvious consequences:
  - In general, inference is *biased*.
  - Confidence intervals in the second stage do not reflect the *uncertainty* of the first stage.
- Some papers tackle these problems, but more work is needed:
  - The structural topic model (Roberts et al., *JASA*, 2014)
  - Party embeddings (Rheault and Cochrane, *PA*, 2019)
  - Embeddings regression (Rodriguez et al., *APSR*, 2023)

Validating Algorithmic Output

- Text as data is often used to *measure* outcomes of interest.
- Some examples:
  - Economic policy uncertainty, geopolitical uncertainty, racial and gender bias, micro-narratives, political frames, etc.
- How should we validate these measures?
  - No clear consensus.
  - Data-driven metrics do not correlate well with human judgement.
  - Humans are not necessarily the best judges (they are biased, too, for instance).

Interpreting Algorithmic Output

- Complex methods perform better at most natural language processing tasks, but this generally comes at the cost of *interpretability*.
- Many text features are actually *confounders* for the purpose of economic analysis.
- Some examples:
  - "Texas" might be an accurate and influential predictor of right-wing ideology but is not a term structurally related to a belief system.
  - Attention-based models capture many small details of language that can give away the labels, but that are potentially of no interest to economists (e.g., writing style, use of specific characters, etc.).
- Two current solutions:
  - Use straight-forward methods (e.g., dictionaries, text regressions).
  - Model explanation methods to provide interpretable diagnostics on the features that algorithms rely on (still a burgeoning literature).

Thanks for listening!

For code implementations of the models discussed:

GitHub Repository

Source: Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*.

Text as Data for the Social Sciences

Germain Gauthier

ETH Zürich
Ecole Polytechnique, CREST

IOEA – May 16, 2023