THE BIRDS EYE:

LANDSCAPING LLMs AND PLANNING YOUR RESEARCH PROJECT

Hauke Licht and Lisa Wierer

OUR MISSION

CREATING A WEEK IN WHICH WE **EXPERIENCE** TEXT ANALYSIS

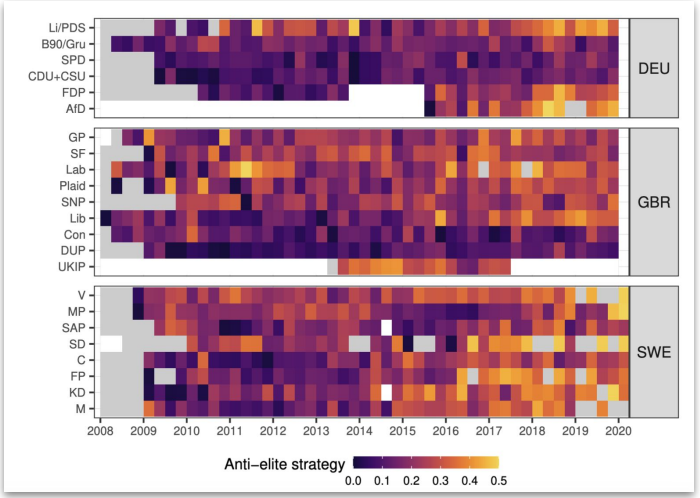
ABOUT US

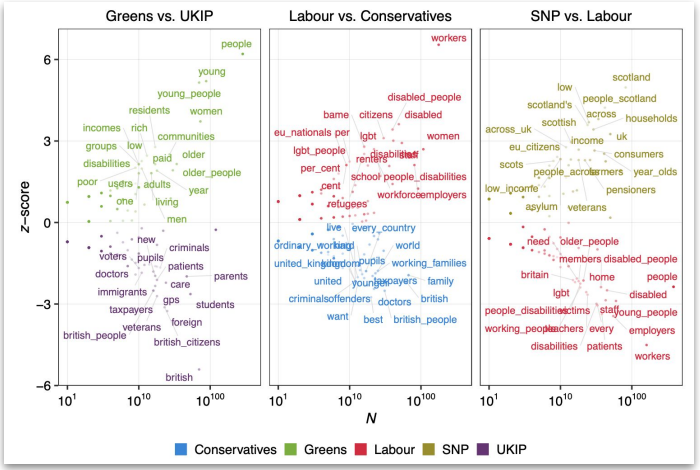
HAUKE

Assistant Professor of
Computational Political Science
at the Department of Political Science and the Digital Science Center (DiSC),
University of Innsbruck

HAUKE

- got interest in NLP when wanting to classify 500K tweets written in ~16 different languages (paper)
- research focus
  - developing text analysis methods to study elites' strategic use of political rhetoric
  - multilingual text analysis
  - NLP and LLMs more generally





LISA

- Assistant Professor for Methods and Methodology in Political Science
- Own business: helps companies communicate data that triggers emotions (business purpose communication)

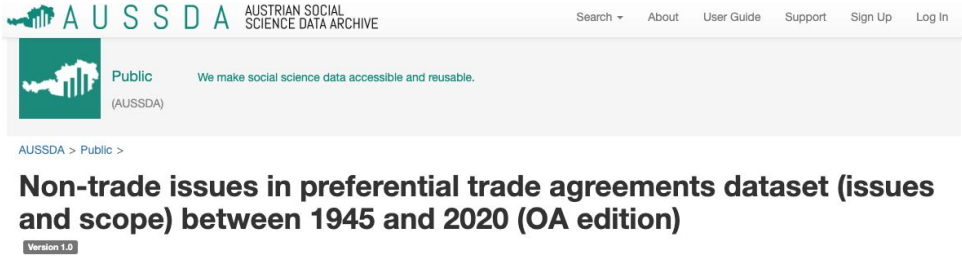
LISA

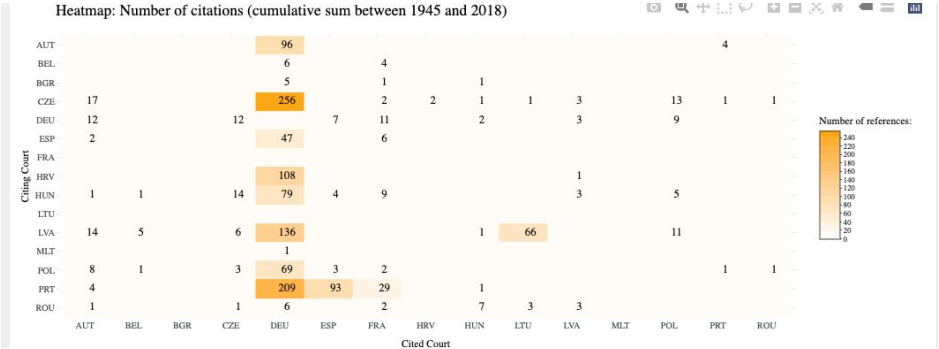
REASONS FOR LEARNING TEXT ANALYSIS:

- frustrated after having double-coded 812 preferential trade agreements
- no money as Phd student to validate coding
- then…starting to having fun with this

RESEARCH FOCUS:

- Text network analysis
- multilingual text analysis
- citation networks
- legal documents (from PTAs over BTT to constitutional court decisions)





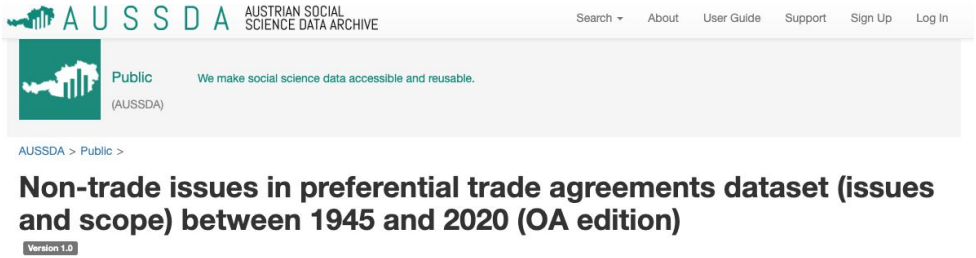
LISA

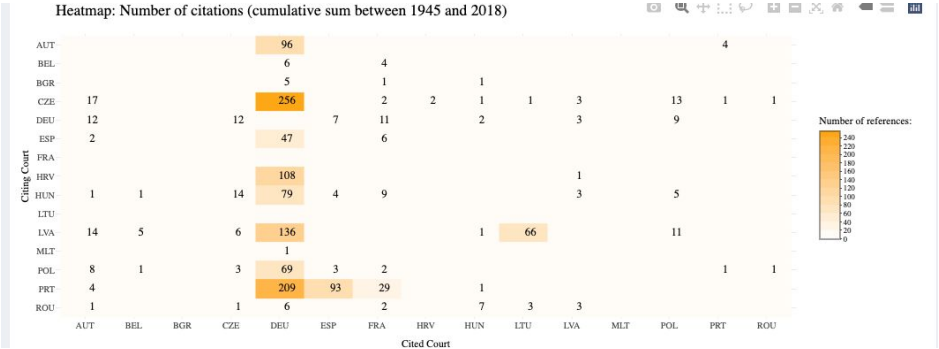
REASONS FOR LEARNING TEXT ANALYSIS:

- frustrated after having double-coded 812 preferential trade agreements
- no money as Phd student to validate coding
- then…starting to having fun with this

RESEARCH FOCUS:

- Text network analysis
- multilingual text analysis
- citation networks
- legal documents (from PTAs over BTT to constitutional court decisions)





and she programs everything in R

Overview of our schedule for this week

Day-by-day schedule

Today

how did we get here? Brief history

developing/elaborating/clarifying your research design

checking set-up

Day-by-day schedule

Tuesday

morning

- static word embeddings
- hands-on programming exercises
  - word vector similarity

afternoon

- intro to contextualized word embedding, attention, and transformer models
- hands-on programming exercises
  - understanding attention
  - sentence and document embedding

Day-by-day schedule

Wednesday

morning

- understanding transformer finetuning
- hands-on programming exercises

afternoon

- evaluating classification performance
- open exercises
    - multilingual classification
    - other classification tasks
    - (maybe) BERTopic

Day-by-day schedule

Thursday

morning

- intro to LLMs
- hands-on exercises with OpenAI’s GPT models
- zero-shot prompting for text classification + exercises

afternoon

- few-shot prompting + exercises
- advanced topics and use cases

Day-by-day schedule

Friday

multilingual aspects

new thinking triggered by LLMs

WE ALL KNOW,
BUT LET’s SAY IT ONCE AGAIN

TEXT DATA = …

DEEPER CONTENT

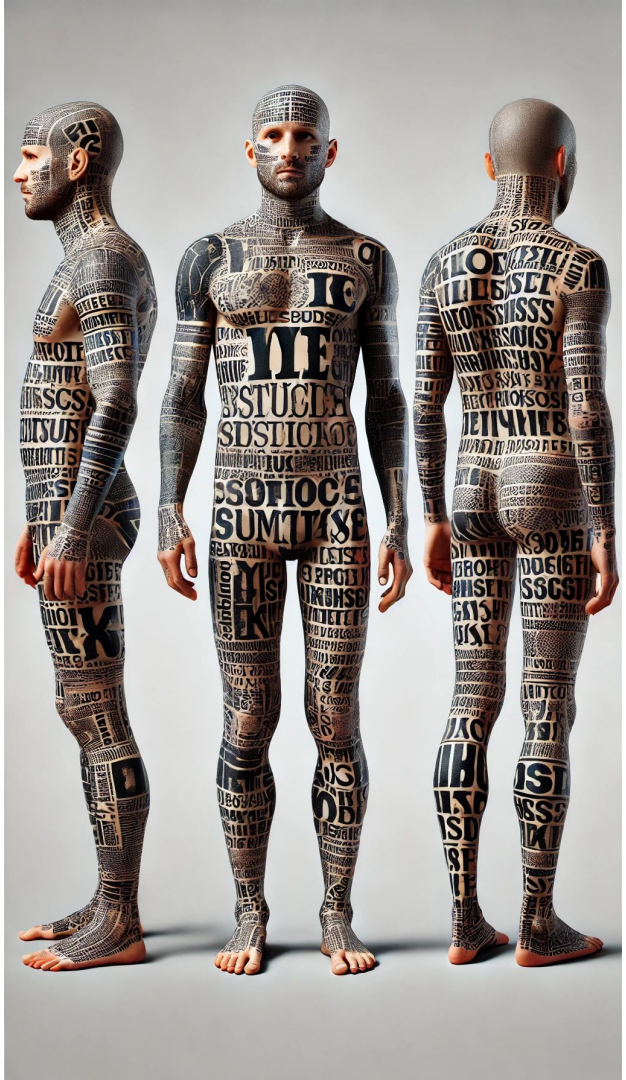
EMOTIONS

OPINIONS

CONTEXT

WIDER DISTRIBUTION

- Social Media
- Blogs
- Websites
- Treaties
- Customer feedback
- …



FASTER

- Real time
- Constant

WHAT CAN YOU DO BETTER WITH YOUR PROJECT THAN OTHERS DO, just by using text data?

WHY ARE WE HERE

- BECAUSE, we are convinced that the meaning of words matters. (And now you might think, this is the entire reason why we use language. But some time ago, we ignored this in QTA)
- BECAUSE, we are no longer happy with ignoring the context.
- BECAUSE, we want to understand people and not numbers.
- BECAUSE, we want to learn and create from the massive existing data.

- ….. and maybe, because we are all a bit of nerds ;)

SO HOW DID WE GET HERE?

BRIEF HISTORY

1948 N-GRAM

This week is going to be fantastic.

This week is going to be fantastic.

Thisweek is going to be fantastic.

This week is going to be fantastic

This week week is is going going to to be be fantastic

This week is week is going is going to going to be to be fantastic

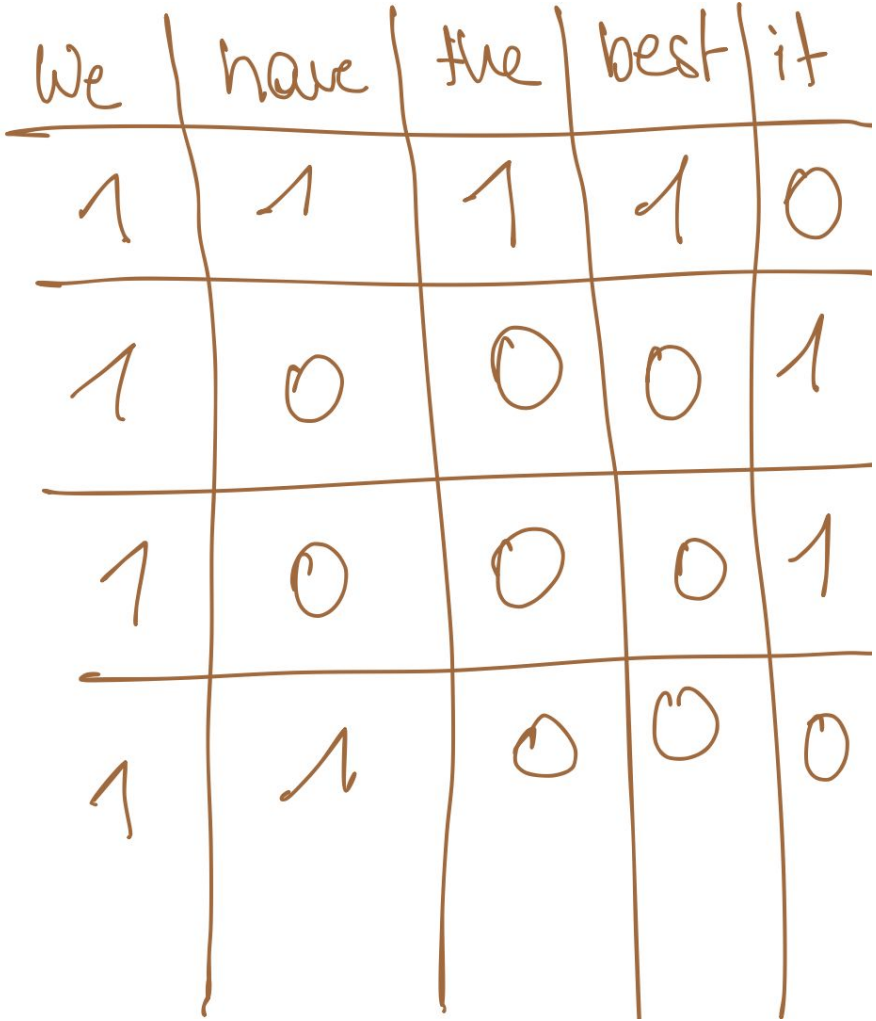
1954 BAG OF WORDS

We have the best time.

We understand it all.

We rock it.

We have fun.



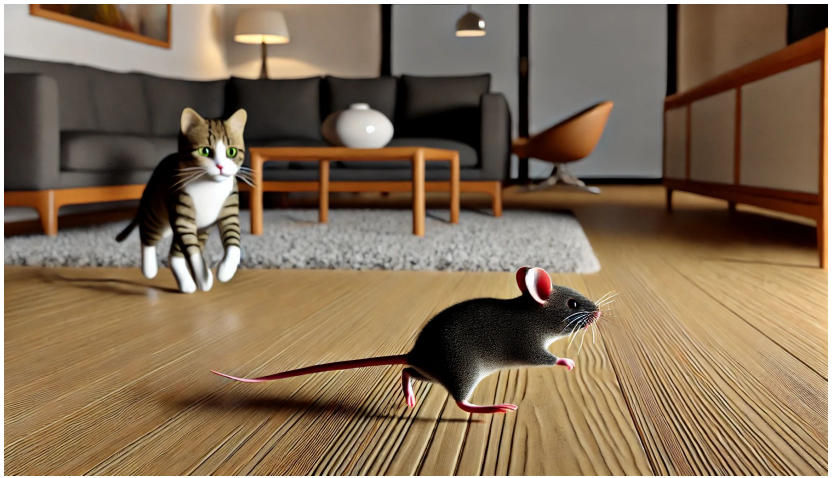
Local context and order → No context and no order

Inflexible → Highly flexible

Computationally inefficient → Computationally efficient

BoW CHALLENGE Nr 1: WORD MEANING

WORD ORDER



BoW CHALLENGE Nr 1: WORD MEANING

CONTEXTUAL MEANING

giving a heartfelt toast

morning toast

BoW CHALLENGE Nr 2: SPEED

BoW CHALLENGE Nr 2: SPEED

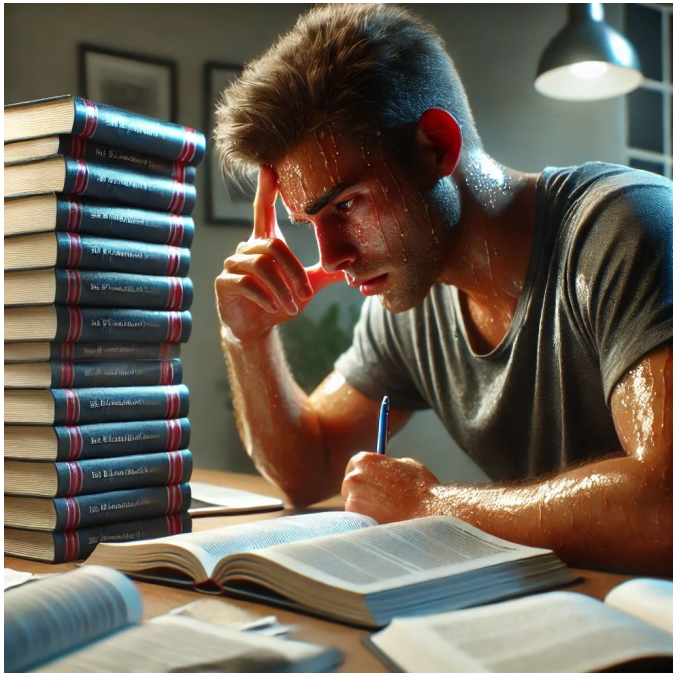
SPARSITY

high-dimensionality

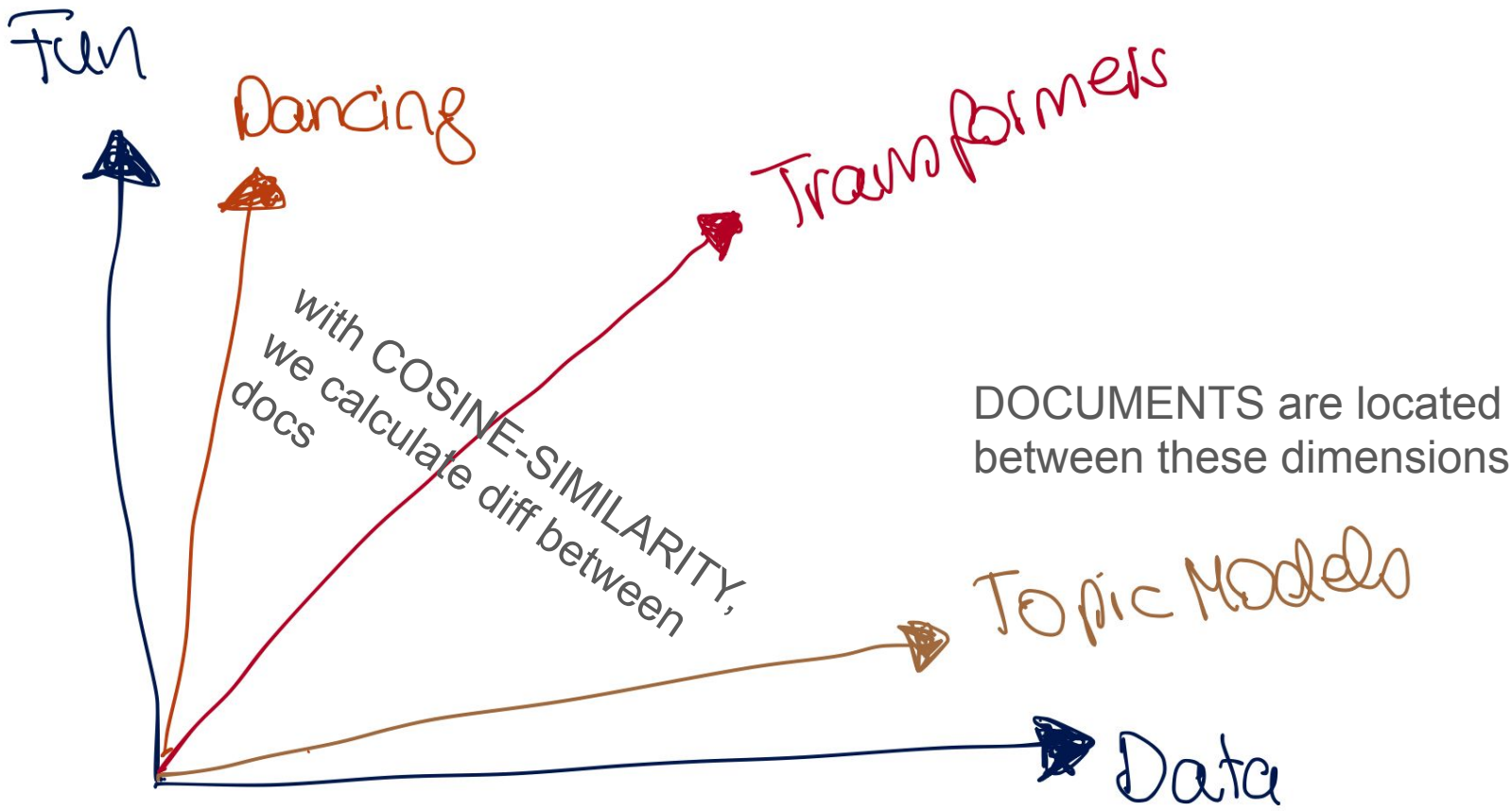
	fascinating	is	language
1	0	0	0
2	1	1	0
3	0	1	1

BoW CHALLENGE Nr 2: SPEED

BoW: start from scratch each time



1986 VECTOR MODEL



DOCUMENTS are located between these dimensions

WORDS are DIMENSIONS

BAG OF WORDS + term weight = VECTOR MODEL

- binary value
- term-frequency value
- TF-IDF value

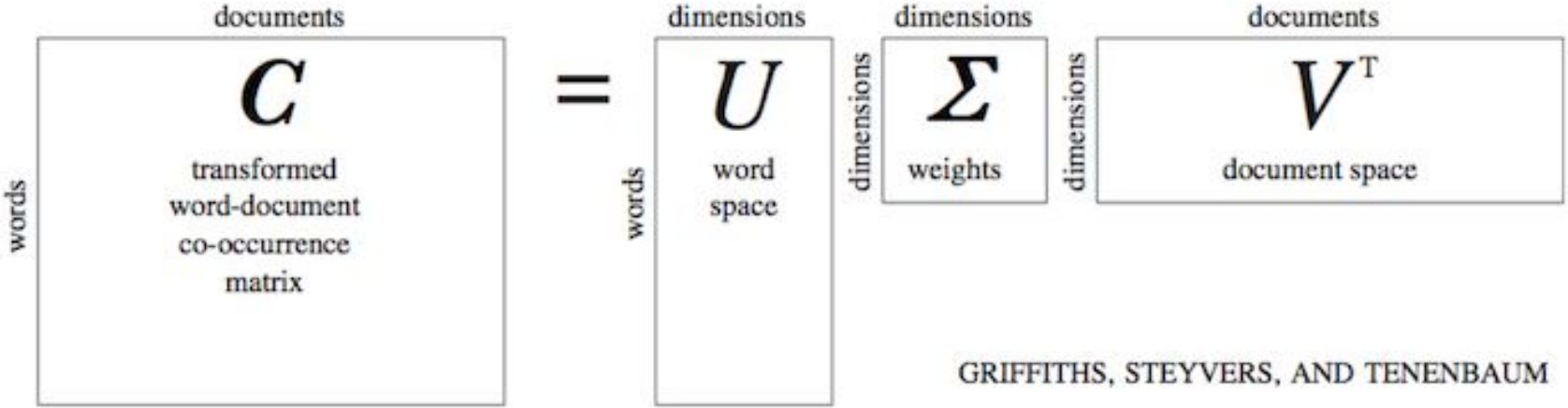
In essence, vector models are upgraded BoW models.

1988 LATENT SEMANTIC ANALYSIS

Deerwester et. al (1988)

SVD (Singular Vector Decomposition) - LSA: IDEA

- Established method.
- Used for signal processing, image compression, machine learning.
- Early application synonym extraction (Deerwester et al. 1990, 2003)
- Decomposes matrix in three parts.



GRIFFITHS, STEYVERS, AND TENENBAUM

2003 NEURAL LANGUAGE MODELS

Bengio et. al (2003)

NEURAL LANGUAGE MODELS

GROUNDWORK FOR WHAT IS NEXT….

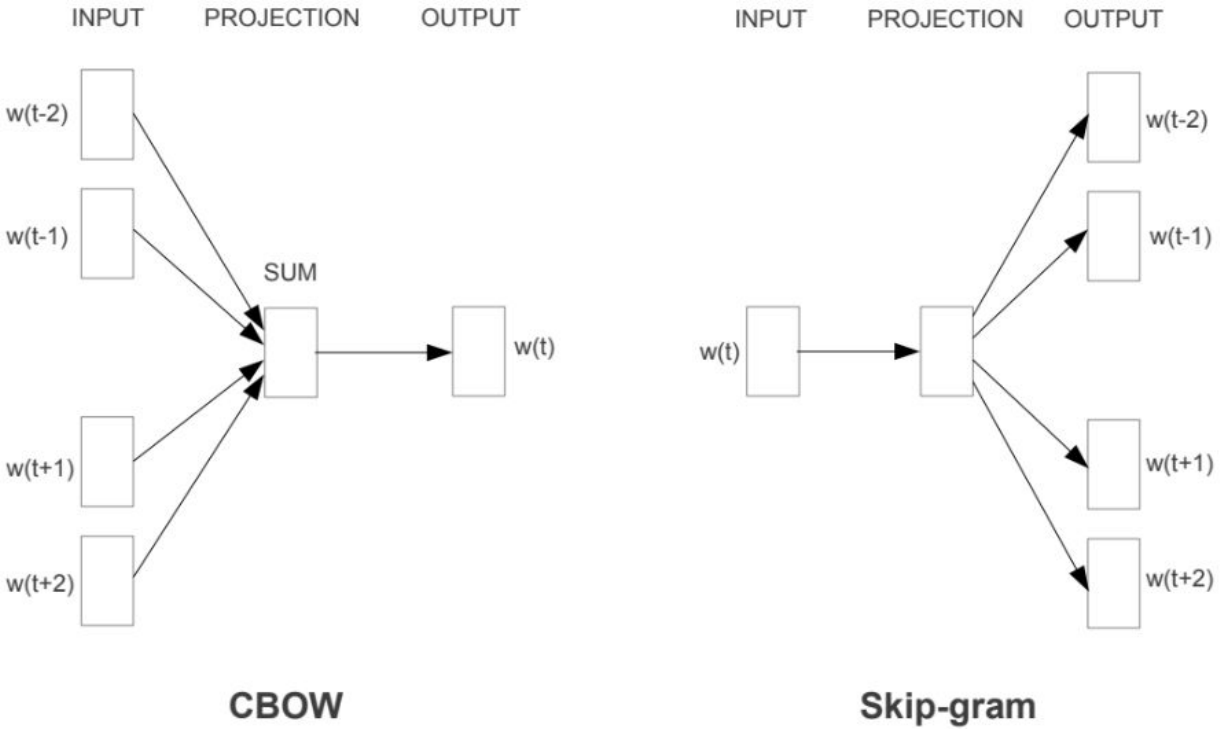
HELLO,

THIS IS A WORKSHOP ON LLMs, IN WHICH WE ____ TO _____.

WHAT IS ____?

2013 WORD2VEC

Mikolov (2013)



Continuous bag of words

Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

SKIP-Gram (predict context with centre word)

1. **Prediction Task**: The model tries to guess which words usually appear near a given word. For example, if the word is "king," it might predict nearby words like "queen" or "royal."
2. **Raw Scores**: The model first gives each word a number (score) based on how likely it thinks that word should be near the target word.
3. **Softmax Function**: These scores are turned into probabilities, which are easier to work with. The probabilities for all possible words add up to 1.
4. **Training Goal**: The model keeps adjusting itself to get better at predicting the right words in the right context. It tries to improve its guesses over time.

2014 Glove

Pennington et. al (2014)

FIRST: OBSERVE CO-OCCURENCE OF WORDS

The model learns by looking at how often words appear near each other in a sentence. For example, the word "dog" might frequently appear near words like "bark" or "pet."

OBSERVE

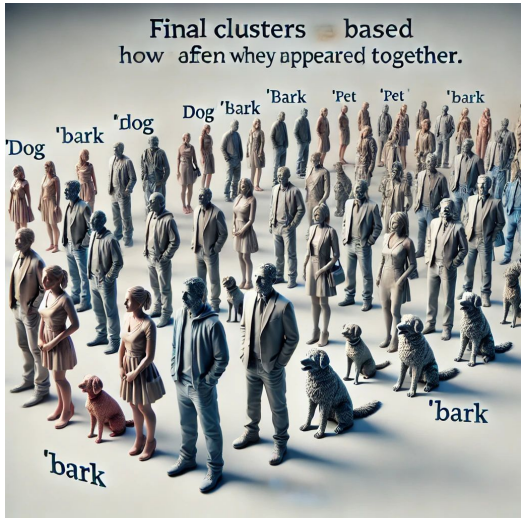
SECOND: PREDICTING NEIGHBORS

The model focuses on predicting a word's neighbors (context) based on the word itself. So, it tries to learn what words are likely to show up next to the word "dog" by seeing many examples of sentences with that word.

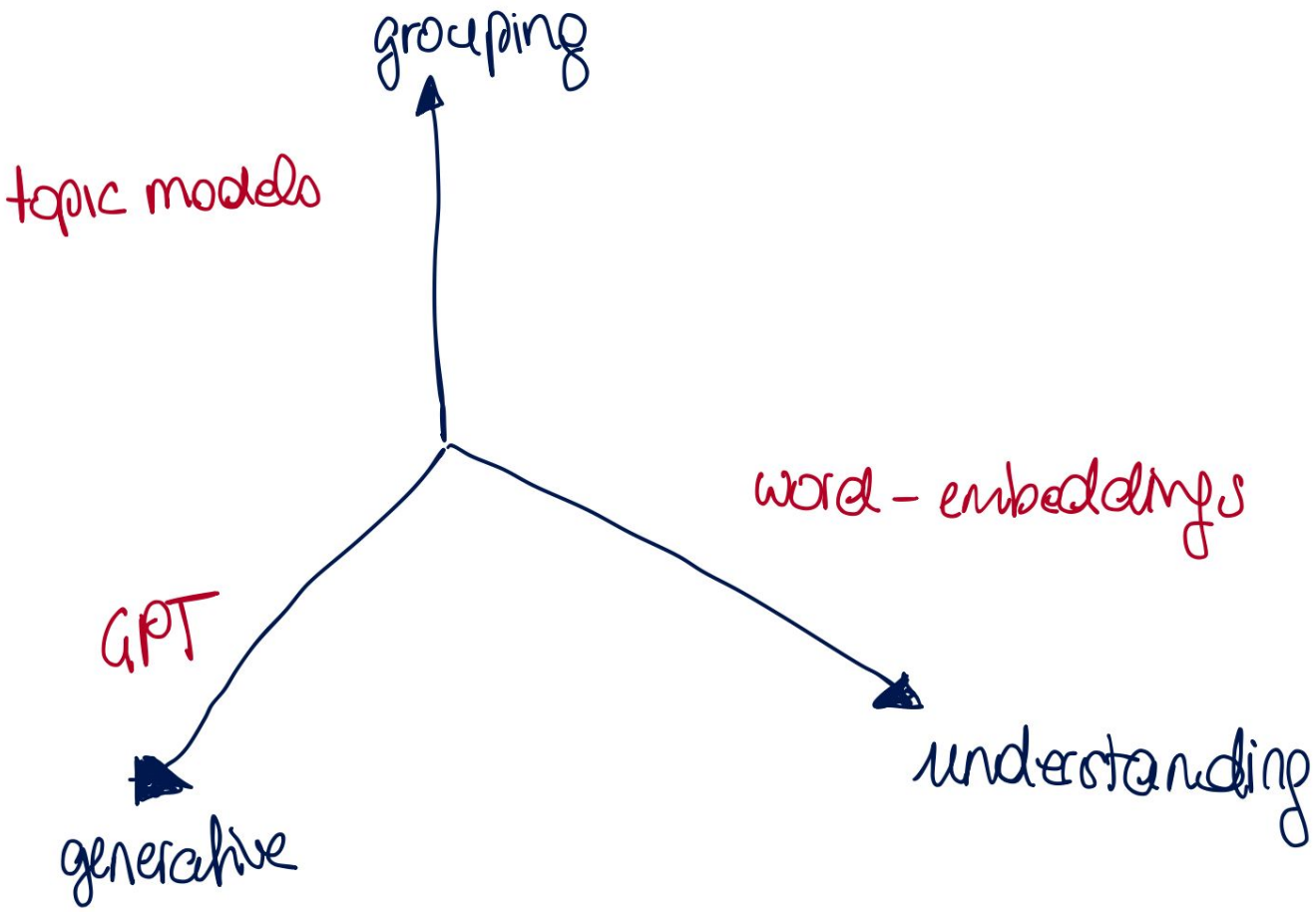
PREDICT

SECOND: OPTIMIZATION SO THAT SIMILAR WORDS HAVE SIMILAR VECTORS

Word2Vec defines a goal (cost function) that helps the model learn. It tries to adjust the word representations (called vectors) so that similar words end up with similar vectors. The cost function tells the model how wrong its predictions are, and the model adjusts its learning to make better predictions.



OPTIMIZE


grouping
topic models
word - embeddings
GPT
understanding
generative

VECTOR SPACE MODELS

Focus on Documents

WORD EMBEDDINGS

Focus on Words

DIFFERENCE BOW AND EMBEDDINGS

MEANING ADJUSTS GIVEN CONTEXT AND ORDER

BoW

=

toast = toast

Word Embeddings

≠

toast ≠ toast

CONDITION - TRAINED ON DIFFERENT CORPORA

REFLECTION TIME

- HOW RELEVANT IS ORDER with respect to YOUR research question?
- HOW RELEVANT IS CONTEXT with respect to YOUR research question?

Task 1: Highlight 2-4 Keywords in the Document

Read the sample document and highlight keywords or phrases that are central to its meaning. Pay special attention to words that might have different meanings depending on context (e.g., "bank").

Task 2: Identify Important Phrases or Sentences

Underline or mark phrases where word order is important for understanding the meaning. For example, does the order of words or phrases change the meaning if rearranged?

Task 3: Analyse Word Meaning in Context

Could it be that the word means something different in different contexts (feel free to ask Chat-GPT)

Task 4: Analyse Word order

How critical is the word order with respect to these highlight words?

BoW CHALLENGE Nr 2: SPEED

SPARSITY

high-dimensionality

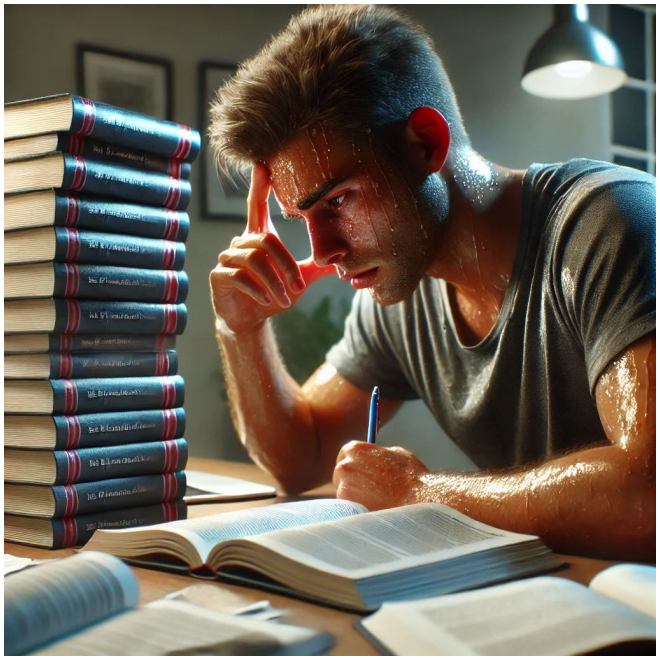
	fascinating	is	language
1	0	0	0
2	1	1	0
3	0	1	1

low-dimensionality

	Dimension 1	Dimension 2
fascinating	0.4	0.9
is	0.1	0.1
language	0.5	0.7

BoW CHALLENGE Nr 2: SPEED

BoW: start from scratch each time



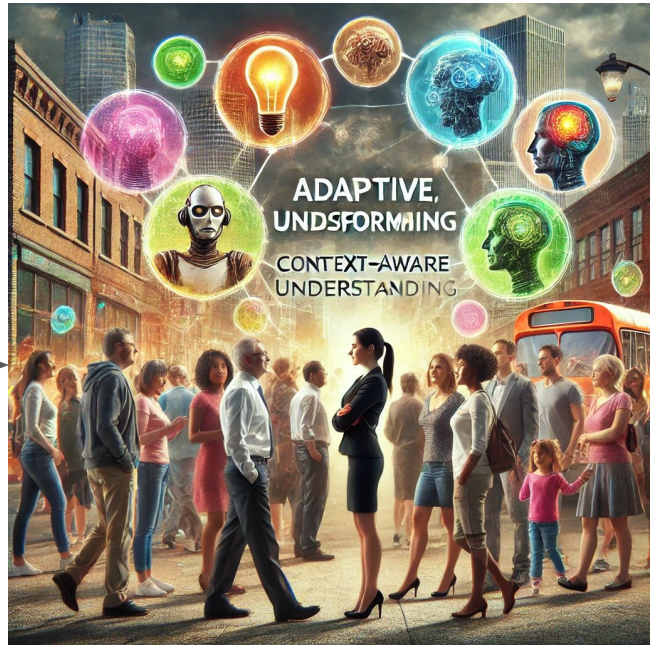
Word-Embeddings: OFTEN Build on pretrained models

2018 TRANSFORMER MODELS (BERT, GPT)

Static Word Embeddings

TRANSFORMERS

Contextualized Embeddings



LET’S MAKE A PLAN

RESEARCH DESIGN

RESEARCH QUESTION

start here

LEVEL/UNIT OF ANALYSIS

WHAT DO I WANT TO MEASURE?

DESCRIPTIVE AND PREDICTIVE ANALYSIS?

SOURCE

DOCUMENTS

Average text length:

Document format (pdf, html, json, ...):

Language(s):

Structure and tidiness level:

Which preprocessing steps are absolutly necessary?

DOCUMENT METADATA

Meta-data I want to collect:

AVAILABILITY

Data gaps:

How to deal with it?

MEASURMENT STRATEGY

HOW DO I CONCEPTUALIZE IT?

LEVEL OF MEASUREMENT

- Full document, paragraph, sentence, (span of) word(s)
- individual text
- pairs of text

WHAT'S THE MEASUREMENT TASK?

- **classification**: texts => categories
- **scoring**: texts => scale/spectrum
- **comparison**: [text, text] => score

INDUCTIVE OR DEDUCTIVE?

classification: you know the categories to which texts can belong? ○yes ○no

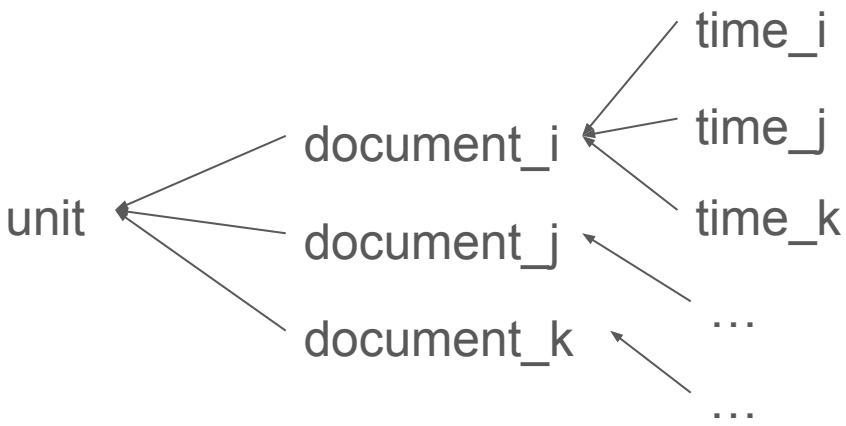
scoring/comparison: you know the underlying dimension? ○yes ○no

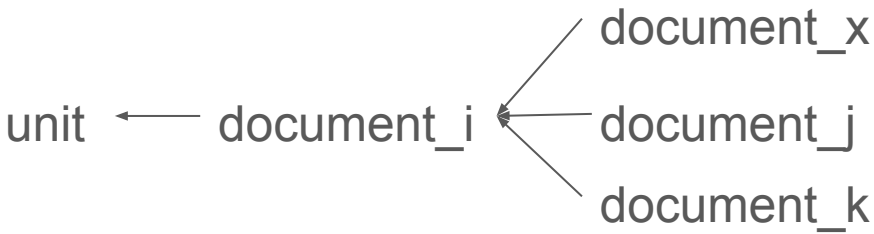
EVALUATION | RELIABILITY | VALIDATION

RESEARCH DESIGN

RESEARCH DESIGN

UNIT OF ANALYSIS





RESEARCH DESIGN

WHAT DO YOU WANT TO ASSESS?

WHAT IS THE LATENT CONCEPT?

EXAMPLES: Political Trust, social capital, public opinion, policy efficacy, authoritarianism, social cohesion, identity, political ideology,...

RESEARCH DESIGN

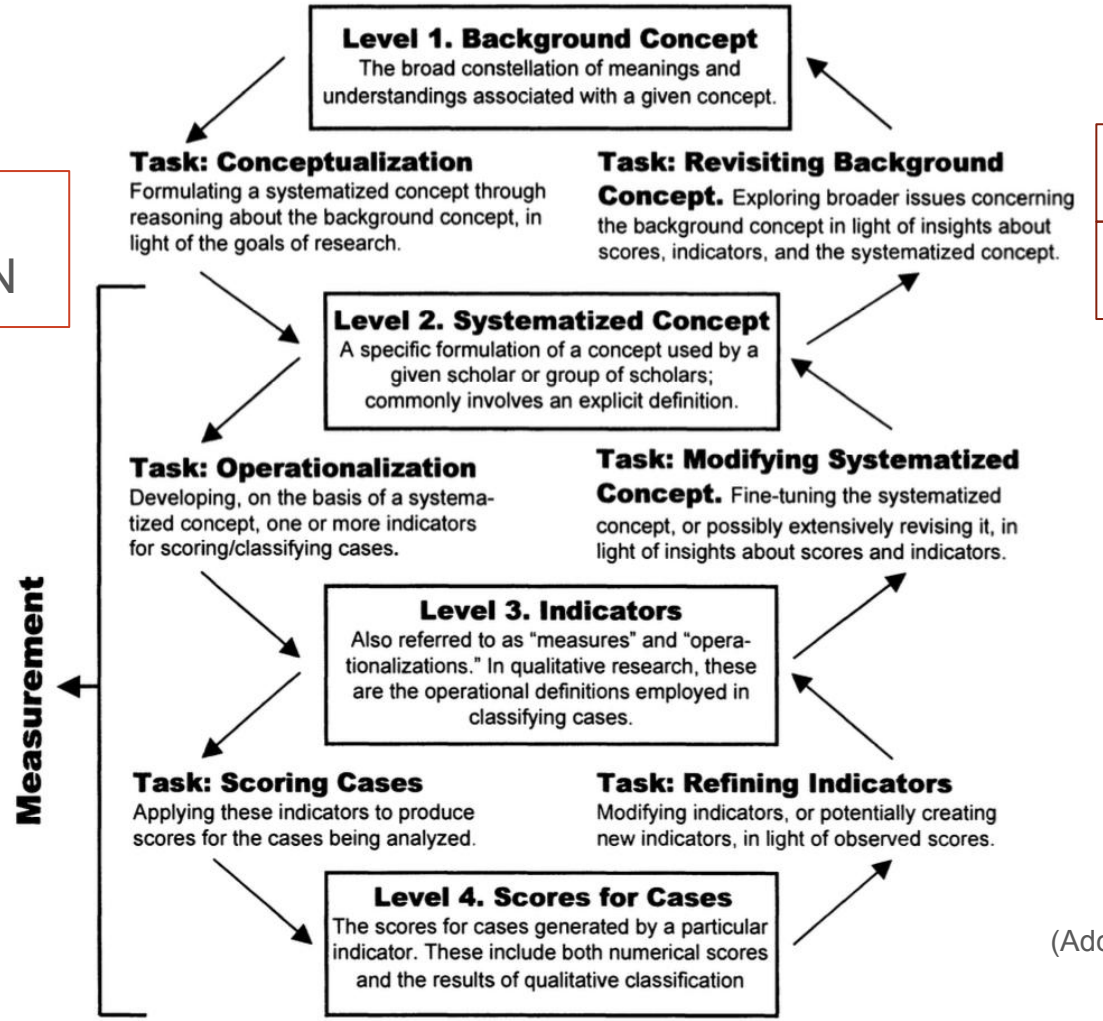
HOW DO YOU TRANSLATE YOUR LATENT CONCEPT INTO A MEASUREMENT?

SPECIFIC FORMULATION

LITERATURE

YOUR BRAIN

FIGURE 1. Conceptualization and Measurement: Levels and Tasks



(Adcock and Collier, 2001)

RESEARCH DESIGN

DESCRIPTIVE ANALYSIS

https://yourstatstory.com/

PREDICTIVE ANALYSIS

PA

Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora

Ludovic Rheault[1] **and Christopher Cochrane**[2]

[1] *Assistant Professor, Department of Political Science and Munk School of Global Affairs and Public Policy, University of Toronto, Canada. Email: ludovic.rheault@utoronto.ca*
[2] *Associate Professor, Department of Political Science, University of Toronto, Canada. Email: christopher.cochrane@utoronto.ca*

Abstract

Word embeddings, the coefficients from neural network models predicting the use of words in context, have now become inescapable in applications involving natural language processing. Despite a few studies in political science, the potential of this methodology for the analysis of political texts has yet to be fully uncovered. This paper introduces models of word embeddings augmented with political metadata and trained on large-scale parliamentary corpora from Britain, Canada, and the United States. We fit these models with indicator variables of the party affiliation of members of parliament, which we refer to as party embeddings. We illustrate how these embeddings can be used to produce scaling estimates of ideological placement and other quantities of interest for political research. To validate the methodology, we assess our results against indicators from the Comparative Manifestos Project, surveys of experts, and measures based on roll-call votes. Our findings suggest that party embeddings are successful at capturing latent concepts such as ideology, and the approach provides researchers with an integrated framework for studying political language.

Keywords: word embeddings, parliamentary corpora, text as data, political ideology, natural language processing

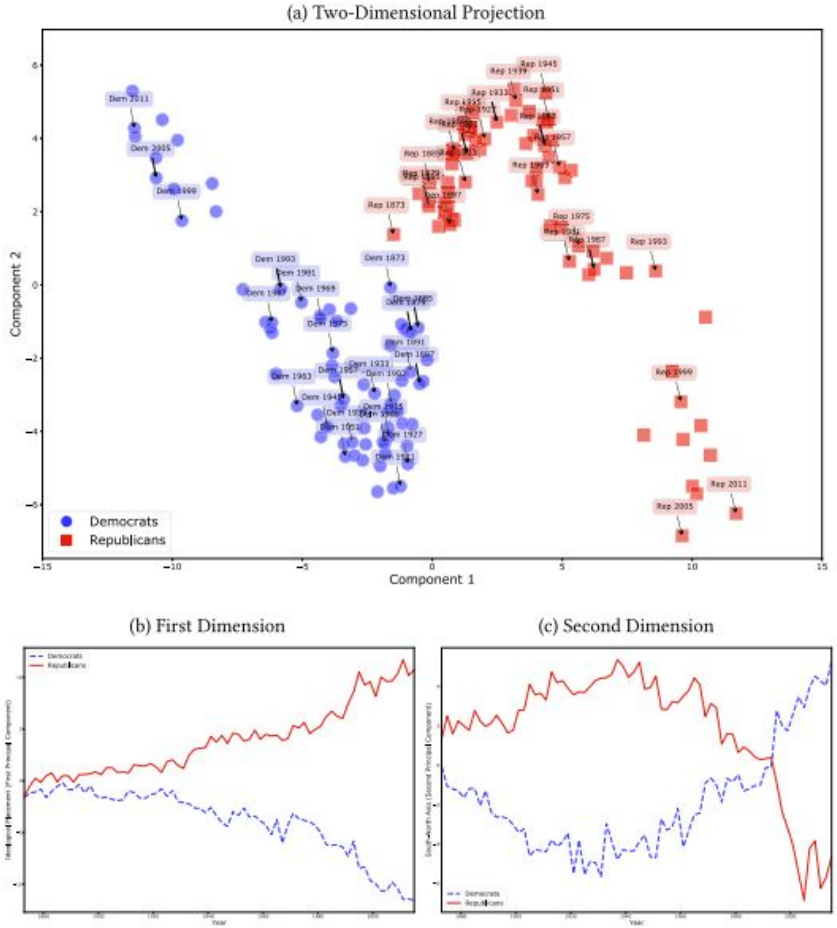


Figure 2. Party placement in the US House (1873–2016). The figure shows a two-dimensional projection of the two principal components of party embeddings for the US House of Representatives (a) and time-series plots for each of the two components separately in (b) and (c).

American Political Science Review (2021) 1–15
doi:10.1017/S0003055421000356

Playing to the Gallery: Emotive Rhetoric in Parliaments

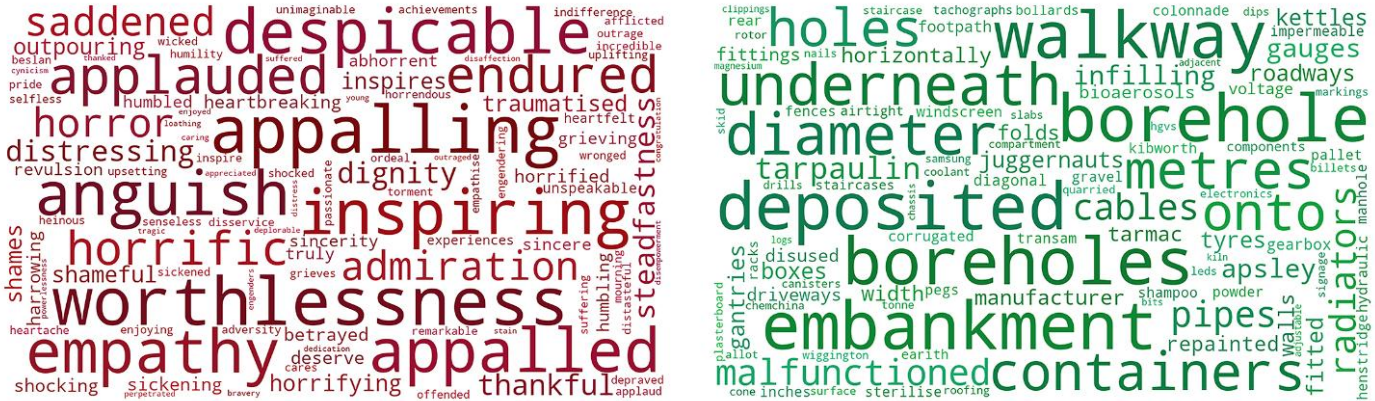
MORITZ OSNABRÜGGE *Durham University*
SARA B. HOBOLT *London School of Economics and Political Science*
TONI RODON *Universitat Pompeu Fabra*

Research has shown that emotions matter in politics, but we know less about when and why politicians use emotive rhetoric in the legislative arena. This article argues that emotive rhetoric is one of the tools politicians can use strategically to appeal to voters. Consequently, we expect that legislators are more likely to use emotive rhetoric in debates that have a large general audience. Our analysis covers two million parliamentary speeches held in the UK House of Commons and the Irish Parliament. We use a dictionary-based method to measure emotive rhetoric, combining the Affective Norms for English Words dictionary with word-embedding techniques to create a domain-specific dictionary. We show that emotive rhetoric is more pronounced in high-profile legislative debates, such as Prime Minister's Questions. These findings contribute to the study of legislative speech and political representation by suggesting that emotive rhetoric is used by legislators to appeal directly to voters.

INTRODUCTION

When and why do legislators use emotive rhetoric? In today's political landscape many prominent politicians, such as Donald Trump, rely heavily on emotional appeals, both positive ("Make America Great Again") and fear-inducing ("They're bringing drugs. They're bringing crime. They're rapists") (Nai and Maier 2018). However, we not only policy substance but also more emotive language. This argument builds on two strands of literature. First, recent work on political speech has studied how politicians adapt the style and the tone of their speeches to appeal to the electorate. For example, Spirling (2016) argues that cabinet members began using more simple language after the extension of the franchise in the Second Reform Act in Britain in 1867, because

FIGURE 1. Word Clouds of Emotive and Neutral Words

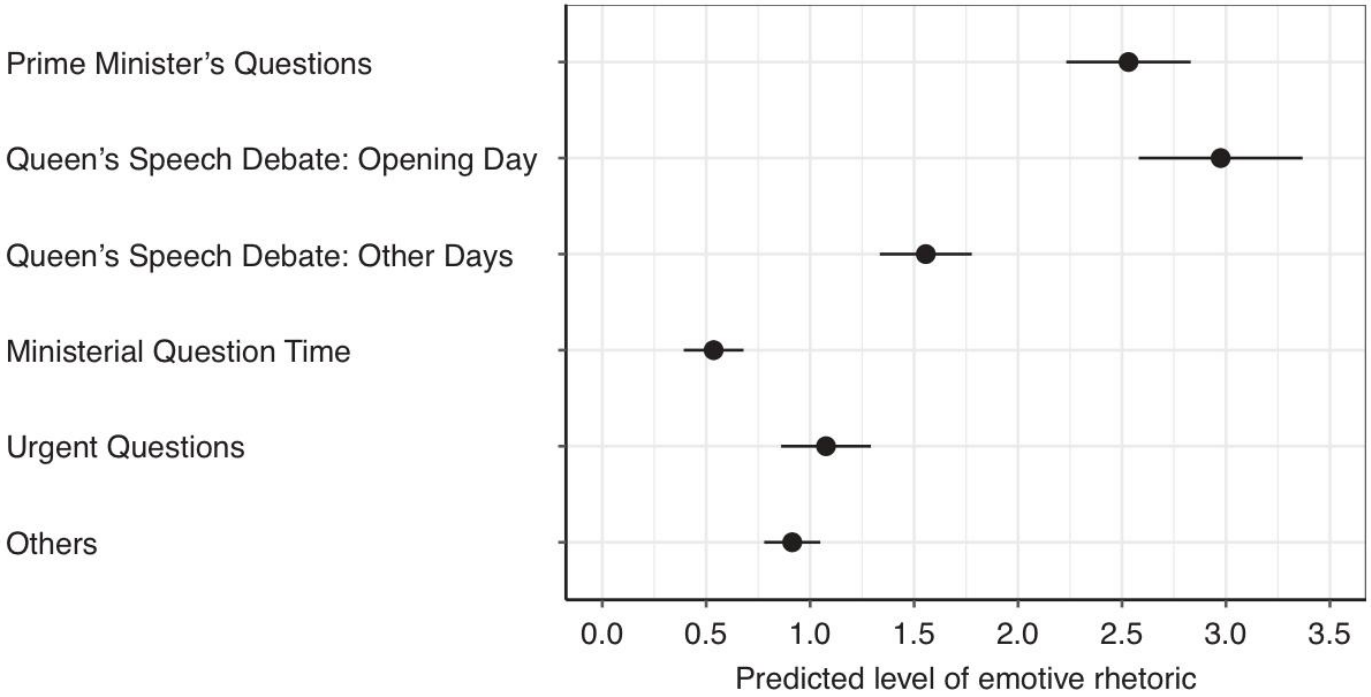


(a) Emotive (b) Neutral

TABLE 2. Examples: Emotive and Neutral Speeches

Score	Text	Speaker
43	Evil happens when good people stand by and do nothing. There is evil running through and infiltrating the Labour party, but it is full of good people and they are trying to do something about it. I commend them, appreciate them and have nothing but respect for them.	Alec Shelbrooke, MP
–25	When used with old-fashioned copper wires, 10 megabits can become a lot less than that. We need a superfast fibre infrastructure instead of copper wires.	Geoffrey Clifton-Brown, MP

FIGURE 3. Predicted Level of Emotive Rhetoric by Type of Debate and 95% Confidence Intervals



Note: The predictions are predictive margins computed based on Model 4.

SOURCE

DOCUMENT LENGTH

1 von 631

631 PAGES

COMPREHENSIVE AND PROGRESSIVE AGREEMENT

FOR

TRANS-PACIFIC PARTNERSHIP

PREAMBLE

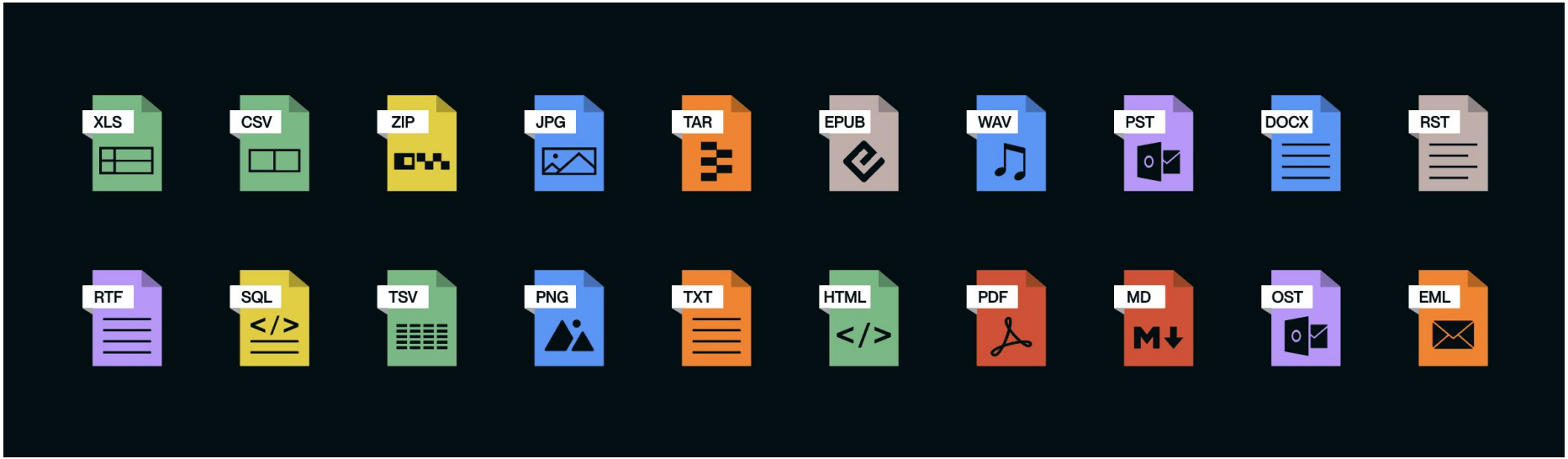
The Parties to this Agreement, resolving to :

REAFFIRM the matters embodied in the preamble to the Trans-Pacific Partnership Agreement, done at Auckland on 4 February 2016 (hereinafter referred to as "the TPP");

REALISE expeditiously the benefits of the TPP through this Agreement and their strategic and economic significance;

280 CHARACTERS

DOCUMENT FORMAT



DOCUMENT FORMAT

Format	Common Use Cases	Python Packages
PDF	Documents, reports	PyPDF2, pdfminer.six, fitz
HTML	Websites, blogs	BeautifulSoup, lxml, html5lib
JSON	APIs, structured data	json, simplejson
TXT	Plain text, logs	io, open()
CSV	Tabular data, databases	csv, pandas
XML	Data interchange, configuration	xml.etree.ElementTree, lxml
DOCX	Word documents	python-docx
EPUB	E-books	ebooklib, beautifulsoup
YAML	Configuration files	PyYAML
Markdown	Documentation, README files	markdown
Parquet	Analytical datasets, data storage	pyarrow, pandas

PREPROCESSING

PREPROCESSING

“The AIM of PREPROCESSING is to make the inputs to a given analysis less complex in a way that does not adversely affect the interpretability or substantive conclusions of the subsequent model.” (Denny and Spirling 2018)

PREPROCESSING

In general preprocessing should be done based on theory and not simply copy-paste what others have done in the past!

PREPROCESSING

- BASIC: removing HTML tags, removing numbers, and lowercasing.
- PUNCTUATION REMOVAL: “@%*()& .:,; +“
- SPELLCHECK
- NEGATION: from “not sad” to “happy“
- PARTS-OF-SPEECH: Keep only NN, VB, JJ and RB
- STOPWORDS removal
- STEMMING
- LEMMATIZATION

PREPROCESSING ROLE CHANGES

Bow

Figure 1. Wordfish results for the 128 different preprocessing possibilities. Each row of the plot represents a different specification. A white bar implies that the manifesto for that year is in the correct place as regards our priors. A black bar implies it was misplaced.

"It has been proven that the time spent on preprocessing take from
50% to 80% of the entire classification process." (Morik and Scholz 2004)

Word Embeddings

Transformers

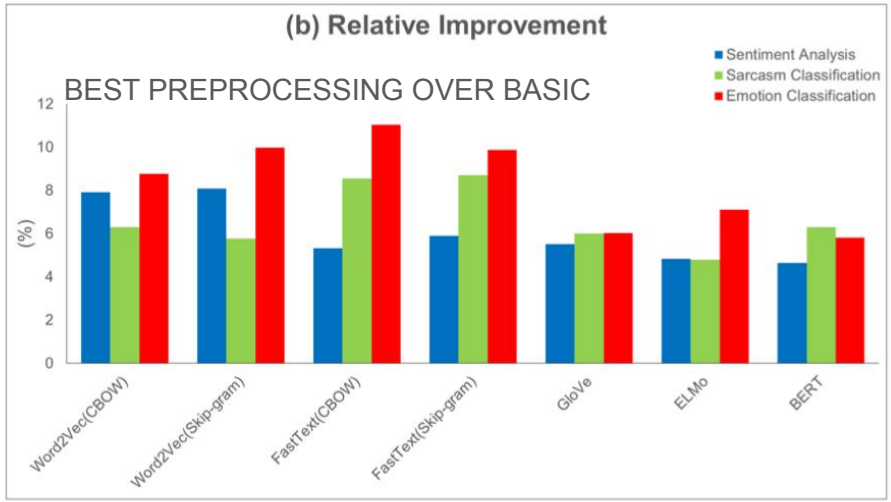


Fig. 2. Average F-scores vs. relative improvement

Babanejad et. Al 2015

PREPROCESSING ROLE CHANGES

Bow

GOAL:
- create a clean DFM with which we can calculate

STRATEGY:
- Common words and rare words are removed.
- Words “standardized”

Word Embeddings

GOAL:
- capture word meaning in a lower dimensionality vector

STRATEGY:
- Keep words that are context relevant (often punctuation are removed, sometimes stop words are dropped, etc.)

Transformers

GOAL:
- capture contextual meaning at word, sentence and text level

STRATEGY:
- Keep words/subwords and sentence with original structure.Do not loose any meaningful indicator.

OPEN-AIs Preprocessing pipeline

HIGH QUALITY TEXT COLLECTION

- remove duplicates
- remove very short texts
- remove overly structured text (HTML tags)

SUBWORD STRUCTURE (BYTE PAIR ENCODING)

- to handle rare words, typos, out of vocab terms
- efficient computation

MAINTAIN

- case
- whitespace
- punctuation

ADD

- [CLS] sentence and document tags
- [SEP] separate sequences
- [PAD] fill up to achieve certain byte length

ALSO

- unicode normalization
- creation of domains

SYNTHETIC DATA GENERATION

MEMORY AUGMENTATION

- handling of very long text

SHUFFLE

- randomize data (duplicate, shuffle) to minimize bias

REMOVE UNWANTED DATA

- code, HTML tags
- harmful or risky content

WHEN TO DO PREPROCESS

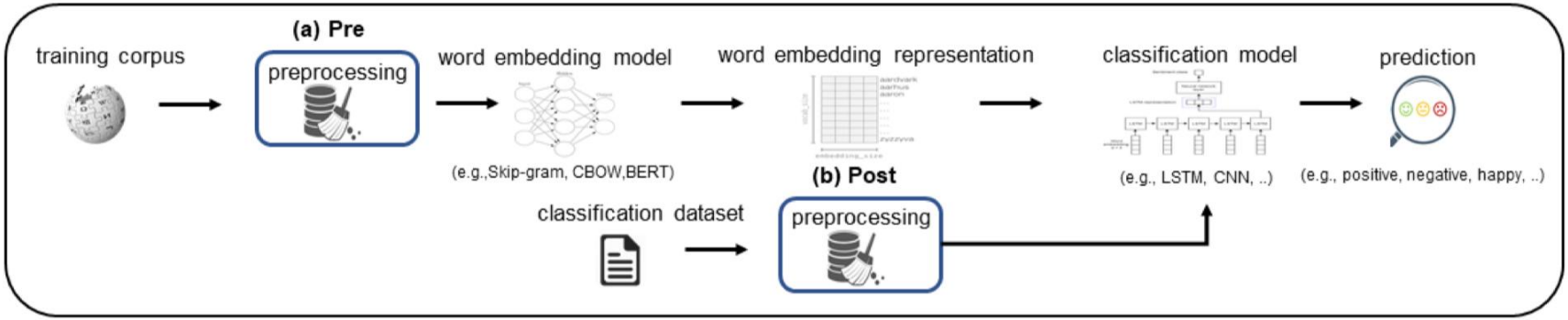


Fig. 1. Framework of applying pre-processing in different stages in the affect prediction system: (a) Pre and/or (b) Post.

Babanejad et. Al 2015

WHEN TO DO PREPROCESS

TABLE 9
Evaluating the effect of pre-processing `(Set B)` at embedding-training corpus (Pre) vs. at classification datasets (Post) and (Both) (F-score)

Embedding-Training Corpus	Processing	IMDB	Semeval	Airline	IAC	Onion	Reddit	Alm	ISEAR	SSEC
Word2Vec (CBOW)	Pre	**90.67**	**62.74**	**74.33**	**73.08**	**76.52**	**69.15**	**53.18**	**66.19**	**60.51**
	Post	87.30	60.04	72.20	68.27	73.61	66.80	48.25	61.29	55.00
	Both	88.13	61.70	72.69	70.08	74.12	68.48	50.23	65.37	58.00
Word2Vec (Skip-gram)	Pre	**89.91**	**62.73**	**73.69**	**72.85**	**76.31**	**69.24**	**52.84**	**64.80**	**59.28**
	Post	88.01	60.22	70.25	71.13	74.28	67.45	50.62	62.00	55.70
	Both	88.57	61.85	73.20	71.08	75.00	69.00	50.74	63.12	57.21
FastText (CBOW)	Pre	**80.71**	**71.90**	**73.70**	**63.17**	**66.24**	**66.71**	**53.00**	**33.25**	**56.49**
	Post	77.30	70.10	71.27	56.80	65.30	63.78	52.67	30.27	54.18
	Both	78.69	71.25	71.69	61.38	65.84	64.37	52.73	32.80	54.80
FastText (Skip-gram)	Pre	**82.93**	**72.00**	**74.15**	**63.57**	**66.80**	**66.79**	**55.38**	**32.29**	**56.63**
	Post	78.20	67.84	70.33	59.67	63.80	61.30	51.27	30.69	54.70
	Both	79.06	70.60	73.12	62.81	65.30	64.80	54.25	30.39	55.00
GloVe	Pre	86.73	**73.41**	**74.00**	**74.23**	**74.27**	**68.40**	**59.80**	**66.85**	**60.11**
	Post	85.12	70.00	71.45	72.64	71.69	65.10	56.48	64.23	57.29
	Both	**87.29**	73.00	73.14	73.45	73.59	67.48	58.29	66.70	58.76
ELMo	Pre	**90.40**	**73.20**	**85.03**	71.19	**75.27**	**69.87**	**65.38**	**69.81**	**71.80**
	Post	86.25	70.33	82.20	69.48	73.02	66.40	63.14	67.40	69.28
	Both	90.33	72.60	83.20	**72.68**	74.11	68.00	64.21	67.62	70.37
BERT	Pre	93.67	**74.00**	**94.88**	**79.00**	**79.84**	**66.00**	**61.18**	**70.28**	**70.33**
	Post	91.83	70.12	92.00	74.04	76.81	62.71	58.02	67.90	66.80
	Both	**94.03**	72.19	92.20	76.39	77.19	63.77	60.03	68.34	67.61

METADATA

Document-Level Metadata

- Title
- Author
- Publication Date
- Source/URL
- Document Type
- Language
- Word Count
- Genre or Category
- Publisher

Technical Metadata

- File Format
- Encoding
- Document Size
- File Path

Temporal Metadata

- Last Modified Date
- Creation Date
- Access Date

User Interaction Metadata (for Online Text)

- Views/Reads
- Shares/Likes
- Comments/Annotations

HANDLING METADATA IN R

HANDLING METADATA IN PYTHON

MISSING DATA

WHAT IS MISSINGNESS IN TEXT DATA?

INCOMPLETE DOCUMENT COLLECTION

TRUNCATED DOCUMENTS

MISSING FEATURES

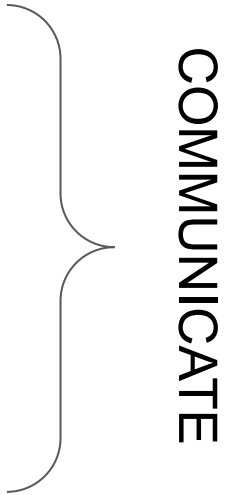
MISSING DATA

3 OPTIONS:

ACCEPT

REVISE AGGREGATION LEVEL

IMPUTE MISSING DATA



MEASUREMENT STRATEGY


OPERATIONALIZATION
FROM CONCEPT
INDICATOR

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

Word embeddings are a powerful machine-learning framework that represents each English word by a vector. The geometric relationship between these vectors captures meaningful semantic relationships between the corresponding words. In this paper, we develop a framework to demonstrate how the temporal dynamics of the embedding helps to quantify changes in stereotypes and attitudes toward women and ethnic minorities in the 20th and 21st centuries in the United States. We integrate word embeddings trained on 100 y of text data with the US Census to show that changes in the embedding track closely with demographic

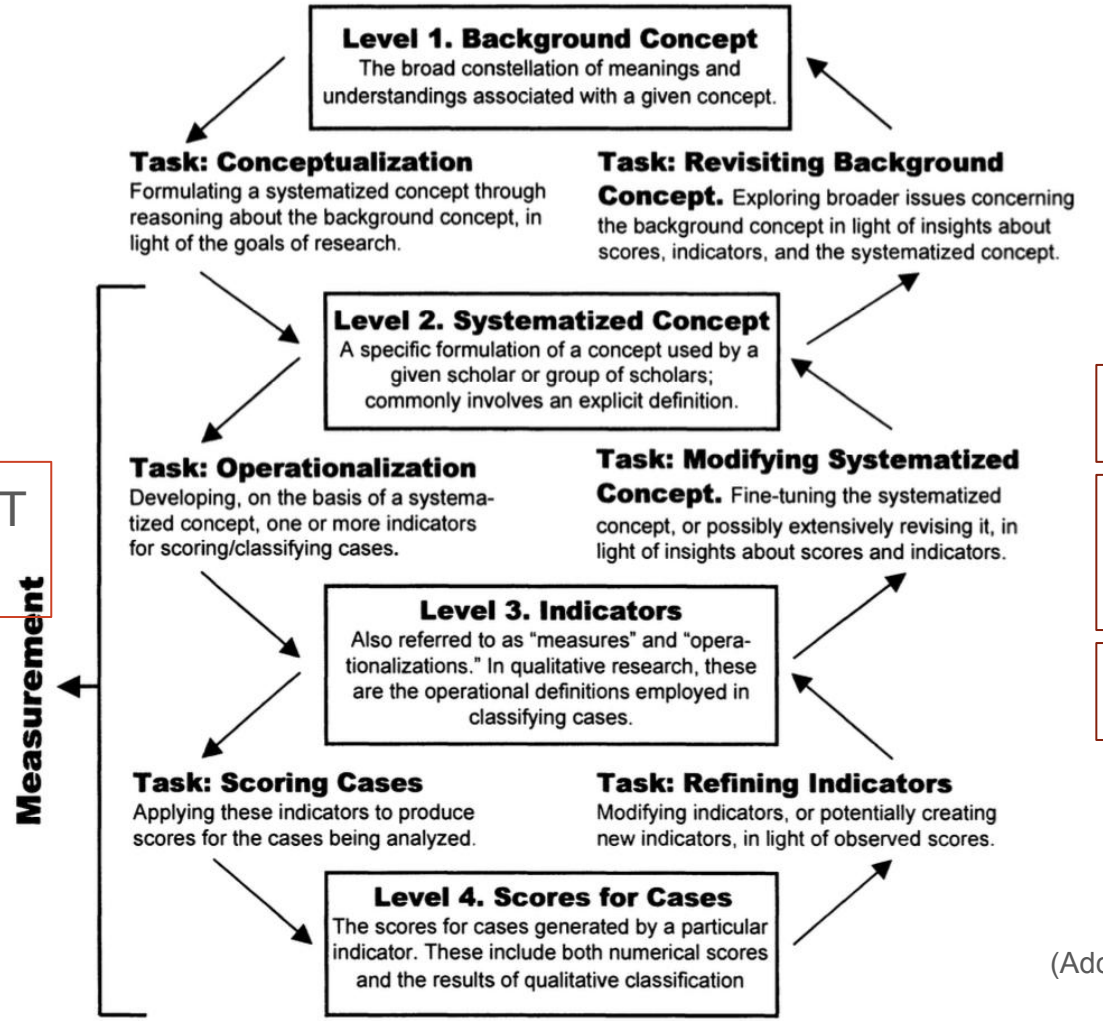
in the large corpora of training texts (20–23). For example, the vector for the adjective honorable would be close to the vector for man, whereas the vector for submissive would be closer to woman. These stereotypes are automatically learned by the embedding algorithm and could be problematic if the embedding is then used for sensitive applications such as search rankings, product recommendations, or translations. An important direction of research is to develop algorithms to debias the word embeddings (20).

In this paper, we take another approach. We use the word embeddings as a quantitative lens through which to study histor-

Gender Stereotypes →

Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

FIGURE 1. Conceptualization and Measurement: Levels and Tasks



MEASUREMENT SYSTEM

LITERATURE

YOUR DOCUMENTS

YOUR BRAIN

(Adcock and Collier, 2001)

What does **validity** mean?

GPT 4o's summary of your pre-course survey responses

1. **Measuring the Intended Concept**: Several responses emphasize that validity is about "measuring what we mean to measure" or "whether the (text-based) thing you are measuring is actually representative of the phenomenon you are claiming to measure." This is the core of **construct validity**.
2. **Accuracy of Classification or Prediction**: Responses like "whether the classes identified by a model in a classification task actually correspond to the 'real' underlying classes" and "does our measurement classify what we intend to classify" focus on whether the classification or prediction aligns with reality, relating to **content validity**.
3. **Generalizability**: One student mentions "to what extent is the method generalizable to another context," addressing **external validity**, i.e., how well the findings apply beyond the studied context.
4. **Model Performance and Logic**: Responses about the "degree in which results are logically found and cohesive" and the idea of evaluating if "the model actually does what it is supposed to" hint at **internal validity** and the logic behind the measurements.
5. **Method and Representation**: Some students focus on "coding schemes" and "how well a method measures what it's supposed to" or whether it represents concepts in the ways humans would interpret them. This touches on the methodological aspect of ensuring valid representation

CONTENT VALIDITY

- are all ingredients in the cake that should be there?
- are ingredients that are not part of the recipe excluded?

CONVERGENT/DISCRIMINANT VALIDATION

- similar dishes correlate with respect to texture and sweetness level
- different dishes are different in terms of texture and sweetness level

Were the differences and similarities as we expected them to be?

NOMILOGICAL/CONSTRUCT VALIDATION

HYPOTHESIS: “Kids do prefer burger over asparagus”

proven by “Kids Digestive Happiness Measure”

Do we find the same result using our “Tiny Tummy Joy Meter“?

Assume the hypothesis, evaluate the measure

LEVEL OF MEASUREMENT

- Full document
- Sentence
- Paragraph
- Pair of documents

LEVEL OF MEASUREMENT

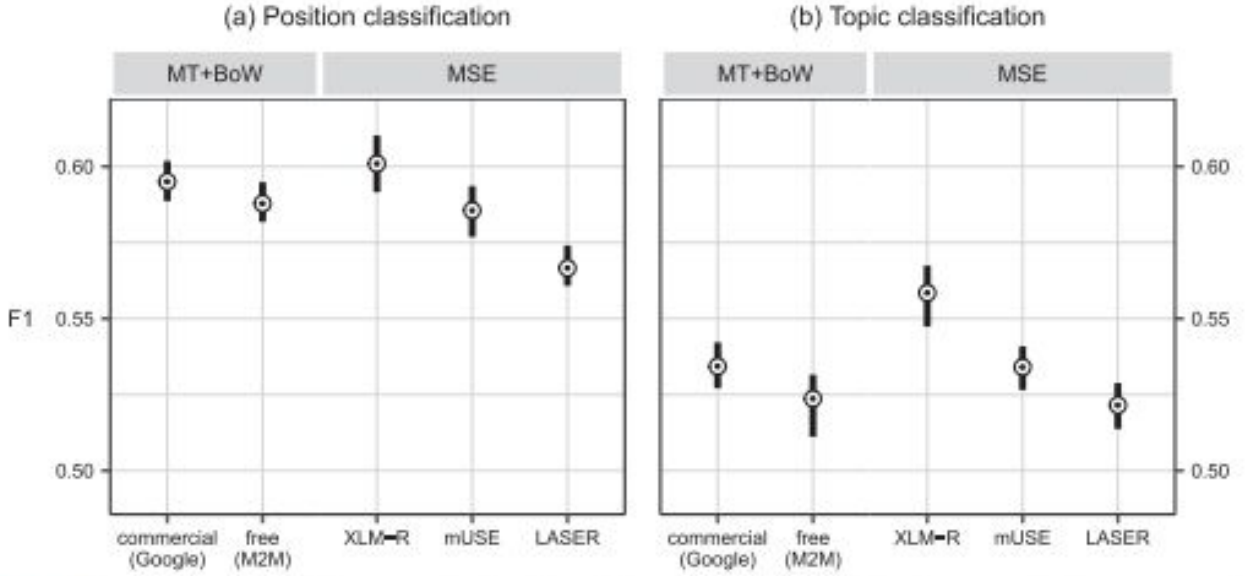


Figure 2. Cross-class mean F1 scores of classifiers trained using different text representation approaches: bag-of-words obtained from machine-translated texts (MT+BoW) and multilingual sentence embeddings obtained from original texts (MSE). Panel (a) reports results for classifying manifesto sentences' positions; panel (b) for classifying their topic. Data plotted summarize 50 bootstrapped cross-class mean F1 scores (excluding the `uncoded` category) for five classifiers per task and approach. Points are averages of bootstrapped estimates, and vertical lines span the 95% most frequent values.

Hauke, 2023

LEVEL OF MEASUREMENT

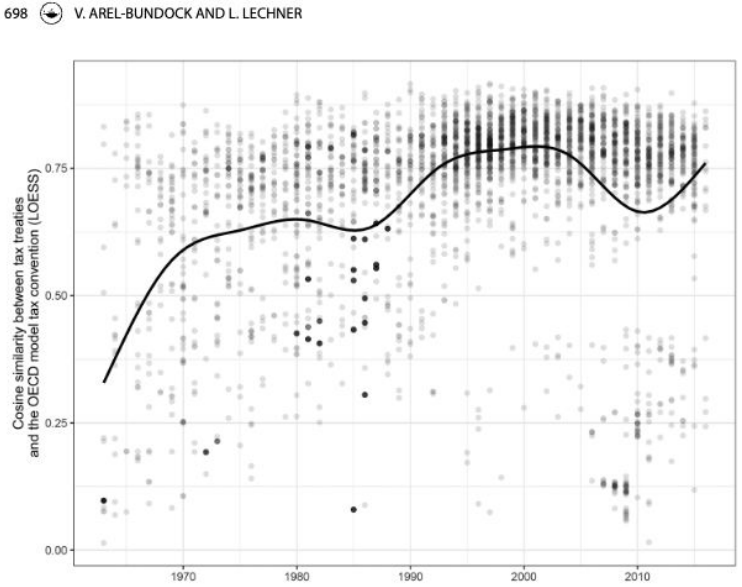
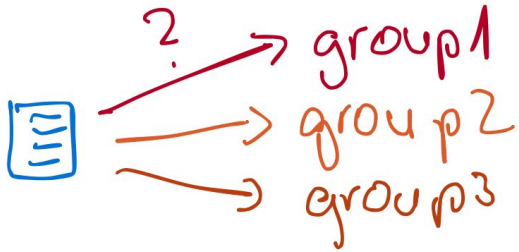


Figure 4. Similarity of bilateral tax treaties to the OECD Model Tax Convention.

MEASUREMENT TASK

CLASSIFICATION



SCALING

COMPARISON

ROLE OF WORD EMBEDDINGS FOR TEXT SCALING - CURRENT OPINION

- Static embeddings can amplify old techniques like dictionary analysis.
  - e.g. Gennaro & Ash (2022), Rice & Zorn (2021)
- Transformer models can fine-tuned/prompted to
  - score texts directly (e.g., O'Hagan & Schein 2023, Mens & Gallego 2024,)
  - to compare pairs of texts for producing data for Bradley-Terry scaling (e.g., Wu et al. 2023)

2018 Fourth International Conference on Information Retrieval and Knowledge Management

FROM THIS…

Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis

Eissa M.Alshari
Computer and Information Techonology
Ibb university, Yemen
eissa.alshari@student.upm.edu.my

Azreen Azman*, Shyamala Doraisamy,
Norwati Mustapha and Mostafa Alkeshr
Universiti Putra Malaysia, Serdang, Malaysia
{azreenazman,shyamala,norwati}@upm.edu.my,
mostafa.alksher@student.upm.edu.my

***Abstract*—Recently, many researchers have shown interest in using lexical dictionary for sentiment analysis. The SentiWordNet is the most used sentiment lexical to determine the polarity of texts. However, there are huge number of terms in the corpus vocabulary that are not in the SentiWordNet due to the curse of dimensionality, which will limit the performance of the sentiment analysis. This paper proposed a method to enlarge the size of opinion words by learning the polarity of those non-opinion words in the vocabulary based on the SentiWordNet. The effectiveness of the method is evaluated by using the Internet Movie Review Dataset. The result is promising, showing that the proposed Senti2Vec method can be more effective than the SentiWordNet as the sentiment lexical resource.**

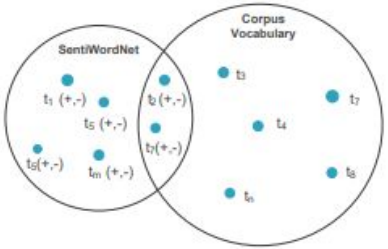
***Index Terms*—Sentiment analysis, Word2Vec, Word embeddings, SentiWordNet**

I. Introduction

Recently, there is an explosive number of user reviews or comments on products and services available on the Web and social media. It has become the source of information for users in making everyday decision, especially on choosing a product

VERB, which can lead to noise for the sentiment classification [6].

In addition, the SentiWordNet does not include all terms in the corpus vocabulary as depicted in Fig. 1. The terms that will be included as the features for sentiment classification reside within the intersection of the two sets. As such, this can be the limitation to the performance of any sentiment analysis approach. It is assumed that by enlarging the size of the intersection will lead to a more effective sentiment analysis.



…TO THIS.

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Back to common sense: Oxford dictionary descriptive knowledge augmentation for aspect-based sentiment analysis

Weiqiang Jin [a], Biao Zhao [a], Liwen Zhang [c], Chenxing Liu [a], Hang Yu [b,*]

[a] School of Information and Communications Engineering, Xi'an Jiaotong University, Innovation Harbour, Xi'an, Shaanxi, 710049, China
[b] School of Computer Engineering and Science, Shanghai University, Shangda Road No. 99., Baoshan, Shanghai, 200444, China
[c] CNBU, Lenovo Future Communication Technology (Chongqing) Company Limited, Yubei District, Chongqing, 401100, China

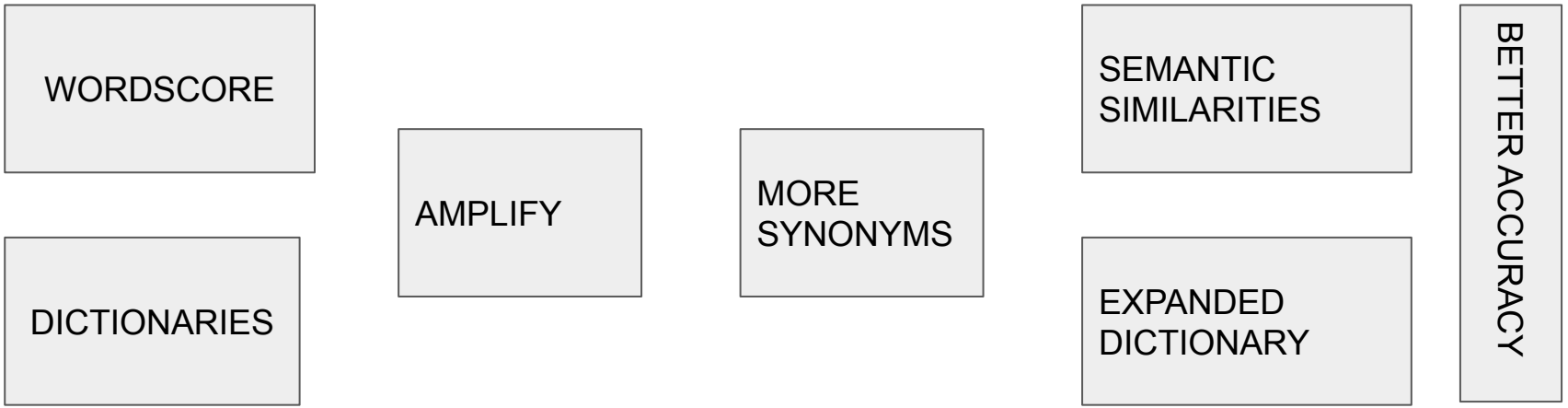
ARTICLE INFO

Keywords:
Natural language understanding
Aspect-based sentiment analysis
Knowledge infusion mechanisms
Pre-trained language models
Model hot-plugging technique

ABSTRACT

Aspect-based Sentiment Analysis (ABSA) is a crucial natural language understanding (NLU) research field which aims to accurately recognize reviewers' opinions on different aspects of products and services. Despite the prominence of recent ABSA applications, mainstream ABSA approaches inevitably rely on large-scale supervised corpora, and their final performances is susceptible to the quality of the training datasets. However, annotating sufficient data is labour intensive, which presents a significant barrier for generalizing a high-quality sentiment analysis model. Nonetheless, humans can make more accurate judgement based on their external background knowledge, such as factoid triples knowledge and event causality. Inspired by the investigations on external knowledge enhancement strategies in other popular NLP research, we propose a novel knowledge augmentation framework for ABSA, named the Oxford Dictionary descriptive knowledge-infused aspect-based sentiment analysis (DictABSA). Comprehensive experiments with many state-of-the-art approaches on several widely used benchmarks demonstrate that our proposed DictABSA significantly outperforms previous mainstream ABSA methods. For instance, compared with the baselines, our BERT-based knowledge infusion strategy achieves a substantial 6.42% and 5.26% absolute accuracy gain when adopting BERT-SPC on SemEval2014 and ABSA-DeBERTa on ACLShortData, respectively. Furthermore, to effectively make use of dictionary knowledge we devise several alternative knowledge infusion strategies. Extensive experiments using different knowledge infused strategies further demonstrate that the proposed knowledge infusion strategies effectively enhance the sentiment polarity identification capability. The Python implementation of our DictABSA is publicly available at https://github.com/albert-jin/DictionaryFused-E2E-ABSA.

ROLE OF WORD EMBEDDINGS FOR TEXT SCALING



INDUCTIVE

(1) **Observation**
(2) **Another observation**
(3) **A third observation**
(4) **Logical conclusion**

DO YOU EXPLORE PATTERNS?

YOU USE UNLABELLED DATA

DEDUCTIVE

(1) **Knowing of a fact**
(2) **Observation**
(4) **Logical conclusion**

DO YOU MEASURE THE PRESENCE/ABSENCE OF PATTERNS?

YOU NEED LABELLED DATA

Chain-of-Thought (CoT) prompting

Deductive Verification of Chain-of-Thought Reasoning

Zhan Ling[1*] **Yunhao Fang**[1*] **Xuanlin Li**[1] **Zhiao Huang**[1] **Mingu Lee**[2]
Roland Memisevic[2] **Hao Su**[1]
[1]UC San Diego, [2]Qualcomm AI Research[†]

Abstract

Large Language Models (LLMs) significantly benefit from Chain-of-Thought (CoT) prompting in performing various reasoning tasks. While CoT allows models to produce more comprehensive reasoning processes, its emphasis on intermediate reasoning steps can inadvertently introduce hallucinations and accumulated errors, thereby limiting models' ability to solve complex reasoning tasks. Inspired by how humans engage in careful and meticulous deductive logical reasoning processes to solve tasks, we seek to enable language models to perform *explicit and rigorous deductive reasoning*, and also ensure the *trustworthiness* of their reasoning process through self-verification. However, directly verifying the validity of an entire deductive reasoning process is challenging, even with advanced models like ChatGPT. In light of this, we propose to decompose a reasoning verification process into a series of step-by-step subprocesses, each only receiving their necessary context and premises. To facilitate this procedure, we propose **Natural Program**, a *natural language-based* deductive reasoning format. Our approach enables models to generate precise reasoning steps where subsequent steps are more rigorously grounded on prior steps. It also empowers language models to carry out reasoning self-verification in a *step-by-step* manner. By integrating this verification process into each deductive reasoning stage, we significantly enhance the rigor and trustfulness of generated reasoning steps. Along this process, we also improve the answer correctness on complex reasoning tasks. Code will be released at `https://github.com/lz1oceani/verify_cot`.

Chain-of-Thought (CoT) prompting

Verification Method	Reasoning Correctness	GSM8k	AQuA	MATH	AddSub	Date	Last Letters	Overall
CoT Two-shot	Correct	98%	96%	100%	92%	100%	96%	97%
	Incorrect	2%	4%	0%	6%	26%	6%	7%
	(Average)	50%	50%	50%	49%	63%	51%	52%
Natural Program One-shot	Correct	84%	72%	70%	95%	90%	96%	85%
	Incorrect	84%	62%	76%	40%	56%	6%	54%
	(Average)	**84%**	**67%**	**73%**	**68%**	**73%**	51%	**69%**

Table 3: Comparison of deductive verification accuracy of reasoning chains for GPT-3.5-turbo (ChatGPT). We compare two approaches: (1) verifying entire reasoning chains generated by Chain-of-Thought prompting; (2) verifying reasoning chains generated in the Natural Program format with step-by-step decomposition. In the latter case, when we verify each reasoning step s_i, we only keep the necessary subset of premises $\bar{p}_i \subseteq p_i$. To calculate verification accuracy, for each dataset, we randomly sample 50 reasoning chains that are deductively valid and 50 reasoning steps exhibiting incorrect reasonings.

APPENDIX

Python packages we use

GENSIM

unsupervised topic modeling & NLP

- document = object of the text sequence type (f.i one tweet, one paragraph, one contract text)
- corpus = collection of documents
- streaming corpus = stream documents one at a time (not all must be loaded, wuhu)

SETUP-CHECKLIST

- GOOGLE COLAB
- VS CODE
- OLLAMA
- OPEN AI

OLLAMA

- **Llama 3.1:** A versatile and powerful model known for its strong performance across various tasks, including text generation, translation, and question-answering.
- **Mistral:** A model renowned for its creative writing abilities, excelling in generating diverse text formats such as poems, scripts, and musical pieces.
- **Code Llama:** A specialized model tailored for coding tasks, assisting developers with code generation, debugging, and understanding complex programming concepts.
- **LLaVA:** A multimodal model capable of processing both text and images, opening up possibilities for creative and visual applications.
- and…

The most capable model

Llama 3 represents a large improvement over Llama 2 and other openly available models:

- Trained on a dataset seven times larger than Llama 2
- Double the context length of 8K from Llama 2
- Encodes language much more efficiently using a larger token vocabulary with 128K tokens
- Less than $\frac{1}{3}$ of the false "refusals" when compared to Llama 2

Google Gemma 2

June 27, 2024

Gemma 2

Google Gemma 2 is now available in three sizes, 2B, 9B and 27B, featuring a brand new architecture designed for class leading performance and efficiency.