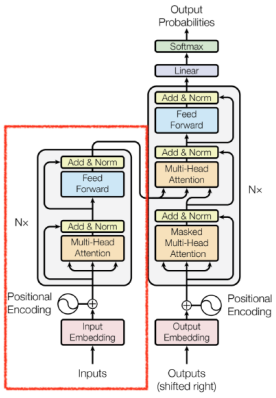
# Lecture 3: Transformers and encoder-based models in Social Science

Miriam Hurtado Bodell

LINKÖPING UNIVERSITY

# Recap transformers/encoders



## Encoders (transformers) in the social sciences

How can we use encoder part of transformer models?

Discovery

- In contrast to WE models, individual vectors and their relations may not be the main interest – but similarly to WE you can study relations between words/sentences to explore cultural associations (which can handle polysemy better!)

Measurement

- The main task to use encoder (so far) is classification of the presence of relevant sociologically relevant concepts in a text

## Discovery: encoders (transformers) in the social sciences

What does vectors from (some different hidden layer(s) in) the encoder in transformer models represent?

- Technical answer: Numerical representation of a word's position in a vector space, taking context into consideration
- Sociological answer: ?

## Discovery: encoders (transformers) in the social sciences

What does vectors from (some different hidden layer(s) in) the encoder in transformer models represent?

- Technical answer: Numerical representation of a word's position in a vector space, taking context into consideration
- Sociological answer: The contextualized meaning of a word

→ meaning is again relational, but also contextual!

## Discovery: encoders (transformers) in the social sciences

No consensus on which layers best capture the specific contextualized semantic meaning of a word

- Input: static embeddings
- First layer: probably contains very limited contextual info
- Last layer: may be overfitted for the training task
- Second to last layer may be best to use, or a combination of the last couple of layers

## Measurement: encoders (transformers) in the social sciences

What type of sociologically relevant concept can we use encoders to measure? For example;

- Frames
  “Protecting our borders and upholding the Second Amendment are essential to preserving our American values and ensuring the safety and security of our citizens.”

  "Ensuring that every American has access to affordable healthcare and quality education is fundamental to our nation’s progress and prosperity"

## Measurement: encoders (transformers) in the social sciences

What type of sociologically relevant concept can we use encoders to measure? For example;

- Frames
- Sentiment
- Political ideology/"isms"

Most importantly; pre-trained encoder can be used for new task by fine-tuning them

Measurement: encoders (transformers) in the social sciences: why is transfer learning/fine-tuning important for social scientists?

The main disadvantage of supervised models previously were that they used to be costly

Pre-trained models *may* encode information about language/culture

→ Models can learn supervised classification tasks with less labeled training data

→ Researcher doesn't have to (or pay others to) annotate as much training data

Fine-tuning = target task-specific output layer is put on top of the pretraining architecture

Example fine-tuning for **measurement**: Le Mens et al. (2023)

Aim: Measure typicality of literary genres using BERT

What is typicality? Related to the literature on categorization (large lit. within for example organization science) = (mental) process of deciding in X an instance of concept Y; e.g. is *Avatar* a sci-fi movie?

Commonly though about as a binary classification task; yes or no but typicality measure the degree to which X can be seen as an instance of concept Y

Example fine-tuning for **measurement**: Le Mens et al. (2023)

Data: Descriptions of books from 36 different genres (partly focus on Mystery and Romance) + humanly annotated prediction set

Method: Feed text to pre-trained BERT model. Use BERT embeddings as input in neural net to classify what genre books belong to; add softmax layer to get probability $p_c \rightarrow \tau_c = log\frac{p_c}{p(c)}$

Compare $\tau_c$ with human annotator asked to answer how typical a book is to a genre (0-100)

Example fine-tuning for **measurement**:

| Typicality measure | Correlations between model-based typicalities and human typicality ratings: Romance and Non-Romance books | Romance books | Non-Romance books | Model components: Language representation | Categorization | Similarity between text document and concept |
|---|---|---|---|---|---|---|
| BERT fine-tuned / categorization probability (baseline BERT typicality) | 0.86 | 0.54 | 0.72 | BERT | Probabilistic binary | Typicality (Eq. [4]) |
| BERT fine-tuned / correlation with prototype | 0.85 | 0.56 | 0.70 | BERT | Probabilistic binary | Cosine |
| BERT pre-trained / categorization probability | 0.70 | 0.40 | 0.55 | BERT | Probabilistic binary | Typicality (Eq. [4]) |
| BERT pre-trained / correlation with prototype | 0.07 | 0.11 | 0.04 | BERT | None | Cosine |
| BERT fine-tuned / categorization probability 36 | 0.85 | 0.51 | 0.67 | BERT | Probabilistic 36 categories | Typicality (Eq. [4]) |
| GloVe fine-tuned / categorization probability | 0.77 | 0.53 | 0.54 | GloVe | Probabilistic binary | Typicality (Eq. [4]) |
| GloVe fine-tuned / correlation with prototype | 0.59 | 0.27 | 0.47 | GloVe | Probabilistic binary | Cosine |
| GloVe pre-trained / categorization probability | 0.76 | 0.45 | 0.56 | GloVe | Probabilistic binary | Typicality (Eq. [4]) |
| GloVe pre-trained / correlation with prototype | 0.30 | 0.30 | 0.22 | GloVe | None | Cosine |
| Word frequencies / categorization probability | 0.62 | 0.34 | 0.42 | BoW term frequencies | Probabilistic binary | Typicality (Eq. [4]) |
| TF-IDF / categorization probability | 0.76 | 0.48 | 0.54 | BoW TF-IDF | Probabilistic binary | Typicality (Eq. [4]) |

Example fine-tuning for **measurement/inference**: Vicinanza et al. (2023)

Aim: Answer where prescient ideas come from; the core or the periphery of different domains? And how does it relate to success?

Prescient ideas = first challenge conventions, then become widely accepted (e.g. civil rights). (1) novel & (2) future-looking

Data: Politics (speeches from floor), Law (court rulings), and Business (Quarterly earnings calls)

Example fine-tuning for **measurement/inference**: Vicinanza et al. (2023)

Method: Fine-tune one BERT model per time unit (year or presidential term)

Calculate contextual novelty (CN)= word-level perplexity in a sentence; how "surprised" is the model from $t = 0$ and $t = 1$ to the a word in a sentence

Calculate sentence prescience $= \frac{CN_{t0} - CN_{t1}}{CN_{t0}}$

Example fine-tuning for **measurement/inference**: Vicinanza et al. (2023)

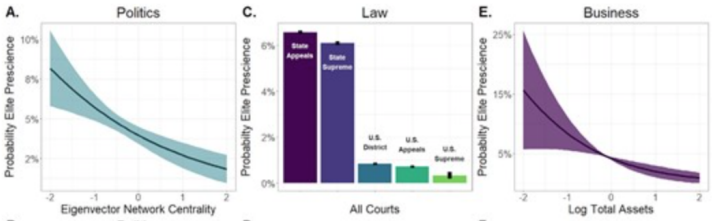



Figure: Prescient ideas come from the periphery in all domains

Example fine-tuning for **measurement/inference**: Vicinanza et al. (2023)

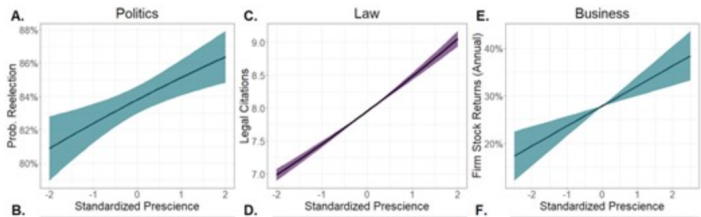



Figure: Prescient actors are more successful

## What encoder to use?

Depend on the task you want them to preform and your data

- Different models are available for different languages
- Did they train on data similar to yours? (e.g. BERT books, Roberta also includes web)
- How long are your texts; different models can deal with different context sizes (e.g. BERT 512 tokens, Longformer 4096 tokens)

No general guidelines, study model performance across tasks that are relevant to your study!

*) Wankmüller, S. (2022); Timoneda, J. C., Vallejo Vera, S. (2024)

## Validation

The main validation technique is to compare with human annotation or human responses

- Annotate random sample and compare with classification output
- Use found labeled data (e.g. reviews)
- Conduct a survey and compare with the classification output

## Validation

Encoder models should not replace social scientists, but augment them! (Do et al. 2023)

- The quality of measurement (classification) still depends on having high-quality labels to fine-tune/validate your model
- Who does the labeling matter; the more expert knowledge the better the model performance in less time

## Limitations

Pre-trained models may "bleed" unwanted information from pre-training to measurement/inference task

- Your validation data is part of training data
- Measurement/inference about subgroups/individuals will be influenced by others in training data

## Summary

Encoder-only transformers has shown great ability in both simple and complex measurement tasks

But their ability still depends on human annotation; the better the training/validation labels the better the model performance

Important to consider how training data may “spill” unwanted information into your inference

# Thank you!

www.liu.se

LINKÖPING UNIVERSITY