Political Text Analysis with Embedding Regression

From Multilingual to Cross-lingual Application

Yen-Chieh Liao[†], Winnie Xia[‡], Chen Zeng[§] and Slava Jankin[†]

[†]University of Birmingham, [‡]Aarhus University, and [§]Kings' College London

2025 APSA

10 September 2025

Overview

Extension of Current Embedding Regression Techniques to Cross-lingual Applications

Political communication occurs in crosslingual settings (e.g., EU Parliament and UN), but the current ConText R package (Rodriguez, Spirling and Stewart, 2023) relies primarily on static embeddings with limited multilingual capabilities.

Limitations

Current methods miss sequential and bidirectional context.

Survey and Evaluation

Three embedding frameworks tested on EU Parliament data: **static**, **sequential**, and **dynamic models**. 36 MEPs, 6 languages, human-coded speeches.

Main Findings

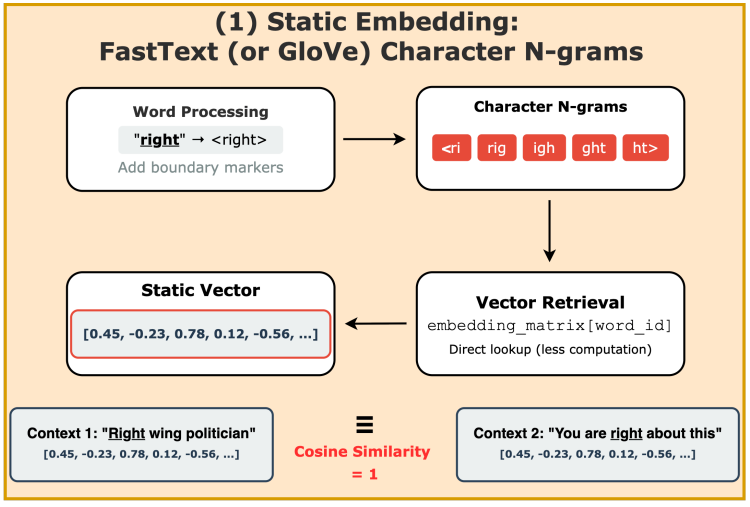
XLM-RoBERTa and sequence models perform consistently, but BPE offers the best accuracy-efficiency balance. Open source package coming soon.

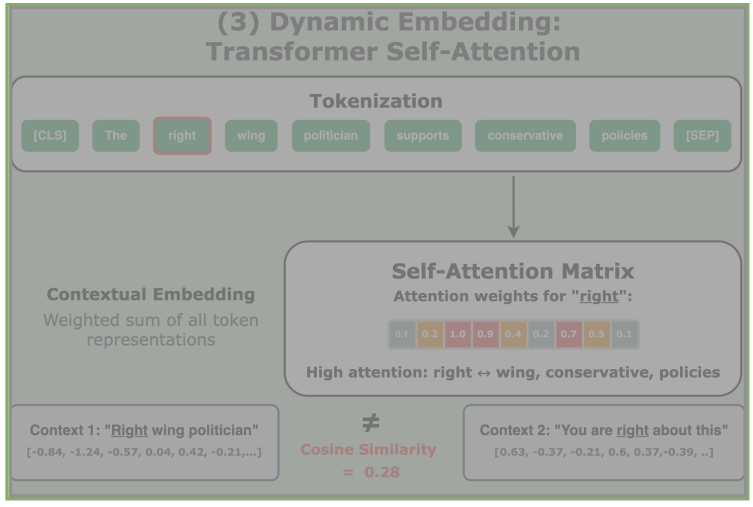
Proof of Concept (1/4)

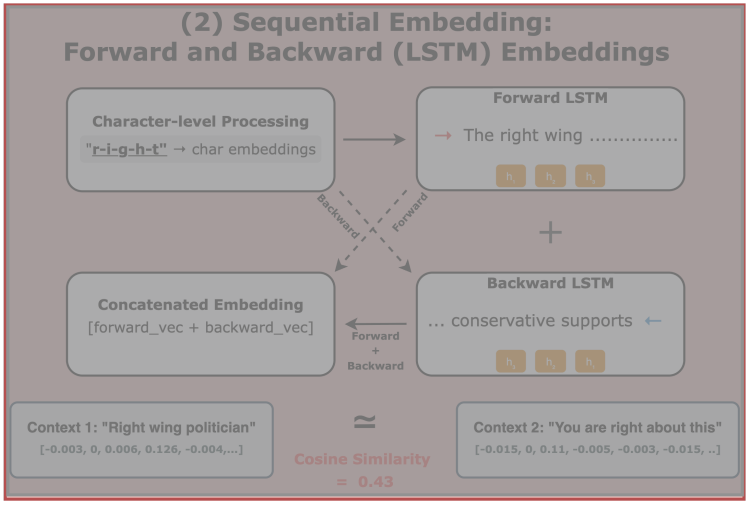
Same Words in Different Context

> "Right wing politician"
>
> v.s
>
> "You are right about this"

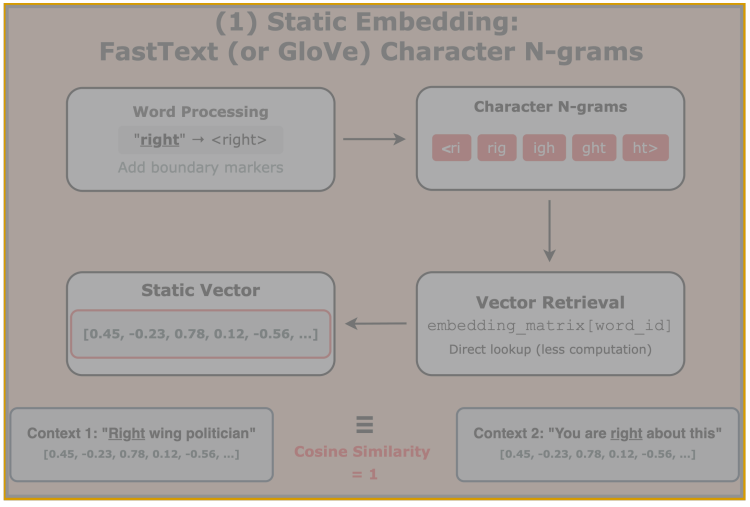
Proof of Concept (2/4)

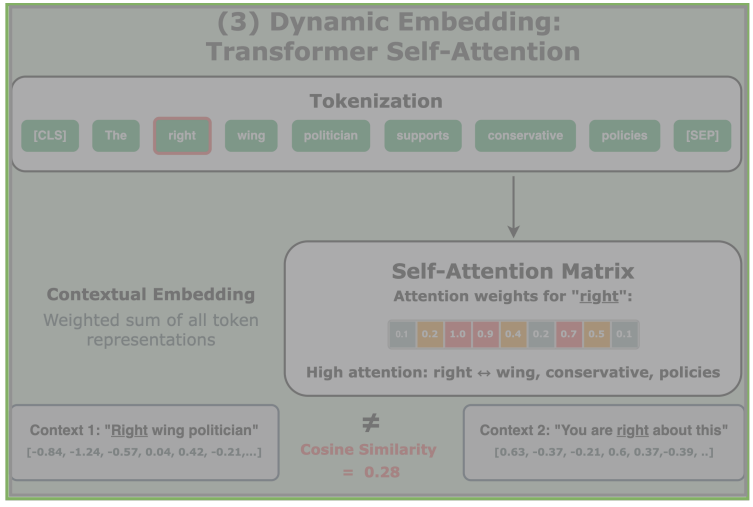


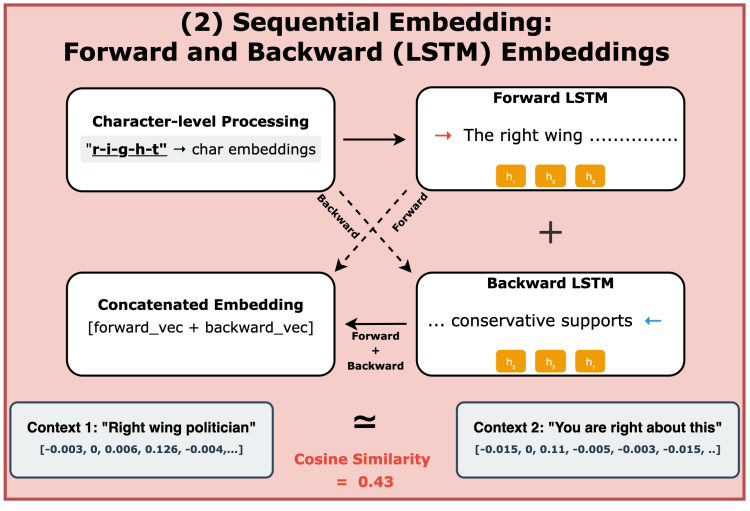




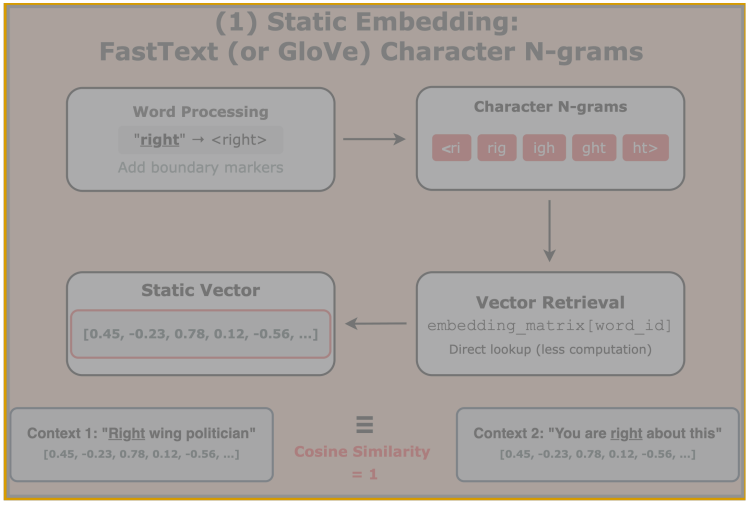
Proof of Concept (2/4)

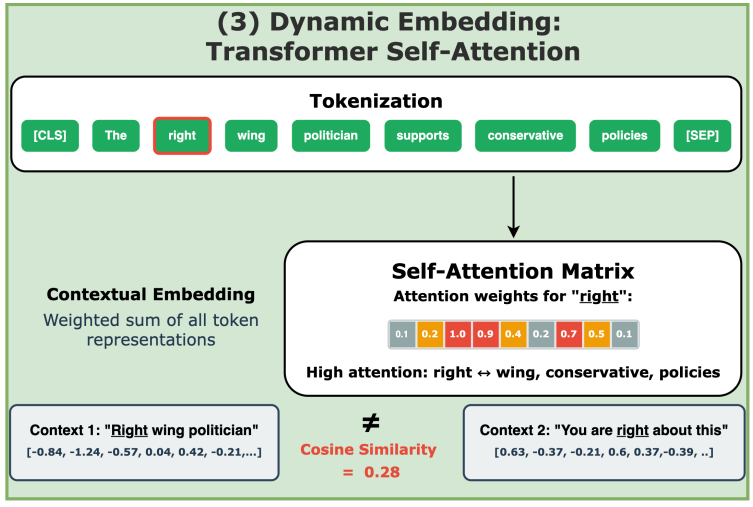


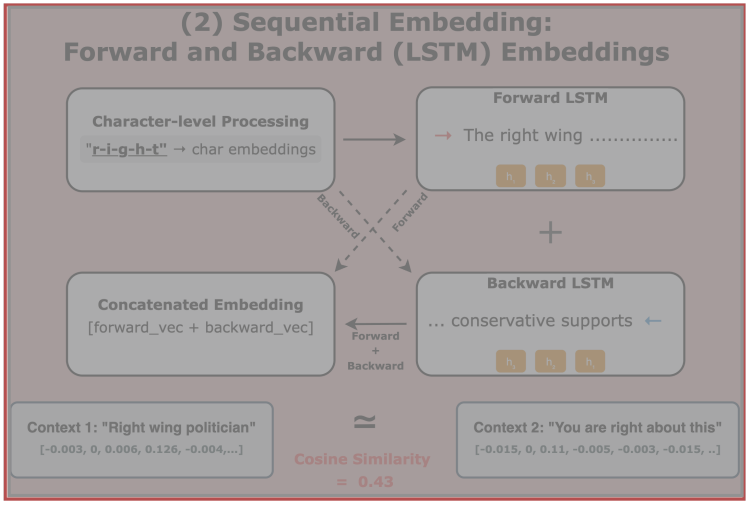




Proof of Concept (2/4)







Proof of Concept (2/4)

Table 2: Contextual Disambiguation Performance Across Embedding Architectures

Method	**Context 1**	**Context 2**	***Diff*Δ**	**Disambiguation**
Static (GloVe)	0.0139	0.0139	0	1.000
Sequential (LSTM based Forward + Backward)	0.0018	0.0016	0.0001	0.434
Dynamic (BERT)	-0.0005	-0.0014	0.0008	0.280

Note: Context 1 and Context 2 represent vectors for "**right**" extracted from "*Right wing politician*" and "*You are right about this*", respectively. The ***Disambiguation*** score represents the cosine similarity between these two context vectors. ***Diff*Δ** represents the vector difference between two contexts.

Proof of Concept (3/4)

Table 1: Comparison of Embedding Method Techniques

Characteristics	Static	Sequential	Dynamic
Architecture and Tokenization	Local window (Word2Vec and fastText), global co-occurrence (GloVe) and Subword-based (BPE)	LSTM	Transformer
Language Support	fastText (157); BPE (275)	Forward (12); Backward (12)	XLM-RoBERTa (100); Multilingual BERT (102)
Model Type	Multilingual by concatenating language-specific models	Cross-lingual through sequential modeling	Cross-lingual through attention mechanisms
Contextual Information	*Word-level embeddings or subword tokenization*	*Character-level sequential modeling with contextual embeddings*	*Subword-level with global self-attention mechanism*
Representative Methods	fastText (Bojanowski et al., 2016), GloVe (Pennington, Socher and Manning, 2014); BPE (Sennrich, Haddow and Birch, 2015)	Flair Embedding (Akbik, Blythe and Vollgraf, 2018)	BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019)

Proof of Concept (4/4)

Table 1: Comparison of Embedding Method Techniques

Characteristics	Static	Sequential	Dynamic
Architecture and Tokenization	Local window (Word2Vec and fastText), global co-occurrence (GloVe) and Subword-based (BPE)	LSTM	Transformer
Language Support	fastText (157); BPE (275)	Forward (12); Backward (12)	XLM-RoBERTa (100); Multilingual BERT (102)
Model Type	Multilingual by concatenating language-specific models	Cross-lingual through sequential modeling	Cross-lingual through attention mechanisms
Contextual Information	*Word-level embeddings or subword tokenization*	*Character-level sequential modeling with contextual embeddings*	*Subword-level with global self-attention mechanism*
Representative Methods	fastText (Bojanowski et al., 2016), GloVe (Pennington, Socher and Manning, 2014); BPE (Sennrich, Haddow and Birch, 2015)	Flair Embedding (Akbik, Blythe and Vollgraf, 2018)	BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019)

Embedding Model Architectures and Specifications

Table 4: Embedding Model Architectures and Specifications

Type	Embedding Models	Original Dimensions	Final Dimensions	Architecture	Language Support
Static	FastText	300 (per language)	1,800 (6×300)	Word2Vec-based	6 languages concatenated
Static	BytePair Encoding	600	600	Subword-based	275 languages
Sequential	Flair Backward	4,096	4,096	Backward LSTM	12 languages
Sequential	Flair Forward	4,096	4,096	Forward LSTM	12 languages
Sequential	Flair Backward+Forward	4,096 (per direction)	8,192	BiLSTM	12 languages
Dynamic	Multilingual BERT (MultiBERT)	768	768	Transformer	104 languages
Dynamic	XLM-RoBERTa (base)	768	768	Transformer	100 languages

Note: The FastText model utilizes Flair NLP's `StackedEmbeddings` function to concatenate 300-dimensional FastText embeddings from six languages (English, German, Spanish, Italian, Polish, and Greek), resulting in 1,800 total dimensions. The Flair models include three variants: Backward (backward LSTM), Forward (forward LSTM), and Backward+Forward (bidirectional LSTM combining both directions). Each direction contributes 4,096 dimensions, with the bidirectional model totaling 8,192 dimensions. The Dynamic models (MultiBERT and XLM-RoBERTa) use Transformer architectures with contextual embeddings at 768 dimensions each.

Dataset: MEPs Coal Mining Debates (Benoit et al. 2016)

Captures Fundamental Divide

- **Purpose**: End subsidies by 2014 vs. extension to 2018+
  - **Competing Interests**: Environmental goals vs. local employment
  - **Ideological Divide**: Market mechanisms vs. government intervention

Source: Spiegel International

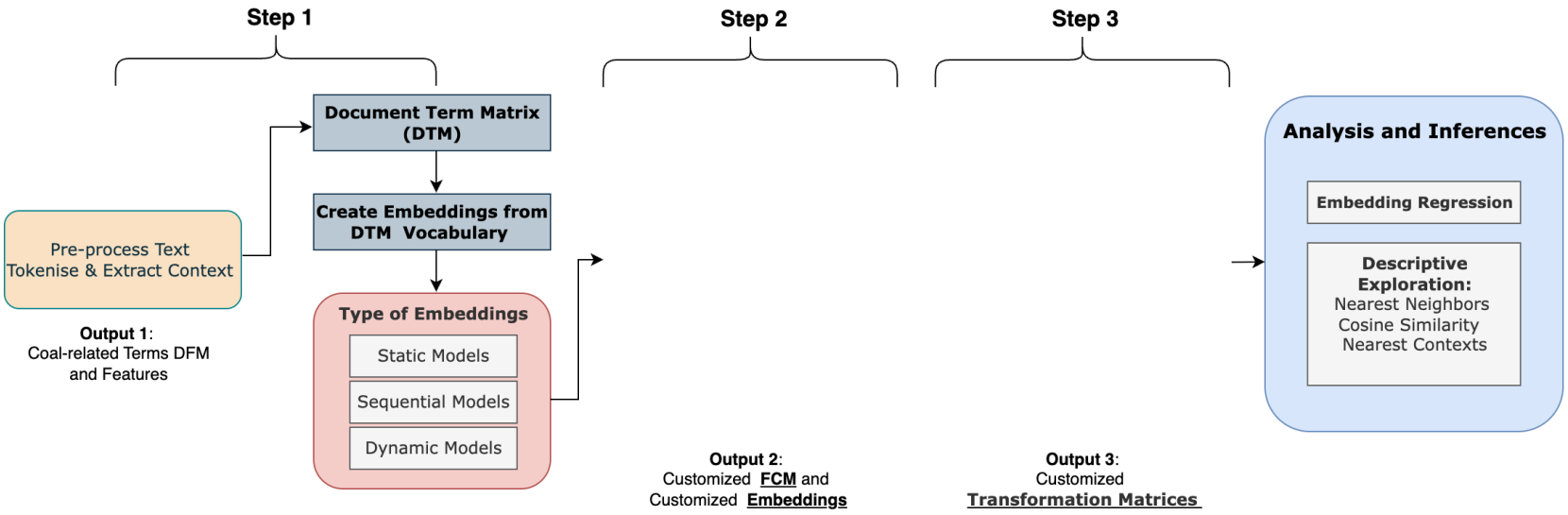
- **36 MEPs, 6 languages**
  - **Human-annotated**: +1 (support) to -1 (oppose), 3-5 coders per sentence
  - **Real voting outcomes**: Debate followed by actual recorded votes

Table 3: The Statistics of Coal Debate Corpus

Language	English	German	Spanish	Italian	Greek	Polish
Sentence N	414	455	418	349	454	437
Total Judgments	3,545	1,855	2,240	1,748	2,396	2,256

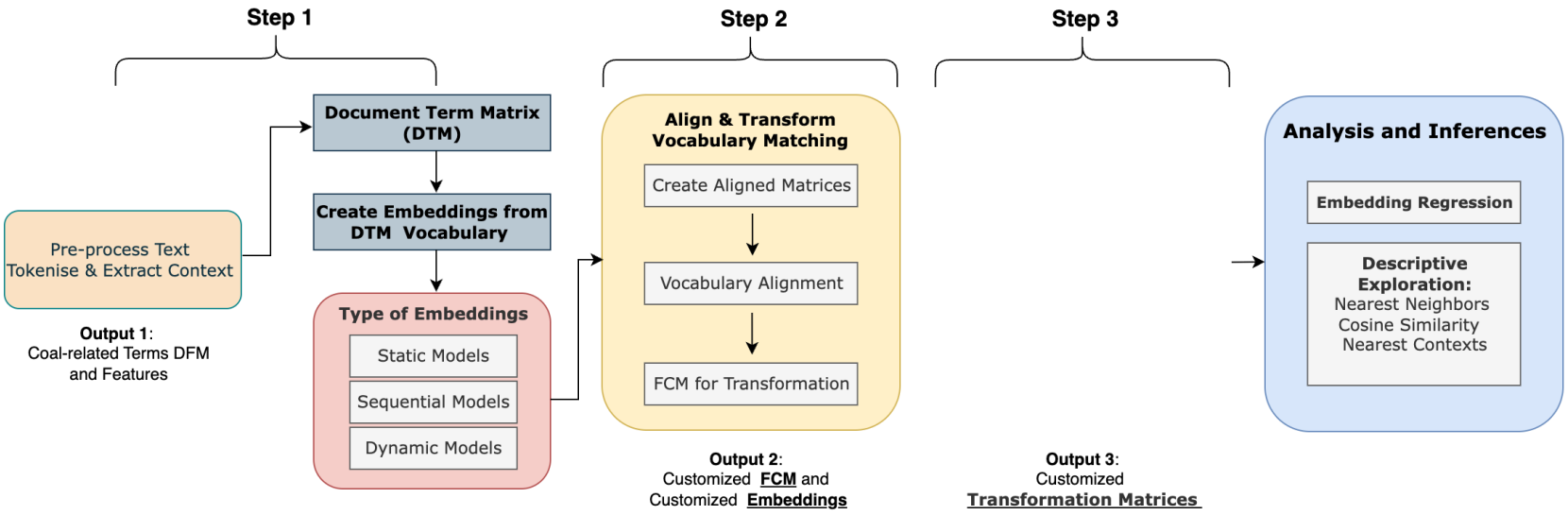
Source: Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016)

Building Context-Aware DEM (1/3)

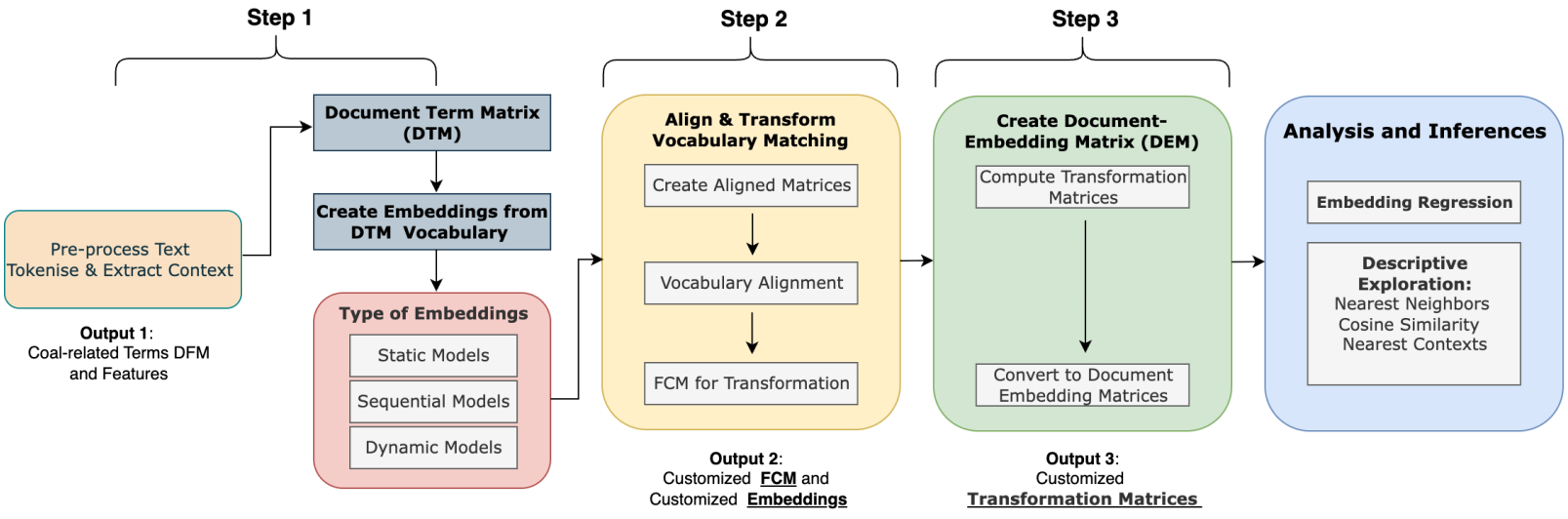


Context-aware extraction: Identify sentences with coal policy terms ("coal", "energy", "mining", "power", "subsidies") within 6-token windows across languages.

Building Context-Aware DEM (2/3)



Building Context-Aware DEM (3/3)



Three Output Matrices → Embedding Regression

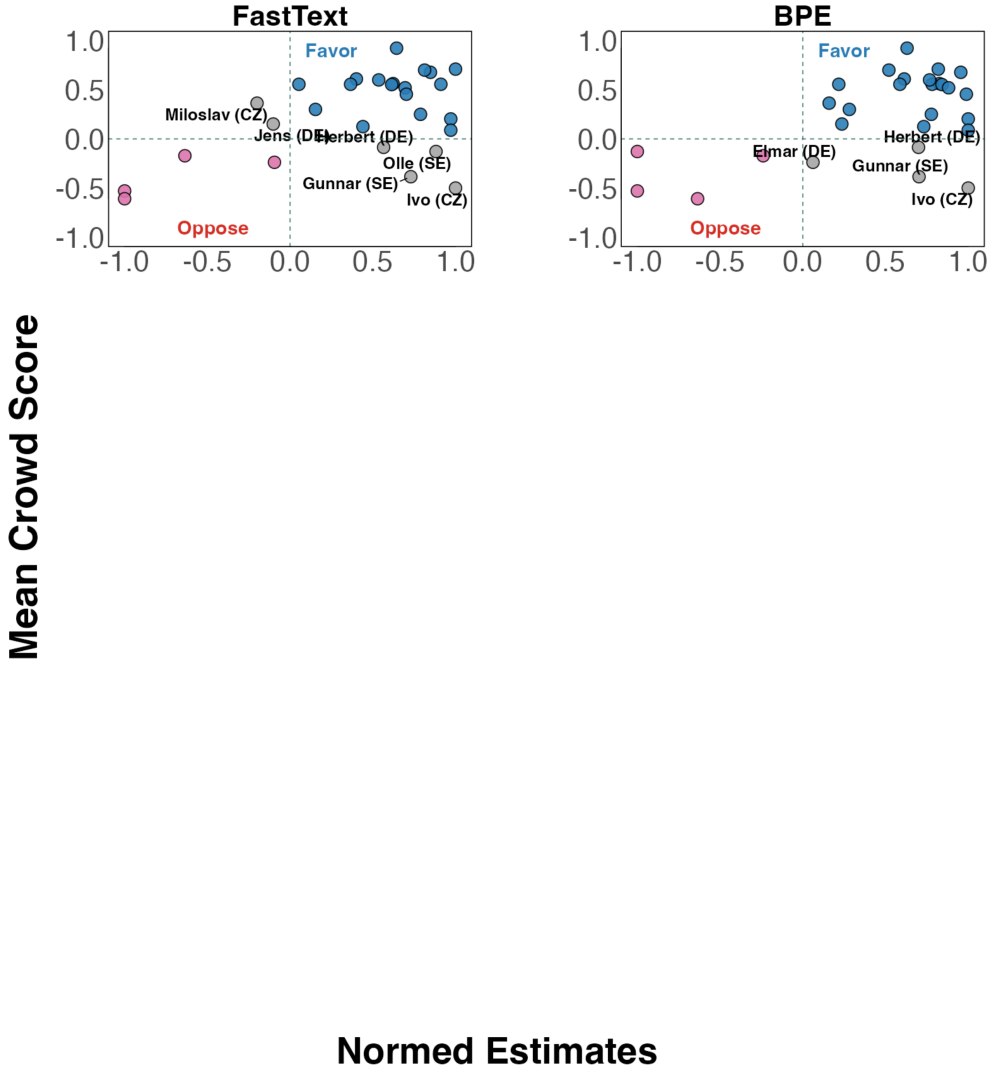
Input Matrices

- **DFM**: Document-Feature Matrix (bag-of-words)
- **FCM**: Feature Co-occurrence Matrix (word patterns)
- **DEM**: Document Embedding Matrix (dense vectors)

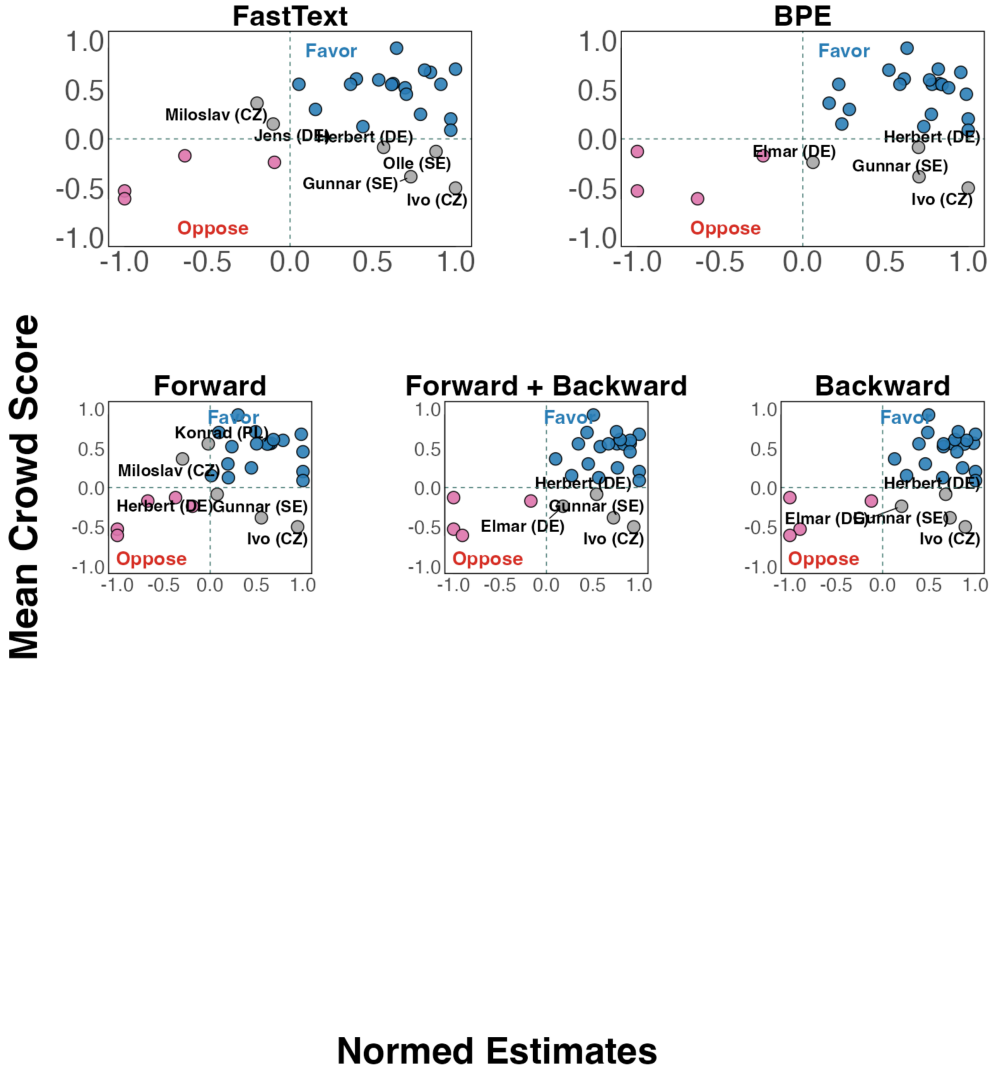
Process

- Target word: **"coal"** with ±6 token window
- Context-aware sentence extraction across languages

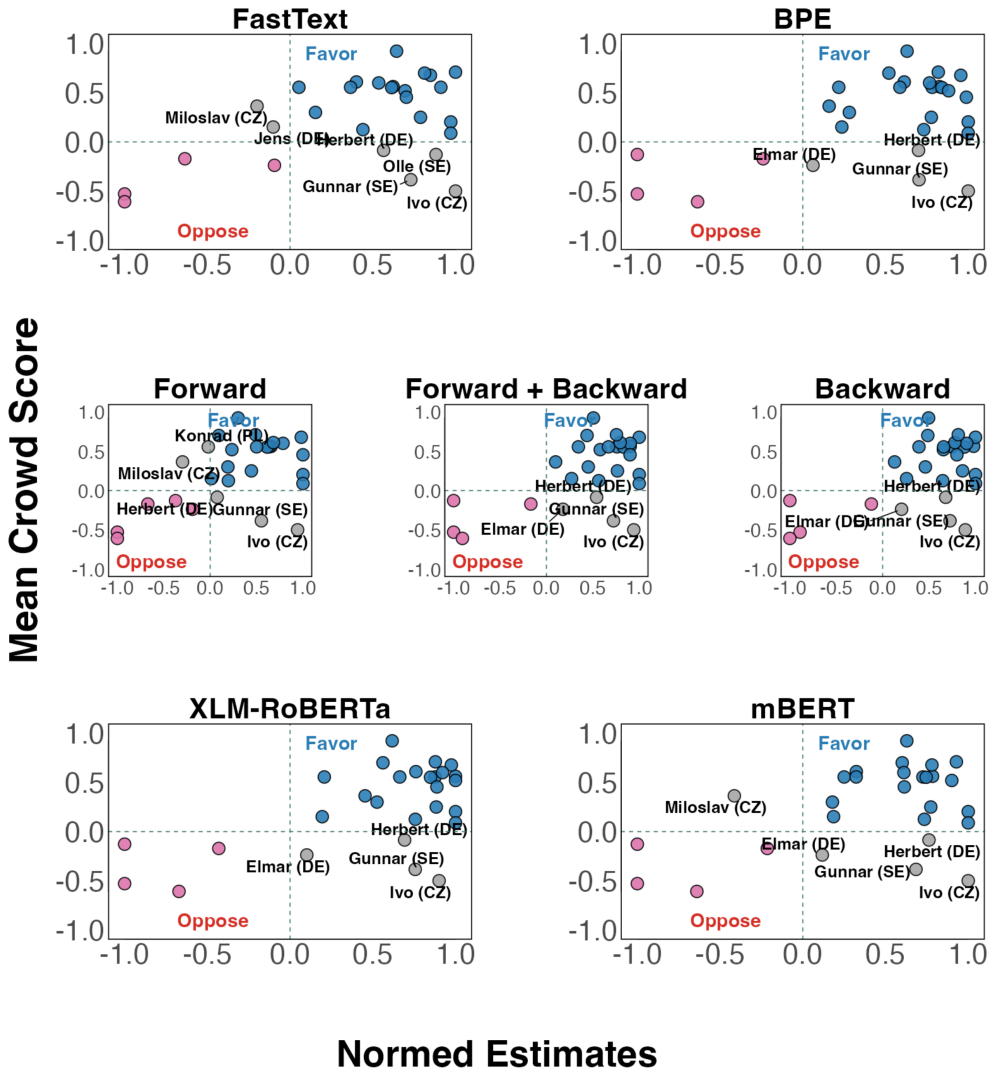
Finding I: Comparison of Esitmates and Crowd Scores (1/5)



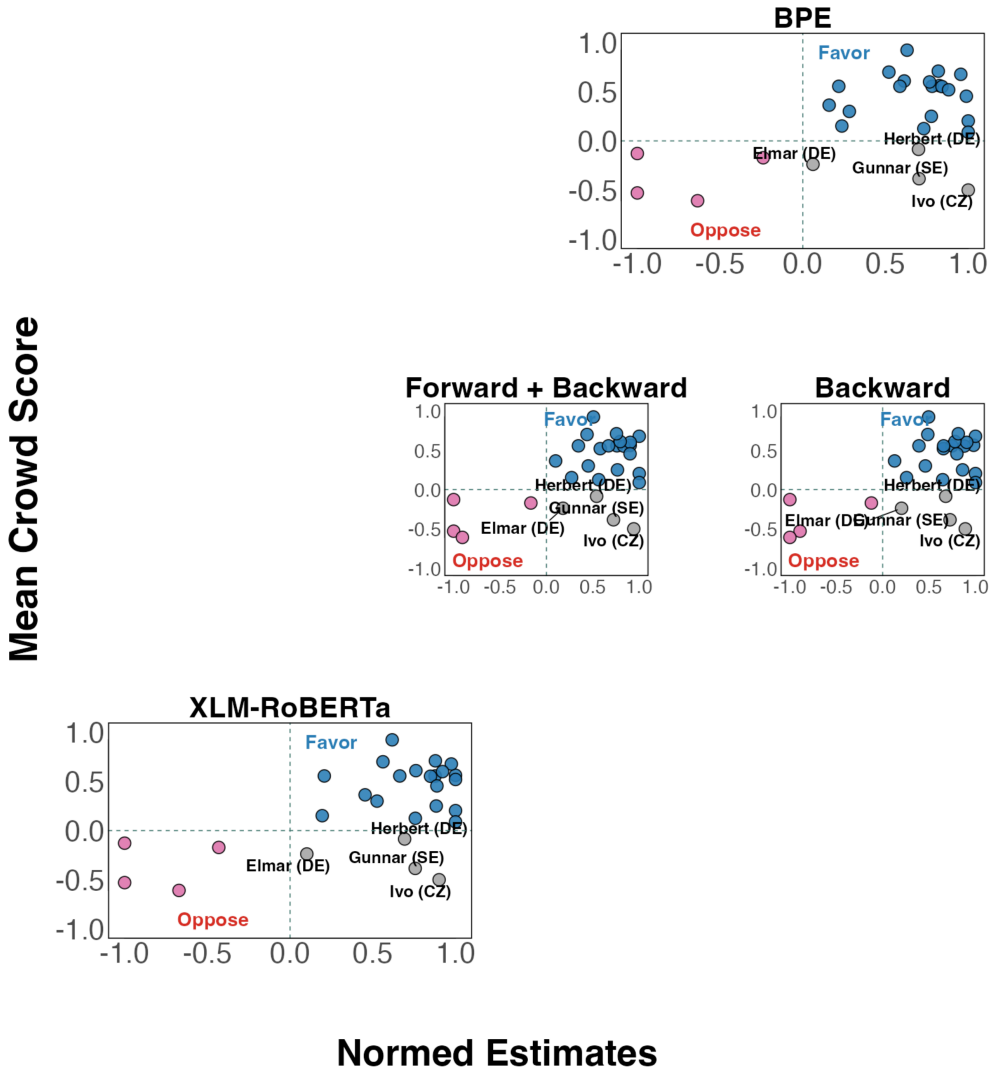
Finding I: Comparison of Esitmates and Crowd Scores (2/5)



Finding I: Comparison of Esitmates and Crowd Scores (3/5)



Finding I: Comparison of Esitmates and Crowd Scores (4/5)



Finding I: Comparison of Esitmates and Crowd Scores (5/5)

Appendix A MEPs with Inconsistent Classifications: Notes from (Benoit et al., 2016)'s Replication Files

Appendix Table A.1: Inconsistent MEPs: Voting Behavior vs. Embedding-Based Estimates and Crowd Scores

MEPs	Vote	XLM-RoBERTs | BPE | Backward	Mean Scores	Original Notes
Herbert Reul (DE)	Yea	Pro Favor	Pro Oppose (-0.086)	*positive but questions Commission's proposal*
Ivo Strejek (CZ)	Yea	Pro Favor	Pro Oppose (-0.500)	*positive but negative about subsidies*
Elmar Brok (DE)	Yea	Pro Favor	Pro Oppose (-0.238)	*positive about proposal but speech focuses more on criticizing the Liberals' position*

Note: DE = Germany, CZ = Czech Republic; CDU = Christlich Demokratische Union Deutschlands; ODS = Obanská demokratická strana. Original notes are from Benoit et al. (2016) replication files.

Finding I: Comparison of Esitmates and Crowd Scores (5/5)

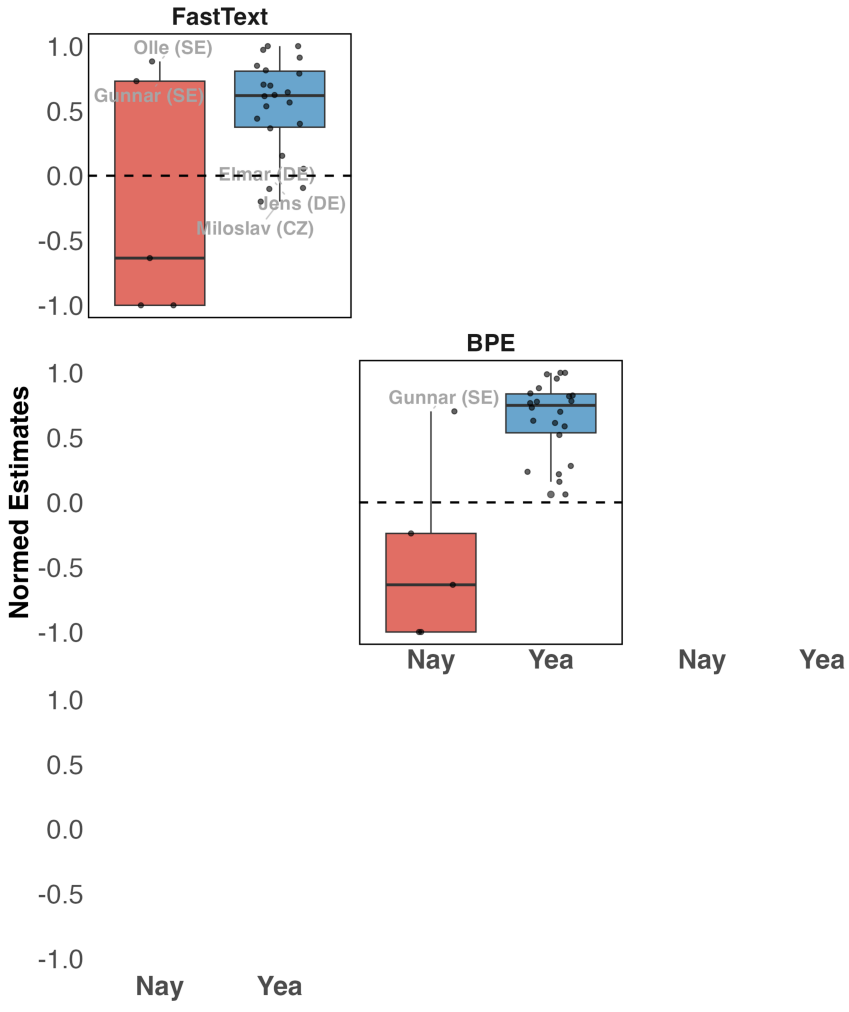
Appendix A MEPs with Inconsistent Classifications: Notes from (Benoit et al., 2016)'s Replication Files

Appendix Table A.1: Inconsistent MEPs: Voting Behavior vs. Embedding-Based Estimates and Crowd Scores

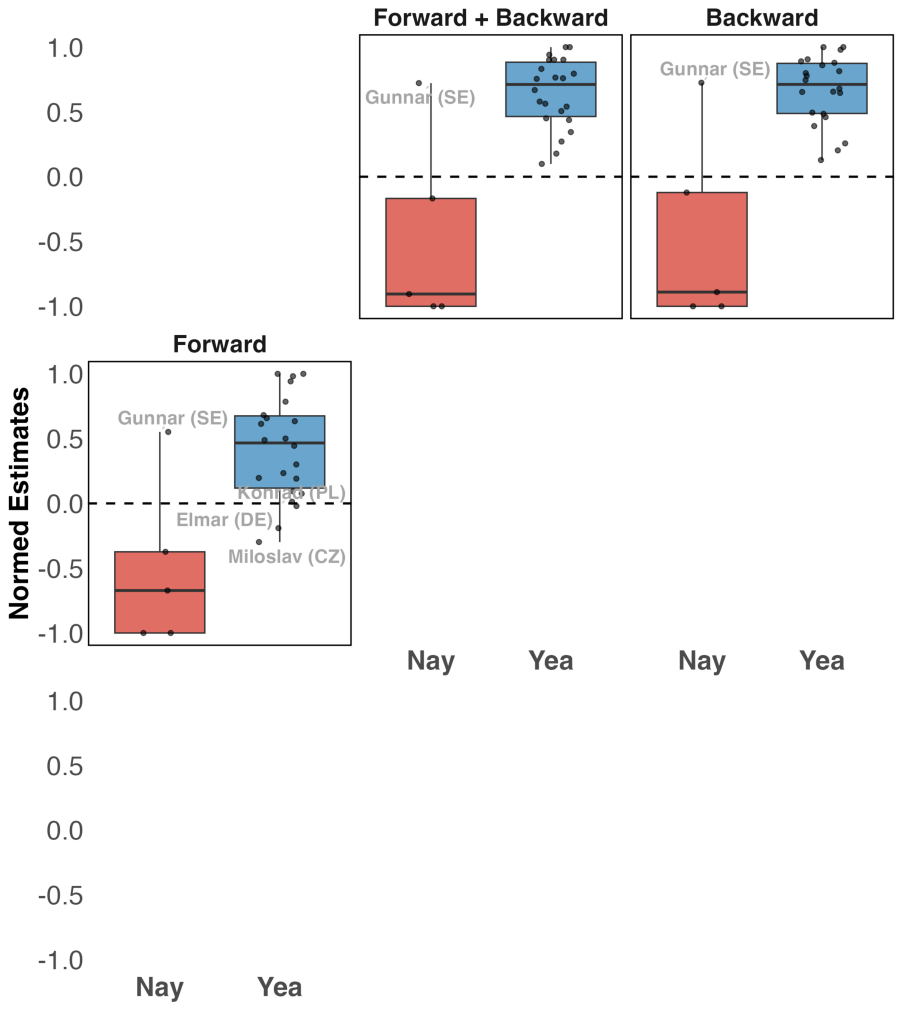
MEPs	Vote	XLM-RoBERTs | BPE | Backward	Mean Scores	Original Notes
Herbert Reul (DE)	Yea	Pro Favor	Pro Oppose (-0.086)	*positive but questions Commission's proposal*
Ivo Strejek (CZ)	Yea	Pro Favor	Pro Oppose (-0.500)	*positive but negative about subsidies*
Elmar Brok (DE)	Yea	Pro Favor	Pro Oppose (-0.238)	*positive about proposal but speech focuses more on criticizing the Liberals' position*

Note: DE = Germany, CZ = Czech Republic; CDU = Christlich Demokratische Union Deutschlands; ODS = Obanská demokratická strana. Original notes are from Benoit et al. (2016) replication files.

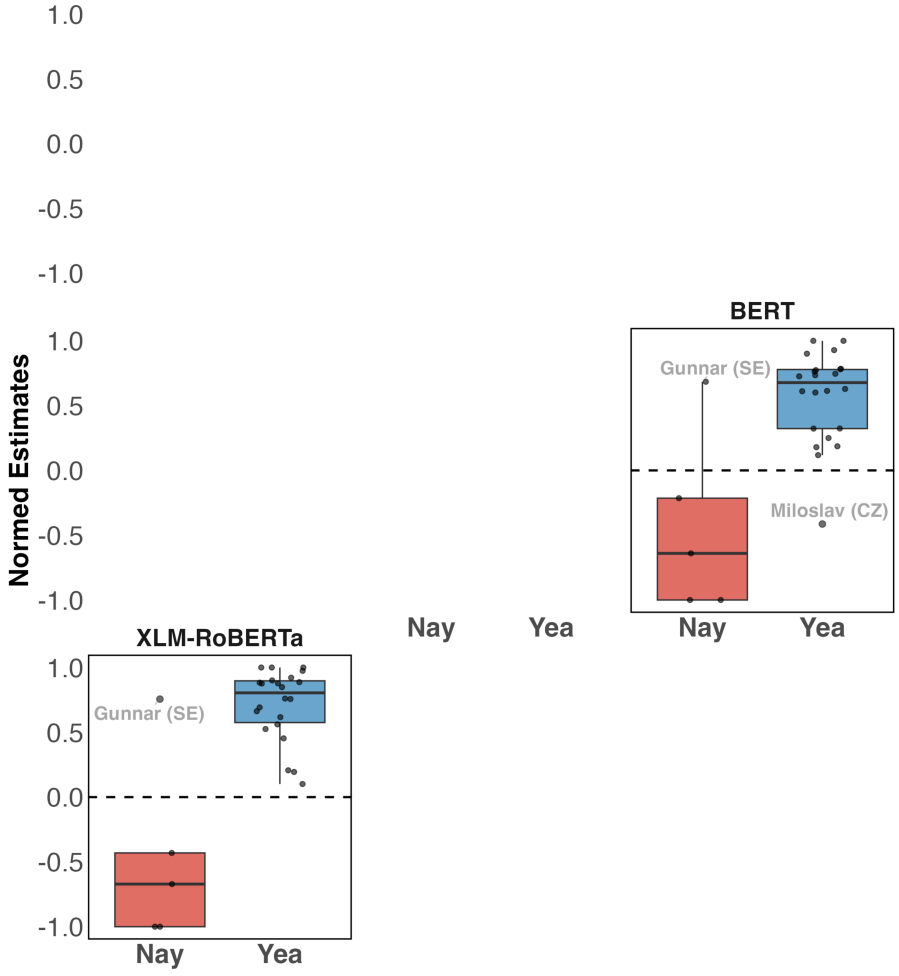
Finding II: Comparison of Esitmates and Legislative Votes (1/4)



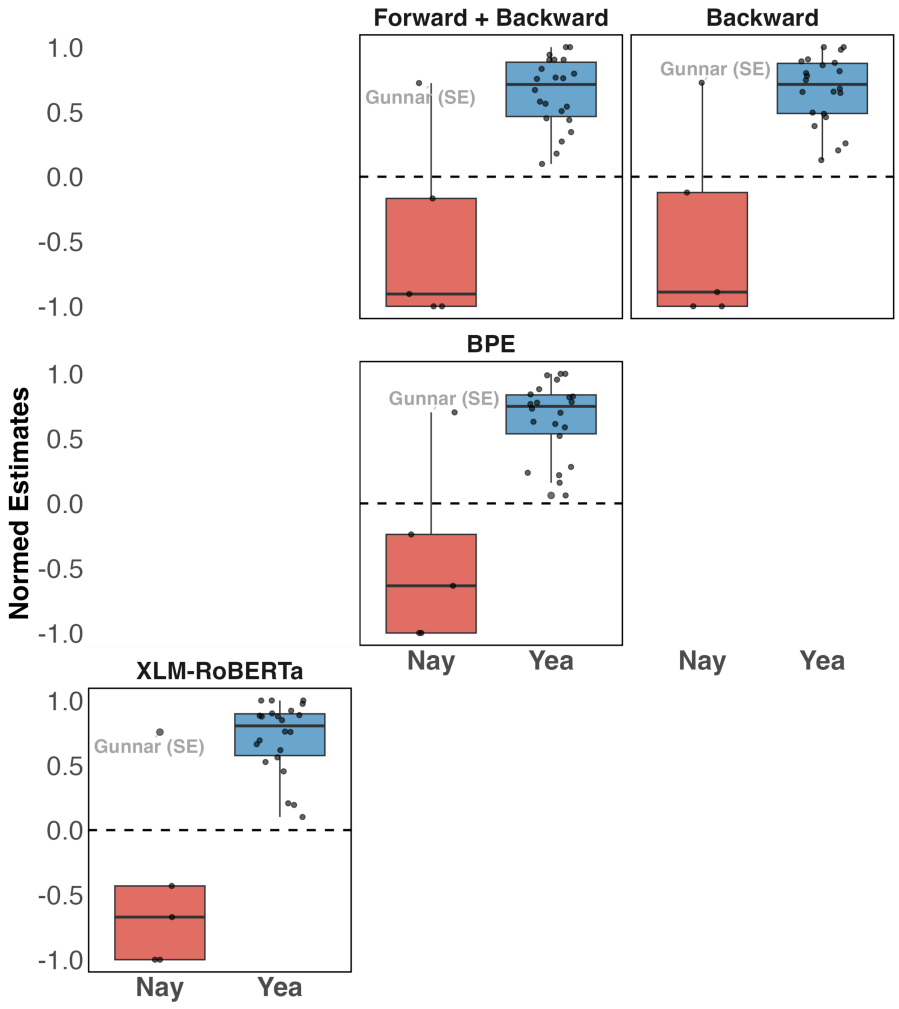
Finding II: Comparison of Esitmates and Legislative Votes (2/4)



Finding II: Comparison of Esitmates and Legislative Votes (3/4)



Finding II: Comparison of Esitmates and Legislative Votes (4/4)



To Wrap-up

Application Guidelines

- **UN/EU Parliament**: Use XLM-RoBERTa for critical cross-lingual consistency
- **Limited resources**: BPE provides balanced multilingual capability
- **Large-scale monolingual**: fastText/LSTM
  - Wirsching et. al (2025)

Important Caveats

- Domain expertise crucial for seed word selection

Next Steps

- Future: Fine-tuned models (i.e., stance detection model fine-tuned by Liat & Müller or Burnham 2024, PSRM) & decoder-based LLM comparison
- Open-source package + tutorials coming soon

Thank You