Who is talking?

Discover what lays behind political speeches.

Motivation

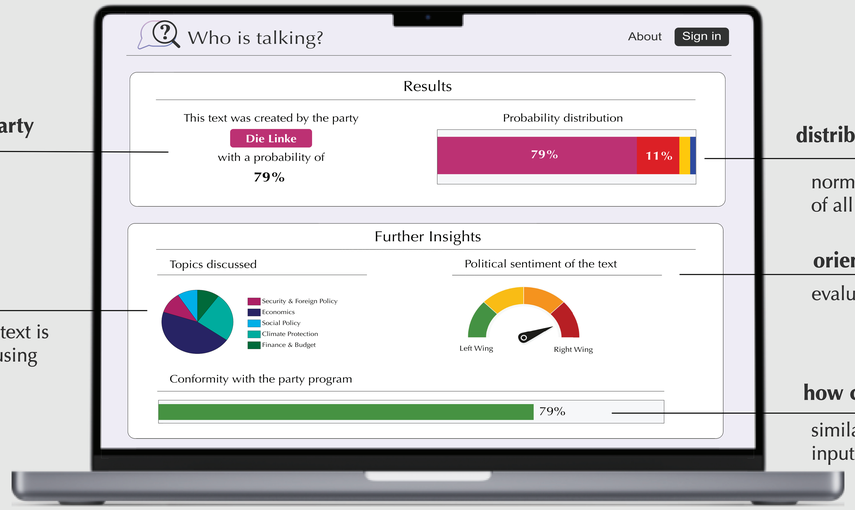
Wouldn't it be interesting to know what political affiliation can be inferred from a text? And that from a more neutral observer than humans?

“Who is talking?” aims to create a tool that is able to analyze language in terms of its political orientation and reliably estimate its political direction.

Aside from the personal amusement of analyzing the last message in the family chat and sending the result back without comment, the following broader benefits could emerge:
- uncovering the political bias of articles and news
- analyzing propagandistic language and determining its characteristics
- analyzing the evolution of the parliament’s parties sentiments towards topics

This Project represents the first stages of the application. Many of the envisioned developments (see Outlook) are yet to be implemented and remain as future goals.

Results



probability of the most likely party

Which topics are discussed?

indicates how strongly the input text is associated with a specific topic using topic modeling

distribution of probabilities

normalized distribution of probabilities of all 6 parties through SoftMax

orientation of the text

evaluated from topic sentiment analysis

how conform is it with the party`s program?

similarity of topic sentiments from the input text with the party’s official program

Methods

Phase 1: Setting up the data

Collect data

Sources:
- bundestag transkripts
- Ana Calotescu, who did an NLP analysis of parlamentary speeches

Clean data

Remove irrelevant data and clean the text from meaningless expressions

Prepare and preprocess

Oversampling,
Tokenization,
Stemming &
Normalization

data_merged_revised.csv		
text	date	legislative_period
speaker_name	speaker_gender	speaker_role
speaker_party	comments	id
speaker_id	speech_length	

speaker_party (Class):
“CDU/CSU”, “SPD”, “FDP”, “BÜNDNIS 90/DIE GRÜNEN”, “DIE LINKE”, “AfD”

Size: ~1.5 million speech samples

Phase 2: Machine Learning

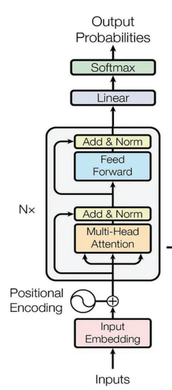
Build ML-Pipeline

BERT-base-uncased:
12-layer
768-hidden-nodes
12-attention-heads
110 mio. parameters

Train model

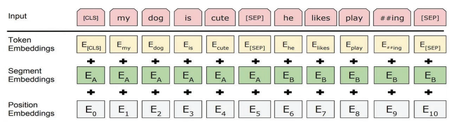
BERT-Classifier

multi-layer bidirectional **transformer** encoder that represents the input text in a high-dimensional space
-> includes the entire context of each word in a sentence



https://medium.com/@alexmriggio/bert-for-sequence-classification-from-scratch-code-and-theory-fbd805138000a

Token Embedding:
The final Token Embedding is passed to the self-attention mechanisms to add contextual information



https://medium.com/@alexmriggio/bert-for-sequence-classification-from-scratch-code-and-theory-fbd805138000a

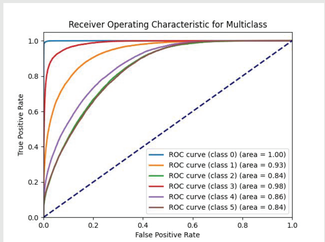
BERT transformer - workflow
The input text is tokenized, embedded, and enriched with positional encodings to capture word order. These embeddings pass through multiple transformer layers, where multi-head attention and feed-forward networks help the model understand contextual relationships. The [CLS] token is then classified via a linear layer and softmax to produce output probabilities for each class.

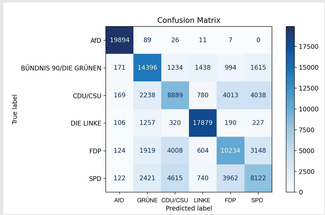
Phase 3: Deployment & Integration

Build application with live prediction

Phase 4: Prepare for further development

Evaluation






```
Accuracy: 0.6617833333333333
F1-Score: 0.6556660903638116
Precision: 0.6517570637266652
Recall: 0.6617833333333333
ROC AUC: 0.9091904324284251
```

Outlook

Expanding the model to include **multiple languages** and dialects will enable international scalability.

Integrating **topic sentiment analysis** could provide deeper insights into the political discourse. It could help identify **how different parties frame topics** like healthcare or immigration, and how sentiment around these topics correlates with political bias.

Implementing **real-time text evaluation**, such as for political debates or social media, through a user-friendly API could further increase accessibility and **feedback loops** provide performance gains.

This program could be extended to **assess political bias in content from public media institutions and evaluate audience engagement**, such as analyzing sentiment in YouTube comments. This could provide valuable insights into media influence and public opinion.

References

1 https://www.bundestag.de/services/opendata
2 https://zenodo.org/records/10895276
3 Ewonus, B., McCann, B. and Roth, N., CS 229 Final Project-Party Predictor: Predicting Political Affiliation.
4 Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 79–84 Brussels, Belgium, October 31, 2018. 2018 Association for Computational Linguistics
5 Biessmann, F., Lehmann, P., Kirsch, D., Schelter, S.: Predicting political party affiliation from text. PolText 2016 14, 14 (2016)

still in progress...

Hochschule Düsseldorf
University of Applied Sciences
HSD

Marie-Juliette Yek

University of Applied Sciences Düsseldorf

Advances in Intelligent Systems

Prof. Dr. Dennis Müller

Fachbereich Medien
Faculty of Media