Putin's Talks SOTA Analysis & POC

Project for Natural Language Processing Course

Łukasz Grabarski
Łukasz Lepianka
Marta Szuwarska

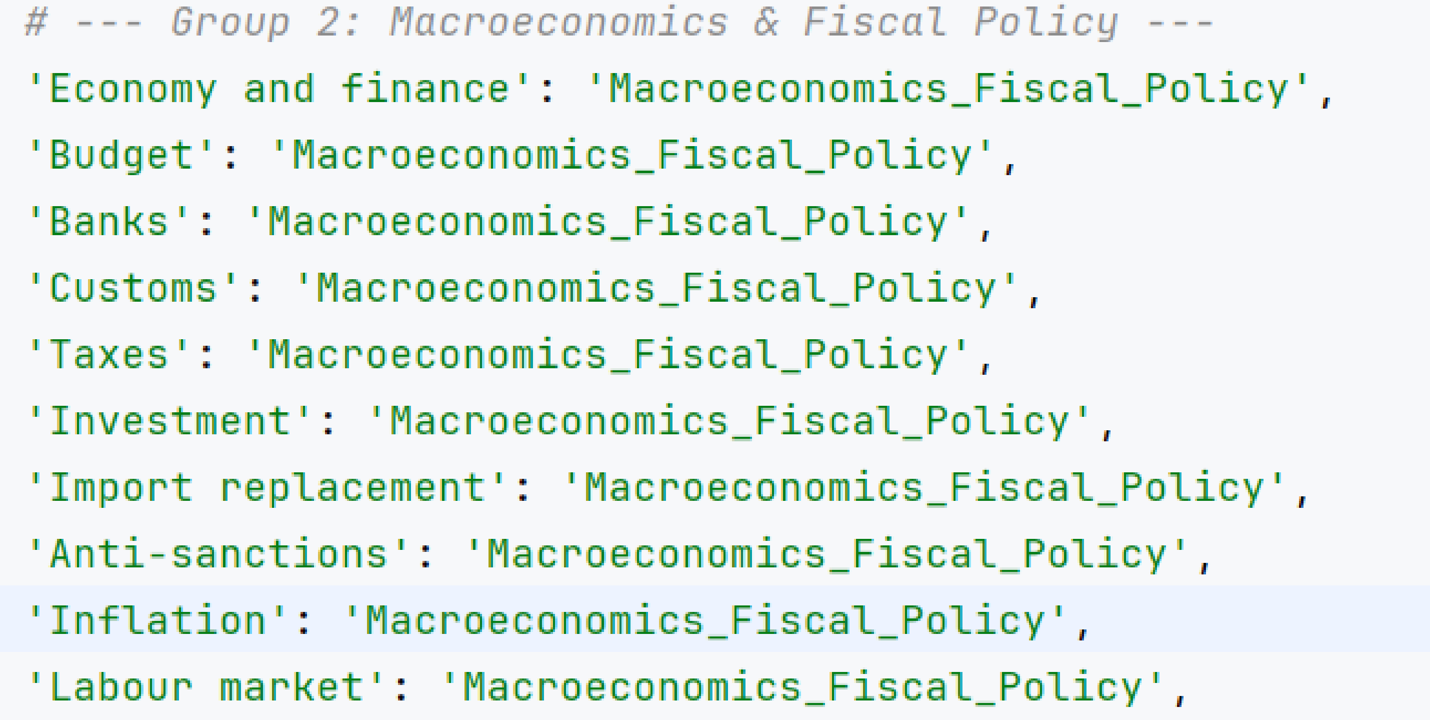
November 2025

Introduction - lots of questions regarding Putin's talks

- How many times do the words “Poland”, “Ukraine”, ... appear in the entire database?
- In which years did Putin most often mention ...?
- Which country, apart from Russia, appears most frequently in his speeches?
- In how many speeches does the word “democracy” appear?
- In which years do the most references to World War II occur?
- In what context does “Poland” most often appear? (enemy, partner, ally, neighbour)
- Is Russia more often described as a “victim”, a “leader”, or a “defender”?
- How often does he mention NATO expansion - before and after 2004?
- How often does he speak about “threats” before and after 2014?
- At what point does Putin start talking about a “multipolar world”?
- List all the countries mentioned in the speech from date Y
- What adjectives or terms mostoften accompany the word “Ukraine”?

Dataset – EDA

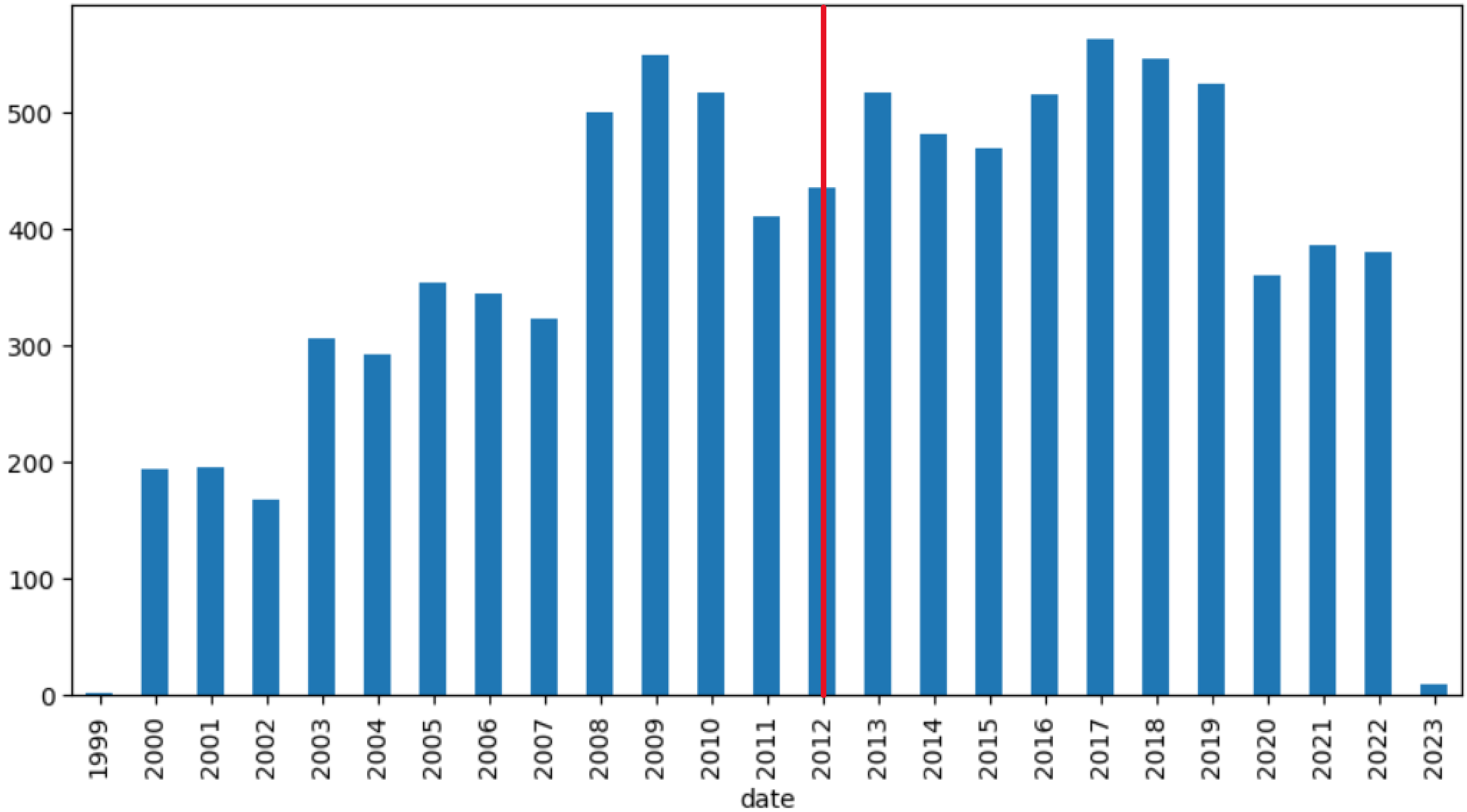
- Over 9000 speeches
- 86 tags -> 13 grouped tag categories
- Data filtered to presidency



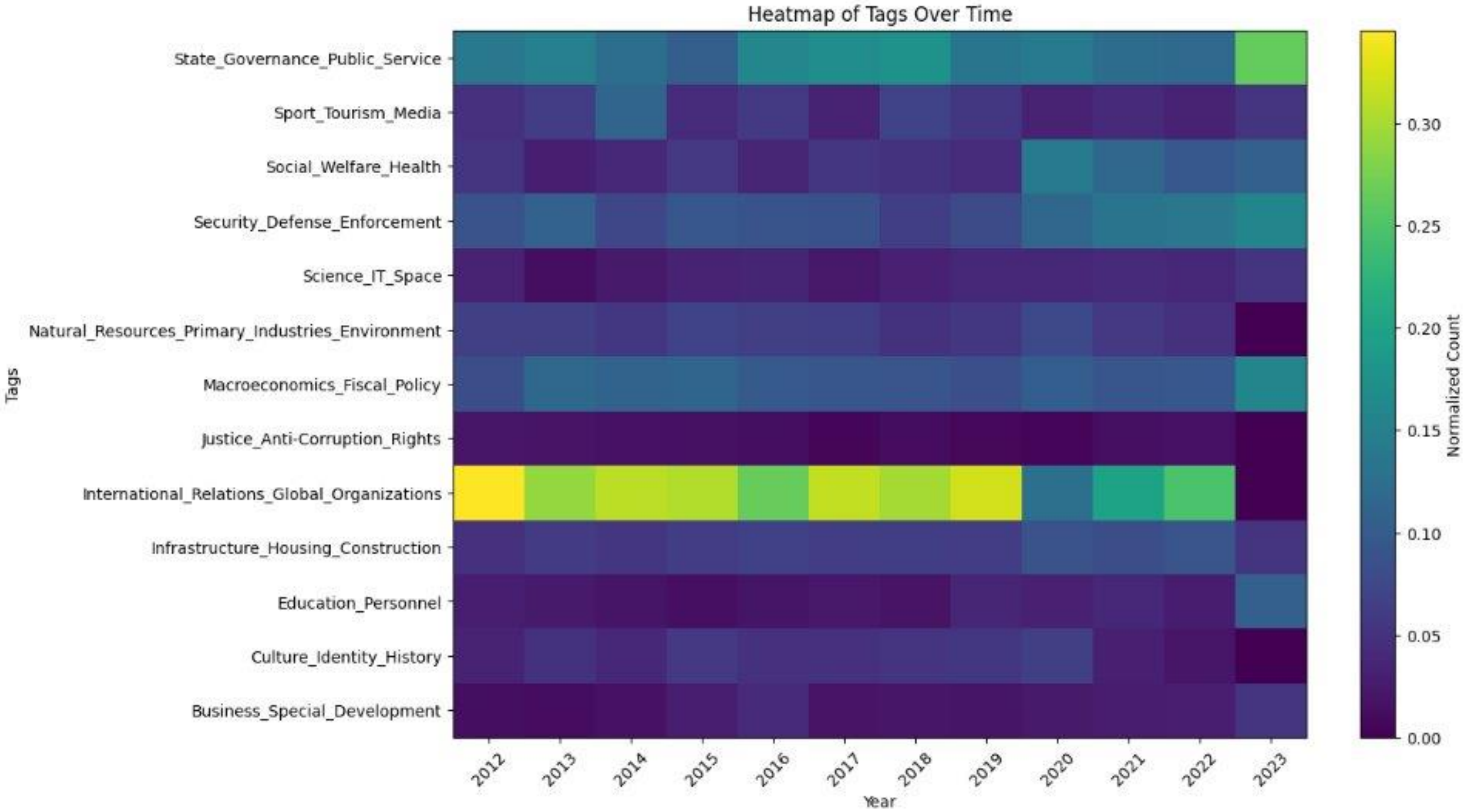
```
# --- Group 2: Macroeconomics & Fiscal Policy ---
'Economy and finance': 'Macroeconomics_Fiscal_Policy',
'Budget': 'Macroeconomics_Fiscal_Policy',
'Banks': 'Macroeconomics_Fiscal_Policy',
'Customs': 'Macroeconomics_Fiscal_Policy',
'Taxes': 'Macroeconomics_Fiscal_Policy',
'Investment': 'Macroeconomics_Fiscal_Policy',
'Import replacement': 'Macroeconomics_Fiscal_Policy',
'Anti-sanctions': 'Macroeconomics_Fiscal_Policy',
'Inflation': 'Macroeconomics_Fiscal_Policy',
'Labour market': 'Macroeconomics_Fiscal_Policy',
```

Dataset - EDA

Putin became president second time on 07.05.2012



Dataset - EDA



Heatmap of Tags Over Time
Tags
State_Governance_Public_Service
Sport_Tourism_Media
Social_Welfare_Health
Security_Defense_Enforcement
Science_IT_Space
Natural_Resources_Primary_Industries_Environment
Macroeconomics_Fiscal_Policy
Justice_Anti-Corruption_Rights
International_Relations_Global_Organizations
Infrastructure_Housing_Construction
Education_Personnel
Culture_Identity_History
Business_Special_Development
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
Normalized Count
0.30
0.25
0.20
0.15
0.10
0.05
0.00

Lexical Statistics – SOTA

Tokenization

- First task – tokenization of the text (splitting it into words or subwords):
  - Classical
    - Rule-based – **spaCy**
    - Regex-driven - **WordPunctTokenizer, TreebankWordTokenizer (NLTK)**
  - Modern
    - **BPE (GPT-2),**
    - **WordPiece (BERT),**
    - **Unigram,**
    - **SentencePiece**
    - Are they reliable?
      - E.g. BPE may create „Ukraine” as a 1 token in one model and in another it takes 3 different tokens to create „Ukraine”.

Key Citations: Schmidt et al. (2024)

Lexical Statistics – SOTA

Token Statistics

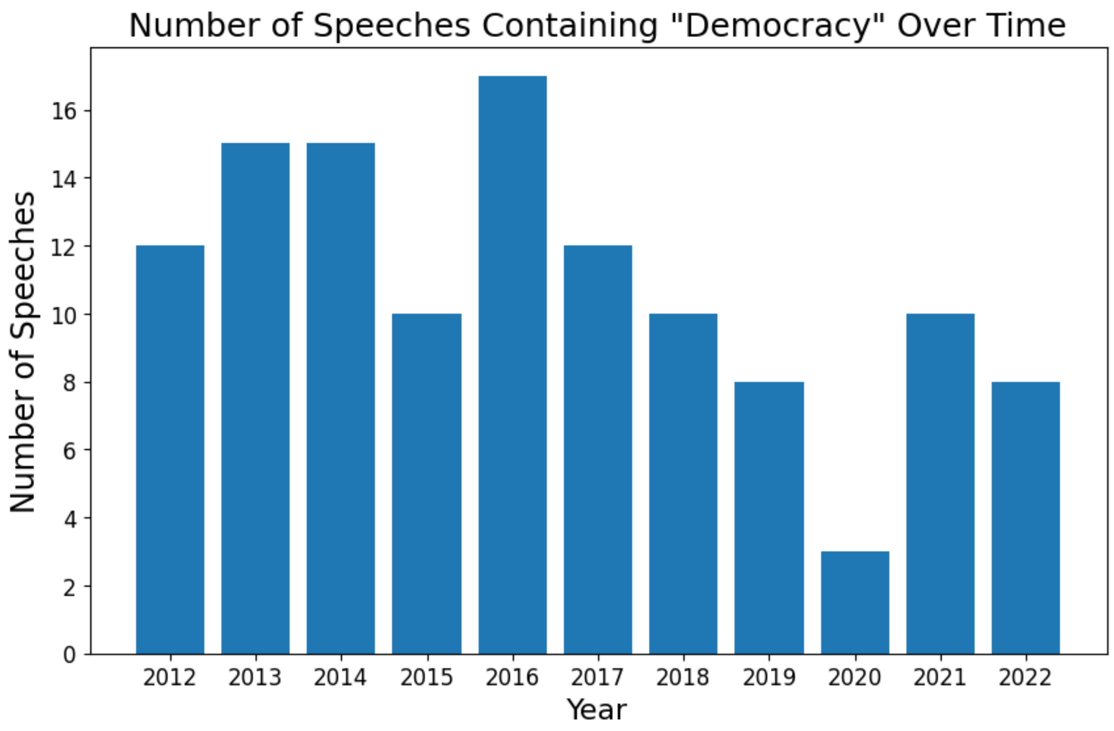
- Core statistical units:
  - Term Frequency - count of a term in a document
  - TF-IDF - highlights informative vocabulary across documents
- Classical tools:
  - Analysis using: pandas, NLTK, spaCy, CountVectorizer -> reliable
- Modern:
  - Queries to LLM’s.
  - Very flexible
    (no need to fine tuning)

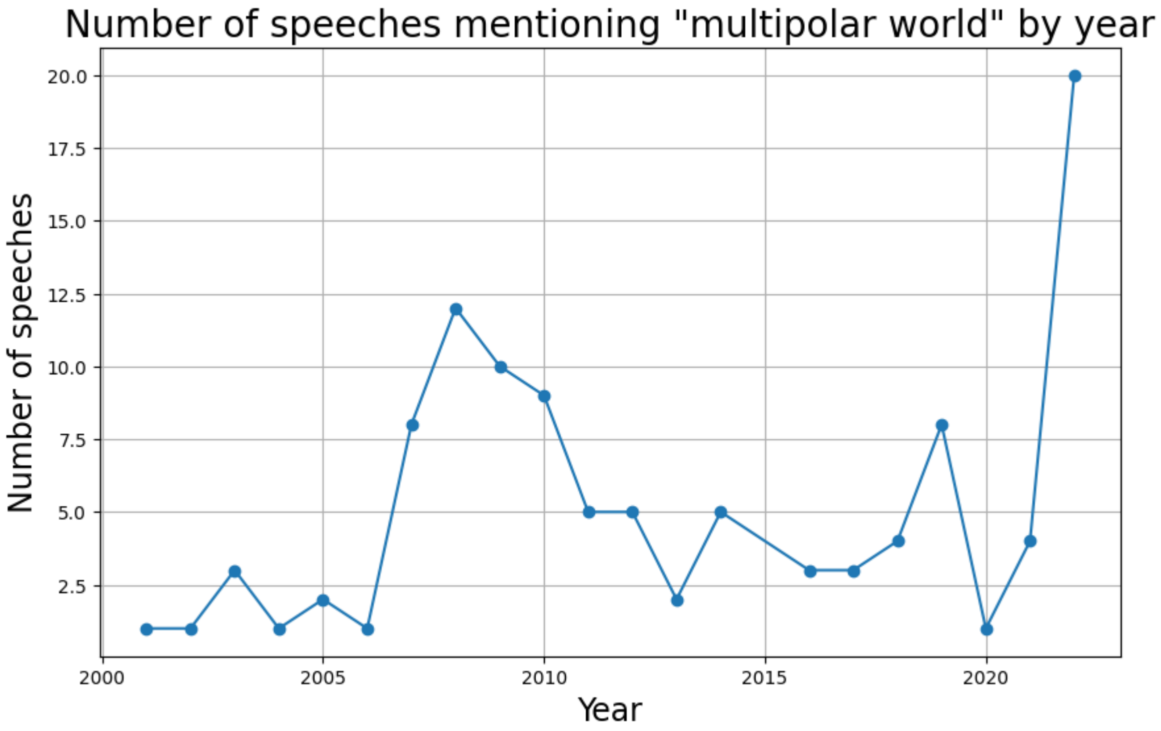
Why LLM’s are not well suited for this task?

- They don’t “count” - they predict plausible numbers
- This results in:
  - hallucinations
  - nondeterminism (no guaranteed reproducibility)

Lexical Statistics - POC

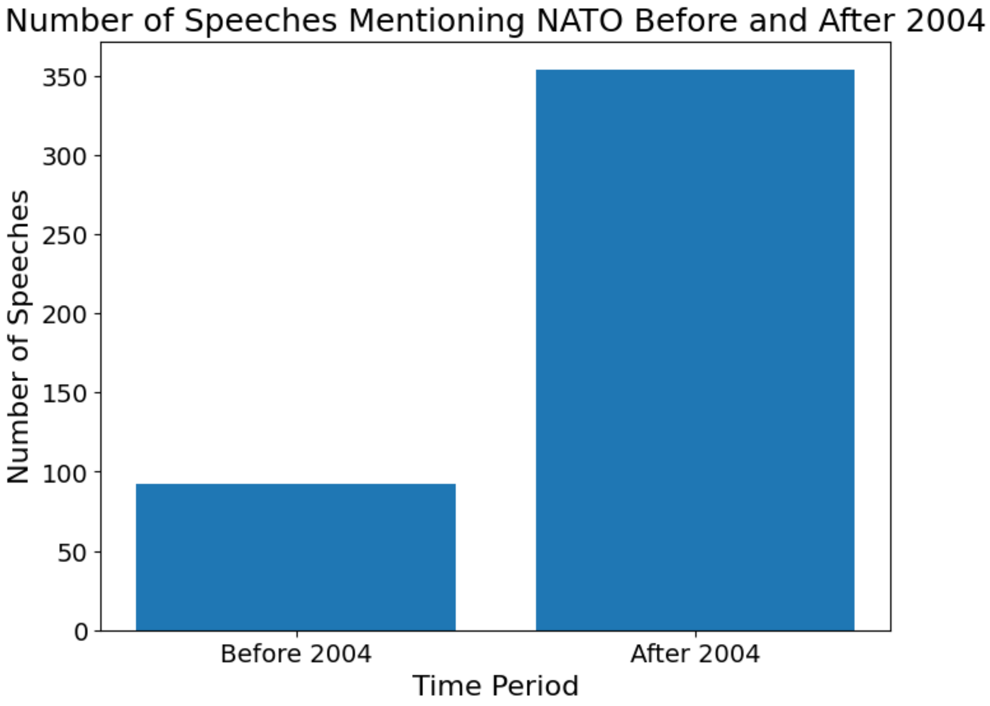
- In how many speeches does the word “democracy” appear? (120)
- In which years did Putin most often mention …?
- At what point does Putin start talking about a “multipolar world”?

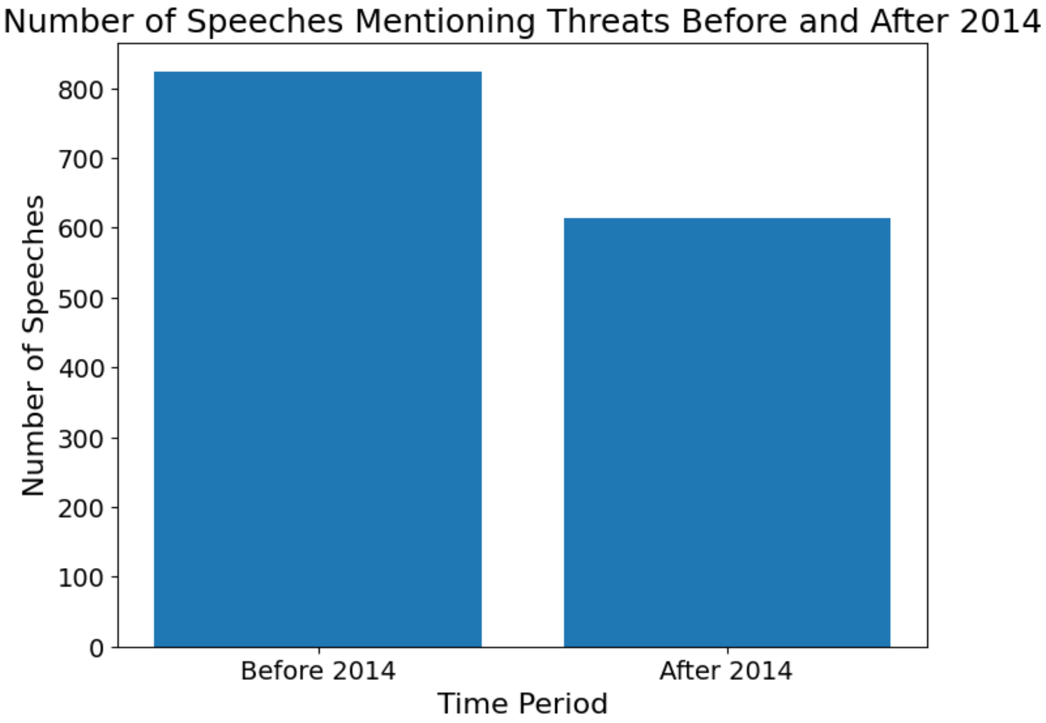




Lexical Statistics - POC

- How often does he mention NATO - before and after 2004?*
- How often does he speak about “threats” before and after 2014?*





*Here dataset **is not** restricted to speeches from 2012.

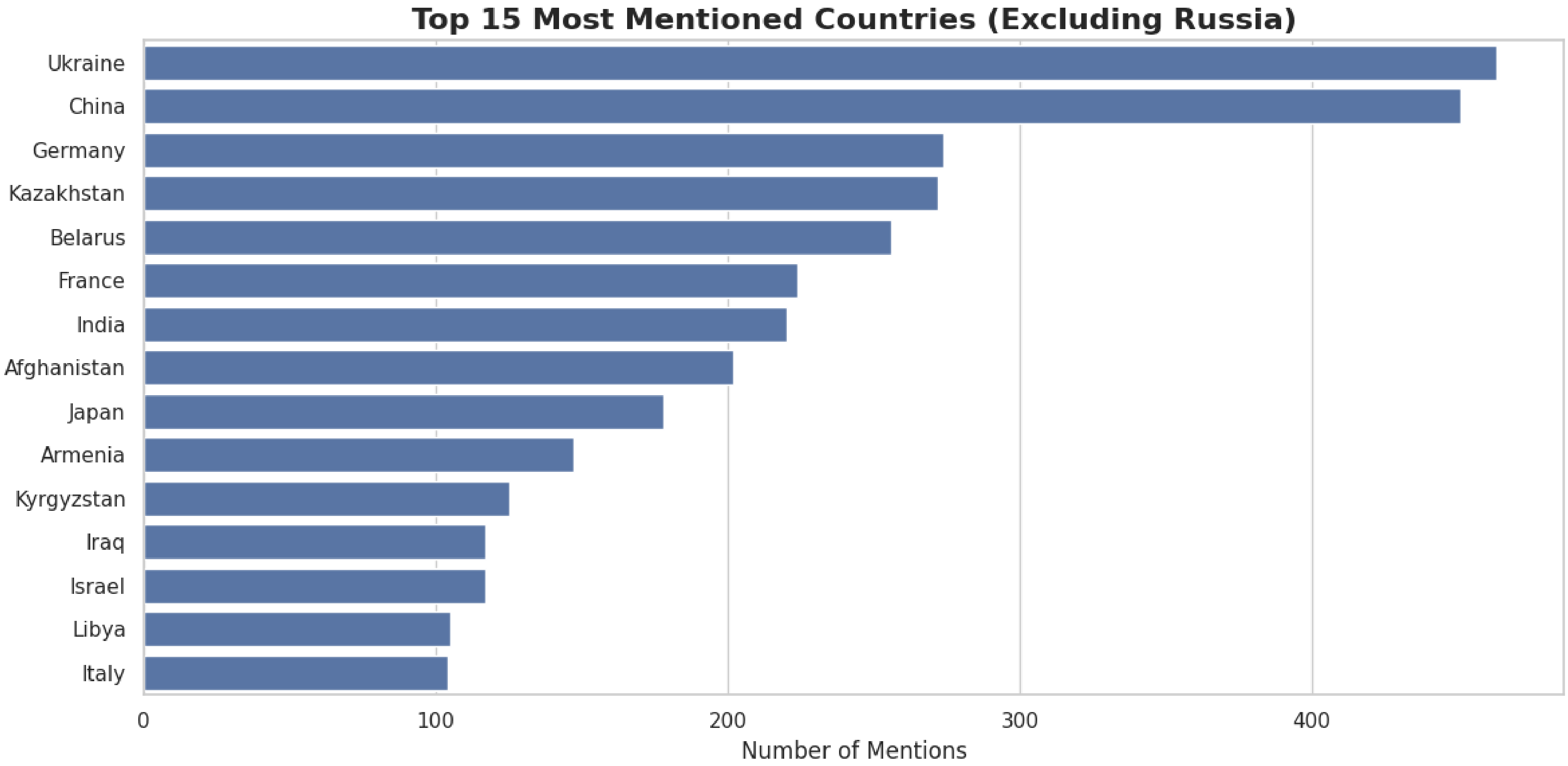
Named Entity Recognition - SOTA

Methodology	Specific Tools/Models	Pros	Cons
Transformer Encoders	**dslim/bert-base-NER**, spaCy (en_core_web_trf)	High precision, structured output	Fixed set of entities (LOC, PER, ORG etc.)
Neural Entity Linking	BLINK, REL	Resolves ambiguity (e.g. Paris, TX vs. Paris, France)	Requires more memory
Heuristic Linking	**Dictionary mapping**	Zero latency, fully customizable	Requires manual curation

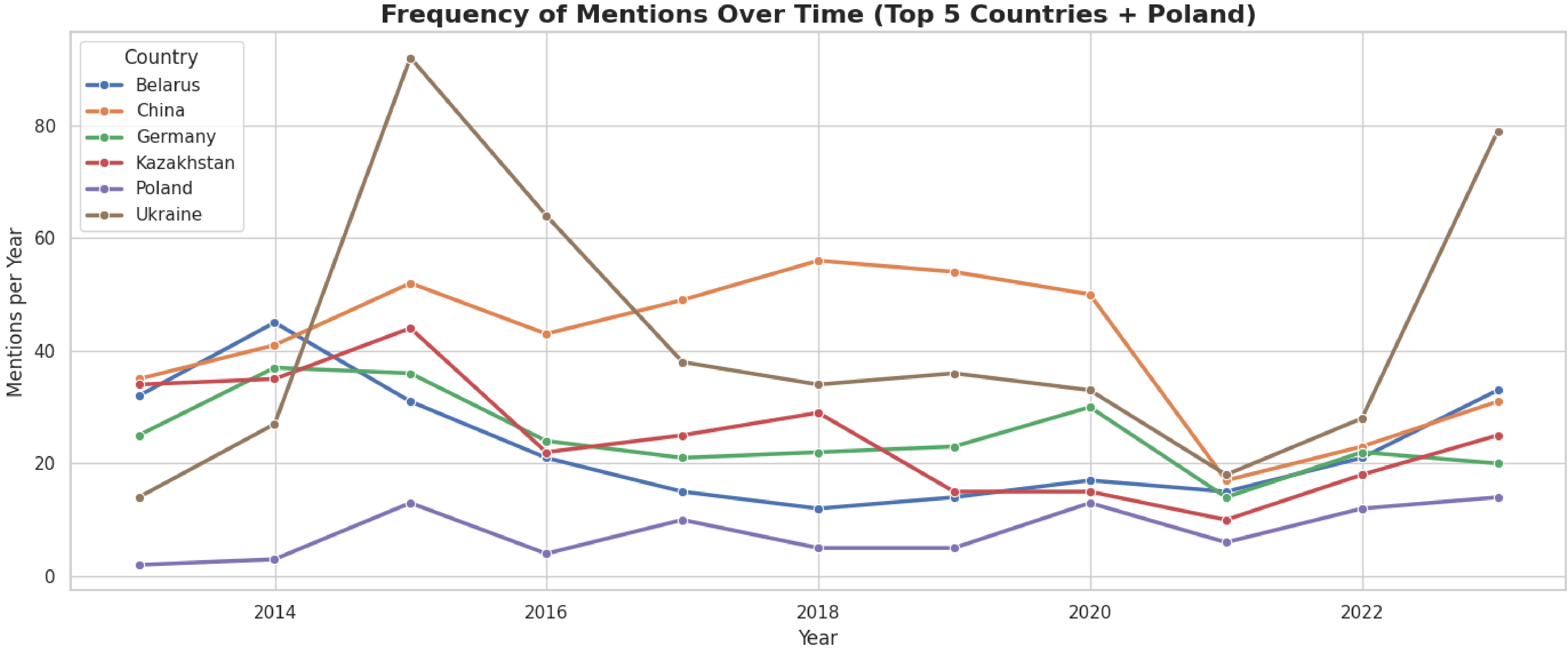
Key Citations: Devlin et al. (2019), Wu et al. (2020), van Hulst et al. (2020)

How many times do the words “Poland”, “Ukraine”, … appear in the entire database?

Which country, apart from Russia, appears most frequently in his speeches?



In which years did Putin most often mention …?

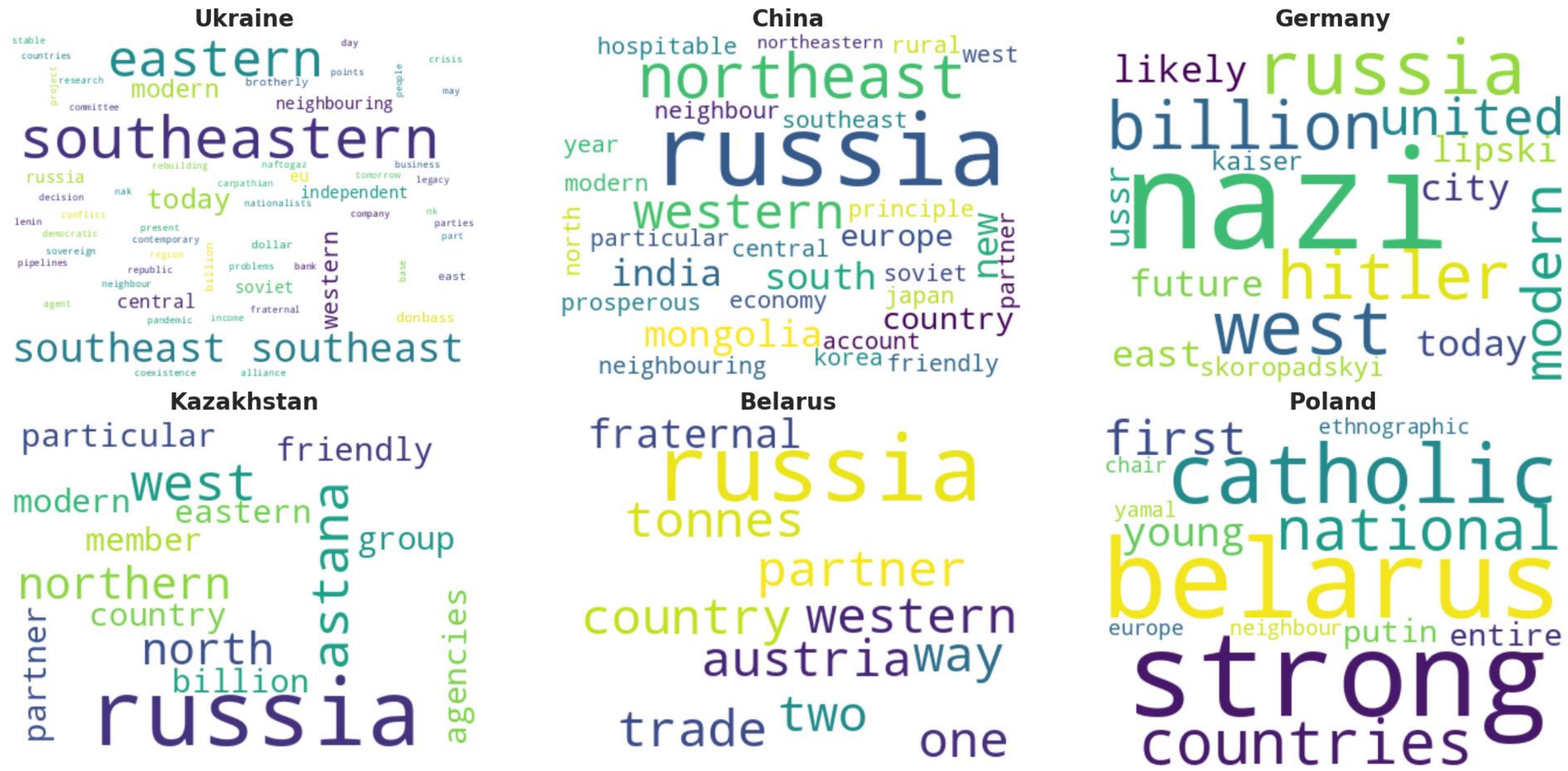


Semantic Framing & Context - SOTA

Methodology	Specific Tools/Models	Pros	Cons
Statistical Association	**Dependency parsing (spaCy)**, PMI	Grammatically precise, identifies direct modifiers	Surface-level only, misses sarcasm & manipulation
Zero-Shot Classification	**facebook/bart-large-mnli**, MoritzLaurer/DeBERTa-v3	Captures implied moral stance, no training needed, good accuracy	Slower inference, sensitive to wording of target words
Generative LLMs	GPT-4, LLaMA-3	State-of-the-art nuance, handles complex reasoning	High cost per document

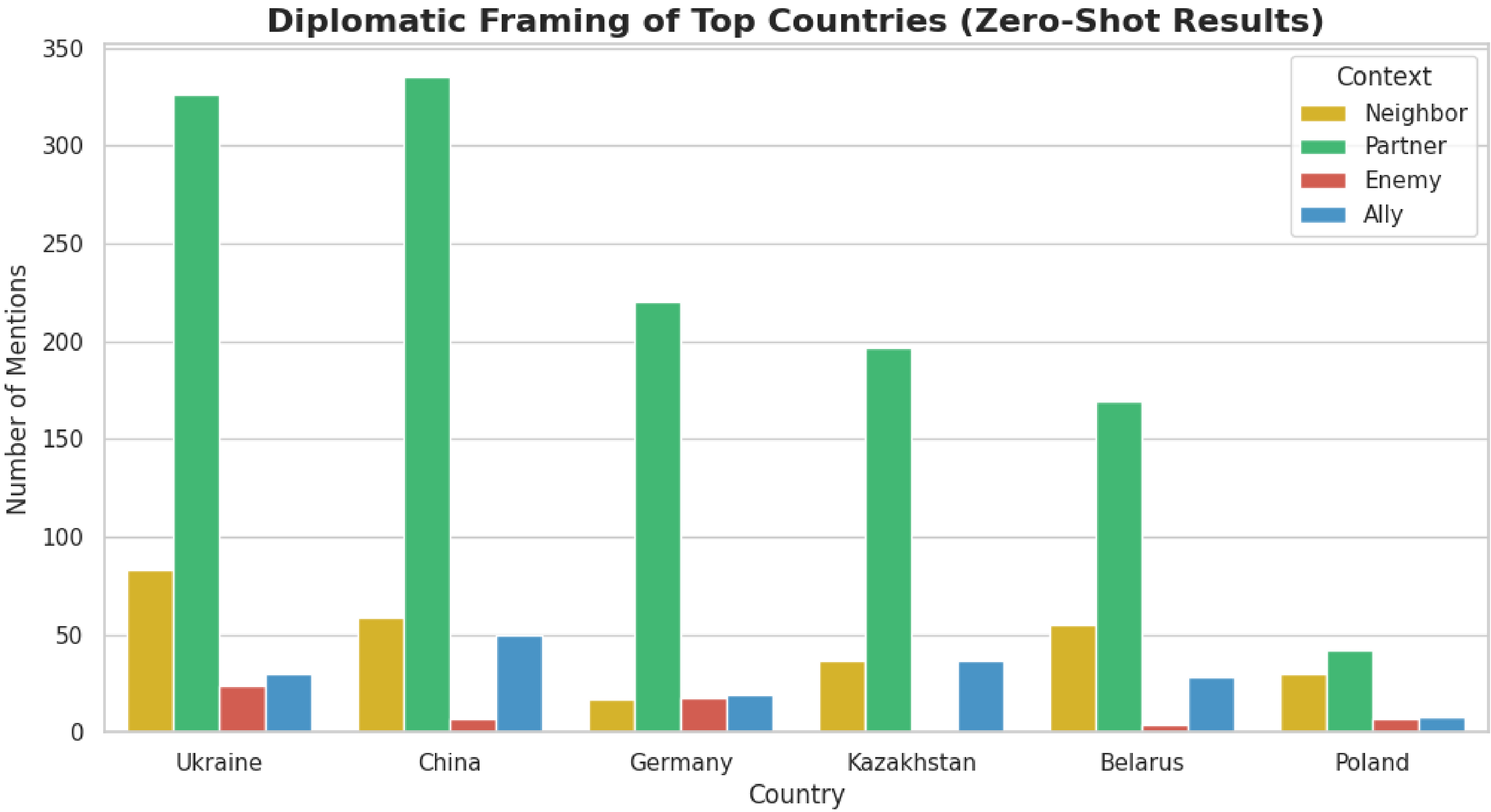
Key Citations: Qi et al. (2020), Honnibal et al. (2020), Lewis et al. (2020), Yin et al. (2019), Laurer et al. (2024), OpenAI (2023), Touvron et al. (2023)

What adjectives or terms most often accompany the word “Ukraine”?

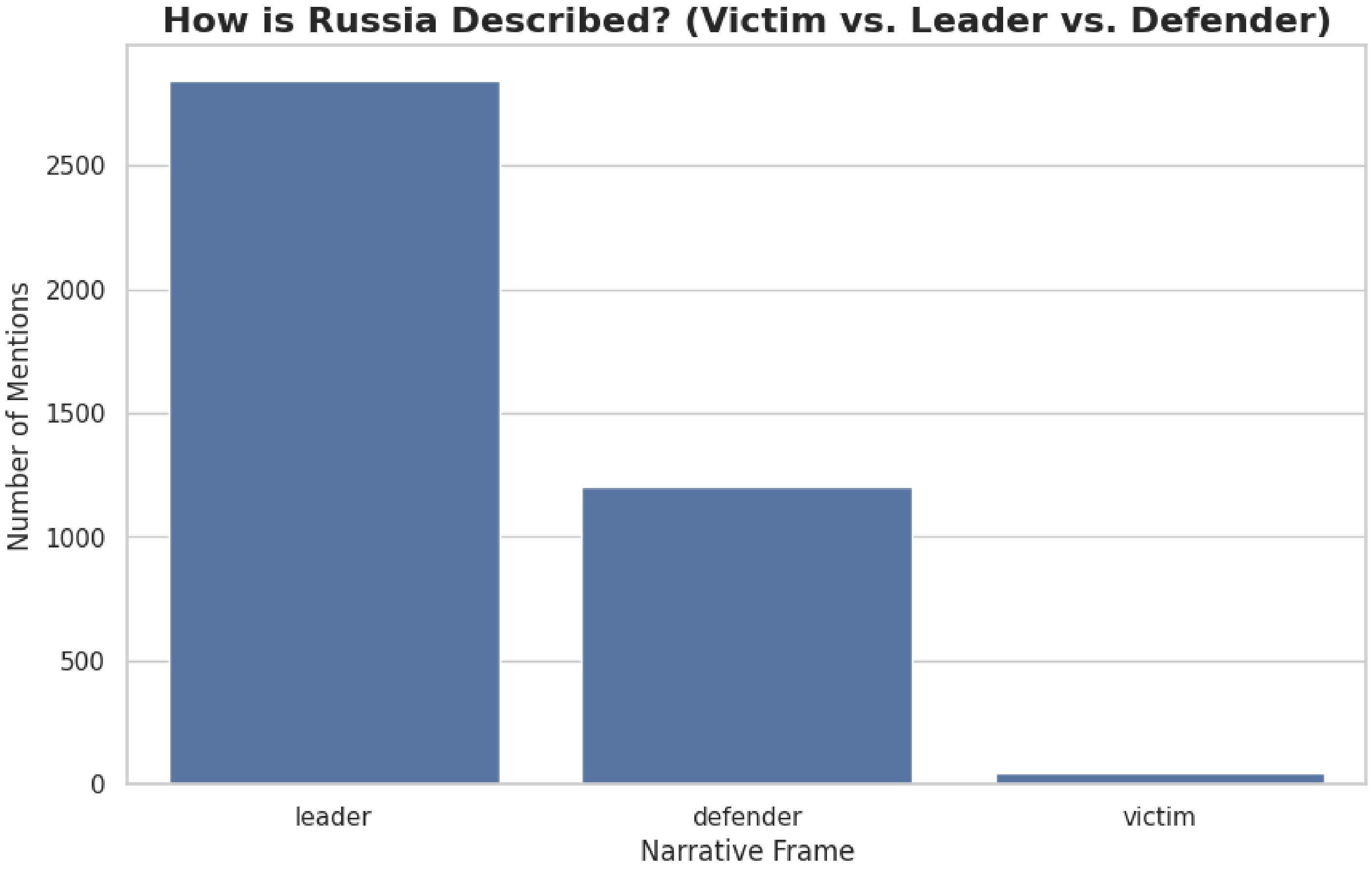


Contextual Word Clouds: Terms Accompanying Top Countries

In what context does “Poland” most often appear? (enemy, partner, ally, neighbour)



Is Russia more often described as a “victim”, a “leader”, or a “defender”?



What's next?

- Answering remaining questions
- Including all data for the analysis (also speeches before 2012)
- Conducting the analysis using bigger models
- Comparison between "old" methods and GenAI tools
- Topic modelling vs given tags

References

- [Schmidt et al. , 2024] Craig W. Schmidt et al. Tokenization Is More Than Compression.
- [Devlin et al., 2019] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.
- Honnibal et al., 2020] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.
- [Laurer et al., 2024] Laurer, M., van der Wal, W., Bonneau, F., & Gurevych, I. (2024). Less is More: Pre-train a Strong Text Encoder for Dense Retrieval Using a Weak Encoder. arXiv preprint.
- [Lewis et al., 2020] Lewis, M., et al. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL.
- [Qi et al., 2020] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. ACL.
- [van Hulst et al., 2020] van Hulst, J. M., Hasibi, F., Dercksen, K., Balog, K., & de Vries, A. P. (2020). REL: An Entity Linker Standing on the Shoulders of Giants. SIGIR.
- [Wu et al., 2020] Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Scalable Zero-shot Entity Linking with Dense Entity Retrieval. EMNLP.
- [Yin et al., 2019] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. EMNLP.