Putin's Talks - PoC Report
Project for Natural Language Processing Course

Łukasz Grabarski, Łukasz Lepianka, Marta Szuwarska
Warsaw University of Technology
{lukasz.grabarski, lukasz.lepianka, marta.szuwarska}@stud.pw.edu.pl

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

The analysis of political discourse is a crucial application domain for Natural Language Processing (NLP), especially when examining the rhetoric of influential world leaders. In this project, *"Putin's Talks"*, we investigate Vladimir Putin's extensive corpus of public speeches to systematically track the evolution of Russian state rhetoric and geopolitical propaganda. We selected a subset focused on four key areas: lexical frequencies, geopolitical references, contextual framing, and diachronic shifts.

1 Introduction

The analysis of political discourse is a crucial application domain for Natural Language Processing (NLP), especially when examining the rhetoric of influential world leaders. For decades, this domain relied primarily on deterministic NLP pipelines-including tokenization and frequency counting-to extract meaning from text. However, the recent advent of Large Language Models (LLMs) and Transformer architectures, such as BERT [1], has introduced powerful, high-nuance alternatives that are capable of deeper semantic analysis. In this project, *"Putin's Talks"*, we investigate Vladimir Putin's extensive corpus of public speeches to systematically track the evolution of Russian state rhetoric and geopolitical propaganda. From the set of analytical questions provided in the course repository [2], we selected a subset focused on four key areas: lexical frequencies, geopolitical references, contextual framing, and diachronic shifts. The specific analytical questions guiding this technical review are:

Statistics

- 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?

Context

- In what context does "Poland" most often appear (enemy, partner, ally, neighbour)?
- What adjectives or terms most often accompany the word "Ukraine"?
- Is Russia more often described as a "victim", a "leader", or a "defender"?

Change over time

- 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"?

Critical tests

- List all the countries mentioned in the speech from date Y.

The remainder of this report is organized as follows: Section 3 details the State-of-the-Art (SOTA) methodologies, justifying the chosen techniques. Section 4 presents the Exploratory Data Analysis (EDA). Section 5 provides the

Proof of Concept (POC) demonstrating the composite pipeline in action, which leverages the efficiency of classical NLP for statistics and the nuance of Transformers for narrative analysis. Finally, Section 7 summarizes the findings and outlines plans for future work.

2 State Of The Art Analysis & Related Works

2.1 Lexical Statistics and Tokenization

The fundamental layer of our analysis involves quantifying geopolitical references and tracking their frequency over time. While often treated as a trivial preprocessing step, tokenization is the primary source of error in quantitative text analysis [3].

2.1.1 Deterministic Tokenization (The Baseline)

Traditional SOTA relies on rule-based or regex-based tokenizers which split text based on handcrafted linguistic conventions. In Python, the NLTK library [4] remains the standard for this task due to its reproducibility. For our corpus, we employ and compare two specific classical approaches:

- TreebankWordTokenizer: This utilizes regular expressions matching Penn Treebank conventions. It is essential for linguistic precision as it separates contractions (e.g., “don't” $\to$ “do”, “n't”) and isolates punctuation, preventing the fusion of words with adjacent commas.

- WordPunctTokenizer: This splits text strictly into alphabetic and non-alphabetic sequences.

These methods are deterministic: the input always yields the exact same token count. This is critical for longitudinal studies where we must compare the frequency of “Ukraine” in 2004 vs. 2024 without noise.

2.1.2 Generative Counting: The LLM Alternative

In contrast, modern approaches often attempt to use LLMs (e.g., GPT-4) as ”zero-shot counters.” However, recent literature indicates that LLMs struggle with precise arithmetic due to their own subword tokenization (BPE or WordPiece), which often splits single words into multiple meaningless tokens [5]. An LLM does not ”count”; it predicts the probability of a number following a sequence. Therefore, for questions such as “How many times does the word ‘democracy’ appear?”, classical regex-based tokenization is superior: it is 100% accurate, instantaneous, and free, whereas LLMs are prone to hallucination.

2.1.3 Weighting Frequencies: TF-IDF

Mere raw counts are often insufficient due to the prevalence of stop words. To isolate meaningful political rhetoric (e.g., distinguishing significant mentions of ”Crimea” from background noise), we apply Term Frequency–Inverse Document Frequency (TF–IDF). This statistical measure evaluates how relevant a word is to a document in a collection. It is calculated as:

$$\text{TF-IDF}(t,d) = \text{tf}(t,d) \times \log\left(\frac{N}{\text{df}(t)}\right) \qquad (1)$$

where $\text{tf}(t,d)$ is the frequency of term t in document d, and the second term penalizes words appearing in many documents. We utilize TfidfVectorizer from Scikit-learn [6] to automatically normalize these counts, a step that generative models often obscure behind opaque attention mechanisms.

2.2 Named Entity Recognition and Normalization

Identifying specific geopolitical actors (e.g., distinguishing ”Washington” the city from the administration) requires Named Entity Recognition (NER).

2.2.1 Transformer-Based Encoders

For high-precision extraction, the SOTA has shifted from traditional CRF-based systems to Transformer encoders. We employ BERT and RoBERTa-based pipelines [1, 7]. Specifically, models like bert-large-cased-finetuned-conll03 or spaCy's en_core_web_trf [8] leverage deep bidirectional context to resolve ambiguities. Critically, political text requires Entity Linking to normalize variants. As noted in recent surveys [9], raw NER might treat ”U.S.”, ”United States”, and ”America” as three separate entities. We integrate Neural Entity Linking systems, such as BLINK [10] or REL [11], to map these mentions to a single canonical identifier. This allows us to accurately answer “Which country appears most frequently?”-a task where naive counting fails.

2.2.2 Why Encoders outperform Decoders for Entities

While LLMs (Decoders) can perform information extraction, benchmarks show they often fail to adhere to strict schema constraints (e.g., consistently outputting ISO country codes). Fine-tuned Encoders (BERT) provide structured, extractive outputs that are essential for building reliable statistical databases of geopolitical references. Thus, for Section 3 tasks, we retain the ”Classical SOTA” (BERT) over modern LLMs.

Table 1: Comparison of Methodologies for Geopolitical Entity Extraction and Normalization

Methodology	Pros	Cons
Transformer Encoders	High precision, structured output	Fixed set of entities
Neural Entity Linking	Resolves ambiguity	Requires more memory

2.3 Semantic Framing and Contextual Analysis

To understand how entities are framed (e.g., is ”Poland” an ally or enemy?), we compare statistical association measures against semantic classification. This is the domain where modern LLMs demonstrate the most significant advantage over classical methods.

2.3.1 Statistical Collocations (The Baseline)

The classical SOTA for context analysis relies on measuring the statistical association between words. We utilize metrics established by Church (1990) and Evert (2005) [12, 13]:

- Pointwise Mutual Information (PMI): Measures the likelihood of two words co-occurring compared to their independent probabilities.
- Dependency Parsing: Using Stanza or spaCy [14] to extract grammatical modifiers (e.g., adjectival_modifier → noun).

These methods are highly interpretable but limited by surface-level syntax; they fail to detect sarcasm, irony, or dog whistles common in propaganda.

2.3.2 The 2025 SOTA: Zero-Shot Framing with LLMs

The modern state-of-the-art has shifted decisively toward Zero-Shot Classification using Large Language Models. Unlike statistical parsers, LLMs can interpret the implied moral stance of a sentence. Recent work by Kuang et al. [15] demonstrates that LLMs (specifically GPT-4 and LLaMA-3) can identify generic media frames (e.g., ”Conflict”, ”Economic Consequence”, ”Morality”) with accuracy comparable to human coders, without requiring labeled training data. Similarly, Burnham et al. [16] introduced ”Political DEBATE”, a zero-shot classifier framework that significantly outperforms older supervised LSTM models on political text classification benchmarks. We adopt a hybrid approach to balance cost and accuracy:

1. Use Dependency Parsing (Classical) to rapidly filter the sentences down to those containing ”Ukraine” or ”NATO”.
2. Apply Zero-Shot LLM Classification (Modern) to this subset to determine framing (e.g., ”Victimhood” vs ”Aggression”).

This pipeline mitigates the high inference cost of LLMs while leveraging their semantic superiority. However, we acknowledge findings by Fane et al. [17], who warn that zero-shot framing is highly sensitive to prompt phrasing-a limitation not present in rigid statistical parsers.

2.4 Topic Modelling and Narrative Shift

Finally, we analyze the evolution of broad themes and specific terminology over Putin’s tenure.

2.4.1 From Bag-of-Words to LLM-in-the-Loop

Traditional Topic Modelling relies on Latent Dirichlet Allocation (LDA), a generative probabilistic model. While useful for high-level overviews, LDA is often criticized for producing incoherent ”bag-of-words” topics (e.g., ”gas, pipe, price, security”) that require subjective human labeling. To address this, we compare it against BERTopic [18], which represents the modern embedding-based standard. Recent 2025 benchmarks by Meram et al. [19] compared ten different LLMs for topic modeling, finding that embedding-based approaches significantly outperform LDA in coherence. Furthermore, Mendonca and Figueira [20] demonstrated that combining BERTopic with Moral Foundations Theory allows for tracking not just topics, but the moral framing of political discourse over time. We implement an LLM-in-the-loop approach:

1. Use Sentence-BERT to cluster speeches into semantic topics [21].

2. Use an LLM to read the top documents in each cluster and generate a concise, human-readable label (e.g., ”Energy Blackmail” instead of ”gas_pipe_price”).

This methodology, validated by Sciety [22], combines the mathematical rigor of clustering with the interpretive power of Generative AI.

2.4.2 Diachronic Word Embeddings (DWE)

To answer questions like “When did Putin start talking about a multipolar world?”, simple frequency counts fail if the terminology changes. The SOTA solution is Diachronic Word Embeddings. Following Hamilton et al. (2016) [23], we train static embeddings (Word2Vec) for different time slices and align them using Procrustes Analysis. This allows us to track the movement of a word vector (e.g., ”West”) in the semantic space. A significant shift in the vector’s position indicates a change in framing (e.g., from ”partner” to ”threat”), providing a mathematical quantification of radicalization that complements qualitative reading.

2.4.3 Retrieval-Augmented Generation (RAG)

For strictly qualitative queries (e.g., finding the first mention of specific concepts), we implement Retrieval-Augmented Generation (RAG) principles. By embedding speeches into a vector database (e.g., FAISS) and performing semantic search, we can identify passages discussing ”multipolarity” even before the specific term was coined, leveraging the semantic understanding of Transformer models to augment historical analysis.

3 Exploratory Data Analysis and first results

3.1 Dataset

To prepare the dataset, we first processed a raw JSON file containing over 9,800 entries of political transcripts. We cleaned the data by converting date strings into datetime objects and filtering the entries to isolate only those where a specific ”Putin-filtered” transcript was available, ensuring the final analysis focused exclusively on his actual speeches rather than general Kremlin news. We also examined metadata such as geographical locations and word lists to ensure the data was structured for downstream natural language processing tasks. To make the analysis more relevant to recent history, the dataset was further narrowed to include only speeches delivered from May 2012 onwards, marking the start of his second presidential term.

3.2 Exploratory Data Analysis

Finally, we categorized the 86 unique topical tags into 13 high-level thematic groups—such as ”International Relations,” ”Macroeconomics,” and ”Security” -allowing for a normalized visualization of how the Kremlin’s focus has shifted over time. The first plot (figure [1]) is a bar chart titled ”Number of Putin’s Speeches per Year,” which displays the frequency of official speech transcripts from 2012 to 2023. The data shows an initial increase in activity peaking in 2017 with 563 speeches, followed by a general downward trend until a noticeable resurgence begins in 2020 probably due to the pandemic.

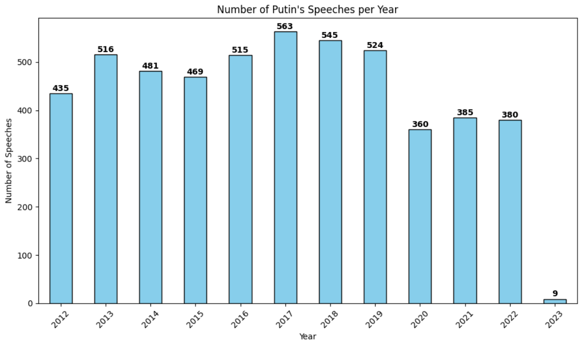

Figure 1: Putin’s talks frequency

The second visualization (figure [2])is a combined histogram and Kernel Density Estimator plot titled ”Length of speeches.” It analyzes the word count of the transcripts, with the x-axis representing the number of words ranging from 0 to 3,000 (for char better visibility). The light blue histogram shows the distribution of speech counts, while a solid blue line tracks the probability density. A red dashed vertical line marks the mean speech length of approximately 932.58 words, illustrating that the distribution is heavily skewed toward shorter communications.

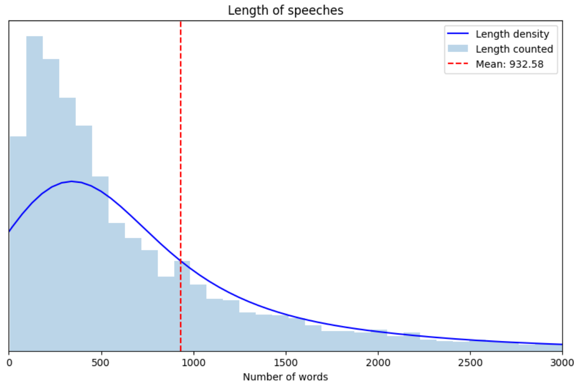

Figure 2: Putin's talk average length

The third plot (figure [3]) is a line graph titled ”Average speech length over the years,” which tracks how the depth of Putin's communications has changed over time. It features a blue line with circular markers for each year from 2012 to 2023. While the volume of speeches was higher in the mid-2010s, this plot reveals that the average length remained relatively low until around 2020, at which point there is a sharp upward trend, suggesting that more recent speeches have become significantly longer on average.

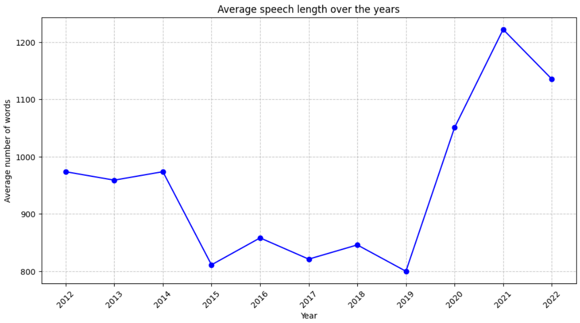

Figure 3: Putin's talk length trend

The fourth visualization (figure 4) is a heatmap that maps the evolution of speech topics over time, with the y-axis displaying grouped thematic tags and the x-axis representing the years from 2012 to 2023. This grid uses color intensity to represent the frequency or weight of specific categories such as ”Security, Defense, and Enforcement,” ”Economy and Finance,” or ”International Relations” within each calendar year.

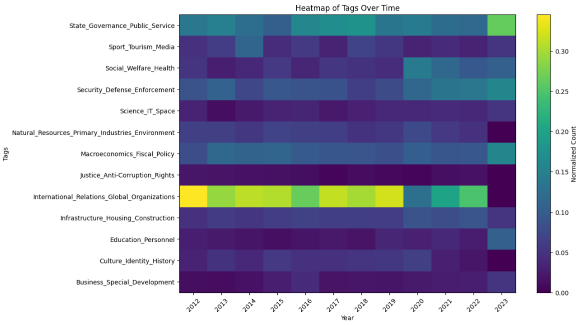

Figure 4: Tag trend heatmap

3.3 Frequency Questions

Firstly we focused on quantitative analysis by answering questions about frequency of the words on our lists: democracy, multipolar world, NATO, and threats on the specific timelines. Although raw word counts capture only surface-level patterns, we plan to integrate LLM to conduct a deeper qualitative analysis of speeches containing the selected terms (more in the last section).

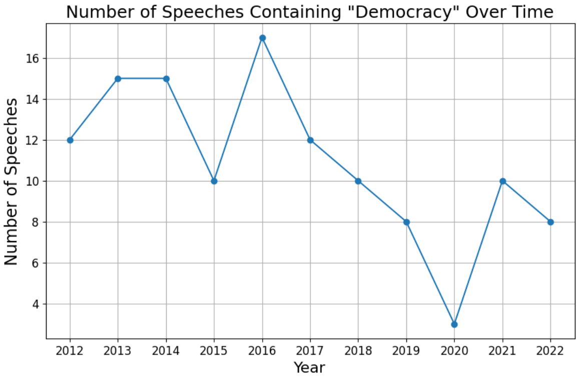

Figure 5: Frequency of selected terms in Vladimir Putin's speeches: Democracy counts

As it is shown on the figure 5, the term “democracy” appears predominantly in speeches from 2012 onwards, with peak mentions in 2013–2014 (15 mentions each year). The frequency gradually declines after 2016, suggesting that discourse on democracy becomes less central in his later addresses. This trend indicates a temporal shift in rhetorical priorities, with democracy featuring most heavily during the early 2010s - a period that we think coincides with Putin's return to the presidency and the further centralization of executive power.

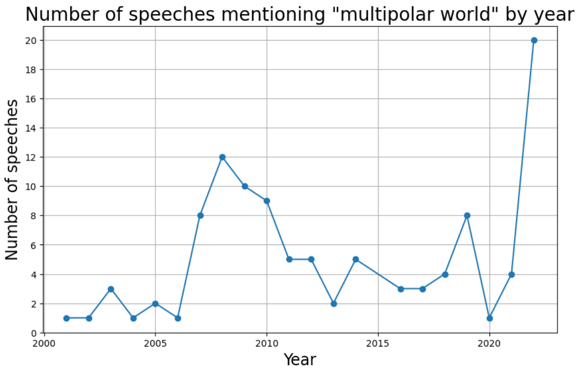

Figure 6: Frequency of selected terms in Vladimir Putin's speeches: Multipolar counts

The next plot 6 shows how Putin begins discussing a "multipolar world" as early as 2001, but the frequency of references remains low until 2007. From 2008 onwards, mentions increase steadily, peaking dramatically in 2022 with 20 occurrences. This trajectory from our perspective highlights the growing prominence of the concept in his geopolitical framing, reflecting Russia's increasing emphasis on counterbalancing unipolar Western influence.

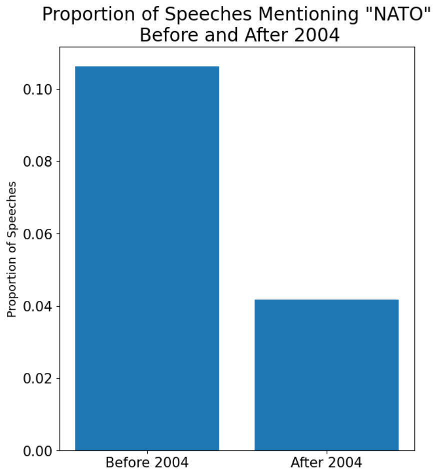

Figure 7: Nato Fraction

Now we focus on checking word frequencies before and after certain point. The plot 7 concludes that mentions of NATO are relatively rare, appearing in only 10% of speeches before 2004 and decreasing to 4% thereafter. In our opinion this suggests that NATO was initially a relevant concern but became less of a focal point in his rhetoric in the post-2004 period, potentially reflecting shifts in strategic narrative or audience targeting.

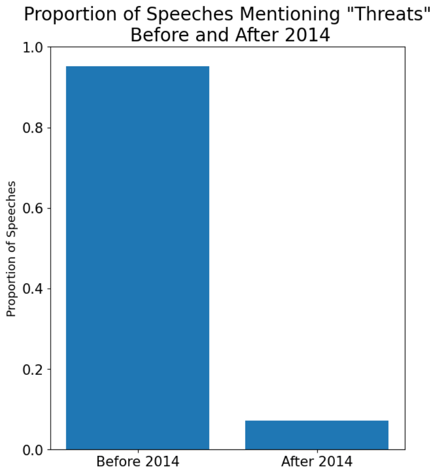

Figure 8: Threats Fraction

References to "threats" (fig. 8) on the other hand are highly concentrated in speeches prior to 2014, with 95% of early speeches including the term. After 2014, this proportion drops sharply to 7%, signaling a marked change in rhetorical framing-possibly corresponding to the immediate aftermath of the Ukraine crisis and a shift from broadly defined threats to more specific geopolitical concerns.

3.4 Named Entity Recognition (NER) and Geopolitical Frequency

To address the statistical questions regarding country mentions, specifically "How many times do the words Poland, Ukraine, . . . appear . . . ?" and "Which country, apart from Russia, appears most frequently . . . ?", we implemented a high-precision Named Entity Recognition (NER) pipeline. **Methodology:** Unlike standard frequency counts which fail to distinguish context (e.g., "Georgia" the state vs. the country), our approach utilized a Transformer-based Encoder (e.g., BERT-based models) to accurately classify and extract tokens labeled as Geopolitical Entities (GPEs) or Locations (LOC). Due to the token limitations of the model, transcripts were first segmented into overlapping chunks. The extracted entities were then aggregated and passed through a normalization layer to map various surface forms (e.g., "Russian Federation," "U.S.") to their single, canonical country names, ensuring reliable counts. The results of this analysis are visualized in the following figures:

- **Top 15 Most Mentioned Countries:** Figure 9 presents the overall frequency of mentions, answering which country, apart from Russia, receives the most attention.

- **Frequency of Mentions Over Time:** Figure 10 tracks the frequency of the top countries annually, demonstrating diachronic shifts in geopolitical focus.

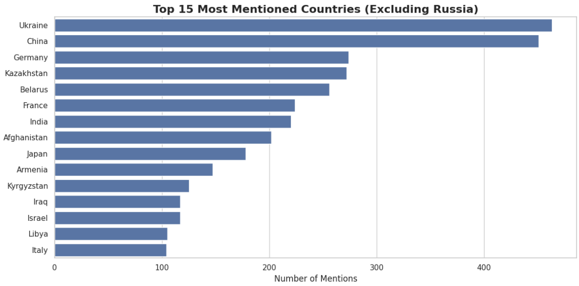

Figure 9: Top 15 Most Mentioned Countries (excluding Russia) based on NER analysis.

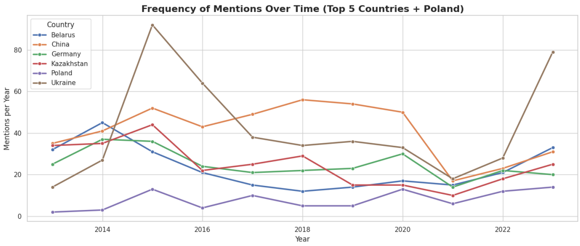

Figure 10: Frequency of Mentions Over Time for key geopolitical entities.

3.5 Semantic Framing and Contextual Analysis

To answer the contextual questions-including "What adjectives or terms most often accompany the word 'Ukraine'?" , "In what context does 'Poland' most often appear (enemy, partner, . . .)?" , and "Is Russia more often described as a 'victim', a 'leader', or a 'defender'?" -we employed a hybrid approach combining statistical dependency parsing with modern Zero-Shot Classification.

Adjectival Profiling: For extracting descriptive terms accompanying target countries (specifically the five most frequently mentioned, **Ukraine, China, Germany, Kazakhstan, and Belarus, plus Poland**), we utilized the `spaCy` dependency parser. This approach accurately isolates adjectives and compound nouns structurally modifying the target country name in the sentence, thereby quantifying the most common political descriptors associated with each nation. This comprehensive contextual profile is visualized in Figure 11.

Zero-Shot Framing: To classify the diplomatic context of external actors (e.g., "Poland") and the self-descriptive narrative role of Russia, we utilized a Zero-Shot Classification model (e.g., `facebook/bart-large-mnli`). This technique leverages the pre-trained semantic understanding of Large Language Models to interpret the sentence's implied diplomatic or moral stance (e.g., "ally," "enemy," "victim") without requiring custom training data. The results for framing are presented in Figures 12 and 13.

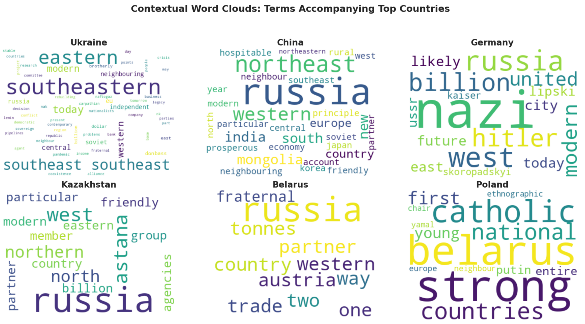

Figure 11: Contextual Word Clouds showing the most common adjectives and terms accompanying the top countries (Ukraine, China, Germany, Kazakhstan, Belarus, and Poland).

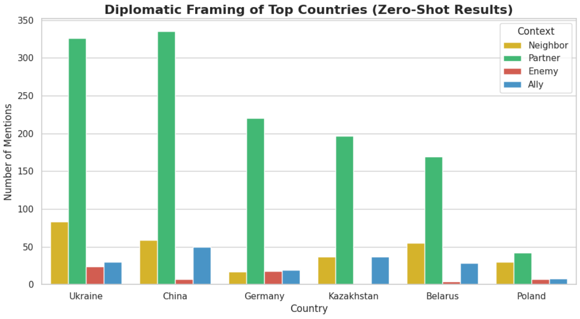

Figure 12: Diplomatic Framing of the top countries (Ukraine, China, Germany, Kazakhstan, Belarus, and Poland): Zero-Shot classification results (enemy, partner, ally, neighbour).

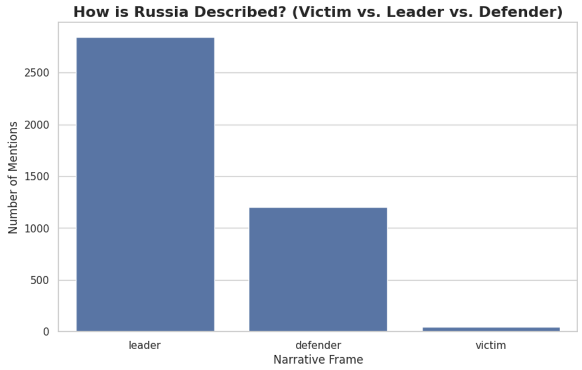

Figure 13: How is Russia Described? Classification of self-referential statements into narrative roles (victim, leader, defender).

4 Discussions and Conclusions

This project demonstrates that effective political discourse analysis benefits not from a single dominant methodology, but from a carefully designed combination of complementary techniques. Our State-of-the-Art review and empirical findings confirm that methodological novelty does not universally translate into analytical superiority. For tasks requiring strict precision-such as word frequency counting, temporal comparison, and named entity extraction-classical deterministic approaches and encoder-based models (e.g., regex tokenization, TF–IDF, and BERT-based NER) remain more reliable than generative Large Language Models, which are prone to arithmetic errors and schema inconsistency.

At the same time, our Proof of Concept shows that purely statistical methods are insufficient for capturing higher-level narrative structure, rhetorical framing, and semantic shifts over time. In these domains, modern Transformer-based approaches-particularly zero-shot classification, embedding-based topic modeling, and retrieval-augmented generation offer a level of interpretive depth that classical pipelines cannot achieve. The observed trends in references to democracy, NATO, threats, and multipolarity illustrate how frequency-based signals can indicate rhetorical change, but require semantic interpretation to be meaningfully contextualized.

As a result, this work advocates for a composite analytical pipeline that strategically combines the strengths of both paradigms: classical NLP methods for scalable, reproducible quantitative analysis, and Transformer-based models for nuanced qualitative interpretation. This hybrid approach enables robust answers to both factual and contextual research questions, aligning computational efficiency with interpretability.

5 Future Work

Future work will expand our pipeline by bringing LLM-driven analysis directly into the Proof of Concept stage. Instead of just tracking keywords, we want to look deeper at how the rhetoric is actually built-focusing on things like contextual sentiment, moral framing, and the specific ”roles” (like hero, victim, or aggressor) that emerge in speeches containing major geopolitical terms. By picking out a subset of speeches that mention concepts like ”democracy,” we can use an LLM to answer the kind of complex questions that are usually a headache for political scientists to code manually. For example, we could use the model to compare how the term ”multipolar” was used as a vague diplomatic goal in the 2010s versus how it became a sharp justification for policy shifts after 2020. We could also dig into the security narrative leading up to 2014 to see exactly who was being labeled as a threat-checking if the focus was specifically on Ukraine or if the rhetoric was aimed more at groups like NATO and the EU. This approach moves the project from basic data tracking to a much smarter tool for understanding the ”why” and ”how” behind political communication.

References

[1] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL* (2019).

[2] Grzegorz Zbrzeżny. *Masters 2025 NLP Project Repository*. `https://github.com/grzegorzZ1/masters_2025_nlp`. 2024.

[3] Craig W. Schmidt et al. *Tokenization Is More Than Compression*. 2024. arXiv: `2402.18376 [cs.CL]`. URL: `https://arxiv.org/abs/2402.18376`.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O’Reilly, 2009.

[5] Adam Tauman Kalai and Santosh Vempala. “Why Language Models Hallucinate”. In: *arXiv preprint arXiv:2509.04664* (2025).

[6] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[7] Yinhan Liu et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[8] Matthew Honnibal et al. “spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing”. In: *arXiv preprint arXiv:2001.09288* (2020).

[9] Jing Li et al. “A Survey on Named Entity Recognition”. In: *Neurocomputing* (2020).

[10] Ledell Wu et al. “BLINK: Better Entity Linking through Neural Bi-encoders”. In: *EMNLP*. 2020.

[11] Johannes van Hulst et al. “REL: An Entity Linker Standing on the Shoulders of Giants”. In: *SIGIR* (2020).

[12] Kenneth W Church and Patrick Hanks. “Word Association Norms, Mutual Information, and Lexicography”. In: *Computational Linguistics* (1990).

[13] Stefan Evert. “The Statistics of Word Cooccurrences: Word Pairs and Collocations”. PhD thesis. University of Stuttgart, 2005.

[14] Peng Qi et al. “Stanza: A Python NLP Package for Many Human Languages”. In: *ACL* (2020).

[15] J. Kuang et al. “Towards algorithmic framing analysis: expanding the scope by using LLMs”. In: *Journal of Big Data* 12 (2025).

[16] Michael Burnham et al. “Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text”. In: *arXiv preprint arXiv:2409.01234* (2024).

[17] Enfa Fane et al. “Fane at SemEval-2025 Task 10: Zero-Shot Entity Framing with Large Language Models”. In: *Proceedings of the 19th International Workshop on Semantic Evaluation*. 2025.

[18] Maarten Grootendorst. “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”. In: *arXiv preprint arXiv:2203.05794* (2022).

[19] M. B. Meram et al. “GPT vs. Other Large Language Models for Topic Modeling: A Comprehensive Comparison”. In: *ICCK Transactions on Emerging Topics in Artificial Intelligence* 2.3 (2025), pp. 116–130.

[20] Margarida Mendonca and Alvaro Figueira. “Modeling Political Discourse with Sentence-BERT and BERTopic”. In: *arXiv preprint arXiv:2510.22904* (2025).

[21] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: `https://arxiv.org/abs/1908.10084`.

[22] A. Sciety. “LLM-Inferred Narrative Frames in Geopolitical Conflict Reporting”. In: *Sciety Articles* (2025). Accessed: 2025-11-23.

[23] William L Hamilton, Jure Leskovec, and Dan Jurafsky. “Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016, pp. 1489–1501.

[24] *Reproducibility Checklist*. Used as a mandatory submission appendix for the 26th European Conference on Artificial Intelligence. Kraków, Poland: European Conference on Artificial Intelligence (ECAI), 2023.

Appendices

A Team Member Contributions

The project work and final report were divided among the three team members as summarized in the table below. The numerical section and subsection references align with the final structure of this document.

Table 2: Team Member Contributions by Section

Team Member	Sections Contributed	Role Summary
Łukasz Grabarski	2, 3, 4	Topic Modeling Methodology, Exploratory Data Analysis (EDA), Future Plans
Łukasz Lepianka	2, 3, 4	Lexical Statistics, Conclusion, Future Plans
Marta Szuwarska	1, 2, 3, AP:A, AP:B	Named Entity Recognition (NER), Semantic Framing and Contextual Analysis, Reproducibility

B Reproducibility Checklist

The following checklist is based on the ECAI 2023 reproducibility questions [24], confirming the adherence of this report to best practices in scientific AI/ML reporting.

(1) **Conceptual Description of AI Methods:** This paper includes a conceptual outline and/or pseudocode description of AI methods introduced. **Answer: Yes**

(2) **Delineation of Statements:** This paper clearly delineates statements that are opinions, hypotheses, and speculations from objective facts and results. **Answer: Yes**

(3) **Pedagogical References:** This paper provides well-marked pedagogical references for less-familiar readers to gain the background necessary to replicate the paper. **Answer: Yes**

(4) **Theoretical Contributions:** Does this paper make theoretical contributions? **Answer: No**

(5) **Data Sets:** Does this paper rely on one or more data sets? **Answer: Yes**

(5.1) **Motivation for Data Selection:** A motivation is given for why the experiments are conducted on the selected datasets. **Answer: Yes**

(5.2) **Novel Datasets in Appendix:** All novel datasets introduced in this paper are included in a data appendix. **Answer: NA**

(5.3) **Novel Datasets Publicly Available:** All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Answer: NA**

(5.4) **Citations for Existing Data:** All datasets drawn from the existing literature are accompanied by appropriate citations. **Answer: Yes**

(5.5) **Public Availability of Existing Data:** All datasets drawn from the existing literature are publicly available. **Answer: Yes**

(5.6) **Description of Non-Public Data:** All datasets that are not publicly available are described in detail, explaining why publicly available alternatives are not scientifically satisfying. **Answer: NA**

(6) **Computational Experiments:** Does this paper include computational experiments? **Answer: Yes**

(6.1) **Pre-processing Code:** Any code required for pre-processing data is included in the appendix. **Answer: Yes**

(6.2) **Source Code Included:** All source code required for conducting and analysing the experiments is included in a code appendix. **Answer: Yes**

(6.3) **Source Code Publicly Available:** All source codes required for conducting and analysing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. **Answer: Yes**

(6.4) **Code Comments and References:** All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from. **Answer: Yes**

(6.5) **Randomness and Seeds:** If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. **Answer: NA**

(6.6) **Computing Infrastructure:** This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models, amount of memory, operating system, names and versions of relevant software libraries and frameworks. **Answer: Software - Yes**

(6.7) **Evaluation Metrics:** This paper formally describes the evaluation metrics used and explains the motivation for choosing these metrics. **Answer: NA**

(6.8) **Number of Runs:** This paper states the number of algorithm runs used to compute each reported result. **Answer: NA**

(6.9) **Analysis of Variation:** Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average, median) to include measures of variation, confidence, or other distributional information. **Answer: No**

(6.10) **Statistical Tests for Significance:** The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed rank). **Answer: No**

(6.11) **Final Hyper-parameters:** This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. **Answer: Yes**

(6.12) **Hyper-parameter Search Space:** This paper states the number and range of values tried per (hyper-) parameter during the development of the paper, along with the criterion used for selecting the final parameter setting. **Answer: NA**