Putin's Talks

Preliminary report

Authors: Katarzyna Leniec, Jan Pogłód, Cyprien Fourcroy, Michał Puścian

1. Introduction

In this project, we want to use Natural Language Processing methods to analyze political speeches, especially those of Vladimir Putin. NLP gives us tools to automatically study topics, emotions, propaganda rhetoric, and how rhetoric changes over time. In this part of the report, we review the literature, available models and tools, and open-source datasets.

2. Proposed Analysis Questions

Our investigation is structured around a set of research questions grouped into six categories (Table 1). The **STATISTICS** category includes descriptive measures of the speeches (counts, proportions, etc.). The **CONTEXT** category addresses external factors such as events, audience, and setting. The **CHANGE OVER TIME** category considers temporal trends in the discourse. The **COMPARISON** category involves contrasting Putin's speeches with other speeches or contexts. The **INTERPRETATIONS** category focuses on the substantive meaning and implications of observed patterns. Finally, **CRITICAL TESTS** involve validation checks and robustness analyses. Table 1 lists all proposed analysis questions under these categories and proposition of methods and models to answer them.

STATISTICS

1. Which three economy-related terms appear most often? (Word frequency analysis, economy lexicon)

2. How many times does the word “modernization” appear in a military context? (Keyword co-occurrence + embeddings similarity, BERTopic / custom military lexicon)

3. How often does he speak about “friendship” in relation to China? (Keyword co-occurrence + topic detection, BERTopic)

4. In which years do the most references to World War II occur? (Named Entity Recognition, spaCy)

CONTEXT

1. How does Putin characterize the USA – more as a threat or as a potential partner? (Sentiment analysis, RoBERTa sentiment classifier)

2. What metaphors does Putin use toward the USA, Ukraine, Poland, and Germany? (Metaphor detection, Transformer-based metaphor model / embedding similarity)

3. In references to Poland, are historical or contemporary contexts more frequent? (Topic modeling, BERTopic)

4. In what context does “Poland” most often appear? (enemy, partner, ally, neighbor) (Zero-shot classification / context classification, BART or RoBERTa)

CHANGE OVER TIME

1. How does the image of Ukraine (Poland, Germany, USA) change in his speeches from 2000 to 2024? (Sentiment analysis + temporal aggregation, RoBERTa)

2. When do numerous references to the “Russian world” (russkiy mir) begin to appear? (Keyword frequency over time + NER, spaCy / HuggingFace NER)

3. When do themes of “sovereignty” begin to dominate foreign policy discourse? (Topic modeling over time, BERTopic)

4. In which years do economic topics appear most frequently? (Topic modeling + word frequency, BERTopic / economy lexicon)

COMPARISON

1. How does the narrative toward Poland differ from that toward Germany? (Sentiment analysis, RoBERTa)

2. Does he speak more positively about China than about India? (Sentiment analysis, RoBERTa)

3. How does the tone toward the USA compare with that toward the European Union? (Sentiment analysis + topic modeling, RoBERTa / BERTopic)

4. How does the image of Germany differ between 2003 (Iraq War) and 2014 (Ukraine crisis)? (Sentiment analysis + topic comparison, RoBERTa / BERTopic)

INTERPRETATIONS

1. What are Putin's most common arguments for strengthening the army? (Topic modeling, BERTopic)

2. How does he construct the image of the "enemy"? (Topic modeling + embedding clustering, BERTopic / Sentence-BERT)

3. What historical events does he use to legitimize actions toward Ukraine? (Topic modeling + NER, BERTopic / spaCy NER)

4. What elements of the "Great Russia" myth recur in his speeches? (Word statistics + manual inspection, Bag-of-Words / frequency counts)

5. What are the three main ways he describes the West? (Topic modeling + important words extraction, BERTopic)

6. How does he portray Russia's role in the world – as a defensive or expansionist power? (Sentiment analysis + topic context, RoBERTa / BERTopic)

CRITICAL TESTS

1. Summarize the speech from date X (Abstractive summarization, BART / T5)

2. Has Putin ever spoken about event Z? (Keyword search + sentence extraction, Regex / spaCy)

3. Provide quotes where he describes Poland in historical terms. (Keyword search + NER + sentence extraction, spaCy / Regex)

4. Has he ever used the term "democratization" in a positive context? (Zero-shot sentiment classification, RoBERTa)

5. What three different arguments does he invoke when speaking about sanctions? (Topic modeling + co-occurrence, BERTopic / embedding clustering)

6. List the passages in which he refers to Lenin or the USSR. (Keyword search + sentence extraction, Regex / spaCy)

3. Literature and Research on NLP in Political Discourse

- **Analysis of Putin's rhetoric:**
  - In a paper "Analyzing Russia's propaganda tactics on Twitter using mixed methods network analysis and natural language processing: a case study of the 2022 invasion of Ukraine" from 2024, Alieva et al. examine Russia's propaganda discourse on Twitter during the 2022 invasion of Ukraine. They construct a pipeline to identify topics, influential actors, and examine the most impactful messages in spreading disinformation narrative. In the pipeline, they use many methods including network analysis, NLP techniques and qualitative analysis.
- **Topic modeling:**
  - In the report *"Unpacking Russian Presidential Speech Patterns with Machine Learning"*, researchers used LDA (Latent Dirichlet Allocation) on Putin's speeches to identify main topics, such as energy or international relations.
- **Propaganda detection:**
  - In a paper from Martino et al. (2020) [1] authors describe competition that was about detecting propaganda in news articles. They got 44 submissions and in this paper they analyze findings about architectures, methods and results obtained. This can help us avoid pitfalls and give a starting point for our model development.
  - In NLP literature, online propaganda is a growing topic. For example, the HQP dataset contains manually labeled propaganda texts, which can be used to train models that classify propaganda - "*Large Language Models for Propaganda Detection".*

4. Tools and Open-Source Models

To analyze Putin's speeches, we plan to use these tools and pre-trained models:

- **Hugging Face Transformers -** A library that gives access to many models (BERT, RoBERTa, XLM-R, etc.). It can be used for classification, tokenization, embeddings, and more.
- **BERTopic** - A topic modeling tool based on embeddings (e.g., BERT). It is good for finding semantic topics in speeches and tracking how they change over time.
- **spaCy-Transformers -** Useful for basic text processing (tokenization, lemmatization)

5. Useful Pre-trained Models for This Project

Below we listed some examples of pre-trained models for political speech analysis:

- **Sentiment / emotions:**

    - Models like RoBERTa fine-tuned for sentiment analysis can be used to detect emotional tone in speeches (positive, negative, neutral).
- **Propaganda detection:**
    - Fine-tuning transformer models (e.g., RoBERTa) on propaganda datasets (like SemEval-2020) allows classification of speech fragments by propaganda techniques (e.g., loaded language, emotional appeal).
- **Framing / metaphors:**
    - Token-classification models (BERT or XLM-R) can be fine-tuned to detect metaphors or narrative frames in text.
- **Diachronic analysis:**
    - BERTopic with dynamic topic modeling can show how topics evolve over time in the speech corpus.

6. Datasets (Open Datasets)

For this project, we will use several open datasets suitable for political speech analysis:

Dataset	**Description / Why We Use It**
Putin Corpus (2012–2022) [8]	Corpus of Vladimir Putin's speeches from kremlin.ru. It is useful for analyzing topics, rhetoric, and propaganda.
Chronorhetorics Corpus (until 2025) [9]	A corpus designed for temporal rhetoric analysis — how politicians refer to past and future to legitimize power. Can be used as a comparison or for temporal analyses. Features many countries, including Russia.
Ideology & Power in Parliamentary Debates [10]	Parliamentary debate data (ParlaMint) for the "Ideology and Power Identification" task (CLEF 2025). Useful for training models to analyze ideology, rhetoric, and comparing to Putin's speeches.

HQP – Online Propaganda [11]	Large labeled propaganda dataset (~30k examples). Ideal for training and testing propaganda classifiers.

7. Data Preprocessing - Putin Corpus dataset.

The project is based mostly on a provided dataset in the form of a single .json file. This file contains a list of objects, where each object represents a single speech transcript. The key fields within each JSON object are:

- `date:` The date and time of the speech.
- `transcript_unfiltered:` The full, unfiltered transcript of the event.
- `transcript_filtered:` A cleaned version of the transcript, focusing on the core content of the speech.
- `wordlist:` A list of pre-processed (lemmatized) words from the transcript.
- `title`: The title of the event.
- `kremlin_id:` The document's identifier from the source system.
- `place`: The location where the speech took place.
- `persons, teasers, tags`: Additional metadata describing the document.

The dataset was loaded from a JSON file into a DataFrame. Missing values in the text columns were filled and converted to strings. The date column was converted to datetime, and rows with missing dates were removed. Only speeches containing “Putin” before a colon were retained, and the speaker prefix was removed. Year and month features were extracted. The main text was selected from the filtered transcript or, if missing, from the unfiltered transcript. Text was split into sentences and tokenized. Tokens were lowercased, filtered to keep only alphabetic words, stopwords were removed, and lemmatization was applied. Additional statistics were computed, including token counts, unique tokens, lexical diversity, and average sentence length. Rhetorical features such as the number of exclamation marks, question marks, and ellipses were also extracted.

8. EDA - Putin Corpus dataset.

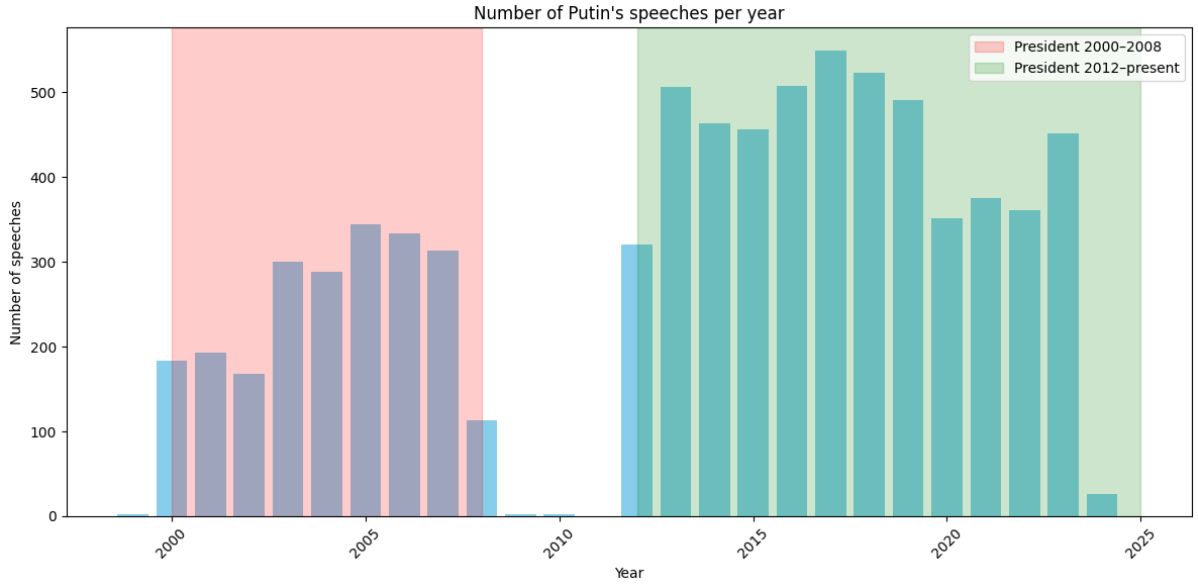
Basic metrics:

- Total Number of Speeches: 7629
- Average Speech Length: 921.04
- Average Lexical Diversity: 0.47
- Average Word Count: 921.04
- Minimum Word Count: 6

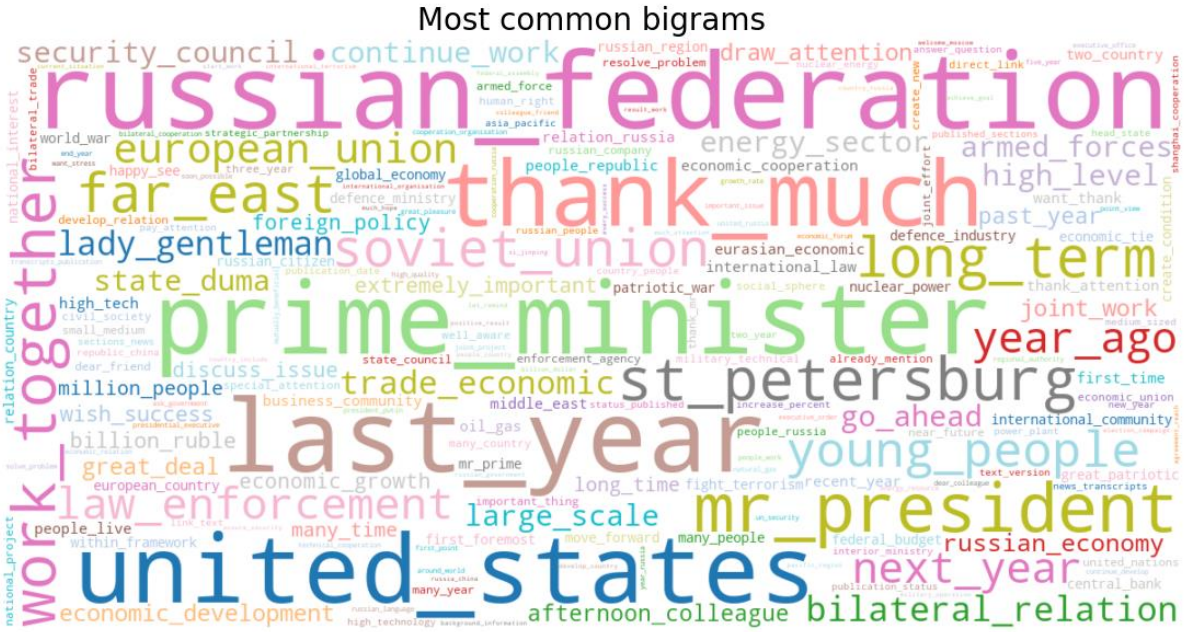
- Maximum Word Count: 26634

More analysis with plots:

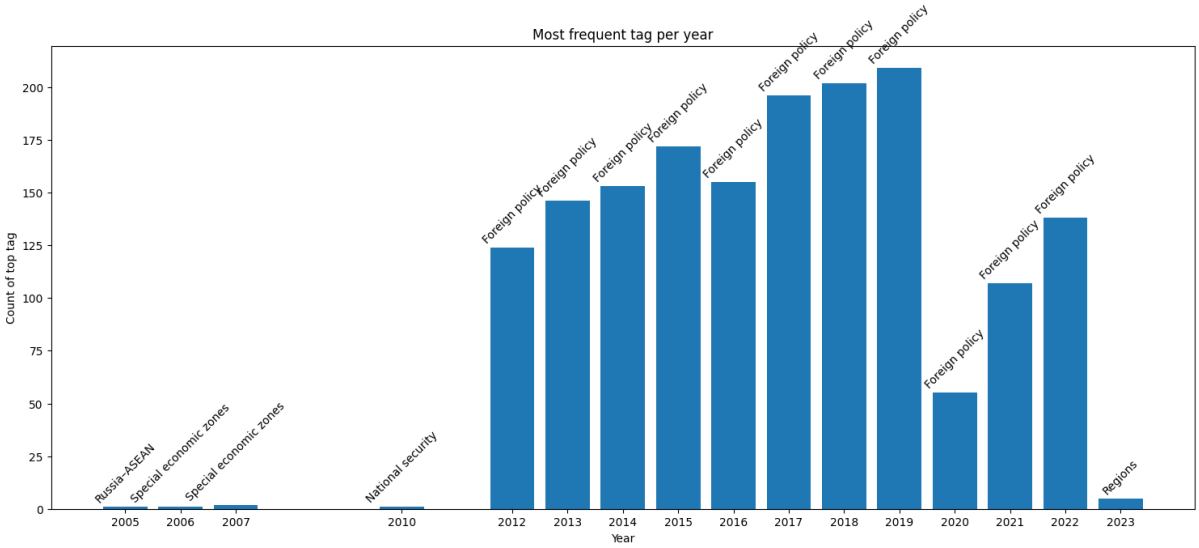
1. The chart shows the number of speeches delivered by Putin each year. The data were grouped by year and displayed as bars. Additionally, his presidential terms are highlighted: 2000-2008 in red and 2012-2025 in green, making it easier to visually relate speech activity to the periods he held office.



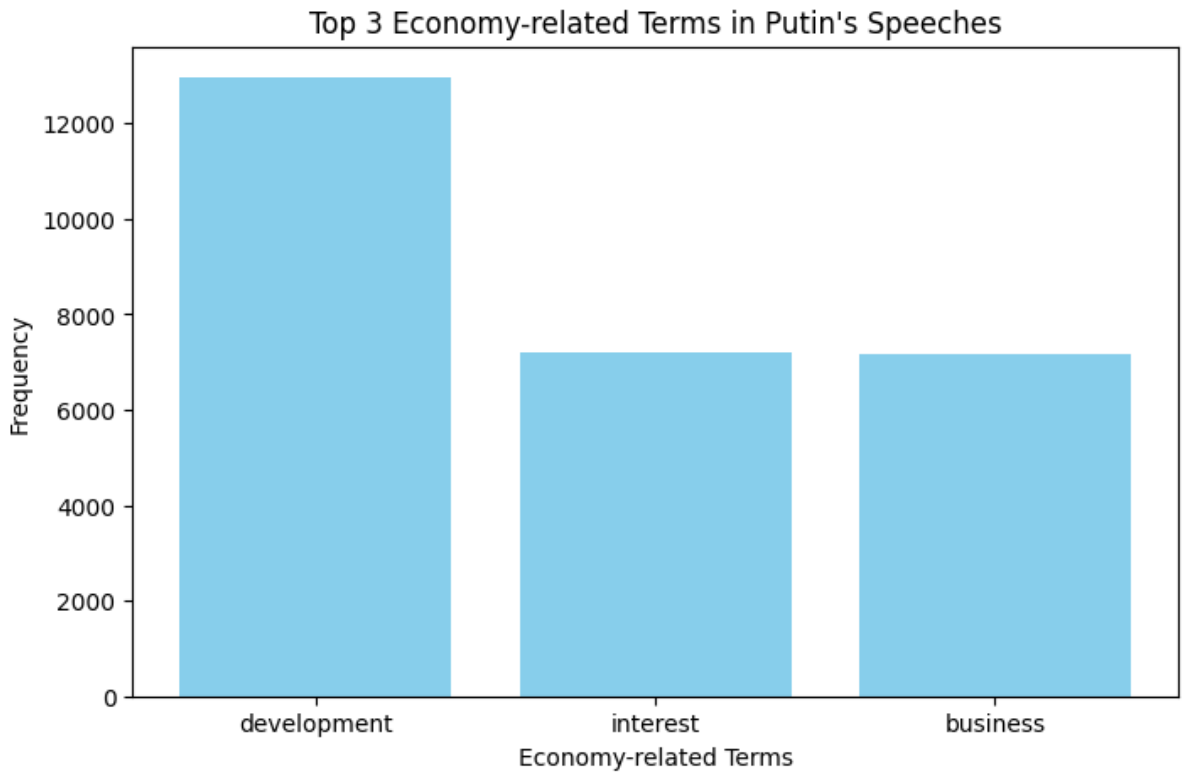
2. We extracted all bigrams (pairs of consecutive words) from the speeches after removing stopwords. To focus on meaningful content, we filtered out bigrams containing “weak” words such as would, like, or know. The top 20 most frequent bigrams were then identified and visualized as a word cloud, highlighting the phrases that appear most often in Putin’s speeches.



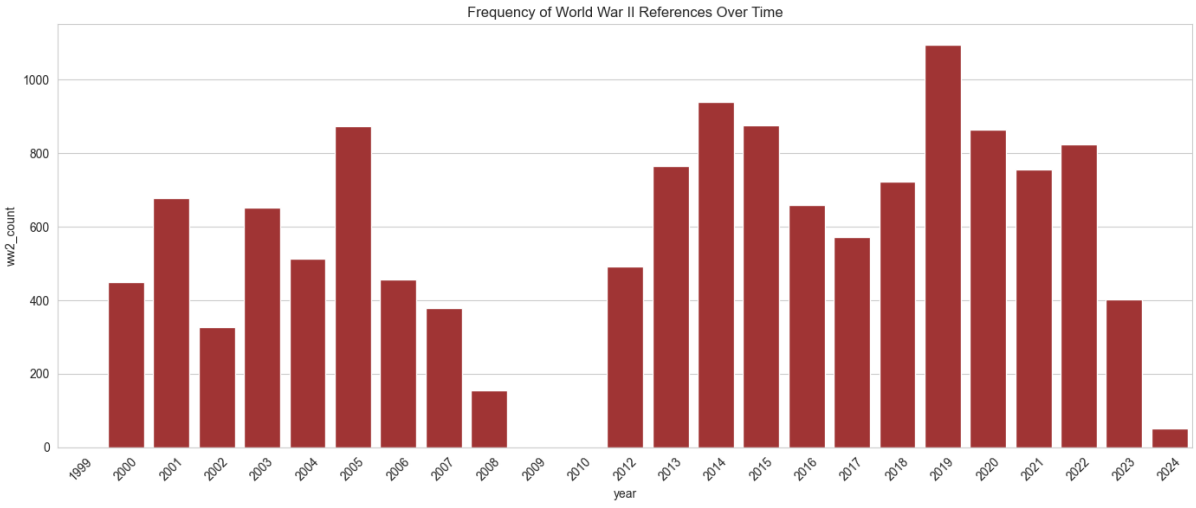
3. We analyzed the tags associated with each speech to identify the most frequent tag per year.



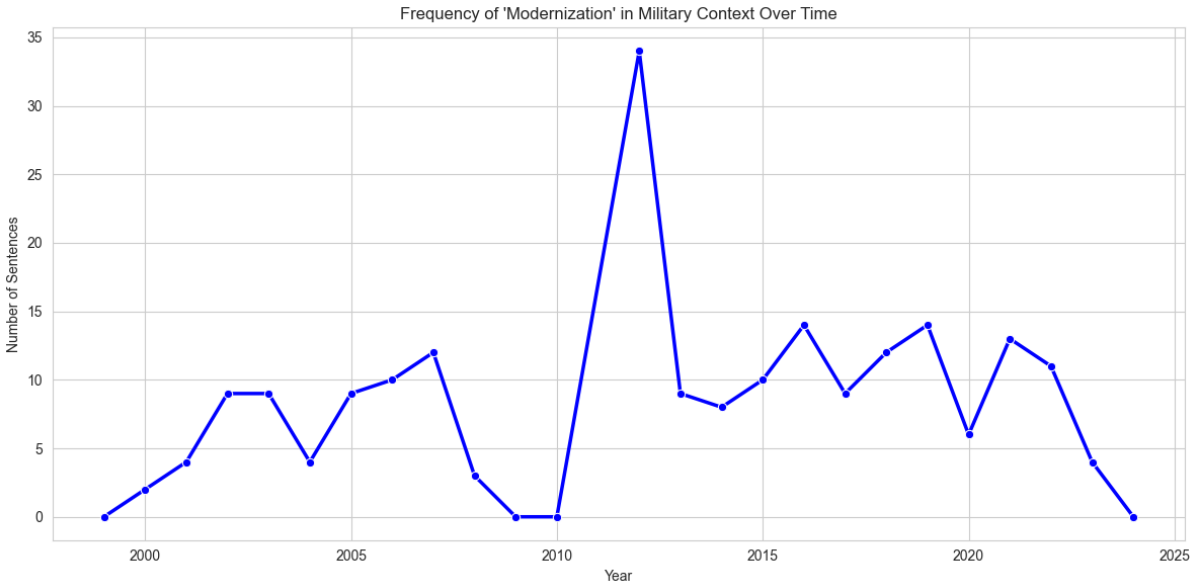
4. We identified the most frequently used economy-related terms. First, we created a comprehensive list of keywords related to economics, finance, and business. Then, we counted how often each term appeared in the speeches (ignoring stopwords). Finally, we extracted the top three most frequent economy-related terms.



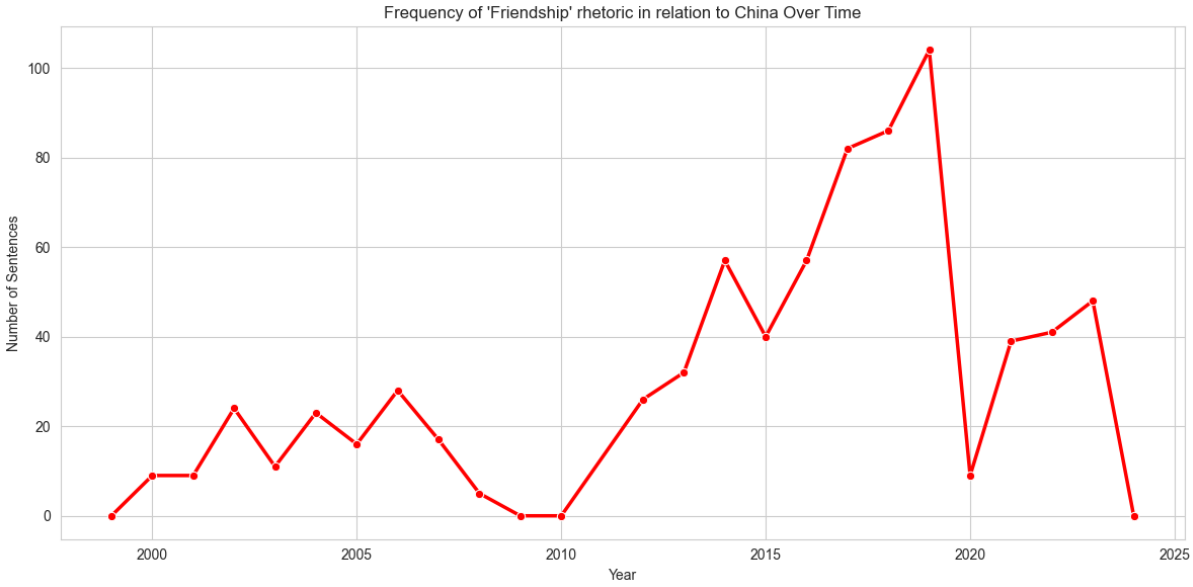
5. We identified references to World War II. First, we created an extensive list of WW2-related keywords, including key events, figures, and dates. Then, for each speech, we counted how many of these keywords appeared. We aggregated the counts per year to see trends over time.



6. We tracked mentions of "modernization" specifically in a military context. First, we defined two keyword lists: one for modernization-related terms (e.g., modernization, upgrade, innovation) and one for military-related terms (e.g., army, armed forces, weapon, missile). For each sentence in the speeches, we checked if it contained both a modernization keyword and a military keyword.
   We then aggregated these counts per year to observe trends over time.



7. We tracked sentences that mention both China and friendship-related concepts. Two keyword lists were defined: one for China (e.g., *China*, *Beijing*, *PRC*) and one for friendship/partnership terms (e.g., *friendship*, *ally*, *cooperation*, *close ties*). For each sentence, we counted it if it contained at least one keyword from both lists. These counts were then aggregated per year to observe trends over time.



9. Data Preprocessing - Chronorhetorics-corpus

This Chronorhetorics-corpus dataset has similar schema and json structure as the previous dataset, so we performed the same processing on it. The most important metadata are:

"title" – for example, "Russian-Chinese talks have confirmed the strategic nature of the two countries' bilateral relations"

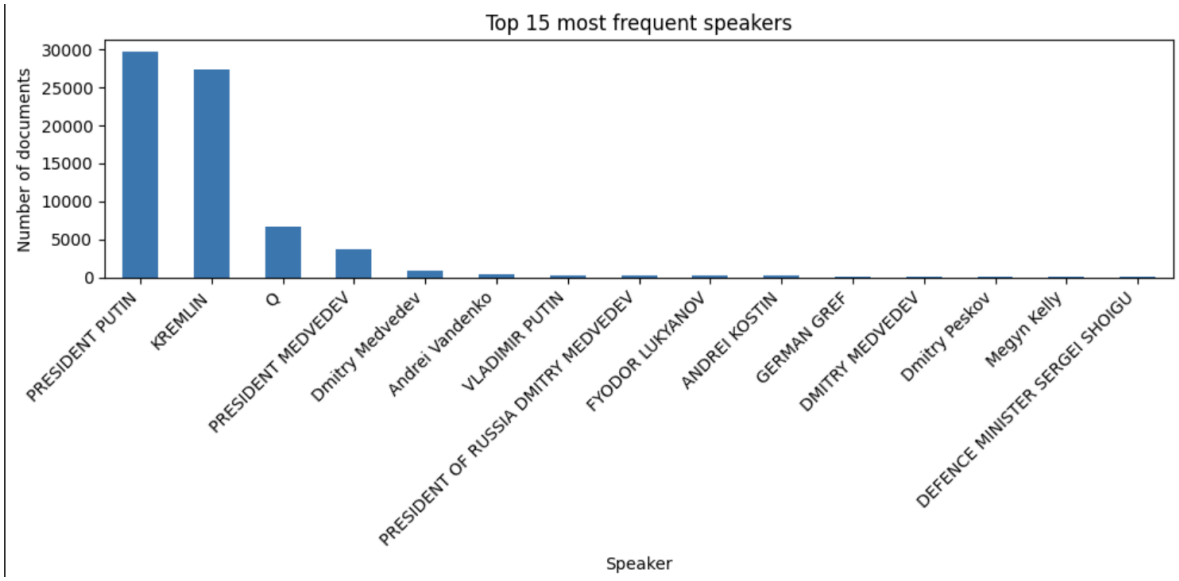
"language": (mostly English),

"date" in the format "2007-03-26,"

"location": for example, "Moscow,"

"speaker": for example, "KREMLIN" / "Putin"

"text": "The Russian and Chinese leaders met six times in 2006....



We can see that mostly there are two speakers with the biggest number of talks – Putin and Kremlin. We can also observe different names for Putin as in the pre processing we will ensure all of the Putin's speeches are combined to the Putin person.

10. EDA - Chronorhetorics-corpus

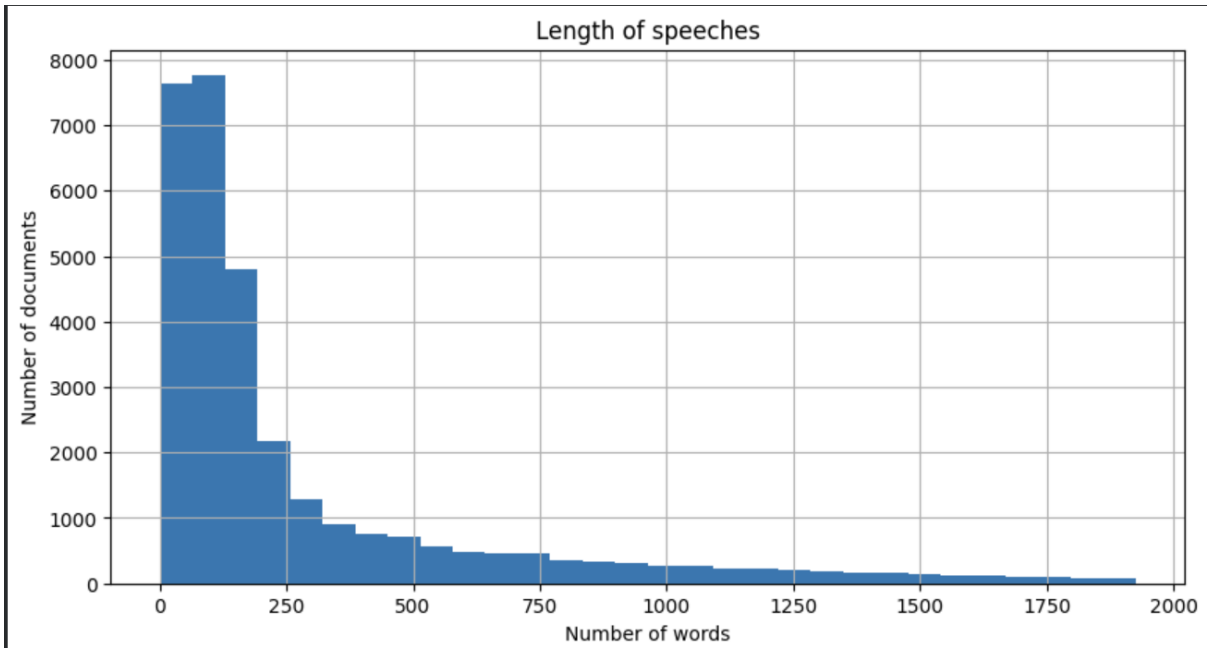
The dataset of Vladimir Putin's talks in this dataset contains:

Documents analyzed: 33,129

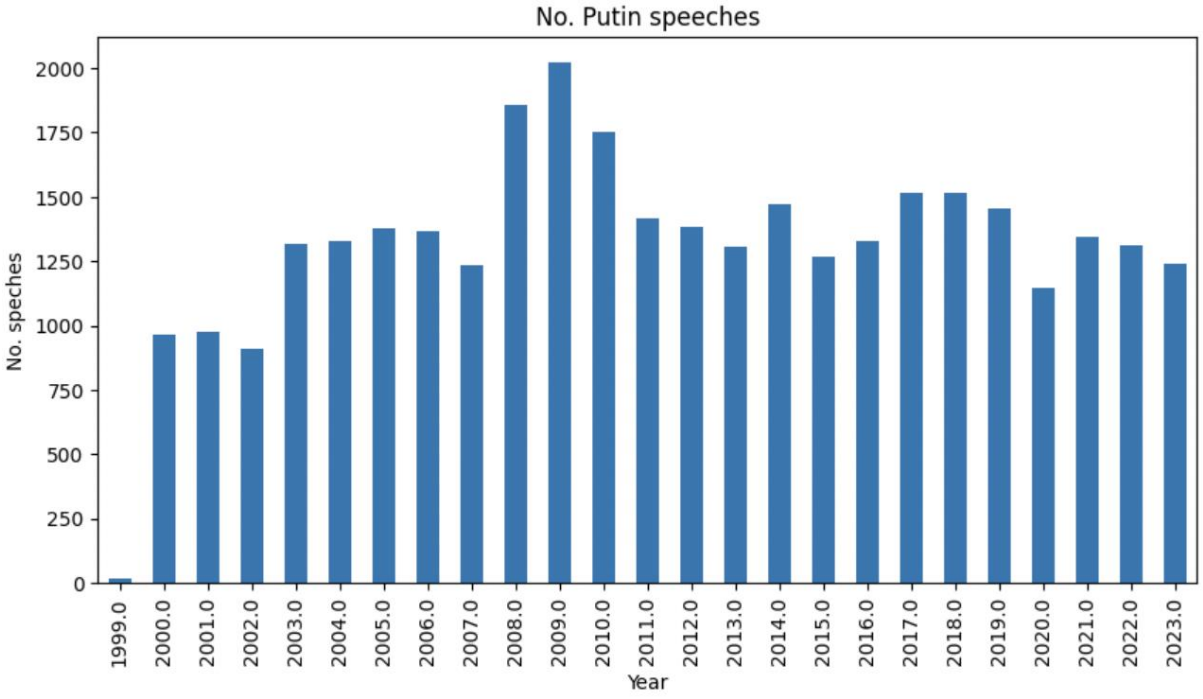
Average speech length: 486 words

Median speech length: 140 words

In addition, the collection includes statements of a very diverse nature: from short, technical answers to long ideological narratives. In most cases we can observe the texts were a maximum of 250 words long, so they were rather short speeches.

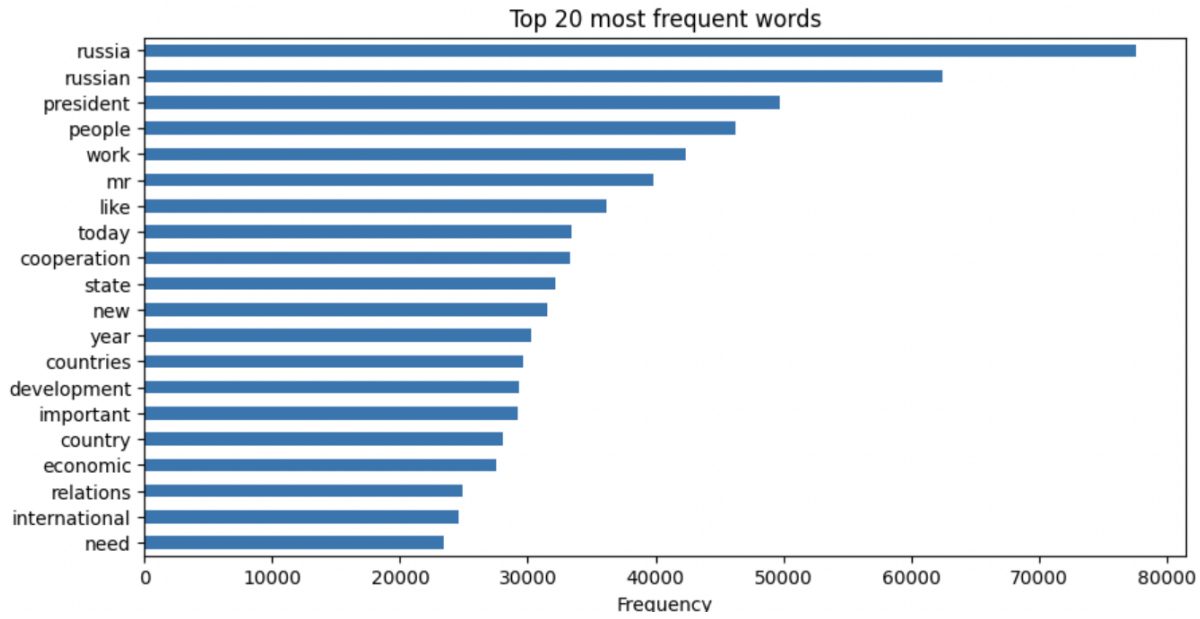


Below we can observe the distribution of speeches between 1999 and 2023. We can see that between 2008 and 2010 Putin spoke a little more than in other years.

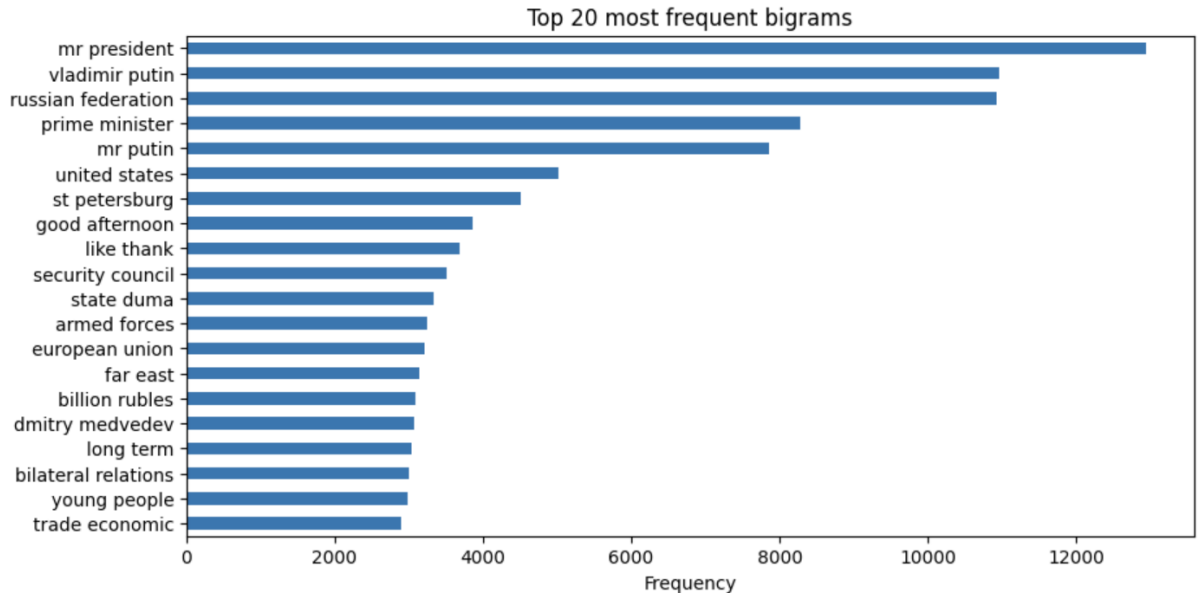


Based on the two frequency charts (single words and bigrams), we can describe the main topics of Vladimir Putin's speeches in the following way.

First, the most frequent single words clearly show that Putin often talked about Russia and the Russian state. Words like "Russia", "Russian", "country", and "state" appear very often, which suggests a strong focus on national identity and sovereignty.



Another important group of words refers to people and society, such as “people”, “work”, “development”, and “important”. This indicates that many speeches addressed social issues, economic development, and the role of citizens. There is also a strong presence of political and institutional language. Words like “president”, “relations”, “international”, and “economic” may show that foreign policy and international cooperation were frequent topics. Time-related words such as “today” and “year” suggest that speeches were often connected to current events and annual summaries.



The bigrams give more concrete context. Bigram such as “Russian Federation”, “state duma”, and “security council” show that he often spoke within official political settings. International relations are also visible through phrases like “United States”, “European Union”, and “bilateral relations”. In addition, some bigrams point to economic and regional topics, for example “billion rubles”, “far east”, and “long term”.

Overall, the charts show that Putin most often spoke about Russia, state power, international relations, economic development, and the role of people. His language is formal, institutional, and focused on governance, stability, and national interests.

10. References

1. Martino, G., Barrón-Cedeno, A., Wachsmuth, H., Petrov, R., & Nakov, P. (2020). *SemEval-2020 Task 11: Detection of propaganda techniques in news articles*. arXiv:2009.02696.
2. Abdullah, U., & Alahmadi, D. (2022). *Detecting propaganda techniques in English news articles using pretrained Transformers*. (EMNLP workshop paper.)
3. Grootendorst, M. (2022). *BERTopic: Embedding-based topic modeling with BERT*.
4. Alieva, Iuliia, Ian Kloo, and Kathleen M. Carley. "Analyzing Russia's propaganda tactics on Twitter using mixed methods network analysis and natural language processing: a case study of the 2022 invasion of Ukraine." *EPJ Data Science* 13.1 (2024): 42.
5. Mendonça, M., & Figueira, A. (2025). *Modeling Political Discourse with Sentence-BERT and BERTopic*. arXiv:2510.22904.
6. Mochtak, M., Rupnik, P., & Ljubešić, N. (2024). *The ParlaSent Multilingual Training Dataset for Sentiment Identification in Parliamentary Proceedings*. LREC-COLING 2024.8. Sprenkamp, K., Jones, D. G., & Zavolokina, L. (2023). *Large Language Models for Propaganda Detection*. arXiv:2310.06422.
7. Wolf, T., Debut, L., Sanh, V., *et al.* (2020). *Transformers: State-of-the-Art Natural Language Processing*. EMNLP Demonstration.
8. https://github.com/levshina/Putin_Corpus
9. https://github.com/HCDH-Uni-Heidelberg/chronorhetorics-corpus
10. https://touche.webis.de/clef25/touche25-web/ideology-and-power-identification-in-parliamentary-debates.html
11. Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2024. HQP: A Human-Annotated Dataset for Detecting Online Propaganda. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6064–6089, Bangkok, Thailand. Association for Computational Linguistics.
12. https://arxiv.org/pdf/2406.12614