The analysis of Russian strategic communication in years 2000-2024 using Natural Language Processing techniques
Project Proposal for NLP Course, Winter 2024

Grzegorz Zbrzeżny
Warsaw University of Technology
01161645@pw.edu.pl

Izabela Telejko
Warsaw University of Technology
01161635@pw.edu.pl

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

Abstract

TBA after whole report is completed.

Introduction

The project aimed to check the hypothesis that with the use of Natural Language Processing (NLP) based methods we are able to do an in-depth, rigorous, and statistically sound analysis of Russian strategic communication addressed to and about a specific actors – Poland and whole Europe. We want to visualize the evolution of the Russian political language related to this states, to understand its statistical prominence in the corpus, to prepare topic modelling, and to examine prevailing sentiment and its evolution throughout the years. This study applies NLP techniques to key sources for analysis of the political language of Russian state: the speeches of Vladimir Putin, authoritarian leader who has been in power for the last 25 years (since 2000), as the president (2000-2008 and since 2012 until now) and prime minister (2008-2012). Using NLP techniques we wanted to look back in order to put evolution of political language against the background of political changes and to explore its predictive value, namely to what extent are political speeches ex-post reactions to events, and to what extent do they announce incoming actions. Our project is conducted in collaboration with `Centrum Dialogu im. Juliusza Mieroszewskiego`, which provides guidance and directions for analysis tailored to their research needs.

Related Work

In this section, we present an overview of the current state-of-the-art solutions relevant to the subject of our study. Our goal is to analyze sentiment trends regarding Poland in Russian political speeches and their evolution over time. Therefore, we emphasize reviewing articles and methodologies that highlight the most efficient techniques for sentiment analysis, document classification as well as strategies for conducting insightful analyses of political data.

Document Classification

The exponential growth of online text has made manual text analysis impractical, necessitating automated methods like document classification (Tsirmpas et al., 2024). It is a machine learning approach aimed at categorizing documents into predefined labels or categories based on their content. This technique is applicable in areas like spam detection, sentiment analysis, and topic modeling, enabling effective analysis and comprehension of textual data.

This task can be approached using various methods, ranging from baseline techniques, such as Naive Bayes, to more advanced models like LSTM, which often yield better performance in this context (Ranjan and Prasad, 2023). Additionally, modern architectures such as BERT (Devlin et al., 2018) have demonstrated superior performance compared to simpler methods, albeit at the expense of significantly higher resource consumption (Taha et al., 2024).

Text Preprocessing

Humans are capable of extracting key information from texts to prepare insightful analyses, but machine learning (ML) models are less proficient at this task. To bridge this gap, raw texts require preprocessing before they can be effectively used for classification or other ML tasks. Fundamental preprocessing steps include transformations like lowercasing, removing stop words, and applying lemmatization or stemming (Siino et al., 2024).

Since ML models operate on numerical vectors rather than words, text must also be converted into this form using word or sentence em-

beddings. Approaches for this conversion range from basic techniques like Bag of Words (BOW) (Qader et al., 2019) to more advanced ones like Word2Vec (Mikolov et al., 2013). However, these methods face challenges with out-of-vocabulary (OOV) words, which can be addressed using techniques such as MorphoRNN (Wang et al., 2020), based on N-grams, or fastText (Joulin et al., 2016), which leverages a Continuous Bag of Words (Xia, 2023) (CBOW) framework.

Additionally, words often have multiple meanings depending on their context. To capture this nuance, advanced models like ELMo (Peters et al., 2018) and OpenAI GPT (Radford and Narasimhan, 2018) generate embeddings that are context-dependent, enabling a more accurate representation of the text's semantics.

Political Text Analysis

For many years, political text analysis depended on manual efforts, where humans examined speeches, policy documents, and other texts to uncover patterns, sentiments, and key themes. While this approach provided valuable insights, it was highly time-consuming and lacked scalability, particularly given the increasing volume of political discourse. With the integration of computational methods and natural language processing (NLP), this process can be greatly improved, enabling large-scale analysis and deeper insights into patterns and trends. Tools such as word clouds, lexical dispersion plots, and time series visualizations (Katre, 2019) can be used to track changes in narrative and keyword usage over time. Additionally, machine learning models (Efat et al., 2023) can identify the context of political statements and the emotions they aim to evoke, offering understanding of political communication.

Sentiment Analysis

Sentiment analysis (Dudhabaware and Madankar, 2015) is an NLP task trying to determine whether the provided text is positive, negative or neutral. This task can be quite complicated, taking into account that sometimes humans themselves have a problem with distinguishing sentiment, and it is affected, among others, by such a detail as the order of the sentence used (Fang and Zhan, 2015). Therefore we can not expect any model to be 100% accurate with its responses, as the labels are subjective and dependent on the person who labelled the data.

One of the most intricate things to identify in the written text is the use of irony (Hernandez Farias et al., 2015). It is quite a subtle thing, as it could be impossible to determine if the irony was used for the sentence out of context. Also, the sentence might appear completely fine on paper, and only the manner of expression can betray the use of irony, which makes it even harder for textual data. Sometimes, the following sentences may indicate earlier use of irony by incorporating contradictory content.

The accepted types of sentiment labels may also vary. Instead of a three-point scale, there can be used a more detailed one. Beside the sentiment, we may also want to define the intensity of sentiment (strong or weak).

Named Entity Recognition

Named Entity Recognition (NER) is a fundamental task in NLP that focuses on identifying and classifying entities such as names, locations, amd dates within a text. Earlier approaches to NER predominantly relied on rule-based systems and statistical methods like Hidden Markov Models (Eddy, 1996) and Conditional Random Fields (Sutton and McCallum, 2010), which required extensive feature engineering and linguistic expertise. With recent advancements in NLP, transformer-based models such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) have revolutionized the field. These models utilize contextual embeddings and self-attention mechanisms, significantly outperforming traditional methods and achieving state-of-the-art accuracy across diverse applications.

Dataset

We utilize a dataset comprising transcripts of Vladimir Putin's speeches, translated into English and publicly accessible via the official website of the Russian president[1]. The dataset is structured as a list of JSON objects, where each JSON represents an individual speech along with its associated metadata. Key features available in the dataset include the `date` and `place` of the speech, `tags`, `title`, and both `filtered` and `unfiltered transcripts`. The dataset encompasses 9,838 speeches delivered between the years 2000 and 2024.

[1] `kremlin.ru`

Proposed Solution

Our objective is to analyze the sentiment and its evolution in speeches that reference Poland, Europe or familiar entities identified using NER. Given the raw text nature of the dataset, initial preprocessing steps are necessary before the data can be used for modeling. Following preprocessing, we will conduct exploratory data analysis (EDA). Finally, we will vectorize the dataset and utilize several pretrained models, comparing their performance to determine the most effective approach for sentiment analysis.

Preprocessing

The data processing workflow begins by removing stop words using the list of English stop words provided by `nltk.corpus`, ensuring that common, non-informative words are excluded. For each speech, the date and filtered transcript are extracted for further analysis. Punctuation and unnecessary symbols, such as #, ..., and ', are removed to maintain consistency and clarity in the text.

Next, the speeches are tokenized and lemmatized using `spaCy`, breaking the text into individual words and reducing them to their base forms for uniformity (e.g., "running" becomes "run"). The processed data is then organized into a dictionary where the keys represent each month in the format `YYYY-MM` (e.g., "2001-01") and the values are lists of the processed speeches corresponding to that month.

Exploratory Data Analysis

As part of the exploratory data analysis, we computed basic text statistics following the preprocessing stage. The average length of the speeches, measured in token count, was found to be 489 tokens. Additionally, the average token length was calculated as 6.41 characters. We also analyzed the most common tokens in the speeches, as summarized in Table 1. The results highlight expected patterns, with the speeches predominantly focusing on Russians themselves and their relationships with other countries

Next, we extended the analysis to examine the same statistics on a monthly basis and created an animation (sample frame shown in Figure 1) to visualize how the subjects of the speeches evolved over time. While terms like Russian or Russia dominate in most months, certain months stand out with other tokens, such as work or submarine, taking prominence. This provides valuable insights into the shifting focus of the speeches, potentially reflecting significant historical events. By comparing these trends with a timeline of important events, we can explore how the topics of the speeches change before and after key moments in Russian politics, offering a deeper understanding of the interplay between political discourse and historical context.

Term	Count
russian	229
russia	227
relation	156
people	151
germany	128
minister	123
republic	119
romania	119
country	117
president	112

Table 1: Terms and their counts.

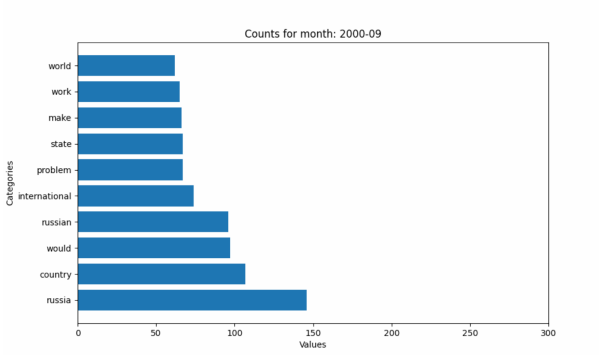

Figure 1: The most common terms in given month

Given our focus on Poland and Europe, we prepared diagrams (Figure 2 for Poland and Figure 3 for Europe) showing the monthly occurrences of terms related to Poland (`poland`, `polish`) and Europe (`europe`, `eu`, `european`). In these visualizations, we highlighted the months with the highest peaks. By analyzing the fluctuations in these token counts, we can trace the varying importance of these two entities in Russian strategic communication over time, offering insights into their evolving roles in the discourse.

Finally, we calculated the Term Frequency–Inverse Document Frequency (TF–IDF) for each

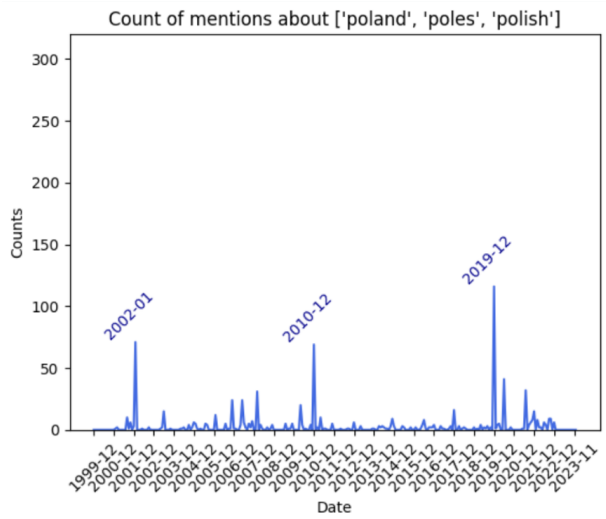

Figure 2: The most common terms in given month

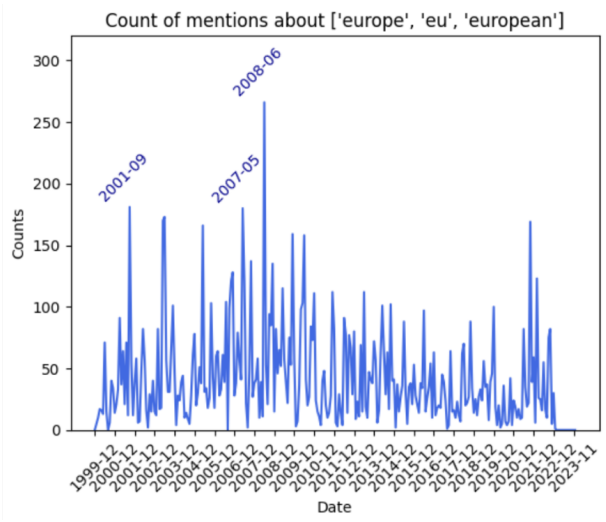

Figure 3: The most common terms in given month

token in the corpus, treating each speech as a single document. This allowed us to identify terms with the high scores as those were widely used within individual speeches but rare across the corpus. However, we filtered out highest–scoring terms as they were in majority meaningless and they appeared only in one or two speeches without contributing meaningful context. Among the remaining high–scoring terms were examples such as `disarmament`, `brussels`, and `sailor`. In addition, we want to analyze terms with low TF–IDF scores, which offer limited predictive value due to their prevalence across the corpus.

Proof of Concept

In this section, we present the results of our work. We developed a Proof of Concept (PoC) that leverages pre-trained models to analyze and track basic sentiment changes within the speeches.

Models

To perform sentiment analysis on our unlabeled dataset, we will leverage pretrained, open-source models. Our first choice is the DistilBERT model (`distilbert/distilbert-base-uncased-finetuned-sst-2-english` [2]) which is a streamlined transformer equipped with a sequence classification/regression head. Fine-tuned on the SST-2 dataset, this lightweight model achieves an accuracy of 91.3% on the development set, performing slightly below the base BERT model, which reaches 92.7%.

The second model used for sentiment analysis task is the Sentiment RoBERTa model (`siebert/sentiment-roberta-large-english` [3]), a fine-tuned version of RoBERTa-large, which is trained and evaluated across 15 diverse datasets to ensure strong generalization. This approach enables it to significantly outperform a DistilBERT-based model trained solely on SST-2, achieving 93.2% accuracy compared to DistilBERT's 78.1%, a margin exceeding 15 percentage points.. Finally, we will experiment with GPT-4o (OpenAI, 2024), leveraging manually created prompts to extract predicted sentiment from predefined sentiment categories for each given text as in (Telejko, 2023).

Experiments

Sentiment change in speeches about Europe.

To track how the sentiment of speeches related to Europe evolved over time, we filtered speeches containing Europe–related terms. We then wanted to perform a binary classification of sentiment, categorizing the content as either positive or negative. Each speech was split into chunks of 256 tokens to ensure compatibility with the BERT model. For each speech, the final sentiment was determined by averaging the numeric scores for the classes across all its chunks and selecting the class with the higher average score. The model used for this analysis was

[2] `https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english`

[3] `siebert/sentiment-roberta-large-english`

`distilbert/distilbert-base-uncased-finetuned-sst-2-english`, along with its corresponding tokenizer, fine-tuned for optimal performance with this model. The results are shown in Figure 4

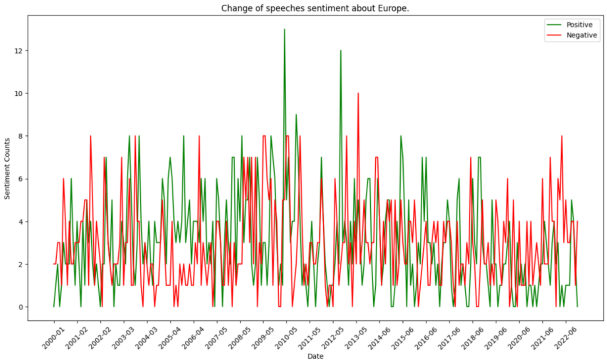

Figure 4: Sentiment evolution in speeches about Europe.

Sentiment change in fragments about Poland.

In the second experiment, we aimed to conduct a more detailed analysis of sentiment by implementing a naive Aspect-Based Sentiment Analysis (ABSA) to track how sentiment toward Poland evolved over time. For each occurrence of a Poland-related token, we extracted a sequence consisting of the token itself along with the five tokens preceding and five following it. These sequences were then fed into the model to predict whether the sentiment was positive or negative. For each month, we calculated the number of positive and negative phrases about Poland within the speeches (Figure 5). Additionally, we computed the ratio of negative to positive sentiment for each quarter (Figure 6. The model used for this analysis was `siebert/sentiment-roberta-large-english`, paired with its dedicated tokenizer fine-tuned for optimal performance.

Next Steps

To improve the analysis, we want to collaborate with experts from the Mieroszewski Center to align anomalies observed in our diagrams with significant historical events. This might help in contextualizing deviations in token frequencies and sentiment trends with real–world events. We also plan to employ more sophisticated methods of Aspect-Based Sentiment Analysis to explore sentiment dynamics specifically related to Poland and

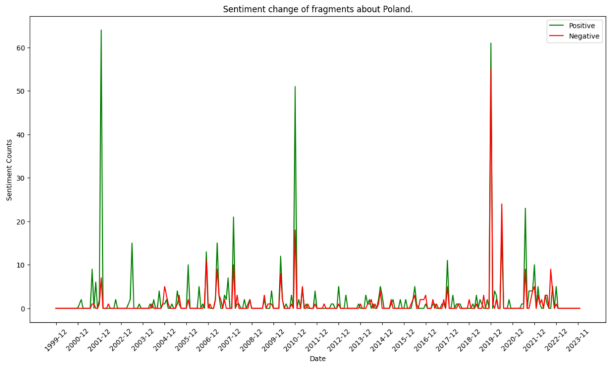

Figure 5: Sentiment evolution in fragments about Poland.

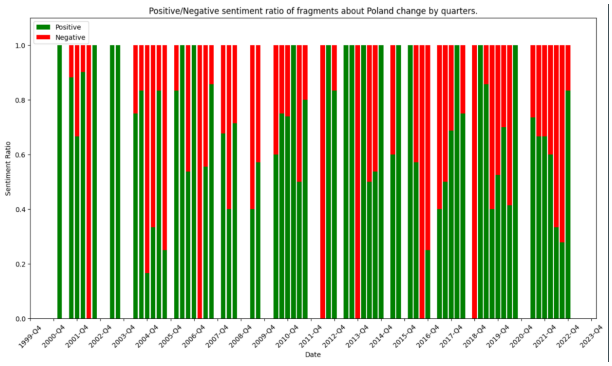

Figure 6: Negative/Positive Sentiment ration evolution in fragments about Poland.

Europe. By analyzing sentiment at a lower level, ABSA can help to discover shifts in tone and message of the speeches. Finally, using sentiment diagrams and historical context, we want to develop predictive models to forecast future events. These models would leverage trends in sentiment and token frequency to identify patterns that could signal the likelihood of upcoming significant events.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Rahul Dudhabaware and Mangala Madankar. 2015. https://doi.org/10.1109/ICCIC.2014.7238427 Review on natural language processing tasks for text documents. *2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014*.

Sean R Eddy. 1996. https://doi.org/10.1016/S0959-

440X(96)80056-X Hidden markov models. *Current Opinion in Structural Biology*, 6(3):361–365.

Azher Ahmed Efat, Asif Atiq, Abrar Shahriar Abeed, Armanul Momin, and Md. Golam Rabiul Alam. 2023. Empoliticon: Nlp and ml based approach for context and emotion classification of political speeches from transcripts. *IEEE access*, 11:1–1.

Xing Fang and Justin Zhijun Zhan. 2015. https://api.semanticscholar.org/CorpusID:16177937 Sentiment analysis using product review data. *Journal of Big Data*, 2:1–14.

Delia Hernandez Farias, José-Miguel Benedí, and Paolo Rosso. 2015. https://doi.org/10.1007/978-3-319-19390-8-38 Applying basic features from sentiment analysis for automatic irony detection. pages 337–344.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification.

Paritosh D. Katre. 2019. Nlp based text analytics and visualization of political speeches. *International journal of recent technology and engineering*, 8(3):8574–8579.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

OpenAI. 2024. https://arxiv.org/abs/2410.21276 Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. https://doi.org/10.18653/v1/N18-1202 Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Wisam A. Qader, Musa M. Ameen, and Bilal I. Ahmed. 2019. https://doi.org/10.1109/IEC47844.2019.8950616 An overview of bag of words;importance, implementation, applications, and challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204.

Alec Radford and Karthik Narasimhan. 2018. https://api.semanticscholar.org/CorpusID:49313245 Improving language understanding by generative pre-training.

Nihar M. Ranjan and Rajesh S. Prasad. 2023. A brief survey of text document classification algorithms and processes. *Journal of Data Mining and Management*, 8(1):6–11.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information systems (Oxford)*, 121:102342.

Charles Sutton and Andrew McCallum. 2010. An introduction to conditional random fields.

Kamal Taha, Paul D. Yoo, Chan Yeun, Dirar Homouz, and Aya Taha. 2024. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Computer science review*, 54:100664.

Izabela Telejko. 2023. Aspect-based sentiment analysis of reviews. Bachelor's thesis, Warsaw University of Technology.

Dimitrios Tsirmpas, Ioannis Gkionis, Georgios Th Papadopoulos, and Ioannis Mademlis. 2024. Neural natural language processing for long texts: A survey on classification and summarization. *Engineering applications of artificial intelligence*, 133:108231.

Shirui Wang, Wenan Zhou, and Chao Jiang. 2020. A survey of word embeddings based on deep learning. *Computing*, 102(3):717–740.

Haowen Xia. 2023. Continuous-bag-of-words and skip-gram for word vector training and text classification. *Journal of physics. Conference series*, 2634(1):12052.