DECODING SIGNALS IN KAZAKHSTAN'S PRESIDENTIAL SPEECHES: INSIGHTS FROM TOPIC AND SENTIMENT ANALYSIS

By

Nurbek Bektursyn

Submitted to
Central European University
Department of Network and Data Science

In partial fulfillment of the requirements for the degree of
Master of Science in Social Data Science

Supervisor: Prof. Mark Wittek
Associated Supervisor: Prof. Ivan Savin

Vienna, Austria

2025

Author's Declaration

I, the undersigned, **Nurbek Bektursyn**, candidate for the MS degree in Social Data Science declare herewith that the present thesis is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright. I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 22 May 2025

Nurbek Bektursyn

Signature

Copyright Notice

[1]

[1] Icon by Font Awesome: https://fontawesome.com/

Abstract

The State of the Nation Address (SONA) is a key policy speech by the President of the Republic of Kazakhstan. It articulates the president's position on the country's central issues and presents the mid-term action plan to manage them. I use structural topic modeling and sentiment analysis to examine the interplay between the content and sentiment of presidential speeches and the macroeconomic state of the country. Using 28 SONAs from the President's official website and multiple macroeconomic indicators such as real GDP and FDI, I determine 9 main topics covering areas from Education & Healthcare to Foreign Affairs. Notably, Industrial Development is the most prevalent topic across all addresses and increased from 0.03% in 1999 to over 42% in 2023. There are also notable differences in the topics discussed by the presidents. Tokayev prioritized the discussion of topics like Reforms & Democracy, while Nazarbayev paid more attention to National Identity. I also uncover that real GDP is of great importance for presidents, having a statistically significant association with the prevalence of most topics. Additionally, sentiment analysis reveals that presidents generally follow a positive-neutral-positive pattern when delivering the addresses and have similar opening and closing remarks.

Acknowledgements

I deeply thank my supervisor, Professor Mark Wittek, for his valuable comments on improving this thesis. His class on the Ethics of Big Data gave me a new perspective on the use and impact of AI.

I also thank my second supervisor, Professor Ivan Savin from the ESCP Business School, for his willingness to collaborate with me, for his guidance throughout the writing of this thesis, and for tolerating my tendency to overlook details.

The thesis is dedicated to three people. First, to my mom, whose unconditional love and belief in me have helped me navigate life's ups and downs. Second, to my dearest friend, Askar Yeslyamgaliyev. Thank you for our late-night conversations, which were full of fun and laughter. They made me feel less homesick. Finally, to my supervisor at the IAEA, Minori Hara, who generously offered me the internship and was flexible with the work arrangements.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Presidential addresses are vital components of political discourse. Their diverse content encompasses security, international relations, social policy, and economic development. The primary aim of presidential addresses is to inform the public, governing bodies, and businesses about the nation's current state and establish key priorities for the forthcoming period. Broadcasted on television or radio, they provide an opportunity for the president to elaborate on the proposed course of action for the mid-term period, thereby strengthening the bond between the president and the nation while demonstrating commitment to the country's unity and interests. The president may also propose new initiatives and reforms deemed necessary. Presidential address is a powerful tool in shaping a nation's political consciousness, allowing the public and political actors to develop a specific political vision (Fairclough, 1989). Using various rhetorical techniques, such as references to religion and history, as well as figurative language, presidential addresses serve not only an informational purpose but also a persuasive one. Thus, the versatile content of presidential addresses makes them a powerful tool for government communication and public sentiment management.

Several studies have investigated the topics emphasized by presidents in various political contexts. In the United States, presidents increasingly discuss the economy, national security, and democracy (Hart et al., 2013; Rule et al., 2015; Wood, 2004; Campbell & Jamieson, 2008; Card et al., 2022). As Wood (2004) notes, contemporary US presidents discuss the economy more frequently than their predecessors. In Europe, the speeches of political leaders reflect the complexities of regional integration (Schumacher et al., 2016; Maerz & Schneider, 2020). Maerz and Schneider (2020) observe that the heads of government in European democracies often discuss the EU crisis, collective memory, and the public sector. Meanwhile, Russian presidents emphasize sovereignty, religion, and patriotism/nationalism in their speeches (Laqueur, 2015; Drozdova & Robinson, 2019; Oleinik, 2023). Drozdova and Robinson (2019) argue that Vladimir Putin's rhetoric places great importance on a centralised and strong state, reinforcing the narratives of stability, freedom, patriotism, and loyalty towards the state.

Kazakhstan presents an intriguing case in this context, as its presidential rhetoric bears certain similarities in some aspects to that of Russian presidents, whilst also reflecting the country's domestic priorities and distinct geopolitical position. Situated in Central Asia and gained independence from the Soviet Union in 1991, Kazakhstan's market-oriented economy and substantial reserves of natural resources, including oil, gas, and gold, have enabled it to emerge as an important regional player. As of 2023, Kazakhstan's GDP amounted to approximately $262 billion, the highest among Central Asian countries (World Bank, 2023). It also managed to attract substantial foreign direct investments (FDI), an essential driver of a country's growth. In 2022, 64% of FDI inflows in Central Asia went to Kazakhstan (United Nations Conference on Trade and Development, 2024). These achievements would not have been possible without a strategic development trajectory developed and executed by the country's leadership. The country's first president (1991-2019), Nursultan Nazarbayev, was one of the key figures responsible for it throughout his tenure. In 1996, he gave his first State of the Nation Address (hereafter, SONA), outlining his vision of the country's future and mission. Since then, it has become an annual tradition. Kazakhstan's current President, Tokayev (2023), articulated SONA's importance as follows:

> *"At this juncture, representatives from all branches of Government convene. We outline key directions for the medium-term, issue specific instructions, and set new objectives. This event breathes new life into the work of Parliament, the Government, and other authorized bodies, playing a pivotal role in the smooth and effective functioning of our state apparatus."*

Therefore, SONA is a critical platform intended to increase the effectiveness of state bodies and a pivotal national occasion that can bring the country together, inspire support for a common goal, and strengthen shared values (Sikanku, 2022). Through such addresses, the president can articulate his vision and main priorities for the upcoming year, expanding the connectedness between the leadership and the population.

Despite being an essential tool for the president to communicate his agenda to the public and various stakeholders, little research has been done on presidential communication in the post-Soviet realm, particularly Kazakhstan (Oleinik, 2023; Lenton & Karibayeva, 2024). Studies involving NLP, such as that conducted by Savin and Teplyakov (2022) on an example of nationwide phone-ins of Russian President Putin, have shown that topics appearing in the addresses are associated with the levels of public support and the state of the economy. In SONA, the President has no direct interaction with the public as in the Russian national phone-ins and, therefore, can deliver the address more concisely and uninterruptedly. This thesis aims to offer a comprehensive and structured analysis of the content of presidential addresses and examine the relationships between content and macroeconomic factors in a new context. It also contributes to the literature by looking at the post-Soviet region and enhances our understanding

of presidential rhetoric.

Presidential addresses in Kazakhstan are worthy of study for several reasons. As the most critical figures in the country, Nursultan Nazarbayev and Kassym-Jomart Tokayev are often viewed as those who possess privileged information and are informed by expert counsel. As a result, their addresses hold immense symbolic value and weight, offering substantial insights to the general public, entrepreneurs, and policymakers. Depending on the content, presidents can influence the country's economy, financial markets, and the public's economic expectations (T. P. Dybowski et al., 2016; Maligkris, 2017; Cinelli et al., 2021). As T. P. Dybowski et al. (2016) demonstrate in their work on presidential tax communication, a positive tone of the president in tax-related statements substantially affects consumer confidence, private consumption, and economic activity. When a crisis hits, presidents are the focal point for everybody to turn to. As such, they have to be involved in using a crisis rhetoric effectively. Presidential crisis rhetoric consists of convincing the public that a crisis exists, communicating relevant details regarding the incident, and calling citizens to support the solution and overcome the challenges faced jointly, as noted in the March 2022 SONA following the January 2022 unrest (Davis & Gardner, 2012). Thus, presidential speeches have great potential to unite citizens in moments of crisis. Finally, Kernell (2007) notes that presidents "go public" to promote themselves and their policies, especially when their popularity declines and as a response to highly negative press coverage. The addresses, therefore, seem to influence the public agenda. For all of these reasons, studying presidential addresses is worthwhile, as it allows us to understand their impact on society as a whole.

This thesis poses the following research questions:

- RQ1: What are the primary topics discussed in SONAs, and how can they be characterised?
- RQ2: How does the prevalence of these topics change over time?
- RQ3: How are macroeconomic indicators like FDI and real GDP associated with the prevalence of the topics?
- RQ4: To what extent does the tone of presidents correlate with macroeconomic indicators?

Methodologically, I make use of Topic Modeling (TM) to identify latent themes (topics) in SONAs. Specifically, I use structural topic modeling that allows the use of textual metadata such as time of the SONA and the associated characteristics of Kazakhstan's economy (like GDP and FDI) to form more interpretable topics. Subsequently, I use sentiment analysis (SA) to classify the presidential addresses as positive, negative, or neutral. There have already

been attempts to measure the impact of business cycles, characterised by economic expansion and contraction periods in the economy, on presidential rhetoric (Hart et al., 2013; Hoffman & Howard, 2010; Wood, 2004). To obtain more accurate sentiment scores, I rely on the sentiment analysis classifiers, specifically fine-tuned for political texts. The need for a fine-tuned model comes from the formal nature of the addresses since they are not so much emotionally driven, like social media comments. Thus, sentiment scores are unlikely to be highly positive or negative. This way, I can gain insight into Kazakhstan's presidents' communication strategies and their evolution during varying economic conditions. Finally, to go beyond a descriptive analysis, I use statistical tests to analyse the interplay between topic prevalences and macroeconomic indicators in the face of real GDP, FDI, and inflation. As described earlier, this particular choice was motivated by the fact that Kazakhstan has attracted substantial FDI since its independence. GDP was chosen because of its importance to politicians, especially during political campaigns, and their call to increase it unconditionally (Van den Bergh, 2009).

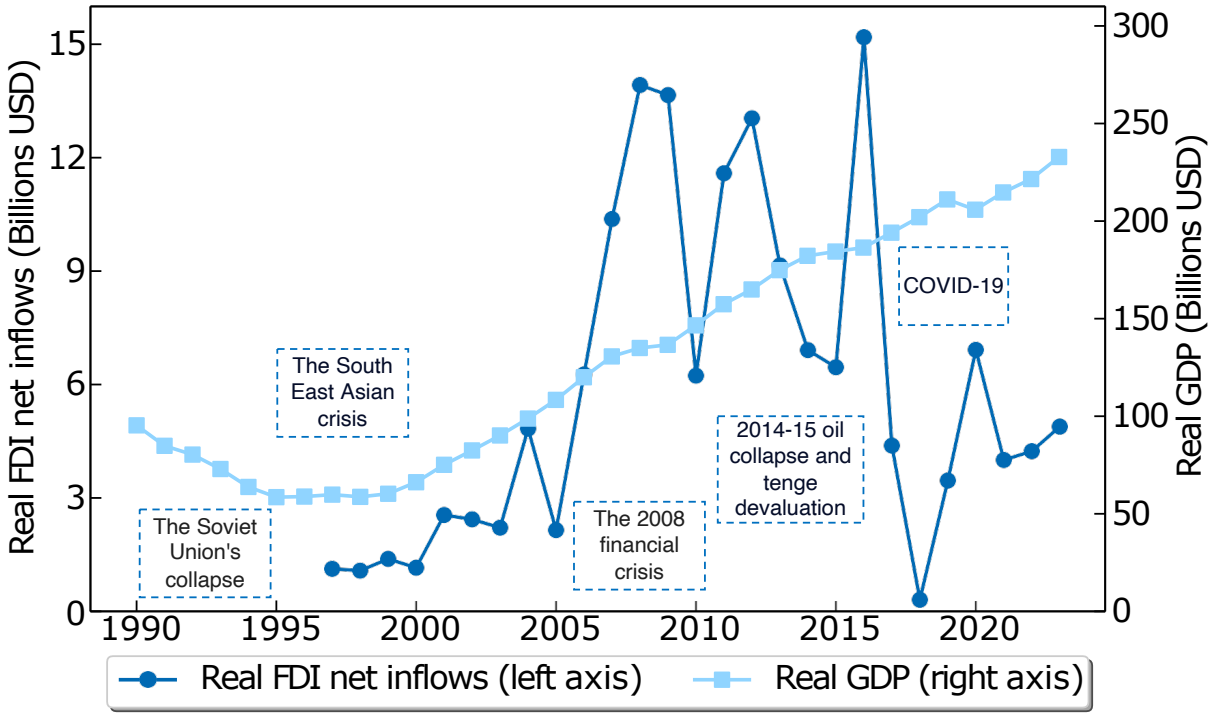

***Figure 1**: Dynamics of Real GDP (1990 - 2023) and Real FDI (1995 - 2023). Source: GDP deflator - Bureau of National Statistics of Agency for Strategic Planning and Reforms of the Republic of Kazakhstan (2023); Real GDP and FDI - World Bank Open Data (2023).*

Figure 1 describes the trends in Kazakhstan's real GDP and real FDI. The World Bank's real GDP data was already adjusted for inflation and currency fluctuations using 2015 constant prices, whereas FDI was adjusted by me using a GDP deflator with the same 2015 base year. The inflow of FDI is in line with the results by Serikkyzy et al. (2024). They found that Kazakhstan's FDI volumes were strongly correlated with the global trends. The country's economy has been negatively affected by several external shocks, like the 1997 South East Asian crisis.

The first years of independence were not so successful for the economy, as the country had to set up its institutions and introduce national currency, the tenge. While real GDP shows a gradual rise throughout the years, the real FDI exhibits a lot of fluctuations, with a peak in 2016, after the 2014-2015 oil prices collapsed because of OPEC's early 2000s strategy of oil supply restrictions (Baffes et al., 2015). Since then, net FDI inflow levels have not yet reached their peak.

To my knowledge, no study has been conducted on measuring the topics and tone of Kazakh presidents as a function of macroeconomic performance. In my thesis, I employ an approach similar to Wood (2004) but with covariates such as real GDP, FDI, inflation, and a binary variable representing whether the address belongs to Tokayev or Nazarbayev. Incorporating these indicators will provide a clearer and more comprehensive understanding of the relationship between presidential rhetoric and macroeconomic performance.

The rest of the thesis is organized as follows: Chapter 2 gives a literature overview, Chapter 3 describes the data and the methods used, Chapter 4 contains the results of the study, while the final Chapter 5 concludes.

Chapter 2

Literature Review

2.1 Inferring Topics of Presidential Addresses

The initial attempts to analyse presidential addresses employed manual qualitative methods. Researchers needed to read speeches closely, sentence by sentence, and rate them to identify recurring topics and themes (Donley & Winter, 1970; Mahoney et al., 1984; Hart et al., 2013). By employing graduate students as raters to assess the inaugural speeches of US presidents for the presence of predefined values such as economy, equality, or freedom, Mahoney et al. (1984) identified the themes that were consistent over time. Although these labour-intensive methods allow for a nuanced understanding of presidential rhetoric, they are costly to apply to a large amount of textual data, while the necessity to use many human coders runs the risk of inconsistent evaluation.

Recent improvements in processing power and computational efficiency of computers have made it increasingly feasible to analyse textual data using methods from natural language processing (NLP), which lies at the intersection of linguistics, machine learning, and computer science. NLP methods for text analysis typically fall into two categories: supervised and unsupervised. The primary distinction lies in the use of labelled data. TM and semantic networks are two common approaches to text analysis that belong to unsupervised learning, as they do not need a prior human annotation. TM is based on the simple and intuitive concept that certain words co-occur together more frequently than they otherwise would when a specific topic is being discussed. Its purpose, then, is to reveal these word patterns in documents and to illustrate how words are distributed across topics, and how topics are represented in different documents (Mohr & Bogdanov, 2013). Blei et al. (2003) introduced the simplest and most popular TM approach called Latent Dirichlet Allocation (LDA). Early studies in TM primarily relied on LDA to explore latent topics in the large text corpora (Griffiths & Steyvers, 2004; Rosen-Zvi et al., 2012). Later, a family of new extensions of LDA has been proposed, like the Correlated Topic Model, Weighted Topic Model, and Structural Topic Model (STM). TM has 20 years of history

and proven past success records in social science research (Dugoua et al., 2022). One of the prominent uses of the Correlated Topic Model is in T. Dybowski and Adämmer (2018) article, where they combined TM with a sentiment analysis. They found that positive news from the president about tax policy positively correlated with the GDP and investment.

Another unsupervised alternative to TM is the semantic networks approach. It is based on the concept that language can be structured as a collection of words and the relations (or absence thereof) among the words. Although it is less popular than TM, scholars still rely on this method due to the lack of necessity to choose a number of topics (Rule et al., 2015; Fuhse et al., 2020). Rule et al. (2015) used this method to analyze the State of the Union addresses from 1790 to 2014.

The supervised approach to TM includes BERT (short for Bidirectional Encoder Representations from Transformers), a pre-training technique developed by Google researchers (Ma et al., 2022). In this approach, the model does not discover hidden topics in the text documents but makes predictions based on a pre-specified set of topics. Specifically, more recent research has employed BERT and the models developed from it, such as BERTopic and RoBERTa (Abercrombie et al., 2019; Ebeling et al., 2023; Mendonça & Figueira, 2024; Aryani et al., 2024; C. Timoneda & Vallejo Vera, 2024). A comparative analysis of BERT, XLM-RoBERTa, and DeBERTa in classifying the multilingual speeches of political leaders by C. Timoneda and Vallejo Vera (2024) showed that XLM-RoBERTa considerably outperforms DeBERTa and BERT in a political science context. The authors noted its advanced capabilities in handling multiple languages and long texts, proposing that RoBERTa's high classification accuracy results from more training data, parameters, and tokens.

Much of the current literature on TM pays particular attention to the speeches of American Presidents (Crockett & Lee, 2012; Ruhl et al., 2018; Kiyama, 2018; Liu & Lei, 2018). Having analyzed almost 4000 presidential direct action documents, Ruhl et al. (2018) found that computational topic modeling could effectively assist in designing and validating topic models of presidential direct action texts. They also highlighted that combining computational and conventional topic modeling, where the former relies on human judgment to read, code, and categorize the documents by topics, and the latter detects topic patterns automatically, will lead to the most secure conclusions about the topics. Although the study's primary target audience seems to be legal researchers, I want to build on their research and explore how the combination of computational and conventional topic modeling will perform on presidential addresses.

2.2 Sentiment Analysis

Numerous studies have investigated the tone of the speeches of American presidents (Card et al., 2022; Hoffmann, 2018; Allen & McAleer, 2019; Liu & Lei, 2018). Hoffmann (2018) examined the 2016 US Presidential Election speeches. Consistent with previous media reports, Donald Trump's language was more negative than that of Hillary Clinton, containing emotions of anger, fear, sadness, and disgust. Allen and McAleer (2019) examined the State of the Union addresses of Presidents Obama and Trump, comparing them to Hitler's Proclamation in Berlin from 1933. The investigation results have shown that the emotional tone of the two US presidents turned out to be more positive than Hitler's, with the dominance of emotions such as "anticipation" and "joy". Furthermore, all speakers preferred to end their speeches on a positive note. The authors attribute both findings to the structure of political speeches intended to win the approval of listeners. I am interested to see if there are similar speech patterns in the case of Kazakh presidents, given the differences in the background of the presidents. Having an engineering background, Nazarbayev quickly climbed the Communist Party ladder during the Soviet Era, greatly shaping his Soviet-style, centralised governance structure. In contrast, coming from an international relations and diplomatic background, Tokayev's governance style is more procedural and globally oriented.

Beyond the U.S. context, sentiment analysis has been conducted on the SONAs of Philippine Presidents (Miranda & Bringula, 2021), Finnish presidents' New Year's parliamentary speeches and messages (Kujanen et al., 2025), and Jordan's president's speeches (Salah, 2020). By comparing sentiment scores from the addresses of the 13 past Philippine Presidents between 1935 and 2016, Miranda and Bringula (2021) identified not only which president had the highest and lowest sentiment scores but also the changes in sentiments from their first to last delivered SONAs. They highlight that external factors influenced presidents' sentiment scores at the moment of delivering an address, such as the Martial Law period. In addition, the authors discovered a consistent cyclical pattern: the emotional tone becomes less positive in the second year of the term, and more positive in the third year compared to the second. This happens because presidents initially hope for favourable changes that their programs will bring, but in the following year, the administration tends to report the challenges faced. Kujanen et al. (2025) reported that the speeches of Finnish presidents are generally more positive than negative, and external factors, such as inflation and unemployment, had no statistically significant effect on the sentiment and content of the speeches. The authors interpret this as the need for presidents to maintain "statespersonlike" behavior, projecting socioeconomic and political stability.

The relationship between presidential rhetoric and a country's macroeconomic performance is not entirely understood. It is a well-known phenomenon that the public tends to evaluate the

president based on economic performance. If the economy is doing well, the president is perceived as performing well, and vice versa. As a result, presidents often emphasize and showcase their successful policies in their speeches as much as possible during times of economic prosperity to enhance their perceived competence. However, during an economic downturn, they may still use positive rhetoric strategically to instill confidence and encourage spending, mitigating the adverse effects of the crisis. Using empirical data from the Public Papers of American Presidents, Wood (2004) found that each 1% increase in the inflation rate and the federal deficit relative to GDP results in 2 and 5 additional statistically significant presidential comments about inflation and the federal deficit per month, respectively. It is noteworthy that presidents made more remarks about the inflation and deficit when they were high, and these remarks tend to be generally positive. As the economy's rhetorical leaders, the presidents' public statements seem to influence people's economic behaviour. Research suggests that the positive remarks of the presidents on the economy alter the people's spending behaviour, which in turn affects the macroeconomic performance (Wood et al., 2005; T. P. Dybowski et al., 2016).

2.3 Approaches to Analyzing Presidential Addresses

Lately, there has been an increase in the volume of literature about the presidents' speeches from post-Soviet countries (Savin & Teplyakov, 2022; Oleinik, 2023; Lenton & Karibayeva, 2024). Lenton and Karibayeva (2024) examined 152 post-Soviet leaders' New Year addresses through a manual content analysis and compellingly demonstrated essential insights into their regime legitimation and political communication styles. They identified that the presidents favour mentioning national identity in their addresses, reinterpreting historical narratives, and outlining a future vision.

Relying on a dictionary-based approach and text mining software, Oleinik (2023) offers a comparative analysis of political discourses between North American (the United States and Canada) and post-Soviet countries (Russia, Ukraine, and Kazakhstan) through Michel Foucault's concept of governmentality. Foucault distinguishes three types of governmentalities: security, sovereignty, and discipline. Foucault explains that sovereignty operates within defined territorial borders, discipline targets and regulates individuals, and security functions at a broader population level (Foucault, 2007). One of his findings, obtained after getting the word frequencies from WordStat and QDA Miner, was that the concepts of "security", "market", and "freedom" were mentioned the most in the presidential speeches and their analogues of the post-Soviet leaders compared to North American ones.

Lastly, based on the 1938 responses from President Putin during his national phone-in meetings, Savin and Teplyakov (2022) identified 16 major topics, including but not limited to various areas like healthcare, social security, and foreign affairs. The topics were identified with

the help of STM. They also demonstrated that Putin adapts the topics he raises depending on the public support levels, and their prevalence is strongly correlated with some of the country's macroeconomic indicators.

Chapter 3

Data and Methods

3.1 Data Collection

The primary data source is the official website of the President of the Republic of Kazakhstan (akorda.kz), which contains all the SONAs spanning almost 30 years in Kazakh, Russian, and English languages. The addresses were scraped using BeautifulSoup, a popular Python package for web scraping. I encountered several language-related issues while working on the addresses. For 1999, 2002, 2007, 2009, and 2010, the addresses in English were completely missing. For 2003 and 2004, the addresses presented in English were summaries of the original addresses delivered in a mix of Kazakh and Russian languages. For 2009 and 2010, the addresses included a mix of Kazakh and Russian languages. Two leading machine translation tools, such as Google Translate and DeepL Pro, were utilized to tackle these issues, which are common in social sciences (Mate et al., 2023; Licht et al., 2024). For the years with missing addresses in English, but available in Russian, I translated all the texts from Russian to English using DeepL Pro because it offers better translation quality than Google Translate (Hidalgo-Ternero, 2021; Aguilar, 2023). For the summarized transcripts, the complete addresses were available in Russian, so the procedure remained the same. For the addresses with a mixture of Russian and Kazakh, I first translated some sections, typically the beginning, from Kazakh to Russian using Google Translate because DeepL Pro does not support Kazakh. I then translated the resulting texts to English through DeepL Pro. As a result, the final dataset contained 28 SONAs, with seven belonging to President Tokayev and the rest to President Nazarbayev. Given my working proficiency in these languages, I was able to verify the accuracy and reliability of the translations.

To measure the macroeconomic situation in the country, I relied on several economic covariates: real FDI net inflows, real GDP, and inflation. Refer to Table A1 in the Appendix for a detailed description of the covariates, their calculation formulas, and corresponding sources.

The left chart in Figure A1 illustrates the distribution of addresses. Presidents traditionally deliver SONAs once a year, yet they were not delivered in 2013, 2015, and 2016. Nazarbayev had a varying schedule for delivering addresses throughout the year, giving SONA anywhere between September and April. In contrast, Tokayev has a more consistent approach, aiming to deliver his addresses in September, which is considered the beginning of a political season. Occasionally, presidents delivered SONAs twice a year, like in 2012, 2018, and 2022. The deviation in 2012 happened because Nazarbayev wanted to introduce a new "Kazakhstan-2050" strategy to replace the old "Kazakhstan-2030" strategy of 1997. In a new strategy, Nazarbayev set a goal of having Kazakhstan ranked among the top 30 developed countries by 2050 (Aitzhanova et al., 2014). In 2018, the deviation from the norm was due to the upcoming presidential election campaign. The SONA was full of promises of future social welfare benefits. In 2022, the deviation occurred for similar reasons, as Tokayev proposed initiating an unscheduled presidential election in the autumn. Thus, presidents deliver the SONA multiple times if they wish to initiate an early election or introduce a new strategic initiative.

3.2 Data Cleaning

Data cleaning is an essential step for obtaining high-quality and relevant topics, reducing the feature space. It consists of the following steps:

1. Tokenization. The text is broken down into distinct small units called tokens to ease the subsequent analysis, and capital letters were converted into lowercase letters (e.g., toponyms and proper names).

2. Lemmatization. This step converts words to their base or dictionary form, referred to as a "lemma." Unlike stemming, another popular cleaning technique in NLP, lemmatization gives back the vocabulary form of the word. Since we require semantically meaningful words for topic modeling, lemmatization is preferred.

3. Removal of punctuation, numbers, and special characters.

4. Removal of stop words. In addition to common stopwords in the libraries (e.g., "a", "the"), I extended the list with some custom words (e.g., "kazakhstani", "president", "congratulation"). For example, although an ordinary opening sentence of an address like "Congratulations to all of you on the opening of the regular session" contains rhetorically important words, they are not informative for identifying the underlying topics as they often occur across the addresses. So, removing them improves topic coherence.

5. Bi-grams. Individual words that predominantly occurred together were connected with underscores (e.g., "european_union") and selected using the Normalized (Pointwise) Mutual Information score developed by Bouma (2009).

6. Removal of rare words. The words that appeared less than 3 times across the addresses were removed.

The right chart in Figure A1 shows the lengths of addresses before and after pre-processing. As of SONA's length, before text processing, the longest SONA was the very first in 1997, with a total word count of 17869, while the shortest was in 1999 and contained 1891 words. The average word count of original addresses is 7152 words. After text processing, the longest SONA was from 2012, with a total word count of 7914, while the shortest remained the same and included 698 words. The average word count of processed addresses is 3202 words.

3.3 Topic Modeling

The topics of the presidential addresses are obtained with the assistance of the method called Topic Modeling (TM). The technique belongs to the family of Natural Language Processing methods, allowing for the automatic detection of latent topics within a collection of documents. It is a non-supervised machine learning technique, implying one does not need to tell the computer how many topics and what kind of topics we are looking for. Instead, we infer this information from the textual data directly. In simple words, TM clusters words into topics based on their co-occurrence across multiple documents. The more often the words co-occur, the more likely they will be attributed to the same topic. For instance, if words like "housing", "pension", and "wage" appear in the topic labeled as "Social Welfare," it indicates that the words appeared more frequently and exclusively alongside other words related to this topic.

Formally speaking, it employs a Bayesian inference model, in which each document is linked to a probabilistic distribution of topics, and each topic is represented as a probability distribution of words. The model is built on the assumption that each word in the documents is generated through multiple iterations of a two-step process: first, each document has a topic distribution, and one topic is chosen at random from it. Second, each topic has a word distribution, from which a word is randomly selected for the topic chosen in the first step. Put differently, it works by backtracking from the documents to infer the topics that are assumed to create them. Its advantage over a simple keyword counting is viewing words not in isolation but in the context of other words, giving a richer and structured understanding of the content. As a result, each SONA is represented as a mix of topics, each appearing in varying degrees.

For analyzing addresses, I use an algorithm known as Structural Topic Modeling (STM), developed by Roberts et al. (2014), because it allows for including metadata about the addresses to form more coherent and interpretable topics. In my situation, metadata involves years of addresses and real GDP, FDI, and inflation. To reduce multicollinearity, I excluded the continuous measure of time (since SONAs are occasionally delivered multiple times in a year)

and unemployment, because they had strong correlations with real GDP. Therefore, rather than assuming topical prevalence (i.e., the extent to which a particular topic represents an individual document) and the specific words used to describe the topic to be the same, I treat this information as updated Bayesian priors. Additionally, one of the advantages of STM is in allowing the topics to be correlated, giving a realistic representation of addresses. This is consistent with Blei et al. (2003) findings that people are more likely to associate specific topics with one another. For instance, individuals discussing the market economy often address global trends and production levels.

Since my initial dataset included only 28 documents, applying STM to them directly would have resulted in broad topics that are difficult to assign to a specific subject. This is because TM treats each document, no matter how long it is, as a single observation. To mitigate this issue, in line with Savin et al. (2025), I split the text into 300-word chunks. In the computational linguistics literature, it is one of the common approaches to split long documents into shorter texts (Devlin et al., 2019). This resulted in 318 texts in total, which have been used to train STM.

One of the challenging tasks in TM is to identify the optimal number of topics (k). The primary aim is to choose such a number that maximizes the held-out log-likelihood, the exclusivity of topics, and their semantic coherence. The first metric indicates how accurately the model predicts words from a sample that was not included (held-out) in the model's training phase. The second metric evaluates the likelihood of observing a topic based on the words, particularly whether the most frequent words do not overlap much. The third metric measures how often the words that belong to the same topic appear together in responses. Figure A2 illustrates how model performance varies across these three metrics depending on the number of topics, which ranges from 3 to 20. I selected 9 topics since they achieved the highest predictive performance with reasonable exclusivity and topic coherence.

3.4 Sentiment Analysis

For sentiment analysis, I utilized VADER and the Multilingual Parliament sentiment regression model XLM-R-ParlaSent, a fine-tuned XLM-R-Parla model from Hugging Face. This model is a domain-adapted version of XLM-RoBERTa-large. One advantage is that VADER and XLM-R-ParlaSent require minimal preprocessing. The texts were cleaned of numbers, newline characters, and bullet points to ensure text consistency, as they provide little meaning. At the same time, the case was lowered, intentionally preserving stopwords and punctuation, as sentiment analysis tools can handle those efficiently.

VADER is a popular lexicon-based sentiment analysis method that uses a predefined list of words with corresponding sentiment scores developed by Hutto and Gilbert (2014). This choice is based on its ability to account for intensifying adverbs and adjectives like "very well", which will be rated higher than just "well". In addition, it analyzes words within a three-word context window, which allows for accounting for contrastive conjunctions like "but," "although," and "though" and adjusting the weight of sentences accordingly (Bestvater & Monroe, 2023). For a given sentence, it outputs four scores: the positive, neutral, negative, and compound. The compound is the most popular among those to use, which adds up all the valence scores of words in a sentence and applies adjustment rules, scaling it between -1 (the most negative) and +1 (the most positive). Yet, one of the limitations of the lexicon-based sentiment analysis method is its inability to understand the context. For example, a sentence from the 2018 SONA states: "Poverty declined 13-fold, and the unemployment rate fell to 4.9%." VADER will classify it as a negative sentence due to negative words like poverty, decline, and unemployment, even when it is clearly positive. To account for this limitation, I complement VADER with the XLM-R-ParlaSent model developed by Mochtak et al. (2024).

The XLM-R-ParlaSent model is a deep learning model that understands the context and is based on the texts of parliamentary proceedings, greatly improving the classifier's accuracy. Its parent model, XLM-R-Parla, was pre-trained for 8000 steps, where each step was based on analyzing about half a million tokens (1024 sequences * 512 tokens). It was further pre-trained on a combined corpus of EuroParl (Koehn, 2011) and ParlaMint 3.0 (Erjavec et al., 2023) datasets. Together, they cover 30 languages and include approximately 1.7 billion words of text (Mochtak et al., 2024). To improve its classification capabilities, Mochtak et al. (2024) fine-tuned the model using the ParlaSent dataset, which included parliamentary proceedings from 7 countries, annotated with 6 different sentiment labels (see Table 2). The model works as follows: It receives the sentence as input, converts it into tokens, runs it through the XLM-RoBERTa model, and provides the predicted score and the corresponding label for the input. For the predicted scores, the values typically range between 0 and 5, where 0 is negative and 5 is positive. Since the model was set up as a regressor, predictions can occasionally fall outside this range (i.e., below 0 or above 5).

Chapter 4

Results

4.1 Main Topics and Macroeconomic Indicators

Table 1 provides an overview of topic labels and the top 10 words based on their frequency and exclusivity. The words are arranged in decreasing order based on their FRequency and EXclusivity (FREX) scores. For example, the word "tax" in the third topic related to Economic management appears more frequently and is more specific to that topic than the word "regulation." Topic labels (column 2 of Table 1) are based on my interpretation following the model's output. After a thorough review of each topic's most frequent and exclusive words and their share in the addresses, the labels were chosen in such a way as to reflect the content of the addresses clearly and concisely.

#	Topic label	Top 10 words by usage frequency and exclusivity	Overall topic proportion	SONA with the highest prevalence of that topic
5	Industrial Development	agricultural, electric, processing, plant, gas, water, agro, industrial, transport, agriculture	15.48%	September 2023
8	Legislation & Justice	executive, judicial, court, civil, party, local, election, judge, criminal, servant	14.46%	March 2022
7	Economic Management	tax, bank, business, company, financial, entrepreneur, holding, enterprise, economy, regulation	13.75%	March 2006
4	Social Welfare	housing, pension, wage, billion, salary, disabled, million, average, payment, income	13.60%	March 2004
6	National Identity	crisis, join, global, challenge, nation, harmony, thanks, achievement, homeland, success	12.30%	March 2009
1	Education & Healthcare	education, medical, educational, school, university, teacher, language, healthcare, sport, profession	9.67%	January 2018
9	Foreign Affairs	military, extremism, threat, terrorism, treaty, security, drug, religious, border, cooperation	9.59%	September 1999
3	Reforms & Democracy	woman, event, democracy, fair, violence, minister, sacred, democratization, majilis, prime	7.05%	September 1998
2	Development Strategy	perspective, thereof, though, side, settle, realization, afford, hardly, suffer, utilization	4.08%	October 1997

***Table 1**: Overview of topics (the topics are ordered in decreasing order of their prevalence).*

To better showcase the words' frequency and exclusivity and go beyond the aggregated FREX measure, word clouds in Figure A3 were constructed. Here, font darkness indicates exclusivity, and font size indicates frequency. In the third topic, for example, one can observe that the word "political" is more exclusive to this topic than the word "state", meaning the latter word tends to occur more frequently in other topics.

I proceed to examine topic prevalence, which is defined as the proportion of text in each SONA assigned to a particular topic in the fourth column of Table 1. The most frequently occurring topic (T5: Industrial Development) is nearly four times more likely to be discussed than the least common one (T2: Development Strategy). This is likely due to Kazakhstan's heavy reliance on oil and gas. Therefore, discussions on industrial development are the president's attempts to promote economic diversification. It is also worth noting that topics related to domestic governance (T1, 4-8) have the highest prevalence, whereas forward-looking and external engagement topics (T2-3, 9) have the lowest. This is logical, as SONAs mainly discuss internal issues of the country. The last column of the same table showcases the time of the SONA where each topic was the most prevalent. For example, in SONA that was delivered in January 2018, the dominant topic was T1 about Education & Healthcare.

I now turn to how the share of topics evolved over the time. Figure A4 summarizes changes in the proportions of topics, highlighting which topics gained popularity and which did not.

I estimated a linear model for each topic (indexed by i) that was used to construct the STM model to identify any time trends:

$$\text{Topic Prevalence}_i \sim \text{Constant}_i + \text{Year} + \text{Residual}_i$$

T2 on the development strategy, T3 on reforms and democracy, and T9 on foreign affairs were most prevalent from 1997 to 1999. However, over time, they decreased radically, showing statistically significant declines over time at 5% for Topic 2 and 1% for Topic 9. T5 on industrial development saw the largest increase among all topics, starting from approximately 0.03% in 1999 to over 42% in 2023. Also, topic 1 on education & healthcare, topic 5 on industrial development, and topic 8 on legislation & justice show statistically significant increases in their proportions throughout the years at 0.1%, 1%, and 10%, respectively. The increase in these topics can be attributed to the growing public demand and concerns about changes in these areas as society has evolved. Presidents, therefore, addressed these topics more frequently to show their alignment with people's problems. Although these topics were important in the early years of the state's formation, presidents, as time progressed, decided to prioritize topics directly impacting people's lives.

I consider several macroeconomic indicators: real GDP, net FDI inflows, and inflation as covariates to build a better topic model (see Section 3.3). Real GDP is calculated by dividing GDP in current terms by the GDP deflator and multiplying the result by 100. Net FDI inflows represent the flow of direct investment equity into the country. It comprises reinvested earnings, equity capital, and other forms of capital. A direct investment relationship is established when the investor owns at least 10% of the voting shares in the foreign enterprise. Year-over-year inflation represents the average change in prices paid by consumers and compares the current month's inflation to the previous year's. I selected the months during which the addresses were delivered to better reflect the macroeconomic situation in the country.

Figure 2 presents the results of the effect of macroeconomic indicators on the share of topics. To establish the relationship, I estimated the following linear model using a set of economic and political predictors:

$$\text{Topic Prevalence}_i \sim \text{Constant}_i + \text{GDPreal} + \text{FDI} + \text{Inflation\_yoy} + \text{Tokayev} + \text{Residual}_i$$

The political predictor here is the 'Tokayev' variable, which was coded as 1 if Tokayev delivered the address, and 0 if delivered by Nazarbayev. Out of 27 (9 topics * 3 macroeconomic indicators) regression estimates, only 9 show a statistically significant relationship between shares of topics and macroeconomic indicators. I also found that most topics (6 out of 9) are significantly associated with real GDP. Topic 1 (Education & Healthcare) and Topic 6 (National

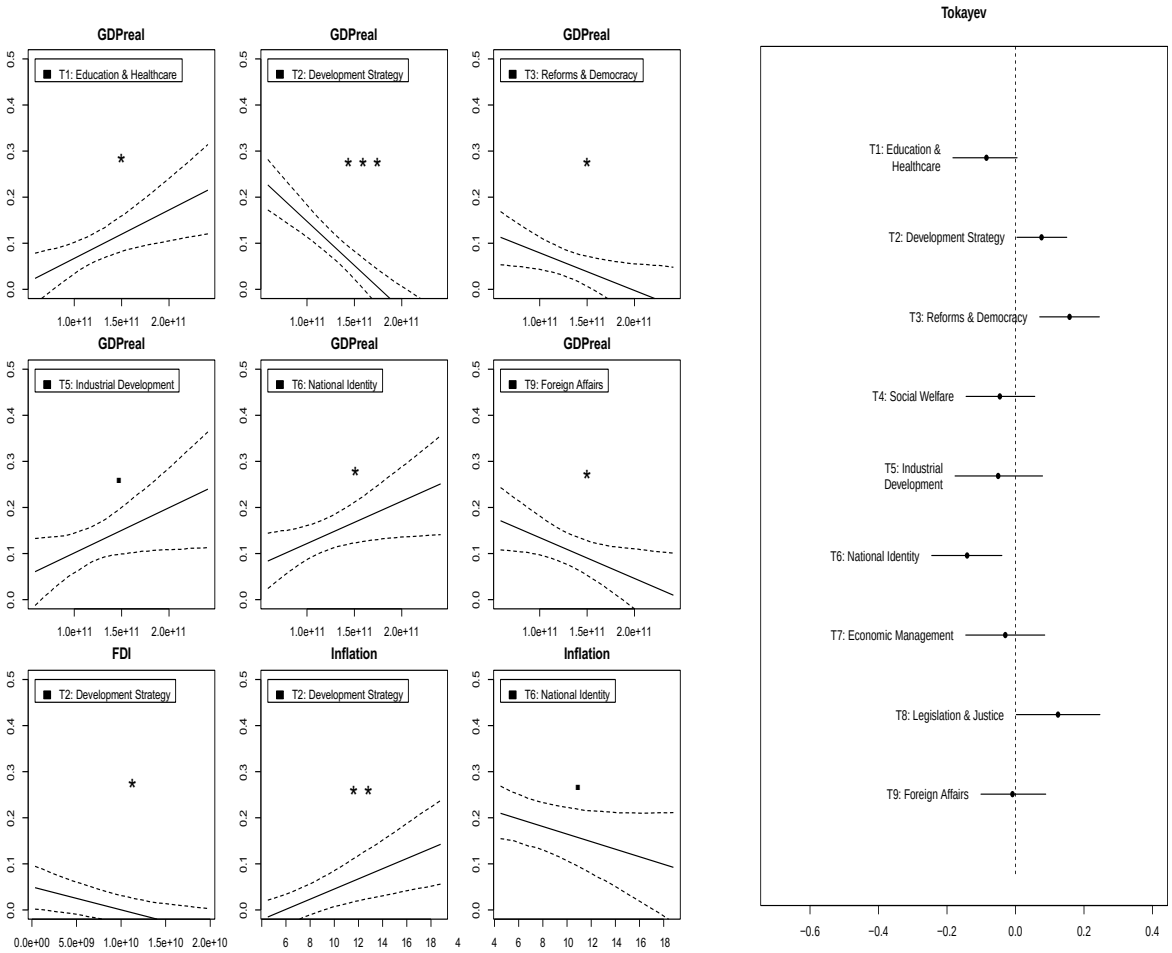

Figure 2*: The effect of covariates on a set of topics. •, *,**, and *** mean 10%, 5%, 1%, and 0.1% significance levels, respectively. Note: The left figure shows only statistically significant regression results.*

Identity) have a positive relationship with real GDP, meaning if this macroeconomic indicator increases, the presidents emphasize these topics in their addresses. In contrast, Topic 2 (Development Strategy), Topic 3 (Reforms & Democracy), and Topic 9 (Foreign Affairs) have a negative relationship with real GDP, meaning if this macroeconomic indicator increases, the presidents put less emphasis on these topics in their addresses. In addition to real GDP, it is also worth considering the effect of FDI and inflation on the share of topics. Both FDI and inflation seem to have a significant effect only on Topic 2 (Development Strategy). Specifically, FDI has a negative effect, while inflation has a positive effect.

The right graph of Figure 2 shows the estimated mean differences in topic proportions between Nazarbayev and Tokayev. Positive values indicate a greater emphasis by Tokayev, while negative values suggest that Nazarbayev raised the topic more frequently. Tokayev is more likely to discuss Topic 2 (Development Strategy), Topic 3 (Reforms & Democracy), and

Topic 8 (Legislation & Justice), while both equally stress the remaining topics, except for Topic 6 (National Identity), which is discussed mostly by Nazarbayev. The pattern is likely influenced by the period during which the presidents took office: Nazarbayev came right after the collapse of the Soviet Union, and the ethnically diverse country was in need of social cohesion, forcing him to place more emphasis on National Identity. As a successor to Nazarbayev, Tokayev committed to advancing the reforms and foundations established by Nazarbayev to build a "New Kazakhstan", a visionary concept developed by him in making the state administration more effective and just.

4.2 The Sentiment Arc of Presidential Speeches

Table 2 provides an overview of the numerical outputs of XLM-R-ParlaSent and the corresponding labels. It goes beyond the usual positive, neutral, and negative sentiment labels, distinguishing between pure and ambiguous sentiments. This annotation schema allows for a more nuanced understanding of sentiments.

Score range	Sentiment category
5	Positive
4	Mixed Positive
3	Neutral Positive
2	Neutral Negative
1	Mixed Negative
0	Negative

***Table 2**: XLM-R-ParlaSent's sentiment labels and numerical outputs.*

Table A2 presents the sentences with the minimum (most negative) and maximum (most positive) sentiment scores measured with VADER and XLM-R-ParlaSent. VADER emphasizes sentiment-laden words and phrases, rating sentences with the highest number of overly emotional words the highest. For example, a sentence from the 2018 SONA, "Recent tragic events have also revealed the problem of poaching, as a most dangerous form of organised crime" was rated the lowest by VADER because of the abundance of negative words like "tragic", "problem", "poaching", "dangerous", and "crime". In contrast, XLM-R-ParlaSent is more sophisticated and captures a more nuanced contextual meaning. A sentence from the same year, "The foundations of the global security system and international trade rules that seemed unshakable are now crumbling," was rated the lowest by XLM-R-ParlaSent despite containing fewer overly negative words due to the contextual understanding of the implicit negativity of "unshakable … now crumbling", going beyond surface-level words.

Figure A5 in the Appendix shows the emotional valence, measured by VADER's compound

score, of Nazarbayev and Tokayev's first and last speeches, as of 2025. In line with Mah and Song (2024), a loess (local weighted regression) smoothing was applied to remove the noise and establish the overall trend. To make the texts comparable, narrative time was normalized to represent the sentence's relative position in the speech, scaling it between 0 and 1. Most of the sentiments are clustered around 0, which is the result of using a compound score that sums the valence of each word in the sentence. According to the loess smoothing, the overall sentiment in the speech is neutral. The averages of compound sentiment scores for Nazarbayev ranged between -0.000327 (1999) and 0.32 (2007), while for Tokayev, they were between 0.11 (2020) and 0.21 (2023). As a result, Tokayev had a narrower sentiment range.

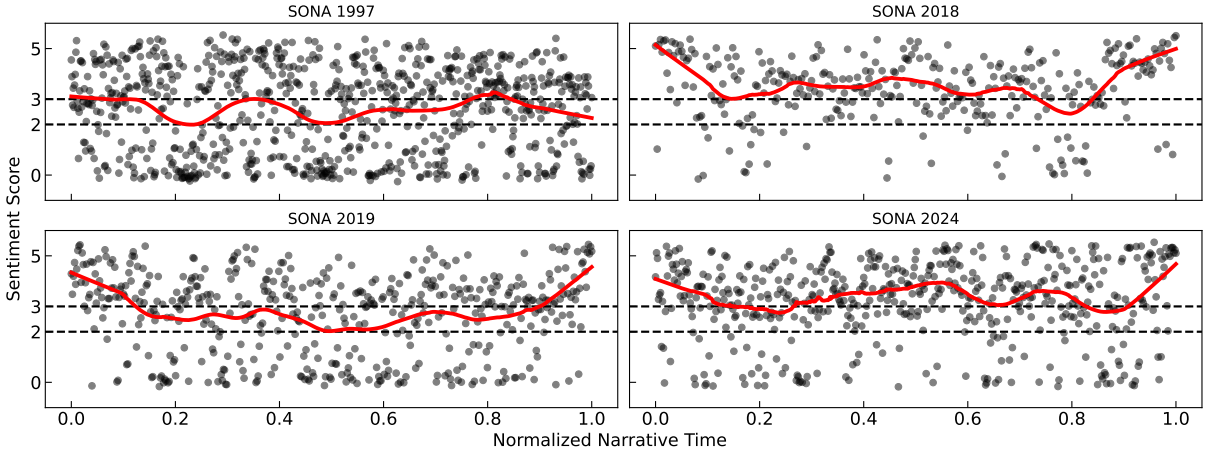

***Figure 3**: The emotional valence of Nazarbayev and Tokayev's first and last SONAs (as of 2025) measured using XLM-R-ParlaSent. Sentiment scale: 0 = Negative, 1 = Mixed Negative, 2 = Neutral Negative, 3 = Neutral Positive, 4 = Mixed Positive, 5 = Positive. The red line represents the LOESS fit. Each dot represents a sentence's sentiment. The time of an address was normalized by representing it as the relative position of sentences, scaled between 0 (start of an address) and 1 (end of an address).*

Figure 3 depicts emotional valence, measured using a state-of-the-art XLM-R-ParlaSent model. Due to its more diverse output range of predictions (between 0 and 5), the sentiment scores are not clustered around a particular value like they used to be with VADER. Nazarbayev's average sentiment scores, as determined by the XLM-R-ParlaSent model, ranged between 1.80 (1999) and 3.67 (2010), while Tokayev's ranged between 2.71 (2020) and 3.22 (2024). Consistent with VADER, Tokayev maintained a narrower emotional range. The overall results suggest that presidents typically start their speeches on a positive note, followed by neutral, and then end on a positive note again. Specifically, Tokayev prefers to begin with "Dear Compatriots" or "Dear Members of Parliament!", while Nazarbayev favours beginning with "Dear Kazakh citizens!" or "Dear people of Kazakhstan!". These greetings are not only similar in meaning, but they both aim to establish a sense of collective consciousness with the listeners. After that, both discuss the past year's accomplishments, usually in a positive light. Tokayev ends his addresses with a variation of "I wish all of you well-being and success", "May our

people be prosperous!", and "The bright future of our sacred homeland is in our hands!". Apparently, Tokayev acknowledges that citizens are not passive observers but active participants in implementing the policies. Nazarbayev concludes similarly with an emphasis on transformation and trust, evident from "I believe in you. I believe that this historic chance will not be lost", "We will turn Kazakhstan into a more prosperous country for our descendants!", and "I am confident that we shall justify the people's trust and reach the goals we have set for ourselves." However, there are some interesting nuances. In 1998 and 2000, Nazarbayev started his speech in a neutral negative tone. In 1998, for example, SONA began with a discussion of instability in Russia, market crisis in Asia, nuclear tensions between India and Pakistan over Kashmir, and regional challenges affecting Kazakhstan. In 2000, SONA commenced with a reflective and somber tone, emphasizing the irreversibility of the past and a 10-year journey filled with achievements and mistakes. In 1999 and 2001, noticeable dips occurred in the mid-speech, where the president discussed historical tragedies experienced by Kazakhstan and criticized the legislative sphere for applying laws too harshly, respectively.

Figure 4 illustrates the proportion of positive sentences in the addresses over the years. The XLM-R-ParlaSent's sentence predictions were classified as negative if below 1.5, neutral if between 1.5 and less than 3.5, and positive if 3.5 or higher. This was followed by the calculation of the ratio of positive, neutral, and negative sentences. There is a noticeable increase in the proportion of positive sentences between 2000 and 2011, and a decline thereafter. The peak in the proportion of positive sentences (approximately 60%) in 2011 is attributed to a significant milestone: the 20th anniversary of Independence. The anniversary called for a highly positive tone, making the president Nazarbayev highlight the achievements made since 1991. The fewest proportion of positive sentences (approximately 27%) occurred in 1999 due to Nazarbayev's discussion of the challenges and demands of the 21st century.

Figure A6 summarizes the correlation between the share of positive, neutral, and negative sentences and the proportions of topics. Among the nine topics, Topic 1 (Education & healthcare), T5 (Industrial development), T7 (Economic management), and Topic 6 (National identity) are more likely to be discussed positively, while Topic 9 (Foreign affairs) tends to be addressed more negatively. For example, in 2018, for Topic 1, Nazarbayev stated that "It is necessary to strengthen the retraining of teachers, to attract foreign managers to universities, and to open campuses of world universities", and in 1998, for Topic 6, he assured that "With this program, Kazakhstan will demonstrate to our own business community, to our trading partners and to the world's financial institutions and analysts that we have the will and the ability to act vigorously in times of economic crisis." Given the multicultural society and the importance of economic messaging, these areas are critical for the well-being of any nation, and promoting them strengthens national pride and unity. In contrast, Topic 9 is discussed more critically, due to concerns about extremism and an unstable geopolitical situation. For instance, in 2021,

Tokayev reported, ”The situation in Afghanistan and the general growth of global tensions has put before us the task of rebooting the military-industrial sector and the Military Doctrine.” Tokayev perceives the international environment here as deteriorating and advocates for the need to be prepared for potential conflict.

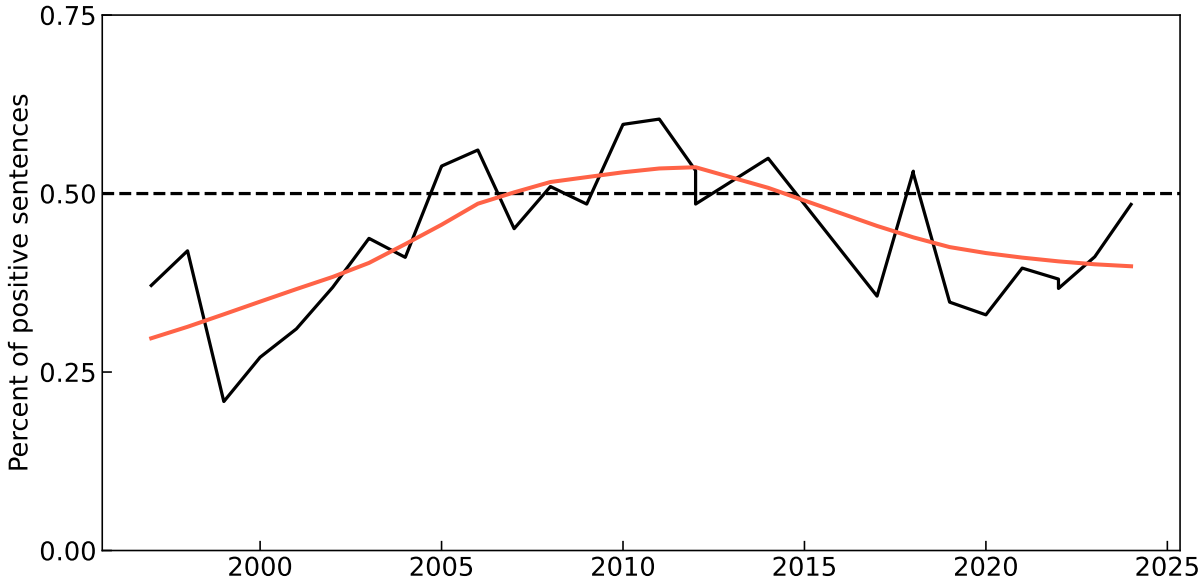

***Figure 4**: Proportion of positive sentences in SONAs over the years. The red line is a LOESS fit.*

Figure A7 depicts the correlations between the share of positive, neutral, and negative sentences and the macroeconomic indicators. FDI shows a statistically significant positive correlation (at the 1% level) with the positive share of statements In SONAs and a statistically significant negative correlation (at the 5% level) with the negative share, suggesting that FDI tends to be discussed positively. As an example, in 2023, Tokayev reported,“... we attracted a record-breaking $28 billion in foreign direct investment.” Furthermore, real GDP has a statistically significant negative correlation with the share of negative sentences (at the 5% level) only, meaning that the increase in real GDP is associated with presidents’ sentiments becoming less negative. For example, in 2010, Nazarbayev announced, “As early as in 2008, in comparison with the year 2000, two years earlier than planned, we realised the volume of the republic’s GDP and effectively fulfilled the social obligations of the state.” While year-over-year inflation is not significantly correlated with the sentiment scores of SONAs, this result contrasts Wood (2004) findings that rising inflation causes presidents to speak more positively about the economy. One reason for this could be that in the context of Kazakhstan, a transition economy with relatively high inflation rates compared to those analyzed by Wood (2004), some moderate inflation peaks are not so critical for the economy or the public as long as the economy is growing in real terms and unemployment is maintained at low levels.

Chapter 5

Conclusion

This thesis applied topic modeling and sentiment analysis to 28 State of the Nation Addresses (SONA) in Kazakhstan from 1997 to 2024. The SONA is crucial since it affects the general population, businesses, and civil servants. For example, socioeconomic initiatives (e.g., an increase in wages and pension savings) are of particular interest to citizens. Thus, it is crucial to analyze it, particularly regarding which topics appear, their proportions, the factors that influence the emergence of specific topics, and how presidents communicate them.

Firstly, I analyzed SONAs using structural topic modeling and document-meta data, which included macroeconomic indicators such as real GDP, FDI, and year-over-year inflation. After considering held-out log-likelihood, the exclusivity of topics, and their semantic coherence, I chose 9 topics circulating across SONAs. Presidents mainly discuss topics related to domestic socioeconomic processes: Industrial Development, Legislation & Justice, and Economic Management. In the late 1990s, Development Strategy, Reforms & Democracy, and Foreign Affairs were the topics that presidents prioritized the most. Yet, now they prefer to discuss Education & Healthcare, Industrial Development, and Legislation & Justice.

The estimated regressions between topic proportions and the aforementioned macroeconomic indicators suggest that most of the topics (6 out of 9) are significantly associated with the real GDP. Only two topics, Development Strategy and National Identity, are significantly associated with inflation. Development Strategy is the only topic that shows a significant relationship with FDI. A comparison between the two presidents revealed that Tokayev prefers to discuss Reforms & Democracy, Legislation & Justice, and Development Strategy, while Nazarbayev demonstrates a preference for addressing National Identity.

Secondly, I examined SONA through the lens of sentiment analysis both within and across the speeches. For this, I utilized a baseline VADER and state-of-the-art XLM-R-ParlaSent models. XLM-R-ParlaSent was more informative and showed a higher accuracy in revealing the sentiments within the speeches, suggesting that presidents usually start and end speeches

positively. The president's opening and closing remarks show notable similarity, where both appeal for a shared national identity and instill trust in implementing the changes to the listeners. Interestingly, Tokayev's emotional range is narrower than Nazarbayev's, likely due to Tokayev's diplomatic background. As a former career diplomat, he is known for strict adherence to diplomatically refined statements and protocol. Overall, the sentiment of SONAs increased steadily until 2011, followed by a modest decline. Although most sentiments are still positive, neutral and negative sentiments have become more popular since then.

The correlations between the share of sentences with negative, neutral, and positive sentences and the topic proportions showed that Education & Healthcare, National Identity are traditionally discussed more positively, Foreign Affairs is discussed more negatively, and Industrial Development and Economic Management are discussed more neutrally. A similar correlation analysis, but with macroeconomic indicators, allowed me to find that FDI and real GDP are generally addressed in a positive light.

While previous studies exploited dictionary-based sentiment analysis to study the link between presidential remarks and economic effects, transformer models that are able to interpret text meaning and estimate their sentiment score were not utilized because they were unavailable back then (Wood, 2004; T. Dybowski & Adämmer, 2018). I fill this gap by combining an STM and a transformer-based model to study SONAs. The findings are relevant to scholars interested in the emotional valence of presidential addresses since they give insights into the rhetorical strategies presidents employ to build connection and persuade the audience (Salah, 2020; Miranda & Bringula, 2021; Kujanen et al., 2025). The findings on the main topics in SONAs address researchers interested in the political communication of leaders (Savin & Teplyakov, 2022; Oleinik, 2023; Lenton & Karibayeva, 2024). My discovery that real GDP is significantly associated with the share of many topics suggests a great importance of this macroeconomic factor to presidents. The finding offers a starting point for future research on the importance of macroeconomic factors in influencing presidents' approval ratings or election success.

The thesis has 4 main limitations. Firstly, a model fine-tuned for parliamentary proceedings was used instead of presidential speeches, as it was the only viable alternative available at the time of writing this thesis. Although similar in style, parliamentary proceedings are more adversarial than presidential addresses, while presidential rhetoric focuses more on unity and is more ceremonial. Therefore, the sentiment scores generated by the model do not accurately reflect the emotions and context as they could be. Secondly, the overall amount of SONAs available is relatively limited due to their rare frequency. Future research could explore the topics and sentiment score of all the President's speeches directed to the general public, parliament, business leaders, and so on. This, however, will make the texts a lot more heterogeneous, not only in length but also in content and sentiment. Thirdly, I first translated the texts into

English before doing topic modeling and sentiment analysis. Analysing original texts might provide more insights, but it is challenging given the multilingual format of SONAs. Fourthly, although I looked exclusively at the effects of macroeconomic indicators, it is worth exploring the extent to which SONA is responsive to media and public discussions (Yao et al., 2020). Nevertheless, I hope this thesis will inspire further research into how presidential addresses, macroeconomic factors, and sentiments interact.

Bibliography

Abercrombie, G., Nanni, F., Batista-Navarro, R., & Ponzetto, S. P. (2019). Policy preference detection in parliamentary debate motions. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Aguilar, A. P. (2023). Challenging machine translation engines: Some spanish-english linguistic problems put to the test. *Cadernos de Tradução*, *43*, e85397.

Aitzhanova, A., Katsu, S., Linn, J. F., & Yezhov, V. (Eds.). (2014). *Kazakhstan 2050: Toward a modern society for all*. Oxford University Press. https://EconPapers.repec.org/RePEc:oxp:obooks:9780199450602

Allen, D. E., & McAleer, M. (2019). Fake news and propaganda: Trump's democratic america and hitler's national socialist (nazi) germany. *Sustainability*, *11*(19), 5181.

Aryani, D., Kharisma, I. L., Sujjada, A., & Kamdan, K. (2024). Topic modeling of the 2024 election using the bertopic method on detik. com news articles. *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, *9*(2), 171–180.

Baffes, J., Kose, M. A., Ohnsorge, F., & Stocker, M. (2015). The great plunge in oil prices: Causes, consequences, and policy responses. *Consequences, and Policy Responses (June 2015)*.

Bestvater, S. E., & Monroe, B. L. (2023). Sentiment is not stance: Target-aware opinion classification for political text analysis. *Political Analysis*, *31*(2), 235–256.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, *30*, 31–40.

C. Timoneda, J., & Vallejo Vera, S. (2024). Bert, roberta or deberta? comparing performance across transformers models in political science text. *The Journal of Politics*, *87*. https://doi.org/10.1086/730737

Campbell, K. K., & Jamieson, K. H. (2008). *Presidents creating the presidency: Deeds done in words*. University of Chicago Press.

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., & Jurafsky, D. (2022). Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the*

National Academy of Sciences, *119*(31), e2120510119. https://doi.org/10.1073/pnas.2120510119

Cinelli, M., Ficcadenti, V., & Riccioni, J. (2021). The interconnectedness of the economic content in the speeches of the us presidents. *Annals of Operations Research*, *299*(1), 593–615. https://doi.org/10.1007/s10479-019-03372-2

Crockett, S., & Lee, C. (2012). Does it matter what they said? a text mining analysis of the state of the union addresses of usa presidents. *2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, 77–82.

Davis, K. M., & Gardner, W. L. (2012). Charisma under crisis revisited: Presidential leadership, perceived leader effectiveness, and contextual influences. *The Leadership Quarterly*, *23*(5), 918–933. https://doi.org/https://doi.org/10.1016/j.leaqua.2012.06.001

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Donley, R. E., & Winter, D. G. (1970). Measuring the motives of public officials at a distance: An exploratory study of american presidents. *Behavioral Science*, *15*(3), 227–236.

Drozdova, O., & Robinson, P. (2019). A study of vladimir putin's rhetoric. *Europe-Asia Studies*, *71*(5), 805–823.

Dugoua, E., Dumas, M., & Noailly, J. (2022). Text as data in environmental economics and policy. *Review of Environmental Economics and Policy*, *16*(2), 346–356.

Dybowski, T. P., Dybowski, J. N., & Adämmer, P. (2016, May). *The economic effects of u.s. presidential tax communication* (CQE Working Papers No. 4916). Center for Quantitative Economics (CQE), University of Muenster. https://ideas.repec.org/p/cqe/wpaper/4916.html

Dybowski, T., & Adämmer, P. (2018). The economic effects of u.s. presidential tax communication: Evidence from a correlated topic model. *European Journal of Political Economy*, *55*, 511–525. https://doi.org/https://doi.org/10.1016/j.ejpoleco.2018.05.001

Ebeling, R., Nobre, J., & Becker, K. (2023). A multi-dimensional framework to analyze group behavior based on political polarization. *Expert Systems with Applications*, *233*, 120768. https://doi.org/https://doi.org/10.1016/j.eswa.2023.120768

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., et al. (2023). The parlamint corpora of parliamentary proceedings. *Language resources and evaluation*, *57*(1), 415–448.

Fairclough, N. (1989). *Language and power*. Longman.

Fuhse, J., Stuhler, O., Riebling, J., & Martin, J. L. (2020). Relating social and symbolic relations in quantitative text analysis. a study of parliamentary discourse in the weimar

republic [Discourse, Meaning, and Networks: Advances in Socio-Semantic Analysis]. *Poetics*, *78*, 101363. https://doi.org/https://doi.org/10.1016/j.poetic.2019.04.004

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hart, R. P., Childers, J. P., & Lind, C. J. (2013). *Political tone: How leaders talk and why*. University of Chicago Press.

Hidalgo-Ternero, C. (2021). Google translate vs. deepl: Analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monografías de Traducción e Interpretación*, 154–177. https://doi.org/10.6035/MonTI.2020.ne6.5

Hoffman, D. R., & Howard, A. D. (2010). The presidential rhetoric of hard times. *APSA 2010 Annual Meeting Paper*.

Hoffmann, T. (2018). "too many americans are trapped in fear, violence and poverty": A psychology-informed sentiment analysis of campaign speeches from the 2016 us presidential election. *Linguistics Vanguard*, *4*(1), 20170008. https://doi.org/10.1515/lingvan-2017-0008

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, *8*(1), 216–225.

Kernell, S. (2007). *Going public: New strategies of presidential leadership*. CQ Press.

Kiyama, N. (2018). How have political interests of us presidents changed?: A diachronic investigation of the state of the union addresses through topic modeling. *English Corpus Studies*, *25*, 79–99.

Koehn, P. (2011). European parliament proceedings parallel corpus 1996–2011 [Accessed March 21, 2025].

Kujanen, M., Koskimaa, V., & Raunio, T. (2025). Confrontational or 'statespersonlike'style? examining finnish and french presidents' public speeches and messages, 2000–2020. *Political Studies Review*, *23*(1), 14–32.

Laqueur, W. (2015). *Putinism: Russia and its future with the west*. Macmillan.

Lenton, A., & Karibayeva, A. (2024). Dear compatriots: New year speeches as sites for post-soviet political communication. *Communist and Post-Communist Studies*, 1–28.

Licht, H., Sczepanski, R., Laurer, M., & Bekmuratovna, A. (2024). *No more cost in translation: Validating open-source machine translation for quantitative text analysis* (tech. rep.). ECONtribute Discussion Paper.

Liu, D., & Lei, L. (2018). The appeal to political sentiment: An analysis of donald trump's and hillary clinton's speech themes and discourse strategies in the 2016 us presidential election. *Discourse, context & media*, *25*, 143–152.

Ma, T., Pan, Q., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. (2022). T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, *9*(3), 879–890. https://doi.org/10.1109/TCSS.2021.3088506

Maerz, S. F., & Schneider, C. Q. (2020). Comparing public communication in democracies and autocracies: Automated text analyses of speeches by heads of government. *Quality & Quantity*, *54*(2), 517–545.

Mah, A., & Song, E. (2024). Elite speech about climate change: Analysis of sentiment from the united nations conference of parties, 1995–2021. *Sustainability*, *16*(7), 2779.

Mahoney, J., Coogle, C. L., & Banks, P. D. (1984). Values in presidential inaugural addresses: A test of rokeach's two-factor theory of political ideology. *Psychological Reports*, *55*(3), 683–686.

Maligkris, A. (2017). Political speeches and stock market outcomes. *30th australasian finance and banking conference*.

Mate, A., Sebők, M., Wordliczek, L., Stolicki, D., & Feldmann, Á. (2023). Machine translation as an underrated ingredient? solving classification tasks with large language models for comparative research. *Computational Communication Research*, *5*(2), 1. https://doi.org/https://doi.org/10.5117/CCR2023.2.6.MATE

Mendonça, M., & Figueira, Á. (2024). Topic extraction: Bertopic's insight into the 117th congress's twitterverse. *Informatics*, *11*(1), 8.

Miranda, J. P. P., & Bringula, R. P. (2021). Exploring philippine presidents' speeches: A sentiment analysis and topic modeling approach. *Cogent Social Sciences*, *7*(1), 1932030.

Mochtak, M., Rupnik, P., & Ljubešić, N. (2024, May). The ParlaSent multilingual training dataset for sentiment identification in parliamentary proceedings. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 16024–16036). ELRA; ICCL. https://aclanthology.org/2024.lrec-main.1393/

Mohr, J. W., & Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter [Topic Models and the Cultural Sciences]. *Poetics*, *41*(6), 545–569. https://doi.org/https://doi.org/10.1016/j.poetic.2013.10.001

Oleinik, A. (2023). Governmentality in north american and post-soviet political discourses: An analysis of presidential speeches and their analogues in the united states, canada, russia, ukraine, and kazakhstan delivered from 1993 to 2021. *International Journal of Communication*, *17*, 24.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Ruhl, J., Nay, J., & Gilligan, J. (2018). Topic modeling the president: Conventional and computational methods. *Geo. Wash. L. Rev.*, *86*, 1243.

Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, *112*(35), 10837–10844.

Salah, Z. (2020). Intelligent framework for long-text political speeches summarization and visualization using sentiment lexicons: A study directed at king abdullah ii discussion papers. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*, 2393–2407. https://doi.org/10.30534/ijatcse/2020/225922020

Savin, I., King, L. C., & van den Bergh, J. (2025). Analysing content of paris climate pledges with computational linguistics. *Nature Sustainability*, 1–10.

Savin, I., & Teplyakov, N. (2022). Topics of the nationwide phone-ins with vladimir putin and their role for public support and russian economy. *Information Processing & Management*, *59*(5), 103043.

Schumacher, G., Schoonvelde, M., Traber, D., Goyal, T., & De Vries, E. (2016). Euspeech: A new dataset of eu elite speeches.

Serikkyzy, A., Bakirbekova, A., Baktymbet, S., Yelshibayev, R., & Baktymbet, A. (2024). Foreign direct investment and economic development: An international perspective. *ECONOMICS-Innovative and Economics Research Journal*, *12*(2), 97–111.

Sikanku, G. E. (2022). Presidential discourse, the public and recurring themes: A political communication analysis of the 2019 state of the nation address in ghana. *Communication and the Public*, *7*(4), 176–187. https://doi.org/10.1177/20570473221129652

Tokayev, K.-J. (2023, September). State of the nation address: Economic course of a just kazakhstan [Accessed: 2025-02-21].

United Nations Conference on Trade and Development. (2024). World investment report 2024: Investment facilitation and digital government – overview [Accessed: 2025-03-20]. https://unctad.org/system/files/official-document/wir2024_overview_en.pdf

Van den Bergh, J. C. (2009). The gdp paradox. *Journal of economic psychology*, *30*(2), 117–135.

Wood, B. D. (2004). Presidential rhetoric and economic leadership. *Presidential Studies Quarterly*, *34*(3), 573–606. Retrieved May 19, 2025, from http://www.jstor.org/stable/27552614

Wood, B. D., Owens, C. T., & Durham, B. M. (2005). Presidential rhetoric and the economy. *The Journal of Politics*, *67*(3), 627–645. https://doi.org/10.1111/j.1468-2508.2005.00332.x

World Bank. (2023). Gdp (current us$) - world bank data [Accessed: 2025-03-20]. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Yao, Q., Liu, Z., & Stephens, L. F. (2020). Exploring the dynamics in the environmental discourse: The longitudinal interaction among public opinion, presidential opinion, media

coverage, policymaking in 3 decades and an integrated model of media effects. *Environment systems and decisions*, *40*, 14–28.

Appendix

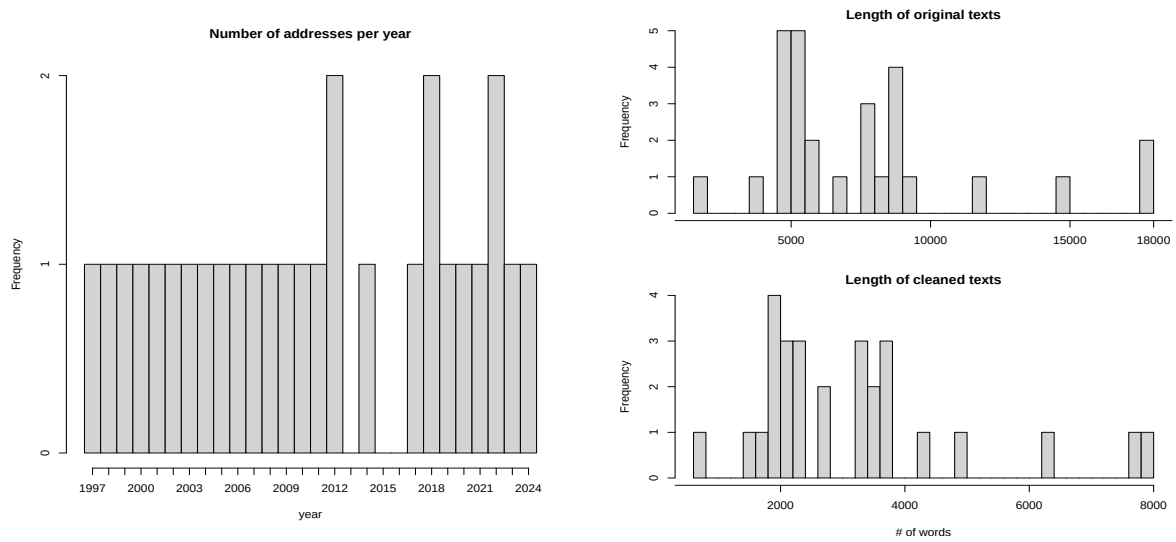

Figure A1*: Number and length of addresses (both original and processed) by presidents across the years.*

Independent Variables	**Calculation Formulas for the Variables**	**Data Source**
Real GDP	$\frac{\text{Nominal GDP}}{\text{GDP Deflator}} \times 100$	The World Bank
FDI net inflows	$\text{Equity Capital}_{\text{in}} + \text{Reinvestment of Earnings}_{\text{in}} + \text{Other Capital}_{\text{in}}$	The World Bank
GDP Deflator	$\frac{\text{Nominal GDP}}{\text{Real GDP}} \times 100$	The Bureau of National Statistics
Real FDI net inflows	$\frac{\text{FDI Net Inflows (Current USD)}}{\text{GDP Deflator}} \times 100$	The World Bank The Bureau of National Statistics
Year-over-Year Inflation	$\frac{P_t - P_{t-12}}{P_{t-12}} \times 100$	Zakon.kz (1997–2020) Trading Economics (2021–2024)

Table A1*: Independent variables, their calculation formulas, and data sources.*

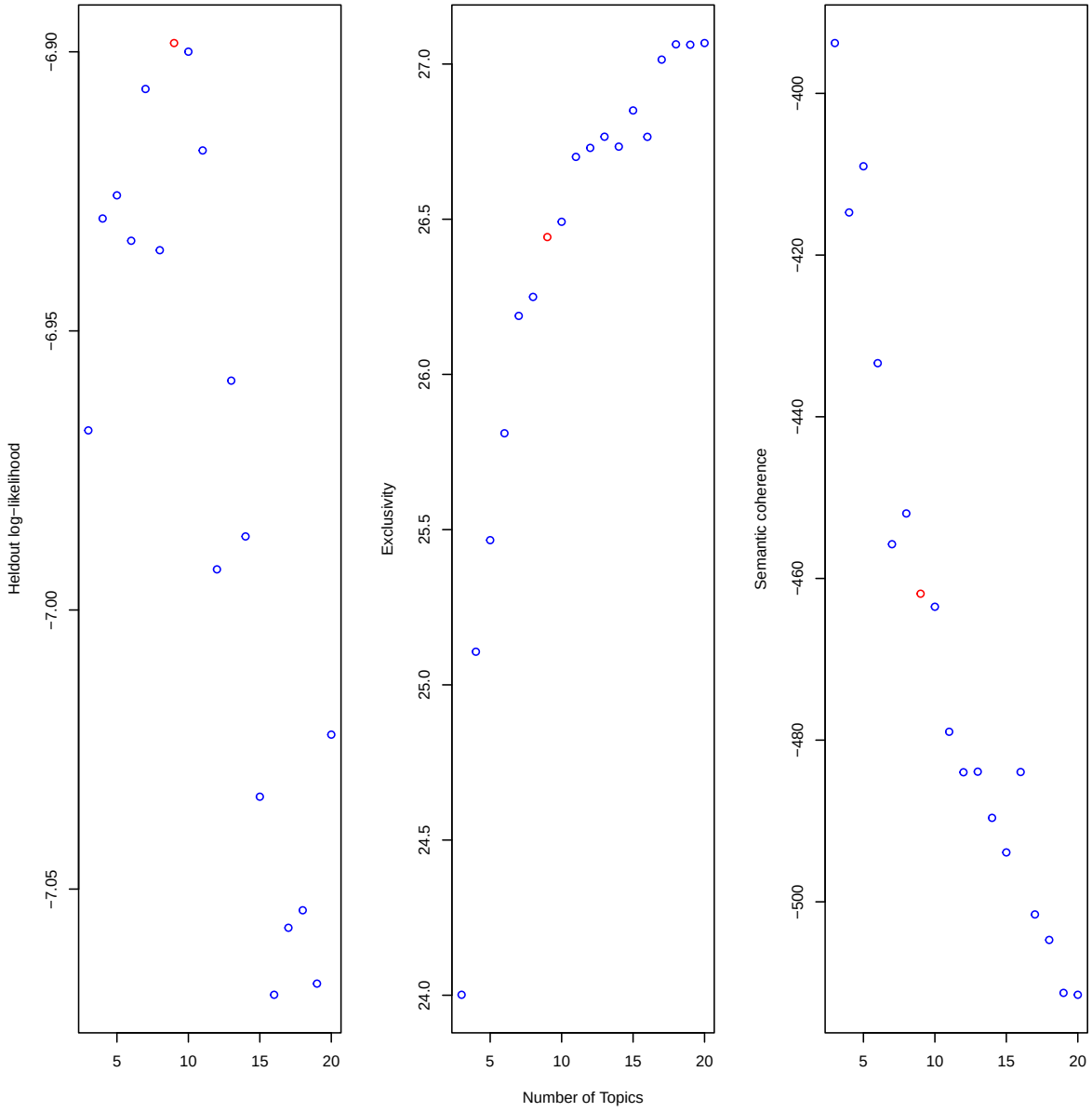

***Figure A2**: Comparison of model performance for different topic numbers.*

***Figure A3**: Visual summary of topics in a word cloud format.*

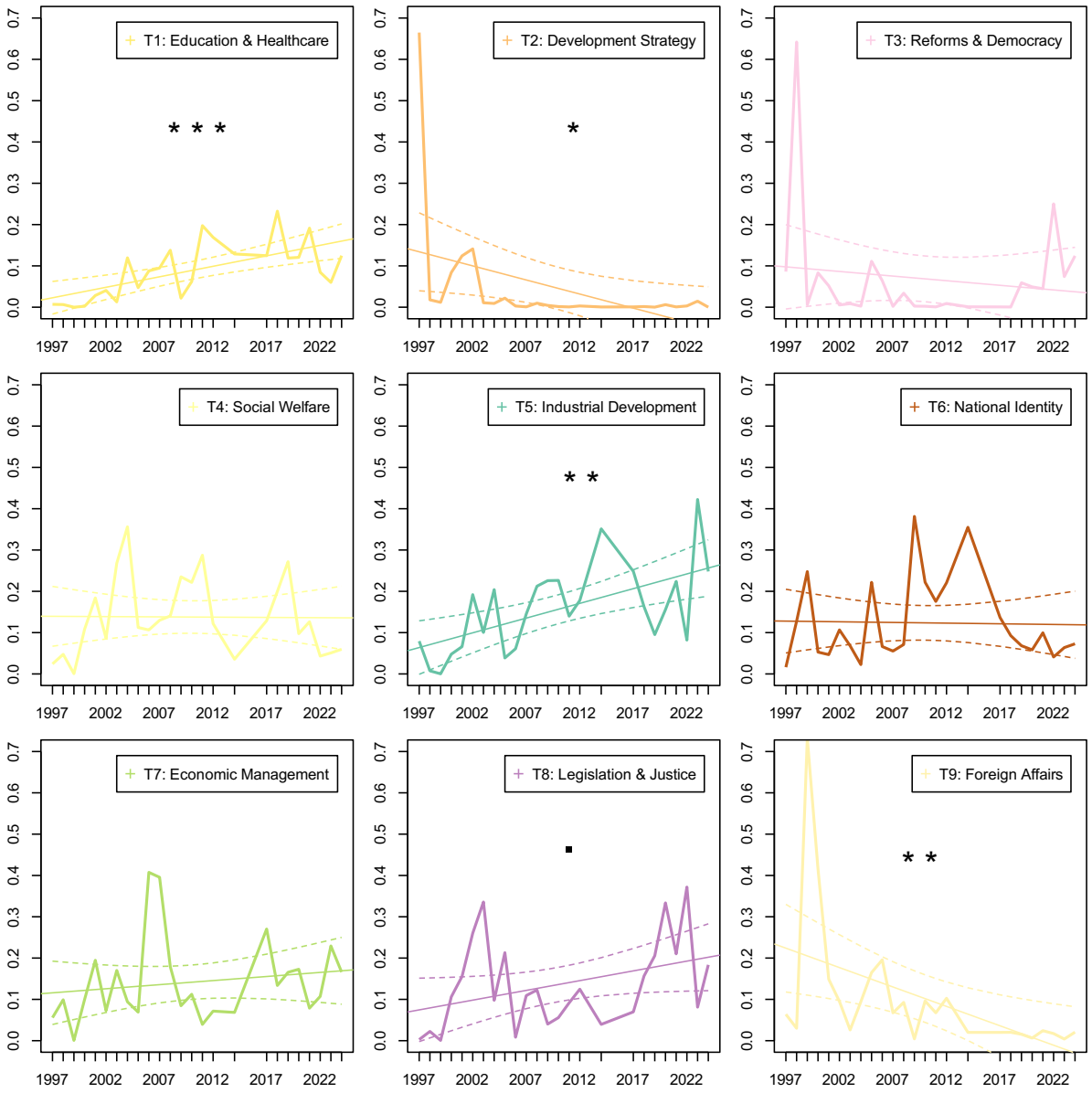

***Figure A4**: Trend of topics over time. ***, ** , *, • mean 0.1%, 1%, 5%, and 10% significance level, respectively.*

Year	Sentence with minimum sentiment score (VADER/RoBERTa)	Sentence with maximum sentiment score (VADER/RoBERTa)
1997	VADER: Even today, in the end of the xxth century, after our learning the lessons of the world war ii and the cold war, we have not yet parted with the threat of the world being split up into blocks and alliences.	VADER: Our strategy of healthy economic growth rests on a strong market economy, an active part played by the state and attraction of significant foreign investments thereto.
	RoBERTa: Actually, the whole of the production sector suffers from the crisis of payment system which is the result of the actions of incompetent or corrupt chiefs of enterprises, who are not accountable or poorly accountable to related owners.	RoBERTa: He will be ever proud of their progress and achievements.
2018	VADER: Recent tragic events have also revealed the problem of poaching, as a most dangerous form of organised crime.	VADER: On top there will be new champions, and at its base a healthy and active youth and, ultimately, a strong nation.
	RoBERTa: The foundations of the global security system and international trade rules that seemed unshakable are now crumbling.	RoBERTa: There is nothing greater than this noble goal!
2019	VADER: This document, which criminalises torture, needs to be brought in line with the provisions of the international convention against torture and other cruel, inhuman or degrading treatment or punishment.	VADER: We must find a positive answer to the growing public demand for a fairer distribution of benefits arising from the growth of national income and for effective social lifts.
	RoBERTa: However, numerous serious crimes still occur in the country.	RoBERTa: Harmony and unity, wisdom and mutual understanding help our nation move forward.
2024	VADER: The ministry of internal affairs must take decisive action against all offences, from petty hooliganism and vandalism to illegal immigration and serious criminal activities.	VADER: Congratulations to all of you on the opening of the regular session, and i wish you success in your activities for the benefit of the country!
	RoBERTa: Currently, some entrepreneurs consume large amounts of electricity to the detriment of the economy while not paying taxes in full.	RoBERTa: We are grateful to our allies and partners for their support.

***Table A2**: Illustrative statements for each model.*

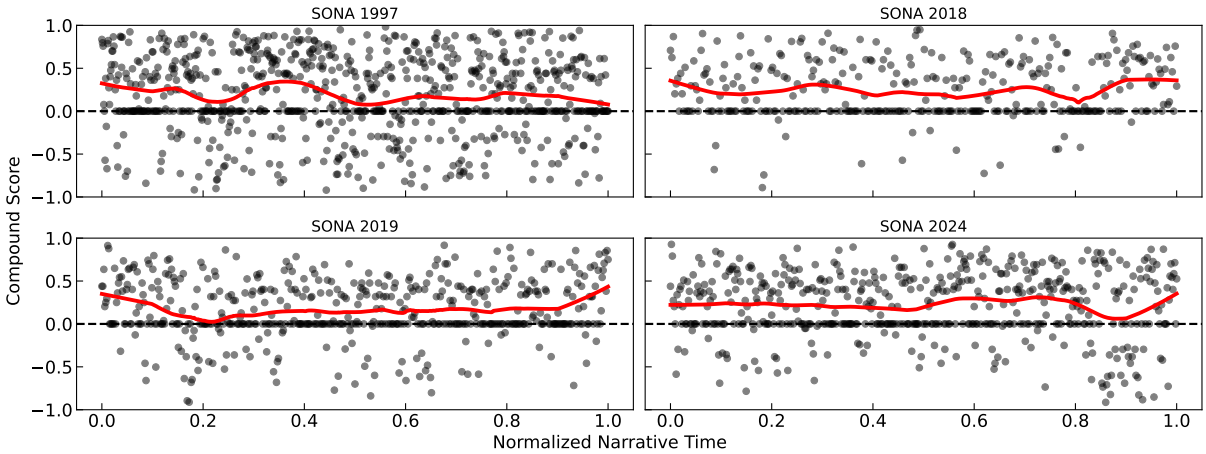

***Figure A5**: The emotional valence of Nazarbayev and Tokayev's first and last SONAs (as of 2025) measured using VADER. The red line represents the LOESS fit. Each dot represents a sentence's sentiment. The time of an address was normalized by representing it as the relative position of sentences, scaled between 0 (start of an address) and 1 (end of an address).*

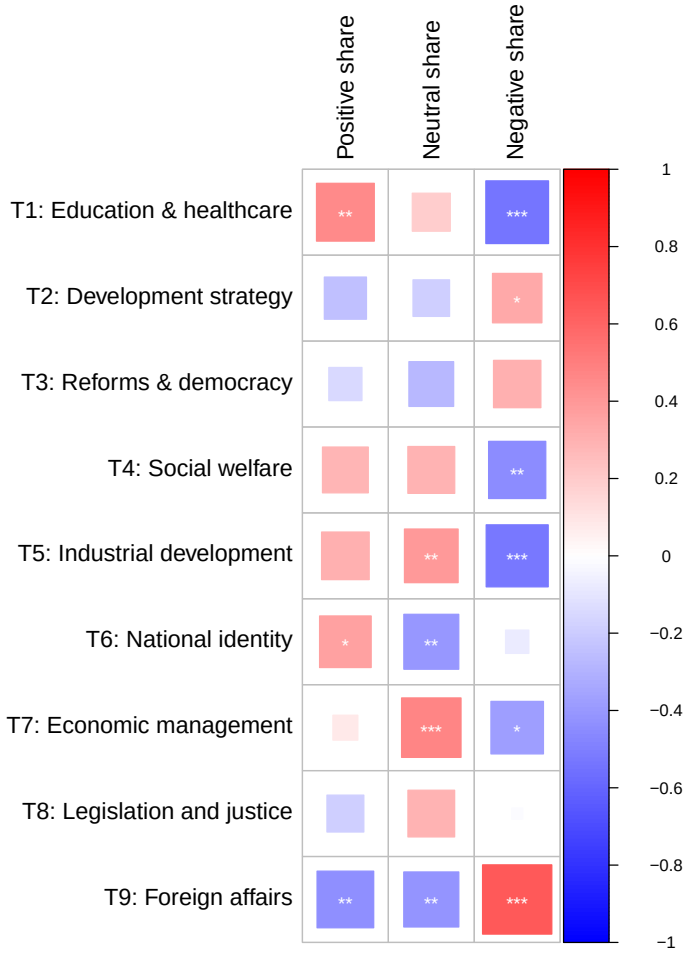

***Figure A6**: Pearson's correlations between topic prevalence and the share of positive, negative, and neutral sentences. ***, ** and * mean 0.1%, 1% and 5% significance level, respectively.*

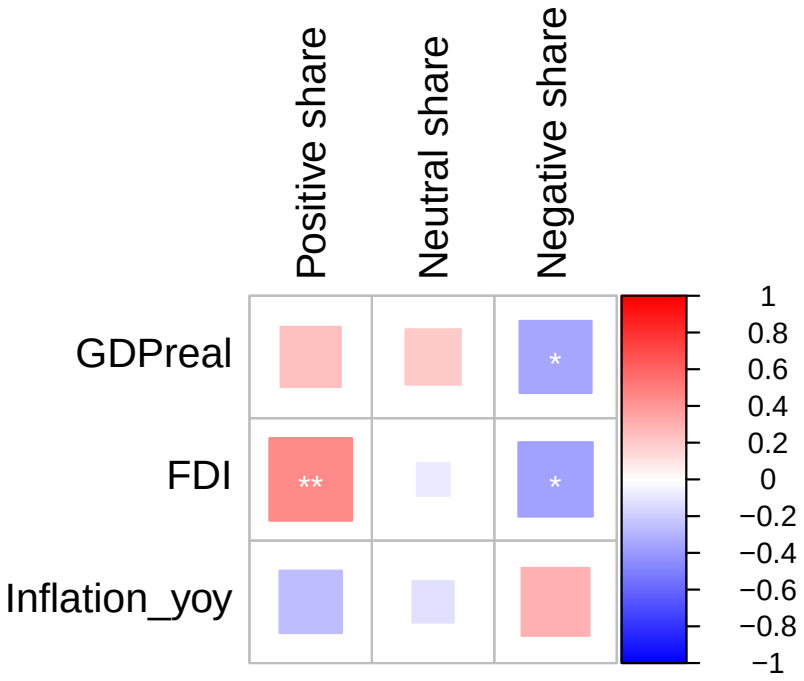

***Figure A7**: Pearson's correlations between macroeconomic indicators and the share of positive, negative, and neutral sentences. ** and * mean 1% and 5% significance level, respectively.*