Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

DISSERTATION

FORECASTING INFLATION WITH MACHINE LEARNING

VLADISLAVA PILETSKA

JANUARY - 2025

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
DISSERTATION

FORECASTING INFLATION WITH MACHINE LEARNING

VLADISLAVA PILETSKA

SUPERVISION:
JOÃO AFONSO BASTOS
ADRIANA CORNEA-MADEIRA

JANUARY - 2025

GLOSSARY

ADF – Augmented Dickey-Fuller test

API – Application Programming Interface

AR – Autoregressive models

AR(4) – Autoregressive model of order 4

BoW – Bag of Words

CPI – Consumer Price Index

DM – Diebold-Mariano test

DTM – Document-Term Matrix

ENet – Elastic Net

EUR to RUB – Euro to Russian Ruble exchange rate

GDP – Gross Domestic Product

KPSS – Kwiatkowski-Phillips-Schmidt-Shin test

L1 – Lasso regularization

L2 – Ridge regularization

LASSO – Least Absolute Shrinkage and Selection Operator

LDA – Latent Dirichlet Allocation

NLP – Natural Language Processing

PLS – Partial Least Squares

Q1 – First Quarter

RF – Random Forest

RMSE – Root Mean Squared Error

rRMSE – Relative Root Mean Squared Error

USD to RUB – US Dollar to Russian Ruble exchange rate

WSJ – The Wall Street Journal

ABSTRACT

This dissertation explores the integration of machine learning techniques and narrative-based data for enhancing inflation forecasting, with a focus on the Russian economy. Traditional econometric models often fail to capture nonlinear dynamics and shifting macroeconomic conditions, particularly in volatile contexts. To address these limitations, this study employs a hybrid approach combining macroeconomic indicators with narrative data extracted from Russian news articles using Latent Dirichlet Allocation (LDA). Predictive modeling is conducted using machine learning algorithms, including Random Forest, LASSO, and Elastic Net.

Macroeconomic data, sourced from platforms such as RosStat and Investing.com, and narrative data, obtained via web scraping from Lenta.ru, were preprocessed to ensure consistency and stationarity. The analysis reveals that integrating text-derived features with economic indicators improves forecasting accuracy across multiple horizons. Random Forest consistently outperforms other models, particularly in short-term forecasts, underscoring its ability to leverage both structured and unstructured data.

Key contributions include the development of a scalable framework for inflation forecasting in non-Western economies, validation of narrative-based predictors, and incorporation of geopolitical factors like sanctions into the analysis. The results show that combining macroeconomic data with narrative-based information leads to better predictions than using either type of data alone, highlighting the value of qualitative insights in understanding and managing economic fluctuations.

KEYWORDS: Big data; Inflation forecasting; Machine learning; Economic narratives; Textual analysis; Macroeconomics.

JEL CODES: C22; C53; C55; E31.

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my professors João Afonso Bastos and Adriana Cornea-Madeira for their invaluable guidance and support throughout this project. Their expertise, encouragement, and insightful feedback have been instrumental in the completion of this work.

I would also like to thank my family for their constant support and encouragement. Their love and understanding have been a continuous source of motivation throughout the course of this work.

1. INTRODUCTION

Accurately forecasting inflation is a critical task for policymakers, businesses, and financial institutions, as inflation directly impacts economic stability, investment decisions, and monetary policy. This study focuses on the integration of machine learning techniques and narrative-based data to enhance inflation forecasting accuracy. The approach is particularly relevant in the context of the Russian economy, where inflation dynamics are shaped by unique economic features and significant geopolitical events. Understanding and predicting these trends is essential for effective decision-making in such a complex environment.

The existing literature highlights the challenges of conventional approaches, such as autoregressive models, random walk models, and unobserved component models. While these methods have been effective in stable conditions, they often struggle to capture nonlinear relationships and dynamic shifts in macroeconomic environments. Recent advancements in machine learning and the inclusion of alternative data sources, such as economic narratives and textual analysis, have shown promise in addressing these gaps. Studies by Yongmiao Hong et al. (2024) and Bianchi et al. (2021) demonstrate that narrative-based forecasting models significantly improve performance, particularly during volatile periods. Similarly, Babii et al. (2021) highlight the advantages of machine learning models in leveraging large and complex datasets to capture nonlinear dependencies that traditional models miss.

This study adopts a hybrid approach by combining Russian macroeconomic indicators with text-derived data from Russian news articles to improve both short- and long-term inflation forecasts. Key methodologies include Latent Dirichlet Allocation (LDA) for extracting narrative topics and machine learning algorithms such as Random Forest, Partial Least Squares, LASSO, and Elastic Net for predictive modelling. To ensure robust evaluation, the Diebold-Mariano (DM) test was employed to compare the accuracy of these models against traditional benchmarks.

The primary contribution of this work lies in demonstrating how integrating narrative-based predictors with traditional macroeconomic indicators enhances forecasting accuracy in the context of the Russian economy. By focusing on Russian data, this study addresses a significant gap in the literature, where many forecasting models are tested

predominantly on Western economies. Additionally, the use of Python enables efficient and reproducible workflows, making the analysis both scalable and adaptable. This research not only validates the utility of machine learning in economic forecasting but also introduces a framework for models tailored to the unique dynamics of the Russian economy.

The next chapters proceed as following. Chapter 2 reviews the relevant literature, discussing traditional models, machine learning techniques, and the role of economic narratives in forecasting. Chapter 3 outlines the data sources, preprocessing steps, and methodologies used for analysis, with an emphasis on Russian economic features, news text data, and Python implementation. Chapter 4 presents the results, including model comparisons and key insights from the forecasts. Finally, Chapter 5 concludes with a summary of findings and directions for future research.

2. LITERATURE REVIEW

The accurate forecasting of inflation is a critical task for policymakers, financial institutions, and economic stakeholders, as it directly influences monetary policy, market expectations, and decision-making processes. Traditional econometric approaches have demonstrated effectiveness in stable environments but often fail during periods of rapid economic change. Recent advancements in machine learning and the utilization of alternative data sources, such as macroeconomic narratives and textual analysis, have opened new pathways for improving forecasting accuracy.

Economic narratives have become a critical addition to forecasting tools, as they capture the sentiment, expectations, and events driving economic trends. For instance, one study utilized over 880,000 articles from the Wall Street Journal to extract topics through Latent Dirichlet Allocation (LDA), demonstrating that narrative-based forecasts outperformed traditional benchmarks, particularly during recessions (Yongmiao Hong et al., 2024). Similarly, Gu et al. (2020) highlighted how the inclusion of narrative data improved forecast accuracy, especially when combined with macroeconomic variables.

Machine learning has been a transformative force in inflation forecasting. Methods such as Random Forests, LASSO regression, and Elastic Net enable the analysis of large datasets while capturing nonlinear relationships. Medeiros et al. (2019) demonstrated that ML models outperform traditional autoregressive and random walk models, achieving error reductions of up to 30%. Furthermore, Babii et al. (2021) showed that ML methods are particularly effective in data-rich environments, where traditional models fail to adapt to the dynamic nature of economic variables.

Traditional econometric models, such as Autoregressive (AR) models, Random Walk models, have long been benchmarks for inflation forecasting. These models rely on linear relationships and historical data but often struggle to adapt to sudden economic changes or incorporate large, complex datasets (Atkeson and Ohanian, 2001; Stock and Watson, 2007). These limitations have driven the shift toward machine learning techniques, which excel in handling nonlinear relationships and diverse data sources.

The Diebold-Mariano's (1995) test is a widely used statistical tool for comparing the machine learning forecasts in out-of-sample forecasting contexts (Diebold and Shin 2019; Gu, Kelly, and Xiu, 2020; Babii, Ghysels, 2021). It evaluates whether the differences in

forecast errors between models are statistically significant, making it particularly valuable for assessing whether one model consistently outperforms another. Importantly, the DM test was designed for comparing forecasts rather than models themselves. Diebold (2015) emphasized this distinction, noting that its misuse for direct model comparisons can lead to incorrect conclusions.

Combining macroeconomic indicators with text-derived data has further improved forecasting capabilities. This hybrid approach allows models to capture both the structured historical trends of economic indicators and the real-time dynamics of economic narratives. For example, Yongmiao Hong et al. (2024) integrated macroeconomic and text data to improve both short- and long-term inflation forecasts. Their findings showed that that narrative-based forecasts have better relative performance in the long run. Narrative-based forecasts have much better forecasting performance than several benchmark models.

Khemani and Adgaonkar (2021) explored the application of natural language processing (NLP) tools in their study. The authors utilized the Natural Language Toolkit (NLTK) to process and analyze textual data from Reddit news headlines, focusing on extracting meaningful insights from user-generated content. This paper highlights the effectiveness of NLTK for tasks such as tokenization, sentiment analysis, and text classification, demonstrating its potential in handling large volumes of social media data.

In addition, for improving model accuracy, these innovations have important implications for policymakers and financial institutions. Narrative-based and ML-enhanced models provide timely forecasts, crucial for effective monetary policy and risk management. Goulet Coulombe et al. (2022) highlighted how these methods allow for better identification of emerging risks and trends, offering a significant advantage in periods of economic uncertainty.

The literature demonstrates the transformative potential of integrating machine learning and narrative-based approaches with traditional econometric methods. These advancements enable the capture of complex and evolving economic dynamics, ultimately leading to more accurate and robust inflation.

3. METHODOLOGY

3.1. Data Collection

For this work, an approach was adopted by combining key macroeconomic features of Russia with textual data sourced from the well-known online publication, Lenta.ru. The macroeconomic features provided a quantitative representation of Russia's economic landscape, offering critical insights into indicators such as inflation, interest rates, Consumer Price Index (CPI), average monthly salaries, unemployment rates, and Gross Domestic Product (GDP). Meanwhile, the textual data from Lenta.ru contributed a qualitative dimension, capturing narratives, sentiment, and contextual discourse related to economic events.

3.1.1. Macroeconomic Data

Macroeconomic data plays a key role in assessing the economy of a country. It helps politicians, businesses, and investors make informed decisions. It provides information on indicators such as GDP, inflation, unemployment, and interest rates, allowing governments to develop effective economic policies. For businesses and investors, this data helps forecast trends, manage risks, and make strategic decisions. Accurate macroeconomic data is essential for economic stability and sustainable growth.

The dataset for macroeconomic data comprised key macroeconomic indicators, including Inflation, Interest Rate, Consumer Price Index (CPI), Average Monthly Salary, Unemployment Rate, and Gross Domestic Product (GDP).

Additionally, it was incorporated financial data, such as the EUR to RUB exchange rate, USD to RUB exchange rate, and the prices of natural gas, oil, and gold.

Other variables in the dataset included Exports of Goods and Services, Imports of Goods and Services, Russia's External Debt, and the prices of essential goods from the basic consumer basket.

For the implementation of this project, all macroeconomic data was collected from open sources. The main portion of the data was obtained from the official resource of the Russian Federation named RosStat.

Additionally, macroeconomic data was collected from the website Investing.com using Parsing Techniques. This platform provides access to financial market data in real-

time and references official sources of the Russian Federation. Supplementary, financial data was retrieved from the public online source called Yahoo Finance with use of Python Library yfinance with integrated API.

The information regarding the publication of new sanctions was obtained from the official source Eur-Lex. Data on sanctions has been utilized in this work as dummy variables to indicate whether sanctions were imposed on Russia in a given month.

Table 1 presents the macroeconomic features with their description and granularity used in this work, which serve as key indicators for analyzing and forecasting inflation trends.

TABLE 1

MACROECONOMIC FEATURES

Feature	Description	Granularity
Inflation Rate	Numeric value of inflation (percentage).	monthly
Interest Rate	Numeric value of the minimum interest rate at which the Central Bank of Russia (CBR) lends to commercial banks.	monthly
CPI	Customer Price Index (CPI) – an economic indicator of consumer inflation, reflecting changes in the purchasing power of the national currency (RUB).	monthly
Average monthly salary	Monthly average salary in Russia (RUB).	monthly
Unemployment Rate	Unemployment rate – the percentage of unemployed individuals relative to the total working-age population.	daily
GDP	Gross Domestic Product (GDP) – the total value of all final goods and services produced in the country annually.	quarterly

Fx EUR to RUB	EUR exchange rate to RUB.	daily
Fx USD to RUB	USD exchange rate to RUB.	daily
Oil Price	Oil price (USD per barrel).	daily
Gas Price	Gas price (USD per million British thermal units (MMBtu)).	daily
Gold Price	Gold price (USD per troy ounce).	daily
Sanctions	Sanctions were used as a dummy variable. Dummy variables indicate whether sanctions were imposed on Russia in a given month: 1 if sanctions were imposed, and 0 otherwise.	daily
Chicken eggs 10pcs	The price of chicken eggs (10 pcs.) in Russia per month.	monthly
Granulated sugar 1kg	The price of granulated sugar (1 kg.) in Russia per month.	monthly
Wheat flour 1kg	The price of wheat flour (1 kg.) in Russia per month.	monthly
Exports	Exports of goods and services.	monthly
Imports	Imports of goods and services.	monthly
External Debt	External debt for Russia.	quarterly

3.1.2. Text Data

The next step of data collection involves the acquisition of textual data. For this purpose, articles were gathered from a well-known online publication, Lenta.ru. This platform serves as a Russia-focused, economy-oriented news source (WSJ) comparable in its focus and scope to the Wall Street Journal (WSJ). The articles from Lenta.ru were obtained using a web scraping method.

The web parser was designed to extract news headlines based on a specific topic, in this case, macroeconomic, over a selected period. This was achieved by leveraging two Python libraries: Requests and BeautifulSoup. The Requests library facilitated access to the website's content, while BeautifulSoup enabled the structured parsing and extraction of information from the HTML code.

Ultimately, 151.215 articles were collected. These articles span a date range from January 2005 to November 2024. Below, Table 2 presents the summary statistics of the collected news data.

TABLE 2

NUMBER OF ARTICLES PER MONTH

Mean	632.89
Std	410.36
Min	174.00
Max	2418.00

Figure 1 illustrates the distribution of the monthly count of economic articles (Lenta.ru) per year over the time period spanning from January 2005 to November 2024.

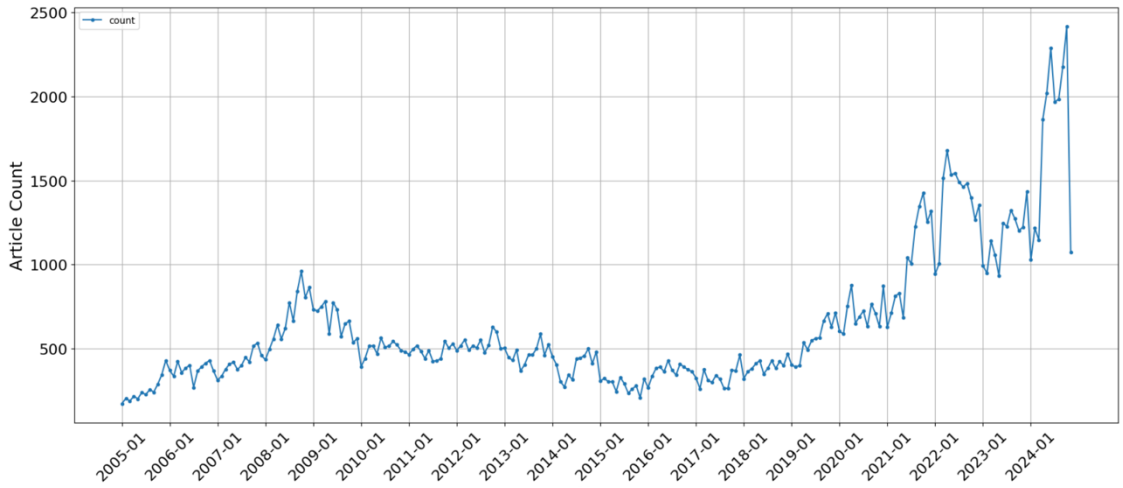

FIGURE 1 – Monthly count of economic articles on Lenta.ru per year.

This graph clearly demonstrates a sharp rise in media interest in economic topics during the period of 2023-2024. This trend indicates a pattern where interest in economic

headlines increases during times of political unrest in the country. Subsequently, in 2024, there is a noticeable decline in media attention to news headlines related to the economy.

3.2. Data Preprocessing

3.2.1. Macroeconomic Dataset Creation

Since the macroeconomic data was collected from various sources, it had to be transformed into a unified format to create a working dataset for further use. All subsets of features were standardized to a common format, ensuring consistent granularity and numerical representation. These subsets were then merged using a shared column, "date". As a result, the final dataset was created, containing the "date" column and columns with numerical values for each feature.

Most of the collected data consisted of monthly records, which is why the granularity for the resulting dataset was chosen to be monthly. This approach also provided a sufficient number of observations in the final dataset. To achieve this, several additional transformations were applied.

For those periods where only quarterly data was available, the same value was assigned to all months within the respective quarter. For example, if data was available for Q1, the same value was assigned to January, February and March.

For features with daily granularity, time-series were transformed into monthly average values and added into the final dataset.

As a result, the final dataset consists of macroeconomic indicators with a total of 151 rows (observations) and columns including the "date" column, 16 macroeconomic features (including Inflation Rate as the target variable) and dummy variables for Sanctions. Sanctions data has been incorporated as dummy variables to indicate whether sanctions were imposed on Russia during a specific month. The variable is assigned a value of 1 if sanctions were imposed and 0 otherwise. The final macroeconomic dataset spans the period from March 2011 to December 2023.

Additionally lag of size 1 month was added for each of numeric features. This should allow the model to understand current timestamp value with respect to the previous one, capture dramatic changes and their impact on target variable. It worth mentioning the t-1 inflation rate is also considered as a feature.

After all transformations, the dataset was checked for missing values. All features contained complete data without any missing values or duplicates.

3.2.2. Text Dataset Creation

Following the data collection process, a thorough preprocessing of the textual data was conducted for the work. A series of preprocessing steps were applied to each individual article, as outlined below:

- Duplicates were removed;

- Missing Values were dropped to maintain the dataset's integrity;

- Punctuation marks and numerical values were eliminated using lists from the NLTK library;

- Common stop words, such as prepositions and articles, which do not contribute to the semantic meaning, were removed. This was achieved using available stop word lists for Russian language provided by the NLTK library and GitHub;

- All words were converted to lowercase;

- Lemmatization was applied to bring words to their base or dictionary form. Since the textual data used in this study was in Russian, the Natasha library was utilized for this step. Natasha is a well-known library tailored specifically for processing Russian-language text, similar to NLTK but designed for Russian linguistic structures;

Additionally, to ensure consistency in the use of textual data, the dataset was trimmed to align with the same date distribution as the macroeconomic features. As a result, the final dataset contained 96,199 articles spanning the period from March 2011 to December 2023.

This thorough preprocessing ensured that the text data was cleaned, normalized, and prepared for the subsequent application of the Latent Dirichlet Allocation (LDA) model to extract topic time series. Finally, the text dataset has no missing values or duplicates.

3.2.3. Text Tokenization

Tokenization is the process of breaking text into smaller parts, called tokens, to make it easier for machine learning models to analyze and understand human language. In the

field of Natural Language Processing (NLP) and machine learning, these tokens can range from individual characters to entire words. This step is crucial because it simplifies complex text, enabling machine learning models to process and interpret language more effectively.

In this work, word-based tokenization was employed as a foundational step in text processing. For this purpose, it was used a Bag of Words (BoW) approach. The Bag of Words model is a widely used approach in Natural Language Processing (NLP) for structuring text data to make it suitable for machine learning algorithms. It transforms text into a numerical representation by occurrences, ignoring grammar and word order. In the context of the Bag of Words model, were utilized both unigrams and bigrams.

Unigrams refer to single words treated as individual tokens, while bigrams involve pairs of consecutive words, capturing relationships between neighboring terms in the text. The combination of unigrams and bigrams allows for a richer representation of the text, as it not only considers individual word occurrences but also their sequential associations. This dual approach improves the understanding of linguistic structures and provides a more comprehensive foundation for model fitting. As a result of the tokenization process, 8,920 unigrams and 10,609 bigrams were obtained, which together account for a total of 19,521 tokens.

3.2.4. Topic Extraction

Topic modeling is a common method in natural language processing and machine learning used to identify topics within a set of documents. A collection of documents is called a corpus. One of the most popular algorithms for these purposes is Latent Dirichlet Allocation (LDA) which is used for this work. LDA is a probabilistic model that treats documents as a combination of topics. It assumes that every document is made up of multiple topics, and each topic has its own unique distribution of words. Each document within the corpus consists of words or tokens. LDA operates on the assumption that there are a fixed number of topics in the corpus, and each topic is represented as a probability distribution over words. The words in a document are thought to be generated from this mixture of topics.

In this study, the raw text data was transformed into a Document-Term Matrix (DTM), where each row represents a document, and each column represents a unique word,

showing the frequency of that word in the document. However, this matrix is highly dimensional and sparse, as not every word appears in every document.

The LDA algorithm reduces this complexity by converting the DTM into a smaller, more manageable matrix that represents topic distributions. In this new matrix, each row corresponds to a document, and each column indicates the extent to which the document relates to a particular topic.

To construct a high-quality Latent Dirichlet Allocation (LDA) model, it is crucial to determine the optimal number of topics. This is typically achieved using the metrics Coherence and Perplexity, which are widely regarded as standard measures for evaluating topic models. Coherence score measures the interpretability of the topics generated by the model. Higher coherence scores indicate that the topics are more meaningful and easier for humans to understand. On the other hand, Perplexity evaluates how well the model predicts unseen data, with lower perplexity values indicating better generalization and model performance. To identify the optimal number of topics, Coherence and Perplexity were calculated for a range of topic numbers, from 10 to 90, with increments of 5.

For this work, the number of topics in the LDA model was set to 30. This decision was based on the observation that at this level, the Coherence score reaches its maximum, ensuring the most interpretable topics, while Perplexity decreases almost linearly, reflecting a well-performing model.

Figure 2 below illustrates the behavior of these two metrics across the tested topic numbers, supporting the selection of 30 topics for the final LDA model.

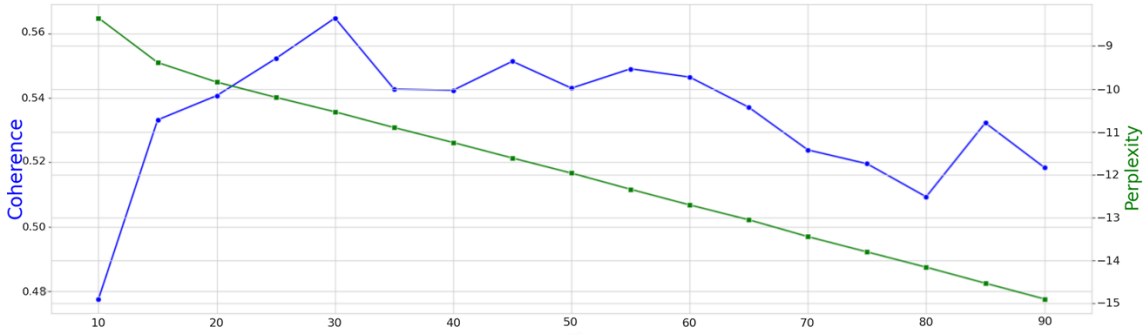

FIGURE 2 – Coherence and Perplexity evaluation.

The trained LDA model was then applied to infer the topic distributions for these aggregated time periods.

3.3. Data Exploration

3.3.1. Macroeconomic Data

First, the target variable, the Inflation Rate, was examined, along with the Interest Rate and the ideal inflation value (4%) set by the Central Bank of Russia. In Figure 3 below, we can observe a strong correlation between these two variables (Inflation Rate and Interest Rate). Additionally, it is evident that both values do not closely approach the target value of 4%, which is the ideal inflation rate set by Russia's Central Bank. Furthermore, based on the collected data, we can identify key critical points in the Russian economy during 2015 and 2022. In 2015, the economic trends suggest the significant impact of the annexation of Crimea at the end of 2014. Similarly, in 2022, the data reflects the effects of the aftermath of the COVID-19 pandemic.

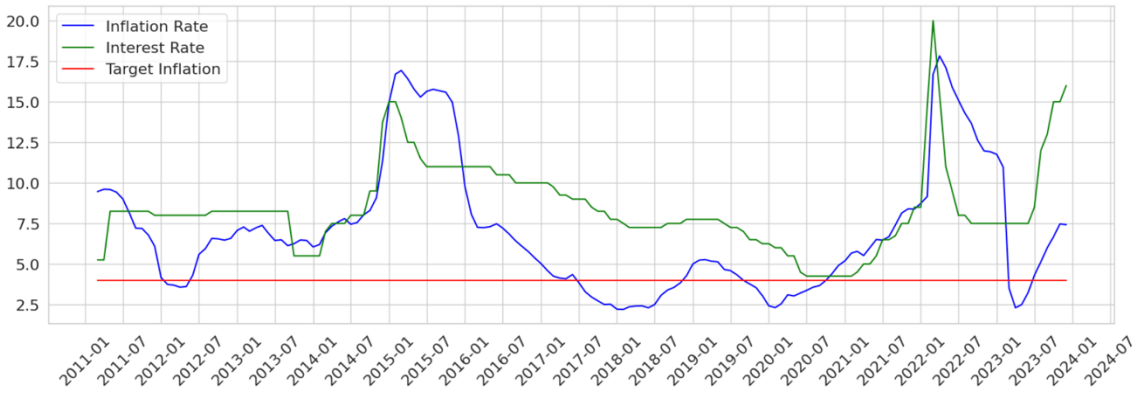

FIGURE 3 – Evolution of the monthly inflation and interest rates during the period covered by the analysis.

Figure 4 illustrates the primary macroeconomic features utilized in this study over time. The visual representation highlights that these features exhibit clear trends, indicating non-stationarity in their original form. To effectively apply certain machine learning models used in this study, such as AR(4), it is crucial to ensure that each time series feature is stationary. Stationarity, which requires consistent statistical properties such as mean and variance over time, is a fundamental requirement for accurately modeling time series data.

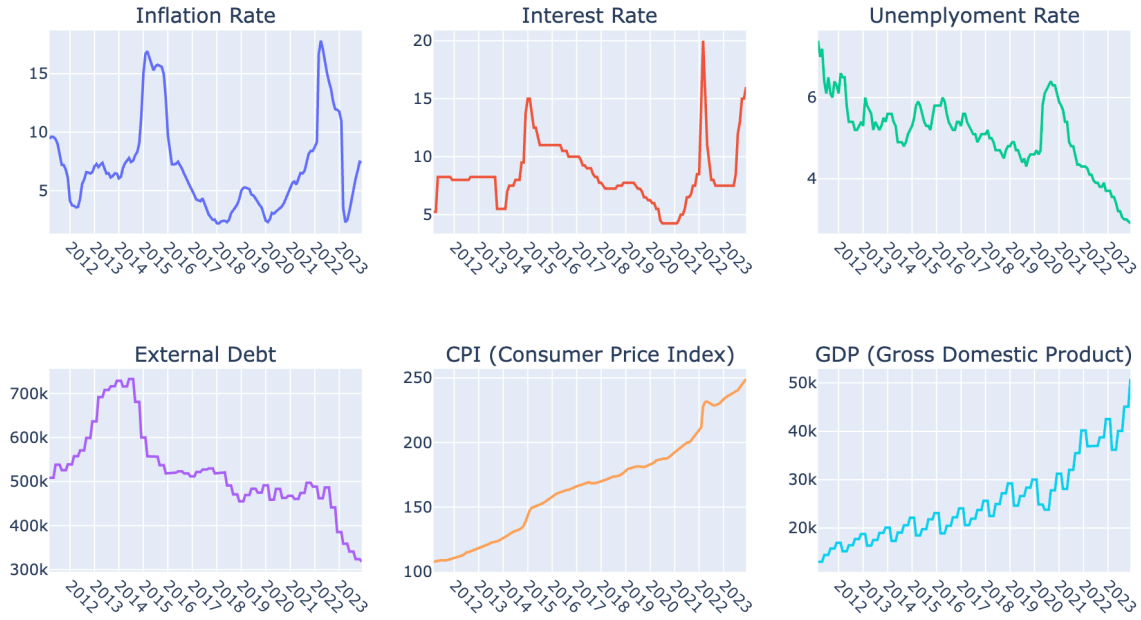

FIGURE 4 – The primary macroeconomic features utilized in this study.

To achieve stationarity, a differencing transformation was applied. An order of differencing check was performed using two widely used statistical tests, the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

The Augmented Dickey-Fuller (ADF) test is a statistical unit root test used to determine whether a time series contains a unit root, which would indicate that the series is non-stationary. A unit root indicates non-stationarity in a time series, but it does not always signify the presence of a trend component.

The KPSS test is also a type of unit root test, particularly useful when a series is close to having a unit root. Unlike the ADF test, its null hypothesis assumes stationarity. It assesses whether a series is stationary around a deterministic trend. While conceptually similar to the ADF test, the KPSS test is not interchangeable with it. Relying on only one of these tests may lead to incorrect conclusions about stationarity.

To mitigate these risks, both tests were applied in this study. A custom function was implemented to conduct both the ADF and KPSS tests for each feature in the dataset, identifying the number of differencing transformations required to achieve stationarity. This process was applied to all macroeconomic features and textual data (topics). The maximum number of differencing steps was capped at two, as this is typically sufficient for most time series.

The results of the differencing transformations can be observed in Figures 5 and 6 below, which visually demonstrate the impact of the transformation on the dataset. By applying the differencing transformation, the distribution of the features is normalized, reducing skewness and improving the stability of the data. This, in turn, will lead to more accurate results in subsequent modeling stages.

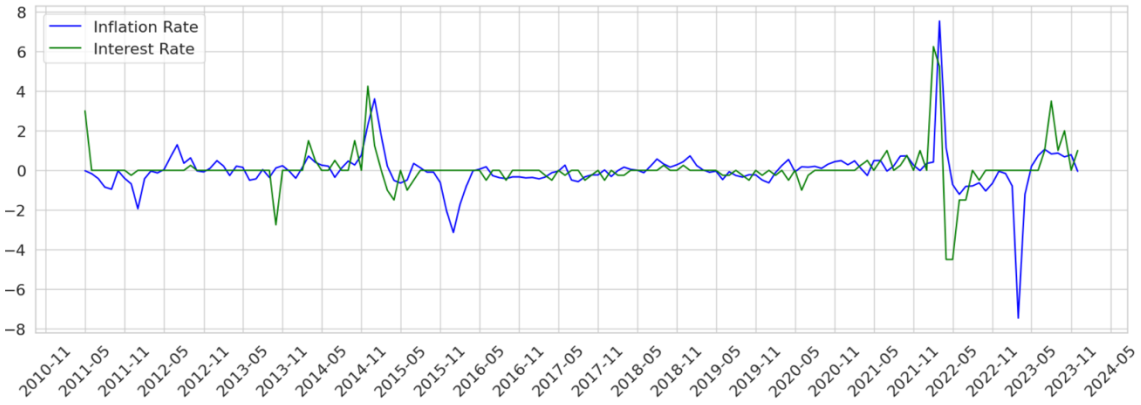

FIGURE 5 – First-order Differencing of Inflation and Interest Rate

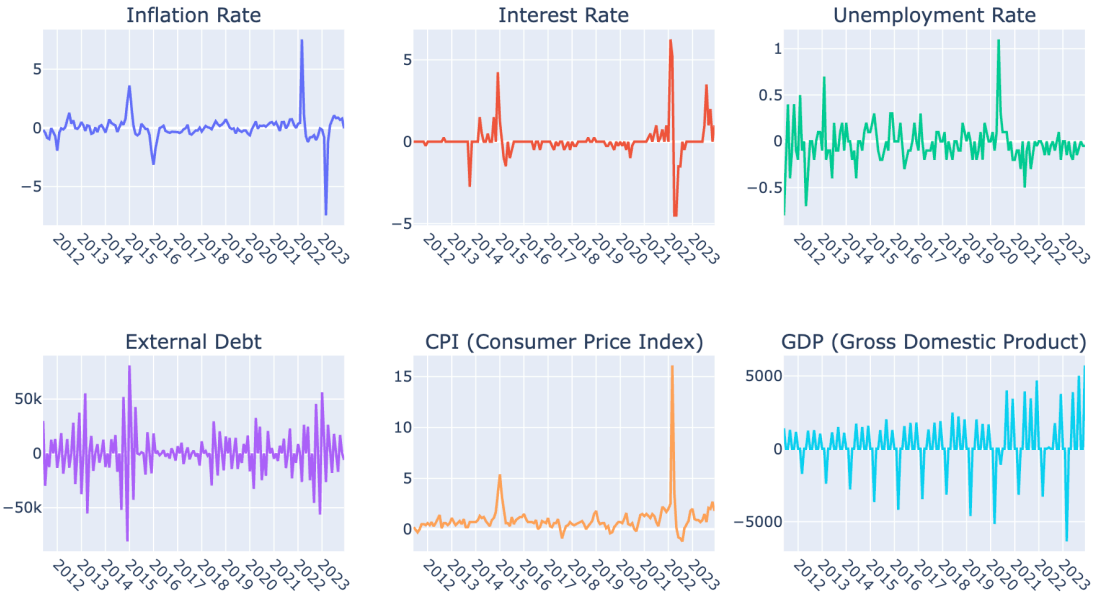

FIGURE 6 – Differencing of Macroeconomic Features.

In the next step, we examine the correlations between features and the target variable before making them stationary. The heatmap presented below (Figure 7) illustrates the correlation between each macroeconomic feature in the dataset. This visual representation provides a clear understanding of how these variables interact with the target variable, the

Inflation Rate. Notably, the features with the highest correlation to the Inflation Rate are the Interest Rate, which historically has a strong influence on inflation, the price of gas and the price of granulated sugar (price per kilogram).

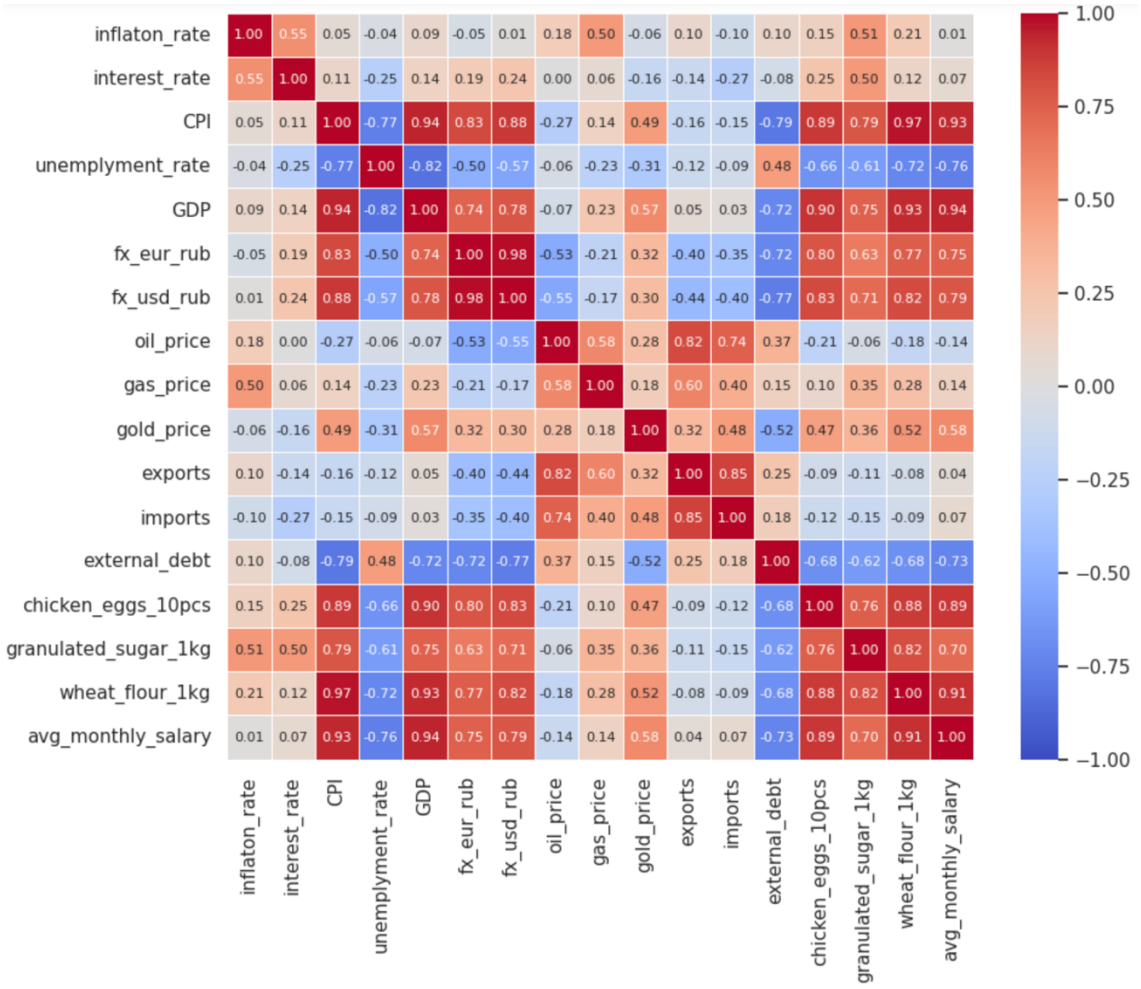

FIGURE 7 – Correlation matrix for macroeconomic features.

Furthermore, the heatmap reveals an interesting interdependencies among product prices and other features. Surprisingly, there is no strong negative correlation between Unemployment Rate and Inflation as it's often discussed in macroeconomic studies, however there is a strong negative correlation with prices on products, like wheat and sugar. The strongest correlation between target variable is with the following predictors: interest_rate, granulated_sugar_1kg, gas_price (0.55, 0.50, 0.51 respectfully). The least correlated features are avg_montly_salary, fx_usd_rub and unemployment_rate (0.01, 0.01, -0.04 respectfully).

3.3.2. Text Data

As previously discussed, we aim to expand the list of macroeconomic predictors by incorporating time series generated through Topic Modelling (LDA). For each calendar month and across 30 topics, the proportion of news dedicated to each topic was calculated by dividing the number of news articles on the topic by the total number of news articles for the same month. This approach allows us to track how news trends evolved over time.

In Figure 8, the distributions of the most frequently discussed topics throughout the study period are presented. These include:

- Topic 1: Discussions on sanctions and limitations.
- Topic 9: Pension payment amounts.
- Topic 24: Economic relations between Russia and the United States.

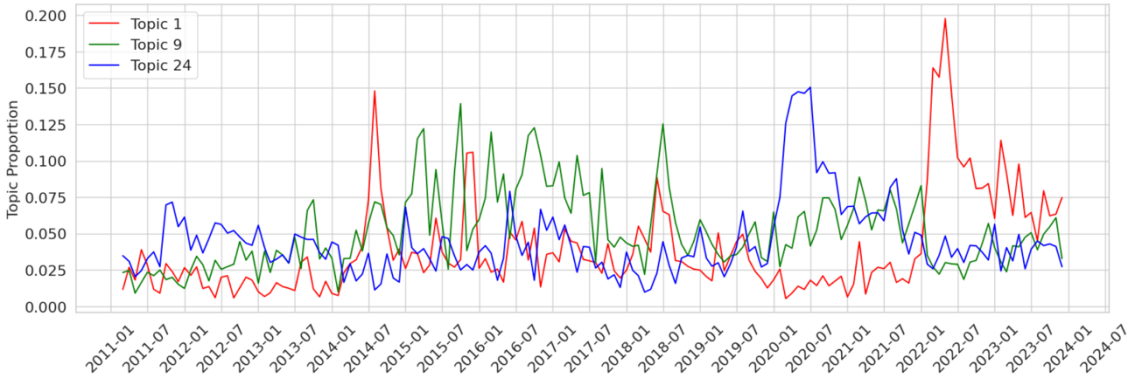

FIGURE 8 – Top-3 the most discussed topics frequency over time.

It is important to note that the news time series also exhibit noticeable trends. Therefore, similar to the macroeconomic indicators, we applied stationarization using the ADF and KPSS statistical tests. As a result, all numerical predictors were transformed into stationary time series.

Additionally, it was decided to analyze the correlations between the news topic time series and the target variable – the Inflation Rate. Based on the correlation matrix in Figure 9, it can be observed that the strongest relationships with the Inflation Rate are associated with 3 topics which are:

- Topic 1: Discussions on sanctions and limitations.
- Topic 20: Price of Rubble & Exchange Rate
- Topic 28: Russian President Putin

Notably, except for five topics (topics 9, 28, 22, 14, and 3), there are few topics whose correlation with the target variable is close to zero. This indicates that the identified topics effectively capture the relationship with inflation over time.

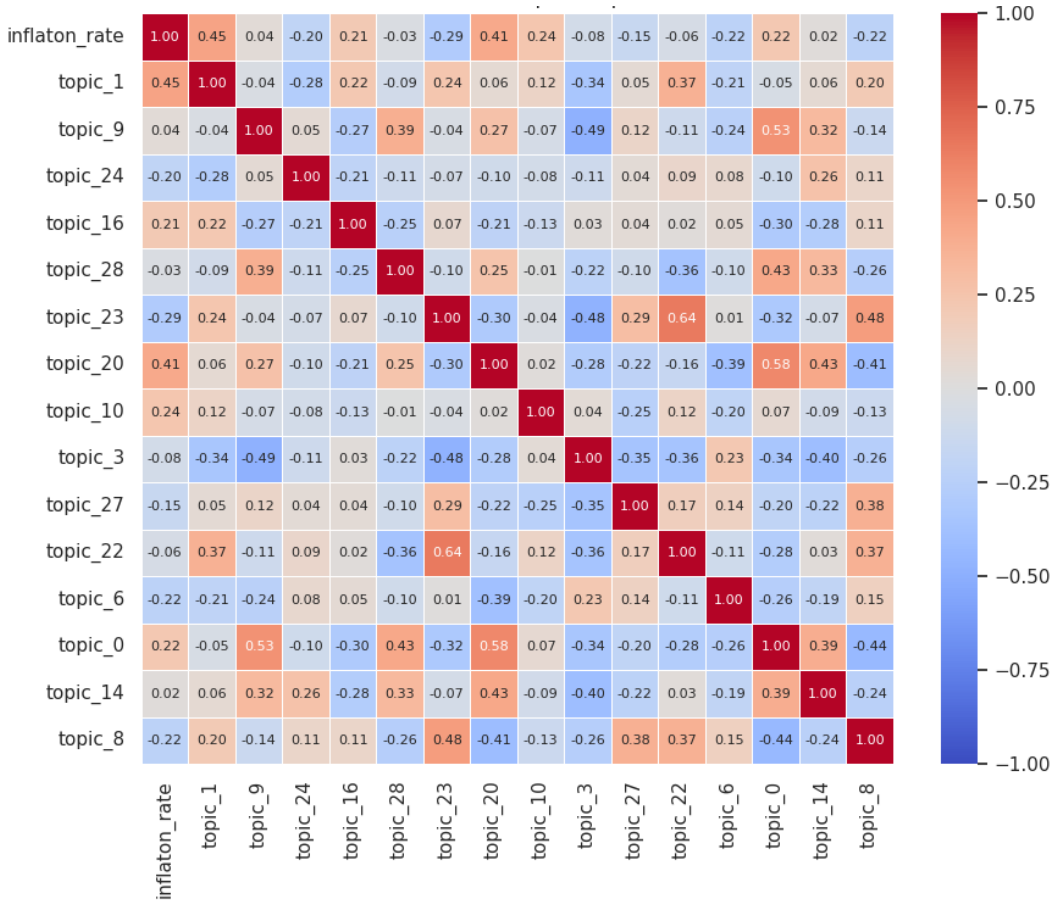

FIGURE 9 – Correlation matrix for the top-15 news topic time series features and the Inflation Rate.

To sum up, data exploration showed necessity of stationarity transformation for most of the Time Series Features. With use of statistical tests, number of differencing procedures was calculated for each of the predictor. Correlation analyses revealed the most and least important features inside two sets: Macroeconomic data and Texts data. The least important features, with 0 correlation with Inflation rate may be excluded from the final dataset on the next step.

3.3. Inflation Prediction

In this study, the following narrative-based forecasting model was utilized:

$$\pi_{t+1,t+h} = G(\theta_t) + \varepsilon_{t+1,t+h}, \quad (1)$$

where $\pi_{t+1,t+h}$ is the accumulated inflation from month $t+1$ to month $t+h$; θ_t is a vector consisting of 30 dimensions that represents the narrative topic distribution at time; $G(.)$ is the mapping from narratives to future inflation; $\varepsilon_{t+1,t+h}$ is a forecast disturbance; h is the forecast horizon.

In this study, five models were considered for the mapping $G(.)$, including Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net (ENet), Partial Least Squares (PLS), and the Autoregressive model of order 4 (AR(4)). These models will be discussed further in the text below.

I first examine the Random Forest (RM) model. It works by generating a set of decision trees, each built by selecting a random subset of features and data samples, with each tree making its own prediction. The final forecast is determined by aggregating the predictions from all the trees, helping to ensure that the model is not overly influenced by any single variable or anomaly. As a robust and user-friendly machine learning algorithm, Random Forest typically delivers strong results without the need for extensive hyperparameter tuning, making it a reliable and efficient choice for various prediction tasks. Let B represent the number of samples. For each sample b, a decision tree is constructed using the economic data estimated by $\pi_{t+1,t+h}$ and a randomly selected subset of the feature vector θ_t. The final mapping between the economic narratives and the inflation rate is derived by averaging the predictions of all B decision trees, as follows:

$$\hat{G}(\theta_t, B, L, N) = \frac{1}{B}\sum_{b=1}^{B} \widehat{g_b}(\theta_t; L, N), \quad (2)$$

where L represents the depth of the trees, N is the number of features randomly selected at each split and $\widehat{g_b}$ is the prediction function of the b-th decision tree in the Random Forest model.

The Least Absolute Shrinkage and Selection Operator (LASSO) is a linear regression model that incorporates a penalty term to achieve both regularization and feature

selection. This penalty term adds the sum of the absolute values of the regression coefficients to the loss function, effectively shrinking some coefficients to zero while retaining others. As a result, a model only includes the most important features, making it easier to interpret and reducing the risk of overfitting. In light of this, let's consider $G(\theta_t) = \beta'\theta_t$, and the parameter vector β' can be estimated by

$$\hat{\beta} = \arg\min \Big(\sum_{t=1}^{T} (\pi_{t+1,t+h} - \beta'\theta_t)^2 + \lambda \sum_{k=1}^{K} |\beta_k| \Big), \tag{3}$$

where λ is a tuning parameter.

The Elastic Net (ENet) is a linear regression model that combines both Lasso (L1) and Ridge (L2) regularization methods. It is particularly useful when there are highly correlated features or when the number of features is greater than the number of observations. By balancing the benefits of both Lasso and Ridge, Elastic Net tends to perform well in a wide range of prediction tasks, especially when dealing with datasets that might have many variables and collinearity among them. Its β can be estimated by

$$\hat{\beta} = \arg\min \Big(\sum_{t=1}^{T} (\pi_{t+1,t+h} - \beta'\theta_t)^2 + \alpha\lambda \sum_{k=1}^{K} \beta_k^2 + (1-\alpha)\lambda \sum_{k=1}^{K} |\beta_k| \Big), \tag{4}$$

Partial Least Squares (PLS) is a model that is used to describe the relationship between a set of independent variables (predictors) and one or more dependent variables (responses). In PLS, the goal is to find a small number of latent variables (also known as components) that explain both the variance in the predictors and the response variables. These latent variables are linear combinations of the original features, constructed in such a way that they capture the most significant directions of variance in both the predictors and responses. Partial Least Squares is a powerful tool for prediction, especially when dealing with high-dimensional, collinear, or noisy data, making it a solid choice for modelling inflation based on economic narratives. It is assumed that all predictors are driven by some latent factors, namely

$$\theta_t = \Lambda f_t + e_t, \tag{5}$$

where f_t is an n-dimensional vector of principal components from θ_t and n is much smaller than the number of predictors. The factor f_t can be estimated by maximizing the

shared variation among the predictors. Consequently, with a linear mapping $G(\theta_t)$, equation (1) can be expressed as follows:

$$\pi_{t+1,t+h} = \gamma' f_t + \varepsilon_{t+1,t+h}. \quad (6)$$

The aforementioned machine learning algorithms include several hyperparameters. In the LASSO and ENet methods, the penalty parameter λ serves as a key hyperparameter. Additionally, for ENet, there is an extra hyperparameter that specifies the balance between L1 (Lasso) and L2 (Ridge) regularization. For the factor model (PLS), the number of factors is a hyperparameter. In the RF model, the depth of the trees (L) is a hyperparameter, while N is set to the total number of features.

These hyperparameters were determined using a grid search technique with cross-validation. Grid search is a traditional method for hyperparameter tuning in machine learning, where every combination of provided hyperparameter values is exhaustively tested to identify the best-performing model.

Cross-validation is a widely used technique in machine learning to evaluate model performance on unseen data. In this case, 5-fold cross-validation was applied. It involves splitting the dataset into multiple folds or subsets, with one fold used as the validation set and the remaining folds used for training the model. This process is repeated multiple times, each time using a different fold as the validation set. The results from each validation step are then averaged to produce a more robust estimate of the model's performance.

By combining these methods, performance metrics were calculated, and the best combination of hyperparameters for each model was identified.

3.4. Feature Selection

Before selecting features for building inflation forecasting models, it was necessary to identify the key indicators typically used to predict inflation dynamics. Based on the correlation heatmap, features with a low correlation (absolute value below 0.05) with the target variable (Inflation Rate) were excluded. It is worth noting that not all lag features demonstrated a correlation above the threshold, leading to the automatic removal of some lagged variables. Lastly, Lasso, Enet and RF employed in this study incorporate internal feature selection mechanisms based on the predictive power of each variable.

3.5. Evaluating Forecasting Models

For the comparison of predictive models, the Autoregressive (AR) model is used in this work as a benchmark, with the lag length fixed at four lags. The Autoregressive model of order 4 (AR4) is a time series model that represents the current value of a variable as a linear combination of its previous four values and a stochastic error term. This model is particularly effective for capturing temporal dependencies in data, where the past behaviour of a variable influences its future values. To begin with, the individual forecast for each model was estimated, serving as the basis for subsequent evaluation, as $\pi_{t+h|t} = \hat{\theta}_{0,h} + \hat{\theta}_{1,h\pi_t} + \cdots + \hat{\theta}_{4,h\pi_{t-3}}$, and then the forecasts were aggregated to obtain the accumulated predictions.

The performance of the narrative-based inflation forecasts was evaluated using the Root Mean Squared Error (RMSE) for model m, which was defined as:

$$RMSE_{m,t+1,t+h} = \sqrt{\frac{1}{T-T0+1}\sum_{t=T0}^{T}(\hat{e}^{m}_{t+1,t+h})^2}, \quad (7)$$

where $\hat{e}^{m}_{t+1,t+h} = \pi_{t+1,t+h} - \hat{\pi}^{m}_{t+1,t+h}$ is the forecasting error; $\hat{\pi}_{m,t+1,t+h}$ is the forecasting value for the next *h*-month inflation rate made by model *m*.

Root Mean Squared Error (RMSE) is a standard metric used to evaluate the accuracy of a model's predictions. It measures the square root of the average squared differences between the predicted values and the actual values. RMSE provides an indication of how well the model fits the data, with lower values representing better predictive accuracy. It is particularly useful for comparing the performance of different models.

For comparison, the relative RMSE (rRMSE) was calculated as the ratio of the RMSE of a given model to that of the AR model. It was defined as:

$$rRMSE_{m,h} = \frac{RMSE_{m,h}}{RMSE_{AR,h}} \quad (8)$$

where $RMSE_{m,h}$ represents the RMSE of one of the observed models.; $RMSE_{AR,h}$ is the RMSE of the AR model in forecasting the next *h*-month inflation rate.

Root Relative Mean Squared Error (rRMSE) is a metric used to evaluate the performance of predictive models by comparing their error to the scale of the actual values. It normalizes the RMSE by the mean of the true values, offering a relative measure of prediction accuracy. This makes RRMSE useful for assessing model performance, as it provides a clearer picture of how well the model performs in relation to the actual data.

To compare the forecast performance of different models, the Diebold-Mariano (DM) test was employed. The Diebold-Mariano (DM) test is a statistical method used to compare the forecast accuracy of two models. The test evaluates whether the forecast errors of two competing predictions differ significantly, which helps determine if one forecasting method is superior to the other.

In this work the DM test compares the predictive accuracy of each model (RF, LASSO, ENet,PLS) against the AR4 model, which serves as the benchmark. The test evaluates the differences in forecast performance by analyzing the loss differentials, calculated as the differences in squared errors between the model predictions and the benchmark. To ensure robust statistical inference, adjustments are applied for serial correlation in the forecast errors using the Newey-West correction.

Firstly, forecast errors were computed as the difference between true values and predicted values for each model and forecast horizon (3, 6, 9, and 12 months). Then for each model (excluding AR(4)), the loss differential was calculated as the squared difference between the forecast errors of the model and those of the AR(4) benchmark. These loss differentials form the basis for testing predictive accuracy differences.

The loss differential was defined as:

$$d_t = e_{m,t}^2 - e_{AR4,t}^2 \tag{9}$$

where d_t is the loss differential at time t; $e_{m,t}$ is the forecast error of the model being evaluated; $e_{AR4,t}$ is the forecast error of the AR4 benchmark.

Using the mean and variance of the loss differentials, the DM test statistic was calculated. To ensure robust inference, the Newey-West adjustment was applied to account for any serial correlation and heteroskedasticity in the forecast errors.

Based on the DM test statistic, one-sided p-values were computed to evaluate the null hypothesis that the model's predictive accuracy is equivalent to the AR(4) benchmark.

Statistical significance was denoted as follows: one star symbol for $p < 0.05$ (5% significance level), two for $p < 0.01$ (1% significance level), and three for $p < 0.001$ (0.1% significance level).

4. RESULTS

After completing all the preliminary steps and determining the metrics and evaluation methods for the models, a Recursive Learning Loop was implemented. The concept is straightforward: using 85% of the dataset (130 observations) as the training set, four models—LASSO, PLS, ElasticNET, and Random Forest Regressor—were trained with use of rolling window approach.

The prediction process followed a specific logic. A set of predictors was taken for a single observation, and each model produced a prediction for the Inflation Rate for t+3, t+6, t+9 and t+12 horizons. Subsequently, the process was repeated till the end of test set. Between each prediction, a re-optimization of parameters was conducted using grid search and cross validation for each of the four models. In this way, each model predicted one step ahead, adjusted its parameters based on the real observed Inflation rate value prediction, and proceeded to the next forecasting step. This approach was used for three types of models: Macroeconomic Data only (M), Narratives data only (N), Macroeconomic + Narratives (M+N).

4.1. Forecasting with Macroeconomic Data

In this type of modeling approach, only macroeconomic features and their lag_1 values were utilized. This methodology can be referred to as a classical approach, as it relies on openly available numerical data representing the overall economic conditions of a country. The best-performing model, based on the rRMSE metric, was the Random Forest, achieving scores of 0.202, 0.209, 0.211, and 0.611 for 3, 6, 9, and 12-month horizons, respectively. Notably, the Diebold-Mariano (DM) test indicates a high level of confidence in these results, particularly for the 3-month horizon, where the p-value is below 0.01. However, the predictive accuracy diminishes for the 6- and 9-month horizons.

Conversely, the poorest-performing model was the Partial Least Squares (PLS) regression, with rRMSE scores of 0.529, 0.483, 0.467, and 0.469 for the respective horizons. None of the prediction samples generated by the PLS model passed the DM test. A more detailed breakdown of the performance metrics can be found in Table 3.

TABLE 3

model	Train Size	horizon (months)	RMSE	rRMSE	DM-test	Significance
RF	130	3	1.036	0.202	-5.350	***
RF	130	6	0.805	0.209	-2.526	**
RF	130	9	0.693	0.211	-2.199	*
RF	130	12	2.235	0.611	-2.203	*
LASSO	130	3	1.017	0.242	-1.908	*
LASSO	130	6	0.763	0.231	-2.153	*
LASSO	130	9	0.692	0.241	-2.227	*
LASSO	130	12	0.604	0.240	-2.012	*
ENET	130	3	1.319	0.314	-1.818	*
ENET	130	6	0.962	0.291	-2.135	*
ENET	130	9	0.836	0.292	-2.243	*
ENET	130	12	0.731	0.291	-2.026	*
PLS	130	3	2.217	0.529	-1.401	
PLS	130	6	1.594	0.483	-1.980	*
PLS	130	9	1.340	0.467	-2.237	*
PLS	130	12	1.180	0.469	-2.032	*

4.2. Forecasting with LDA Topics

For this approach, only Time Series data extracted from naratives and corresponding lag_1 features were used as a predictors set. Again best-performing model, based on the rRMSE metric, was the Random Forest, achieving scores of 0.281, 0.327, 0.362, and 0.698 for 3, 5, 6, and 9-month horizons, respectively. Notably, the Diebold-Mariano (DM) test indicates a high level of confidence in these results, particularly for the 3-month horizon, where the p-value is below 0.01. However, the predictive accuracy for the 6- and 9-month and 12-month horizons are stsatistically significant on 5% significance level only. This suggests that relying solely on textual data does not significantly compromise

the accuracy of inflation estimation. Moreover, it indicates that incorporating textual data has the potential to enhance the performance of the first model, which is based on macroeconomic indicators.

Conversely, the poorest-performing model was the Partial Least Squares (PLS) regression, with rRMSE scores of 1.080, 1.011, 0.981, and 0.977 for the respective horizons. None of the prediction samples generated of other models (LASSO, ElasticNet, PLS) passed the DM test. A more detailed breakdown of the performance metrics can be found in Table 4.

TABLE 4

model	Train Size	horizon (months)	RMSE	rRMSE	DM-test	Significance
RF	130	3	1.442	0.281	-4.470	***
RF	130	6	1.260	0.327	-2.298	*
RF	130	9	1.186	0.362	-1.981	*
RF	130	12	2.552	0.698	-1.666	*
LASSO	130	3	4.299	1.025	0.071	
LASSO	130	6	3.120	0.945	-0.273	
LASSO	130	9	2.594	0.905	-0.637	
LASSO	130	12	2.256	0.897	-0.810	
ENET	130	3	4.298	1.025	0.070	
ENET	130	6	3.119	0.945	-0.273	
ENET	130	9	2.594	0.905	-0.638	
ENET	130	12	2.255	0.897	-0.811	
PLS	130	3	4.531	1.080	0.226	
PLS	130	6	3.339	1.011	0.055	
PLS	130	9	2.814	0.981	-0.123	
PLS	130	12	2.456	0.977	-0.180	

4.3. Forecasting with Combined Data (Macroeconomic and Narratives data)

This stage of the research can be considered its culmination, as it is here that macroeconomic indicators are enriched with data derived from news texts over the entire study period. The initial hypothesis—that topic modeling and the resulting text-based

time series could enhance the predictive power of the models—was confirmed. Notably, this approach yielded significant improvements in the performance metrics of the Random Forest (RF) and Partial Least Squares (PLS) models.

For the RF model, the rRMSE score (Table 5) decreased to 0.192 for the 3-month horizon, 0.205 for the 6-month horizon, and 0.209 for the 9-month horizon, though there was almost no change for the 12-month horizon. This indicates that news headlines contribute more effectively to shorter-term predictions (3, 6, and 9 months).

It is worth mentioning that models such as LASSO and ElasticNet did not demonstrate any improvements. This suggests that the automatic feature selection in these models disregarded the characteristics derived from the text data, which is a surprising outcome.

TABLE 5

model	Train Size	horizon (months)	RMSE	rRMSE	DM-test	Significance
RF	130	3	0.984	0.192	-5.172	***
RF	130	6	0.790	0.205	-2.496	**
RF	130	9	0.684	0.209	-2.183	*
RF	130	12	2.258	0.618	-2.131	*
LASSO	130	3	1.017	0.242	-1.908	*
LASSO	130	6	0.763	0.231	-2.153	*
LASSO	130	9	0.692	0.241	-2.227	*
LASSO	130	12	0.604	0.240	-2.012	*
ENET	130	3	1.319	0.314	-1.818	*
ENET	130	6	0.962	0.291	-2.135	*
ENET	130	9	0.836	0.292	-2.243	*
ENET	130	12	0.731	0.291	-2.026	*
PLS	130	3	1.590	0.379	-1.557	
PLS	130	6	1.369	0.415	-1.791	*
PLS	130	9	1.129	0.394	-2.060	*
PLS	130	12	1.023	0.407	-1.881	*

In summarizing the results, it is important to emphasize once again that standalone models based solely on macroeconomic indicators or textual data are outperformed by models that combine these two types of predictors. The best-performing algorithm across all three tests was the Random Forest, reaffirming the findings of previous studies. Side by side comparison of the model can be found in Table 6.

Given the significant volatility of inflation in the Russian Federation, particularly during periods of military conflict, this approach has statistically proven to be effective. However, there remain a number of ideas and potential improvements that should be considered in future research efforts.

TABLE 6

model	horizon (months)	rRMSE (M)	rRMSE (N)	rRMSE (M+N)	% Improvement
RF	3	0.202	0.281	0.192	-4.960
RF	6	0.209	0.327	0.205	-1.930
RF	9	0.211	0.362	0.209	-1.280
RF	12	0.611	0.698	0.618	1.030
LASSO	3	0.242	1.025	0.242	0.000
LASSO	6	0.231	0.945	0.231	0.000
LASSO	9	0.241	0.905	0.241	0.000
LASSO	12	0.240	0.897	0.240	0.000
ENET	3	0.314	1.025	0.314	0.000
ENET	6	0.291	0.945	0.291	0.000
ENET	9	0.292	0.905	0.292	0.000
ENET	12	0.291	0.897	0.291	0.000
PLS	3	0.529	1.080	0.379	-28.270
PLS	6	0.483	1.011	0.415	-14.090
PLS	9	0.467	0.981	0.394	-15.770
PLS	12	0.469	0.977	0.407	-13.330

5. CONCLUSION

This study developed models capable of making statistically significant inflation predictions, achieving the initial objective of combining macroeconomic indicators with text-derived data. The comparison of three approaches revealed that while standalone models based solely on macroeconomic or textual data perform reasonably well, their predictive accuracy is outperformed by the hybrid approach that combines both data types. This finding underscores the importance of integrating textual narratives with economic indicators for more robust forecasting, particularly in the context of the Russian economy.

A key contribution of this project was the inclusion of sanctions as a variable, recognizing its critical impact on forecasting in the context of Russia. Given the country's unique economic conditions, including the ongoing state of conflict and the aggressive economic policies imposed against it, the incorporation of this factor was essential for creating realistic and relevant models.

The Latent Dirichlet Allocation (LDA) methodology enabled the extraction of topics from textual data, which were subsequently used to enrich macroeconomic indicators. This integration significantly improved the forecasting performance, aligning with findings from previous research. Among the tested models, Random Forest Regression consistently outperformed others, confirming its superiority in predictive tasks. In contrast, the other models demonstrated relatively weaker performance, emphasizing the importance of choosing appropriate algorithms for this type of analysis.

While the results are promising, there is significant potential for further improvements. Future research could expand the training dataset, incorporate additional features, and utilize texts from diverse sources to reduce bias in the narrative data. Moreover, exploring other more complex machine learning models, or experimenting with neural networks could further enhance the forecasting capabilities. These steps represent a natural progression for refining and extending the current framework.

In summary, this study not only confirmed the benefits of combining text and macroeconomic data but also provided a foundation for future research to build upon, with promising directions to improve both the scope and accuracy of inflation forecasting.

REFERENCES

Atkeson A., Ohania L. E. (2001). *Quarterly review*. Federal Reserve Bank of Minneapolis.

Babii, A., Ghysels, E., and Striaukas, J. (2021). *Machine Learning Time Series Regressions with an Application to Nowcasting*. Journal of Business and Economic Statistics.

Coulombe P. G., Leroux M., Stevanovic D., Surprenant S. (2022). *How is Machine Learning Useful for Macroeconomic Forecasting*. Journal of Applied Econometrics.

Diebold, F. X., Mariano, R. S. (1995). *Comparing Predictive* Accuracy. Journal of Business and Economic Statistics.

Diebold, F. X. (2015). *Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold–Mariano Tests*. Journal of Business and Economic Statistics.

Diebold, F. X., Shin, M. (2019). *Machine Learning for Regularized Survey Forecast Combination: Partially-Egalitarian Lasso and its Derivatives*. International Journal of Forecasting.

Gu, S., Kelly, B. and Xiu, D. (2020). *Empirical asset pricing with machine learning*. Working paper, University of Chicago.

Khemani B., AdgaonkarA. (2021). *A Review on Reddit News Headlines with NLTK Tool*. Journal of Business and Economic Statistics.

Medeiros M., Vasconcelos G., Veiga A., Zilberman E. (2019). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods. Journal of Business and Economic Statistics

Stock J. H., Watson M. W. (2007). *Why Has U.S. Inflation Become Harder to Forecast?* Journal of Money, Credit and Banking.

Yongmiao H., Fuwei J., Lingchao M., Bowen X. (2024). *Forecasting Inflation Using Economic Narratives.* Journal of Business and Economic Statistics.