Bank of Russia

The Impact of Negative News on Public Perception of Inflation

WORKING PAPER SERIES

No. 111 / February 2023

Alina Evstigneeva
Daniil Karpov

Alina Evstigneeva
Monetary Policy Department, Bank of Russia
E-mail: EvstigneevaAG@cbr.ru

Daniil Karpov
Volga-Vyatka Main Branch of the Bank of Russia
E-mail: dankarpov90@gmail.com

The authors would like to thank Ksenia Yudaeva, Alexander Morozov, Vadim Grishchenko, Henry Penikas, anonymous reviewers and participants of internal Bank of Russia research seminars for their helpful comments and suggestions.

Bank of Russia Working Paper Series is anonymously refereed by members of the Bank of Russia Research Advisory Board and external reviewers.

Photo: Maria Bagaeva

12 Neglinnaya Street, Moscow, 107016
+7 499 300-30-00, +7 495 621-64-65 (fax)
Bank of Russia website: www.cbr.ru

Contents

Abstract

This study presents a novel approach to distinguishing the news that has the greatest impact on households' perception of inflation. Narrowing down the long list of all news items to only the strongly negative requires taking into account the concept of rational inattentiveness by the implementation of a 'too costly to ignore' principle. To accurately determine important negative news, we use data from Russian Public Opinion Research Center (VCIOM) polls about the threats to and fears of Russians.

We obtain 10 negative news topics time series with the frequencies of the topics of Poverty, Inflation, Food crisis, Debt problems, Economic crisis, Geopolitics, Ruble devaluation, COVID, Domestic instability, and Unemployment. Additionally, we have also included the topic of Bank of Russia's communication measured as intensity of its communication.

We conduct feature importance analysis using four algorithms: Lasso, Random Forest, XGBoost, and Bayesian Structure Learning. Our results indicate that three main news topics contribute the most to both household inflation expectations and perceived inflation: Inflation, Economic crisis, and Ruble devaluation. Bayesian Structure Learning presents a deeper picture with possible differences in the formation of households' inflation expectations and perceived inflation. In answering the survey about price changes in the next 12 months, consumers may actually be answering a question about their confidence in the economy in general, which in Russia is closely connected to the external situation. On the other hand, in answering the question about inflation today and yesterday, consumers focus on actual prices and the current situation in the country.

We also report differences in higher and lower income households' perception of inflation. Namely, the subgroup of respondents with savings tends to pay greater attention to the dynamics of the ruble exchange rate. The subgroup of respondents without savings pays more attention to news about local price increases, probably due to the higher costs of these events for family budgets. The intensity of central bank communication, apparently, has no significant effect on households' perception of inflation.

The results are robust to a noise check and to alternative measures of inflation expectations and perceived inflation.

Key words: monetary policy, text analysis, inflation expectations.
JEL classification: C83, E52, D83.

1. Introduction

Household inflation expectations have attracted a lot of attention in recent years. Inflation expectations are one of the key variables affecting macroeconomic outcomes, and they have also been a significant object of central bank communication since monetary authorities recognised the power of communicating with the wider public. Better understanding of how households form their expectations might help to improve communication strategies and practices, and offer new dialogue patterns.

To contribute to the work devoted to better understanding the nature of households' inflation expectations, we start with three strong streams of research.

First, the well-known epidemiological model of inflation expectations by Carroll (2003), which changed economists' view on what really drives inflation expectations. Carroll introduces a strong doubt about the rationality of all agents' expectations and shows that households form their expectations by observing professional forecasters' opinions in the news media. Pfajfar and Santoro (2013) argue that most consumers who update their expectations do not revise them towards professional forecasts and that news on inflation plays only a small role in the accuracy of their expectations. As has been shown in later works (see Larsen et al, 2019), the news media are good predictors of inflation expectations. We are also influenced by a large corpus of papers focussed on the kinds of prices that drive household inflation expectations. The most important are prices for oil (Sussman et al., 2022), food, and energy (Arora et al., 2013).

Second, we consider papers related to the rational inattention of households (the idea, proposed by Christopher A. Sims (Sims, 2003), that economic decision makers cannot absorb all the available information but can choose which pieces of information to process). The most complete review of this subject is presented in a recent work by Maćkowiak et al. (2021). The theory of rational inattention postulates that agent can choose what kind and how much information to absorb. If the costs of not having information are small, it can be ignored. This situation is typical for developed countries with long and successful experience of inflation targeting: people pay only partial attention to inflation statistics, because consuming such information requires considerable effort (Mankiw et al., 2004; Carroll, 2001), and the potential financial cost of ignoring such information is negligible for most households in low-inflation countries (Cavallo et al., 2017). For developing countries situation is different. In particular, Arora et al. (2013) show that high and unanchored expectations in Peru are caused by a lack of knowledge about monetary policy measures and goals, about inflation dynamics, and an overreaction to news about inflation.

Further exacerbating the situation is the fact that central banks speak a language that requires special knowledge and is hard for the wider public to understand. As noted by Haldane (2017b), the linguistic complexity of central bank publications renders them accessible to only about 5-10% of the population.

Third, we consider the growing body of machine learning algorithms that have made it possible to extract public opinion from news, social media, and services such as Google Trends. Dealing with inflation expectations specifically, we can mention recent papers by Sahu and Chattopadhyay (2020), Tilly and Livan (2021), and Angelico et al (2022). The findings of these works which are most relevant to our approach are: that different news has a different impact on expectations, that expectations may be formed even by non-inflationary news (news that does not contain explicit 'inflation' terms), and that negative news has a greater impact on expectations (consumers are more receptive to negative news).

News-driven models of household inflation expectations can be improved by reducing the noise of information which is nonessential to consumers. Since noise reduction is one of the main problems for all big data models, in this study, we attempt to show the role of negative news in household inflation expectations, news which is too costly to ignore (i.e., news containing the terms 'crisis', 'recession', 'depreciation', etc.)

To test our hypothesis, we use Russian news media and data on household inflation expectations. We create a news database with 7,779 million news items collected from 28 Russian media agencies over a period of about nine years.

Our paper advances the literature in few ways. First, we propose a new approach to news selection to improve the quality of news-driven models by reducing the noise. This contributes to the growing body of machine learning based models for inflation expectations. Second, we improve the understanding of consumer behaviour by giving deeper evidence of how they form their perception of the economic environment and of why households pay too little attention to central bank communication.

This study advances the existing research on news-driven inflation expectations and contributes to the literature in few ways. First, we introduce a new approach to filtering news that is significant for economic agents. Second, we provide more evidence and shed more light on the concepts of rational inattentiveness and the epidemiological model of inflation expectations. Third, we develop a large news database of leading Russian media agencies trained on economic news.

2. Data

2.1. News database

We create a news database for the purposes of our research, as we have been unable to find a sufficiently large database with wide coverage and with major Russian regional media agencies included. The process of creating the big Russian news database consists of two parts:

- data mining;
- topic modelling.

2.1.1. Data mining

We start by compiling a list of 28 key federal media agencies and the largest regional media agencies with audience coverage close to that of the federal media. We use Medialogia[1] data for this task. Medialogia publishes monthly ratings of the most popular Russian media agencies, including regional coverage. These ratings give us the

[1] Medialogia is a Russian company that develops an automatic system for monitoring and analyzing media and social networks in real time.

most complete picture possible of the preferences of Russians in the consumption of information. Our list of 28 is composed of 26 federal media agencies and 2 regional agencies. We then extract references to the news items of all 28 agencies for the period between 01.2014 and 08.2022. These references include the name of the agency, the news headline, the time of news publication, a link to the news (URL), and the category of the news (where available). More than 91% of these references (for 23 out of 28 agencies) are obtained using the Feedly API.

It is important to note that we initially upload only economic news from three media agencies (namely Interfax, Regnum, and RBC, thanks to the RSS settings of these websites), which enables us to compile a vocabulary of economic terms. Also, as there are categories for certain news items, we are able to apply supervised learning algorithms in the second stage.

The next stage is to retrieve the text of the news items via the links. For this process, we analyse the HTML codes of the media websites to identify the text of the news in the common code. The resulting news texts then cleaned of ‘rubbish’ (characters other than punctuation marks, letters, and numbers) and saved to the database. After the first stage, the database consists of more than 7,779 million news items from 28 news sources over a period of about nine years. The media agency names, their residences (federal or regional), and the number of news items included are presented in Appendix 1.

2.1.2. Topic modelling

The second stage is to highlight economic news. This can be accomplished in two main ways, through unsupervised and supervised learning (using premarked news).

Generally, unsupervised learning models, in particular, Latent Dirichlet Allocation (LDA, described by Blei et al, (2003)), are used to select different topics from an array of news items by means of so-called ‘soft clustering’. However, as there are categories for some news items, it is possible to mark up parts of the data and isolate the economic theme from them for further supervised learning. This potentially improves the quality of topic modelling.

We take both approaches and choose the best one. Before the topic modelling itself, we carry out standard pre-processing of the text, which includes tokenisation (i.e., splitting the text into individual words and punctuation marks) and lemmatisation (the reduction of words to their basic forms). Then, depending on the task, numerical lemma metrics are derived: for supervised learning - TF-IDF, and for unsupervised learning – Count Vectoriser.

TF-IDF (TF means 'term frequency' and IDF means 'inverse document frequency') is a statistical measure that is used to assess the importance of a word in the context of a document that is a part of a collection or corpus of documents (Salton, 1988):

$$TF-IDF = \frac{n_t}{\sum_k n_k} \log\left(\frac{1+n}{1+df_t}\right), \text{ where} \quad (1)$$

n_t – is the number of times word t appears in the document

$\sum_k n_k$ – is the number of words in the document

n – is the number of documents in the dataset

df_t – is the number of documents in which word t occurs

TF-IDF provides information on the importance of a particular word in determining the differences between groups of texts. This statistical measure carries more information than a simple 'bag of words' (a set of words with quantitative characteristics without regard to grammar, word order, etc.) (Harris, 1954)), but it is much faster than word2vec[2] and other neural network-based methods, which is important when working with a large data set. This measure is used for the supervised learning models. TF-IDF filtering is also performed: a word must occur at least 100 times in the corpus (this is how we filter out highly specialised terms and misspellings) and in 80% of documents at most (this is how we filter out the most common words, such as prepositions and conjunctions, which are not important for determining the topic of a text). Filtering increases the speed of model training by reducing the dimensionality of the resulting matrix.

For the purpose of pre-processing in unsupervised learning (we use LDA), we apply Count Vectoriser, a method that converts a pool of documents into a matrix, in which the columns correspond to particular words and the rows correspond to texts.

Next, we use a modification of the LDA model – Guided LDA (or Seeded LDA, which is a representation of classic LDA and is described by Jagarlamudi et al. (2012)). The essence of this model is that the vocabulary for a topic is used to induce a shift in the right direction, and to shift the selection of topics in the document to the topics related to the seed words. Guided LDA is thus suitable for highlighting the economic theme in a news corpus.

The dictionary of economic words is compiled on the basis of the vocabulary of the words used most frequently in the original corpus of economic news from the three media agencies mentioned previously. Additionally, the vocabulary is expertly cleaned of words not related to economics. The resulting vocabulary is used to form theme 0. The other themes have no initial shift in vocabulary, so they are generated by standard LDA algorithms. The training process is performed on a sample of 100,000 news items from 25 sources, which are equally represented in the sample (excluding the three media agencies mentioned previously to ensure that the distribution of themes is not skewed towards economic news).

The marked sample for the supervised learning models is formed on the basis of the top news categories (which cover almost half of all news items in the database), grouped into six main themes (see Appendix 1), one of which is devoted to the economy. The training is done on a sub-sample of 120,000 news items, with each of the six themes presented in equal shares.

[2] A family of algorithms that use neural networks to produce vector representations of words.

The methods used for supervised learning are: support vector machine (SVM) (Boser et al., 1992), softmax regression, random forest (Breiman, 2001) and gradient boosting (XGBoost) (Chen and Guestrin, 2016). This choice of models is based on the fact that the results of softmax regression can be interpreted as a baseline solution, the support vector machine is a classic method for text analysis problems, the random forest is the main ensemble method over decision trees, and gradient boosting is currently the most popular method for classification tasks and often yields the best results.

A total of five models are evaluated: four 6-class models and one 2-class model to compare the effect of the number of themes on the value of the metric. For the 2-class model, a specification is used in which 0 represents economic news and 1 represents the remaining categories, and respectively the test part for this model is also represented by two classes. All supervised learning models are trained on cross-validation (model evaluation in which the data are divided into k equal parts, learning occurs in part k-1, and quality evaluation is conducted in part k, with each of the k parts serving for both quality evaluation and learning, the procedure is repeated k times) with 5 splits and premixing.

The quality of the Guided LDA model is tested on a marked test sample for the 2-class supervised learning model. In order to assess the metrics of the economic theme, the Guided LDA themes are also grouped into group 0 for the economy, and group 1 for other themes. In this case, the use of classic metrics for thematic modelling tasks (e.g., coherence) does not make sense, as there is a marked dataset for quality checks. The main quality metrics chosen are precision, recall, and F-score for the economic theme.

The Guided LDA models with seven and ten themes perform the best on the F-score, and the model with seven themes is taken for further comparison. The trained supervised learning models perform better than Guided LDA, which confirms that reinforcement learning models yield better results.

The SVM and Gradient Boosting (XGBoost) models for six classes yielded the highest F-scores on economic topics, showing an equal result of 0.94 on the test sample (see Appendix 1 for details). However, Gradient Boosting yields the best result for precision and also, importantly, produces forecasts much faster than SVM, so the former model is chosen for further work. A total of 1.567 million economic news items from 28 media outlets over the period from 01.2014 to 08.2022 are represented in the corpus.

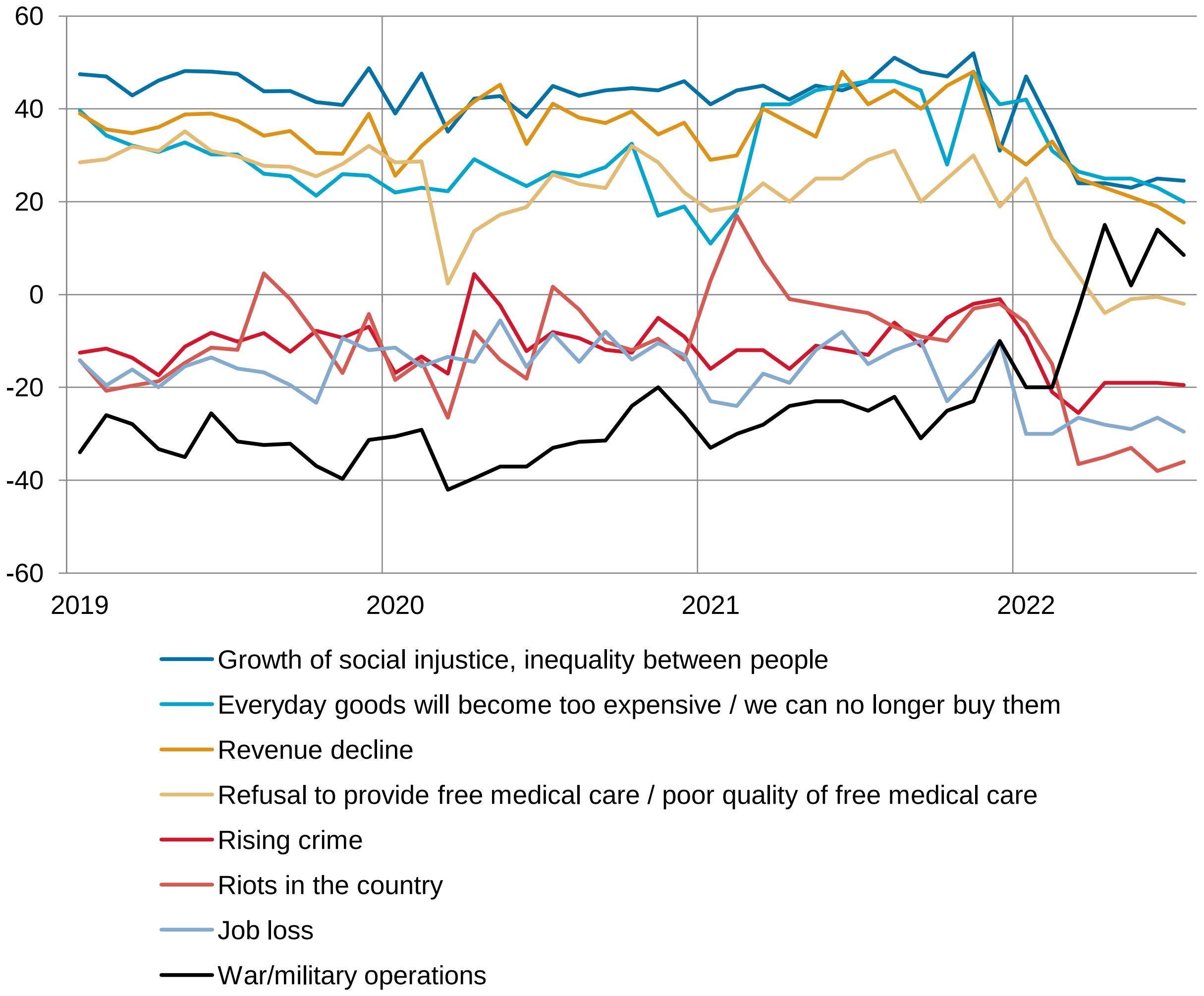
2.2. Defining negative semantic load

Determining the type of news that has the greatest effect on households' decisions may be one of the most difficult challenges for researchers (Larsen et al., 2019). As shown by Coibion et al. (2019), specific monetary policy news is potentially ineffective, because households are not typically interested in central bank policies. We confirm the findings of this study in the following chapters.

Our hypothesis proceeds from the idea that people may form their expectations on more general and strongly negative news, and therefore our search is focussed on the news items that are too expensive to ignore. This approach potentially allows us to bypass the restrictions of rational inattention. To accurately determine important negative news, we use data from VCIOM[3] polls about the threats to and fears of Russians. The survey results are presented in figure 1.

For negative news topic modelling, we use the Guided LDA described by Jagarlamudi et al. (2012). We defined the anchor terms from the VCIOM survey. These are subjectively chosen nouns that in our opinion concentrate the greatest semantic load from VCIOM's categories.

Figure 1. Fears of and threats to Russians according to VCIOM survey data[4]



[3] Russian Public Opinion Research Center (VCIOM). VCIOM is the oldest polling institution in post-Soviet Russia and one of Russia's leading sociological and market research companies.

[4] The index is measured in points and can range from -100 to 100. The answer 'I'm quite sure it will happen' is assigned a coefficient of 1, the answer 'It is likely to happen' is assigned a coefficient of 0.5, the answer 'It is not likely to happen' is assigned a coefficient of -0.5, and the answer 'I'm quite sure it won't happen' is assigned a coefficient of -1. The higher the value of the index, the likelier Russians think the problem to be. The data are presented on the basis of Sputnik all-Russian telephone polls. The marginal error is 2.5\%, and the sample size is 1600 respondents from at least 80 regions of the Russian Federation.

2.3. Seeding negative tokens

As the next step, we use the anchor terms from the previous step as seeded words (unigrams and bigrams) for Guided LDA. We slightly extend the given terms from 8 categories to 10 blocks for seeding. We do so to improve the model's explanatory ability and due to the high concentration of topics in several of the raw categories (i.e., there are three topics in the second line: prices, debt, and food crisis). Additionally, we allocate ruble devaluation as an independent topic due to the great importance of this variable to Russians[5]. Finally, we merge the topics of crime and political instability. As result, we get a list of 10 topics with probability distributions of the terms. Also, in order to assess the possible role of communication of the Bank of Russia on the topic of monetary policy, we added a topic corresponding to communication. We defined it as news in which individual members of the Board of Directors or the Bank of Russia comment on inflation.

2.4. Topic time series

After defining the topics list, we calculate the frequency of each topic in our dataset as follows:

$$Freq_{i,t} = \frac{N_{i,t}}{\sum_k N_{k,t}} \text{, where} \qquad (2)$$

$\mathrm{Freq}_{i,t}$ – is the frequency of topic i in month t

$\mathrm{N}_{i,t}$ – is the number of news items containing topic i in month t

$\sum_k N_{k,t}$ – is the total number of news items at time t

Figure 2 shows the time series (topic frequencies) obtained. The more often a topic occurs during a month, the higher the frequency score obtained.

This approach differs from those proposed by Larsen et al. (2019) and other researches, where media attention for a given topic is measured by the weight of the LDA unit in the whole corpus as one document. In our opinion, this approach may overestimate the weight of topics due to the specific structure of the news. Specifically, a topic usually appears in a single news item several times with different details, but the media consumer would not perceive this as several news items. The news item is usually perceived as one whole unit, especially if only the headline is read. There is evidence that people have a strong preference for skim-reading the headlines only, as shown by work by Gabielkov et al. (2016), a survey conducted by the Media Insight Project, an initiative of the American Press Institute (API) and the Associated Press-

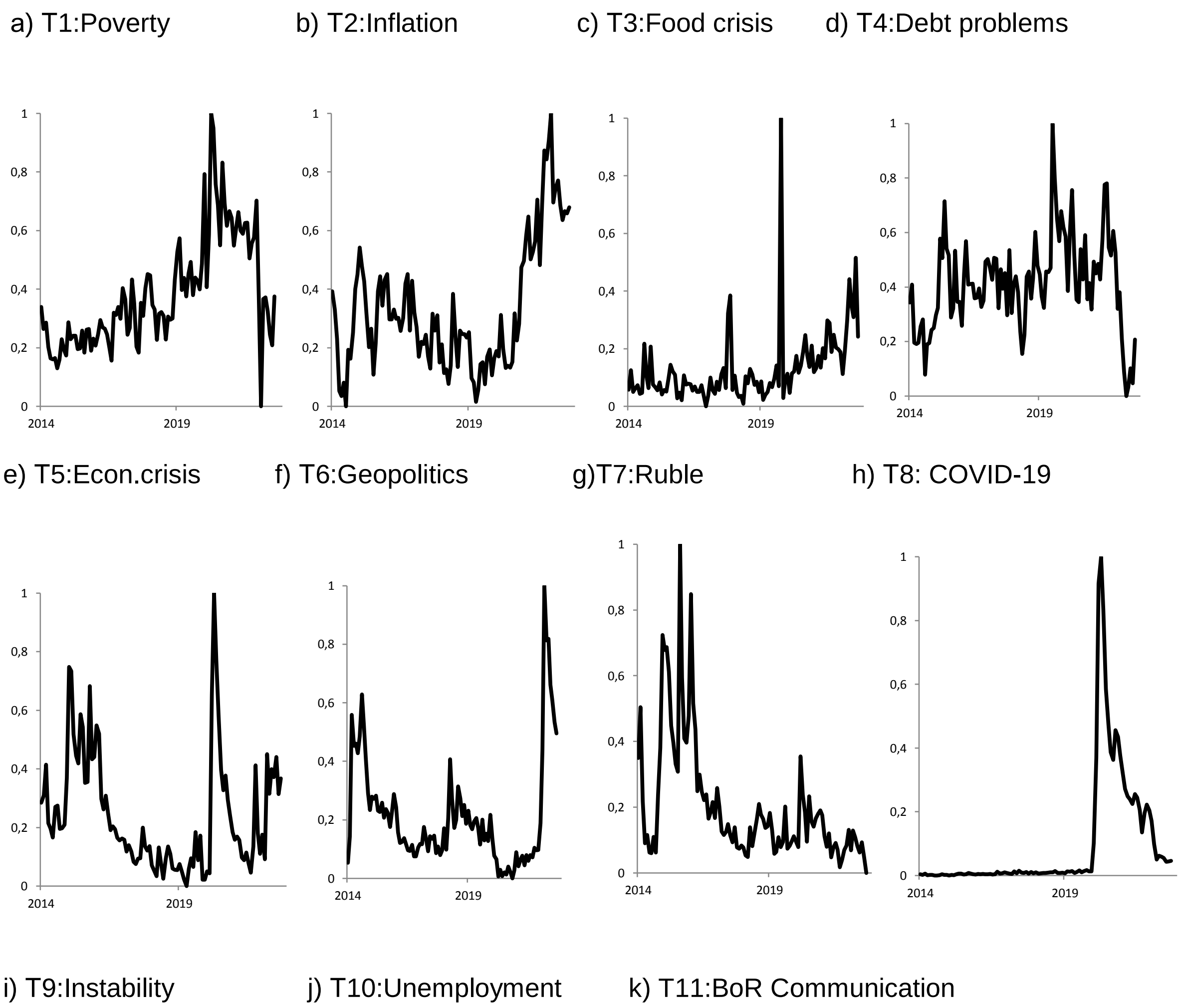
[5] According to FOM surveys, up to 18% of Russians base their inflationary expectations on fluctuations of the ruble. For further details, see https://www.cbr.ru/analytics/dkp/inflationary_expectations/#a_102610

NORC Center for Public Affairs Research (2014) for the USA, and the Nielsen Norman Group's report 'How People Read Online: The Eyetracking Evidence'.

Following Martin and Johnson (2015), we also manually concentrated the nouns in the Guided LDA results by reducing the occurrence of other parts of speech in order to reduce the noise in the data.

Figures 2-12 and the correlation matrix in Table 1 show that there are several correlated variables, such as T8: COVID and T10: Unemployment (0.81 Pearson correlation coefficient), T1: Poverty and T8: COVID (0.76), and T1: Poverty and T10: Unemployment (0.64). The common trends in the graphs appear even in spite of the fact that there are no common terms in those topics. Our interpretation of is this is that, during the COVID-19 pandemic, especially in the acute phase of early 2020, Russia, like many other countries, locked down many public places, so many people engaged in the service sector were afraid for the security of their jobs. Thus, topics related to COVID, poverty, and unemployment appear together in the news (but only over the short term of the first half of 2020). Despite these strong correlations, we do not merge topics to provide better explanatory ability to the models.

Figures 2-12. Topic time series[6]



[6] The time series are normalised.

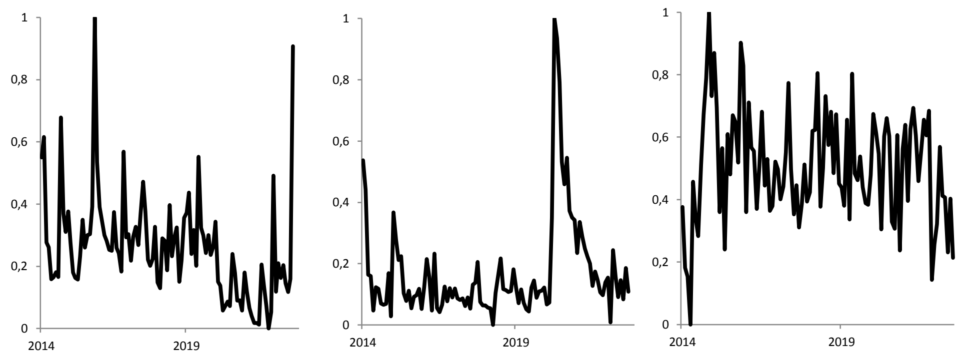


Table 1. Topic correlations

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
T1	1,00	0,22	0,21	0,43	0,04	-0,46	-0,30	**0,76**	-0,36	**0,64**	0,05
T2	0,22	1,00	0,36	-0,13	0,18	0,19	0,04	0,10	-0,14	0,01	0,19
T3	0,21	0,36	1,00	-0,06	0,05	0,18	-0,22	0,13	-0,18	0,11	-0,11
T4	0,43	-0,13	-0,06	1,00	-0,08	-0,63	-0,02	0,25	-0,16	0,17	0,13
T5	0,04	0,18	0,05	-0,08	1,00	0,16	0,57	0,39	-0,02	0,57	0,20
T6	-0,46	0,19	0,18	-0,63	0,16	1,00	-0,01	-0,37	0,15	-0,27	-0,23
T7	-0,30	0,04	-0,22	-0,02	0,57	-0,01	1,00	-0,10	0,20	0,06	0,36
T8	0,76	0,10	0,13	0,25	0,39	-0,37	-0,10	1,00	-0,43	**0,81**	0,03
T9	-0,36	-0,14	-0,18	-0,16	-0,02	0,15	0,20	-0,43	1,00	-0,34	0,03
T10	0,64	0,01	0,11	0,17	0,57	-0,27	0,06	0,81	-0,34	1,00	0,09
T11	0,05	0,19	-0,11	0,13	0,20	-0,23	0,36	0,03	0,03	0,09	1,00

2.5. Predicted variables

We attempt to predict household inflation expectations on the horizon of 12 months ahead. In Russia, regular surveys of inflation expectations have been conducted since 2009 as part of a joint project between the Bank of Russia and the Public Opinion Foundation (FOM). The FOM surveys include a long list of questions concerning perceptions of prices, cost expectations, consumer preferences, etc. FOM publishes perceived inflation and household inflation expectations 12 months ahead. To quantify these variables, FOM uses the survey questions 'How do you think prices will generally change over the next 12 months (year)?' and 'How do you think prices have generally changed over the past 12 months (year)?'. To quantify expectations, FOM has implemented the probability method described by Carlson and Parkin (1975) and Berk (1999).

2.6. Data pre-processing

First, we exclude calendar factors in the topic time series by dividing all variables by the number of days in each month. Second, we apply Box-Cox transformation (Box and Cox, 1964) and clean the data of seasonality. Third, we apply Min-Max Scaling to give equal weight to the independent variables. All the topic time series are one month lagged (so we estimate inflation expectations and perceived inflation at time t using the news at time t-1).

For further analysis, we apply the augmented Dickey Fuller test to 11 explanatory variables and 2 predicted variables. Stationarity is rejected at 5% significance for all explanatory variables and for all predicted variables, except for Bank of Russia's communication which is stationary. To address the issue of non-stationarity and to avoid spurious dependencies between the variables, we transform all non-stationary variables into first differences. Further, for the Bayesian Structure Learning algorithm, we transform these first differences from numeric to discrete data using constant subinterval width.

For the machine learning tree-based algorithms, we use raw data as suggested by recent studies. Ahmed et al. (2010) explain that, for many time series, the absolute level is a useful piece of information. A number of papers (see the work of Kim (2003) for support vector machines, and the work of Kondratyev (2018) for neural networks) have shown that machine learning algorithms work better with non-stationary data. Finally, Petelin at al. (2022) have proved that common methods of transforming nonstationary data, such as differencing and log transformation, may not always be a good idea for machine learning models and that many of them show the best quality with raw data.

All data are bootstrapped for machine learning algorithms. Bootstrapping resamples the dataset to create many simulated samples. This idea was introduced by Efron (1979). The technique allows the bypassing of the restrictions of small datasets and the calculation of crucial statistics such as standard errors.

3. Methods

The challenge we try to solve is determining the features of news topics that make an important contribution to the creation of households' inflation expectations and perceived inflation. Valid methods to solve this challenge are regression algorithms with the option of defining feature importance or establishing cause-effect relationships.

3.1. Lasso

Least Absolute Selection and Shrinkage Operator (Lasso) was proposed by Tibshirani (1996) and is now one of the most popular regularisation techniques to select subsets of variables. The cost function can be written as:

$$\sum_{i=1}^{n}(y_i - \sum_j x_{ij}\ \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad \text{, where} \qquad (3)$$

n – is the number of instances

p – is the number of features

λ – controls the strength of the regularisation penalty

Trying to minimise the cost function, Lasso discards redundant features and selects useful features. Discarding a feature makes its coefficient equal to 0, while useful features retain positive coefficients. Lasso is a convenient starting point due to its transparency and explainability.

3.2. Random Forest

A Random Forest is an ensemble technique capable of performing regression tasks with the use of multiple decision trees. The algorithm was introduced by Breiman (2001). We use a bootstrapped Random Forest Regression with 50 trees, and a minimum number of samples required to split an internal node of 2. We use K-Fold Cross Validation to avoid overfitting with k = 10. For feature importance statistics, we use permutation feature importance. This method measures the increase in the prediction error of the model after permuting a feature's values. If the error increases, then the model relies on the feature for its predictions.

3.3. XGBoost и Kernel SHAP

Extreme Gradient Boosting (XGBoost) was developed by Chen and Guestrin (2016). It is a parallel decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. It works well with small and medium structured and tabular datasets. For our model, we use 30 boosting iterations, the maximum depth of a tree is 2, and the squared error is a loss function. For feature importance scores, we use Kernel SHAP (SHapley Additive exPlanation, see the work of Lundberg and Lee (2017)). The method uses a special weighted linear regression to compute the importance of each feature. The computed importance values are Shapley values from game theory and also coefficents from a local linear regression. The Shapley values can be estimated as follows:

$$\varphi_i(v) = \sum_{C\subseteq N-i} \frac{|C|!(n-|C|-1)!}{n!}[v(N\backslash C) - v(C)] \text{, where} \qquad (4)$$

φ – is the Shapley value

n – is the number of features

C – represents any subset of features that does not include the i – th feature

$|C|$ – is the size of this subset

$\frac{|C|!(n-|C|-1)!}{n!}$ can be interpreted as the probability that in any permutation, the members of C are ahead of distinguished player i.

3.4. Bayesian Structure Learning

Bayesian Structure Learning (BSL) stands apart from the other algorithms discussed above. It is a probabilistic graphical model which returns a directed acyclic graph (DAG) with features as nodes and their dependencies as edges. A lack of edges between nodes indicates the independence of the variables. Causal Bayesian networks are one approach to capturing causal relationships (see the work of Pearl (1988) as one of the first related papers). We apply several learning algorithms (PC algorithm for learning Gaussian Bayesian networks (Spirtes and Glymour, 1991), Grow-Shrink for learning Markov network structures (Margaritis D., 2003), and Incremental Association Markov Blanket (Tsamardinos et al., 2003)) to find the graph composition that maximises the data likelihood. We use the BIC with the smallest value to choose the algorithm. The software we use is the bnlearn package for R (for technical details, see work of Scutari (2009)).

4. Results

Table 2 presents the feature importance scores for all the selected models: Lasso, Random Forest, XGBoost, and the presence or absence of an edge with the inflation expectations or perceived inflation node in the Bayesian Structure Learning graph. As feature importance scores, we use the Lasso weights associated with each feature term, the permutation feature importance for the Random Forest, and the SHAP values for XGBoost. All feature statistics are scaled to sum to one for ease of comparison. Bold font highlights the top 3 features for each model. Despite different pre-processing techniques, almost all the models chose 3 topics as contributing the most to both household inflation expectations and perceived inflation: 'Inflation', 'Economic crisis', and 'Ruble devaluation'.

Adjusted R^2 is used as summary statistics for Lasso, Random Forest, and XGBoost. Mean absolute error (MAE) is used for all the methods (with predicted variables as testing nodes for BSL). Random Forest show higher R^2 for both inflation expectations and perceived inflation variables. The lowest MAE for inflation expectations was obtained using BSL, and for perceived inflation the lowest MAE was obtained using Random Forest. The findings show that both inflation expectations and perceived inflation might be shaped by three strong news streams: current dynamics of inflation, information about economic crises and ruble devaluation. The most significant news in the inflation expectations model was the ruble exchange rate and inflation, while in the observed inflation model it was inflation and the economic crisis.

Table 2. Feature importance scores for all models

Features	Inflation expectations 12m				Perceived inflation			
	Lasso	RF		Lasso	RF		Lasso	RF
Poverty	0,00	0,02	0,05	False	**0,24**	0,06	0,11	False
Inflation	**0,32**	**0,25**	**0,34**	False	**0,14**	**0,20**	**0,25**	**True**

Food crisis	0,00	0,02	0,02	False	0,03	0,02	0,01	False
Debt problems	0,00	0,02	0,01	False	0,00	0,04	0,05	False
Economic crisis	**0,29**	0,02	0,05	False	**0,18**	**0,18**	**0,16**	False
Geopolitics	0,00	0,02	0,00	False	0,13	0,08	0,06	False
Ruble devaluation	**0,39**	**0,42**	**0,26**	**True**	0,00	**0,27**	**0,22**	False
COVID	0,00	0,15	**0,25**	False	0,04	0,09	0,10	False
Domestic instability	0,00	0,02	0,01	False	0,00	0,03	0,02	**True**
Unemployment	0,00	0,05	0,01	False	0,13	0,03	0,01	False
BoR Communication	0,00	0,01	0,00	False	0,12	0,01	0,01	False
Adjusted R^2	0,23	0,74	0,83		0,43	0,81	0,67	
MAE	0,03	0,06	0,06	0,02	0,01	0,01	0,01	0,03

We present two graphs as result of the BSL model: Figure 13 for household inflation expectations and Figure 14 for perceived inflation. The BSL model has difficulties in determining the direction of some edges, but it groups news items according to their connections. We interpret this lack in the model's work as evidence of events and news happening at the same time (at least within the 1-month horizon). Thus, BSL can not determine causal relationships for these simultaneously occurring processes.

Looking at these results, we can assume that in answering the question about the next 12 months' inflation, households take into account news about economic conditions in general, and in Russia, they are historically closely related to geopolitics and the ruble exchange rate. In other words, people may not perceive the question about future inflation as a question about prices in the strictest sense, but as a broader question about economic outlook. It forms their 'premonition of inflation'. This is further illustrated by the lack of edge between 'Inflation' and 'Inflation expectation' nodes.

As for perceived inflation, in answering that question, people may reflect more on the current situation with prices, poverty, and domestic instability. These three variables probably reflect the best households' perception of current material living conditions.

The variable reflecting the intensity of the presence of the Bank of Russia in the news flow did not have a significant impact on the perception of inflation by the households in any of the models.

Figure 13. Causal Bayesian network structure with 12m expectations

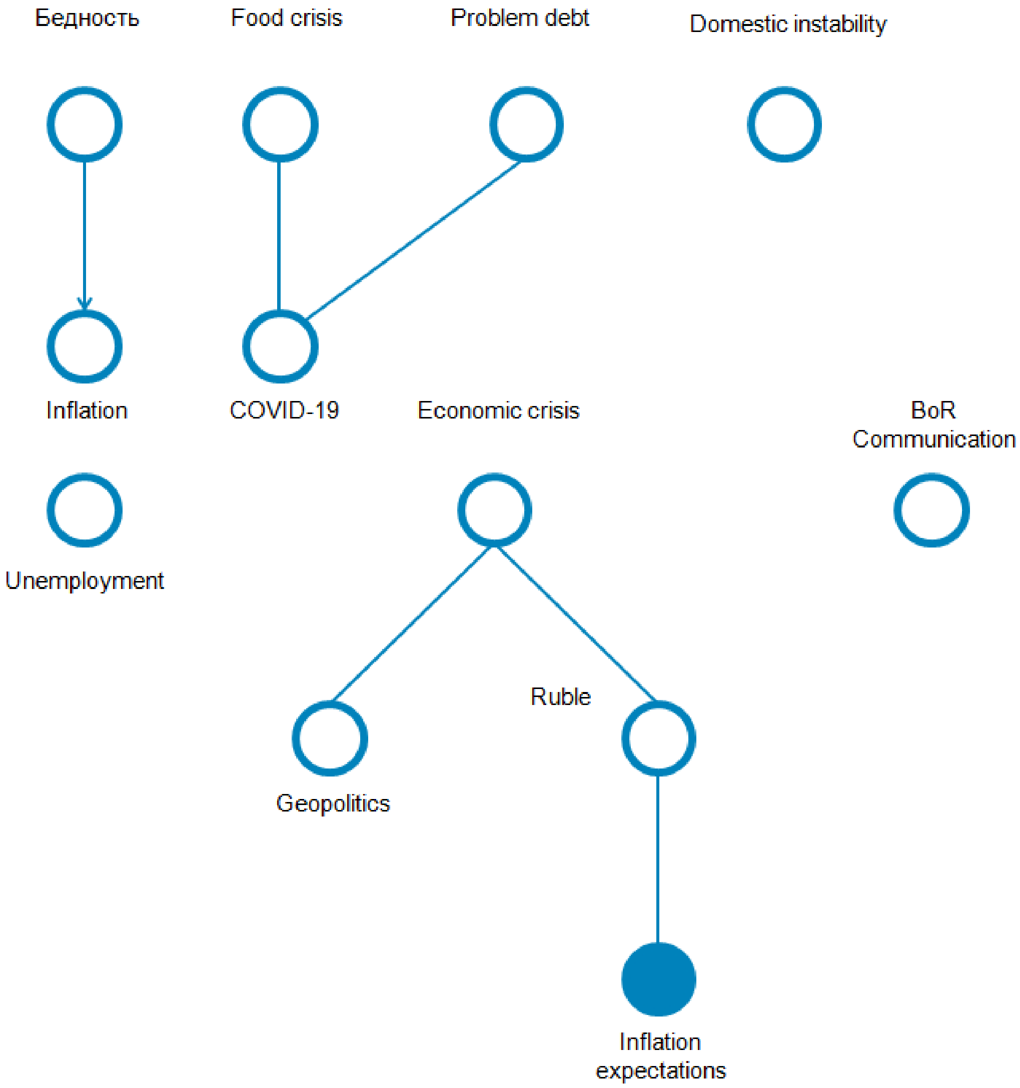
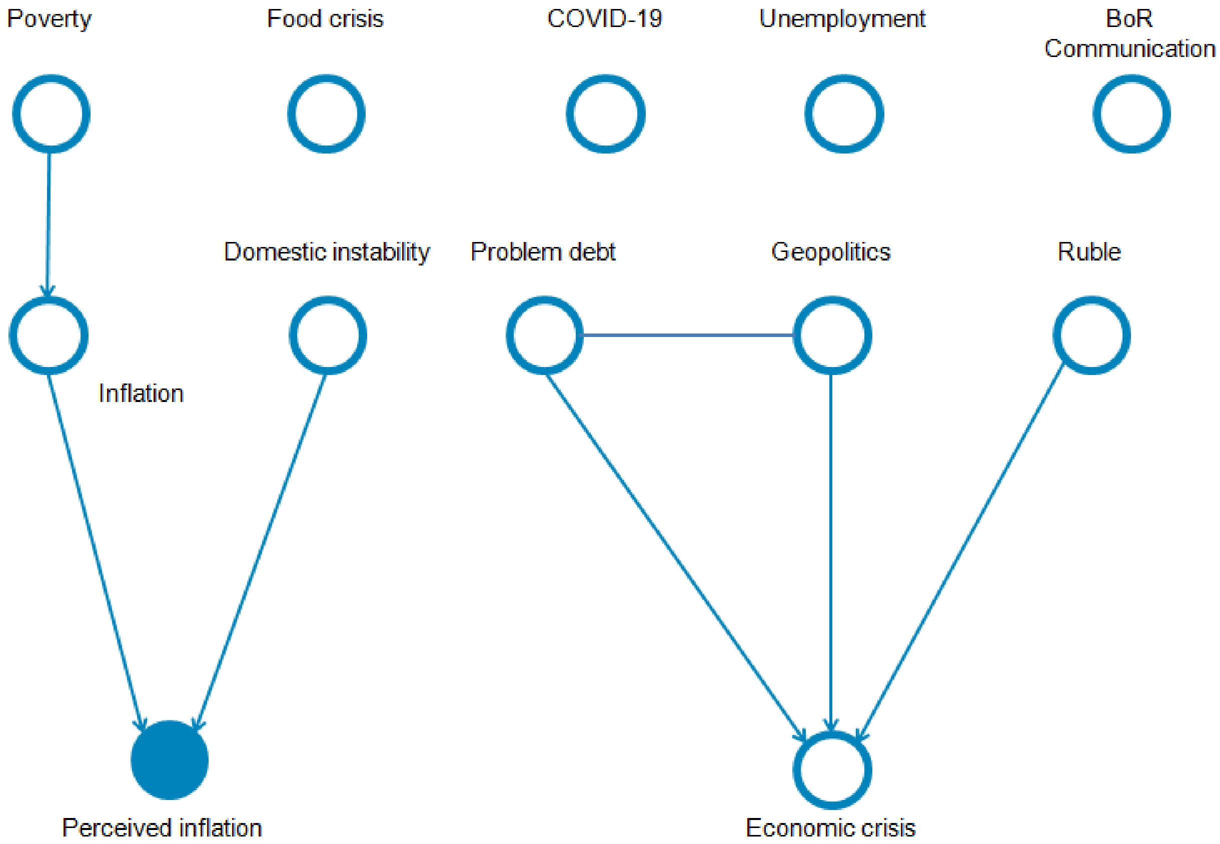


Figure 14. Causal Bayesian network structure with perceived inflation



5. Robustness

In this section, we conduct robustness analysis to check the validity of our findings. To test the robustness of our results, first we perform a set of alternative regressions, which are presented in Appendix 3. The data we use for these checks is the original data with 3 additional generated white noise variables as the independent variables (ARIMA (0, 0, 0)). For models with non-stationary data we have added a random trend. Through this experiment, we check if significant variables in our models remain significant. Random Forest and Lasso pass the robustness check: all of the top 3 most significant topics remain. As for XGBoost, the 'Covid' topic replaces the 'Economic crisis' topic for this model with perceived inflation in the test check. The alternative BSL model does not create edges between the dependent variable and the white noise variables.

Second, we perform a set of regressions with alternative predicted variables. We use data for two subgroups of respondents: those with savings and those without savings. In this way, we obtain four new dependent variables: inflation expectations from respondents with savings, perceived inflation from respondents with savings, inflation expectations from respondents without savings, and perceived inflation from respondents without savings. Almost all conclusions remain stable. There are minor changes for 1) the Lasso perceived inflation of households without savings (geopolitics has become a much more significant factor), 2) the Lasso perceived inflation of households with savings (the 'Inflation' topic fell out of the top 3 news topics, ‘Unemployment’ joined the top-3 topics), 3) the Lasso inflation expectations of households without savings (communication has become a significant factor, but with a minimal coefficient), 4) the XGBoost perceived inflation of households without savings ('Covid' was replaced with 'Poverty'). At the same time, the significance estimates for the subgroups with and without savings changed somewhat, which can be attributed to differences in the perception of price dynamics between those who have savings and those who do not.

6. Discussion

As a potential limitation of our study, we can note first of all the restrictions of the small datasets of inflation expectations in Russia. Regular surveys have been conducted only since 2014. We take those restrictions into account in the modelling, but longer time series and data from single respondents may provide more robust results.

For future work, we consider it expedient to compare news-driven factors with real world price data to distinguish the degree of influence of both factors. Second, it would be interesting to validate the negative news approach using data from developed countries, which may be especially promising during the acceleration of inflation in 2022.

Conclusions

In this study, we propose a new approach to the selection of the news that contributes the most to households' price perception based on the public's fears. We confirm the conclusions of previous studies about the huge impact that public information has on inflation expectations. In a continuation of this theme, we narrow the range of the news to strongly negative news. We define the key words for negative news topics from survey data about the public's fears and build topic time series using Guided LDA from 7,779 million news items collected from 28 Russian media agencies over a period of about nine years.

We obtain 10 negative news topics time series with the frequencies of the topics of Poverty, Inflation, Food crisis, Debt problems, Economic crisis, Geopolitics, Ruble devaluation, COVID, Domestic instability, and Unemployment. Additionally, we have added a variable with news, reflecting the intensity of the presence of the Bank of Russia in the news flow.

We conduct feature importance analysis using four algorithms: Lasso, Random Forest, XGBoost, and Bayesian Structure Learning. Our results indicate that three main news topics contribute the most to both household inflation expectations and perceived inflation: Inflation, Economic crisis, and Ruble devaluation. Bayesian Structure Learning presents a deeper picture with possible differences in the formation of households' inflation expectations and perceived inflation. In answering the survey about price changes in the next 12 months, consumers may actually be answering a question about their confidence in the economy in general, which in Russia is closely connected to the external situation (the 'Geopolitics' and 'Ruble devaluation' topics). On the other hand, in answering the question about inflation today and yesterday, consumers focus on actual prices and the current situation in the country.

The results are robust to a noise check and to alternative measures of inflation expectations and perceived inflation.

We report four main findings.

First, strongly negative news may represent a sufficient condition for the increase of both inflation expectations and perceived inflation.

Second, we find differences in 1) the perception of inflation by higher and lower income households and 2) the formation of expected and perceived inflation. Namely, the subgroup of respondents with savings tends to pay greater attention to the dynamics of the ruble exchange rate. The subgroup of respondents without savings pays more attention to news about geopolitics and poverty. As for differences between expected and perceived inflation, respondents may look at the question about the next 12 months' inflation as a question about economic outlook in general and at the question about the past 12 months' inflation as a question about the most memorable price changes.

Third, we provide evidence that households may forecast the economic situation in general (for Russia, the main factors in such analysis have historically been geopolitics and the ruble exchange rate) and, depending on this, give a forecast for inflation. Comparing the impact of news indices and marker products could be an important topic for future research.

Fourth, The intensity of the informational presence of the Bank of Russia in the media, apparently, does not have a significant impact on the perception of inflation by the households.

This study advances the existing research on news-driven inflation expectations and contributes to the literature in few ways. First, we show that strongly negative news about the most important economic subjects can determine the price perceptions of households in developing countries as Russia. Second, we develop a large news database of leading Russian media agencies trained on economic news. Third, we provide more evidence and shed more light on the concepts of rational inattentiveness and the epidemiological model of inflation expectations.

With these results, we contribute to a better understanding of the nature of household inflation expectations in Russia.

References

1. Ahmed, N.K., Atiya, A.F., Gayar, N.E., and El-Shishiny, H. 2010. 'An empirical comparison of machine learning models for time series forecasting'. Econometric reviews, 29(5-6), pp.594-621.

2. Angelico, C., Marcucci, J., Miccoli, M. and Quarta, F., 2022. Can we measure inflation expectations using Twitter?. Journal of Econometrics, 228(2), pp.259-277.

3. Arora, V., Gomis-Porqueras, P. and Shi, S., 2013. The divergence between core and headline inflation: Implications for consumers' inflation expectations. Journal of Macroeconomics, 38, pp.497-504.

4. Berk, J.M., 1999. Measuring inflation expectations: a survey data approach. Applied Economics, 31(11), pp.1467-1480.

5. Bird, S., Klein, E. & Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc."

6. Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

7. Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers." Proceedings of the fifth annual workshop on Computational learning theory. 1992.

8. Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

9. Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.

10. Carlson, J.A. and Parkin, M., 1975. Inflation expectations. Economica, 42(166), pp.123-138.

11. Carroll, C.D., 2001. The epidemiology of macroeconomic expectations.

12. Carroll, C.D., 2003. Macroeconomic expectations of households and professional forecasters. the Quarterly Journal of economics, 118(1), pp.269-298.

13. Cavallo, A., Cruces, G. and Perez-Truglia, R., 2014. Inflation expectations, learning and supermarket prices (No. w20576). National Bureau of Economic Research.

14. Coibion, O., Georgarakos, D., Gorodnichenko, Y. and Van Rooij, M., 2019. How does consumption respond to news about inflation? Field evidence from a randomized control trial (No. w26106). National Bureau of Economic Research.

15. Chen, T., & Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

16. Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, Arnaud Legout. Social Clicks: Whatand Who Gets Read on Twitter?. ACM SIGMETRICS / IFIP Performance 2016, Jun 2016, AntibesJuan-les-Pins, France. hal-01281190

17. Efron, B., 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM review, 21(4), pp.460-480.

18. Del Giovane, P. and Sabbatini, R., 2004. Lï¿ ½introduzione dellï¿ ½euro e la divergenza tra infl azione rilevata e percepita (No. 532). Bank of Italy, Economic Research and International Relations Area.

19. Jagarlamudi, J., Daumé III, H. and Udupa, R., 2012, April. Incorporating lexical priors into topic models. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 204-213).

20. Haldane, A., 2017. A little more conversation, a little less action. Bank of England-Speech.

21. Harris, Z. S., 1954. Distributional structure. Word, 10(2-3), 146-162.

22. Kim, K.J., 2003. Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), pp.307-319.

23. Kondratyev, A. (2018). Learning curve dynamics with artificial neural networks. Available at SSRN 3041232.

24. Larsen, V.H., Thorsrud, L.A. and Zhulanova, J., 2021. News-driven inflation expectations and information rigidities. Journal of Monetary Economics, 117, pp.507-520.

25. Lundberg, S. M., & Lee, S. I., 2017. A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

26. Lyziak, T., 2010. Measurement of perceived and expected inflation on the basis of consumer survey data (No. 5). Bank for International Settlements.

27. Maćkowiak, B., Matějka, F. and Wiederholt, M., 2021. Rational inattention: A review.

28. Mankiw, N.G., Reis, R. and Wolfers, J., 2003. Disagreement about inflation expectations. NBER macroeconomics annual, 18, pp.209-248.

29. Margaritis D., 2003. Learning Bayesian Network Model Structure from Data. Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA. Available as Technical Report CMU-CS-03–153.

30. Martin, F. and Johnson, M., 2015, December. More efficient topic modelling through a noun only approach. In Proceedings of the Australasian Language Technology Association Workshop 2015 (pp. 111-115).

31. Pearl, J., 1988. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann.

32. Petelin, G., Cenikj, G. and Eftimov, T., 2022. Towards understanding the importance of time-series features in automated algorithm performance prediction. Expert Systems with Applications, p.119023.

33. Pfajfar, D. and Santoro, E., 2013. News on inflation and the epidemiology of inflation expectations. Journal of Money, Credit and Banking, 45(6), pp.1045-1067.

34. Rehurek, R. & Sojka, P., 2011. Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2).

35. Sahu, S. and Chattopadhyay, S., 2020. Epidemiology of inflation expectations and internet search: an analysis for India. Journal of Economic Interaction and Coordination, 15(3), pp.649-671.

36. Salton, G., & Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5), 513-523.

37. Sedova A., \& Mitrophanova P., 2018. Topic Modelling of Russian Texts based on Lemmata and Lexical Constructions. Computational Linguistics and Computational Ontologies. 132-144. 10.17586/2541-9781-2017-1-132-144.

38. Scutari, M., 2009. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817.

39. Sims, C.A., 2003. Implications of rational inattention. Journal of monetary Economics, 50(3), pp.665-690.

40. Sussman, N. and Zohar, O., 2022. Have Inflation Expectations Become Un-anchored? The Role of Oil Prices and Global Aggregate Demand. International Journal of Central Banking, 18(2), pp.149-192.

41. Spirtes, P. and Glymour, C., 1991. An algorithm for fast recovery of sparse causal graphs. Social science computer review, 9(1), pp.62-72.

42. Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

43. Tilly, S. and Livan, G., 2021. Predicting market inflation expectations with news topics and sentiment. arXiv preprint arXiv:2107.07155.

44. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R. and Statnikov, E., 2003, May. Algorithms for large scale Markov blanket discovery. In FLAIRS conference (Vol. 2, pp. 376-380).

Appendices

Appendix 1. News database and topic modelling

Table 3. Composition of news database

Media agency name	Residence	Number of news items included
Kommersant	federal	334754
Arguments and Facts	federal	18957
E1.ru (Ekaterinburg News)	regional	117911
Vedomosti	federal	58536
Utro.ru	federal	182579
Moskovsky Komsomolets	federal	1784184
Forbes	federal	28090
Banki.ru	federal	183399
BBC Russia	federal	80841
Interfax	federal	86913
Lenta.ru	federal	593999
Finam.ru	federal	74907
VZ.ru	federal	492891
Echo.msk	federal	562774
Komsomolskaya Pravda	federal	414135
Parliamentary newspaper	federal	252769
Meduza.io	federal	96575
Regnum	federal	44204
Pravda.ru	federal	236773
Republic.ru	federal	45491
360tv.ru	federal	537391
Svpressa.ru	federal	205910
Finanz.ru	federal	272249
RG.ru	federal	586372
Svoboda.org	federal	95596
Finmarket.ru	federal	11368
Fontanka.ru	regional	363596
РБК	federal	16106

Table 4. Aggregation of keywords into topics

Sub-topics	Topic tag	Number if news items
Economy	0	298 343
Economy and finance	0	5 248
Business	0	83 082
Finance	0	19 297
Construction	0	6 263
Real estate	0	8 138
Finance and investment	0	2 906
Politics	1	400 108
International politics	1	10 015
Politics in Russia	1	6 059
In the world	1	211 104
World	1	161 736
News of the world	1	18 230
Society	2	901 518
News – society	2	11 931
State and society	2	8 682
Culture	3	115 380
Incidents	4	663 231
Sports	5	220 919

Table 5. Outcomes for supervised learning models

		2 classes	6 classes			
		XGBoost		**SVM**	**Logistic regression (softmax)**	**Random Forest**
Economic news	precision	0.9381	0.9363	0.9267	0.8658	0.8886
	recall	0.9064	0.9482	0.9548	0.9249	0.9240
	F-score	0.9219	0.9422	0.9405	0.8944	0.9060
Other news	precision	0.9598	0.9388	0.9344	0.8757	0.9006
	recall	0.9472	0.9390	0.9347	0.8765	0.9013
	F-score	0.9534	0.9389	0.9345	0.8758	0.9006

Appendix 2. Model's residuals and Dickey-Fuller tests

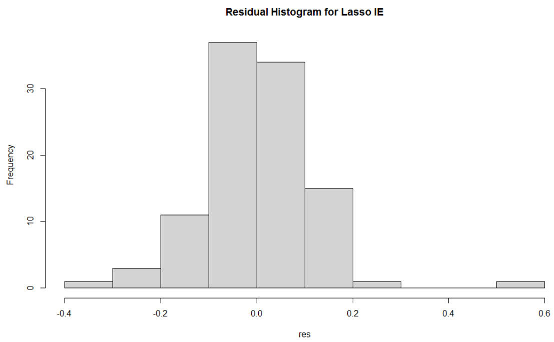
Table 6. Dickey-Fuller tests for time series in Lasso and BSL

Time series	Dickey-Fuller	p-value
Poverty	-6,1192	<0,01
Inflation	-3,5895	0,03751
Food crisis	-3,9019	0,01665
Debt problems	-4,8159	<0,01
Economic crisis	-4,0759	<0,01
Geopolitics	-4,2891	<0,01
Ruble devaluation	-4,0910	<0,01
COVID	-3,5200	0,04372
Domestic instability	-5,6691	<0,01
Unemployment	-3,9125	0,01614
inflation expectations	-5,0091	<0,01
perceived inflation	-4,2700	<0,01

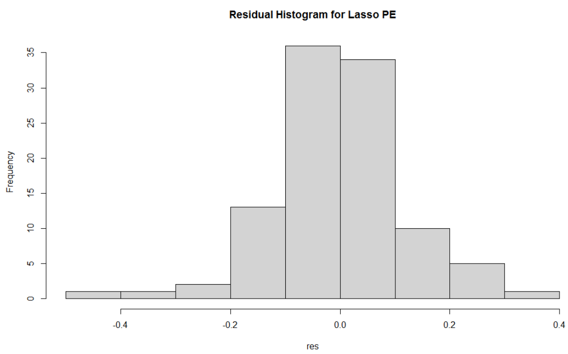
Figures 14-19. Residuals

1.) Lasso

a) inflation expectations

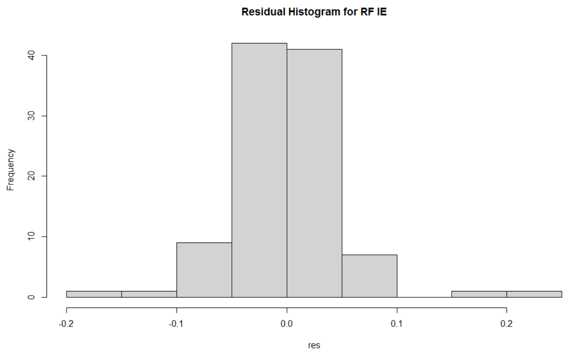


b) perceived inflation

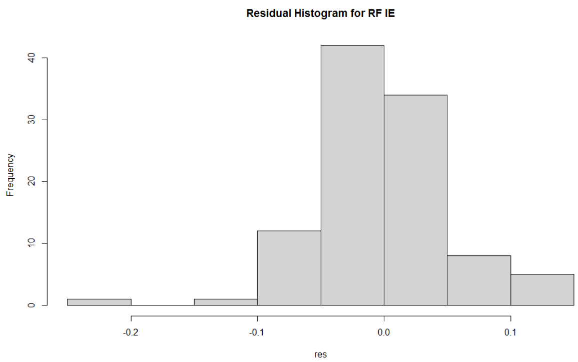


2.) Random Forest

a) inflation expectations

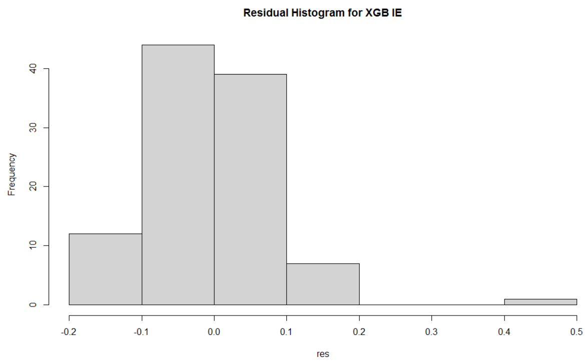


b) perceived inflation

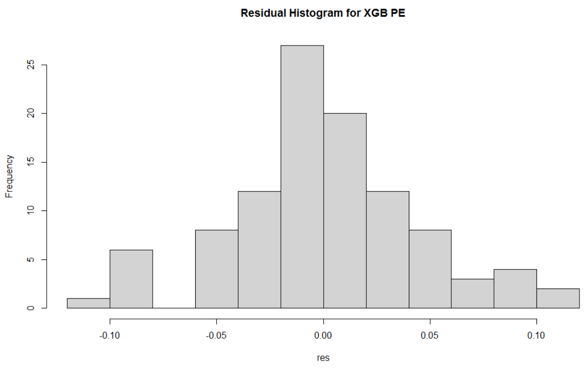


3.) XGBoost

a) inflation expectations



b) perceived inflation



Appendix 3. Robustness check

Table 7. Models with noised variables

Feature	inflation expectations				perceived inflation			
	Lasso	RF		Lasso	RF		Lasso	RF
Poverty	0,00	0,02	0,05	False	0,18	0,05	0,09	False
Inflation	0,11	0,24	0,34	False	0,20	0,21	0,30	True
Food crisis	0,00	0,01	0,02	False	0,01	0,01	0,00	False
Debt problems	0,00	0,02	0,01	False	0,00	0,03	0,02	False
Economic crisis	0,42	0,02	0,05	False	0,18	0,15	0,10	False
Geopolitics	0,00	0,03	0,00	False	0,13	0,07	0,07	False
Ruble devaluation	0,46	0,40	0,26	True	0,00	0,30	0,29	False
COVID	0,00	0,16	0,25	False	0,05	0,08	0,12	False
Domestic instability	0,00	0,02	0,01	False	0,00	0,02	0,01	True
Unemployment	0,00	0,05	0,01	False	0,14	0,02	0,00	False
BoR Communication	0,01	0,01	0,00	False	0,12	0,01	0,00	False
Noise 1	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Noise 2	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Noise 3	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Adjusted R^2	0,17	0,72	0,83		0,43	0,72	0,66	

Table 8. Models with alternative inflation expectations variable

Feature	inflation expectations of households with savings				inflation expectations of households without savings			
	Lasso	RF		Lasso	RF		Lasso	RF
Poverty	0,07	0,09	0,05	False	0,00	0,04	0,06	False
Inflation	0,13	0,15	0,25	False	0,27	0,34	0,38	True
Food crisis	0,00	0,02	0,02	False	0,00	0,01	0,00	False
Debt problems	0,00	0,03	0,03	False	0,00	0,01	0,01	False
Economic crisis	0,20	0,08	0,02	False	0,69	0,02	0,05	False
Geopolitics	0,00	0,02	0,00	False	0,00	0,01	0,00	False
Ruble devaluation	0,49	0,27	0,24	True	0,00	0,31	0,20	False
COVID	0,00	0,18	0,30	False	0,00	0,16	0,26	False
Domestic instability	0,00	0,04	0,04	False	0,00	0,01	0,02	True
Unemployment	0,00	0,10	0,03	False	0,00	0,07	0,03	False
BoR Communication	0,11	0,01	0,03	False	0,04	0,01	0,00	False
Adjusted R^2	0,25	0,56	0,72		0,19	0,70	0,83	

Table 9. Models with alternative perceived inflation variable

Новостной индекс	**perceived inflation of households with savings**				**perceived inflation of households without savings**			
	Lasso	**RF**	**XGBoost**	**BSL**	**Lasso**	**RF**	**XGBoost**	**BSL**
Feature	0,30	0,08	0,08	False	0,15	0,06	0,14	False
	0,12	0,13	0,24	True	0,20	0,21	0,33	True
Poverty	0,00	0,01	0,00	False	0,00	0,01	0,00	False
Inflation	0,00	0,08	0,02	False	0,00	0,01	0,03	False
Food crisis	0,20	0,15	0,10	False	0,23	0,23	0,14	False
Debt problems	0,12	0,04	0,05	False	0,15	0,15	0,04	False
Economic crisis	0,00	0,28	0,33	False	0,00	0,21	0,24	False
Geopolitics	0,00	0,10	0,15	False	0,05	0,07	0,07	False
Ruble devaluation	0,00	0,03	0,02	False	0,00	0,02	0,02	False
COVID	0,15	0,05	0,00	False	0,09	0,02	0,00	False
BoR Communication	0,12	0,06	0,00	False	0,13	0,01	0,00	False
Adjusted R^2	0,33	0,74	0,60		0,46	0,77	0,64	