THE GRADUATE INSTITUTE GENEVA

INSTITUT DE HAUTES ÉTUDES INTERNATIONALES ET DU DÉVELOPPEMENT
GRADUATE INSTITUTE OF INTERNATIONAL AND DEVELOPMENT STUDIES

Graduate Institute of International and Development Studies

International Economics Department

Working Paper Series

Working Paper No. HEIDWP13-2021

Role of the Media in the Inflation Expectation Formation Process

Tetiana Yukhymenko
National Bank of Ukraine

June 2021

Chemin Eugène-Rigot 2
P.O. Box 136
CH - 1211 Geneva 21
Switzerland

Role of the Media in the Inflation Expectation Formation Process

Tetiana YUKHYMENKO
National Bank of Ukraine

Abstract

This research highlights the role played by the media in the inflation expectations formation process of different types of respondents in Ukraine. Using a large news corpus and machine learning techniques I constructed news-based measures transforming text into quantitative indicators, which reflect news topics relevant to inflation expectations. As such, I found evidence that the different news topics have an impact on inflation expectations and can explain part of their variance. Thus, my results can help understand inflation expectations, especially as anchoring inflation expectations remains a key challenge for central banks.

Keywords: inflation expectations, natural language processing, textual data, machine learning.

JEL: C55, C82, D84, E31, E58

I would like to express my gratitude to Dr Rina Rosenblatt-Wisch from Swiss National Bank for valuable comments and guidance. I am grateful to SECO, and the Graduate Institute in Geneva for funding the coaching program under the Bilateral Assistance and Capacity Building for Central Banks (BCC).
The views expressed in the paper are those of the author and do not necessarily represent the views of the National Bank of Ukraine.

1.Theoretical framework, literature overview, motivation

Anchoring inflation expectations remains a key challenge for central banks, especially in developing countries. The process of inflation expectations formation matters for understanding macroeconomic dynamics and optimal policy design. For many years the rational expectations hypothesis has dominated the macroeconomic literature. However, this hypothesis is modified by accounting for information rigidities in a growing body of research. Thus, expectations could be rational but in a more realistic environment agents may be inattentive to relevant information due to the costs of acquiring and processing information. The two leading models of information rigidities are the sticky information model of Mankiw and Reis (2004) and the noisy information model developed by Woodford (2004) and Sims (2009), as well as Mackowiak and Wiederholt (2009).

Coibion and Gorodnichenko (2012) proved that information rigidities have a large impact on the macroeconomic variables, thus they should be integrated into modern macroeconomic policies to execute the optimal monetary policy. Also, they found that despite common wisdom, there is no significant difference in the degree of information consumption across agents – the speed of information processing of consumers is no lower than for other agents. Among other things this can be explained by noisy information model. Similarly, Coibion and Gorodnichenko (2015a) find that inflation expectations of professional forecasters from the US Survey can be modeled with imperfect information models due to the existence of information frictions. Coibion and Gorodnichenko (2015b) also research economic agents' expectations in Ukraine on the basis of survey data on inflation and exchange rate expectations. The survey also shows that there is a strong positive correlation between the evolution of Ukrainian economic agents' expectations of inflation and exchange rates. While some correlation might be expected from the pass-through of exchange rates into prices, a more likely rationale is the use of the exchange rate as a straightforward proxy by households of broader price movements within the economy, very much like households within the US do with gasoline prices.

It can be assumed that respondents are also influenced by uncertainties regarding tax, tariffs, spending, monetary and regulatory policy. These effects, however, are hard to detect because uncertainty is unobservable. One of the most popular approach to solve this problem is a news-based method. Typically, people obtain their views about the future path of the economy from the news media, directly or indirectly. So, news-based methods could be used to investigate the impact of the media environment on the formation of respondents' expectations.

Christopher D. Carroll (2003) tested an epidemiological model of expectations in which information diffuses over time from professional forecasters to households. In his work he describes that typical people obtain their views about the future path of the economy from the news media, directly or indirectly. Pfajfar and Santoro (2013) complement this epidemiological model with a measure of the actual perception of new information about prices. As a news measure they used a question from survey where participants have to indicate whether they have heard about positive or negative changes. Hearing news related to prices increases the probability of adjusting inflation expectations, while quality of forecasts is not likely to improve. Similarly, Coibion et al (2019) researched how central bank communications impact expectations. Thus, they compare answers of respondents after receiving eight

different forms of information regarding inflation. They concluded that these messages to the public influence expectations by economically significant magnitudes, however its effectiveness significantly decreases when channeled via news media. Mazumder (2021) proved that newspaper mentions of the Fed bring consumer and professional inflation forecasts closer, although this effect can vary depending on which newspaper it was published and how the topic was covered by the author.

Dräger and Lamla (2017) also found evidence on the role of media in inflation expectation formation. They analyzed rotating panel dimension of the microdata in the University of Michigan Survey of Consumers and found evidence that respondents are more likely to adjust their expectations if they have heard news on inflation.

Measuring the impact of news and constructing relevant indexes requires novel sources of information and processing methods, as well as significant computational resources. Consequently, researchers are replacing these indexes with alternative indicators which could be related to news measures. For example, Bauer (2015) used macroeconomic data surprises cumulated over the monthly or quarterly observation windows as an economic news measure. Thus, the data are macroeconomic indicators collected from traditional statistical sources, but their interpretation is somewhat different from the usual time series. Bauer investigated that several different survey measures of inflation expectations respond significantly to macroeconomic surprises. He also concluded that better anchoring of long-term inflation expectations could reduce sensitivity of inflation expectations to macroeconomic news, and variability of nominal rates as well. Garcia and Werner (2018) confirmed a significant impact of early inflation releases on long-term inflation expectations together with weakening of anchoring of inflation expectations in EU over recent years. Nautz et al (2017) also found that euro area inflation expectations anchoring undermined after fall 2011. They discovered that long-term inflation expectations respond significantly to macroeconomic news. As a news measure a set of macroeconomic variables was used, including CPI, PPI, unemployment, GDP, trade balance etc. D'Acunto et al. (2017) additionally found the relationship between frequency and size of price changes.

Larsen et al. (2020) used a more sophisticated approach. They applied machine learning algorithms to a large news corpus and examined the role of the media in the expectation formation process of households. It turned out that the news topics in the media are a good predictor of both inflation and inflation expectations. Also, they found that the degree of information rigidity among households varies across time, which can be explained by relevant media coverage. Goloshchapova and Andreev (2017) proposed to use mining information from text available online with the machine learning techniques to measure inflation expectations. They developed two indicators based on term frequency and sentiments of readers' comments on news related to inflation in major online economic media for the years 2014 - 2016. These indicators turned out to be close to household inflation expectations with a lead of one month. Angelico et al. (2021) used a similar approach to build real-time measures of consumers' inflation expectations from tweets. They combined unsupervised machine learning techniques with a dictionary-based approach to construct indices. Twitter-based indicators appear to be highly correlated with traditional measures of inflation expectations while having an advantage in time.

In this work I focus on the analysis of news and their impact on the formation of inflation expectations. To this end, I explore approaches to transform text into quantitative indicators which can then be further used for traditional econometric analysis. Our indicators should reflect news topics relevant to inflation expectations and properly assess their intensity. These measures also should be easily interpretable as they aimed to explain the impact of news on the formation of inflation expectations. All these challenges can be accomplished with text-mining techniques.

All measurement methods that are based on text-mining can be divided into two groups: 1) so-called naïve methods and 2) more complex methods, which are based on machine learning. Naïve methods are parsimonious, easy to use as they do not require much computational power, and are recognized worldwide due to their simplicity. They are based mostly on term frequency and document frequency. For example, Baker et al. (2016) investigated the relationship between economic policy uncertainty and investment rates, output, and employment growth. For these purposes the authors developed an index of economic policy uncertainty (EPU) based on a monthly count of articles that contain specific terms. Findings demonstrate that EPU is a reasonable proxy for different types of important macroeconomic variables and results are consistent with theories that highlight negative economic effects of uncertainty shocks.

However, these approaches could underestimate the actual level of uncertainty because they require qualitative expertise and human resources. For example, most naïve methods involve dictionary formation. This problem can be resolved with the more complex methods, which are based on machine learning techniques. Despite their relative complexity, according to the empirical results, machine learning approaches have larger predictive power than the naïve methods. The fastest and easiest way is to use unsupervised machine learning techniques.

One of the most popular unsupervised tool of natural language processing is the Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003). This generative statistical model allows dividing a collection of texts into subgroups, where each subgroup is characterized by keywords, associated with a topic. This method estimates the likelihood of the probability of words for a different number of topics. The results indicate the most likely number of topics. LDA is an unsupervised machine learning technique, which does not require training dataset. However, model results are unpredictable and need careful analysis. The methodology has been heavily applied in the machine learning literature and for textual analysis. Surprisingly, in economics, it has so far only a small number of successful applications, e.g., Larsen (2020) or Azqueta-Gavaldon (2017). Tobback et al. (2016) went another way and tried to improve the first EPU-index designed by Baker by applying supervised machine learning. Thus, they developed a classification model based on support vector machines (SVM) and labeled articles in two classes – related to economic policy uncertainty or not. Further, they constructed EPU SVM indicator based on this classification and include it in different macroeconomic models. This helps to improve accuracy of the different economic variable forecasts of these models in the short term.

However, unsupervised models as well as naïve methods also have disadvantages – absence of sentiment analysis. This could be fixed with machine learning and lexicon-based techniques foreseeing a predefined vocabulary and assessing the relative frequency of these words in the text. For example, Taboada et al.

presented Semantic Orientation CALculator (SO-CAL). This model uses dictionaries of words annotated with their semantic polarity and strength featuring intensification and negation. SO-CAL can be used in completely unseen data. VADER (Valence Aware Dictionary for sEntiment Reasoning) is another successful example of lexicon-based sentiment analysis tool. In order to develop it, Hutto and Gilbert (2014) constructed a list of lexical features and combined them with general rules that embody grammatical and syntactical conventions for expressing sentiment intensity. As a result, VADER outperformed many other highly regarded sentiment analysis tools. However, the main downturn of lexicon-based techniques is a lack of trained dictionaries in languages other than English.

A significant breakthrough in the sentiment analysis was an introduction of a new language representation model called BERT (Bidirectional Encoder Representations from Transformers), developed by Google researchers (Devlin et al., 2018). Like many other recent works in pre-training contextual representations BERT makes use of an attention mechanism that learns contextual relations between words (or sub-words) in a text. Unlike many other models it is designed to pre-train deep bidirectional representations from unlabeled text (treating on both left and right context). As result, BERT distinguishes even the same word as different ones taking into account the context for each occurrence of a given word. Pre-trained BERT model can be fine-tuned for a wide range of tasks, including classification. Pre-trained versions of BERT are available in a wide range of languages (including Ukrainian and Russian texts).

To narrow down the scope of this paper, I focus on the simpler naive methods and unsupervised machine learning and leave the sentiment analyses for future research. Starting from simplest naïve methods, I will continue with more complex machine learning methods of text classifications such as LDA. Thereafter, I develop an econometric model to assess the impact of the constructed indices on the formation of inflation expectations.

The paper is organized as follows. The next section presents data characteristics divided in two parts: text corpus of economic news and inflation expectations in Ukraine. Section 3 describes the construction and results of news-based indices and presents their statistiscal properties. Section 4 analyzes empirical specifications of the models and describes the results. Finally, Section 5 offers some concluding remarks and future steps. Additional information and results can be found in the Appendices.

2. Data characteristics

2.1. News corpus

The news corpus is webscraped from four Ukrainian online newspapers listed in the most popular online media in Ukraine. In particular, I used data from Ukrainian Pravda, Liga, Finance.ua and UNIAN. The general important requirements in selection were the availability of a fairly long archive (at least for the last ten years) and the possibility to parse data from the web. These newspapers have mainly economic orientation and are not subject to the explicit influence of individual political forces. However, some of the sources cover also non-economic topics.

The official language in Ukraine is Ukrainian, however, the Russian language is also very popular, so most national media publish materials in two languages. At the same time, the natural language processing infrastructure for the Russian language is developed slightly better and contains richer libraries and modules, which can improve the results of the research. That is why it was decided to process the news articles in Russian.

Our news corpus consists of more than 2 000 000 articles published online covering a sample of 20 years from January 2000 to December 2020. Our dataset contains the full text of the article and the available metadata, which include, for instance, the date, the link, and the title. Some sources also have a subtitle or general topic. All work with textual data, starting from the web-scraping and preprocessing and ending with the index construction was carried out using the Python programming language.

The textual data is presented in a non-traditional format which makes statistical inference challenging. Thus, it is really essential to preprocess corpus data into a readable machine format. Preprocessing includes some steps to clean and reduce the raw dataset before estimation.

First, it is essential to lowercase all of the characters to avoid any case-sensitive process. This should help to clean the dataset at least in two cases:

- words with the uppercase letter may not be detected as a stopword because all the stopword lists are lowercased. Stopwords are the words that have no significant contribution to the meaning of the text. For example, the most common stopwords are conjunctions and prepositions;
- the same word can be treated as different due to position in the sentence due to grammar. For example, first word of the sentence is always uppercased, but does not have to be a proper name. As a result, one word may be divided on two different values.

Second, but fundamental, step in NLP methods is tokenization. Tokenization is a method of breaking a piece of raw text into smaller units called tokens and converting them into a list. Tokens can be words, characters, or subwords. A token is a unit that NLP tools can easily convert to a value suitable for further machine processing.

Third, I removed non-essential information like stopwords from the text to simplify data processing. NLTK library in Python has rich corpora for stopwords in different languages, including Russian. Additionally, I removed non-ASCII characters, links, and punctuations by using Regular Expression (Regex).

The fourth step of text preprocessing is text normalization. The most common normalization techniques for Natural Language Processing are stemming and lemmatization. Stemming is a technique that chops off the ends of words. Due to this approach, the words having the same meaning but have some variations according to the context or sentence are normalized. Lemmatization usually refers to morphological analysis of words, normally aiming to return the base or dictionary form of a word (lemma). Russian, as well as Ukrainian, is a morphologically rich language, characterized by free word order and different word forms. Almost all language parts are marked for many characteristics as number, gender, case, tense, aspect, or person, and should be agreed grammatically with each other (Rozovskaya and Roth, 2019). Therefore, despite the longer processing time and the need for larger computing capacity,

lemmatization is more preferable than stemming for Russian text normalization. For this purpose, I used morphological analyzer pymorphy2, which returns the dictionary form of a word (Korobov, 2015).

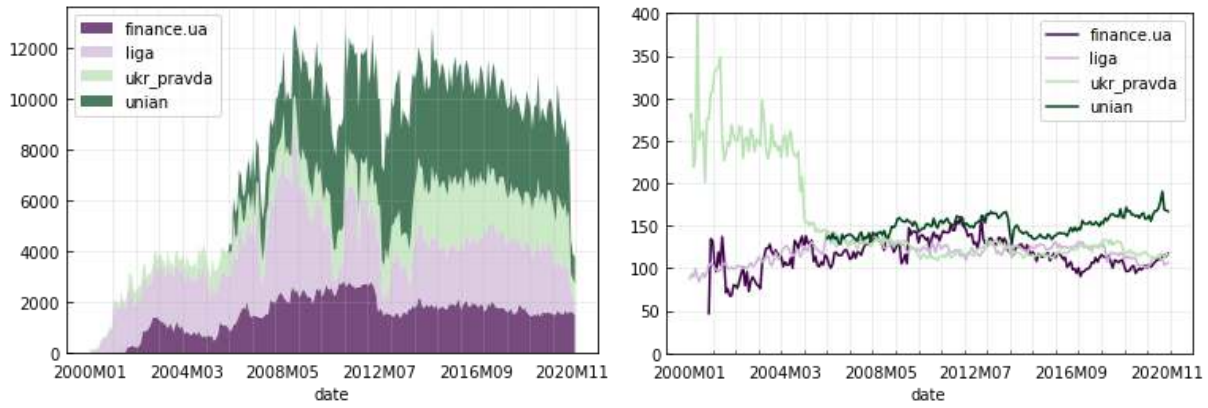


Figure 1. Number of news (a) and average size of articles (b) per month

Finally, I received a cleaned and normalized textual dataset consisting of around 300 million words and near 800 000 unique tokens. On average, as can be seen in Figure 1 article size did not change significantly during the observed period (100-140 words per article). However, the number of articles grew considerably from few articles per month in the early 2000s to 8-14 thousand per month since 2008 while the distribution of articles quantity between online sources has also changed (for more details about news corpus see Appendix A). Increasing number of articles in mid 2000's is related to rapid growth of internet penetration in Ukraine. Thus, according to State statistics service of Ukraine, number of active internet users in Ukraine exceeded 1 million people for the first time in 2007, gradually increasing from 200 000 in 2000. Since that time, number of active users skyrocketed to more than 23 million, and accordingly, internet penetration to 56% in 2019. This forced news media to move from traditional paper form to online versions. As a result, the media not only fully transferred their articles online, but also expanded the content of websites with additional materials that are not traditionally placed in newspapers.

2.2. Inflation expectations

The National Bank of Ukraine has been running surveys of inflation expectations for the next 12 months for several types of agents: households, banks, businesses, and professional forecasters. Prior to adopting inflation targeting regime by the NBU in 2015, Coibion and Gorodnichenko (2015b) widely reviewed these surveys and discussed their limitations. In our paper I will briefly describe the characteristics of inflation expectations of all groups of respondents (Figure 2).

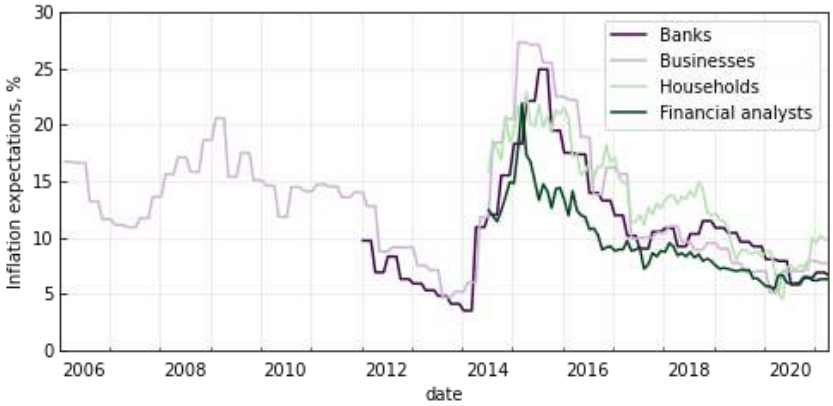


Figure 2. Inflation expectations for the next 12 months, %

Source: NBU, GfK Ukraine, Info Sapiens.

Banks. The survey of banks covers at least 90% of the banking system's assets excluding insolvent banks and banks in the process of liquidation. The NBU started to survey banks in 2012 and the data is available on a quarterly basis. Banks are surveyed during the the first weeks of the quarter.

Businesses. This survey includes answers from about 700 non-financial sector enterprises. Enterprises are selected by the quota principle corresponding to the structure of Ukraine's economy, which ensures the representativeness of the sample. Similar to banks businesses surveys conducted by the NBU on a quarterly basis since February 2006, however it happens during the second month of the quarter.

Financial analysts. The NBU commenced the survey of professional forecasters in July 2014 on a monthly basis (during second and third weeks of each month). Since November 2019 the frequency of this survey reduced to eight times a year and connected to the schedule of the Monetary Policy Committee meetings. Answers of financial analysts are collected one week prior to the meeting date. The number of professional forecasters varies over time – from 6 to 12.

Households. Simultaneously with the surveys of financial analysts, a survey of households was launched in July 2014. Unlike other surveys, the household survey is run monthly by the third-party company Info Sapiens (until 2019 it was run by GfK Ukraine). Every second and third weeks of the month approximately 1'000 consumers are surveyed about their inflation expectations and many other different social and economic issues. The sample is nationally representative and changes each month.

Banks, businesses, and households choose an interval of expected inflation for the next 12 months (more details in appendix B). The resulting estimate is the weighted average of middle points of those intervals. The answer "hard to answer" is also available for households, and these answers were excluded from calculation of average expectations. At the same time, financial analysts provide their point inflation forecasts (actual number, not interval estimate), and their expectations are the simple average of these estimates. The latter can lead to periodic bias, as the number of experts in the survey is not constant.

Table 1. Statistical properties of inflation expectations

	Banks		Businesses		Households	Financial analysts
	Full sample	Since July 2014	Full sample	Since July 2014	Full sample (Since July 2014)	Full sample (Since July 2014)
Count (quarters or months)	38	28	61	27	81	75
Mean, %	10.66	12.12	13.07	13.43	13.78	9.88
std, p.p.	4.99	4.88	5.39	6.92	4.68	3.43
min, p.p.	3.50	5.8	4.70	5.1	4.51	5.34
25%, p.p.	6.89	9.15	9.00	7.8	9.79	7.20
median, p.p.	9.92	10.65	12.76	10	13.55	8.80
75%, p.p.	12.00	14.33	15.80	18.65	17.14	12.18
max, p.p.	24.90	24.9	27.30	27.3	22.89	21.90
Skewness	1.074	1.045	0.731	0.811	0.108	1.082
Kurtosis	0.653	0.556	0.132	-0.699	-1.002	0.852

Table 1 provides brief statistical information about inflation expectations in Ukraine. Historically, expectations of professional forecasters are lower than any other respondents (mean is 2-4 pp lower than banks, businesses and household expectations). However, they do not provide much more accurate forecasts, as forecast error fluctuates in different directions (figure 3), similarly to other respondents. Thus, RMSE of expectations of financial analysts is 12.0 p.p., which is higher than households' and businesses' expectations RMSE (11.4 p.p. and 11.3 p.p. respectively for the same period since July 2014). Banks expectations show the worst results of forecasting power with RMSE of 13.1 p.p.

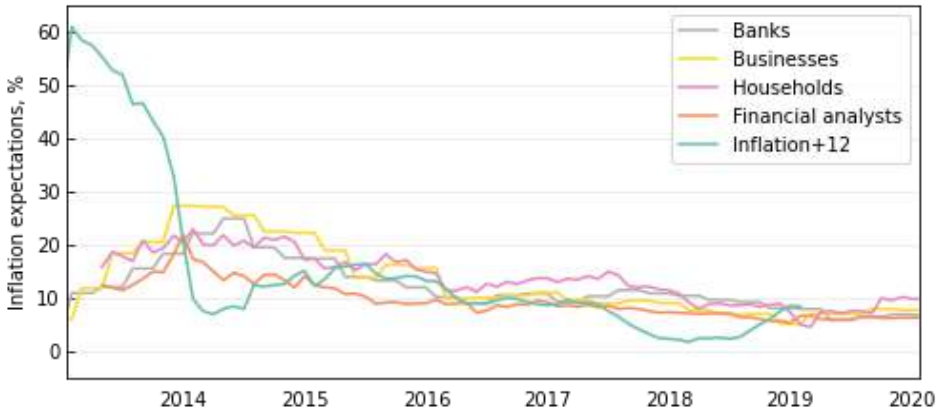


Figure 3. Inflation expectations for the next 12 months and actual inflation (+12 months), %

Source: State Statistic Service of Ukraine, NBU, GfK Ukraine, Info Sapiens.

All the expectations have a positive skew, which means that right tails are quite longer. Meanwhile, household expectations are almost symmetrical, having only small right-skewed tail. The distribution of household inflation expectations is flatter than normal, while all other expectations are more peaked.

3. Constructing aggregate news indexes

News content in our corpus is related mainly to economic, social and political topics. Thus, our sample includes news associated not only with inflation developments or expectations (prices, supply of certain goods, tariffs, statistical information, forecasts etc.). To focus only on the factors determining inflation expectations, I take out the news related to these topics. I apply two different approaches to filter out the

news. First, I use a dictionary-based approach to build a set of indexes based on the raw count of news. Second, I implemented a topic analysis using Latent Dirichlet Allocation (LDA) according to Blei et al. (2003).

Both approaches do not consider the sentiments of news content. However, it may not be a huge problem, because usually news is biased negative. Thus, Hester & Gibson (2003) found that economic news was written in negative tone more often than in positive. Additionally, they proved that negative news was a significant predictor of consumer expectations about the future economic developments. Damstra & Boukes (2018) well explain this negative bias of news with few main reasons:

- free media perform a crucial role in government control, so negative events receive more attention while positive ones do not meet such a need;
- in the process of judging the newsworthiness of real-world events negativity can be a key value, consequently "bad" news story is more likely to be selected by journalists;
- negative events have stronger impact than good ones.

Moreover, Soroka et al. (2019) defines negative tone of news as an essential feature, while good news, in contrast, may be considered with the absence of news. Therefore, to construct simple indexes, it is possible to assume that as a tendency news has a negative impact on perception and expectations.

3.1. Dictionary-based approach

Dictionary-based approach is the simplest approach to estimate the impact of news on various macroeconomic indicators. These indices are calculated as a share of articles related to the topic or more common denotation "document frequency". Intuition behind these indices is that the more alarming the topic – the more articles would be written on the subject, for example in times of crisis.

Document frequency (*df*) is the fraction of the documents that contain the word and obtained by dividing the number of documents containing the term by the total number of documents:

$$df(t, D) = \frac{d}{N} \ , \qquad (1)$$

where N is the total number of documents in the corpus D, and number of documents d where the term appears.

Dictionary-based approach to construct news indices requires good expertise aiming to select relevant keywords. In this case, to determine which prices worry Ukrainians the most, I turned to the consumer basket of the average household. Ukrainians spend the most on food. In different periods, the share of spending on food and soft drinks was 40-60% for the period from 2000 to 2020, slightly decreasing in recent years. Accordingly, it is important to select news that contains mentions of basic foods: bread, meat, dairy, vegetables, fruits etc.

Another essential component of household expenditure is utilities. Although the share of this type of spending is much lower than in many other countries, the utility tariffs is critical for Ukrainians and is

often speculated by politicians, and therefore may have a visible impact on expectations. The most important utilities for Ukrainians are electricity and natural gas.

Fuel prices may also have a significant impact on the formation of inflation expectations of households, even though not all people use private transport. For example, Kilian and Zhou (2020) found several episodes since 1990 in the US when household inflation expectations growth could almost entirely be explained by a hike in fuel prices. On the one hand, this is due to the widespread of gas stations and price boards, which allows them to be used for daily price monitoring. On the other hand, everyone is well aware that fuel is a component of the cost of most goods and services, explicitly or implicitly. In this case, I include news not only about fuel but also about oil as a defining cost component.

As was stated in Coibion & Gorodnichenko (2015b) there is a strong positive correlation between the inflation expectations and exchange rate developments, especially for households. In this case, logically, that not only the exchange rate dynamics affects expectations, but also the coverage of this topic in the media.

In addition, I will analyze the index of news related to the word inflation itself, as such news often contains expert forecasts or analysis of the current situation. According to Zholud et al. (2019), inflation expectations in Ukraine are highly linked to current inflation trend, therewith have a future-oriented component. Therefore, it is advisable to check the impact of references to inflation in the media on the formation of expectations.

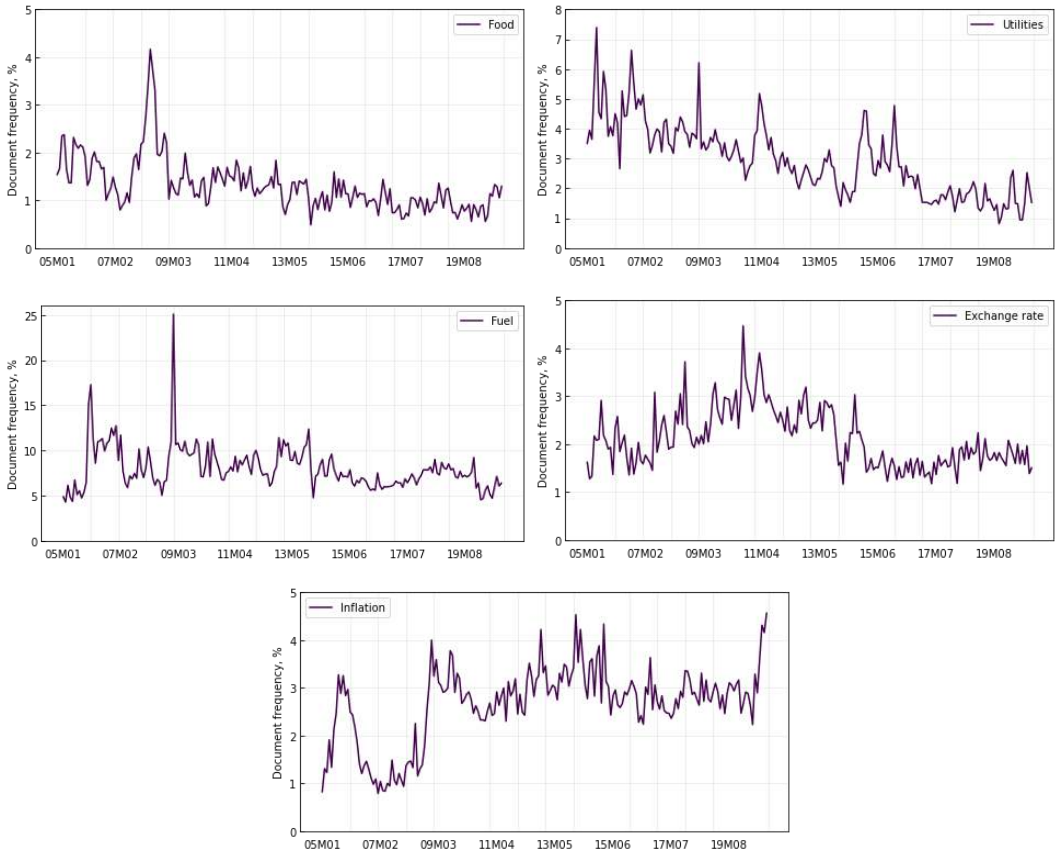


Figure 4. Document frequency of topics relevant to inflation expectations

Figure 4 shows the resulting indices calculated on a dictionary-based approach. Since the inflation expectations of the respondents are available at the earliest since 2006, all the time series of news indices will also be reduced from 2005 (one year backward to assess the lag effects). However, given that the amount of news has been much smaller in the early 2000s, I removed only about 10% of the articles from the corpus and there are about 1 800 000 articles left. News related to food have the smallest share among selected topics, while news about fuel looks the most important. Also, it can be seen that document frequency of news related to utilities in general decreased over time, except in 2015 when there was a significant jump in the importance of this topic as a result of bringing the utility tariffs to market levels. Interesting, that until 2014 topic related to exchange rate movements was mentioned more often in the news, which is probably due to the greater negative consequences of the sharp devaluation observed in Ukrainian history amid fixed exchange rate regime. For more information about indices built with dictionary-based approach see appendix C.1 and C.2.

As inflation expectations of different respondents are collected at different periods and not evenly throughout the month, the impact of some short-lived or even discrete news may be extremely important. Thus, some news may last only for several days and due to the rapid loss of interest in the topic the effect on monthly document frequency can fade away. Therefore, the monthly indices may not reflect real dynamics of the importance of individual events, and applying indices with higher frequency may shed light on this issue. To assess such impact, I additionaly computed similar indices in decade terms of each month – a decade being a third of the month (results are in Appendix C.3).

It is important to assume independence of researched variables, which can be indicated by its stationarity. Stationarity is necessary for the application of many statistical tools and procedures in time series analysis. Indeed, if the data was generated by a stationary process, so it has the properties of a sample generated by such a process. According to the Dickey-Fuller test, monthly time series for share of utilities, exchange rate and inflation news are non-stationary. Simultaneously, with respect to indices in decade terms only time series for share of utilities and inflation news are non-stationary. Also, the autocorrelation is high, and it seems that there is no clear seasonality. Therefore, to get rid of the high autocorrelation and to make all the process stationary in the same way, I take first differences.

3.2. Unsupervised ML approach

One of the important shortcomings of the dictionary-based approach is the availability of quality expertise and selection of texts based on it. In particular, the article may contain keywords, but its topic is devoted to a completely different issue. For example, word “fuel” can be attributed to the topics related to the science and technology or car manufacturing. The solution here is unsupervised topic modelling algorithms. These statistical methods analyze the words of the collection of texts and divide it into subgroups, where each subgroup is associated with a set of keywords. Thus, the model finds combinations of words, but not a single one. In our “fuel” example articles with word combinations “fuel price” and “rocket fuel” would be distinguished. Most machine learning models require the use of a part of a data set in which specially trained people classify information according to a predetermined procedure, and therefore put labels on data. However, there are methods that do not need such labeled

training samples. Latent Dirichlet Allocation (LDA) presented by Blei et al. in 2003 nowadays is a very common example of topic modelling method which uses unsupervised learning algorithm.

I used an extremely efficient implementation of LDA called LightLDA provided in nimbusml Python module (Jinhui Yuan et al., 2014). This state-of-the-art implementation incorporates a number of optimization techniques and can train a topic model on large document sets much faster. For example, our model produces 100 topics on 2 million news dataset in less than within an hour, while LDA at this scale takes days. Figure 5 shows the distribution of topics received with LDA. The popularity of some topics changed over time, while the topics of others remained relevant throughout the observation period.

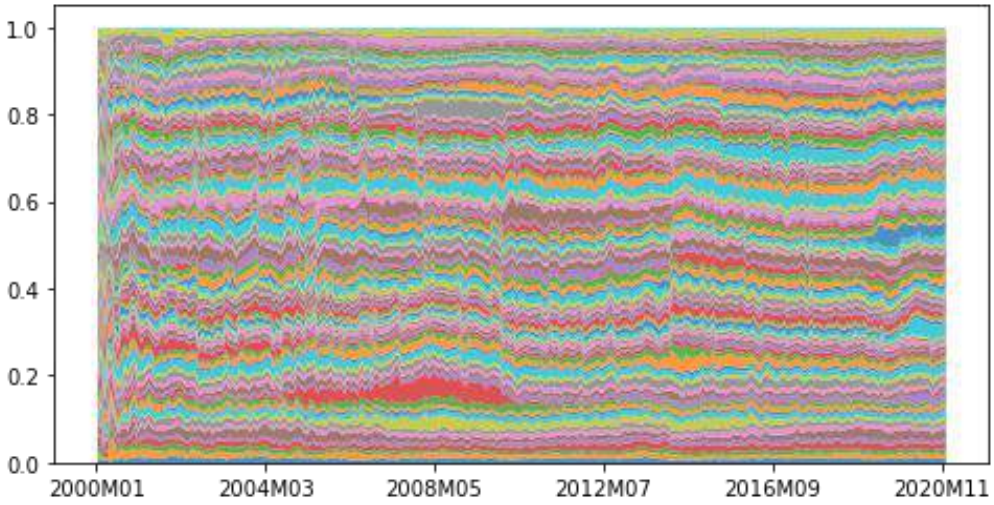


Figure 5. Distribution of topics received with LDA

The number of topics in LDA is not fixed and can be set regarding the task. I experimented with using different number of topics. I observed that with larger number of topics our main results do not change – some topics have very similar content and have to be joined in further analysis. At the same time, interpretation of larger number of topics becomes more complicated. With lower number of topics, it is sometimes difficult to distinguish between different topics which have similar keywords. For example, topics related to exchange rate may include not necessary information as some articles contain similar words, but in a different content.

At this point, human intervention is necessary to analyze and label the topics of the received news clusters. Figure 6 is a graph showing the relationship between the topics distributed by the LDA. Most of the news clusters are expectedly attributed to the politics, international relations, parliament and government. At the same time, some topics are clearly different and can be referred to the economic topics that may affect inflation expectations. Most news topics do not belong to one but to several clusters.

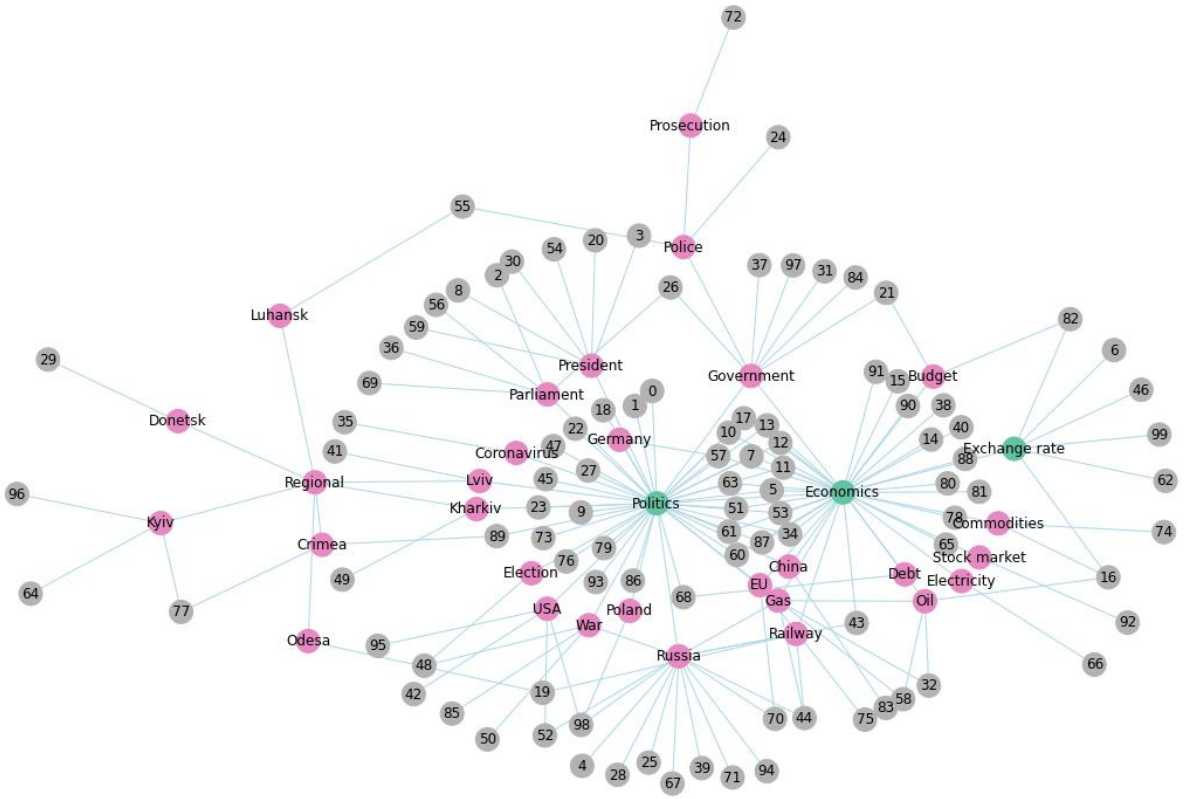
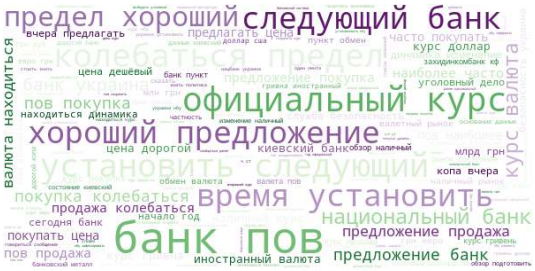


Figure 6. Graph of topics in news corpus. Grey circles refer to topics defined by LDA, pink circles - manually labeled clusters, and green circles – general topics

I managed to identify a news cluster related to exchange rate movements, which includes 6 news topics defined with the help of LDA. As we already know the situation in the foreign exchange market has some influence on the formation of inflation expectations in Ukraine. I also found a cluster related to commodities, including oil and gas. Additionally, topics describing electricity market, budget and government debt can be easily identified. Interestingly, the LDA helped identify a topic related to the spread of coronavirus, where the number of articles unsurprisingly increased since the end of 2019. In addition, the LDA has well defined the period of the conflict between Russia and the Ukraine in 2014 and subsequent years. Instead, the LDA did not group articles related to food prices, utility tariffs, etc., in separate recognizable topics, which can be explained by a similar structure to other articles, as well as the relatively low share of such news. However, increasing number of topics does not fix this. In Figure 7 I provide wordclouds for few of the most relevant topics.

Exchange rate movements

Topic 46



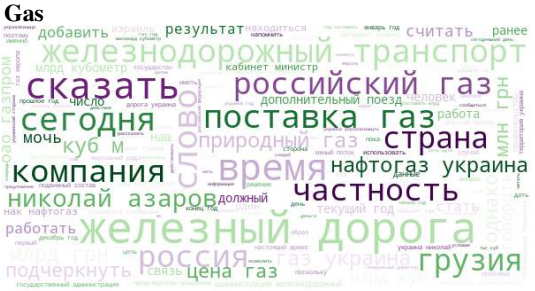
Topic 62

Topic 82

Topic 99

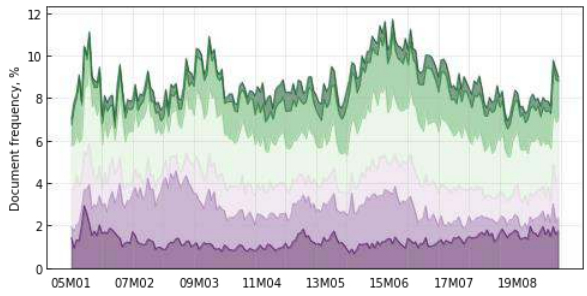
Oil and Gas

Topic 44

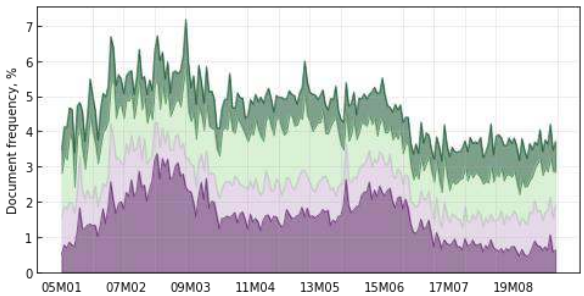


Topic 58

Figure 7. Wordclouds for selected topics



Related to exchange rate movements



Related to oil and gas

Figure 8. Share of topics, identified by LDA

The popularity of certain topics highly corresponds to the historical development of events. In particular, the share of articles on the hryvnia exchange rate, seen in figure 8, increased in 2008, when the hryvnia depreciated rapidly amid the global financial crisis. The next peak was observed in 2014-2015, when due to conflict between Russia and Ukraine and the loss of part of the territory, the economy suffered a significant blow. At this time, the hryvnia also depreciated rapidly. With the transition to a floating exchange rate and stabilization in the foreign exchange market, interest to this topic in news began to wane.

News about gas and oil behaved in the same way. Thus, in 2006-2008, the gas issue was extremely important for Ukraine against the background of complicated relations with Russia. Problems with gas supply were repeated in 2014. Instead, lower energy prices have contributed to less coverage of these topics in the coming years.

I built the indexes on the same principle as in the dictionary-based approach, using equation 1 of document frequency. Thus, monthly indices were calculated to assess long-term impact, while decadal indices were calculated to assess short-term media shocks, which may be important during time of inflation expectation survey and fade throughout the month. More details about statistical characteristics of indices constructed on LDA division are in Appendix D. According to the Dickey-Fuller test, monthly time series for share of energy news are non-stationary, while share of news on exchange rate movements are stationary. Decade time series can be considered as stationary with 95% probability.

4. Estimation results

As noted in previous sections, inflation expectations are largely shaped by past inflation (Zholud, 2018, and Gorodnichenko, 2015). Therefore, for the analysis I used an extrapolative approach to the formation of inflation expectations (Lines, 2010):

$$E\pi_t = \alpha + \beta\pi_{t-1} + \gamma(\pi_{t-1} - \pi_{t-2}) + \varepsilon, \qquad (2)$$

where $E\pi_t$ is expected inflation in period *t*, π_{t-1} denotes inflation in previous period, and $\pi_{t-1} - \pi_{t-2}$ stands for change in inflation, α and γ – coefficients of regression, while ε is an error. I use annual CPI change as a measure of inflation.

In this research I assume that instead of the changes in inflation the formation of inflation expectations (equation 2) is influenced by the media environment:

$$E\pi_t = \alpha + \beta\pi_{t-1} + \delta df_T^m + \varepsilon, \qquad (3)$$

where *df* denotes document frequency of the news topic *m* in period *T*. T may be equal t in case of testing the impact of news on formation of inflation expectations in the same month when survey is conducted. However, some surveys are conducted at the beginning of the month, therefore, I will test the impact of the frequency of news publications in previous three months on the formation of inflation expectations. Accordingly, T can be equal to t-1, t-2 and t-3. I am going to test monthly and decade frequency of *T* as the inflation expectations survey are not conducted for a whole month, but for shorter periods of time. In addition, for different respondents these periods also vary. As quarterly surveys are not conducted throughout the quarter, I used matching months instead of aggregating news indices at the quarterly level. For example, bank surveys take place in the first month of the quarter, thus the same month for the news index was used as base month.

I will also test another variation of extrapolative inflation expectations, assuming that change of respondents' expectations changes in response to changes in current inflation. In this case formula of inflation expectations looks like:

$$E\pi_t - E\pi_{t-1} = \alpha + \gamma(\pi_{t-1} - \pi_{t-2}) + \varepsilon. \quad (4)$$

I expanded formula 4 with changes in media environments in changes in current inflation:

$$E\pi_t - E\pi_{t-1} = \alpha + \gamma(\pi_{t-1} - \pi_{t-2}) + \eta(df_T^m - df_{T-1}^m) + \varepsilon. \quad (5)$$

In this case, all our variables are stationary and we can be sure that their properties do not change over time.

We also suppose that impact of constructed news indices on inflation expectations is linear, so to estimate this effect we use the OLS regression.

I start with analyzing the impact of news on the formation of inflation expectations using the dictionary-based approach. Table 2 presents the coefficients and p-values (in brackets) of news indices built with dictionary-based approach in OLS regressions of inflation expectations of different groups of respondents. This table shows two different approaches: without transformations (equation 3), using all variables as they were computed, and the 1st difference of all variables, presenting extrapolation of change of inflation expectations (equation 5). R^2 for the first type of relationship is expectedly much higher than for estimates of the first difference. However, the relatively low R^2 for such studies is quite normal and typical for studies of human behavior (King, 1986).

As can be seen from the Table 2, all types of inflation expectations are dependent on current inflation trends as coefficients are statistically significant. At the same, only banks and businesses tight changes of inflation expectations to recent changes in inflation, while relationship between changes of household and financial analyst expectations with recent inflation dynamics is insignificant. This is in line with the opinion that well-anchored long-term inflation expectations should not change in response to news about macroeconomic indicators, in particular inflation (Galati et al., 2011). However, it is too early to talk about anchoring inflation expectations, given the difference between the central bank's inflation target and inflation expectations. Therefore, in this case, to the result could be due to information rigidity.

Banks' inflation expectations are virtually independent of the current media environment on inflation. Most indicators are not statistically significant or contradict economic logic. For example, banks' inflation expectations are negatively correlated with food news with 90% probability. That is, the more thrilling this topic for society, the faster it reduces the inflation expectations of banks. This might be explained by the tone and content of the news. However, without a more detailed study of the content of this news, it is impossible to determine.

Table 2. Relationship between monthly news indices and inflation expectations

Respondents	Variables	Without transformations					Variables	1st difference				
		Inflation	Exchange rate	Utilities	Food	Fuel		Inflation	Exchange rate	Utilities	Food	Fuel
Banks	π_{t-1}	0.2949 ***(0.000)***	0.289 ***(0.000)***	0.2994 ***(0.000)***	0.297 ***(0.000)***	0.295 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.1047 **(0.022)**	0.1545 ***(0.005)***	0.1642 ***(0.003)***	0.1658 ***(0.003)***	0.1737 ***(0.001)***
	df^m_t	-1.2029 (0.194)	0.9028 (0.510)	-0.7928 (0.356)	-4.0589 *(0.099)*	0.0736 (0.824)	$df^m_t - df^m_{t-1}$	0.6139 (0.379)	1.5717 (0.189)	0.096 (0.903)	1.6328 (0.460)	0.3832 (0.220)
	df^m_{t-1}	1.1547 (0.360)	-0.7359 (0.662)	2.2787 *(0.099)*	2.9647 (0.104)	-0.3926 (0.346)	$df^m_{t-1} - df^m_{t-2}$	1.5263 (0.123)	-2.2497 (0.153)	1.2997 (0.306)	0.8267 (0.629)	-0.1084 (0.772)
	df^m_{t-2}	1.336 (0.153)	-0.7157 (0.591)	-1.761 (0.143)	-0.4362 (0.840)	-0.7258 (0.113)	$df^m_{t-2} - df^m_{t-3}$	1.614 **(0.031)**	0.5906 (0.624)	-2.2488 **(0.037)**	-3.6356 *(0.065)*	-0.972 **(0.026)**
	df^m_{t-3}	-1.074 (0.264)	-0.6057 (0.675)	0.7886 (0.345)	-1.4298 (0.534)	0.7137 *(0.100)*	$df^m_{t-3} - df^m_{t-4}$	-1.037 (0.158)	-0.5096 (0.689)	0.9001 (0.246)	2.9091 (0.167)	0.8387 **(0.033)**
	C	6.4354 *(0.093)*	9.5205 ***(0.000)***	5.7224 ***(0.000)***	10.1442 ***(0.001)***	9.7139 ***(0.000)***	C	-8.3529 ***(0.006)***	0.9782 (0.546)	-0.4893 (0.664)	-2.2835 (0.388)	-1.1227 (0.573)
	R^2	0.843	0.825	0.84	0.837	0.845	R^2	0.57	0.334	0.372	0.34	0.413
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	***(0.000)***	**(0.030)**	**(0.015)**	**(0.027)**	***(0.006)***
Businesses	π_{t-1}	0.3452 ***(0.000)***	0.3698 ***(0.000)***	0.3173 ***(0.000)***	7.0696 ***(0.000)***	0.3552 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.1434 ***(0.003)***	0.1344 ***(0.004)***	0.1467 ***(0.002)***	0.1326 ***(0.003)***	0.14 ***(0.003)***
	df^m_t	-0.1827 (0.843)	1.2093 (0.245)	-0.2132 (0.768)	0.3255 ***(0.000)***	0.2268 (0.451)	$df^m_t - df^m_{t-1}$	1.0633 (0.122)	0.9183 (0.232)	-0.8341 (0.153)	0.9607 (0.277)	0.0557 (0.799)
	df^m_{t-1}	-2.0502 *(0.073)*	-0.5729 (0.672)	1.5272 **(0.030)**	3.3463 ***(0.005)***	0.2061 (0.293)	$df^m_{t-1} - df^m_{t-2}$	0.3751 (0.645)	0.0937 (0.926)	1.1851 **(0.037)**	-0.5685 (0.659)	0.0689 (0.631)
	df^m_{t-2}	3.107 **(0.026)**	0.0296 (0.981)	-0.327 (0.720)	-5.3502 ***(0.002)***	0.0973 (0.767)	$df^m_{t-2} - df^m_{t-3}$	-0.3006 (0.768)	-0.3382 (0.720)	-0.2001 (0.787)	2.7441 *(0.054)*	0.0326 (0.910)
	df^m_{t-3}	-0.6576 (0.541)	0.9159 (0.384)	0.375 (0.584)	3.692 **(0.050)**	-0.1574 (0.582)	$df^m_{t-3} - df^m_{t-4}$	-0.6195 (0.442)	-0.4283 (0.576)	-0.2405 (0.664)	-2.7642 **(0.012)**	-0.3174 (0.203)
	C	8.1511 ***(0.000)***	4.9233 ***(0.010)***	5.0653 ***(0.000)***	-0.137 (0.924)	5.5163 **(0.020)**	C	-1.5599 (0.178)	-0.6747 (0.578)	-0.0132 (0.988)	-0.7796 (0.349)	1.1141 (0.489)
	R^2	0.696	0.685	0.746	0.736	0.687	R^2	0.227	0.179	0.219	0.274	0.188
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	**(0.016)**	*(0.058)*	**(0.020)**	***(0.004)***	**(0.045)**
Households	π_{t-1}	0.2314 ***(0.000)***	0.2412 ***(0.000)***	0.1257 ***(0.000)***	0.1901 ***(0.000)***	0.2328 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.0403 (0.417)	0.0368 (0.440)	0.0538 (0.795)	0.0486 (0.315)	0.0363 (0.456)
	df^m_t	0.3198 (0.708)	0.7091 (0.609)	1.2605 **(0.034)**	2.2806 (0.146)	0.2741 (0.536)	$df^m_t - df^m_{t-1}$	0.6496 (0.106)	0.9103 (0.155)	-0.3622 (0.268)	0.923 (0.239)	0.229 (0.283)
	df^m_{t-1}	-0.109 (0.905)	0.4006 (0.773)	0.5428 (0.469)	1.8611 (0.253)	0.1923 (0.691)	$df^m_{t-1} - df^m_{t-2}$	-0.592 (0.169)	-0.5324 (0.406)	0.6644 (0.125)	-0.3849 (0.632)	0.0145 (0.951)
	df^m_{t-2}	0.9379 (0.320)	-0.4846 (0.725)	-0.1165 (0.876)	2.8517 *(0.082)*	0.3577 (0.463)	$df^m_{t-2} - df^m_{t-3}$	0.6681 (0.136)	-0.7833 (0.223)	-0.6464 (0.135)	1.3106 (0.106)	0.1944 (0.401)
	df^m_{t-3}	1.1311 (0.209)	1.0076 (0.461)	1.5535 **(0.011)**	1.7897 (0.280)	-0.0602 (0.890)	$df^m_{t-3} - df^m_{t-4}$	-0.0787 (0.861)	1.2966 **(0.040)**	0.3707 (0.261)	-1.3967 *(0.081)*	-0.3634 *(0.080)*
	C	3.3627 (0.305)	7.1208 **(0.011)**	4.9512 ***(0.000)***	2.2403 (0.289)	4.7291 *(0.092)*	C	-1.9764 (0.234)	-1.5633 (0.198)	-0.1267 (0.795)	-0.5184 (0.607)	-0.5716 (0.672)
	R^2	0.626	0.61	0.739	0.669	0.623	R^2	0.107	0.107	0.059	0.09	0.068
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	(0.144)	(0.144)	(0.489)	(0.236)	(0.400)
Financial analysts	π_{t-1}	0.1624 ***(0.000)***	0.1704 ***(0.000)***	0.2542 ***(0.000)***	0.1344 ***(0.000)***	0.1667 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	-0.0152 (0.696)	-0.0159 (0.668)	-0.0117 (0.752)	-0.003 (0.940)	-0.0024 (0.953)
	df^m_t	0.1562 (0.794)	0.0426 (0.967)	0.0672 ***(0.000)***	2.1366 *(0.086)*	0.1165 (0.729)	$df^m_t - df^m_{t-1}$	0.5383 *(0.100)*	0.333 (0.520)	0.3792 (0.148)	0.4323 (0.538)	0.0671 (0.708)
	df^m_{t-1}	0.9969 (0.133)	-0.1376 (0.892)	1.0333 ***(0.004)***	1.7494 (0.156)	-0.1388 (0.728)	$df^m_{t-1} - df^m_{t-2}$	0.6199 *(0.086)*	-0.8697 *(0.092)*	-0.4401 (0.215)	-0.3148 (0.646)	-0.093 (0.667)
	df^m_{t-2}	0.7433 (0.257)	-0.297 (0.765)	0.0352 (0.939)	1.405 (0.259)	0.2002 (0.628)	$df^m_{t-2} - df^m_{t-3}$	-0.832 **(0.021)**	0.0794 (0.875)	0.8993 ***(0.010)***	-0.2915 (0.672)	0.1157 (0.595)
	df^m_{t-3}	1.2712 *(0.052)*	2.324 **(0.022)**	0.7893 *(0.077)*	1.3447 (0.295)	0.1429 (0.677)	$df^m_{t-3} - df^m_{t-4}$	0.4078 (0.284)	1.5838 ***(0.002)***	-0.6829 **(0.013)**	0.6863 (0.333)	0.0082 (0.964)
	C	-2.1422 (0.355)	3.7398 *(0.058)*	1.2051 ***(0.001)***	1.1574 (0.491)	4.8133 **(0.033)**	C	-2.2581 *(0.097)*	-1.9795 **(0.039)**	-0.4214 (0.296)	-0.5979 (0.513)	-0.769 (0.523)
	R^2	0.675	0.629	0.838	0.66	0.598	R^2	0.164	0.175	0.148	0.023	0.01
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	**(0.036)**	**(0.026)**	*(0.058)*	(0.904)	(0.984)

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. Time indicator T of document frequencies is set to t, t-1, t-2 and t-3. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively).

At the same time, it is interesting that banks change their inflation expectations under the influence of changes in the information environment around utility tariffs and fuel, as well as news about inflation in previous periods. Quite a significant lag in 2-3 months can be explained by the time of preparation of macroeconomic forecasts, which are the base for the answers in the survey.

Similar to banks, businesses expectations may be significantly affected by news about past inflation trends and utilities. However, businesses are the most sensitive to the news about foods. This may be explained by a high share of agriculture, food industry, retail and wholesale trade (related to food)

enterprises among the surveyed ones, which also corresponds to the structure of the economy of Ukraine. The importance of food news remains in the estimation of changes in businesses' inflation expectations too.

Households' expectations are the most sensitive to the number of news related to utility tariffs in the reported period and the quarter ago. This can primarily be explained by the high importance of utility tariffs for Ukrainian households. Thus, a significant part of tariffs is regulated by the government or local authorities, and changes in tariffs cause a substantial negative reaction from society. The share of utility tariffs in the CPI is relatively low, which is largely due to non-monetary subsidies that operated in previous periods. However, despite this for the average Ukrainian utility tariffs are one of the most urgent topics related to inflation, which is confirmed, among other things, by the results of our analysis. Households are also slightly sensitive to information about food in previous periods. Interestingly, these results are not confirmed to changes in households' inflation expectations. Thus, citizens change their estimates of future inflation under the influence of changes in the information field about the exchange rate three months ago, while the change in the importance of other topics has little effect on expectations dynamics.

The expectations of professional forecasters respond best to information on utility tariffs in the reporting and previous months, as well as on the exchange rate three months ago. In this case, financial analysts respond to both the amount of information and its change. This may reflect approaches to forecasting for such analysts. As usual, change in the exchange rate and the expected changes in utility tariffs have the greatest impact on updating forecasts.

I also decided to test the hypothesis that shorter-term trends in the media environment may better explain the process of formation of inflation expectations of different respondents. This is in line with the fact that most surveys are conducted in shorter period of time than a month. To this end, I use decadal indices of the frequency of mentions of these topics. Shocks in the news that last for several days can fade within a month, due to the rapid loss of interest in the topic, and therefore the monthly indices may not reflect real dynamics of the importance of individual events. Accordingly, it is important to apply indices with higher frequency. Going to the decimal level, I get a mixed frequency in the OLS, so to switch to one frequency, I just used matching by month. Thus, I compare news indices separately for the first, second and third decades of the reporting month with inflation expectations for the same reporting month. The procedure was repeated for individual news indices for the three decades of the previous month to the inflation expectations of the reporting month, given that respondents also respond to the dynamics of the media environment in previous periods.

Table 3. Relationship between decadal news indices and inflation expectations

Respondents	Variables	Inflation		Exchange rate		Utilities		Food		Fuel	
		Curr.	Prev.	Curr.	Prev.	Curr.	Prev.	Curr.	Prev.	Curr.	Prev.
Banks	π_{t-1}	0.3025 (0.000)	0.3007 (0.000)	0.2962 (0.000)	0.2924 (0.000)	0.2924 (0.000)	0.2767 (0.000)	0.2982 (0.000)	0.3023 (0.000)	0.2959 (0.000)	0.2892 (0.000)
	df^m I	-0.2913 (0.719)	1.0018 (0.272)	-0.5795 (0.593)	-1.7788 (0.030)	0.3746 (0.652)	0.3437 (0.609)	-0.3568 (0.783)	-0.766 (0.452)	-0.5005 (0.287)	0.0193 (0.961)
	df^m II	-0.7864 (0.142)	0.8323 (0.346)	-0.3309 (0.742)	3.6288 (0.007)	0.2311 (0.732)	0.8281 (0.291)	-1.3411 (0.334)	-0.2261 (0.849)	-0.132 (0.648)	-0.3379 (0.225)
	df^m III	0.5547 (0.255)	-0.2232 (0.778)	0.4306 (0.680)	-2.2436 (0.012)	-0.0498 (0.932)	-0.2502 (0.738)	-0.718 (0.637)	1.8529 (0.098)	0.1892 (0.568)	-0.0686 (0.799)
	C	8.6576 (0.002)	2.2805 (0.494)	8.2388 (0.000)	8.228 (0.000)	6.0084 (0.000)	5.2585 (0.000)	9.7445 (0.000)	6.6276 (0.000)	10.4832 (0.000)	10.3404 (0.000)
	R^2	0.825	0.825	0.814	0.867	0.816	0.828	0.82	0.83	0.823	0.829
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Businesses	π_{t-1}	0.3429 (0.000)	0.3468 (0.000)	0.3645 (0.000)	0.359 (0.000)	0.3285 (0.000)	0.3165 (0.000)	0.33 (0.000)	0.3378 (0.000)	0.3517 (0.000)	0.3527 (0.000)
	df^m I	-0.0008 (0.999)	-0.561 (0.409)	-0.0527 (0.940)	0.7648 (0.331)	0.9796 (0.052)	0.7777 (0.142)	1.3742 (0.104)	1.1391 (0.228)	0.722 (0.019)	0.4026 (0.146)
	df^m II	-0.7326 (0.355)	0.0077 (0.991)	1.194 (0.167)	-0.7624 (0.299)	-0.083 (0.908)	0.5752 (0.206)	0.5473 (0.558)	-0.2134 (0.865)	-0.1173 (0.682)	0.0089 (0.955)
	df^m III	0.9513 (0.167)	0.1996 (0.723)	-0.0056 (0.995)	1.2836 (0.110)	0.2448 (0.725)	0.1232 (0.812)	-0.2146 (0.831)	-0.5889 (0.632)	-0.0325 (0.906)	0.0459 (0.853)
	C	8.0788 (0.000)	9.5355 (0.000)	5.9074 (0.001)	5.8469 (0.001)	5.711 (0.000)	4.7702 (0.000)	6.5813 (0.000)	8.1705 (0.000)	3.9404 (0.051)	4.8362 (0.006)
	R^2	0.666	0.659	0.686	0.689	0.725	0.748	0.688	0.664	0.698	0.69
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Households	π_{t-1}	0.2332 (0.000)	0.2338 (0.000)	0.2273 (0.000)	0.2341 (0.000)	0.1569 (0.000)	0.1519 (0.000)	0.2257 (0.000)	0.2177 (0.000)	0.2307 (0.000)	0.2337 (0.000)
	df^m I	0.6075 (0.470)	0.3696 (0.641)	-2.1987 (0.031)	-0.6561 (0.507)	1.9786 (0.003)	2.3171 (0.001)	0.7161 (0.587)	1.3905 (0.280)	0.9265 (0.017)	0.6657 (0.098)
	df^m II	-0.53 (0.317)	-0.4296 (0.421)	0.8797 (0.287)	0.6005 (0.512)	0.6423 (0.249)	0.3836 (0.485)	2.0484 (0.088)	1.55 (0.192)	0.3043 (0.305)	0.4797 (0.083)
	df^m III	0.3656 (0.425)	0.3585 (0.484)	1.42 (0.159)	0.7847 (0.463)	0.1211 (0.810)	0.249 (0.621)	0.2029 (0.851)	0.4941 (0.682)	-0.4318 (0.120)	-0.4599 (0.136)
	C	8.7718 (0.001)	9.1639 (0.000)	10.3816 (0.000)	8.9515 (0.000)	5.5579 (0.000)	5.2581 (0.000)	7.0805 (0.000)	6.793 (0.000)	4.4817 (0.074)	5.1578 (0.043)
	R^2	0.609	0.605	0.648	0.614	0.709	0.718	0.628	0.627	0.647	0.638
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Financial analysts	π_{t-1}	0.1604 (0.000)	0.1623 (0.000)	0.1652 (0.000)	0.1662 (0.000)	0.1028 (0.000)	0.0922 (0.000)	0.1508 (0.000)	0.1492 (0.000)	0.1655 (0.000)	0.1668 (0.000)
	df^m I	1.692 (0.006)	1.4931 (0.011)	-1.1937 (0.134)	-0.8612 (0.263)	1.6138 (0.001)	1.8996 (0.000)	1.7747 (0.095)	1.4521 (0.131)	0.6686 (0.033)	0.5142 (0.100)
	df^m II	-0.1233 (0.775)	-0.0997 (0.805)	0.5855 (0.359)	0.745 (0.299)	0.7149 (0.069)	0.509 (0.175)	0.4292 (0.646)	0.3017 (0.733)	0.0583 (0.796)	0.1352 (0.544)
	df^m III	0.0259 (0.939)	0.4572 (0.218)	0.9114 (0.275)	0.47 (0.568)	-0.0186 (0.958)	0.2341 (0.506)	0.7651 (0.354)	1.1162 (0.228)	-0.2818 (0.190)	-0.3239 (0.172)
	C	2.5993 (0.138)	1.7719 (0.325)	6.7942 (0.000)	6.6295 (0.000)	3.2151 (0.000)	2.7488 (0.000)	4.3551 (0.000)	4.5878 (0.000)	3.9331 (0.051)	4.7127 (0.023)
	R^2	0.634	0.642	0.621	0.611	0.748	0.769	0.627	0.623	0.623	0.613
	p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively). Curr. stands for decades of reported month, Prev. – for previous month. Indicators I, II and III following document frequency indices denote number of decades.

Table 3 presents the results of OLS estimations of impact of decade news indices on formation of inflation expectations. As in previous case, I add latest available inflation data, which is published with delay, so I use actual inflation in period t-1.

Here we have a few interesting outcomes, that vary from our monthly estimations. For example, banks are sensitive to document frequency of news about exchange rate in all decades of previous month, while monthly indices do not show this relationship. At the same time, news about food is more important in last decade of previous month, although monthly indices show significance for current month. Businesses respond more to the news about utilities and fuel in the first decade of the reporting month. As for monthly indices, decade indices on utility tariffs affect the formation of inflation expectations of households. In this case, the most important are the first decades of the reporting and previous months. The expectations of financial analysts also proved to be most dependent on the frequency of news in the

first decades of the reporting and previous months. However, in addition to utilities, they follow the news about inflation (official figures are just published in the first decade of the month) and about fuel.

Another important opinion concerns the fact that the expectations of banks and enterprises meet once a quarter. Therefore, the time period for assessing the impact of news on inflation expectations was increased by applying a three-month moving average. This is especially important considering that the coefficients for the monthly indices are very volatile and even change the sign, depending on the applied lag. Table 4 presents the results of OLS estimations of impact of quarterly news indices (3-month moving average) on formation of inflation expectations.

Table 4. Relationship between quarterly news indices and inflation expectations

Respondents	Variables	Without transformations					Variables	1st difference				
		Inflation	Exchange rate	Utilities	Food	Fuel		Inflation	Exchange rate	Utilities	Food	Fuel
Banks	π_{t-1}	0.2962 ***(0.000)***	0.282 ***(0.000)***	0.2832 ***(0.000)***	0.3004 ***(0.000)***	0.2898 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.0623 (0.154)	0.0903 **(0.033)**	0.1041 **(0.015)**	0.1014 **(0.020)**	0.0895 **(0.033)**
	df^m_t	-1.7001 (0.170)	1.1527 (0.467)	0.1101 (0.918)	1.3103 (0.571)	-0.0158 (0.973)	$df^m_t - df^m_{t-1}$	1.6473 **(0.017)**	1.5458 *(0.094)*	-0.4047 (0.516)	0.3991 (0.766)	0.2467 (0.368)
	df^m_{t-1}	0.7511 (0.669)	-1.7379 (0.440)	1.1679 (0.557)	-1.4879 (0.672)	-0.4129 (0.615)	$df^m_{t-1} - df^m_{t-2}$	-0.7856 (0.425)	-3.4065 ***(0.010)***	1.6354 (0.165)	-1.1473 (0.576)	-0.6507 (0.176)
	df^m_{t-2}	1.5584 (0.669)	-0.1744 (0.938)	-1.2997 (0.512)	-0.6431 (0.855)	-0.4109 (0.616)	$df^m_{t-2} - df^m_{t-3}$	0.0814 (0.933)	1.3132 (0.322)	-2.1069 *(0.075)*	-0.0197 (0.992)	0.1333 (0.782)
	df^m_{t-3}	-1.1049 (0.394)	-0.5921 (0.704)	0.8456 (0.421)	0.0465 (0.984)	0.3926 (0.399)	$df^m_{t-3} - df^m_{t-4}$	0.2624 (0.719)	0.6112 (0.508)	0.9212 (0.140)	0.7249 (0.582)	0.4089 (0.139)
	C	8.6963 ***(0.000)***	9.9577 ***(0.000)***	5.5146 ***(0.000)***	7.9478 ***(0.000)***	10.5852 ***(0.000)***	C	-3.6686 ***(0.006)***	-0.1601 (0.781)	-0.1348 (0.764)	0.0122 (0.987)	-1.07 (0.174)
	R^2	0.812	0.822	0.817	0.81	0.826	R^2	0.159	0.114	0.08	0.059	0.134
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	***(0.003)***	**(0.030)**	(0.127)	(0.287)	**(0.012)**
Businesses	π_{t-1}	0.3567 ***(0.000)***	0.3775 ***(0.000)***	0.3229 ***(0.000)***	0.3442 ***(0.000)***	0.3681 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.1549 ***(0.000)***	0.1513 ***(0.000)***	0.1549 ***(0.000)***	0.1451 ***(0.001)***	0.1468 ***(0.000)***
	df^m_t	-1.3567 (0.305)	1.2189 (0.364)	0.6571 (0.428)	1.4312 (0.387)	0.3357 (0.245)	$df^m_t - df^m_{t-1}$	0.7995 (0.166)	0.8543 (0.162)	-0.3296 (0.416)	1.9479 ***(0.009)***	-0.0014 (0.991)
	df^m_{t-1}	0.2246 (0.914)	-0.788 (0.691)	0.7613 (0.596)	-1.2567 (0.665)	0.1356 (0.757)	$df^m_{t-1} - df^m_{t-2}$	-0.7318 (0.421)	-1.3392 (0.137)	1.38 *(0.051)*	-3.1614 **(0.015)**	0.1989 (0.319)
	df^m_{t-2}	1.4987 (0.470)	-0.2494 (0.899)	-0.3459 (0.807)	-0.6994 (0.809)	-0.0457 (0.916)	$df^m_{t-2} - df^m_{t-3}$	0.8714 (0.335)	0.6937 (0.442)	-1.1423 (0.107)	1.8089 (0.159)	-0.3072 (0.120)
	df^m_{t-3}	0.066 (0.961)	1.5895 (0.233)	0.4552 (0.576)	2.1718 (0.186)	0.1126 (0.676)	$df^m_{t-3} - df^m_{t-4}$	-0.8554 (0.144)	-0.0955 (0.875)	0.0518 (0.898)	-0.461 (0.524)	0.0751 (0.541)
	C	7.4593 ***(0.000)***	4.5674 ***(0.000)***	4.6576 ***(0.000)***	6.6463 ***(0.000)***	4.0647 ***(0.001)***	C	-0.2782 (0.492)	-0.2835 (0.498)	0.0782 (0.798)	-0.213 (0.475)	0.2446 (0.637)
	R^2	0.676	0.695	0.746	0.687	0.696	R^2	0.11	0.088	0.111	0.111	0.094
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	***(0.001)***	***(0.007)***	***(0.001)***	***(0.001)***	***(0.004)***

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. Time indicator T of document frequencies is set to t, t-1, t-2 and t-3 and corresponds to quarters. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively).

As can be seen, the hypothesis that banks and enterprises follow longer trends is mostly not confirmed. At the same time, the long-term change in the information space about inflation and the exchange rate is related to the change in inflation expectations of banks, and the change in the volume of food news affects the inflation expectations of enterprises. However, in both cases, this impact is limited to 1-2 quarters.

I repeat similar procedure to reveal the impact of indices built by LDA on formation of inflation expectations. Table 5 presents the coefficients and p-values (in brackets) of news indices built by LDA in OLS regressions of inflation expectations of different groups of respondents. Similar to simple indices I tested two different approaches: without transformations, using all variables as they were computed, and the 1st difference of all variables, presenting extrapolation of change of inflation expectations.

Table 5. Relationship between monthly news indices built by LDA and inflation expectations

Topic	Variables	Without transformations				Variables	1st difference			
		Banks	Businesses	Households	Financial analysts		Banks	Businesses	Households	Financial analysts
Energy	π_{t-1}	0.2707 ***(0.000)***	0.3437 ***(0.000)***	0.2051 ***(0.000)***	0.1626 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.1731 ***(0.003)***	0.1598 ***(0.001)***	0.0532 (0.291)	-0.0016 (0.970)
	df^m_t	0.4646 (0.766)	-1.0941 (0.413)	-2.7212 **(0.037)**	-0.3453 (0.754)	$df^m_t - df^m_{t-1}$	0.7627 (0.622)	-1.8396 *(0.059)*	-1.1436 (0.111)	0.3425 (0.564)
	df^m_{t-1}	-1.8021 (0.296)	2.4938 *(0.072)*	-1.5848 (0.196)	-0.3938 (0.708)	$df^m_{t-1} - df^m_{t-2}$	-1.0257 (0.538)	1.2913 (0.197)	1.08 (0.116)	0.2248 (0.693)
	df^m_{t-2}	-0.9249 (0.405)	0.5825 (0.657)	-1.7524 (0.151)	-0.4304 (0.677)	$df^m_{t-2} - df^m_{t-3}$	-0.2834 (0.809)	0.2077 (0.833)	0.0709 (0.916)	0.414 (0.455)
	df^m_{t-3}	-0.5554 (0.758)	-2.0173 (0.112)	-1.803 (0.148)	-1.5595 (0.165)	$df^m_{t-3} - df^m_{t-4}$	-0.1006 (0.956)	-0.8172 (0.374)	0.2491 (0.249)	-1.0749 *(0.070)*
	C	15.9023 ***(0.000)***	8.9195 **(0.044)**	32.4739 ***(0.000)***	14.6931 ***(0.000)***	C	1.9027 (0.579)	3.4848 (0.209)	-0.7805 (0.729)	0.1552 (0.936)
	R^2	0.846	0.685	0.718	0.621	R^2	0.276	0.215	0.072	0.054
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	*(0.081)*	**(0.022)**	(0.369)	(0.592)
Exchange rate	π_{t-1}	0.2207 ***(0.000)***	0.2349 ***(0.000)***	0.0553 **(0.028)**	0.0443 **(0.021)**	$\pi_{t-1} - \pi_{t-2}$	0.1404 **(0.011)**	0.1296 ***(0.005)***	0.0376 (0.422)	-0.0148 (0.703)
	df^m_t	0.2808 (0.622)	1.9069 ***(0.008)***	1.2814 ***(0.004)***	0.7794 **(0.022)**	$df^m_t - df^m_{t-1}$	0.25 (0.667)	0.6776 (0.232)	0.6895 **(0.020)**	-0.0101 (0.968)
	df^m_{t-1}	1.0477 (0.254)	-0.6672 (0.334)	0.0629 (0.904)	0.9172 **(0.032)**	$df^m_{t-1} - df^m_{t-2}$	0.9799 (0.265)	-0.1744 (0.770)	-0.7571 **(0.030)**	0.6726 **(0.031)**
	df^m_{t-2}	0.7299 (0.397)	0.4674 (0.571)	0.7218 (0.165)	-0.214 (0.608)	$df^m_{t-2} - df^m_{t-3}$	0.5198 (0.529)	0.4751 (0.509)	0.5903 *(0.088)*	-0.702 **(0.025)**
	df^m_{t-3}	-0.481 (0.520)	0.7308 (0.322)	1.0341 **(0.022)**	0.6904 **(0.047)**	$df^m_{t-3} - df^m_{t-4}$	-1.4949 **(0.033)**	-0.9657 (0.109)	-0.4956 *(0.085)*	0.0254 (0.917)
	C	-5.9118 (0.203)	-11.2109 **(0.015)**	-15.1982 ***(0.000)***	-10.7207 ***(0.000)***	C	-2.2814 (0.441)	-0.229 (0.938)	-0.3152 (0.800)	0.0642 (0.953)
	R^2	0.864	0.756	0.821	0.809	R^2	0.397	0.217	0.147	0.102
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	p-value	***(0.009)***	**(0.021)**	**(0.043)**	(0.209)

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. Time indicator T of document frequencies is set to t, t-1, t-2 and t-3. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively).

According to the results of regressions, I observe a weak correspondence between the news about energy and utility tariffs, which were determined by the LDA, and the formation of inflation expectations. Inflationary expectations of households, as well as changes in business expectations demonstrate significance in reporting month at 5% and 10% levels respectively, but the sign of the coefficients of these variables contradicts economic logic, which can be associated either with emotional coloring, or to reflect the coincidence of circumstances.

Instead, the situation is somewhat different with the exchange rate news set by the LDA. The frequency of such news in the reporting period was significant for the formation of expectations of businesses, households and financial analysts. For households and financial analysts, these indices have also been important in recent months. Financial analysts and households have also been sensitive to changes in the frequency of exchange rate news. However, households are changing their expectations in response to more recent developments, while financial analysts are responding to a longer period.

Similar to simple indices, I identified short-term spot effects on the formation of inflation expectations by estimating decade indices. Table 6 represent the results of this estimation.

Table 6. Relationship between decade news indices built by LDA and inflation expectations

Respondents	Variables	Banks		Businesses		Households		Financial analysts	
		Curr.	Prev.	Curr.	Prev.	Curr.	Prev.	Curr.	Prev.
Energy	π_{t-1}	0.2788 ***(0.000)***	0.2715 ***(0.000)***	0.3554 ***(0.000)***	0.3647 ***(0.000)***	0.206 ***(0.000)***	0.2151 ***(0.000)***	0.1588 ***(0.000)***	0.1605 ***(0.000)***
	df^m I	-2.2081 **(0.038)**	-1.2279 (0.176)	0.8904 (0.394)	-0.1089 (0.914)	-2.8889 ***(0.004)***	-2.5453 **(0.013)**	-1.336 (0.103)	-1.3255 *(0.098)*
	df^m II	0.1589 (0.875)	0.6438 (0.413)	-1.9179 **(0.043)**	0.325 (0.702)	-1.1732 (0.163)	-0.9812 (0.243)	-0.5387 (0.478)	0.1262 (0.858)
	df^m III	-0.1123 (0.891)	-1.8735 **(0.017)**	1.0673 (0.102)	1.2336 (0.136)	-1.1909 *(0.078)*	-0.9189 (0.197)	0.0937 (0.882)	-0.4895 (0.384)
	C	13.8197 ***(0.001)***	14.9455 ***(0.000)***	8.388 **(0.025)**	4.0943 (0.210)	25.1004 ***(0.000)***	22.6818 ***(0.000)***	12.0712 ***(0.000)***	11.8456 ***(0.000)***
	R^2	0.838	0.865	0.687	0.673	0.689	0.665	0.613	0.615
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***
Exchange rate	π_{t-1}	0.2496 ***(0.000)***	0.2161 ***(0.000)***	0.2851 ***(0.000)***	0.3141 ***(0.000)***	0.1237 ***(0.000)***	0.1121 ***(0.000)***	0.0774 ***(0.000)***	0.0704 ***(0.000)***
	df^m I	0.8894 *(0.072)*	0.8441 **(0.049)**	1.007 **(0.014)**	-0.0078 (0.987)	1.8598 ***(0.000)***	1.9123 ***(0.000)***	1.5874 ***(0.000)***	1.2134 ***(0.000)***
	df^m II	0.8664 **(0.038)**	0.4398 (0.371)	-0.0442 (0.917)	0.3258 (0.463)	0.5687 *(0.058)*	0.7552 **(0.012)**	0.5167 **(0.016)**	0.5048 **(0.022)**
	df^m III	-0.6396 (0.110)	0.3476 (0.424)	0.7195 (0.132)	0.4304 (0.328)	-0.2831 (0.373)	-0.4095 (0.184)	-0.3247 (0.132)	0.057 (0.824)
	C	-2.2563 (0.566)	-6.3928 (0.126)	-5.2484 (0.155)	2.4127 (0.520)	-7.5191 ***(0.006)***	-8.434 ***(0.003)***	-7.5665 ***(0.000)***	-7.462 ***(0.000)***
	R^2	0.849	0.864	0.734	0.682	0.752	0.756	0.796	0.772
	p-value	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***	***(0.000)***

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively). Curr. stands for decades of reported month, Prev. – for previous month. Indicators I, II and III following document frequency indices denote number of decades.

Interestingly, for some groups of respondents there is a clear relationship with the indices in the periods when the surveys are conducted. For example, bank surveys are usually conducted at the beginning of the month, and sometimes cover even the week of the previous month. News about energy in the last decade of the previous month and in the first decade of the reporting month turned out to be significant. The situation is similar with businesses and households. At the same time, the sign of the coefficients needs further study in terms of sentiments. Inflation expectations of businesses are formed under the influence of news about the exchange rate in the first decade of the reporting month, while all other respondents follow the news for previous periods.

I repeated the same procedure for determining the longer-term impact of news on the formation of inflation expectations using the current three-month average for banks and corporates. Table 7 shows the main results of the estimations. However, the results indicate the absence of any long-term impact of news on the formation of inflation expectations. Only the expectations of enterprises have a significant connection with the change in the frequency of news about the exchange rate in the current quarter.

Table 7. Relationship between quarterly news indices built by LDA and inflation expectations

Topic	Variables	Without transformations		Variables	1st difference	
		Banks	Businesses		Banks	Businesses
Energy	π_{t-1}	0.2675 ***(0.000)***	0.3566 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	0.1846 ***(0.001)***	0.1466 ***(0.002)***
	df^m_t	-1.3148 (0.744)	1.1802 (0.661)	$df^m_t - df^m_{t-1}$	0.3972 (0.914)	-2.3983 (0.206)
	df^m_{t-1}	-2.459 (0.609)	1.6579 (0.695)	$df^m_{t-1} - df^m_{t-2}$	-5.1964 (0.261)	4.6212 (0.137)
	df^m_{t-2}	1.2399 (0.760)	-2.7387 (0.503)	$df^m_{t-2} - df^m_{t-3}$	2.3077 (0.550)	-3.169 (0.275)
	df^m_{t-3}	-0.5603 (0.775)	0.4886 (0.857)	$df^m_{t-3} - df^m_{t-4}$	2.147 (0.248)	-0.0908 (0.962)
	C	16.6815 ***(0.000)***	6.7721 (0.148)	C	0.9729 (0.758)	3.1107 (0.271)
	R^2	0.844	0.662	R^2	0.341	0.198
	p-value	***(0.000)***	***(0.000)***	p-value	**(0.027)**	**(0.034)**
Exchange rate	π_{t-1}	0.2194 ***(0.000)***	0.2357 ***(0.000)***	$\pi_{t-1} - \pi_{t-2}$	-0.7182 **(0.021)**	0.1224 ***(0.008)***
	df^m_t	1.898 (0.214)	2.8326 *(0.059)*	$df^m_t - df^m_{t-1}$	1.8418 (0.207)	2.6938 **(0.022)**
	df^m_{t-1}	-0.9758 (0.752)	-2.127 (0.399)	$df^m_{t-1} - df^m_{t-2}$	-0.0609 (0.983)	-3.355 *(0.098)*
	df^m_{t-2}	1.3292 (0.665)	0.174 (0.941)	$df^m_{t-2} - df^m_{t-3}$	0.1379 (0.961)	1.801 (0.341)
	df^m_{t-3}	-0.6878 (0.639)	1.4948 (0.375)	$df^m_{t-3} - df^m_{t-4}$	-1.8472 (0.156)	-1.1601 (0.388)
	C	-5.8439 (0.247)	-10.7359 **(0.025)**	C	0.1227 (0.807)	0.037 (0.990)
	R^2	0.859	0.743	R^2	0.432	0.241
	p-value	***(0.000)***	***(0.000)***	p-value	***(0.004)***	**(0.010)**

Notes: The table shows results of OLS regressions where inflation expectations are dependent variable. Time indicator T of document frequencies is set to t, t-1, t-2 and t-3 and corresponds to quarters. First figures in cells indicate coefficients, p-values are shown in parentheses (1%, 5% and 10% significance levels highlighted by underlined bold italic (blue), bold (green) and italic (orange) respectively).

Thus, the formation of inflation expectations of different groups of respondents may depend on the media environment, namely both the volume of published articles and changes in this indicator. It is important that different groups of respondents rely on different topics and different periods when estimating future inflation. It can also be seen that recent news, published during previous month and even decade preceding the survey, is mostly more important in shaping inflation expectations than older ones. This may, among other things, be important for the central bank's communication policy.

Conclusions

In this paper, I analyzed the role of news in the formation of inflation expectations of different types of respondents in Ukraine using textual data. I scraped the news corpus from four Ukrainian online newspapers listed in the most popular online media in Ukraine, which have mainly economic orientation. Using natural language processing and machine learning techniques I cleaned and transformed textual data into news-based quantitative measures reflecting news topics relevant to inflation and inflation expectations.

I applied two different approaches to filter out the news: a dictionary-based approach and Latent Dirichlet Allocation (LDA). Both approaches do not consider the sentiments of news content, which we leave for future research. I computed all news indices as a “document frequency” following the intuition that the more alarming the topic – the more articles would be written on the subject.

I assumed that the impact of the constructed news indices on inflation expectations is linear and estimated this effect with the OLS regression. I tested the impact on the level of inflation expectations as well as on the change thereof. As such, I found evidence that the different news topics have an impact on inflation

expectations. For example, I found a strong relationship between inflation expectations of households and financial analysts with news about utilities, while businesses sensitive to the news about food. Additionally, financial analysts and households have also been sensitive to levels and changes in the frequency of exchange rate news, constructed by LDA.

I also tested the hypothesis that shorter-term trends in the media environment may better explain the process of formation of inflation expectations of different respondents as document frequency may vary during the month and the impact of the short-term news may fade away. I proved that for some groups of respondents there is a clear relationship with the indices within the periods when the surveys are conducted. I also showed that recent news is mostly more important in shaping inflation expectations than older ones.

As a result, the formation of inflation expectations of different groups of respondents may depend on the media environment, namely both the volume of published articles and changes in this indicator. Different groups of respondents rely on different topics and different periods when assessing future inflation. I also found that some events contradict economic logic, which could be a question for future research. In particular, an important issue is the impact of news indices in different periods (during stable, accelerating inflation, or disinflation). Other research questions may include the assessment of the tone of news, their relationship with other macroeconomic indicators, as well as the predictive power of such indices.

Results of this research can help understand inflation expectations, especially as anchoring inflation expectations remains a key challenge for central banks. This may, among other things, be important for the central bank's communication policy and help to articulate clear and effective messages, as well as to design optimal policy.

Appendices

Appendix A. News corpus

Originally, the news corpus consists of 2 030 000 unique articles, however, after cleaning and filtering out items with different types of errors (parsing errors when web-page tags are wrongly placed, empty pages, corrupted symbols etc.), the remaining number of articles decreased by 50 000 items. As this is only 2.5% of the total number of articles, we consider such a reduction quite acceptable and not to affect the overall result.

Table 8. Article size in news corpus (after cleaning)

	finance.ua	liga	ukr_pravda	unian	*Total*
count	389951	620655	339275	634832	*1985143*
mean	120.6	121.5	131.0	151.5	*132.5*
std	96.8	79.6	94.2	123.9	*102.1*
min	0	2	4	3	*0*
25%	63	76	83	88	*78*
50%	99	108	115	127	*113*
75%	151	149	157	182	*162*
max	3832	11540	5842	3986	*11540*
skewness	3.305	24.465	9.911	7.766	*10.072*
kurtosis	32.299	2456.736	228.865	109.719	*394.412*

Articles in corpus differ not only in content but also in writing style, size as measured by word count, and other features (Table 8). Expectedly, different sources of information have some dissimilarities in how news is written which is e.g., revealed in the article size. Unian has the largest articles on average, while finance.ua has the smallest articles. At the same time, the sizes of articles from all sources are very close.

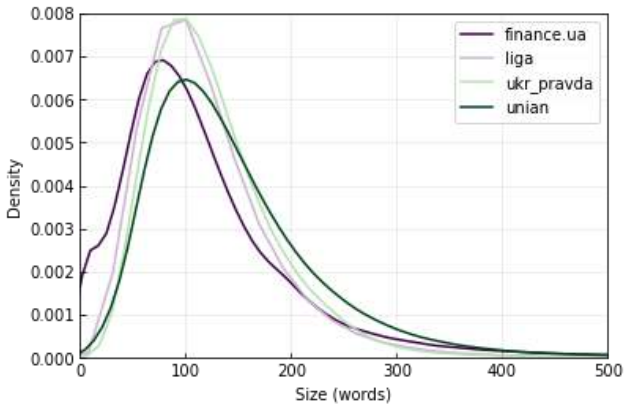


Figure 9. Distribution of article size by sources

The distribution of article sizes for all news sources (figure 9) is highly asymmetrical. All values of the skewness are positive and the tail of the distribution is longer towards the right-hand side of the curve. Though, articles from Liga are the most skewed. At the same time, distributions of article length are leptokurtic, which means they are tall and thin, and so near the mean. For example, the number of articles

with a length of more than 500 words is less than 15 000, which is only 0.7% of the corpus. The number of articles with extremely small size[1] is also negligible (around 2.5% of the total number).

Appendix B. Inflation expectations survey design

In contrast to financial analysts, who are provided to answer open question, banks, businesses, and households are asked to pick from a set of inflation intervals, for example:

"Inflation over the next twelve months will be:

a) less than zero ("prices will fall"),
b) between 0 and X percent,
c) between X and 2X percent,
d) between 2X and 3X percent,
e) between 3X and 4X percent,
f) over 4X percent.

In this example, inflation expectations would be computed by formula:

$$E\pi = w_a \cdot \left(-\frac{X}{2}\right) + w_b \cdot \frac{X}{2} + w_c \cdot \frac{X+2X}{2} + w_d \cdot \frac{2X+3X}{2} + w_e \cdot \frac{3X+4X}{2} + w_f \cdot (4X + \frac{X}{2}) , \qquad (6)$$

where *w* is a share of respondents, who pick respective interval. Size of X as well as number of intervals is not fixed and changes over time to match normal distribution of answers. Thus, in 2015 inflation in Ukraine accelerated drastically, so the maximum bracket was expanded to 50% and respondents selected from 12 intervals. Following disinflation in 2020 maximum bracket was decreased to 10% and number of intervals was cut to 8.

Since January 2018 surveys of households also include question about inflation perceptions. Once a year consumers are asked to answer on open question about perceived inflation over the previous 12 months. Additionally, households are also asked to pick answers from interval question once in a quarter. Construction of this question is similar to inflation expectations question.

Appendix C. Dictionary-based approach

Appendix C.1. Statistical characteristics

We selected topics that may have the biggest impact on formation of inflation expectations. Table 9 contains detailed characteristics of each topic. News related to fuel and oil are the most frequent in our corpus with mean document frequency of 8%, while news about food, can be met less often – its average document frequency is near 1.3%. At the same time, share of news related to fuel is more volatile, while the frequency of food and exchange rate news were more stable. Decade indices show similar tendencies, although with largest extreme value distribution: lower minimum values, higher maximum values and standard deviations.

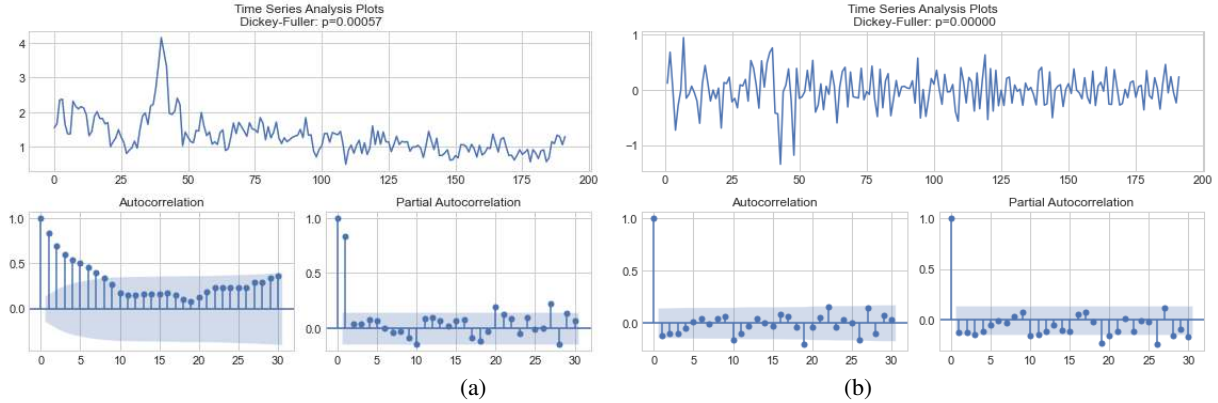
[1] The average sentence ranges from 15 to 20 words, we consider that the smallest article consists of a topic and one sentence ≈ 30 words.

Table 9. Dictionary-based indices characteristics

	Foods	Utilities	Fuel	ER	Inflation
Description	Names of the most popular foods	The most important utilities	Fuel and oil	Exchange rate developments	Inflation review and forecasts
Key words	*'гречка', 'овощ', 'картошка', 'фрукт', 'молоко', 'мясо', 'яйцо', 'мука', 'крупа', 'масло', 'морковь', 'яблоко', 'хлеб', 'сахар'*	*'газ', 'электричество', 'тариф', 'коммунальный'*	*'бензин', 'нефть'*	*'доллар', 'евро', 'курс'* in combination with *'гривна', 'грн'*	*'инфляция'*
Monthly document frequency, %					
mean, %	1.309397	2.958930	8.002455	2.106668	2.661430
min, %	0.484653	0.815411	4.276804	1.164327	0.787697
max, %	4.163525	7.393600	25.084746	4.470135	4.564644
std, p.p.	0.549031	1.205741	2.375635	0.610790	0.802666
ADF p	0.00057	0.64143	0.00000	0.47538	0.10667
ADF p 1st diff	0.00000	0.00002	0.00000	0.00013	0.00000
Decade document frequency, %					
mean, %	1.313320	2.957145	8.028077	2.101207	2.690746
min, %	0.185931	0.227790	2.836879	0.408163	0.454545
max, %	5.170800	10.647737	36.338419	5.577173	10.022779
std, p.p.	0.634849	1.326824	2.691632	0.732638	1.044147
ADF p	0.00190	0.16471	0.00000	0.04130	0.66145
ADF p 1st diff	0.00000	0.00000	0.00000	0.00000	0.00000

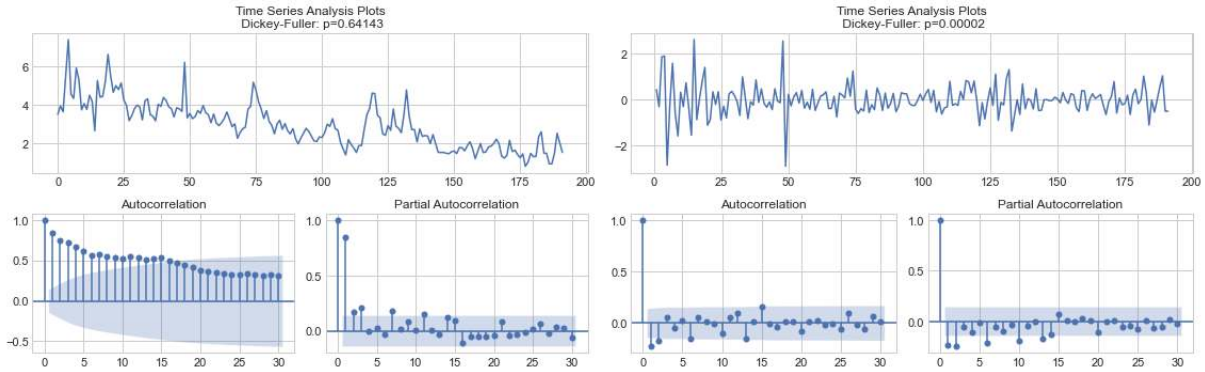
Appendix C.2. Monthly time series

Monthly time series of document frequency are serially correlated (figures 10-14). As autocorrelation should be avoided in order to apply further data analysis more accurately, we applied first difference for each time series. In this case, we got rid of serial correlation and received stationary data.



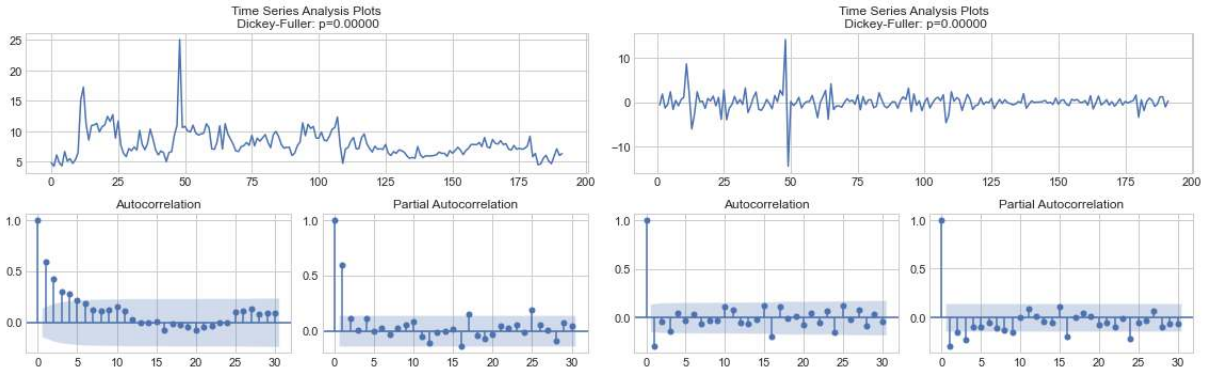
(a) (b)

Figure 10. Share of news related to foods (a) and its 1st difference (b)



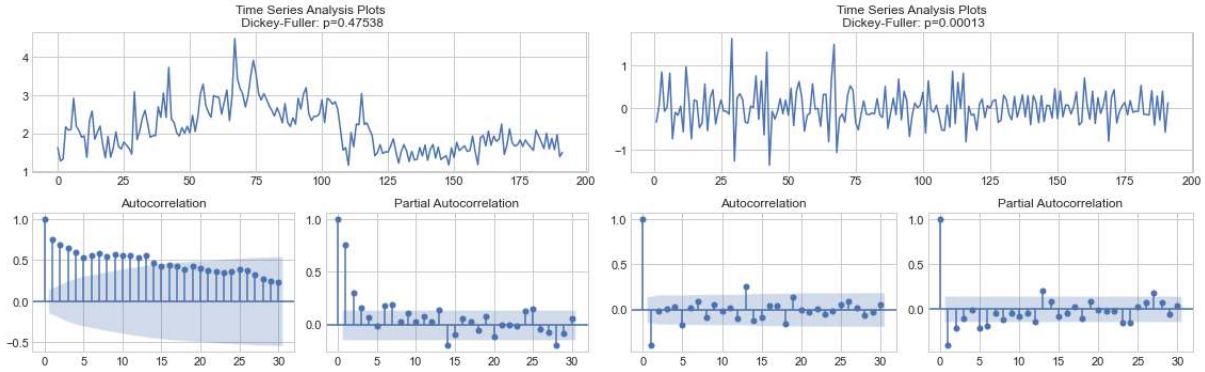
(a) (b)

Figure 11. Share of news related to utilities (a) and its 1st difference (b)



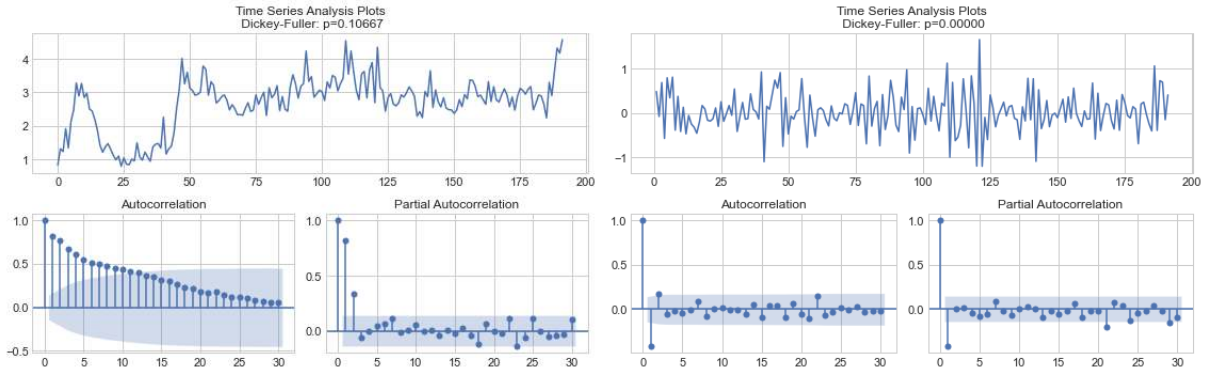
(a) (b)

Figure 12. Share of news related to fuel (a) and its 1st difference (b)



(a) (b)

Figure 13. Share of news related to exchange rate (a) and its 1st difference (b)

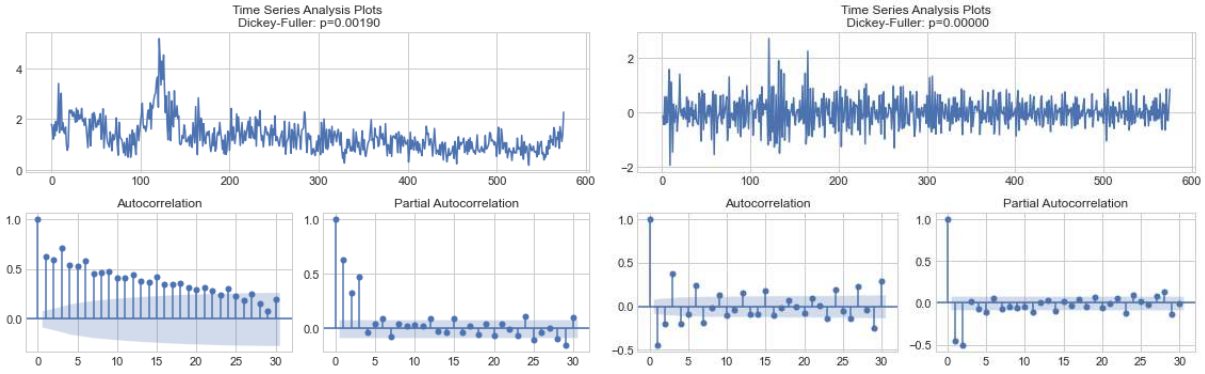


(a) (b)

Figure 14. Share of news related to inflation (a) and its 1st difference (b)

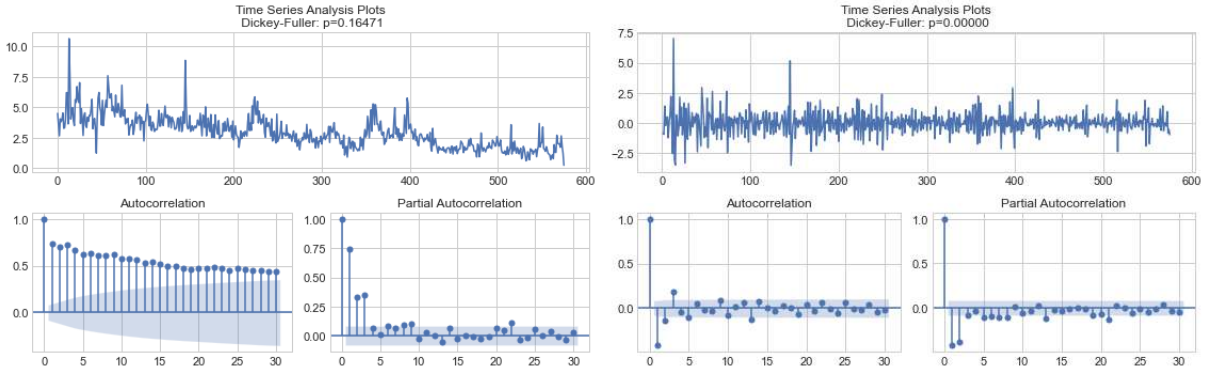
Appendix C.3. Decade time series

Decade time series of document frequency are also serially correlated (figures 15-19). Application of first difference for decade time series similarly helped to avoid serial correlation and receive stationary data.



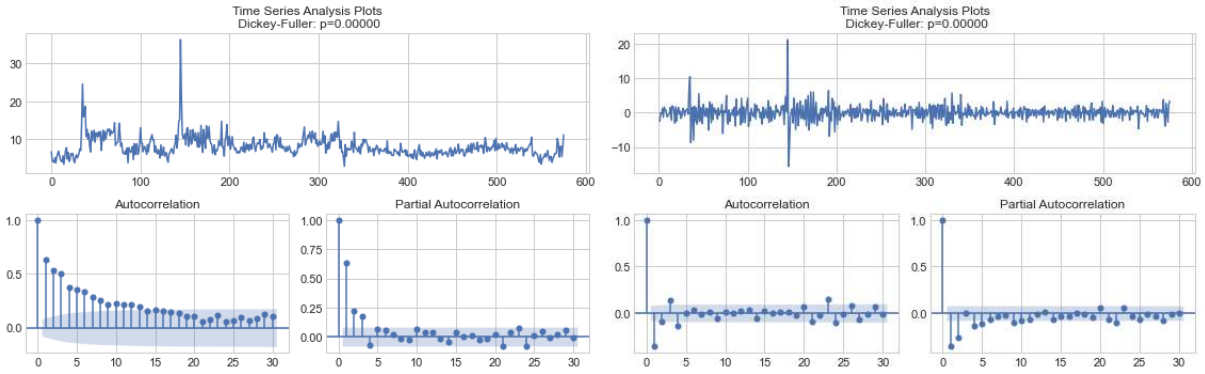
(a) (b)

Figure 15. Share of news related to foods (a) and its 1st difference (b)



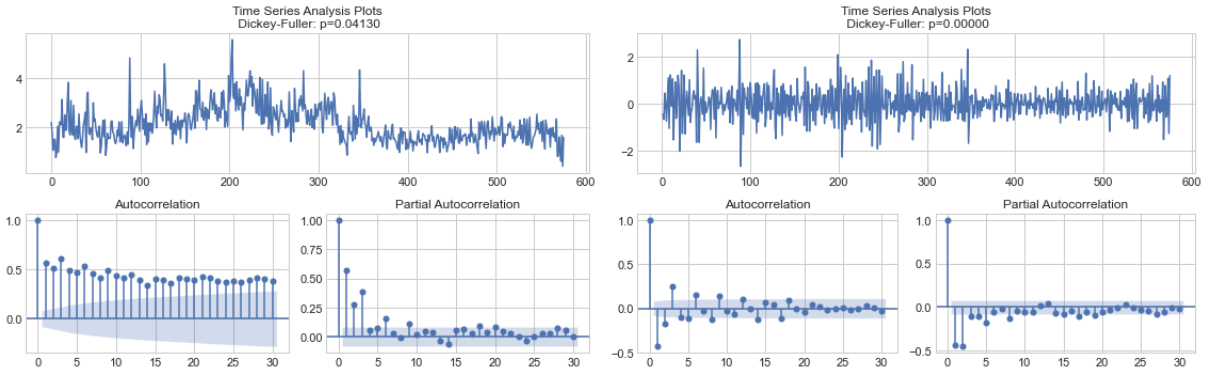
(a) (b)

Figure 16. Share of news related to utilities (a) and its 1st difference (b)



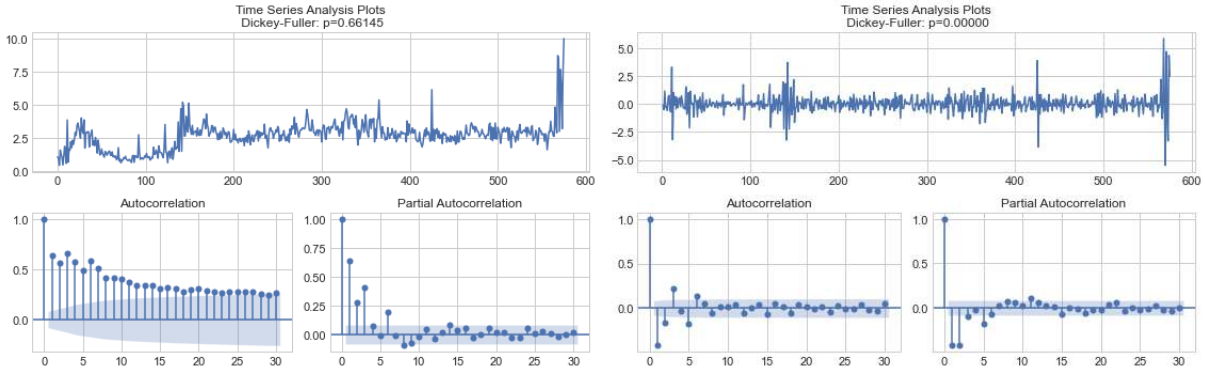
(a) (b)

Figure 17. Share of news related to fuel (a) and its 1st difference (b)



(a) (b)

Figure 18. Share of news related to exchange rate (a) and its 1st difference (b)



(a) (b)

Figure 19. Share of news related to inflation (a) and its 1st difference (b)

Appendix D. LDA approach

Appendix D.1. Additional most meaningful news topics identified by LDA (wordclouds and shares)

LDA helped to identify other different social and political topics. For example, figures 20-22 show topics related to conflict between Ukraine and Russia, the spread of coronavirus and government spending. These topics can be hardly related to formation of inflation expectations, although their existence is evidence of high quality of LDA model. At the same time, these results may be used in further research.

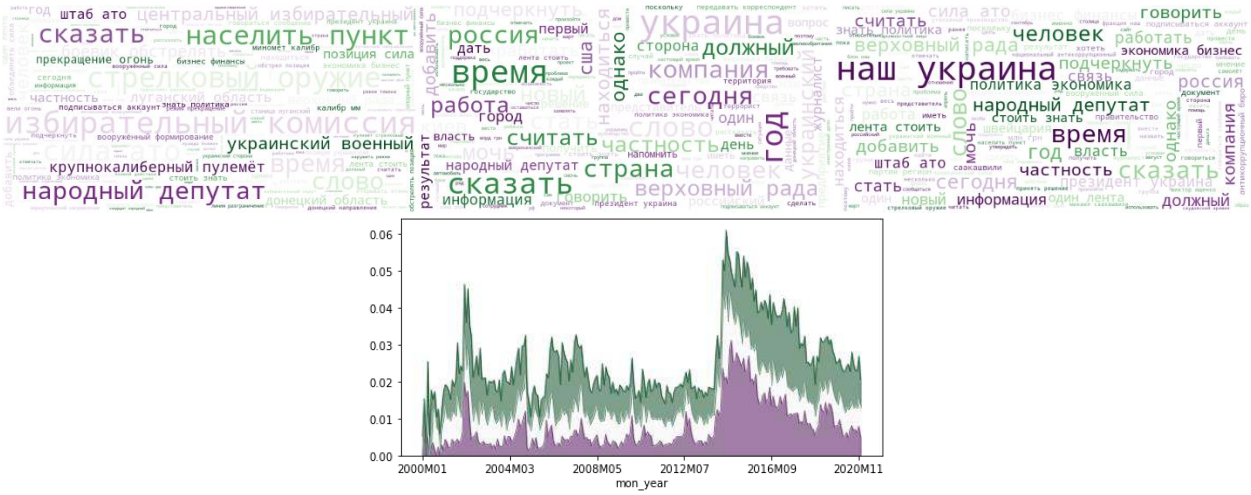


Figure 20. Wordclouds and share of news about Russian-Ukrainian conflict

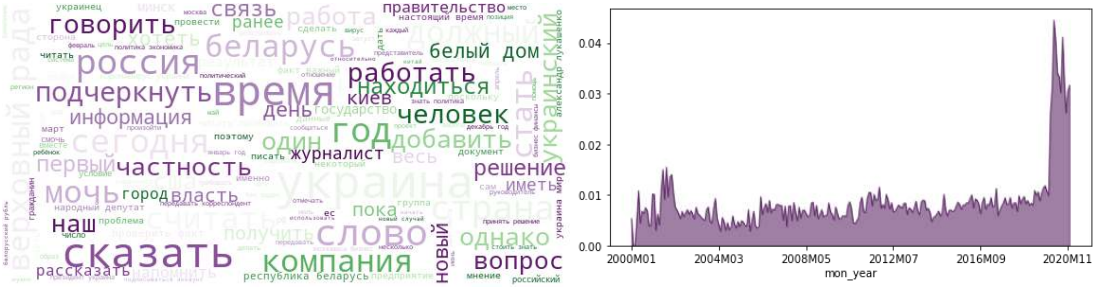


Figure 21. Wordclouds and share of news related to coronavirus

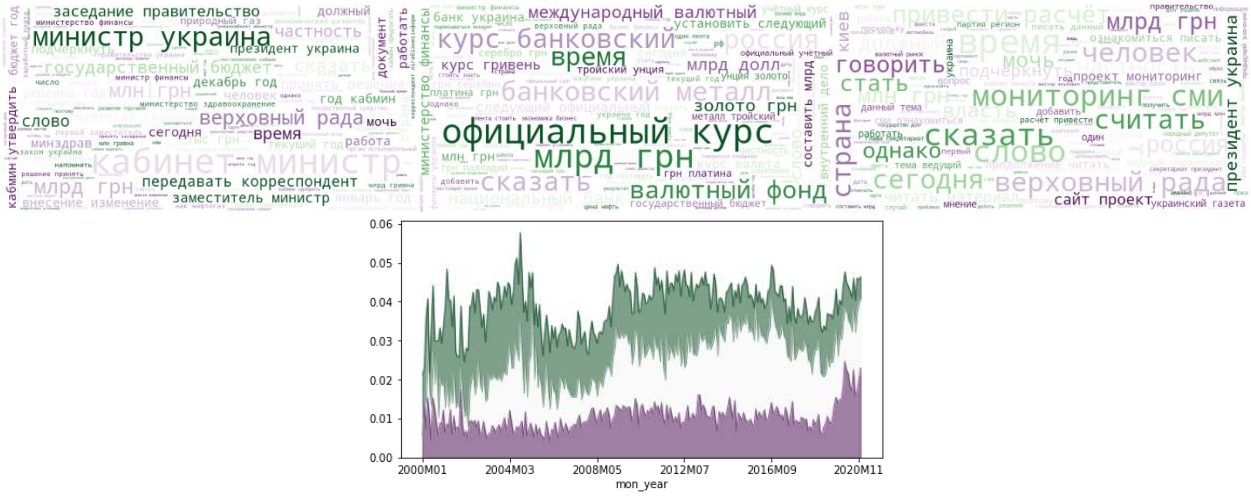


Figure 22. Wordclouds and share of news related to government spending and debt

Appendix D.2. Statistical characteristics

We reviewed topics created by LDA and selected all that can be attributed to inflation expectations. Thus, we had only two clusters of news – about energy and exchange rate. News related to the exchange rate have quite a high frequency in our corpus with mean share of 8.8%, while news about energy occur in 3% of documents (Table 10). Decade indices show similar tendencies, although with largest extreme value distribution: lower minimum values, higher maximum values and standard deviations.

Table 10. LDA indices characteristics

	Energy	Exchange rate
Monthly document frequency, %		
mean, %	3.087231	8.781308
min, %	2.163307	6.932337
max, %	4.762712	11.718750
std, p.p.	0.439327	1.025374
ADF p	0.12235	0.03602
ADF p 1st diff	0.00000	0.00000
Decade document frequency, %		
mean, %	3.095512	8.809923
min, %	1.529637	5.833333
max, %	5.825243	15.141431
std, p.p.	0.599796	1.272625
ADF p	0.03822	0.03197
ADF p 1st diff	0.00000	0.00000

Appendix D.3. Monthly time series

Similarly to indices constructed by dictionary-based approach, monthly time series of LDA document frequency are serially correlated and their first difference are not (figures 23-24).

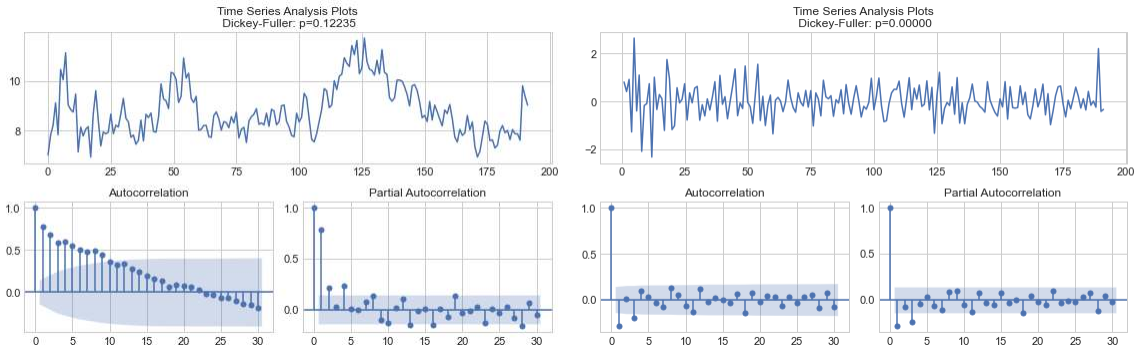


Figure 23. Share of news related to exchange rate (a) and its 1st difference (b)

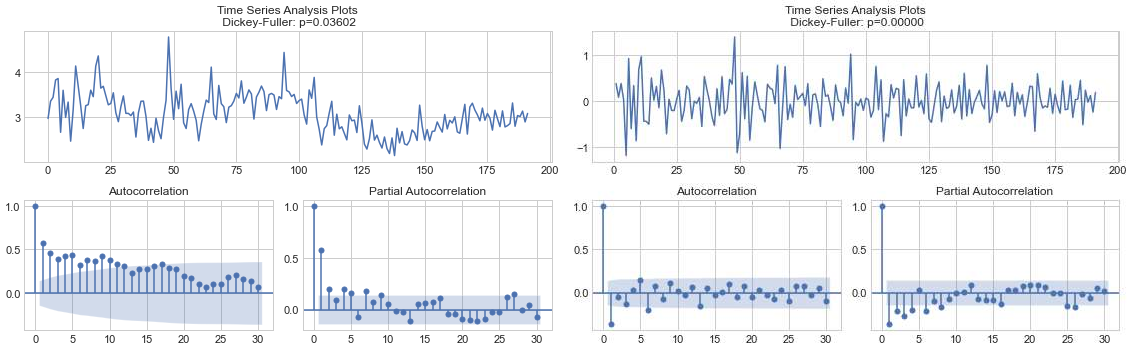


Figure 24. Share of news related to oil and gas (a) and its 1st difference (b)

Appendix D.4. Decade time series

Decade time series of LDA document frequency also are serially correlated and their first difference are not (figures 25-26).

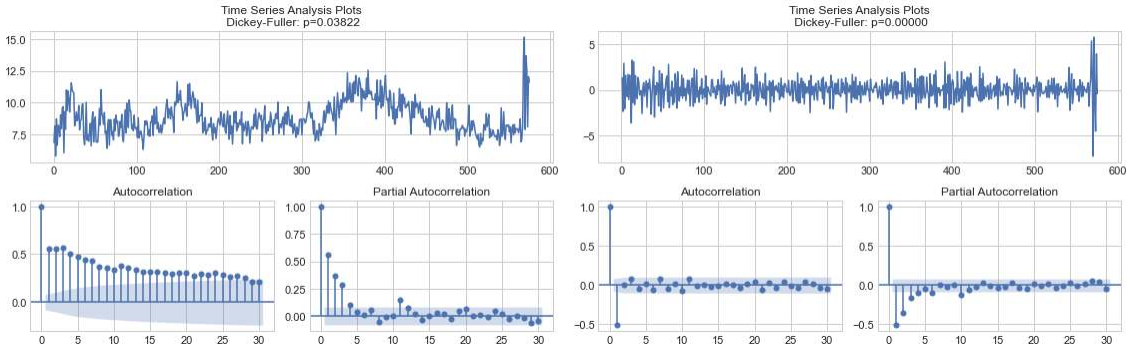


Figure 25. Share of news related to exchange rate (a) and its 1st difference (b)

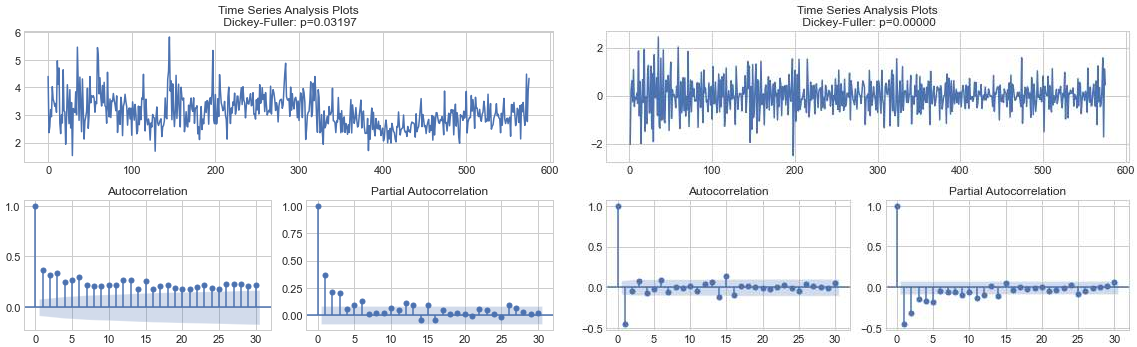


Figure 26. Share of news related to oil and gas (a) and its 1st difference (b)

Literature

Angelico, C., Marcucci, J., Miccoli, M. and Quarta, F., (2021), Can we measure inflation expectations using Twitter?, No 1318, Temi di discussione (Economic working papers), Bank of Italy, Economic Research and International Relations Area, https://EconPapers.repec.org/RePEc:bdi:wptemi:td_1318_21.

Azqueta Gavaldon, Andres. (2017). Developing news-based Economic Policy Uncertainty index with unsupervised machine learning. Economics Letters. 158. 10.1016/j.econlet.2017.06.032.

Baker, Scott & Bloom, Nicholas & Davis, Steven. (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics. 131. qjw024. 10.1093/qje/qjw024.

Bauer Michael D., 2015. "Inflation Expectations and the News," International Journal of Central Banking, International Journal of Central Banking, vol. 11(2), pages 1-40, March.

Blei, David & Ng, Andrew & Jordan, Michael. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research. 3. 993-1022. 10.1162/jmlr.2003.3.4-5.993.

Carroll, C.D., 2003. Macroeconomic expectations of households and professional forecasters. Q. J. Econ. 118 (1), 269–298.

Coibion, O., Gorodnichenko, Y., 2012. What can survey forecasts tell us about information rigidities? J. Polit. Econ. 120 (1), 116–159.

Coibion, O., Gorodnichenko, Y. (2015a). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. The American Economic Review, 105(8), 2644-2678. http://www.jstor.org/stable/43821351

Coibion, O., Gorodnichenko, Y. (2015b). Inflation Expectations in Ukraine: A Long Path to Anchoring?. Visnyk of the National Bank of Ukraine, 233, 6-23. https://doi.org/10.26531/vnbu2015.233.006

Coibion, O., Gorodnichenko, Y., Weber, M. (2019). Monetary Policy Communications and their Effects on Household Inflation Expectations, National Bureau of Economic Research Working Paper Series, No. 25482, https://doi.org/10.3386/w25482, http://www.nber.org/papers/w25482

D'Acunto, F, U Malmendier, J Ospina, and M Weber (2019), "Exposure to Daily Price Changes and Inflation Expectations", NBER working paper 26237.

Damstra, A., & Boukes, M. (2018). The Economy, the News, and the Public: A Longitudinal Study of the Impact of Economic News on Economic Evaluations and Expectations. Communication Research, 009365021775097. doi:10.1177/0093650217750971

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dräger, L. and Lamla, M.J. (2017), Imperfect Information and Consumer Inflation Expectations: Evidence from Microdata. Oxf Bull Econ Stat, 79: 933-968. https://doi.org/10.1111/obes.12189

Gabriele Galati & Peter Heemeijer & Richhild Moessner, 2011. "How do inflation expectations form? New insights from a high-frequency survey," DNB Working Papers 283, Netherlands Central Bank, Research Department.

Garcia, J. A. and Werner, S., Inflation News and Euro Area Inflation Expectations (July 2018). IMF Working Paper No. 18/167, Available at SSRN: https://ssrn.com/abstract=3333718

Goloshchapova I., Andreev M. Measuring inflation expectations ofthe Russian population with the help of machine learning. Voprosy Ekonomiki. 2017;(6):71-93. (In Russ.) https://doi.org/10.32609/0042-8736-2017-6-71-93

Hester, J. B., & Gibson, R. (2003). The Economy and Second-Level Agenda Setting: A Time-Series Analysis of Economic News and Public Opinion about the Economy. Journalism & Mass Communication Quarterly, 80(1), 73–90. doi:10.1177/107769900308000106

Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric P. Xing, Tie-Yan Liu, Wei-Ying Ma: LightLDA: Big Topic Models on Modest Compute Clusters. CoRR https://arxiv.org/abs/1412.1576 (2014)

Kilian, Lutz and Zhou, Xiaoqing, Oil Prices, Gasoline Prices and Inflation Expectations: A New Model and New Facts (November 15, 2020). CFS Working Paper, No. 645, 2020, Available at: http://dx.doi.org/10.2139/ssrn.3731014

King, Gary. 1986. How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. American Journal of Political Science 30:666-687.

Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, pp 320-332 (2015).

Larsen V.H., Thorsrud L.A. and Zhulanova J., News-driven inflation expectations and information rigidities, Journal of Monetary Economics, 2020, https://doi.org/10.1016/j.jmoneco.2020.03.004

Lines, Marji and Westerhoff, Frank, (2010), Inflation expectations and macroeconomic dynamics: The case of rational versus extrapolative expectations, Journal of Economic Dynamics and Control, 34, issue 2, p. 246-257.

Maćkowiak, B., & Wiederholt, M. (2009). Optimal Sticky Prices under Rational Inattention. The American Economic Review, 99(3), 769-803. http://www.jstor.org/stable/25592482

Mankiw, N. Gregory, Reis, Ricardo and Wolfers, Justin, (2004), Disagreement about Inflation Expectations, p. 209-270 in , NBER Macroeconomics Annual 2003, Volume 18, National Bureau of Economic Research, Inc.

Mazumder, Sandeep, 2021. "The reaction of inflation forecasts to news about the Fed," Economic Modelling, Elsevier, vol. 94(C), pages 256-264.

Nautz D., Pagenhardt L., Strohsal T., The (de-)anchoring of inflation expectations: New evidence from the euro area, The North American Journal of Economics and Finance, Volume 40, 2017, Pages 103-115, https://doi.org/10.1016/j.najef.2017.02.002

Pfajfar, D. And Santoro, E. (2013), News on Inflation and the Epidemiology of Inflation Expectations. Journal of Money, Credit and Banking, 45: 1045-1067. https://doi.org/10.1111/jmcb.12043

Rozovskaya, A., Roth, D. 2019. Grammar error correction in morphologically rich languages: The case of Russian. Transactions of the Association for Computational Linguistics, 7:1–17.

Sims, Christopher A., Inflation Expectations, Uncertainty and Monetary Policy (March 1, 2009). BIS Working Paper No. 275, http://dx.doi.org/10.2139/ssrn.1440243

Soroka, S., Fournier, P., Nir, L. Cross-national evidence of a negativity bias in psychophysiological reactions to news. Proceedings of the National Academy of Sciences Sep 2019, 116 (38) 18888-18892; DOI: 10.1073/pnas.1908369116

Taboada, M., J. Brooke, M. Tofiloski, K. Voll and M. Stede (2011) Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37 (2): 267-307.

Tobback, Ellen & Naudts, Hans & Daelemans, Walter & Junque de Fortuny, Enric & Martens, David. (2016). Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting. 10.1016/j.ijforecast.2016.08.006.

Woodford M., 2004. "Inflation targeting and optimal monetary policy," Review, Federal Reserve Bank of St. Louis, vol. 86(Jul), pages 15-42.

Zholud, O., Lepushynskyi, V., Nikolaychuk, S. (2019). The Effectiveness of the Monetary Transmission Mechanism in Ukraine since the Transition to Inflation Targeting. Visnyk of the National Bank of Ukraine, 247, 19-37.