TOPIC CLASSIFICATION OF ECONOMIC NEWSPAPER ARTICLES IN A HIGHLY INFLECTIONAL LANGUAGE – THE CASE OF SERBIA

Mirko Đukić

Available at www.nbs.rs

Economic Research and Statistics Department

NATIONAL BANK OF SERBIA

Belgrade, 12 Kralja Petra Street

Telephone: (+381 11) 3027 100

Belgrade, 17 Nemanjina Street

Telephone: (+381 11) 333 8000

www.nbs.rs

Topic classification of economic newspaper articles in a highly inflectional language – the case of Serbia

Mirko Đukić

Abstract: The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. The application of topic modelling in the Serbian language, using the LDA model, is hampered by the fact that Serbian is a highly inflectional language, where words have a large number of forms which the model recognises as words with a different meaning. In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicability in further economic analyses.

Keywords: textual analysis, topic modelling, Latent Dirichlet Allocation, LASSO model
[**JEL Code]:** C13, C55, E31, E37, E52

Non-technical summary

Newspaper articles are an important source of information about economic developments. They can cover a wide spectrum of economic topics – from the analysis of individual companies to the global economy. The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. A precondition for this type of analysis is that a large number of articles are classified based on the topic covered.

Topic modelling is a textual analysis technique that uncovers the patterns of common occurrence of certain words in a set of documents, interpreted as hidden topics in that set. In this paper, we used the Latent Dirichlet Allocation (LDA) method which is based on the assumption that each topic is a combination of different words, and each article a combination of different topics.

The application of topic modelling in the Serbian language, as well as indeed any other textual analysis method, is hampered by the fact that Serbian is a highly inflectional language, in which words have a large number of forms that the model recognises as words with a different meaning.

In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance (when applying the LDA model for classification into topics) than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicability in further economic analyses.

The analysis was applied to 25,248 articles in the economy section of the Politika daily in the period 2006–2023. The LDA model extracted 40 topics from these articles, and they were later named based on the most frequent words in each topic. In the majority of cases, the topic covered by the article was unambiguously determined, and only in two of the 40 cases were we unable to determine the content of the topic. Some topics in the sample cover broad areas (trade, corporates, economy…), while others are narrow and specific (fuel, petroleum, copper, steel, electricity…).

Lastly, the obtained series of the shares of topics over time were regressed to household inflation expectations using the LASSO model. The estimated model was well able to catch inflation cycles with a high determination coefficient. Of the 40 topics, the model kept 17 as relevant, some of which expectedly so, while for some there was no clear economic interpretation.

Contents:

1 Introduction

Newspaper articles are an important source of information about economic developments. They can cover a wide spectrum of economic topics – from the analysis of individual companies to the global economy. The frequency of certain topics in newspaper articles can be a good indicator of some economic developments. A precondition for this type of analysis is that a large number of articles are classified based on the topic covered.

Topic modelling is a textual analysis technique that uncovers the patterns of common occurrence of certain words in a set of documents, interpreted as hidden topics in that set. The first model for topic modelling was the Latent Semantic Analysis (Dearwester et al. 1990), which grouped documents based on words with a similar semantic structure.

In this paper, we used the Latent Dirichlet Allocation (LDA) model developed by Blei et al. (2003), which is based on the assumption that each topic is a combination of different words, and each article a combination of different topics. The LDA uncovers topics based on the patterns of the common occurrence of words (for each word, there is a certain likelihood that it would occur under a certain topic), and topics occur as the likelihood of occurrence in each document. This method can be used for analysis in different fields, such as law, humanities, classification of scientific papers, etc.

For the purpose of economic analysis, the LDA model can analyse different types of texts and we will list several examples here. Combining the LDA with a dictionary approach, Angelico et al. (2021) created a measure of inflation expectations based on Twitter (now the X platform) and established that it is highly correlated with the usual inflation expectation measures. Gonzales et al. (2018) created a measure of volatility of policies for ten Latin American countries based on the change in the topic of their presidents' speeches and concluded that greater changes in policies lead to smaller economic growth. Yono et al. (2020) measured news-based macroeconomic uncertainty for the purpose of investment decisions. An expanded topic model they proposed attributes a numerical value to each individual text, and the indices obtained in this manner correlate well with market volatility indices. The paper most similar to ours is by Larsen et al. (2021), who concluded that the media coverage of certain topics, some of which initially seemed unrelated to inflation, can play a significant role in predicting consumers' inflation expectations.

The application of topic modelling in the Serbian language, as indeed is the case with any textual analysis model, is hampered by the fact that Serbian is a highly inflectional language, in which words have many forms that the model recognises as words with a different meaning. In this paper, we tried to turn that aggravating circumstance into an advantage by reducing only the economic words to their base form. Thus, we attributed to them a greater relevance (when applying the LDA model for classification into topics) than to non-economic words, which remained in a large number of forms with a lower frequency of occurrence. As the topics classified in this manner were mostly based on economic expressions, it was expected that they would have a greater applicable value in further economic analyses.

The use of textual analysis in economic research is not novel in the National Bank of Serbia. Đukić (2022) created an indicator of inflationary pressures based on counting

expressions related to price changes in newspapers, which was determined to precede inflation movements. In this paper, we analysed newspaper articles from a different, topic standpoint.

Below we will first present a theoretical description of the LDA model. Afterwards, we will explain the way in which we prepared the articles for topic analysis, and then present the results of the LDA model application. Lastly, we will conclude the paper by assessing the impact of the movement of topics on inflation expectations by applying the LASSO model.

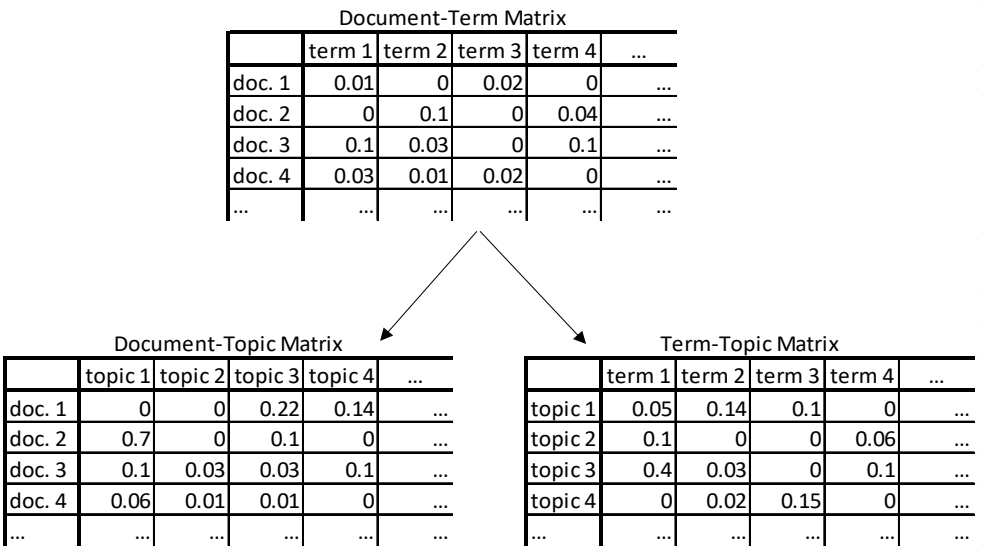
2 LDA model for classification of articles into topics

In this paper we use the Latent Dirichlet Allocation (LDA) model, developed by Blei et al. (2003). It is a widely used tool for detecting hidden topics within large sets of documents (newspaper articles in our analysis). The model assumes that documents are combinations of different topics and that topics are combinations of different words.

The LDA model treats documents as "bags of words", where their order and grammatical meaning play no role. Different forms of words with the same basic meaning are treated as entirely independent terms. The main aim of applying the LDA model is to reveal topics as sets of words that occur together, where some will have greater importance (probability) than others, and to present each document as a combination of different topics, where some topics will also be more important than others. The number of topics into which we wish to classify documents must be defined in advance.

The starting basis for the analysis is the Document Term Matrix (DTM), whose rows represent documents and columns all unique words in all documents. The element i, j in the matrix shows the share of the occurrence of word j in document i. As rows represent the shares of all words in the document, the sum of elements by rows is 1. Thus, the LDA from the DTM assesses the elements of the matrix of the share of topics in each document and the share of words in each topic (Figure 1).

Figure 1 **Matrices connecting words, topics and documents in the LDA model**



The LDA model algorithm is iterative. In the first iteration, one topic is randomly assigned to each word in each document, which is in the following iterations gradually adjusted based on specific criteria until optimal distribution is achieved.

The process is carried out for each "current" word by treating the topics assigned to other words as accurate. For instance, topic *t* in document *d* is assigned to current word *w*. The following probabilities are calculated:

- p_1: proportion of other words in document *d* that are assigned to the same topic *t*.
- p_2: proportion of documents assigned to topic *t* which stem from word *w*.

If a larger number of words from a given document belong to the same topic *t* (high p_1), it is more probable that the current word belongs to that topic. If the current word has a high probability of belonging to topic *t*, all documents containing *w* will be more strongly associated with topic *t* (high p_2). The higher the product of probabilities $p_1 \cdot p_2$, the more probable it is that the current word *w* belongs to topic *t*.

The LDA is run in a large number of iterations by assigning a new topic to word *w* based on the product of probabilities $p_1 \cdot p_2$ until equilibrium is achieved. The outcome are similar words grouped by topics, with the probabilities of each of them to stem from the given topics, and the distribution of the probabilities of shares of individual topics in each individual document. For a detailed mathematical overview of this method see Blei et al. (2003).

3 Text preparation for analysis

Before running the LDA model, texts should be abbreviated and adjusted, rendering them more conducive to faster machine processing. This includes reducing words to their base form, changing Latin letters specific for Serbian, eliminating frequent non-important words and punctuation, and turning uppercase letters into lowercase.

The Serbian language has a large number of forms of the same word, i.e. linguistically speaking it is a *highly inflectional* language, which generally aggravates any type of textual analysis. For instance, the algorithm recognises the words *nafta*, *nafte*, *naftu*, *naftni*, *naftna* (*oil* in different case forms) and similar words as different, although they all relate to the forms of the same base word. In our analysis, this would reduce the importance of the given term for classification of articles into topics, as its different forms will have a smaller frequency of occurrence.

One of the methods of reducing words to the base form is stemming, where suffixes are eliminated, keeping common beginning for all forms of the given word. Another method is lemmatisation, which reduces words to their roots as they would appear in the dictionary. The latter method is much more complex as it requires the existence of a dictionary with all forms of all words in a language. On the other hand, the stemmer is based on the rules whose number does not exceed several hundred, and can be developed much more easily.

Several automatic, programme stemmers have been developed for Serbian. The most famous was developed by Krovetz and Šipka (2008), according to whom the stemmer for Serbian must have eight times more rules for suffix stripping than for English, which is a good

illustration of how highly inflective Serbian language is. The above example contains several forms of the word *nafta* (*oil*), while *oil* is the only form of this word in English.

Instead of using the general stemmer that would treat all words in texts, we decided to reduce to the base form only those words that are relevant for our analysis, i.e. economic terms, whereby we favour their role in classifying texts into topics. Compared to non-economic terms, which remain in many different forms, economic terms reduced to the base form will have a larger share and be more relevant for classification. In this way, the aggravating circumstance of working with a language with a high number of word modifications was turned into an advantage for this type of analysis.

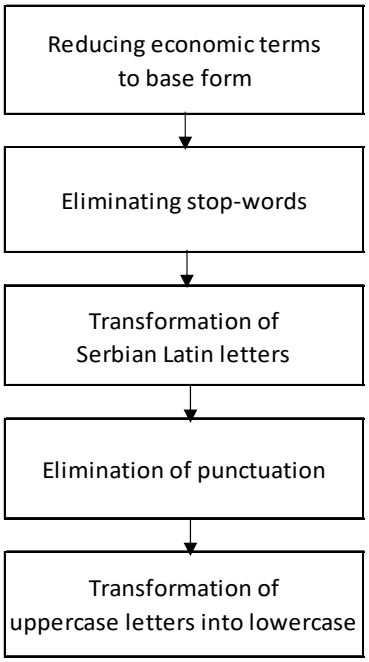
We interpreted the base form in the broadest possible way – unlike the standard stemmers, it includes all adjectival, verb and noun forms of a word. While stemmers for Serbian would treat the words *izvoz*, *izvozni* and *izvoziti* (*export* as a noun, an adjective, and a verb) as different base forms, in our analysis we reduced all forms of these words into the word *izvoz* (in order to attach greater importance to economic terms). Moreover, in some cases we grouped similar words with different prefixes as well, as is the case with *skupo (expensive)* and *poskupljenje (price increase)*.

Words were reduced to the base form in two ways. Where possible, words with the same base forms were identified based on the common beginning of all word forms, the so-called "stem", and were then replaced with the base form. For instance, the words *nafta, nafte, naftu, naftni, naftna*, etc. start with the letters *naft*, which is why we replaced them with the word *nafta*. Under the same principle, we replaced the words starting with *kamat* with the word *kamata* (*interest*), and the words starting with *inflaci* or *inflator* with *inflacija* (*inflation*) etc. This rule could be applied only if the stem for different forms of the base term is not at the same time the stem of another non-similar term.

Contrary to this, for instance, the stem for different forms of the word *cena* (*cenu, cene, cenama, cenovni* …) (*price* in different forms) – *cen* – is at the same time the stem for forms of some other non-similar terms (*centar*, *ceniti*, *cenzura*) (*centre, assess, censorship*). Applying the above rule in this case would unjustifiably transform the other words into the word *cena*, thus changing their meaning. Therefore, in this and similar cases, it was necessary to explicitly define all concrete forms of words that we wish to transform into the base form.

The first method (recognition based on the beginning of the word) is obviously simpler to apply, so we used it in certain cases where there are several terms for the same word beginnings. For example, words beginning with *bank* can be forms of the words *banka* and *bankina* (*bank* and *road barrier fence*), but as the latter term is very rare, if not non-existent in economic texts, there was no reason for concern that its unjustified transformation could disrupt the further analysis.

Figure 2 **Text preparation for the LDA model**



There was a certain degree of arbitrariness in choosing economic words to be reduced to the base form. Certain words, such as *saobraćaj, država, vlada* or *grejanje* (*traffic, state, government* or *heating*), strictly speaking, are not economic, but as we analysed economic texts only, we assumed that these words have an economic connotation, so we included them in the transformation list. In this list, we also included the names of institutions that are often mentioned in newspaper articles (IMF, EPS, Telekom).

While preparing the text for further analysis, we eliminated the stop-words, which are frequent in Serbian but are not essential, such as: *i*, *ili*, *ali*, *koji*, *to*, *od*, *gde* (*and, or, but, who/which, that, from, where*). Due to their frequency, the algorithm may recognise them as crucial in classifying texts into topics, which is not useful for economic or any other type of analysis.

As our text analysis program does not recognise Serbian Latin letters (*č, ć, ž, đ, š*), we transformed them into forms suitable for processing. We combined the base letters (without diacritics) with the letters not used in Serbian (*x*, *y*) as follows: *č→cx, ć→cy, ž→zx, đ→dx, š→sx.*

Punctuation was also eliminated, and uppercase letters were converted into lowercase letters. All these interventions – the elimination of frequent words and punctuation, and reducing frequent economic terms to the common base form make the text more conducive to our analysis, but also reduce the time for text processing, which is not insignificant given that the process may take several hours.

The above transformation rules can be illustrated with the following text:

Vlada ograničila cene osnovnih životnih namirnica
Vlada Srbije na današnjoj sednici donela je odluku da ograniči visinu cena osnovnih životnih namirnica: šećer, brašno tip T-400, suncokretovo ulje, svinjsko meso i dugotrajno mleko sa 2,8 procenata mlečne masti, tako da one ne prelaze nivo cena na dan 15. novembar 2021.
Ograničenje cena utvrđeno je kako bi se otklonile štetne posledice i sprečili poremećaji na tržištu i neće se odnosit i na snižene cene, kao što su rasprodaje, sezonska sniženja ili akcijske prodaje, ukoliko su bila na snazi 15. novem bra, već na redovne, odnosno cene pre sniženja, saopšteno je iz vlade, preneo je Tanjug.
Odlukom, koja će biti primenjivana u trajanju od 60 dana, predviđeno je da proizvođači ove proizvode ne smeju isporučivati u količinama manjim od prosečnih u poslednjih 12 meseci.
Za kršenje navedenih odredaba, predviđene su i novčane kazne u iznosu od 100.000 do dva miliona dinara, kao i zabrana vršenja delatnosti u trajanju od šest meseci do jedne godine.

and its transformed version:

vlada ogranicxila cena osnovnih zxivotnih namirnica vlada srbije danasxnjoj sednici donela odluku ogranicxi visinu cena osnovnih zxivotnih namirnica sxecyer brasxno tip t400 suncokretovo ulje svinjsko meso dugotrajno mleko 28 procenata mlecxne masti one ne prelaze nivo cena dan 15 novembar 2021 ogranicxenje cena utvrdjeno bi otklonile sxtetne posledice sprecxili poremecyaji trzxisxte necye odnositi snizxene cena sx su rasprodaje sezonska snizxenja akcijske prodaje ukoliko su bila snazi 15 novembra redovne odnosno cena snizxenja saopsxteno vlada preneo tanjug odlukom primenjivana trajanju 60 dana predvidjeno proizvod proizvod ne smeju isporucxivati kolicxinama manjim prosecxnih poslednjih 12 meseci krsxenje navedenih odredaba predvidjene su novac kazne iznosu 100000 dva milion dinar zabrana vrsxenja delatnosti trajanju sxest meseci jedne godine

Though for a Serbian reader the first article is certainly easier to understand, the transformed article is more conducive to the LDA method processing for the reasons we have outlined above. The transformed article is also shorter (by around 30%), which is important when processing large quantities of data.

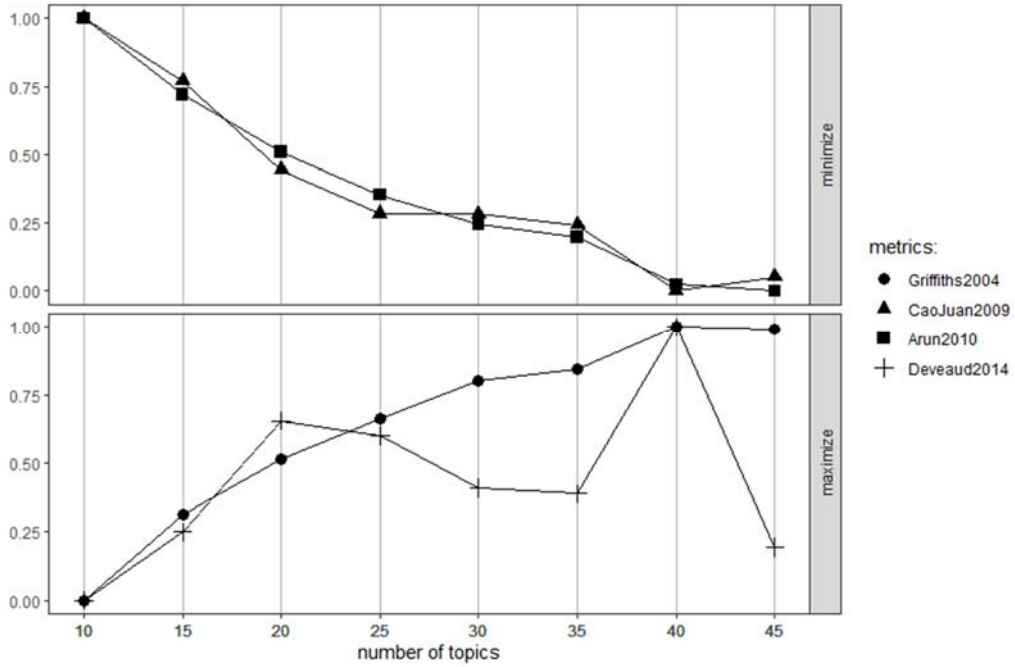
The choice of economic words for transformation and the elimination of frequent superfluous words was not entirely pre-defined, but was supplemented in several steps in parallel with the application of the LDA model. In the first step, the model was run with the initially shorter list of transformed and eliminated words, based on which the texts were classified into topics. Then, by analysing the most frequent words in topics, we extended the list of stop-words and economic terms to be reduced to their base forms. We ran the LDA model again with the new extended lists for word elimination/correction. We repeated the entire process several times, filtering the dominant words in topics. The final list of words for transformation is contained in Table A1 of the Appendix.

4 Text classification into topics with the LDA model

The starting point for the topic classification of texts with the LDA model is the Document-Term Matrix that contains the number of term occurrences per document. In our analysis, the Matrix contains 25.248 · 193.142 elements, where the first dimension represents the number of topics, and the second the number of unique words.

When running the LDA model, it is necessary to specify the number of topics into which we wish to classify the articles from the sample. To determine the optimal number of topics for classification, we used the four criteria developed by Griffiths, et al. (2004), Cao et al. (2009), Arun et al. (2020), and Deveaud (2014). According to three of these four criteria, the optimal number of topics for our sample is 40 (Chart 1).

Chart 1 **Criteria for the selection of optimal number of topics for the LDA model**



As already stated, we estimated the model in several steps, supplementing in each step the list of words for transformation or elimination, based on the analysis of the most frequent words in topics.

Chart A1 in the Appendix shows the final division into topics and the most frequent words. Technically, those are words with the highest estimated β coefficient, which measures the probability of a concrete word stemming from a concrete topic. Note that the model did not specify topics titles on its own (as it only marked them with numbers), i.e. we have done this based on the dominant words per topic.

In most cases, it was easy to determine the topic based on the most frequent words. For instance, the topic where words such as *zaposlenost, radnik, rad, posao* (*employment, employee, labour, work*) were dominant is obviously related to the labour market; the topic with the words *proizvod, poljoprivreda, tržište, tona, pšenice, voća* (*product, agriculture, market, tonne, wheat, fruit*) concerns the agricultural commodities market; the topic with the words *energetika, EPS, elektrika, struja, uglja* (*energy, EPS (Electric Power Industry of Serbia) electrics, electricity, coal*) pertains to electrical energy, etc. Particularly relevant for the central bank and monetary policy are topics related to the banking sector (*kredit, banka, kamata, dinar…*) (*loan, bank, interest, dinar…*) and inflation (*inflacija, odsto, rast, cena…*) (*inflation, percent, growth, price…*). The titles of two topics could not be determined given

that non-economic and unrelated words were dominant in them. We named these topics *UNTITLED*[1]. In some cases, we had several topics dealing with the same area (*ECONOMY, AGRICULTURE, GOVERNMENT, PENSIONS, TOURISM, WORKERS*).

While each topic is a combination of different words, each article is a combination of different topics. Coefficient γ represents the share of individual topics in each document. Articles can be dominantly represented with one topic or as a combination of several topics.

For instance, *GAS* ($\gamma = 0{,}98$) is the dominant topic in the following short article:

The Balkan Stream gas pipeline connected at the border with Hungary *(4 June 2021)*

Horgoš – The public enterprise Srbijagas and the Hungarian company FGSZ connected today the Balkan Stream gas pipeline at the border with Hungary, through which gas will in future be delivered from Turkey through Bulgaria and Serbia to Central Europe.

Dušan Bajatović, general director of Srbijagas, said that in this way, an end was put to the long-term great work by which our country finally solved the issue of gas supply from another direction, not only through Ukraine.

"Nobody in Serbia will be freezing anymore, and the price of gas for households will not be changed in autumn. The increase in the price of oil, which is followed by the price of gas, will spill over to Srbijagas", Bajatović explained.

On the other hand, the article below is a combination of several topics, where none has a majority share (*FUEL 0.42; STOCK EXCHANGE 0.33; OIL 0.24*):

European stock indices mostly down, gold price at a historic high *(4 December 2023)*

NEW YORK/FRANKFURT/MOSCOW – European stock indices are mostly down at the beginning of the week, while the price of gold is at an all-time high today. The DAX index of the Frankfurt Stock Exchange rose to 16,421.45 points at 10:00 a.m. today, while the French CAC 40 fell to 7,333.14 points, as well as the London FTSE 100 – to 7,508.57 points and the Moscow MOEX – to 3,113.05. The value of the American stock index Dow Jones rose to 36,245.50 before today's opening of the stock market in America, as well as the value of the S&P 500 index – to 4,594.63 points and the value of the Nasdaq index – to 14,305.03.

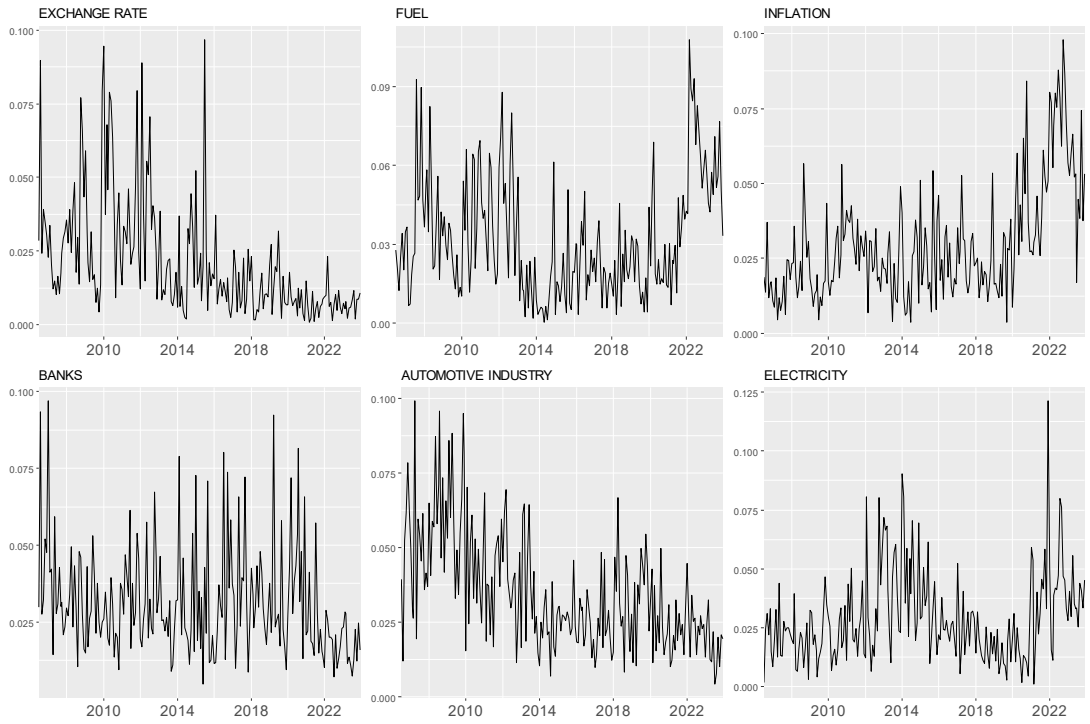
According to stock exchange data, the price of crude oil fell to $72,874 per barrel, as well as the price of Brent oil – to $77,615. European gas futures were sold today at the opening of the TTF stock exchange at a price of EUR 42,750 per megawatt-hour.

The price of gold reached an all-time high of $2,110.8 an ounce early this morning, and by 10:00 a.m. it fell slightly to $2,069.16 an ounce (an ounce equals 28.35 grams). Wheat also rose to $5.7936 a bushel (a bushel equals 27,216 kg). The value of the euro against the dollar was 1.08692, which is approximately the same as on Friday, reports Tanjug.

The time series of the share of topics display exceptionally high volatility (Chart 2). Although the topic classification sample contains a large number of articles (over 25,000) – 120 articles divided into 40 themes, on average, per month, implies a small monthly number of articles per topic.

[1] The titles of topics are written in uppercase letters.

Chart 2 **Frequency of selected topics in newspaper articles** (monthly averages of the γ coefficient)



Despite high volatility, the frequency of topics well reflects developments in the economy: the interest in the exchange rate grew during the periods of its high volatility, and subsided as it stabilised in the past several years; the automobile industry was a frequent topic at the time when Fiat arrived in Kragujevac; fuel and electricity are frequent topics during the periods of their price hikes; the interest in inflation is the highest during inflation cycles.

5 Estimate of the link between inflation expectations and topics

Households form inflation expectations based on the information they obtain from various sources, including the media. A feedback effect is also possible – media reporting can reflect inflation expectations. In any case, writing about specific topics is a potential indicator of inflation expectations.

In this paper, we estimated the link between inflation expectations and topics by the LASSO regression (Tibshirani, R. (1996)), which is suitable in cases when the number of variables in the regression is large, of which some are irrelevant. The key model characteristic is adding the penalty term $\lambda\sum_{j=1}^{p}|\beta_j|$ in the function of the ordinary least squares objective function, which encourages the model to reduce to zero the β_j coefficients of less important variables:

$$min: \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

The λ parameter determines the degree to which we wish to penalize the coefficient retention in the model (higher λ – less non-zero coefficients). In our analysis, we used the λ parameter, which results in the lowest out-of-sample mean forecasting error (Table A3 in the Appendix).

In our case, we regressed inflation expectations π_t^{exp} to 40 variables of the movement of the share of individual topics T_t^i in the period January 2009 – December 2023. To avoid the potential problem of endogeneity (simultaneous impact), we took one-month arrears as independent variables.

$$\pi_t^{exp} = \alpha + \sum_{i=1}^{40} \beta_i \, T_{t-1}^i + \lambda \sum_{i=1}^{40} |\beta_i|$$

The estimated model retained 17 variables as relevant and reduced the coefficients of the other ones to zero (Table 1), with the determination coefficient of $R^2 = 0.67$ (Table A3).

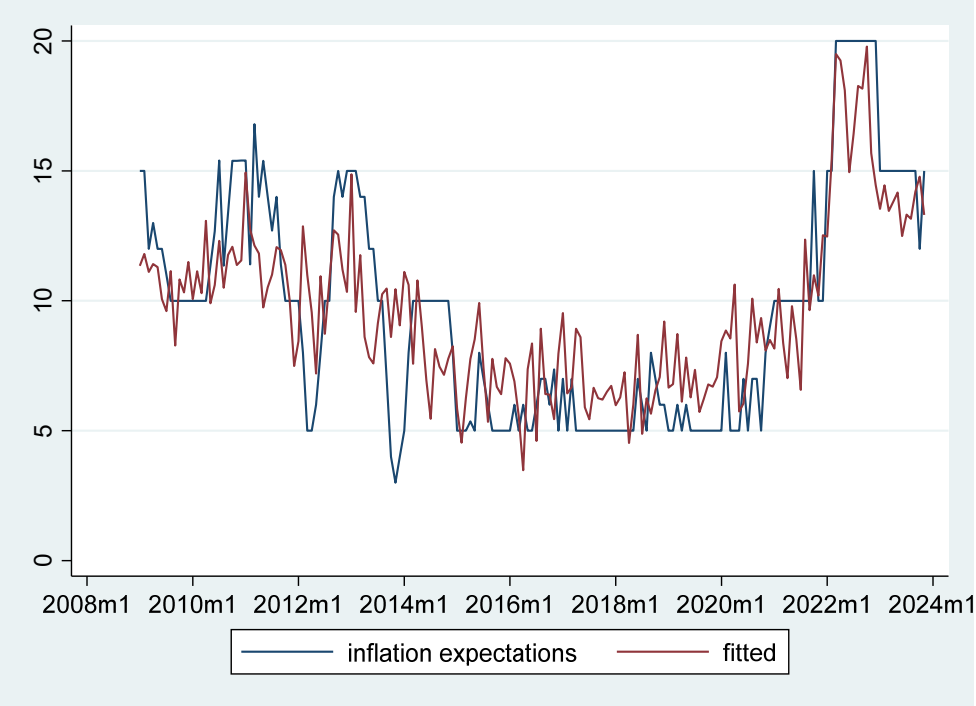
The variables with a positive coefficient include some expected topics: *INFLATION, FUEL, OIL, ELECTRICITY, WAGES-CONSUMPTION.* In these topics on energy products, the movement of their prices is often mentioned. In case of individual variables, however, there is no clear economic interpretation of the coefficient (e.g. a negative coefficient in topics *COPPER, EMPLOYMENT, ROAD CONSTRUCTION…*).

Table 1 **Estimated coefficients for topics from the LASSO model for inflation expectations**

Topic	Coefficient	Topic	Coefficient
1. ECONOMY	0	*21. AUTOMOTIVE IND.*	9.1
2. FX FLOWS	0	*22. LAW-ECONOMY*	-16.6
3. COPPER	-35.1	*23. INTERNATIONAL*	0
4. TOURISM	0	*24. PENSIONS*	0
5. TELEPHONY	0	*25. OIL*	70.3
6. FUEL	36.6	*26. WAGES-CONSUMPTION*	83.0
7. WORKERS	0	*27. EMPLOYMENT*	-30.6
8. EXCHANGE RATE	0	*28. AGGRICLUTURE*	0
9. AGGRICLUTURE	0	*29. LOCAL*	27.2
10. TOURISM	0	*30. ECONOMY*	0
11. TAX	0	*31. TRADE*	0
12. BUDGET	0	*32. AIR TRANSPORT*	-12.0
13. INFRASTRUCTURE	0	*33. GAS*	0
14. GOVERNMENT	0	*34. STOCK EXCHANGE*	0
15. ELECTRICITY	33.3	*35. FLATS*	50.1
16. INFLATION	40.0	*36. GOVERNMENT*	60.0
17. BANKS	0	*37. FIRMS-BANKS*	43.7
18. PENSION-INSURANCE	-18.5	*38. ROAD CONSTRUCTION*	-10.3
19. WORKERS	0	*39. UNTITLED*	0
20. UNTITLED	0	*40. STEEL*	-23.6

Chart 3 shows that the model is good at predicting the inflation expectations cycles, with the estimated series being much more volatile than expectations. This is a natural consequence of high volatility in the movement of individual topics, on the one hand, and relatively stable inflation expectations from the survey, on the other.

Chart 3 **Inflation expectations and their fitted values based on the LASSO model with the movement of topics (in %)**



6 Conclusion

This paper shows the topic classification of 25,248 articles from the economic section of the Politika daily in the 2006–2023 period, by applying the LDA model.

The specificity of our approach lies in the fact that during the text preparation stage, we selectively chose only economic words to be reduced to their base form. By doing so, we provided them with an advantage in topic classification over non-economic words, which remained in a large number of forms with lower frequencies of occurrence. This advantage is particularly pronounced in languages with high inflection, such as Serbian, where words have numerous forms. The selection of words for transformation was not entirely pre-defined, but was supplemented in several steps in parallel with the application of the LDA model, based on the analysis of the most frequent words in topics.

The LDA model classified the modified newspaper articles into economic topics in a satisfactory way. In most cases, the content of the topic was unambiguous, while the content of two topics, with dominantly non-economic terms, was not evident. Some topics from the sample have a relatively large scope (*ECONOMY, INTERNATIONAL*…), while some are specific (*TELEKOM, OIL, COPPER, STEEL, ELECTRICITY*).

The monthly movement of topics displays high volatility, which can be explained by the insufficiently large sample for such a large number of topics. By using the LASSO model, we regressed these series to household inflation expectations. The inflation cycles were well

captured by the estimated model with a high coefficient of determination. Of the 40 topics, the model kept 17 as relevant, some of them expectedly so, while for others there is no clear economic interpretation. To have more reliable econometric analyses, it is probably desirable to expand the sample of documents, which will be one of our objectives going forward.

Appendix

Table A1 **Replacement of the word with its base form based on the beginning of the word**

Beginning of word	Replacement
inflaci	*inflacija*
inflator	*inflacija*
deflaci	*deflacija*
deflator	*deflacija*
poskup	*poskupljenje*
skuplj	*poskupljenje*
pojeft	*pojeftinjenje*
jeft	*pojeftinjenje*
kurs	*kurs*
dinar	*dinar*
dolar	*dolar*
deviz	*deviza*
novc	*novac*
kredit	*kredit*
kamat	*kamata*
bank	*banka*
banc	*banka*
banaka	*banka*
finansi	*finansije*
monetar	*monetarna*
naft	*nafta*
barel	*barel*
goriv	*gorivo*
benzin	*benzin*
dizel	*dizel*
derivat	*derivat*
energ	*energija*
elektri	*elektrika*
struj	*struja*
grejanj	*grejanje*
proizvod	*proizvod*
preduzec	*preduzeće*
kompanij	*kompanija*
fabri	*fabrika*
penzi	*penzija*
zarade	*zarada*
zarada	*zarada*

Beginning of word	Replacement
zaposlen	*zaposlenost*
zapošlj	*zaposlenost*
radn	*radnik*
porez	*porez*
pores	*porez*
budžet	*budžet*
akciz	*akciza*
drzav	*drzava*
guverner	*guverner*
minist	*ministar*
ekonom	*ekonomija*
privred	*privreda*
makroekon	*makroekonomija*
uvoz	*uvoz*
uvezen	*uvoz*
izvoz	*izvoz*
izvezen	*izvoz*
trži	*tržište*
bdp	*bdp*
trgov	*trgovina*
kriz	*kriza*
recesi	*recesija*
investi	*investicija*
poljoprivr	*poljoprivreda*
potroš	*potrošnja*
milion	*milion*
milijard	*milijarda*
infrastrukt	*infrastruktura*
turis	*turizam*
turiz	*turizam*
transakci	*transakcija*
osiguran	*osiguranje*
eps	*eps*
mmf	*mmf*
telekom	*telekom*
berz	*berza*

Table A2 **Replacement of the word with its base form based on the entire word**

Word	Replacement
cene	*cena*
cenu	*cena*
cenama	*cena*
cenovni	*cena*
cenovna	*cena*
cenovne	*cena*
evra	*evro*
evri	*evro*
evru	*evro*
evrima	*evro*
plate	*plata*
platu	*plata*
plati	*plata*
platama	*plata*
gasa	*gas*
gasu	*gas*
gasni	*gas*
gasna	*gas*
gasovod	*gas*
gasovoda	*gas*
gasovodu	*gas*
vlade	*vlada*
vladu	*vlada*
vladi	*vlada*
firmi	*firma*
firme	*firma*
firmu	*firma*
firmama	*firma*
rada	*rad*
rade	*rad*
radu	*rad*

Figure A1-a **The most frequent words in topics** (β coefficients)

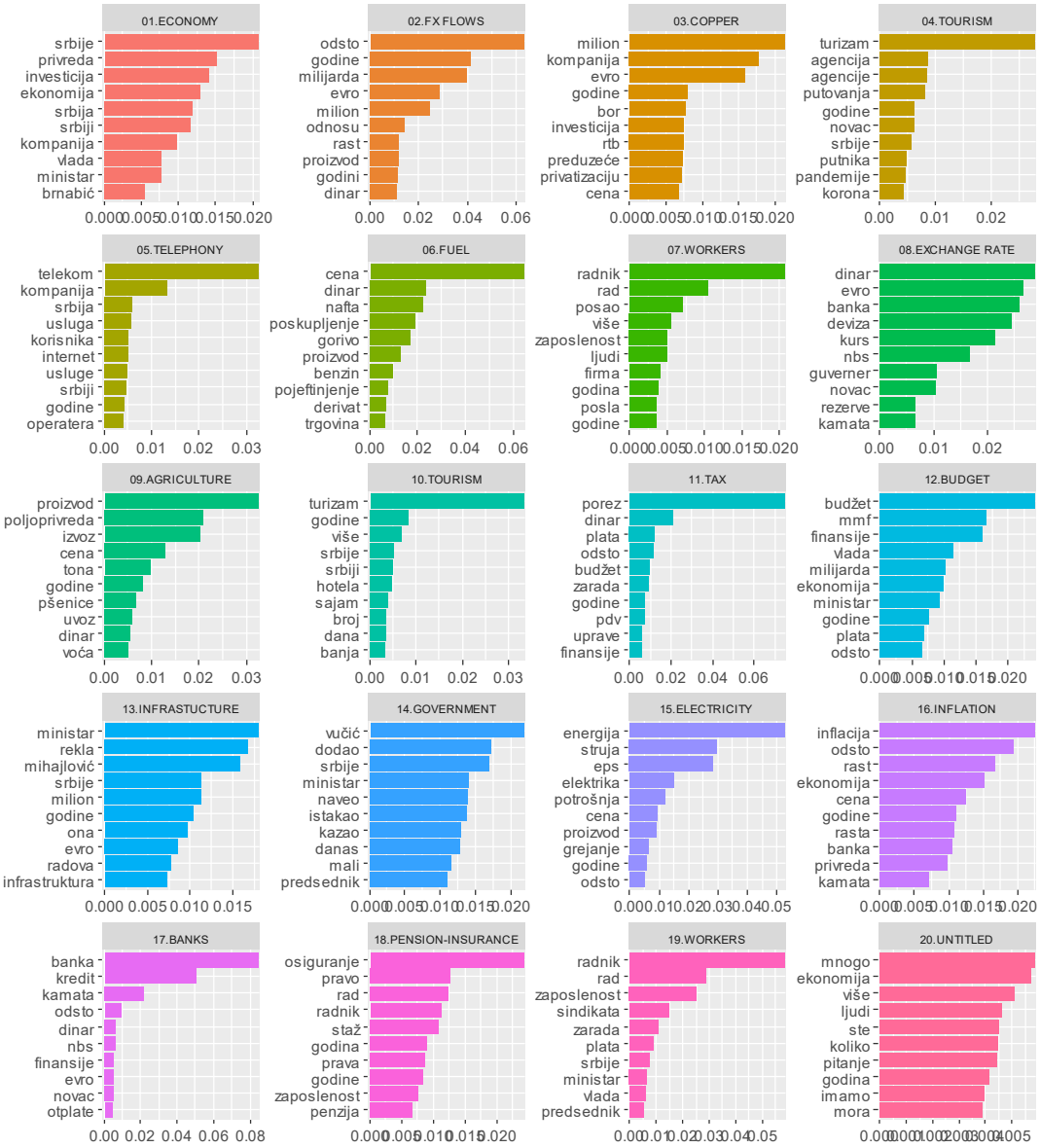


Figure A1-b **The most frequent words in topics** (*β* coefficients)

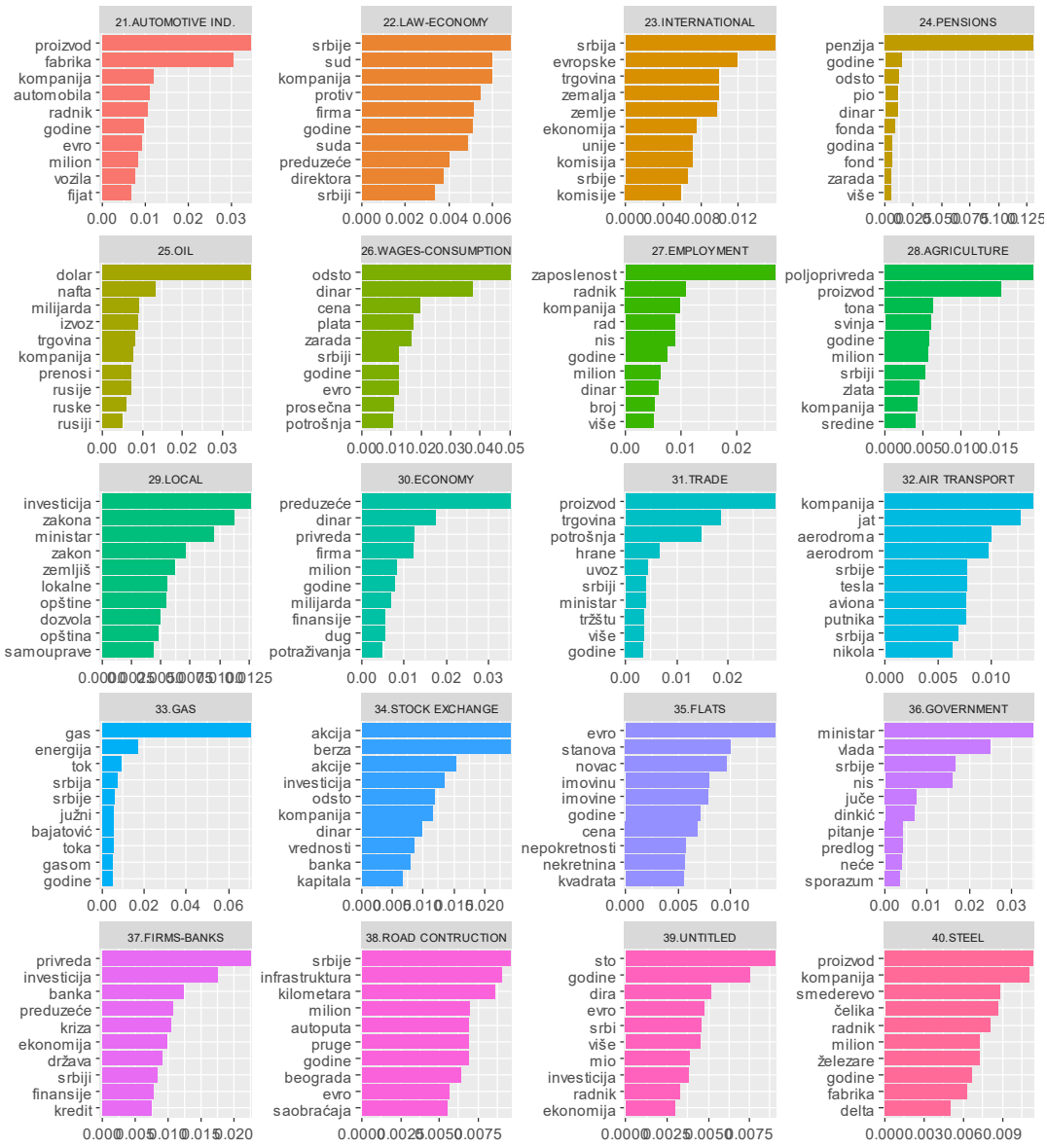


Table A3 **Results of LASSO regression of inflation expectations to topics from newspaper articles**

Lasso linear model — No. of obs = 180
No. of covariates = 40
Selection: Cross-validation — No. of CV folds = 10

ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	2.516326	0	0.0009	19.49777
29	lambda before	.1859746	17	0.5599	8.587741
* 30	selected lambda	.1694532	17	0.5601	8.583828
31	lambda after	.1543994	17	0.5596	8.594913
33	last lambda	.1281851	19	0.5568	8.649741

* lambda selected by cross-validation.

	active
t3	-35.07961
t6	36.57279
t15	33.34805
t16	40.01757
t18	-18.47686
t21	9.089888
t22	-16.58271
t25	70.2822
t26	82.99576
t27	-30.5818
t29	27.18242
t32	-11.97298
t35	50.08295
t36	59.94977
t37	43.70352
t38	-10.29667
t40	-23.55415
_cons	1.135015

Penalized coefficients

MSE	R-squared	Obs
6.447942	0.6696	180

Literature

Angelico C., Marcucci J., Miccoli M. & Quarta F., (2021). "Can we measure inflation expectations using Twitter?," Temi di discussione (Economic working papers) 1318, Bank of Italy.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). "On finding the natural number of topics with latent Dirichlet allocation: Some observations" In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 391–402). Springer.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). "Latent Dirichlet Allocation". Journal of Machine Learning research, 3(Jan), 993–1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). "A density-based method for adaptive LDA model selection". Neurocomputing, 72(7–9), 1775–1781.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). "Accurate and effective latent concept modeling for ad hoc information retrieval". Information Retrieval Journal, 17(2), 175–198.

Đukić, M. (2022). "Assessment of inflationary pressures using newspaper text analysis". Working Papers Bulletin. National bank of Serbia, September 2022. 41–66.

Griffiths, T. L., & Steyvers, M. (2004). "Finding scientific topics". Proceedings of the National Academy of Sciences of the United States of America, 101 Suppl 1(suppl 1), 5228–5235.

Kešelj V. and Šipka D. (2008) "A Suffix Subsumption-based Approach to Building Stemmers and Lemmatizers for Highly Inflectional Languages with Sparse Resources". INFOTHECA, Journal of Informatics and Librarianship, vol. IX, no. 1–2, pp. 23a–33a, 21–31 May 2008.

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). "News-driven inflation expectations and information rigidities". Journal of Monetary Economics, 117, 507–520.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990), "Indexing by latent semantic analysis", Journal of the American Society for Information Science.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso", Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.