ТЕМАТСКА КЛАСИФИКАЦИЈА ЕКОНОМСКИХ НОВИНСКИХ ЧЛАНАКА НА ЈЕЗИКУ С ВИСОКОМ ФЛЕКСИЈОМ – ПРИМЕР СРБИЈЕ

Мирко Ђукић

Доступно на www.nbs.rs

За ставове изнете у радовима у оквиру ове серије одговоран је аутор и ставови не представљају нужно званичан став Народне банке Србије.

Сектор за економска истраживања и статистику

НАРОДНА БАНКА СРБИЈЕ

Београд, Краља Петра 12

Тел.: (+381 11) 3027 100

Београд, Немањина 17

Тел.: (+381 11) 333 8000

www.nbs.rs

Тематска класификација економских новинских чланака на језику с високом флексијом – пример Србије

Мирко Ђукић

Апстракт: Учесталост појављивања појединих тема у новинским чланцима може бити добар индикатор одређених економских кретања. Примена тематског моделирања на српском језику, моделом *LDA*, отежана је чињеницом да је у питању језик с високом флексијом, у коме речи имају велики број облика, које модел препознаје као речи с различитим значењима. У овом раду настојали смо да ту отежавајућу околност претворимо у предност, тако што смо само економске речи свели на основни облик. Тиме смо им дали већи значај у односу на неекономске речи, које су остале у великом броју облика с мањим фреквенцијама појављивања. Како су теме класификоване на тај начин у већој мери базиране на економским изразима, за очекивање је да имају већу употребну вредност у даљим економским анализама.

Кључне речи: текстуална анализа, тематско моделирање, *Latent Dirichlet Allocation*, модел *LASSO*
[JEL Code]: C13, C55, E31, E37, E52

Нетехнички резиме

Новински чланци су значајан извор информација о економским кретањима. Они могу покривати широк спектар економских тема – од анализе појединачних предузећа до светске економије. Учесталост појављивања појединих тема може бити добар индикатор одређених економских кретања. Предуслов за ту врсту анализе јесте да се велики број новинских чланака класификује на основу тема којима се баве.

Тематско моделирање је техника текстуалне анализе која открива обрасце заједничког појављивања одређених речи у скупу докумената, који се интерпретирају као скривене теме у том скупу. У овом раду користимо модел *Latent Dirichlet Allocation* (*LDA*), који се базира на претпоставци да свака тема представља комбинацију различитих речи, а сваки чланак комбинацију различитих тема.

Примена тематског моделирања на српском језику, као уосталом и било ког метода текстуалне анализе, отежана је чињеницом да је у питању језик с високом флексијом, у коме речи имају велики број облика које модел препознаје као речи с различитим значењима.

У овом раду настојали смо да ту отежавајућу околност претворимо у предност, тако што смо само економске речи свели на основни облик. Тиме смо им дали већи значај (при примени модела *LDA* за класификацију на теме) у односу на неекономске речи које су остале у великом броју облика с мањим фреквенцијама појављивања. Како су теме класификоване на тај начин у већој мери базиране на економским изразима, за очекивање је да имају већу употребну вредност у даљим економским анализама.

Анализу смо применили на 25.248 чланака из економске рубрике дневног листа *Политика* у периоду 2006–2023. Модел *LDA* је из чланака екстраховао 40 тема, које смо касније именовали на основу најчешћих речи у њима. У већини случајева било је недвосмислено чиме се одређена тема бави, док само у два случаја од 40 нисмо успели да одредимо садржај теме. Неке теме из узорка покривају широке области (трговина, предузећа, економија…), док су поједине уско специфичне (гориво, нафта, бакар, челик, струја…).

На крају смо добијене серије учешћа тема кроз време, употребом модела *LASSO*, регресирали на инфлациона очекивања становништва. Оцењени модел је добро ухватио инфлационе циклусе, с високим коефицијентом детерминације. Од 40 тема, модел је задржао 17 као релевантне, од којих је за неке то било очекивано, док за неке не постоји јасна економска интерпретација.

Садржај:

1. Увод

Новински чланци су значајан извор информација о економским кретањима. Они могу покривати широк спектар економских тема – од анализе појединачних предузећа до светске економије. Учесталост појављивања појединих тема у новинама стога може послужити као индикатор одређених економских кретања. Предуслов за ту врсту анализе јесте да се новински чланци класификују на основу тема којима се баве.

Тематско моделирање је техника текстуалне анализе која открива обрасце заједничког појављивања одређених речи у скупу докумената, који се интерпретирају као скривене теме у том скупу. Први модел за тематско моделирање био је *Latent Semantic Analysis* (*Dearwester et.al*. 1990), који је документа груписао на основу речи са сличном семантичком структуром.

У овом раду користимо модел *Latent Dirichlet Allocation* (*LDA*), који су развили *Blei et. al*. (2003), а који се заснива на претпоставци да свака тема представља комбинацију различитих речи, а сваки чланак комбинацију различитих тема. *LDA* открива теме на основу образаца заједничког појављивања речи (свака има одређену вероватноћу појављивања у теми), а теме се појављују као вероватноће појављивања у сваком документу. Овај метод може се користити за анализе у разним областима, као што су право, друштвене науке, класификација научних радова итд.

У сврху економских анализа, моделом *LDA* могу се анализирати разне врсте текстова. Овде наводимо неколико примера. *Angelico et. al*. (2021), комбинујући *LDA* с речничким приступом, креирали су меру инфлационих очекивања на основу Твитера (данас платформа Икс) и установили њену високу корелисаност са уобичајеним мерама инфлационих очекивања. *Gonzales et. al*. (2018) израдили су меру волатилности политика за десет латиноамеричких држава на основу промене тема говора њихових председника и закључили да веће промене политика воде мањем привредном расту. *Yono et. al*. (2020) мерили су макроекономску неизвесност на основу вести ради доношења инвестиционих одлука. Проширени тематски модел који су предложили приписује нумеричку вредност сваком појединачном тексту, а овако добијени индекси добро корелирају са индексима тржишне волатилности. Најсличнији нашем раду је рад *Larsen et. al*. (2021), који су утврдили да медијска покривеност појединих тема, од којих су неке на први поглед неповезане са инфлацијом, може у знатној мери предвидети инфлациона очекивања потрошача.

Примена тематског моделирања на српском језику, као и уосталом било ког метода текстуалне анализе, отежана је чињеницом да је у питању језик с високом флексијом, у коме речи имају велики број облика, које модел препознаје као речи с различитим значењима. У овом раду настојали смо да ту отежавајућу околност претворимо у предност тако што смо само економске речи сводили на основни облик. Тиме смо им дали већи значај (при примени модела *LDA* за класификацију на теме) него неекономским речима, које су остале у великом броју облика с мањим фреквенцијама појављивања. Како су теме класификоване на тај начин у већој мери базиране на економским изразима, може се очекивати да ће имати већу употребну вредност у даљим економским анализама.

Употреба текстуалне анализе у економским истраживањима није нова у Народној банци Србије. У раду *Đukić* (2022) креиран је индикатор инфлаторних притисака заснован на бројању израза у вези с променама цена у новинама, за који је утврђено да претходи кретању инфлације. У овом раду новинске чланке анализирамо из другог, тематског угла.

У наставку рада најпре дајемо теоријски опис модела *LDA*. Након тога објашњавамо начин на који смо текстове припремили за тематску анализу, затим приказујемо резултате примене модела *LDA*, да бисмо рад завршили оценом утицаја кретања тема на инфлациона очекивања применом модела *LASSO*.

2. Модел *LDA* за класификацију текстова на теме

Модел *Latent Dirithlet Allocation* (*LDA*), који користимо у овом раду, а који су развили *Blei et. al*. (2003), широко је коришћен алат за откривање скривених тема унутар великих скупова докумената (новинских чланака у нашој анализи). У сржи *LDA* лежи претпоставка да су документи комбинације различитих тема, а теме комбинације различитих речи.

Модел *LDA* третира документе као „вреће речи", где њихов редослед и граматичко значење не играју никакву улогу. Различити облици речи са истим основним значењем третирају се као потпуно независни појмови. Главни циљ примене модела *LDA* јесте да открије теме као скупове речи који се заједно појављују, где ће неке имати већи значај (вероватноће) него друге, те да сваки документ представи као комбинацију различитих тема, где ће такође неке бити значајније од других. Моделу је неопходно унапред задати број тема за класификацију докумената.

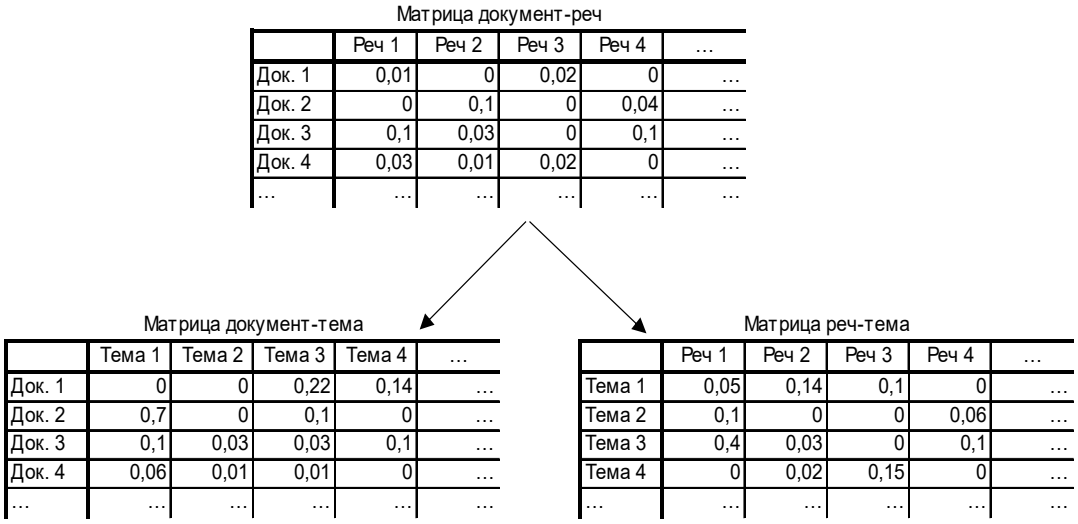
Почетну основу за анализу чини *матрица документ-реч* (*Document Term Matrix – DTM*), чији редови представљају документа, а колоне све јединствене речи у свим документима. Елеменат (*i, j*) у тој матрици показује учешће појављивања речи *j* у документу *i*. Како редови приказују учешћа свих речи у документу, збир елемената по редовима је 1. У складу с претходно реченим, *LDA* из *DTM* оцењује елементе матрица учешћа тема у сваком документу и учешћа речи у свакој теми (Слика 1).

Алгоритам модела *LDA* је итеративан. У првој итерацији, свакој речи у сваком документу насумично се придружује једна тема, што се у наредним итерацијама постепено исправља и прилагођава на основу одређених критеријума све док се не постигне оптимална расподела.

Тај поступак се спроводи за сваку појединачну, „тренутну" реч, тако што се теме приписане осталим речима третирају као исправне. Нека је тренутној речи *w* приписана тема *t* у документу *d*. Рачунају се следеће вероватноће:

- p_1: пропорција осталих речи у документу *d* који су приписани истој теми *t*.
- p_2: пропорција приписаних докумената теми *t* које проистичу из речи *w*.

Слика 1. **Матрице које повезују речи, теме и документа у моделу *LDA***



	Реч 1	Реч 2	Реч 3	Реч 4	…
Док. 1	0,01	0	0,02	0	…
Док. 2	0	0,1	0	0,04	…
Док. 3	0,1	0,03	0	0,1	…
Док. 4	0,03	0,01	0,02	0	…
…	…	…	…	…	…

	Тема 1	Тема 2	Тема 3	Тема 4	…
Док. 1	0	0	0,22	0,14	…
Док. 2	0,7	0	0,1	0	…
Док. 3	0,1	0,03	0,03	0,1	…
Док. 4	0,06	0,01	0,01	0	…
…	…	…	…	…	…

	Реч 1	Реч 2	Реч 3	Реч 4	…
Тема 1	0,05	0,14	0,1	0	…
Тема 2	0,1	0	0	0,06	…
Тема 3	0,4	0,03	0	0,1	…
Тема 4	0	0,02	0,15	0	…
…	…	…	…	…	…

Ако већи број речи из датог документа припада истој теми *t* (високо p_1), вероватније је да и тренутна реч припада тој теми. Ако тренутна реч има високу вероватноћу да припада теми *t*, сви документи који садрже *w* биће у већој мери повезани с темом *t* (високо p_2). Према томе, што је производ вероватноћа, $p_1 \cdot p_2$ виши, вероватније је да текућа реч *w* припада теми *t*.

LDA се спроводи у великом броју итерација тако што се речи *w* приписује нова тема на бази производа вероватноћа $p_1 \cdot p_2$ све док се не постигне равнотежно стање. Крајњи резултат примене овог метода јесте груписане сродне речи по темама, с вероватноћама сваке од њих да проистичу из дате теме, те расподеле вероватноћа учешћа појединачних тема у сваком појединачном документу. За детаљан математички приказ овог метода видети *Blei et. al.* (2003).

3. Припрема текстова за анализу

Пре саме примене модела *LDA*, текстове је пожељно скратити и кориговати тако да буду погоднији, али и бржи за машинску обраду. То укључује свођење речи на основни облик, кориговање латиничких слова специфичних за српски језик, избацивање честих небитних речи, елиминисање знакова интерпункције и претварање великих слова у мала.

За српски језик карактеристично је постојање великог броја облика исте речи, за шта се у лингвистици каже да језик има *високу флексију*, што генерално отежава било коју врсту текстуалне анализе. Примера ради, речи *nafta*, *nafte*, *naftu*, *naftni*, *naftna*[1] и сл. алгоритам ће препознати као различите, иако се све односе на облике исте основне речи.

[1] Како смо текстове обрађивали на латиници, речи и теме које су део ове анализе писаћемо латиничким словима.

У нашој анализи то би смањило значај овог појма за раздвајање чланака на теме, јер ће његови различити облици имати мању фреквенцију појављивања.

Један од метода да се речи сведу на основни облик јесте стеминг (*stemming*), који отклања суфиксе различитих облика исте речи, остављајући њихов заједнички почетак. Други метод је лематизација, која своди речи на корен коришћењем речника. Овај метод је знатно сложенији, јер захтева постојање речника са свим облицима свих речи у језику. Стемер се, с друге стране, базира на правилима чији број не прелази неколико стотина, па га је далеко једноставније развити.

За српски језик је развијено неколико аутоматских, програмских стемера. Најпознатији на који смо наишли у литератури развили су Кешељ и Шипка (2008). Они у свом раду наводе да би стемер за српски језик морао имати осам пута више правила за отклањање суфикса речи у односу на енглески језик, што је добра илустрација високе флексије српског језика. У горњем примеру навели смо неколико облика речи *nafta*, док је *oil* једини облик ове речи на енглеском језику.

Уместо употребе општег стемера који би третирао све речи у текстовима, определили смо се за свођење на основни облик само речи од интереса за нашу анализу, конкретно речи са економским значењем, чиме фаворизујемо њихову улогу у класификацији текстова на теме. Наиме, у односу на неекономске речи, које ће остати у великом броју различитих облика, економске речи сведене на основни облик имаће веће учешће и тиме бити релевантније за класификацију. Тиме смо отежавајућу околност да анализу радимо на језику с великим бројем промена речи, на неки начин, претворили у предност за ову врсту анализе.

Основни облик смо интерпретирали на најшири могући начин тако да он, за разлику од стандардних стемера, укључи све придевске, глаголске и именичке облике речи. Док би стемери за српски језик речи *izvoz*, *izvozni,* и *izvoziti* третирали као различите основне облике, у нашој анализи све облике ових речи сводимо на реч *izvoz* (опет ради давања већег значаја економским изразима). Штавише, у појединим случајевима груписали смо речи и са различитим префиксима, као што је случај са *skupo* и *poskupljenje*.

Свођење речи на основни облик рађено је на два начина. Тамо где је било могуће, речи са истим основним облицима идентификоване су на основу заједничких почетних слова, а затим су замењене тим основним обликом. Примера ради, речи *nafta, nafte, naftu, naftni, naftna* итд. све почињу словима *naft* и њих смо заменили речју *nafta*. По истом принципу, речи које почињу на *kamat* заменили смо речју *kamata*, речи које почињу на *inflaci* или *inflator* речју *inflacija* итд. Ово правило је било могуће применити тамо где заједнички почетак за различите облике основног појма није истовремено заједнички почетак неког другог несродног појма.

За разлику од тога, на пример, заједнички почетак за различите облике речи *cena* (*cenu, cene, cenama, cenovni, cenovnik*…) – *cen* – истовремено је заједнички почетак за облике неких других несродних појмова (*centar*, *ceniti*, *cenzura*). Примена претходног правила у овом случају неоправдано би претварала и те друге речи у реч *cena*, мењајући њихово значење. Због тога је у том и сличним случајевима било неопходно експлицитно дефинисати све конкретне облике речи које желимо да претворимо у основни облик.

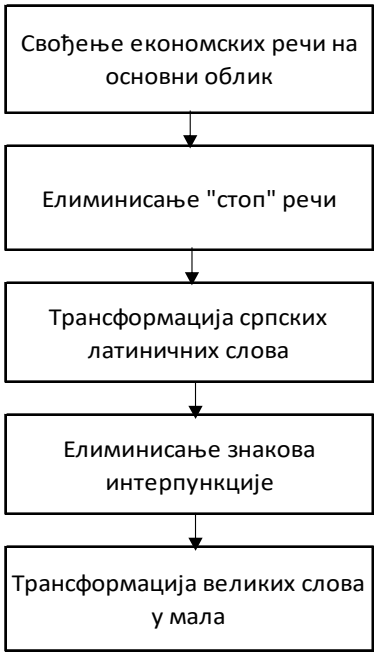
Први метод (препознавање на основу почетка речи) очигледно је једноставнији за примену, па смо га користили и у појединим случајевима у којима постоји више појмова за исте почетке речи. На пример, речи које почињу на *bank* могу бити облици речи *banka* или *bankina*, али како је овај други појам врло редак, ако не и непостојећи у економским текстовима, није било разлога за бојазан да би његово неоправдано трансформисање могло пореметити даљу анализу.

У самом избору економских речи за свођење на основни облик постојао је одређени степен арбитрарности. Поједине речи, попут *saobraćaj, država, vlada* или *grejanje*, стриктно гледано, нису економске, али с обзиром на то да смо анализирали само економске текстове, претпоставили смо да поменуте речи имају економску конотацију, па смо их укључили у списак за корекцију. У овај списак укључили смо и називе појединих институција које се често помињу у новинским чланцима (ММФ, ЕПС, Телеком).

Процес припреме текста за даљу анализу укључио је и елиминисање речи које су честе у српском језику, али саме за себе немају суштинског значаја за анализу текста (тзв. стоп речи). У питању су речи као што су: *i*, *ili*, *ali*, *koji*, *to*, *od*, *gde* и др., чије би задржавање у тексту због своје учесталости могло водити томе да их алгоритам препозна као кључне у раздвајању текстова на теме, што нам не би било корисно за економске, нити било које друге анализе.

Како програм који користимо за анализу текста не препознаје латиничка слова која су специфична за српски језик (*č, ć, ž, đ, š*), било их је неопходно трансформисати у облике погодне за обраду. То је урађено комбиновањем основних слова (без дијактричких знакова) са словима која се не користе у српском језику (*x*, *y*), на следећи начин: *č→cx, ć→cy, ž→zx, đ→dx, š→sx.*

Слика 2. **Поступак припреме текста за модел *LDA***



Додајмо да су из текстова елиминисани и знаци интерпункције, а велика слова претворена у мала. Све ове интервенције – елиминисање честих речи и знакова интерпункције, те свођење честих економских речи на заједнички основни облик – поред тога што текст чине погоднијим за нашу анализу, смањују време за обраду текста, што није небитан фактор када процес траје по неколико сати.

На основу описаних правила трансформације дајемо пример једног чланка:

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 trž ištu i neće se odnositi na snižene cene, kao što su rasprodaje, sezonska sniženja ili akcijske pr odaje, ukoliko su bila na snazi 15. novembra, već na redovne, odnosno cene pre sniženja, sao pšteno 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.

и његове обрађене верзије:

vlada ogranicxila cena osnovnih zxivotnih namirnica vlada srbije danasxnjoj sednici donela o dluku
ogranicxi visinu cena osnovnih zxivotnih namirnica sxecyer brasxno tip t400 suncokretovo ulj e 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 poremec yaji
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 mese ci
krsxenje navedenih odredaba predvidjene su novac kazne iznosu 100000 dva milion dinar za brana
vrsxenja delatnosti trajanju sxest meseci jedne godine

Мада је људском уму свакако разумљивији први чланак, за машинску обраду методом *LDA* погоднији је обрађени чланак из разлога које смо изнели раније. Поред тога, обрађени чланак је и мањи (за око 30%), што је битан фактор у обради великих количина података.

Избор речи за трансформацију економских речи и елиминисање честих сувишних речи није био у потпуности унапред дефинисан, већ је допуњаван у више корака

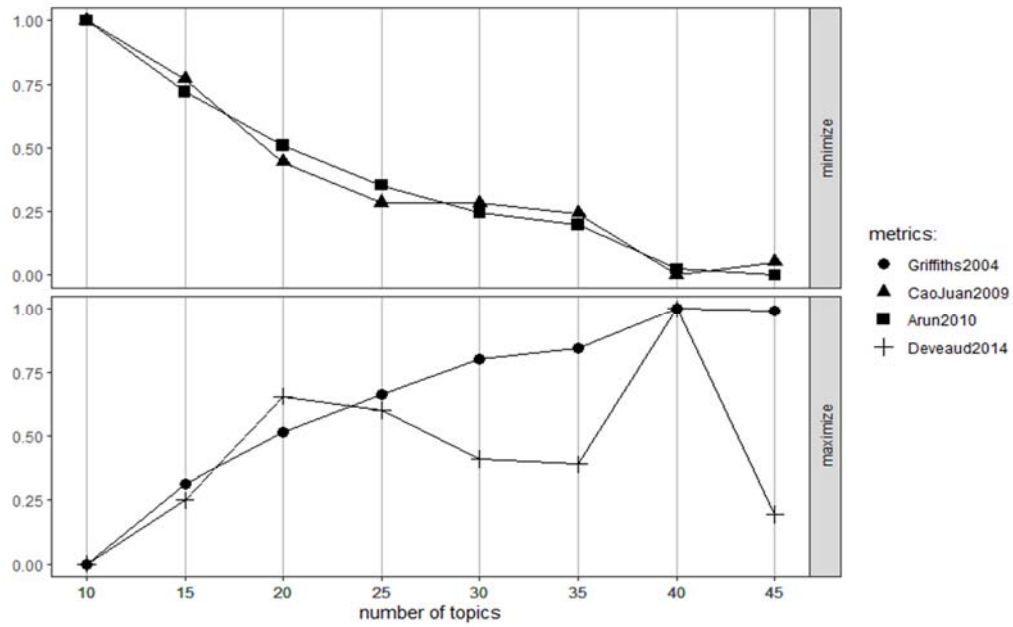
паралелно с применом модела *LDA*. У првом кораку модел је пуштен са иницијалним краћим списком коригованих односно елиминисаних речи, на основу чега је извршена класификација текстова на теме. Анализом најчешћих речи у темама идентификоване су нове стоп речи, као и економске речи које нису сведене на основни облик. У појединим темама доминирале су за даљу анализу небитне речи попут *koji, koje, bio, treba* и др., које у првој итерацији нису биле изостављене из текстова. Поред тога, поједине економске речи појављивале су се у различитим облицима (нпр. *тржиште, тржишни, тржишна*…) а нису се нашле на почетном списку речи за корекцију. На основу ове анализе проширили смо спискове стоп речи и економских речи и поново пустили модел *LDA*. Цео поступак поновили смо неколико пута док доминантне речи у темама нису ишчишћене на жељени начин. Коначан списак речи за корекцију приказан је у Табели А1 у Додатку.

4. Класификовање текстова на теме применом модела *LDA*

Полазну основу за тематску класификацију текстова моделом *LDA* представља матрица документ-реч, чији елементи представљају учешћа појављивања појединачних речи у документима. У нашој анализи ова матрица има 25.248 · 193.142 елемента, где је прва димензија број тема, а друга број јединствених речи.

При пуштању модела *LDA*, неопходно је специфицирати број тема на које желимо да класификујемо чланке у узорку. За одређивање оптималног броја тема за класификацију користили смо четири критеријума (*Griffiths, et. al*. (2004), *Cao et al*. (2009), *Arun et al*. (2020), *Deveaud* (2014)). Према три од наведена четири критеријума, оптималан број тема за наш узорак јесте 40 (Графикон 1).

Графикон 1. **Критеријуми за избор оптималног броја тема за модел *LDA***



Као што је већ речено, модел смо оцењивали у више корака, допуњавајући у сваком кораку списак речи за корекцију односно елиминацију, на основу анализе најчешћих речи у темама.

На Графикону А1. у Додатку приказана је финална подела на теме с приказом најчешћих речи. Технички речено, у питању су речи с највећим оцењеним β коефицијентима, који представљају вероватноће да конкретна реч произлази из конкретне теме. За потребе овог приказа речи вратили смо трансформацију српских латиничких слова (*cx* → *č*...). Напомињемо да сâм модел није именовао теме (само их је обележио бројевима), већ смо то ми урадили на основу доминантних речи по темама.

У већини случајева било је једноставно одредити тему на основу најчешћих речи. На пример, тема у којој су најчешће речи: *zaposlenost, radnik, rad, posao* – очигледно се односи на тржиште рада; тема с речима: *proizvod, poljoprivreda, tržište, tona, pšenice, voća* – тиче се тржишта пољопривредних производа; тема с речима *energetika, EPS, elektrika; struja; uglja* – бави се електричном енергијом (због једноставности смо је назвали *struja*) итд. Из угла централне банке и монетарне политике, од посебног значаја су теме које се тичу банкарског сектора (*kredit, banka, kamata, dinar*…) и инфлацијом (*inflacija, odsto, rast, cena*…). За две теме није било могуће одредити наслов, зато што су у њима доминирале неекономске и неповезане речи. Ове теме смо назвали "*NEODREĐENO*"[2]. У појединим случајевима имали смо више тема које се тичу исте области (*PRIVREDA*, *POLJOPRIVREDA*, *VLADA, PENZIJE, TURIZAM, RADNICI*).

Док свака тема представља комбинацију различитих речи, сваки чланак представља комбинацију различитих тема. Коефицијент γ представља учешће појединачних тема у сваком документу. Чланци могу бити доминантно представљени једном темом или као комбинација више тема.

На пример, у следећем кратком чланку доминантна тема је *GAS (γ = 0,98)*:

***Na granici sa Mađarskom povezan gasovod Balkanski tok** (4.7.2021.)*

Horgoš – Javno preduzeće Srbijagas i mađarska kompanija FGSZ povezali su danas na granici s Mađarskom gasovod Balkanski tok, kojim će ubuduće gas iz Turske preko Bugarske i Srbije stizati do srednje Evrope.

Dušan Bajatović, generalni direktor Srbijagasa, rekao je da je na ovaj način stavljena tačka na dugogodišnji veliki posao kojim je naša zemlja i konačno rešila pitanje snabdevanja gasa iz drugog pravca, a ne samo preko Ukrajine.

"Niko se više u Srbiji neće smrzavati a cena gasa se za domaćinstva neće ni od jeseni menjati. Poskupljenje zbog cene nafte koju prati cena gasa će se preliti na Srbijagas, objasnio je Bajatović.

[2] Имена тема пишемо великим словима.

С друге стране, наредни чланак комбинација је више тема, где ниједна од њих нема већинско учешће (*GORIVO 0,42; BERZA 0,33; NAFTA 0,24*):

***Evropski berzanski indeksi uglavnom u padu, cena zlata na istorijskom maksimumu** (4.12.2023.)*

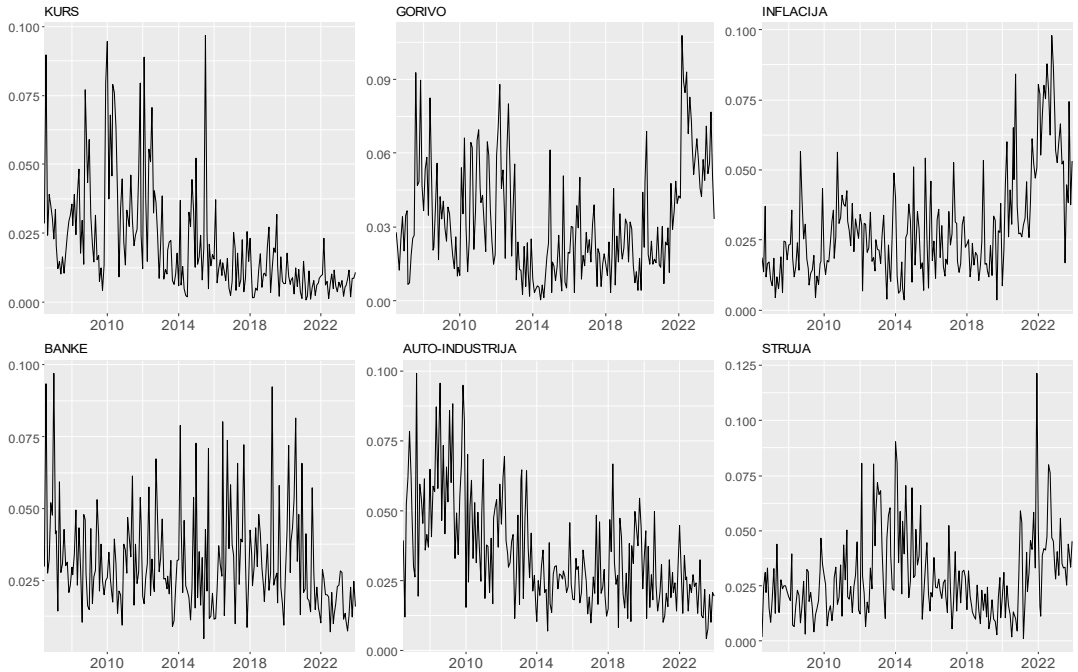
NjUJORK/FRANKFURT/ MOSKVA – Evropski berzanski indeksi su na početku nedelje uglavnom u padu, dok je cena zlata danas na istorijskom maksimumu. Indeks frankfurtske berze DAKS je danas u 10.00 sati porastao na 16.421,45 poena, dok je francuski KAK 40 pao na 7.333,14 poena, kao i londonski FTSE 100 - na 7.508,57 poena i moskovski MOEKS - na 3.113,05 poena. Vrednost američkog berzanskog indeksa Dau Džouns je pre današnjeg otvaranja berzi u Americi porasla na 36.245,50 poena, kao i vrednost indeksa S&P 500 - na 4.594,63 poena i vrednost indeksa Nasdak - na 14.305,03 poena.

Prema podacima sa berzi, cena sirove nafte je pala na 72,874 dolara za barel, kao i cena nafte Brent - na 77,615 dolara. Evropski fjučersi gasa su se danas na otvaranju berze TTF prodavali po ceni od 42,750 evra za megavat-sat.

Cena zlata je danas rano ujutru dostigla istorijski maksimum od 2.110,8 dolara za uncu, a u 10.00 sati je malo pala na 2.069,16 dolara za uncu (unca iznosi 28,35 grama), Pšenica je takođe poskupela na 5,7936 dolara za bušel (bušel iznosi 27,216 kg). Vrednost evra u odnosu na dolar je iznosila 1,08692, što je otprilike isto kao u petak, prenosi Tanjug.

Временске серије учешћа тема показују изузетно високу волатилност (Графикон 2). Иако узорак за тематску класификацију садржи велики број чланака (преко 25.000), разбијено на месечни ниво, то је у просеку 120 чланака, подељених на 40 тема, што у просеку значи мали месечни број чланака по теми.

Графикон 2. **Учесталост појављивања изабраних тема у новинским чланцима** (месечни проседи γ коефицијента)



И поред високе волатилности, може се приметити да кретање учесталости тема добро одсликава одређена дешавања у економији: интересовање за тему девизног курса расло је у периодима његове високе волатилности, да би са стабилизацијом курса последњих година, опадало учешће ове теме; ауто-индустрија била је честа тема у периоду око доласка Фијата у Крагујевац; гориво и струја честе су теме у периодима њиховог поскупљења; док је интересовање за инфлацију највеће у периодима инфлационих циклуса.

5. Оцена везе инфлационих очекивања и тема

Становништво формира инфлациона очекивања на основу информација које добија из разних извора, између осталог, и из медија. Истовремено, могућ је и повратан утицај – писање медија може бити одраз инфлационих очекивања. У сваком случају, писање на одређене теме потенцијалан је индикатор кретања инфлационих очекивања.

У овом раду смо везу инфлационих очекивања и тема оценили регресијом *LASSO* (*Tibshirani*, R. (1996)), која је прикладна у случајевима у којима је број варијабли у регресији велики, од којих су неке нерелевантне. Кључна карактеристика модела је у томе што се у функцију циља обичних најмањих квадрата додаје „казнени" елемент ($\lambda\sum_{j=1}^{p}|\beta_j|$), који подстиче модел да коефицијенте (β_j) мање важних варијабли сведе на нулу:

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

Параметар λ одређује меру у којој желимо да „кажњавамо" задржавање коефицијената у моделу (веће λ – мање ненултих коефицијената). Ми смо у нашој анализи користили параметар λ који резултира у моделу с најмањом ванузорачком средњом грешком прогнозе (Табела А3 у Додатку).

У нашем конкретном случају регресирали смо инфлациона очекивања (π_t^{exp}) на 40 варијабли кретања учешћа појединих тема (T_t^i) на периоду јануар 2009 – децембар 2023. Да бисмо избегли потенцијални проблем ендогености (симултаног утицаја), као независне варијабле узели смо доцње од једног месеца.

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

Оцењени модел је 17 варијабли „задржао" као релевантне, док је коефицијенте осталих свео на нулу (Табела 1), уз коефицијент детерминације R^2 = 0,67 (Табела А3).

Као варијабле с позитивним коефицијентом појављују се неке теме за које би то било очекивано: *INFLACIJA*, *GORIVO*, *NAFTA*, *STRUJA*, *PLATE-POTROŠNJA*. У поменутим темама везаним за енергенте често се помиње кретање њихових цена. У случају појединих варијабли, међутим, недостаје јасна економска интерпретација коефицијената (нпр. негативан коефицијент код тема *BAKAR*, *ZAPOSLENOST*, *GRADNJA PUTEVA,…*).

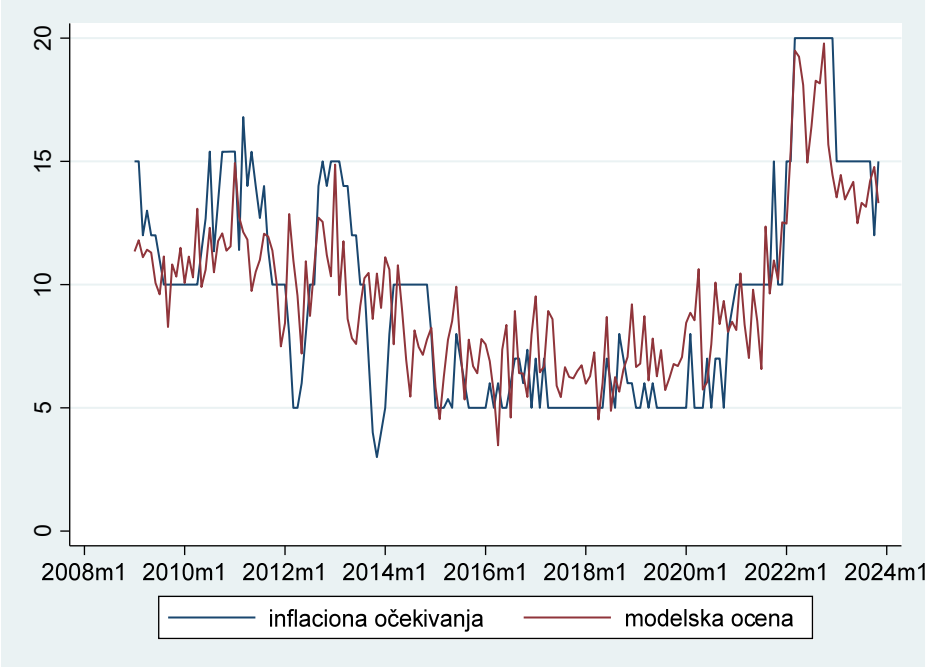
Табела 1. **Оцењени коефицијенти за теме из модела *LASSO* за инфлациона очекивања**

Тема	Коефицијент
1. PRIVREDA	0
2. DEVIZNI TOKOVI	0
3. BAKAR	-35,1
4. TURIZAM	0
5. TELEFONIJA	0
6. GORIVO	36,6
7. RADNICI	0
8. KURS	0
9. POLJOPRIVREDA	0
10. TURIZAM	0
11. POREZ	0
12. BUDŽET	0
13. INFRASTUKTURA	0
14. VLADA	0
15. STRUJA	33,3
16. INFLACIJA	40,0
17. BANKE	0
18. PENZ.OSIGURANJE	-18,5
19. RADNICI	0
20. NEODREĐENO	0

Тема	Коефицијент
21. AUTO-INDUSTRIJA	9,1
22. PRAVO-PRIVREDA	-16,6
23. MEĐUNARODNA	0
24. PENZIJE	0
25. NAFTA	70,3
26. PLATE-POTROŠNJA	83,0
27. ZAPOSLENOST	-30,6
28. POLJOPRIVREDA	0
29. LOKALNA	27,2
30. PRIVREDA	0
31. TRGOVINA	0
32. AVIO SAOBRAĆAJ	-12,0
33. GAS	0
34. BERZA	0
35. STANOVI	50,1
36. VLADA	60,0
37. PRIVREDA-BANKE	43,7
38. GRADNJA PUTEVA	-10,3
39. NEODREĐENO	0
40. ČELIK	-23,6

На Графикону 3 може се приметити да модел добро предвиђа циклусе инфлационих очекивања, с тим да је оцењена серија много волатилнија од самих очекивања. То је логична последица високе волатилности у кретању појединачних тема, с једне стране, и релативно стабилног кретања инфлационих очекивања из анкете, с друге стране.

Графикон 3. **Инфлациона очекивања и њихова оцена на основу модела *LASSO* с кретањем тема (у %)**



6. Закључак

У раду приказали смо тематску класификацију 25.248 чланака из економске рубрике дневног листа *Политика* у периоду 2006–2023. године, применом метода *LDA*.

Специфичност нашег приступа огледа се у томе што смо у фази припреме текста селективно бирали речи које смо сводили на основни облик. То су биле речи са економским значењем, којима смо на тај начин дали предност у односу на неекономске речи, које су остале у великом броју облика с мањим фреквенцијама појављивања. Та предност је нарочито изражена у језику с високом флексијом, какав је српски, у коме речи имају велики број облика. Избор речи за трансформацију није био у потпуности унапред дефинисан, већ је допуњаван у више корака паралелно с применом модела *LDA*, на основу анализе најчешћих речи у темама.

Модел *LDA* је овако кориговане новинске чланке класификовао на економске теме на задовољавајући начин. У већини случајева било је недвосмислено на шта се односи одређена тема, а само код две, у којима су доминирале неекономске речи, нисмо успели да одредимо садржај теме. Неке теме из узорка имају релативно широк опсег *(PRIVREDA, MEĐUNARODNA*…), док су неке уско специфичне (*TELEKOM, NAFTA, BAKAR, ČELIK, STRUJA*…).

Кретање тема на месечном нивоу показује високу волатилност, што се може објаснити недовољно великим узорком за толики број тема. Ове серије смо употребом модела *LASSO* регресирали на инфлациона очекивања становништва. Оцењени модел је добро ухватио инфлационе циклусе, с високим коефицијентом детерминације. Од 40 тема, модел је задржао 17 као релевантне, од којих је за неке то било очекивано, док за неке не постоји јасна економска интерпретација. За поузданије економетријске анализе вероватно је пожељно проширити узорак докумената, што ће нам бити један од циљева у наредном периоду.

Додатак

Табела А1. **Замена речи основним обликом на основу почетка речи**

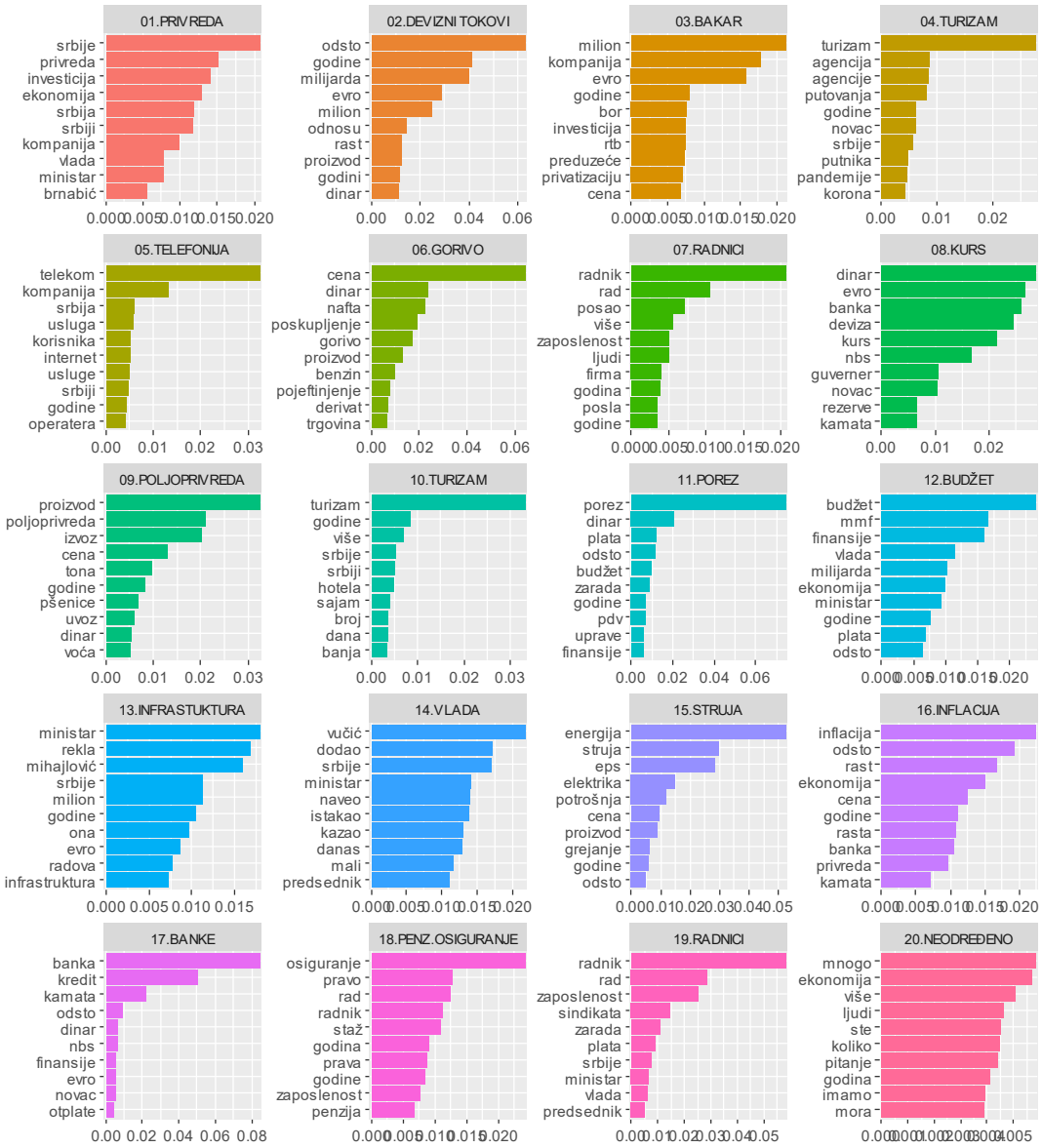
Почетак речи	Замена
inflaci	*inflacija*
inflator	*inflacija*
deflaci	*deflacija*
deflator	*deflacija*
poskup	*poskupljenje*
skuplj	*poskupljenje*
pojeft	*pojeftinjenje*
jeft	*pojeftinjenje*
kurs	*kurs*
dinar	*dinar*
dolar	*dolar*
deviz	*deviza*
novc	*novac*
kredit	*kredit*
kamat	*kamata*
bank	*banka*
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*

Почетак речи	Замена
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*

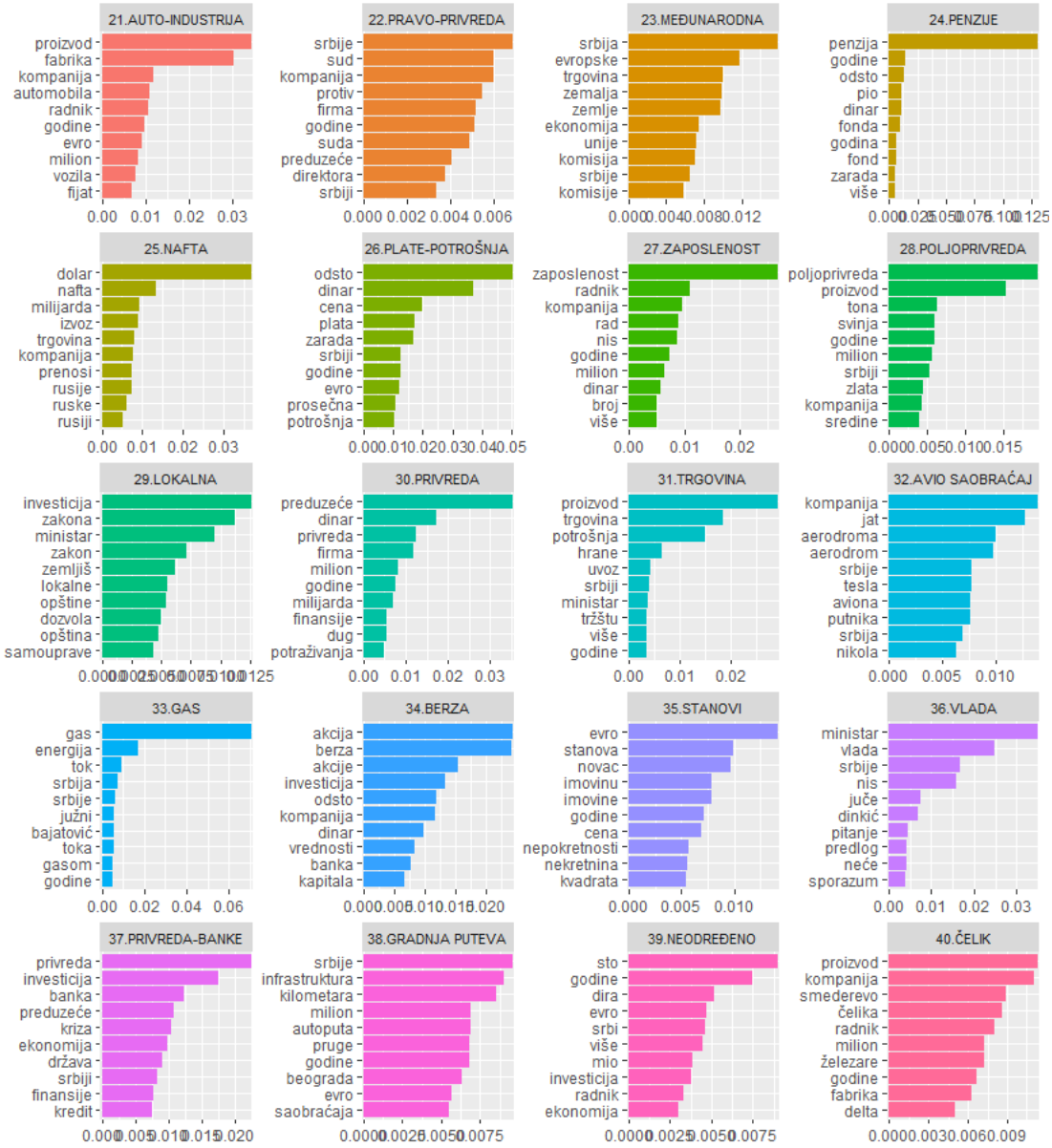
Табела А2. **Замена речи њеним основним обликом на основу целе речи**

Реч	Замена
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*

Слика А1-а. **Најчешће речи у темама** (β коефицијенти)



Слика А1-б. **Најчешће речи у темама** (β коефицијенти)



Табела А3. **Резултати *LASSO* регресије инфлационих очекивања на теме новинских чланака**

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

Литература

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). „Procena inflatornih pritisaka putem analize novinskih tekstova". Zbornik radova. Narodna banka Srbije, Septembar 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 Sipka 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.