СТРУЧНИ РАДОВИ

ПРИМЕНА НАИВНОГ БАЈЕСОВОГ КЛАСИФИКАТОРА У АНАЛИЗИ НОВИНСКИХ ТЕКСТОВА

Ана Вујовић

Доступно на www.nbs.rs

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

Примена Наивног Бајесовог класификатора у анализи новинских текстова

Ана Вујовић

Апстракт: У овом раду представљен је Наивни Бајесов класификатор, један од стандардних модела који се користи за решавање класификационих проблема у текстуалној анализи. Овај модел је изучен у раду прво с теоријског аспекта, а након тога и с практичног. Спроведено је емпиријско истраживање чији је задатак био тематска класификација новинских текстова употребом Наивног Бајесовог класификатора. Резултати нашег истраживања потврђују да овај класификатор има високу предиктивну моћ упркос поједностављеним претпоставкама на којима се базира. Такви налази упућују на то да постоји простор за даљу примену Наивног Бајесовог класификатора када је у питању тематска класификација текстова, што може имати значајне импликације за економска истраживања.

Кључне речи: Наивни Бајесов класификатор, тематска класификација, обрада природног језика
[JEL Code]: C13, E37

Нетехнички резиме

У последње време обрада природног језика постаје све актуелнија тема у економији, поготово након избијања пандемије, у периоду изражене неизвесности, када су уочени значај и потреба за применом модела који користе фреквентније податке. Како новински текстови или коментари на друштвеним мрежама прате актуелна дешавања, анализа ове врсте података је корисна за праћење економских промена у датом тренутку.

С популаризацијом обраде природног језика и са све већом доступношћу текстуалних извора података дошло је до адаптације класичних класификационих модела текстуалним подацима. Један од модела који се у пракси показао као изузетно користан у овој области јесте Наивни Бајесов класификатор.

У овом раду представљен је Наивни Бајесов класификатор (НБК), с теоријског и с практичног аспекта, односно приказано је на који начин се овај модел може применити у текстуалној анализи. Поред тога, један део рада посвећен је и обради текстуалних података, с обзиром на то да су ови подаци нешто захтевнији за моделирање од нумеричких. У раду смо спровели емпиријско истраживање чији је задатак био тематска класификација новинских текстова употребом НБК-а.

Анализа је спроведена на узорку од 2225 новинских текстова са сајта Би-Би-Сија из пет различитих категорија. Како је НБК надгледани метод машинског учења, што подразумева обучавање модела на претходно обележеним подацима, где модел „учи" како да повеже задате инпуте с познатим аутпутом, а затим након обуке обавља исти задатак на необележеним подацима, део података смо искористили за обуку модела, а на другом делу података тестирали смо предиктивну моћ овог модела. Резултати анализе упућују на то да НБК има високу предиктивну моћ, упркос поједностављеним претпоставкама на којима се заснива. Такав налаз отвара простор за даљу примену овог модела у текстуалној анализи, што се у пракси показало као веома корисно за различите економске анализе.

Садржај:

1. Увод

У последње време обрада природног језика постаје све актуелнија тема у економији, поготово након избијања пандемије, у периоду изражене неизвесности, када су уочени значај и потреба за применом модела који користе фреквентније податке. Како новински текстови или коментари на друштвеним мрежама прате актуелна дешавања, анализа ове врсте података је корисна за праћење економских промена у датом тренутку.

Улагање у технике за анализу текстуалних података може да буде исплатива инвестиција за централне банке, с обзиром на то да ове технике омогућавају управљање низом извора података који су важни за процену монетарне и финансијске стабилности и који се не могу квантитативно анализирати другим методама (*Bholat D.*, *Hans, S.*, *Santos, P., & Schonhardt-Bailey, C.*, 2015). Стога је текстуална анализа област којој се у Народној банци Србије придаје све већи значај. Креиран је новински инфлациони сентимент на основу пребројавања израза који се односе на промене цена и на инфлацију, а затим је потврђена статистички значајна веза са инфлацијом, што указује на то да новински инфлациони сентимент може да се користити као показатељ јачине инфлаторних притисака (Ђукић, 2022). Такође, пронађена је значајна веза између учешћа појединих тема у новинским чланцима кроз време и инфлационих очекивања становништва (Ђукић, 2024). Тематска класификација новинских текстова предмет је и овог рада.

Медијска покривеност одређених тема може служити као добар показатељ економских кретања, па су зато новински текстови најзаступљенији извор података за обраду природног језика у економији. Информације из новинских текстова могу бити од користи не само када је у питању инфлација већ и за праћење привредног раста, инвестиција, незапослености или неизвесности, с обзиром на то да је сентимент уско повезан и са овим варијаблама. За краткорочно пројектовање раста реалног бруто домаћег производа у зони евра показатељ сентимента из новинских чланака показао се бољи од званичних пројекција Европске централне банке и од *PMI* индекса (*Purchasing Managers composite index*), поготово током финансијске кризе (2008/09) и у периоду потпуног затварања због пандемије (2020) (*Ashwin, J.*, *Kalamara, E.*, *Saiz*, *L.*, 2021). Када се ради о неизвесности, учесталост понављања појмова попут „економија", „неизвесност" „дефицит" и сл. може се искористити за конструисање индекса неизвесности економске политике (у САД), чије се кретање поклапа са економском теоријом о вези неизвесности и најважнијих макроекономских варијабли (*Baker, S. R., Bloom, N.*, *& Davis*, *S. J.*, 2016). Према нешто новијем истраживању (*Kalamara, E., Turrell, A., Redl, C.*, *Kapetanios*, *G., & Kapadia*, S., 2020), новински текстови садрже релевантније сигнале економског сентимента него неизвесности, иако је литература претежно фокусирана на примену текстуалних података за процену неизвесности.

У овом раду биће приказан Наивни Бајесов класификатор (НБК), један од стандардних модела који се користи за решавање класификационих проблема у анализи новинских текстова. Ово истраживање има за циљ упознавање с НБК моделом и

оцењивање могућности његове примене у анализи текстуалних података, што може имати значајне практичне импликације када су у питању економска истраживања.

Овај рад је организован у неколико поглавља. Након увода, у другом поглављу приказана је уобичајена процедура за припрему и обраду текстуалних података. Након тога је представљен НБК, његове теоријске основе и најважније претпоставке на којима се базира, као и преглед емпиријске литературе о примени овог модела. У трећем поглављу износимо резултате емпиријског истраживања у ком смо употребили НБК за класификацију новинских текстова са сајта Би-Би-Сија[1] у пет категорија којима ови текстови припадају. У последњем поглављу сумирамо наша запажања.

2. Припрема и обрада текстуалних података

С обзиром на то да су предмет анализе овог рада текстуални подаци који захтевају нешто сложенију припрему од нумеричких података, у овом делу рада биће представљени уобичајени кораци обраде и припреме текстуалних података. Конкретно, прво ћемо објаснити кораке сређивања, тј. нормализације текстуалних података, а затим трансформацију текстуалних података у нумеричке податке уз помоћ метода за извлачење карактеристика текстуалних података, као и метод за одабир ових карактеристика.

2.1. Обрада текстуалних података

Уобичајена процедура обраде текстуалних података подразумева неколико фаза: трансформацију великих у мала слова; искључивање свих карактера сем речи; стеминг или лематизација и избацивање стоп-речи.

Да програмски језик не би другачије третирао једну реч написану почетним великим словом јер се налази на почетку реченице од исте те речи написане малим словима, неопходно је претворити сва велика слова у мала. Затим треба искључити знакове интерпункције и остале карактере (попут карактера који се односи на нови ред \n) који нису речи да их програм не би третирао као значајне информације с обзиром на њихово често понављање. Ове трансформације се могу спровести **токенизацијом –** методом која омогућава поделу текста на токене. **Токени** су основна јединица анализе природног језика. Токени могу бити саме речи, реченице, сложенице од више појмова или карактери. У пракси се најчешће као токени користе речи, мада су у неким случајевима потребне и остале врсте токена. Приликом токенизације аутоматски се избацују остали непотребни карактери (односно карактери који нису речи а налазе се у текстовима), а ова функција у многим програмима нуди и могућност трансформације великих слова у мала.

Слично као с проблемом малих и великих слова, програмски језик другачије третира различите облике исте речи попут: игра, играње, играш, играти... Како ове речи преносе

[1] Подаци су преузети са платформе *Kaggle.*

исту или сличну информацију, за анализу је потребно да се овакви облици сведу на свој основни формат – односно корен речи, тј. да се скину суфикси. Постоје два приступа нормализацији речи: стеминг (*stemming*) или лематизација (*lemmatization*).

1. Стеминг подразумева скидање суфикса с речи. Овај метод је релативно једноставан јер подразумева задавање малог броја правила ради нормализације речи. Недостатак стеминга се огледа у томе што у неким случајевима аутпут не буде реч која има значење, што је илустровано на примеру приказаном у Табели 1.
2. Лематизација се заснива на речнику, па је резултат лематицазије увек ваљана реч. Овај метод је захтевнији за извођење, јер подразумева прво дефинисање речника, па обраду у складу с тим речником. Лематизација може бити временски и софтверски исувише захтевна, поготово у случајевима језика или тема где нису доступни унапред дефинисани речници за лематизацију.

На једноставном примеру приказаћемо разлику између ове две методе: у програмском језику *R* дефинишемо списак речи (на енглеском језику због доступности речника), који ћемо прво стемирати, па затим лематизовати. У Табели 1 приказан је резултат стеминга и лематизације на примеру пет речи истог порекла: резултат стеминга[2] у четири случаја није цела реч, него сам корен речи, док је лематизацијом добијен бољи резултат – све речи постоје, а две се третирају као да имају исто значење.

Табела 1. **Разлика између стеминга и лематизације**

Words	*Stemming*	*Lemmatization*
creation	*creation*	*creation*
create	*creat*	*create*
created	*creat*	*create*
creative	*creativ*	*creative*
creatively	*creativ*	*creatively*

Последњи корак обраде текстуалних података јесте избацивање стоп-речи. Стоп-речи су оне речи које се често појављују у тексту а не пружају релевантне информације о садржају текста (*Claudia*, *E*., *Scott B*., 2022). Оне се елиминишу из текста, јер би их у супротном алгоритам означио као релевантне због фреквентног појављивања и нарушиле би анализу текста. Примери стоп-речи су: и, или, да, како, зато, потом и сл. Српски језик не карактеришу чланови, али у енглеском језику је исто тако важно избацити чланове: *the*, *a*, *an* пре спровођења било какве анализе. Већина програмских језика садрже лексиконе стоп-речи, што овај корак чини једноставним.

2.2. Извлачење карактеристика из текста

Текстови садрже велики број информација које је, такорећи, немогуће апстраховати моделима. Зато се пре моделирања примењује извлачење карактеристика из текстуалних података. Циљ овога је трансформација текстуалних података у нумеричке векторе, које даље модели машинског учења обрађују. Постоје различите методе за

[2] За стеминг смо применили Портеров речник.

извлачење карактеристика из текстова, а у пракси се најчешће користе: обележавање делова реченица (*Part of Speech tagging*), идентификација, негација, *Bag of Words* и *Term Frequency-Inverse Document Frequency* (*TF-IDF*). У овом раду примењен је последњи метод, који ћемо и представити у овом делу рада.

TF-IDF вредност указује на релативни значај одређеног токена за документ у ком се налази у односу на целу збирку докумената. Наиме, овај метод помаже у одређивању пондера који треба да имају опште речи које се често појављују кроз збирку докумената у односу на специфичне речи које се ређе појављују, па су самим тим значајније за класирање документа (*Tripathy*, *A*., *Agrawal*, *A*., *Rath*, *S*., *K*., 2015). Једноставно речено, то што се нека реч много пута појављује у одређеном документу не значи да је значајна уколико се, такође, често појављује у свим документима (на пример, речи попут *and, or, if…*). Због тога главна претпоставка овог приступа је да је релативни значај неке речи обрнуто пропорционалан учесталости њеног понављања у свим документима у збирци (*Raschka*, *S*., 2014).

Да би се објаснило извођење овог статистичког показатеља, неопходно је прво објаснити све његове компоненте:

- укупан број докумената у збирци (*N*);
- *Term Frequency* (*TF*) – учесталост појављивања једне речи у документу *d*:

$$TF(t,d) = \frac{koliko\ puta\ se\ reč\ t\ pojavljuje\ u\ dokumentu\ d}{broj\ reči\ u\ dokumentu\ d} \quad (1)$$

- *Document Frequency* (*DF*) – у колико докумената се појављује одређена реч у целој збирци докумената;
- *Inverse Document Frequency* (*IDF*) – мера значајности одређене речи:

$$IDF(t,D) = log\ (\frac{N}{DF}) \quad (2)$$

па је на крају *Term Frequency-Inverse Document Frequency* (*TF-IDF*):

$$TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D) \quad (3)$$

2.3. Елиминисање ретких речи

Једна од уобичајених метода за смањивање броја опсервација када су речи у питању односи се на избацивање оних речи које се не појављују у великом броју докумената. То је зато што те речи могу бити изузетно специфичне, самим тим и не превише значајне за класификацију јер није велика вероватноћа да ће се наћи у делу података за предвиђање. У пракси се обично искључују оне речи које се појављују у мање од пет или десет докумената у целој збирци докумената. За то се користи *Document Frequency* (*DF*), односно задаје се правило да се искључе све опсервације за које важи $DF < 5$ (односно $DF < 10$).

3. Наивни Бајесов класификатор

Генерално, класификациони проблеми се односе на категоризацију опсервација на основу њихових атрибута у одговарајуће категорије. У случају класификационих проблема у области обраде природног језика, текстови се такође групишу у различите категорије на основу својих атрибута, само што су у овом случају атрибути саме речи.

НБК спада у пробабилистичке моделе јер се заснива на вероватноћама, а основу овог модела чини **Бајесово правило**. НБК треба да одреди вероватноћу да ће једна опсервација припасти одређеној класи на основу карактеристика те опсервације. Као инпуте, модел користи **априори** или унапред познате вероватноће да опсервација припада одређеној класи и **условну вероватноћу** о заступљености карактеристика унутар класа. Када су опсервације текстови, на основу вероватноћа да текст припада категорији и условних вероватноћа о карактеристикама текста у оквиру једне категорије, модел треба да одреди **постериор** вероватноћу којој категорији ће текст припасти. Модел додељује текст оној категорији где је ова вероватноћа највиша.

3.1. Бајесово правило

Да би се стекао увид у овај модел, потребно је прво представити Бајесову теорему, коју је формулисао британски статистичар Томас Бајес (1701–1761):

$$P(A\ |\mathrm{B}) = \frac{P(B|A)P(A)}{P(B)} \tag{4}$$

где је:

- $\mathrm{P}(A\ |\mathrm{B})$ – постериор вероватноћа, која представља вероватноћу хипотезе А под условом да се реализовао догађај В који представља доказ;
- $\mathrm{P}(B|A)$ – условна вероватноћа догађаја В под условом да је хипотеза А тачна;
- P(A) – априори вероватноћа (*prior probability*) хипотезе А без икаквог сазнања о догађају В;
- P(B) – вероватноћа догађаја В, односно стања на које указују подаци (*evidence*).

3.2. Примена Бајесовог правила у класификационим проблемима

Након представљања Бајесове формуле у свом општем облику, представићемо је у контексту класификационог проблема:

$$P(C_i|x_1, x_2 \ldots x_q) = \frac{P(x_1,x_2\ldots x_q|C_i)P(C_i)}{P(x_1,x_2\ldots x_q|C_1)P(C_1)+P(x_1,x_2\ldots x_q|C_2)P(C_2)+\cdots+P(x_1,x_2\ldots x_q|C_p)P(C_p)} \tag{5}$$

У овом случају постериор вероватноћа је вероватноћа да опсервација припада класи *i* под условом да има атрибуте *x_1,x_2…x_q*. Атрибути опсервације су нам познати и на основу њих треба одредити којој класи опсервација припада. Циљ је пронаћи постериор вероватноће за све класе (од 1 до *p*) и доделити опсервацију оној класи за коју је ова вероватноћа највећа. Ово правило одлучивања може се представити формулом:

$$dodeljena\ klasa \leftarrow \underset{i=1,2\ldots,p}{\operatorname{argmax}} P(C_i \mid x_1, x_2 \ldots x_q) \quad (6)$$

Условна вероватноћа представља вероватноћу да опсервација има карактеристике *x_1,x_2…x_n* и да припада класи *i*. Априори вероватноћа је вероватноћа да опсервација припадне одређеној категорији. У имениоцу формуле 5 налази се формула тоталне вероватноће за израчунавање вероватноће атрибута опсервације, односно то је вероватноћа да опсервација има одређене атрибуте невезано од класе којој припада. У случајевима када модел наиђе на опсервацију у тест-сету података која се није нашла у тренинг-сету података, овај израз је једнак 0. Како би се избегла ситуација да се израз дели нулом, додаје се **Лапласов фактор заглађивања** у бројилац и исти овај фактор помножен са укупним бројем атрибута у именилац формуле за израчунавање постериор вероватноће.

3.3. Претпоставке модела

Да би се лакше израчунала постериор вероватноћа, уводи се **наивна** претпоставка, на основу које је овај модел и добио атрибут „наивни". Ово поједностављење се односи на то да се претпоставља да су **атрибути међусобно независни**. У реалности ова претпоставка не важи, али то не утиче на перформансе овог модела који се показао добар у многобројним класификационим задацима. Ово се може објаснити чињеницом да оптималност у контексту **класификационе грешке** (губитак 0/1) не зависи од тога у којој мери се подаци поклапају с претпостављеном расподелом вероватноће, већ је важно да реализована и претпостављена дистрибуција одговарају највероватнијој класи (*Rish*, *I*., 2001). Значајнији показатељ прецизности НБК-а је количина информација о класи које се изгубе због претпоставке о независности атрибута (*Rish*, *I*., 2001). Ова претпоставка олакшава израчунавање условних вероватноћа, које се у том случају рачунају на следећи начин:

$$P(x_1, x_2 \ldots x_q \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i) \quad (7)$$

Онда формула за рачунање постериор вероватноће постаје:

$$P(C_1 \mid x_1, x_2 \ldots x_q) = \frac{P(x_1 \mid C_1)\mathrm{P}(x_2 \mid C_1)\ldots P(x_q \mid C_1)P(C_1)}{P(x_1 \mid C_1)P(x_2 \mid C_1)\ldots P(x_q \mid C_1)P(C_1) + \cdots + P(x_1 \mid C_p)P(x_2 \mid C_p)\ldots P(x_q \mid C_p)P(C_p)} \quad (8)$$

За израчунавање условних вероватноћа уводи се додатна претпоставка, па се оне могу израчунати на следећи начин:

$$P(x_k \mid C_i) = \frac{n}{N} \quad (9)$$

То значи да се вероватноћа да опсервација са атрибутом x_k припада класи C_i изражава као количник броја опсервација са атрубутом x_k који припадају класи C_i (у делу података на којима се модел обучава) – n и укупног броја опсервација у класи C_i – N.

3.4. Наивни Бајесов класификатор у текстуалној анализи

Када је у питању примена НБК-а у текстуалној анализи, атрибути x_k су речи, а класе C_i категорије текстова. Условна вероавтноћа $P(x_1|C_1)$ јесте вероватноћа да се одређена реч нађе у задатој категорији. Ова вероватноћа се рачуна као однос броја речи x_i које се појављују у категорији C_i и укупног броја речи у посматраној категорији. Постериор вероватноћа $P(C_1|x_1, x_2 \dots x_q)$ јесте вероватноћа да текст који садржи одређене речи припадне категорији од интереса. Ово ћемо илустровати на примеру реченице: „Централна банка је заоштрила монетарну политику". У овој реченици постериор вероватноћа је вероватноћа да ова реченица припадне категорији економских текстова, а она се израчунава на основу условних вероватноћа да свака од речи из реченице припада овој категорији.

3.5.Мере ваљаности класификационих модела

Пре залажења у примере у којима се користи НБК требало би представити мере ваљаности класификационих модела да би се боље разумеле перформансе овог модела. Ови показатељи користе се за све класификационе моделе и без обзира на тип података.

Да би се извели показатељи ваљаности класификације, неопходно је прво креирати матрицу **класификације** (*confusion matrix*), у којој је приказан однос између тачно и нетачно класификованих опсервација. Ова матрица се може применити и када се класификациони проблем односи на више од две класе, али у табели испод, ради једноставности, приказана је класификациона матрица за бинарни класификациони проблем, где су класе назване позитивна и негативна.

Табела 2. **Матрица класификације**

		Одговарајућа категорија	
		Positive	***Negative***
Додељена категорија	*Positive*	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

Из ове матрице могу се извући четири категорије опсервација:

1. Опсервације које су заиста позитивне и модел их је доделио категорији којој припадају – ***true positive*** (*TP*);
2. Опсервације које су позитивне, али их је модел доделио погрешној категорији – ***false negative*** (*FP*);
3. Негативне опсервације које су додељене негативној категорији – ***true negative*** (*TN*);
4. Негативне опсервације које је модел класификовао као позитивне – ***false positive*** (*FP*).

На основу ове четири категорије опсервација добијених као резултат категоризације модела изводе се мере ваљаности класификације:

a) ***Accuracy*** – тачност. Однос тачно категоризованих опсервација и свих опсервација које су се налазиле у делу података за тестирање. Овај показатељ се најчешће наводи када се говори о успешности одређеног класификатора. Рачуна се помоћу формуле:

$$accuracy = \frac{TP+TN}{TP+TF+FP+F} \tag{10}$$

b) ***Error rate*** (*estimated misclassification rate*) – показатељ супротан од *accuracy*. Изражава однос нетачно класификованих опсервација и свих опсервација које су класификоване:

$$err = \frac{FP+F}{TP+TF+FP+F} \tag{11}$$

Ова два показатеља спадају у опште показатеље и нису довољно информативни у случајевима када је једна класа значајнија од друге. На пример, у медицини може да буде веома опасно ако пацијент има неку болест и тест је позитиван, а модел га класификује као негативан (*false negative*). Зато се рачунају и следећи показатељи:

c) ***Precision*** *– positive predicted value* (*ppp*) – као што назив говори, мери прецизност модела. Односи се на удео тачно категоризованих позитивних опсервација у односу на све опсервације које су категоризоване као позитивне:

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

d) ***Negative predicted value*** (*npp*) – показатељ који се рачуна на исти начин као *precision*, само што се односи на негативне опсервације:

$$npp = \frac{TN}{TN+} \tag{13}$$

e) ***Recall*** (*sensitivity*) – показатељ способности модела да категоризује опсервације класи од интереса, изражава комплетност модела, односно удео тачно класификованих позитивних у односу на све позитивне опсервације:

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

f) ***Specificity*** – сличан показатељ претходном, с тим што се односи на негативне опсервације:

$$Specificity = \frac{TN}{TN+FP} \tag{15}$$

g) ***F*-вредност** класификационих модела рачуна се уз помоћ показатеља *precision* и *recall*:

$$F - vrednost = \frac{2\times precision\times recall}{precision+rec} \tag{16}$$

3.6.Преглед емпиријске литературе о Наивном Бајесовом класификатору

НБК се показао као најпрецизнији у бинарним класификационим проблемима. Објашњење за ово лежи у чињеници да процена класификационог модела зависи од знака оцењене функције (*McCallum*, *A*., *Nigam*, *K*., 1998). Ово објашњава и успешност

овог модела упркос нарушености претпоставке о независности атрибута. Зато није изненађујуће да се НБК најчешће примењује у анализи сентимента, где је углавном циљ класификовати коментаре на позитивне и негативне, мада у неким случајевима постоји и трећа категорија за неутралне коментаре. За потребе компанија су свакако значајнији изразито позитивни и негативни ставови о производима/услугама јер се више употребљивих информација може извући из таквих коментара. Зато компаније неретко примењују овај модел за класификовање рецензија својих производа/услуга, као и за праћење своје репутације. На примеру поделе рецензија филмова на позитивне и негативне НБК је достигао ниво прецизности од 80% за негативне рецензије и 87% за позитивне (*Tripathy*, *A., Agrawal, A., Rath, S., K*., 2015). Ово истраживање не представља изолован случај у ком је НБК прецизнији у класификацији позитивних него негативних рецензија. Сличан резултат је добијен и приликом класификовања критика ресторана, зато је предложена модификована верзија НБК модела, која је смањила јаз између прецизности класификације позитивних и негативних рецензија за 3,6% (*Kang, H., Yoo, S. J., & Han, D*., 2012). Овај недостатак модела не би требало да представља препреку за његову употребу за обраду природног језика, с обзиром на предложену модификацију. С друге стране, у оба случаја се *Support Vector Machine* (*SVM*), линеарни класификациони модел, који функционише по принципу одређивања најбоље линеарне границе (равни) за раздвајање категорија, показао као прецизнији, што значи да је могуће да би комбиновани приступ ова два класификациона модела представљао оптимално решење.

Још једна честа примена НБК модела у класификацији текста односи се на сортирање мејлова и детекцију непожељних мејлова, односно спамова. У таквим случајевима модел се обучава већ сортираним мејловима који се разликују на основу речи које се најчешће појављују у стандардним мејловима наспрам оних које указују на тзв. џанк-мејлове. Принцип по ком функционише је сличан као у претходно наведеним примерима, с тим што се сада не траже позитивне или негативне речи, већ оне које су карактеристичне за различите типове мејлова, тако да анализа треба да буде фокусирана на друге карактеристике текстова. Филтер за категоризацију мејлова заснован на НБК показао је добре перформансе, достигнута је прецизност од 92% за класификацију џанк-мејлова, док за класификацију стандардних мејлова прецизност истог филтера износи 95% (*Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E*., 1998).

Како је оцењивање писаних одговора временски и трошковно захтевно за универзитете, али и остале институције, у последње време на популарности добија аутоматско оцењивање есеја. Ова метода такође спада у класификационе задатке где оцене представљају различите класе. Стога је могуће обучити модел да на основу већ оцењених радова и њихових карактеристика разврстава неоцењене радове по истим критеријумима и додељује оцене. Ова идеја није толико нова, колико с развојем вештачке интелигенције добија на значају у скорије време – још 2002. је рађено истраживање у ком је утврђено да НБК може с тачношћу од 80% да разврстава неоцењене студентске есеје (*Rudner, L. M., Liang*, *T*., 2002).

Област у којој се НБК показао као користан и у којој се примењује већ дуже време јесте медицина. Овај модел се најчешће користи за предвиђање ризика за развијање

одређених болести на основу фактора ризика пацијената. Када су упоређене перформансе НБК и *SVM* модела за предвиђање ризика од развијања болести јетре, *SVM* се показао као прецизнији, а НБК као бржи метод (*Vijayarani, S.*, *Dhayanand, S.*, 2015). С друге стране, модификована верзија НБК-А (пондерисани НБК) показала се као изузетно добра у детекцији рака дојке – овај модел је достигао ниво осетљивости од 99,11%, тачности од 98,54% и *specificity* од 98,25% (*Karabatak*, *M.*, 2015). Слично, примена овог класификатора за предвиђање ризика од поновног јављања или прогреса рака мозга довела је до значајних резултата, који се даље могу користити у практичне сврхе (*Kazmierska*, *J.*, *& Malicki*, *J.*, 2008).

4. Анализа аутора: Класификовање текстова на теме применом НБК модела

Предмет наше анализе је класификација новинских текстова са сајта Би-Би-Сија у пет категорија на основу тема којима се баве коришћењем НБК-а. Новински текстови припадају областима бизниса, забаве, политике, спорта и технологије. Ови подаци преузети су с платформе *Kaggle*. Анализа је спроведена у програмском језику *R*.

Цео сет података садржи 2225 новинских чланака. Како је ово релативно мањи обим података за појмове *data science*, сматрамо да је НБК погодан за класификацију, јер се овај класификатор генерално употребљава када је део података за обуку мањег обима (*Wankhade, M., Rao, A., Kulkarni, C.*, 2022). Истраживање је спроведено кроз следеће етапе:

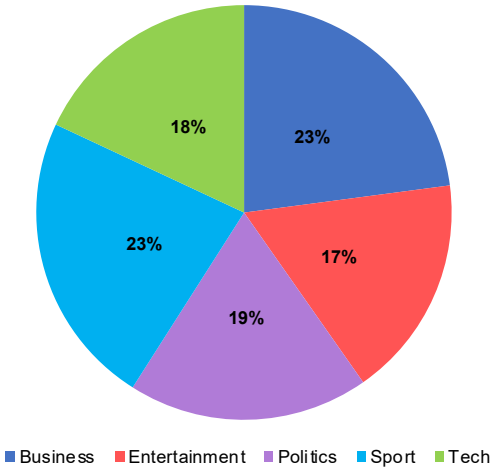
1. Анализа, учитавање и припрема података, која подразумева све кораке обраде текстуалних података представљене у другом поглављу;
2. Извлачење карактеристика из текстова применом методе *TF-IDF*;
3. Подела података на део за обуку и део за тестирање;
4. Примену НБК-а на делу података намењеном за обуку модела;
5. Тестирање модела, што подразумева класификацију текстова из дела података намењених тестирању модела;
6. Испитивање карактеристика модела уз помоћ мера ваљаности класификационих модела које су представљене у претходном поглављу;
7. Утврђивање економског и маркетиншког значаја добијеног модела.

4.1. Анализа података

Како су у новинарству увек неке теме актуелније од осталих, очекивано је да у посматраном периоду није објављен једнак број текстова из свих пет области. Тако у овом корпусу текстова има 510 чланака из бизниса, 386 из области забаве, 417 из политике, 511 из спорта и 401 текст који покрива област технологије. Посматрајући Графикон 1, долазимо до закључка да је овај сет података приближно уравнотежен, с

тим што су најзаступљеније теме бизнис и спорт. Не очекујемо да ове мале разлике у релативном учешћу тема доведу до пристрасности приликом обуке и тестирања модела.

Графикон 1. **Учешће различитих тема у збирци текстова**



4.2. Припрема података

Пошто НБК спада у моделе надгледаног машинског учења, неопходно је, након учитавања базе података, обележити категорије текстова. Када имамо учитану базу с јасно обележеним категоријама, можемо да пређемо на обраду самих текстова.

У досадашњем раду је већ објашњено да текстуални подаци у свом сировом облику нису спремни за моделирање. Припрема података се обавља у три фазе: токенизација, избацивање стоп-речи и стеминг. Иако је из Табеле 1, приказане у другом поглављу, очигледно да се лематизацијом добијају бољи резултати, због комплексности те методе у овом раду примењен је стеминг.

Први корак припреме података је токенизација. Као што је већ објашњено, токени могу бити засебне речи, сложенице (односно *n-gram*-ови) или реченице. За потребе тематске класификације токени од интереса су речи. Овај поступак знатно повећава број опсервација – са 2225 (чланака) на 861.355 (речи). Међутим, у следећем кораку тај број се смањује, јер за класификацију текстова нису потребне оне речи које се често појављују а нису информативне – стоп-речи. Избацивањем стоп-речи број података смањен је са 861.355 на 268.178, што нам указује на значај искључивања ове врсте речи. Последњи корак припреме података је стеминг, који за резултат има речи у основном облику, односно са уклоњеним суфиксима.

Да бисмо илустровали значај овакве обраде текстуалних података, у Табели 2 приказали смо првих 30 опсервација насумично одабраног текста у четири фазе припреме. Прва фаза подразумева најједноставније извлачење речи из текстова где се под појмом речи подразумева све што се налази између два бела поља (*white space*). Та врста раздвајања текстова на речи примењена је само на овом малом узорку, с обзиром на то да токенизацијом већ добијамо текст раздвојен на речи, које су, притом, и смањене

на мала слова, при чему су и беспотребни карактери аутоматски искључени. У овом случају смо приказали и ту фазу само ради поређења корака за обраду текстуалних података.

Табела 3. **Фазе припреме текстова**

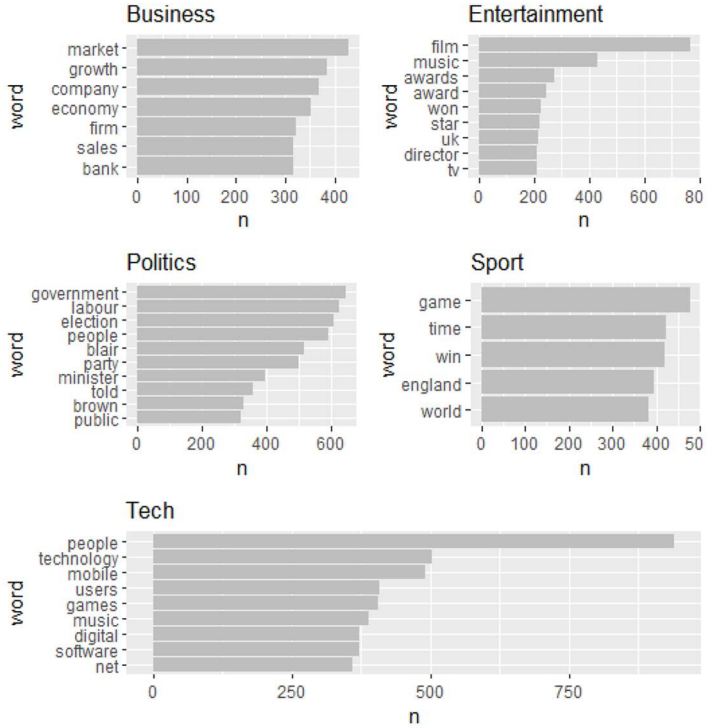
Words	*Tokens*	*No stop words*	*Stemming*
Lennon	*lennon*	*lennon*	*lennon*
brands	*brands*	*brands*	*brand*
Rangers	*rangers*	*rangers*	*ranger*
favourites\n\nCeltic's	*favourites*	*favourites*	*favourit*
Neil	*celtic's*	*celtic's*	*celtic'*
Lennon	*neil*	*neil*	*neil*
admits	*lennon*	*lennon*	*lennon*
Rangers	*admits*	*admits*	*admit*
could	*rangers*	*rangers*	*ranger*
be	*could*	*considered*	*consid*
considered	*be*	*slight*	*slight*
\"slight	*considered*	*favourites*	*favourit*
favourites\"	*slight*	*firm*	*firm*
for	*favourites*	*cis*	*ci*
the	*for*	*cup*	*cup*
Old	*the*	*clash*	*clash*
Firm	*old*	*insists*	*insist*
CIS	*firm*	*win*	*win*
Cup	*cis*	*lennon*	*lennon*
clash,	*cup*	*concedes*	*conced*
but	*clash*	*rangers*	*ranger*
insists	*but*	*form*	*form*
his	*insists*	*moment*	*moment*
side	*his*	*failed*	*fail*
can	*side*	*beat*	*beat*
still	*can*	*celtic*	*celtic*
win.\n\nLennon	*still*	*meetings*	*meet*
concedes	*win*	*rangers*	*ranger*
Rangers	*lennon*	*run*	*run*
are	*concedes*	*recent*	*recent*

Када упоредимо прву и другу колону, јасно је због чега прва фаза није примењена на свим подацима. Токенизација за резултат има целе речи смањене на мала слова, без тачака, знакова навода и карактера који се односе на нови ред (*\n*), односно токенизација је логичан први корак обраде текстова. У следећем кораку су избачене речи попут: *could*, *be*, *but*, *his* и сл. које нису од значаја за класификацију текста. И коначно, након стеминга имамо речи које ће се користити у даљој анализи. Очигледно је да резултат стеминга није идеалан – на пример реч *considered* је скраћена на форму *consid*, што нема никакво значење. Али у већем броју случајева је овај корак значајан – нама је потребно да алгоритам речи *admits* и *admit* третира на исти начин јер имају исто значење. Поред тога, непрецизност стеминга се може занемарити с обзиром на број речи који се анализира.

4.3. Извлачење и одабир карактеристика текстова

Да би се текст класификовао према темама, потребни су атрибути текстова који се користе као инпути модела. Интуитивно је јасно да су најбитнији атрибути речи које се најчешће понављају кроз чланке у оквиру једне теме. Пре извлачења карактеристика из текстова, потребно је прво прегледати које су те речи. На Слици 7 приказане су речи које се најчешће јављају у новинским чланцима из свих пет категорија. Резултати су у складу са очекивањима: у категорији бизниса најчешће се појављују речи ***тржиште***, ***раст***, ***компанија***, док су у категорији забаве то речи ***музика***, ***филм*** и ***награде***. Оно што је изненађујуће да, иако су с речи скинути суфикси применом стеминга, речи *awards* и *award* третирају се као различите иако имају исто значење. Овај проблем може да буде занемарен ако су у питању изоловани случајеви, али, ако се често јавља, може да се деси да доведе до предимензионисаности података, што може неповољно утицати на оцењивање модела. Још један потенцијални проблем тиче се категорија политике и технологије – у обе категорије међу најфреквентнијим речима налази се реч *људи*. Ово би могло да доведе до тога да модел погрешно додели тексту из области политике категорију технологије.

Графикон 2. **Најчешће речи по категоријама**



Као што је већ наглашено у другом делу овог рада, саме фреквенције речи нису толико значајне колико релативно учешће одређених речи у тексту у односу на целу збирку текстова. Зато смо за извлачење карактеристика из текстуалних података применили методу *TF-IDF*. На овај начин је свакој речи додељена *TF-IDF* вредност у интервалу од 0 до 1.

Након тога, ради смањивања броја опсервација, искључили смо све речи које се појављују у мање од десет текстова у целој збирци текстова. Тако је укупан број опсервација смањен са 268.178 на 56.529, чиме смо смањили проблем предимензионисаности података.

4.4. Обука и тестирање модела

Пре обуке модела, неопходно је прво поделити податке на део за обуку и део за тестирање, што је стандардна процедура код свих надгледаних метода машинског учења. Како су подаци обрађени на нивоу речи, а не на нивоу текстова, неопходно је да их раздвојимо по текстовима да не би дошло до преклапања, односно да не бисмо дошли до ситуација да је модел обучен на одређеном броју речи из једног текста, а предвиђа на осталим речима из истог текста. Након што смо податке раздвојили по текстовима, можемо их поделити у складу са уобичајеном праксом да се 70% података искористи за обуку модела, а преосталих 30% за тестирање.

Следећи корак је обука НБК на делу података за обуку. Зависна променљива је категорија, а независне су све остале променљиве (речи и њихове *TF-IDF* вредности). Поред тога, задајемо вредност 1 Лапласовом фактору заглађивања како бисмо избегли нулте вероватноће.

На делу података који смо издвојили за предвиђање (30% опсервација) проверићемо предиктивну моћ овог модела. Резултати тестирања илустровани су у Табели 3, где је приказан узорак од 11 речи, додељена категорија свакој од ових речи, оцењене вероватноће да реч припадне свакој од пет категорија и одговарајуће односно категорије којима текст из ког је одређена реч извучена заправо припада.

Табела 4. **Класификација речи на основу НБК-а**

Реч	Додељена категорија	Вероватноће					Одговарајућа категорија
		Business	*Entertainment*	*Politics*	*Sport*	*Tech*	
parti	*Politics*	0,00	0,00	0,99	0,01	0,00	*Politics*
time	*Sport*	0,00	0,00	0,00	1,00	0,00	*Sport*
game	*Tech*	0,00	0,00	0,00	0,00	1,00	*Tech*
sale	*Entertainment*	0,00	1,00	0,00	0,00	0,00	*Businesss*
film	*Entertainment*	0,00	1,00	0,00	0,00	0,00	*Entertainment*
govern	*Entertainment*	0,00	0,96	0,00	0,00	0,04	*Politics*
product	*Businesss*	1,00	0,00	0,00	0,00	0,00	*Businesss*
campaign	*Entertainment*	0,24	0,76	0,00	0,00	0,00	*Entertainment*
win	*Sport*	0,00	0,00	0,39	0,61	0,00	*Politics*
futur	*Tech*	0,00	0,00	0,00	0,01	0,99	*Tech*
start	*Tech*	0,00	0,00	0,00	0,01	0,99	*Tech*

4.5. Испитивање карактеристика модела

Да бисмо проверили карактеристике модела, изводимо матрицу класификације, на основу које даље рачунамо мере ваљаности класификације. Матрица класификације и табела с мерама ваљаности класификације за добијени модел приказане су испод.

Прегледом матрице класификације изводи се закључак да је НБК добро класирао речи по категоријама. Мера тачности овог модела (*accuracy*) износи **95,77%**, а удео погрешно класификованих опсервација (*estimated misclassification rate*) – **4,23%**.

Табела 5. **Матрица класификације за добијени модел**

		Одговарајућа категорија				
		Business	*Entertainment*	*Politics*	*Sport*	*Tech*
Додељена категорија	*Business*	***4174***	*131*	*0*	*0*	*0*
	Entertainment	*94*	***1774***	*68*	*1*	*0*
	Politics	*0*	*91*	***3423***	*115*	*0*
	Sport	*0*	*0*	*64*	***3158***	*85*
	Tech	*0*	*0*	*2*	*52*	***3407***

Табела 6. **Мере ваљаности класификације**

	Business	*Entertainment*	*Politics*	*Sport*	*Tech*
Sensitivity	0,978	0,889	0,962	0,95	0,976
Specificity	0,989	0,989	0,984	0,989	0,996
Pos. Predicted Value	0,97	0,916	0,943	0,955	0,984
Neg. Predicted Value	0,992	0,985	0,99	0,987	0,993
F - Measure	0,974	0,902	0,952	0,952	0,980

Из Табеле 5 видимо да су најпрецизније класификоване речи из области технологије. Објашњење за ово лежи у чињеници да је ова област најмање повезана са остале четири области. С друге стране, најмања прецизност класификације остварена је у категорији забаве, што је донекле и очекивано, јер је најмањи број текстова из ове области, што је резултирало благом пристрасношћу ка осталим категоријама. Поред тога, највећи број погрешно класификованих речи јесу речи из категорије забаве које су додељене категорији бизниса (од 1996 речи које припадају категорији забаве, 131 реч додељена је категорији бизниса). Слично, од 4268 речи из категорије бизниса, 94 су оцењене да припадају категорији забаве. Ово указује на преклапање кључних речи из ове две области, што је такође утицало на прецизност класификације речи из категорије забаве.

Највећа осетљивост, односно способност модела да додели опсервације одговарајућим категоријама, остварена је за категорију бизниса (97,8%), док је најнижа вредност овог показатеља остварена у категорији забаве (88,9%). Иако је највише грешака остварено у овој категорији, на основу табела 6 и 7 можемо да закључимо да је НБК показао добре перформансе у класификацији речи.

Табела 7. **Матрица класификације на нивоу текстова**

		Одговарајућа категорија				
		Business	*Entertainment*	*Politics*	*Sport*	*Tech*
Додељена категорија	*Business*	***160***	*8*	*0*	*0*	*0*
	Entertainment	*0*	***97***	*0*	*0*	*0*
	Politics	*0*	*3*	***118***	*2*	*0*
	Sport	*0*	*0*	*1*	***162***	*0*
	Tech	***0***	*0*	*0*	*0*	***121***

Табела 8. **Мере ваљаности класификације на нивоу текстова**

	Business	*Entertainment*	*Politics*	*Sport*	*Tech*
Sensitivity	1	0,9	0,992	0,99	1
Specificity	0,984	1	0,991	0,998	1
Pos. Predicted Value	0,952	1	0,959	0,994	1
Neg. Predicted Value	1	0,981	0,998	0,996	1
F - Measure	0,975	0,947	0,975	0,991	1,000

Међутим, циљ овог истраживања је класификација на нивоу текстова, а не на нивоу речи. Зато је потребно поновно спајање речи у текстове и додељивање оне категорије којој највећи број речи из тог текста припада. Иако за класификацију на нивоу текста нисмо креирали засебан модел, у табелама испод приказане су матрица и мере ваљаности класификације за текстове класификоване на овај начин.

Мера тачности класификације на нивоу текстова виша је у односу на класификацију на нивоу речи и износи 97,92%. Такође, добијена је мања грешка класификације за 2,13 процентних поена, односно 2,1%. Такво унапређење класификације постигнуто је јер су прво речи класификоване по категоријама, па су затим текстови класификовани тако што су сумиране класе речи које се налазе у тим текстовима, односно тексту је додељена она категорија којој највећи број речи припада. Класификација текстова је упросечена класификација речи, па је и очекивано да буде тачнија.

5. Закључак

У овом раду приказали смо тематску класификацију 2225 новинских чланака са сајта Би-Би-Сија у пет категорија којима ови текстови припадају употребом НБК. Модел је обучен на већ обележеном делу података, а тестиран је на осталом, необележеном, делу података, што је стандардна пракса за надгледане методе машинског учења. На

овај начин се најједноставније стиче увид о карактеристикама, предностима и недостацима модела.

Најзначајније предности НБК-а јесу то што је једноставан, робустан и што не заузима много меморијског простора. Фаза обуке овог модела је релативно спорија, док је фаза предвиђања бржа. Још један недостатак НБК-а огледа се у томе што се слабије адаптира новим подацима – овај модел се обучава на основу доступног тренинг сета података, па је за евентуално додавање нових података потребно поновно постављање модела. Поред тога, „наивна" претпоставка о независности атрибута не утиче у знатној мери на прецизност предвиђања.

Иако је сет података над којима је спроведена анализа релативно мали, резултати нашег истраживања су корисни јер потврђују предиктивну моћ НБК-а упркос поједностављености претпоставки на којима је заснован – тачност добијеног модела износи 97,9%, односно грешка класификације износи 2,1%. Осим тачности, додатна предност овог модела јесте што се, управо захваљујући својим једноставним претпоставкама, може лако прилагодити различитим врстама података и различитим областима.

Како су за текстуалну анализу у економији најзначајнији извор података новински текстови, у овом раду смо на том извору података и спровели анализу, те утврдили да им се НБК може веома ефикасно прилагодити. Такав увид нам отвара простор за даљу примену НБК-а у анализи новинских текстова у економским истраживањима, што може да буде веома практично, с обзиром на доступност и ажурност ових извора података. У складу с тим, економски новински текстови и сет података на српском језику биће предмет наших будућих анализа.

Литература

Ashwin, J., Kalamara, E., Saiz, L. 2021. Working Paper Series Nowcasting euro area GDP with news sentiment: a tale of two crises. ECB Working Paper No. 2021/2616.

Baker, S. R., Bloom, N., & Davis, S. J. 2016. »Measuring economic policy uncertainty«. The quarterly journal of economics, 131(4), 1593–1636.

Bholat D., Hans, S., Santos, P., & Schonhardt-Bailey, C. 2015. Text mining for central banks. Handbooks, Centre for Central Banking Studies, Bank of England, number 33

Claudia, E., Scott B. 2022. Text Analysis with R. https://cengel.github.io/R-text-analysis/

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. 2020. »Making text count: economic forecasting using newspaper text«. Staff Working Paper No. 865., Bank of England.

Kang, H., Yoo, S. J., & Han, D. 2012. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications, 39(5), 6000–6010.

Karabatak, M. 2015. A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, 32–36.

Kazmierska, J., & Malicki, J. 2008. Application of the Naïve Bayesian Classifier to optimize treatment decisions. Radiotherapy and Oncology, 86(2), 211–216.

McCallum, A., & Nigam, K. 1998. A comparison of event models for naive bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, No. 1, pp. 41–48).

Raschka, S. 2014. Naive bayes and text classification i-introduction and theory. arXiv preprint arXiv:1410.5329. https://arxiv.org/abs/1410.5329

Rish, I. 2001. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41–46).

Rudner, L. M., & Liang, T. 2002. Automated essay scoring using Bayes' theorem. The Journal of Technology, Learning and Assessment, 1(2).

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. 1998. A Bayesian approach to filtering junk e-mail . In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98–105).

Tripathy, A., Agrawal, A., Rath, S., K. 2015. Classification of sentimental reviews using machine learning techniques. Procedia Computer Science 57:821–829.

Vijayarani, S., & Dhayanand, S. 2015. Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), 816–820.

Wankhade, M., Rao, A., Kulkarni, C. 2022. A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review 55:5731–5780.

Ђукић, М. 2022. Процена инфлаторних притисака путем анализе новинских текстова. Зборник радова Народне банке Србије, III.

Ђукић, М. 2024. Тематска класификација економских новинских чланака на језику са високом флексијом – пример Србије. Зборник радова Народне банке Србије, VI.

Извор података: BBC Full Text Document Classification. Преузето 27. апр. 24. са: https://www.kaggle.com/datasets/shivamkushwaha/bbc-full-text-document-classification