Истраживачко-развојни
институт за вештачку
интелигенцију Србије

СИСТЕМ ЗА АУТОМАТСКО ПРЕПОЗНАВАЊЕ ТЕМА У КОРИСНИЧКИМ МЕЈЛОВИМА У ЦИЉУ ПОБОЉШАЊА РАДА КОРИСНИЧКЕ СЛУЖБЕ

ТЕХНИЧКО РЕШЕЊЕ

2025.
НОВИ САД

СИСТЕМ ЗА АУТОМАТСКО ПРЕПОЗНАВАЊЕ ТЕМА У КОРИСНИЧКИМ МЕЈЛОВИМА У ЦИЉУ ПОБОЉШАЊА РАДА КОРИСНИЧКЕ СЛУЖБЕ

Основни подаци техничког решења

Назив техничког решења:	Систем за аутоматско препознавање тема у корисничким мејловима у циљу побољшања рада корисничке службе.
Кључне речи од значаја за техничко решење:	вештачка интелигенција, ненадгледано дубоко учење, моделовање тема, *BERTopic*
Подтип решења:	M82 – Ново техничко решење примењено на националном нивоу
Реализатор:	Истраживачко-развојни институт за вештачку интелигенцију Србије (ИВИ), Нови Сад
Област:	Вештачка интелигенција
Пројекат:	Техничко решење је резултат пројекта Истраживачко-развојног института за вештачку интелигенцију Србије под називом „Развој модела машинског учења за анализу текста“. Пројекат је реализован за компанију Телеком Србија током 2024. године.
Аутори:	Бојана Башарагин, Дарија Медвецки, Горана Гојић, Милена Опарница, Драгиша Мишковић
Година комплетирања решења:	2024.
Почетак примене:	Јун 2024. године
Корисници:	Телеком Србија а.д, Булевар Арсенија Чарнојевића 99б, Београд
Кратак опис:	Систем за аутоматско препознавање тема у корисничким мејловима је нов производ развијен за потребе компаније Телеком Србија а.д. са циљем да се унапреди рад корисничке службе ове компаније. Систем се заснива на примени напредног приступа за моделовање тема са циљем аутоматске анализе корисничких мејлова на српском језику и

	њихове класификације у једну од предефинисаних тема које одговарају пословним потребама компаније. Развијени приступ исправно одређује тему за 96% мејлова без интервенције оператера. Употребом овог система стварају се предуслови за прелазак са ручног рутирања корисничких мејлова које обавља оператер, на потпуно аутоматизовано рутирање. Аутоматизација процеса доприноси остварењу временских и финансијских уштеда за компанију, као и увећању задовољства корисника услед бржег времена одзива корисничке службе на корисничке захтеве.

Опис техничког решења

1. Опис проблема и стање у свету

У данашње време, корисничке службе компанија као што су банке и телекомуникациони оператери примају велики број мејлова који се односе на различите типове питања, проблема или сугестија. Ручна обрада велике количине података је захтевна и подразумева веће финансијске издатке. Због тога сам процес обраде корисничких захтева обично дуже траје, не постоји приоритизација, што се на крају негативно одражава и на корисничко искуство. Једно од решења за овај проблем је аутоматизовање доделе тема корисничким мејловима. Кроз потпуно аутоматску доделу тема пристиглим мејловима стварају се предуслови за бржи одговор корисницима, унапређује се корисничко искуство [1] и смањују се трошкови компаније [2]. Аутоматски систем смањује временско оптерећење запослених, јер се мејлови преусмеравају директно специјализованим тимовима и одељењима. Самим тим, оператери у корисничком центру имају више времена за анализу озбиљнијих или специфичнијих ситуација. Такође, додела тема на основу садржаја мејлова кроз алате за обраду природног језика и машинско учење омогућава компанијама да открију уобичајене обрасце проблема и унапређују своје услуге.

Моделовање тема представља технику ненадгледаног учења која идентификује обрасце и тематске структуре у тексту без претходног ручног означавања. Ова техника се показала учинковитом како на дугом [3], тако и на кратком тексту на српском језику [4]. До скоро су се углавном примењивале традиционалне методе као што су Латентна Дирихлеова алокација (енгл. *Latent Dirichlet allocation - LDA*) [5] и ненегативна факторизација матрице (енгл. *Non-negative matrix factorization*

- *NMF*) [6]. Ове методе не узимају у обзир семантичке односе међу речима, већ текст посматрају као ***врећу речи*** (енгл. *bag-of-words*). ЛДА је метода која на основу дистрибуција вероватноћа класификује речи и документа и на тај начин генерише теме, док се НФМ метода моделовања тема заснива на линеарној алгребри, тј. разлагању матрице која представља речи и документа. Од 2022. године, када се појавила, *BERTopic* [7] архитектура се издваја у односу на остале методе. За разлику од конвенционалних метода које користе статичке векторске приказе текста као основ за кластеровање докумената, *BERTopic* користи претренирани велики језички модел за креирање динамичких векторских приказа, чиме се узимају у обзир контекстуалне зависности међу речима у тексту. Тиме се, на пример, решава проблем различитог значења истих речи у другачијим контекстима и моделу омогућава боље разумевање сложености природног језика. Због тога, *BERTopic* се показао као изузетно ефикасна техника за моделовање тема, посебно у раду са краћим документима какви су мејлови [8], [9], [10].

У недавно спроведеним истраживањима, *BERTopic* архитектура се показала као напредно решење и за српски језик, при чему је утврђено да постиже боље резултате у односу на друге методе за моделовање тема [3], [4]. Систем који је описан у овом техничком решењу такође се заснива на *BERTopic* архитектури. Модел који је реализован у овом техничком решењу додељује теме аутоматски са тачношћу од 96% (Ф1 мера), што значајно превазилази друге методе које обично постижу Ф1 од приближно 91% [11]. Ова метода може се применити на било који језик са минималним изменама, али има посебан значај за флективне језике.

2. Допринос

У овом техничко решењу приказан је систем за аутоматско препознавање тема у долазним корисничким мејловима. Главни допринос овог система представља унапређење рада корисничке службе кроз ефикаснију анализу велике количине долазних мејлова. Конкретно, овај систем омогућава:

1. Аутоматско усмеравање мејлова релевантним секторима без потребе за претходним читањем и ручном категоризацијом;
2. Убрзавање процеса одговарања на корисничке захтеве, чиме се побољшава корисничко искуство;
3. Оптимизацију ресурса у секторима корисничке подршке, који се сада могу усмерити само на питања и проблеме намењене њима;
4. Анализу рада корисничке службе и детектовање кључних проблема у задатом периоду.

Додатни доприноси тимовима компаније ван корисничке службе укључују:

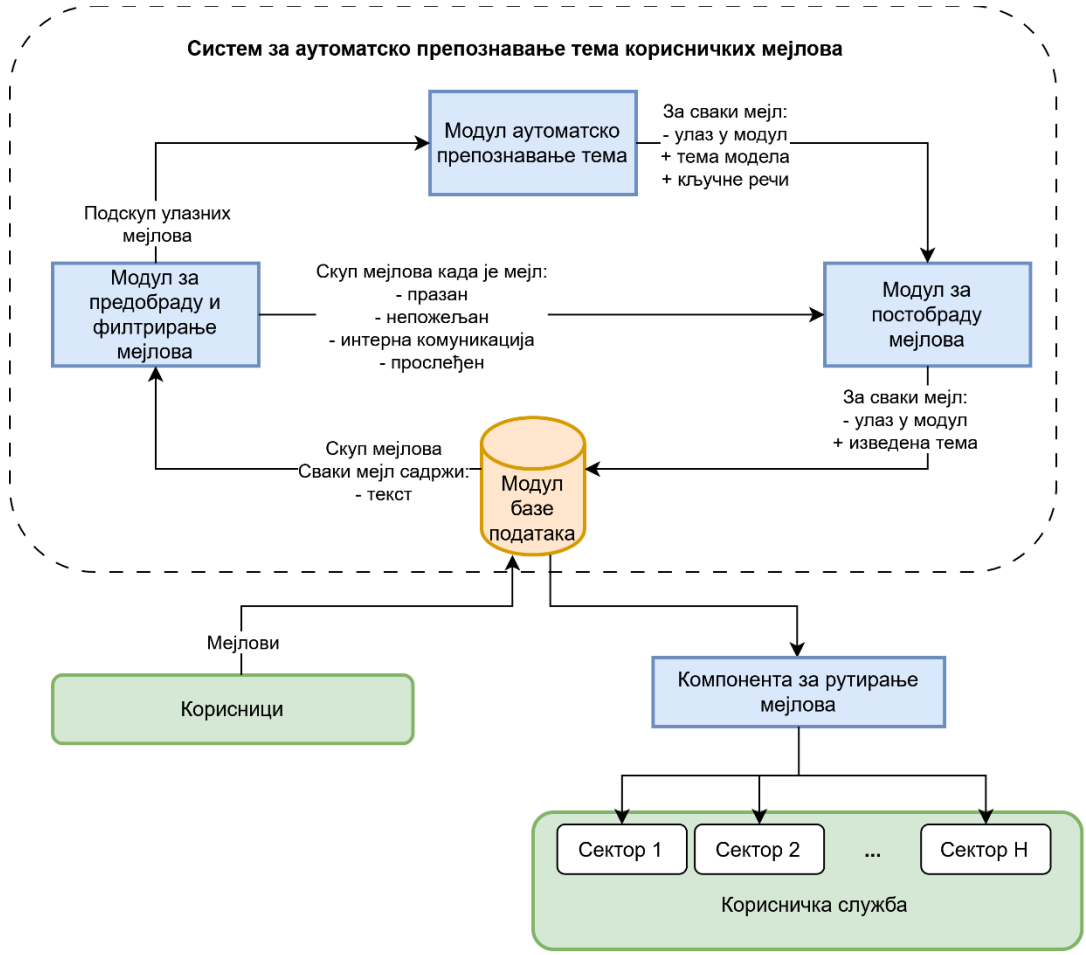
1. Увид у најчешће пријављене проблеме, што омогућава њихово благовремено решавање и смањење незадовољства корисника.

2. Увид у систематичну аналитику типова мејлова на месечном нивоу кроз извештавање, чиме се доприноси оптимизацији интерних процеса, унапређењу квалитета услуге и доношењу стратешких одлука.

3. Објашњење суштине техничког решења и детаљан опис са карактеристикама

Ово техничко решење представља систем за аутоматско одређивање тема корисничких мејлова. Систем је имплементиран као модуларан софтвер, чији су модули и њихова међусобна интеракција приказани на слици Слика 1. Софтвер се покреће периодично и учитава све мејлове пристигле од последњег покретања софтвера из базе података мејлова. Учитани мејлови пролазе процес предобраде, где се филтрирају и трансформишу за потребе обраде сукцесивним модулима. Мејлови који су празни, непожељни (енгл. *spam*) или нису кориснички мејлови (нпр. интерна комуникација или прослеђени мејлови) прослеђују се директно модулу за постобраду, док се остатак мејлова прослеђује модулу за аутоматско препознавање тема. По излазу из овог модула, сваком мејлу се придружује аутоматски одређена тема, односно тема модела, са кључним речима, и мејлови се даље прослеђују модулу за постобраду. Овај модул на основу информације о типу мејла, теми модела и кључним речима тамо где постоје, одређује такозвану изведену тему мејла, која је прилагођена потребама корисничке службе компаније. Циклус обраде мејлова се завршава уписом података о теми модела, кључним речима и изведеној теми у базу података за сваки од обрађених мејлова. Софтверски модули су детаљно описани у поглављима 3.1, 3.2, 3.3 и 3.4.

Примарни корисник система је корисничка служба компаније Телеком Србија. Екстерна компонента система за рутирање мејлова имплементирана од стране компаније користи податке о изведеним темама како би рутирала мејлове у одговарајуће секторе корисничке службе задужене за различите типове корисничких захтева и проблема.



Слика 1. Приказ система за аутоматско препознавање тема корисничких мејлова.

3.1. Модул базе података

Модул базе података обухвата базу података и софтверски подмодул за комуникацију осталих модула са базом података.

Као база података користи се Хадуп (енгл. *Hadoop*) дистрибуирано складиште података, док се за приступ и обраду података користи *IBM DB2 Big SQL,* софтвер који омогућава приступ Хадуп складишту преко интерфејса за релационе базе података. За сваки мејл се у бази података чувају улазни и излазни подаци. Улазни подаци су текст наслова и тела мејла (чувају се у одвојеним пољима), време пријема и формат мејла, *HTML* или обичан текст (енгл. *plain text*). Ови подаци се по учитавању користе у модулима за предобраду, аутоматску доделу тема и постобраду. Излазни подаци обухватају тему модела, кључне речи за тему модела и изведену тему.

Подмодул за комуникацију са базом података имплементиран је у програмском језику Пајтон коришћењем парадигме објектно-оријентисаног мапирања објеката на податке из базе. Подмодул садржи функционалности за учитавање улазних података са могућношћу филтрирања према датуму пријема и формату мејла, као и чување излазних података за селектоване групе мејлова.

3.2. Модул за предобраду и филтрирање корисничких мејлова

Пре употребе модела, кориснички мејлови пролазе процес предобраде и филтрирања.

Мејлови у бази података могу бити сачувани у два формата: обичан текст и *HTML* формат. Мејлови који су у формату обичног текста директно улазе у процес предобраде и филтрирања. С обзиром на сложеност *HTML* формата и присуство различитих структура, за екстракцију релевантног садржаја имплементирано је шест типова парсера, прилагођених различитим варијантама *HTML* кодирања и структурирања текста. Након парсирања, садржај мејлова у *HTML* формату улази у исти процес предобраде и филтрирања.

У преодбраду појединачно улазе и наслов и тело мејла који се након тога спајају како би се добио целовит текст мејла. Одлука да се наслов и тело мејла споје и користе као целина проистиче из запажања да у појединим случајевима пошиљалац мејла сав информативни садржај наводи у наслову, док тело мејла оставља празним, те би у овим случајевима коришћење само једног поља (нпр. тела мејла) било неинформативно или недовољно за одређивање теме мејла.

Предобрада подразумева следеће трансформације корисничког мејла:

1. **Стандардизација писма** – мејлови написани ћириличним писмом преводе се на латинично писмо употребом софтверског пакета *srtools* [12].
2. **Пребацивање у мала слова** – текстови мејла се трансформишу у мала слова.
3. **Уклањање интерпункције, специјалних карактера и бројева** – наведени карактери и симболи се уклањају јер не доприносе одређивању тема. Једини изузетак су бројеви који су део речи попут *5G*, који су значајни за делатност којом се компанија бави.
4. **Уклањање потписа, упозорења о одрицању од одговорности и отпоздрава** – исти отпоздрави (нпр. *све најбоље*, *с поштовањем*), потписи и упозорења о одрицању од одговорности се јављају у великом броју мејлова различитих тема. Како ови сегменти не носе релевантне информације за одређивање теме, уклањају се да би се избегла потенцијално погрешна додела теме од стране модела.
5. **Уклањање делова текста који означавају да је мејл прослеђен или представља одговор на претходну преписку** – овај текст је уклоњен из истог разлога као и у кораку 4.

Филтрирање мејлова одређује који ће мејлови бити прослеђени моделу ради аутоматског одређивања теме и заснива се на два типа критеријума.

Први критеријум обухвата идентификацију мејлова који су на енглеском језику, непожељни или празни. Како би се ови мејлови детектовали, након предобраде њиховог наслова и тела, врши се провера да ли су оба поља празна. У случају да јесу, овакви мејлови аутоматски добијају посебну ознаку и директно улазе у фазу постобраде, прескачући аутоматску доделу тема од стране модела (видети поглавље 3.4). Уколико наслов или тело мејла нису празни, даље се проверава језик мејла. Уколико је утврђено да је у питању енглески језик, мејлу се додељује иста ознака као и у случају празног мејла. Разлог за ово филтрирање произилази из ранијих анализа, које су показале да мејлови на енглеском језику најчешће представљају непожељне мејлове. За детекцију језика мејла коришћен је софтверски пакет *langID* [13].

Други критеријум филтрирања односи се на идентификацију интерне преписке, засновану на дефинисању листе интерних мејл адреса, која се може редефинисати у складу са потребама компаније. Циљ увођења овог филтера је да се осигура да у анализу проблема корисника не улазе мејлови запослених. Поред тога, за мејлове запослених нема потребе за аутоматским усмеравањем. Такви мејлови добијају ознаку за интерну преписку и на основу ње се директно обрађују у фази постобраде, слично као и код првог филтера.

Филтрирани и предобрађени кориснички мејлови представљају улаз за модул за аутоматску доделу тема.

3.3. Модул за аутоматско препознавање тема

Модул за аутоматско препознавање тема, као централни елемент овог система, заснован је на савременој *BERTopic* архитектури [7] за моделовање тема. Једна од кључних карактеристика ове архитектуре је примена ненадгледаног дубоког машинског учења којом се елиминише потреба за постојањем обележеног скупа мејлова за обуку, чиме се значајно смањују трошкови и време потребно за имплементацију метода у поређењу са методима базираним на надгледаном учењу. Значајна предност *BERTopic* архитектуре огледа се у њеној модуларности, која омогућава флексибилно прилагођавање сваког слоја у процесу, у складу са својствима скупа података за обуку и специфичним потребама анализе. Основна идеја је да се текстуални подаци најпре трансформишу у векторске репрезентације, које се затим групишу на основу њихове међусобне удаљености, при чему мања удаљеност указује на већу семантичку сличност. На основу формираних група је могуће изоловати теме које се појављују у самим текстовима. Након што је модел обучен, добија се скуп тема са кључним речима које су одређене на основу садржаја скупа за обуку и које уједно најбоље репрезентују сваку од тема.

Добијени обучени модел, прилагођен специфичностима и захтевима компаније, интегрисан је у систем са улогом аутоматске доделе једне од претходно научених тема новопристиглим мејловима. Тема се новом мејлу одређује на

основу сличности његове векторске репрезентације са векторским репрезентацијама постојећих тема модела.

У наредном потпоглављу биће детаљно приказан процес обуке модела, који обухвата опис скупа за обуку, појединачних слојева архитектуре и њихових улога, као и финалног излаза модела.

3.3.1 Развој модела за аутоматско препознавање тема

Припрема скупа података за обуку модела

За развој модела коришћен је скуп података који је обезбедила служба за подршку корисницима компаније Телеком Србија. У складу са прописима о заштити података и приватности корисника, сви мејлови су достављени у анонимизованој форми, где су лични подаци корисника замењени ознакама које указују на категорију уклоњеног податка (нпр. *PERSON, LOCATION, ORGANIZATION*). Овако припремљен скуп за обуку је у процесу предобраде прошао кроз исте фазе као и новопристигли мејлови, описане у поглављу 3.2, уз одређене додатне кораке у циљу постизања што бољег квалитета скупа. Примењене су следеће трансформације:

1. **Уклањање дуплираних мејлова** – Редунданса у скупу за обуку може негативно утицати на својство генерализације модела [14] због чега се дупликати мејлова уклањају из скупа података.
2. **Стандардизација писма** – Мејлови у скупу за обуку написани ћириличним писмом преведени су на латинично писмо. На овај начин је обезбеђено да модел не учи различите векторске репрезентације за исте појмове када су они написани различитим писмима.
3. **Пребацивање у мала слова** – Ова трансформација је примењена како би се спречило да речи истог значења, а различито записане, имају различите векторске репрезентације и тиме утичу на креирање тема.
4. **Уклањање интерпункције, специјалних карактера и бројева** – Наведени елементи текста, осим бројева који су део речи, уклоњени су јер не доприносе креирању тема.
5. **Уклањање потписа, упозорења о одрицању од одговорности и отпоздрава** – Ова трансформација је вршена јер ови елементи не доприносе процесу моделовања тема, а њихова учесталост у овом типу комуникације може довести и до формирања засебних, нерелевантних тема.
6. **Уклањање делова текста који означавају да је мејл прослеђен или представљају одговор на претходну преписку** – Овај текст је уклоњен јер не доприноси процесу моделовања тема.

7. **Уклањање анонимизацијских ознака** – Ове ознаке се уклањају с обзиром на то да се јављају у свим мејловима, без обзира на тему на коју се односе, а не садрже релевантне информације за процес моделовања тема.
8. **Уклањање кратких мејлова** – Након што су мејлови прошли претходне кораке предобраде, кратки мејлови се уклањају јер не садрже довољно информативних података за одређивање теме. За овде представљен скуп података емпиријски је утврђена дужина од три речи, као минимална дужина коју мејл мора задовољити да не би био уклоњен из скупа података.
9. **Уклањање дугачких мејлова** – Услед техничког ограничења првог слоја *BERTopic* модела, који при трансформисању мејла у његову векторску репрезентацију обрађује улазе дужине до 128 токена, на самом крају процеса предобраде уклоњени су мејлови који премашују тај лимит. Главни разлог за то је што би присуство пресечених мејлова у скупу за обуку могло негативно утицати на квалитет добијених тема.

По завршетку претходних корака предобраде, скуп је пречишћен на основу језика мејлова, с циљем да се задрже само они написани на српском језику. За сваки мејл из скупа података језик је одређен употребом софтверског пакета *langID*, и уколико то није био српски, мејл је уклоњен. Како би се обезбедио висок квалитет скупа података, сви преостали мејлови су додатно ручно прегледани и затим су уклоњени преостали мејлови који представљају аутоматске одговоре и остали непожељни мејлови.

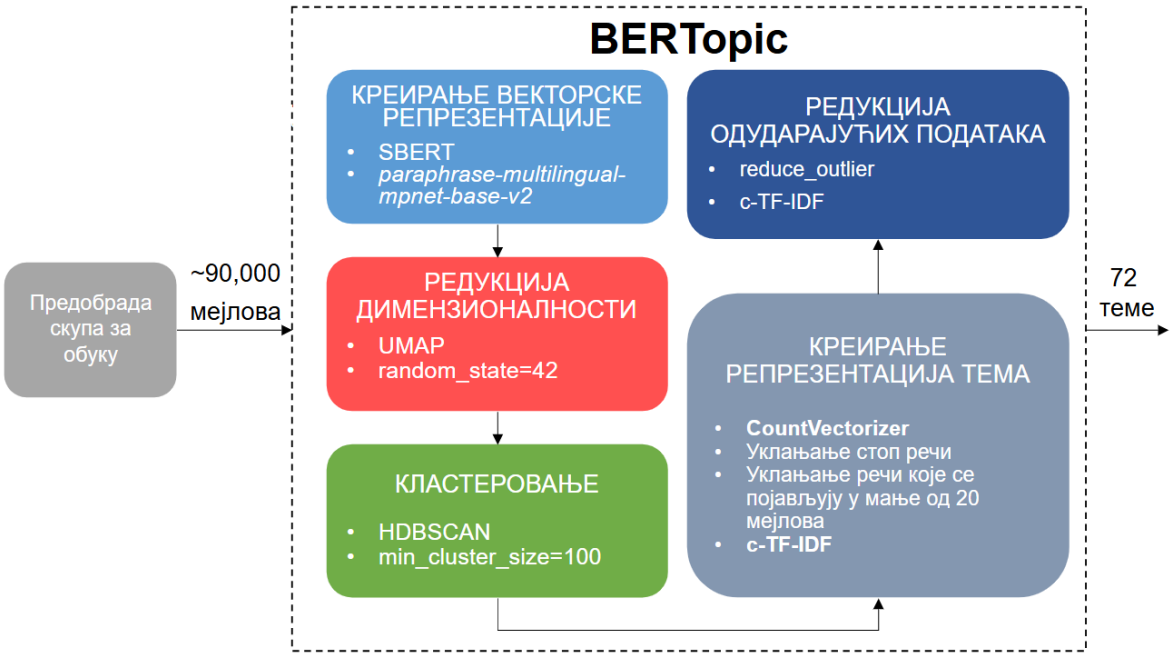
Од полазних ~250.000 мејлова по завршетку предобраде у скупу је остало 89.486 мејлова, од чега 48.954 мејла резиденцијалних корисника и 40.532 мејла пословних корисника. За потребе развоја модела, скуп података је подељен на скуп за обуку (89.386 мејлова) и тест скуп (100 мејлова, по 50 за резиденцијалне и пословне кориснике). Како мејлови из тест скупа садрже анонимизацијске ознаке, а за потребе тестирања модела у реалним условима функционисања, ове ознаке су у свим мејловима замењене фиктивним подацима одговарајуће категорије (нпр. ознака *PERSON* је замењена фиктивним именом Петар Петровић).

Обука модела

У наставку је представљен опис слојева *BERTopic* архитектуре, као и параметара који су коришћени у сваком од њих приликом обуке на припремљеном скупу података.

BERTopic архитектура се састоји из четири основна слоја (1-4. слој), а на њу се могу додавати и опциони слојеви. Овде је представљен опциони слој који је

коришћен у развоју модела (5. слој). Слојеви се у току обуке над текстом примењују секвенцијално почевши од слоја за издвајање векторских репрезентација текста. Поред слојева, у конфигурацији модела могу се подешавати и додатни параметри. Сви дефинисани слојеви и додатни параметри се затим интегришу у једну целину кроз иницијализацију класе *BERTopic,* чиме се формира комплетна архитектура за моделовање тема.



Слика 2. Процес обуке модела за аутоматско препознавање тема

У наставку су дата објашњења слојева и њихове улоге у *BERTopic* архитектури, као и конкретних параметара коришћених у развоју модела имплементираног у овом модулу, које сажето приказује Слика 2.

1. **Слој за креирање векторских репрезентација текста** којим се текст сваког мејла из задатог скупа преводи у одговарајућу векторску репрезентацију (енгл. *embedding*). За генерисање овакве репрезентације *BERTopic* подразумевано користи *SentenceTransformer* (*SBERT*) језички модел [15] оптимизован за одређивање семантичке сличности међу текстовима. Како у периоду имплементације није било доступних модела овог типа обучених искључиво на српском језику, а претходно истраживање [4] показује да се најбољи резултати на предобрађеном тексту без лематизације за српски језик постижу применом модела *paraphrase-multilingual-mpnet-base-v2*[1], одлучено је да се овај модел користи за издвајање векторских репрезентација мејлова. У питању је

[1] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

вишејезички модел са подршком за више од 50 језика, укључујући и српски, који садржи 278 милиона параметара.

2. **Слој за редукцију димензионалности векторских репрезентација текста** којим се високо-димензионалне векторске репрезентације (768) добијене језичким моделом из корака 1 редукују са циљем побољшања перформанси корака 3. У овом слоју коришћен је *UMAP* [16] алгоритам, који је уједно и подразумевани *BERTopic* алгоритам за редукцију димензионалности.

3. **Слој за кластеровање докумената у теме** примењује се након редукције димензионалности како би се векторске репрезентације које су међусобно блиске у простору груписале. На тај начин формирају се групе тј. кластери мејлова са семантички сличним садржајем, што је предкорак у идентификацији теме за сваки од кластера.

   Као алгоритам за кластеровање коришћен је *HDBSCAN* [17], који је и подразумевани алгоритам у *BERTopic* архитектури због његове способности да открива структуре различитих густина у подацима, са подразумеваним вредностима параметара *о*сим параметра *min_cluster_size*, чија је вредност пропагирана из параметра *min_topic_size*, који је објашњен у наставку овог поглавља.

4. **Слој за креирање репрезентација тема.** Након кластеровања мејлова по сличности, за сваки кластер се одређују најрепрезентативније кључне речи које представљају тему. Свака тема је представљена са по 10 кључних речи издвојених унутар кластера. За овај слој коришћена је *CountVectorizer* компонента *BERTopic* архитектуре са *c-TF-IDF* [1][18] алгоритмом за одређивање најрепрезентативнијих кључних речи по кластерима. Док је *c-TF-IDF* коришћен са подразумеваним параметрима *BERTopic* имплементације, у *CountVectorizer* компоненти су дефинисани параметри за филтрирање речи, које претходи одређивању репрезентативног скупа кључних речи за сваку тему. Конкретно, речи су филтиране уклањањем стоп речи и речи које се појављују у мање од 20 мејлова. За уклањање стоп речи коришћена је листа стоп речи за српски језик [19], која је додатно проширена речима које се често појављују у мејловима, а нису информативне за одређивање тема (нпр. *поштовани* или о*браћам*).

5. **Слој за редукцију одударајућих података.** Овај слој је опциони и може се додати у случајевима када постоји велики број одударајућих података (енгл. *outliers*), односно мејлова који нису додељени ниједном од кластера који представљају теме. Најчешће је потребан када се за кластеровање текстова користи *HDBSCAN* алгоритам, који по природи прави посебну

групу са оваквим подацима. У оквиру овог слоја се примењује метод *reduce_outliers* са *c-TF-IDF* стратегијом редукције одударајућих података којом је од око 47.000 мејлова, њихов број смањен на око 470.

Поред дефинисања појединачних слојева *BERTopic* архитектуре, коришћени су и додатни параметри, при чему су за већину преузете подразумеване вредности *BERTopic* архитектуре, осим параметара *min_topic_size* и *seed_topic_list*, који су дефинисани у складу са потребама корисничке службе.

Параметром *min_topic_size* дефинише се минималан број мејлова у кластеру. Емпиријски је утврђено да је оптимална вредност овог параметра 100, односно да се кластер може формирати ако му припада најмање 100 мејлова.

Параметар *seed_topic_list* се користи за вођено моделовање тема кроз дефинисање листе кључних речи које одговарају теми од интереса у конкретној примени. На основу такве листе, модел ће настојати да бар једну тему усмери ка тим кључним речима. Након тестирања на више различитих примера тема, као корисно се показало једино профињење тема модела које се односе на фиксну и мобилну телефонију, уз примену следеће листе: *[["fiksni", "fiksnog", "fiksna", "fiksnoj"]]*.

Обучени модел је на основу полазног скупа података издвојио укупно 72 теме описане са по 10 кључних речи и једну додатну групу са одударајућим подацима, у коју су сврстани сви преостали мејлови.

Примена обученог модела

Како би се новопридошлом мејлу доделила тема, текст мејла пролази кроз слојеве 1-3 и опционо слој 5, након чега се тема мејла одређује спрам сличности векторске репрезентације мејла некој од претходно научених тема. Долазним мејловима се додељује искључиво једна тема.

3.4. Модул за постобраду мејлова

Циљ овог модула је да се 72 теме модела, укључујући и групу одударајућих података, трансформишу у такозване изведене теме. Изведених тема има 12 и оне одговарају различитим категоријама потребним са аспекта пословања компаније (нпр. *Фискална каса, Пакети услуга, Картице, Фиксна и мобилна телефонија*). Улаз у овај модул су сви мејлови учитани у једном циклусу покретања софтвера. У скупу свих мејлова могу се изоловати две подгрупе:

- **Подгрупа 1**: мејлови са додељеним темама модела.
- **Подгрупа 2**: мејлови без теме модела, добијени као излаз модула за предобраду података.

За мејлове из подгрупе 1 се изведена тема одређује на основу мапирања тема модела на изведене теме. Пример овог мапирања за једну тему приказује

Табела 1. Уз мапирање, у табели су приказане и кључне речи које су током процеса обуке *BERTopic* модела изоловане за појединачне теме модела.

Табела 1. Мапирање тема модела на изведену тему Рачуни и фактуре.

Изведена тема	Тема модела	Кључне речи
Рачуни и фактуре	30	['dinara', 'iznos', 'objasnite', 'din', 'cena', 'račun', 'iznosu', 'racun', 'naplaćuje', 'iznosi']
	70	['dobili', 'stigao', 'poštom', 'račun', 'pošaljete', 'mesec', 'račune', 'racun', 'mail', 'dostavite']
	4	['dugovanje', 'dug', 'dugovanja', 'izmirili', 'duga', 'odnosi', 'izmirena', 'račun', 'iznosu', 'perioda']
	60	['dva', 'platila', 'greskom', 'racun', 'platio', 'racuna', 'isti', 'uplatila', 'račun', 'uplatu']
	43	['faktura', 'fakture', 'plaćanje', 'fakturu', 'br', 'iznos', 'sistemu', 'crf', 'sistem', 'računa']
	18	['mesec', 'pošaljete', 'dostavite', 'posaljete', 'kopiju', 'račun', 'račune', 'racune', 'računa', 'firmu']
	71	['naknada', 'pretplata', 'din', 'mesečna', 'mesecna', 'cene', 'iznosi', 'cena', 'naknade', 'rsd']
	67	['pdv', 'prethodnog', 'potpisano', 'om', 'predhodnog', 'din', 'potpisana', 'dinara', 'račun', 'storno']
	57	['placen', 'racun', 'racuna', 'mesec', 'racuni', 'uplaceno', 'uplacen', 'uplata', 'placeni', 'uplatnicu']
	15	['poziv', 'uplata', 'račun', 'uplatu', 'uplate', 'plaćen', 'racuna', 'računa', 'broj', 'racun']
	68	['račun', 'iznos', 'računa', 'iznosu', 'mesec', 'rsd', 'greškom', 'računu', 'prigovor', 'dinara']
	41	['račune', 'šaljete', 'adresu', 'mail', 'ubuduće', 'mejl', 'računi', 'račun', 'računa', 'formi']
	10	['uplati', 'dokaz', 'izvod', 'ukljucenje', 'uplata', 'uplatu', 'računa', 'racuna', 'uplate', 'uključenje']

Сви мејлови подгрупе 2 добијени су издвајањем из почетног скупа оних мејлова који нису адекватни за аутоматску доделу тема. У зависности од ознаке мејла која носи информацију о разлогу искључивања мејла из аутоматског препознавања теме, мејлу може бити додељена једна од специјалних изведених тема. Код ових изведених тема, тема модела је предефинисана вредност која није из скупа вредности научених тема модела. Ознака мејла се шаље као излаз модула за предобраду уз сваки мејл и на основу њене вредности се мејлу додељује једна од следећих изведених тема:

- **Интерна преписка** (-2) – тема се додељује свим мејловима који представљају интерну преписку запослених унутар компаније, односно мејловима који нису кориснички мејлови.

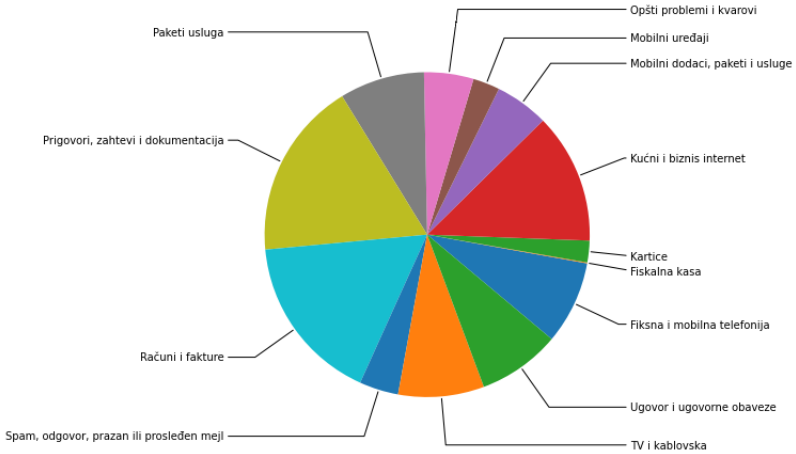
- **Спам, одговор, празан или прослеђен мејл** (-3) – тема се додељује свим примљеним непожељним мејловима, мејловима који после предобраде имају празан наслов и тело, односно у целини представљају прослеђен мејл или садрже само одговор на претходну преписку.

Излаз овог модула је полазни скуп мејлова где је за сваки мејл позната изведена тема, која има посебне вредности -2 и -3 за мејлове из подгрупе 2, уз тему модела за мејлове из подгрупе 1. За све мејлове из подгрупе 1 су доступне и кључне речи. Наведени подаци се серијализују у одговарајућа поља базе података за сваки мејл из улазног скупа података.

4. Корисници и примена

Систем за аутоматско одређивање тема корисничких мејлова описан у овом техничком решењу је у примени од 1.6.2024. године у компанији Телеком Србија а.д. Користи се како би се аутоматски одредиле теме за корисничке мејлове, без интервенције оператера, те како би се обезбедила неопходна инфраструктура за имплементацију екстерне компоненте за рутирање корисничких мејлова одговарајућим секторима корисничке службе. Према захтевима компаније, систем се периодично активира за резиденцијалне и пословне кориснике, при чему се у просеку обради 2310 мејлова дневно (просечно 1921 мејлова резиденцијалних и 300 пословних корисника).

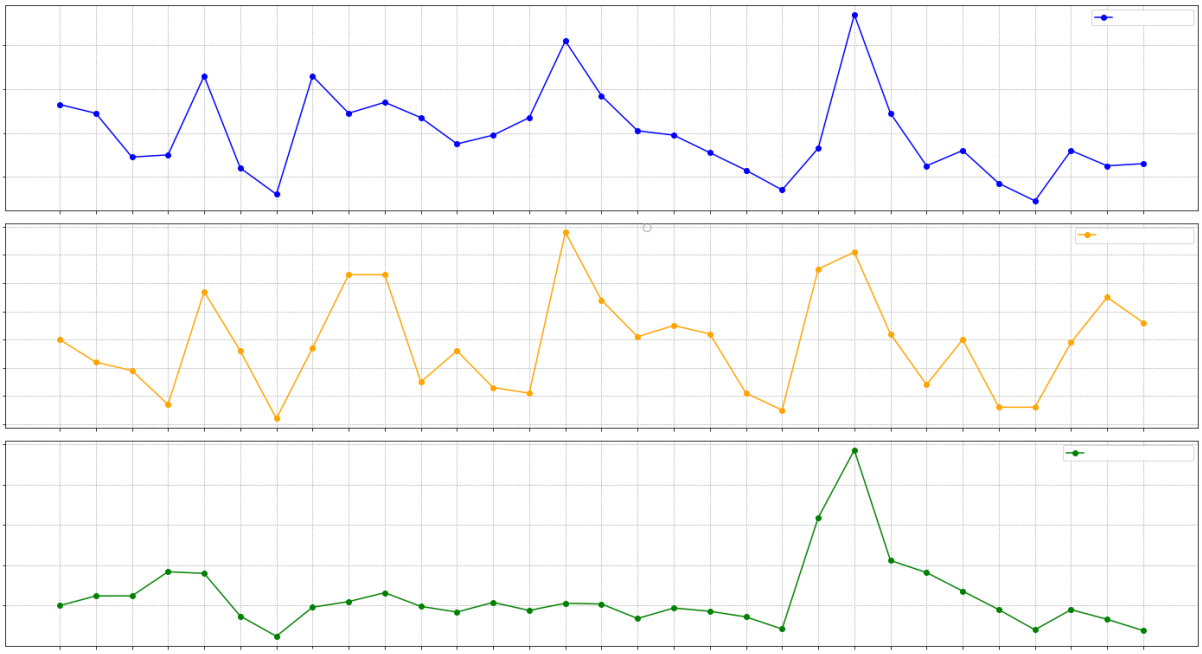
Корисници система предложеног овим техничким решењем укључују запослене у корисничкој служби, тимовима за бригу о корисницима и тимовима за анализу



Слика 3. Пример дијаграма расподеле мејлова по изведеним темама на месечном нивоу.

података компаније. Аутоматском категоризацијом мејлова стварају се предуслови за убрзање и побољшање ефикасности корисничке службе услед елиминисања потребе за прегледом мејла од стране оператера. Тим за бригу о корисницима анализира изведене теме мејлова како би идентификовао најчешће проблеме и трендове у упитима корисника, што ствара предуслове за

континуирано побољшање услуга и корисничког искуства. Поред тога, тим за анализу података користи структурирану класификацију тема у мејловима за дубљу пословну анализу и извештавање, чиме доприноси оптимизацији интерних процеса, унапређењу квалитета услуге и доношењу стратешких одлука. Слике Слика 3 и Слика 4 приказују примере дијаграма које генерише описани систем.



Слика 4. Пример дијаграма тренда примљених мејлова за најзаступљенију тему. Подаци са дијаграма су анонимизовани – један дијаграм је један канал комуникације који показује број примљених мејлова (ордината) за сукцесивне датуме у месецу (апсциса).

5. Верификација

1) Техничко решење је реализовано у компанији Телеком Србија а.д., где га свакодневно користи Сектор за бригу о корисницима, док његово извршавање надгледа и прати Сектор за стратегију и дигитал.

2) Пре примене, делови решења су испитивани и верификовани на развојним системима Истраживачко-развојног института за вештачку интелигенцију Србије. Том приликом је утврђено да модул за аутоматску обраду тема ради с тачношћу од 96% (Ф1 мера) и просечном брзином од 0.041 секунди по мејлу. Ови резултати, као и предложени систем, описани су у раду:

Bašaragin, Bojana, Darija Medvecki, Gorana Gojić, Milena Oparnica, and Dragiša Mišković. "Improving customer service with automatic topic detection in user emails." 15th International Conference on Information Society and Technology (ICIST), Kopaonik, Serbia, 9-12 March 2025 (прихваћено за објављивање у Lecture Notes

in Networks and Systems, Springer Nature Switzerland, препринт доступан на arXiv-у).

Док су основе овог техничког решења постављене у раду:

Medvecki, Darija, Bojana Bašaragin, Adela Ljajić, and Nikola Milošević. "Multilingual transformer and BERTopic for short text topic modeling: The case of serbian." In: Trajanovic, M., Filipovic, N., Zdravkovic, M. (eds) Disruptive Information Technologies for a Smart Society. ICIST 2023. Lecture Notes in Networks and Systems, vol 872. Springer, Cham. https://doi.org/10.1007/978-3-031-50755-7_16

3) Након интерне провере, Научно веће Истраживачко-развојног института за вештачку интелигенцију Србије утврдило је да предложено техничко решење испуњава све услове да буде признато као ново техничко решење примењено на националном нивоу, у складу са Правилником Министарства, и сходно томе је једногласно донело Одлуку на упућивање на даљу евалуацију надлежном матичном одбору. Ова одлука се налази у Прилогу 3.

6. Литература

[1] S. Filippou, A. Tsiartas, P. Hadjineophytou, S. Christofides, K. Malialis, and C. G. Panayiotou, “Improving Customer Experience in Call Centers with Intelligent Customer-Agent Pairing,” 2023. doi: 10.1007/978-3-031-34111-3_19.

[2] S.-T. Pi, C.-P. Hsieh, Q. Liu, and Y. Zhu, “Universal Model in Online Customer Service,” in *Companion Proceedings of the ACM Web Conference 2023*, in WWW ’23 Companion. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 878–885. doi: 10.1145/3543873.3587630.

[3] T. Mihajlov, M. I. Nešić, R. Stanković, and O. Kitanović, “Topic Modeling of the SrpELTeC Corpus: A Comparison of NMF, LDA, and BERTopic,” pp. 649–653, Oct. 2024, doi: 10.15439/2024F1593.

[4] D. Medvecki, B. Bašaragin, A. Ljajić, and N. Milošević, “Multilingual transformer and BERTopic for short text topic modeling: The case of Serbian,” vol. 872, 2024, pp. 161–173. doi: 10.1007/978-3-031-50755-7_16.

[5] D. M. Blei, Ng, Andrew Y, and Jordan, Michael I, “Latent Dirichlet Allocation,” *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[6] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[7] M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” Mar. 11, 2022, *arXiv*: arXiv:2203.05794. doi: 10.48550/arXiv.2203.05794.

[8] Ş. Ö. Birim, “Product Insights from Customer-Generated Data Using Topic Modeling with BERTopic and Sentiment Analysis with XLM-T: An Experiment on Turkish Reviews,” Feb. 26, 2024, *In Review*. doi: 10.21203/rs.3.rs-3981153/v1.

[9] O. I. Babina, “Topic Modeling for Mining Opinion Aspects from a Customer Feedback Corpus,” *Autom. Doc. Math. Linguist.*, vol. 58, no. 1, pp. 63–79, Feb. 2024, doi: 10.3103/S0005105524010060.
[10] C. Kim and J. Lee, “Discovering patterns and trends in customer service technologies patents using large language model,” *Heliyon*, vol. 10, no. 14, p. e34701, Jul. 2024, doi: 10.1016/j.heliyon.2024.e34701.
[11] B. Bašaragin, D. Medvecki, G. Gojić, M. Oparnica, and D. Mišković, “Improving customer service with automatic topic detection in user emails,” Feb. 26, 2025, *arXiv*: arXiv:2502.19115. doi: 10.48550/arXiv.2502.19115.
[12] Radović, Andrej, *Srtools*. (2021). Available: https://pypi.org/project/srtools/
[13] Lui, M., *LangID*. (2016). Available: https://pypi.org/project/langid/
[14] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
[15] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Aug. 2019, doi: 10.48550/arXiv.1908.10084.
[16] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Sep. 18, 2020, *arXiv*: arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426.
[17] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *JOSS*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[18] “cTFIDF - BERTopic.” Accessed: Mar. 17, 2025. [Online]. Available: https://maartengr.github.io/BERTopic/api/ctfidf.html
[19] Marovac, Ulfeta, Avdić, Aldina, and Ljajić, Adela, “Creating a Stop Word Dictionary in Serbian,” *Scientific publications of the state University of Novi Pazar*, vol. 13, no. 1, pp. 17–25, 2021.

7. Прилози

1. Списак раније прихваћених техничких решења за сваког од аутора појединачно.

2. Уговор о имплементацији пројекта „Развој модела машинског учења за анализу текста“ у компанији Телеком Србија а.д. у форми наруџбенице.

3. Одлука Научног већа Истраживачко-развојног института за вештачку интелигенцију Србије.

Прилог 1 - Списак раније прихваћених техничких решења за сваког од аутора појединачно

Бојана Башарагин, научни сарадник, Истраживачко-развојни институт за вештачку интелигенцију Србије

Ново техничко решење примењено на међународном нивоу (М81):

1. Милош Кошпрдић, Никола Продановић, Адела Љајић, Бојана Башарагин, Никола Милошевић, „Метода за препознавање биомедицинских именованих ентитета без примера или са малим бројем примера за учење.“, Матични научни одбор за електронику, телекомуникације и информационе технологије, ТР 07/27 од 06.06.2024.

Драгиша Мишковић, виши научни сарадник, Истраживачко-развојни институт за вештачку интелигенцију Србије

Нови производ или технологија уведени у производњу на међународном нивоу М81:

1. Мирко Раковић, Драгиша Мишковић, Лазар Милић, Андреј Чилаг, Тања Берисављевић, „Аутоматизација фабричких процеса кроз интеграцију мобилног и индустријског робота“, техничко решење, ФТН Нови Сад, 2022.
2. Драгиша Мишковић, Дарко Пекар, Милан Сечујски, Едвин Пакоци „Speech-Label – алат за фонетско и прозодијско лабелирање говорних база”, техничко решење, ФТН Нови Сад, 2015.
3. Бранислав Поповић, Драган Кнежевић, Милан Сечујски, Дарко Пекар, Никша Јаковљевић, Драгиша Мишковић, Рон Хасон „Технологија аутоматске синтезе говора на основу текста на хебрејском језику”, тех. решење, ФТН Н. Сад, 2012.
4. Владо Делић, Никша Јаковљевић, Радован Обрадовић, Дарко Пекар, Драгиша Мишковић, Драган Кнежевић „Технологија аутоматског препознавања говора на српском и њему сродним језицима”, техничко решење, ФТН Нови Сад, 2007.
5. Милан Сечујски, Владо Делић, Дарко Пекар, Драгиша Мишковић, Драган Кнежевић, Марко Јанев „Технологија аутоматске синтезе говора на основу текста на српском и другим јужнословенским језицима”, техничко решење, ФТН Нови Сад, 2006.

Нова производна линија, нови материјал, индустријски протоип, ново прихваћено решење у области макроекономског, социјалног и проблема одрживог просторног развоја уведени у производњу М82:

1. Огњен Кундачина, Владимир Винцан, Милица Шкипина, Илија Каменко, Драгиша Мишковић, Дубравко Ћулибрк „Веб сервис за подршку апликацијама за видео конференције базиран на дубоком учењу“, ИВИ Нови Сад, 2024
2. Драгиша Мишковић, Дарко Пекар, Никша Јаковљевић, Драган Кнежевић, Милана Бојанић „MRCP интерфејс за синтетизатор говора на српском језику”, техничко решење, ФТН Нови Сад, 2012.

Прототип, нова метода, софтвер, стандардизован или атестиран инструмент М85:

1. Драгиша Мишковић, Милан Гњатовић, Владо Делић, Бранислав Боровац, Јовица Тасевски „Когнитивна платформа за управљање конверзационим роботом“, техничко решење, ФТН Нови Сад, 2015.
2. Стеван Остроганац, Дарко Пекар, Драгиша Мишковић, Милан Сечујски, Бранислав Поповић, Владо Делић „Паметна кућа базирана на говорним технологијама за српски језик (anSmartHome)“, ФТН Нови Сад, 2014.
3. Никша Јаковљевић, Драгиша Мишковић, Едвин Пакоци, Роберт Мак, Бранислав Поповић „Декодер за препознавање континуалног говора на великим речницима“, техничко решење, ФТН Нови Сад, 2012.
4. Јовица Тасевски, Драгиша Мишковић, Милан Гњатовић, Милутин Николић, Бранислав Боровац, Владо Делић „Роботски систем са интегрисаним системима за обраду слике и говора и конверзационим агентом“, техничко решење, ФТН Нови Сад, 2012.

ПРИЛОГ 2 - Уговор о имплементацији пројекта „Развој модела машинског учења за анализу текста“ у компанији Телеком Србија а.д. у форми наруџбенице.