# АНАЛИЗА ГРЕШАКА У МЕДИЦИНСКИМ ИЗВЕШТАЈИМА НА СРПСКОМ ЈЕЗИКУ У ЦИЉУ ЊИХОВЕ АУТОМАТСКЕ ДЕТЕКЦИЈЕ И КОРЕКЦИЈЕ

*Алдина Авдић[1] Улфета Маровац [2]Драган Јанковић[3]*

**Резиме:** Развој информационих технологија омогућио је креирање и коришћење медицинских информационих система у којима се свакодневно чувају подаци о здрављу пацијената у виду електронских медицинских извештаја. Применом техника вештачке интелигенције, као што су обрада природног језика и машинско учење, из ових извештаја могу се извући значајни закључци о здрављу појединца или популације. У циљу добијања што прецизнијих закључака, потребно је, пре процесирања текста из медицинских извештаја, извршити њихову адекватну припрему. Један од корака у припреми података јесте детекција и корекција грешака насталих у току писања извештаја. У овом раду анализиран је скуп података електронских медицинских података, дата је класификација типова грешака које су у овом скупу пронађене и дати резултати њихове расподеле. Предложена је архитектура система са предлогом метода за аутоматску детекцију и корекцију грешака у електронским медицинским извештајима на српском језику.

**Кључне речи:** електронски медицински извештаји, истраживање текста, обрада природног језика, аутоматска детекција грешака, аутоматска корекција грешака.

# ANALYSIS OF MISSPELLINGS IN MEDICAL REPORTS IN SERBIAN LANGUAGE FOR THE PURPOSE OF THEIR AUTOMATED DETECTION AND CORRECTION

**Abstract:** The development of information technologies has enabled the creation and use of medical information systems in which data on patient health are stored on a daily basis in the form of electronic medical reports. By applying artificial intelligence techniques, such as text mining and machine learning, significant conclusions about the health of an individual or population can be drawn from these reports. In order to obtain the most precise conclusions, it is necessary, before processing the text from medical reports, to make adequate preparation. One of the steps in data preparation is the detection and correction of misspellings that occurred during the writing of the report. In this paper, a set of electronic medical data is analyzed, a classification of types of misspellings found in this set is given, and the results of their distribution are given. The architecture of the system is proposed with a proposal of methods for automated detection and correction of errors in electronic medical reports in the Serbian language.

**Key words:** electronic medical reports, text mining, natural language processing, automated misspellings detection, automated misspellings correction.

## 1. УВОД

Медицински информациони системи (МИС) имају значајну улогу у савременим здравственим системима [1]. Поред тога што се подаци о здравственом стању пацијената чувају и централизовани су, омогућена је и њихова лакша претрага и свеукупно управљање, а отвара се и простор за анализу ових података. Такође, помоћу ових система може се водити евиденција о медицинском особљу, стању потребних ресурса, смањују

[1] PhD, State University of Novi Pazar, Vuka Karadžića bb, 36300 Novi Pazar, apljaskovic@np.ac.rs

[2] PhD, State University of Novi Pazar, Vuka Karadžića bb, 36300 Novi Pazar, umarovac@np.ac.rs

[3] PhD, Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18106 Niš, dragan.jankovic@elfak.ni.ac.rs

се трошкови администрације, тако да ови системи имају бројне предности и брзо су постали неизоставни део у здравственим системима.

Електронски медицински извештаји (енгл. *Electronic Health Record – EHR*) чувају податке о здравственом стању пацијената, и најчешће их пишу лекари [2]. Они могу имати структурирани, полуструрирани и неструктурирани део. Структурирани део настао је уношењем у означена текстуална поља и њихова структура у потпуности је позната (нпр. датум прегледа, лични број осигураника, код дијагнозе итд.). Полуструктурирани део има делимично познату структуру, за разлику од неструктурираног дела који се састоји од слободног текста, у коме лекар даје додатна запажања о здравственом стању пацијената која се не могу изразити кроз претходна дата поља за унос података. Најчешће су у овом делу белешке о резултатима лабораторијских анализа, претходне или повезане болести, симптоми, дијагнозе, терапије или други подаци од значаја. Због ограниченог времена трајања прегледа, обично је овај део медицниских извештаја подложан грешкама у куцању.

Грешке у медицинским извештајима могу бити кобне, нпр. могу водити до погрешне терапије уколико је назив лека погрешно написан. Постојање грешака у медицинским извештајима отежава њихову анализу, а она је потребна због добијања новог знања. Зато је адекватно припремање ових извештаја пре процесирања основна мотивација овог истраживања. Бројни су примери примене анализе слободног текста из медицинских извештаја, али један од актуелних био би закључивање о промени симптома током времена за одговарајућу болест, имајући у виду тренутну ситуацију са пандемијом корона вируса и појаве разних сојева који са собом носе различите симптоме.

У овом раду примењене су технике обраде природног језика над скупом електронсих медицинских извештаја на српском језику прикупњених од стране информационог система МЕДИС.НЕТ , у циљу детекције и корекције грешака у слободном тексту медицинских извештаја, и приказани су и дискутовани резултати њихове примене [3].

Циљ овог истраживања је анализа грешака које се јављају у слободном тексту у медицинским извештајима на српском језику како би се формирала правила за њихово аутоматско исправљање. Главни допринос рада је класификација врсти грешака, приказ њихове заступљености у анализираним медицинским извештајима и предлог архитектуре система за њихову детекцију и корекцију који је базиран на специјализованим речницима, обради природног језика, тренинг скупу који се састоји из ручно означених извештаја, машинском учењу и правилима.

Рад је организован на следећи начин. У другом поглављу дат је преглед радова који су се бавили сличном тематиком. У трећем поглављу описани су подаци и методе које су коришћени за анализу. Следи приказ класификације и расподеле пронађених грешака у анализираном скупу. Дат је и предлог архитектуре система за откривање и исправљање грешака у медицинским извештајима на српском језику. На крају је дат закључак и правци даљег истраживања.

## 2. ТЕОРИЈСКА ПОЗАДИНА И СТАЊЕ У ОБЛАСТИ

Истраживање података (енгл. *data mining*) и истраживање текста (енгл. *text mining*) се разликују у погледу врсте података које обрађују [4][5]. Док истраживање података обрађује структуриране податке (нпр. базе података), истраживање текста се бави неструктурираним текстуалним подацима (нпр. постови друштвених медија) [6][7]. Ниједно од њих није јединствена технологија, већ користе широк спектар функција за претварање доступних података у знање. Истраживање података комбинује дисциплине

које укључују статистику, вештачку интелигенцију и машинско учење над структурираним подацима. Неке од функција моделирања података су: асоцијација, класификација, кластеровање и регресија.

Истраживање текста захтева додатни корак уз задржавање истог циља као и код истраживања података. Истраживање текста се бави неструктурираним подацима, тако да пре него што се примени било која функција моделирања података или препознавања образаца, неструктурирани подаци морају бити организовани и структурирани на начин који омогућава њихово моделирање и анализу. Овај процес је обично повезан са техником вештачке интелигенције која се назива процесирање природног језика, у даљем тексту НЛП (енгл. NLP – *Natural Language Processing*) и омогућава систему да разуме значење података на људском језику. Крајњи циљ НЛП-а је читање, дешифровање, разумевање и налажење смислености у природном језику [8]. Већина НЛП техника ослања се на машинско учење да би се закључило о значењу података на природним језицима.

Детекцијом и корекцијом грешака у медицинским извештајима бавили су се аутори са разних говорних подручја. У прегледном раду техника провере правописних грешака, идентификована су три подпроблема: откривање грешака које нису речи, исправљање грешака изоловане речи и исправљање грешака зависно од контекста [9] [10]. У н-грам анализи, која се углавном користи у системима оптичког препознавања карактера, необични низови знакова су индикатори за препознавање грешака [11]. Н-грам анализа користи се за корекцију грешака из медицинског домена на персијском језику [12]. Често се системи за корекцију грешака користе речници: свака реч која није у речнику је вероватно погрешно написана.

Већина система за корекцију грешака изолованих речи користи неки облик минималне удаљености за уређивање за генерисање или рангирање предлога. Преко 80% правописних грешака састоји од једне од следећих операција: уметнуто слово, обрисано слово, слово замењено другим или два транспонована или замењена слова [13]. ДЛ удаљеност представља број колико је ових операција потребно да се једна реч трансформише у другу и користи се и за руски језик [14].

Исправљање грешака у зависности од контекста се користи у случајевима када се правилно написана реч замењује другом. Ове технике користе статистичке језичке моделе за откривање лоше обликованих низова речи. Такође се у детекцији и корекцији грешака који користи НЕР, НЛП методе и Шенонов модел за шум у комуникационим каналима [15][16].

Независно од медицинског домена, о детекцији и корекцији грешака у српском језику говори се у дисертацији [17].

## 3. МАТЕРИЈАЛ И МЕТОДЕ

За потребе овог истраживања коришћен је корпус који чини 5261 медицинских извештаја. Ови извештаји написани су на српском језику из 32 здравствене установе који припадају дому здравља Ниш (ДЗ Ниш), а прикупио их је информациони систем МЕДИС.НЕТ [3]. Медицинске извештаје написало је 169 различитих лекара. Овај корпус је направљен према свим етичким стандардима, уз уклањање идентитета пацијената и медицинског особља, као и одржавање веза са припадношћу више извештаја истом пацијенту.

У овом раду су коришћене следеће научне методе: дескрипција, анализа садржаја, и експериментална метода. Метода дескрипције примењена је на постојеће методе за детекцију и корекцију грешака у медицинским извештајима, док је над описаним скупом података најпре извршена анализа садржаја, у циљу проналажења грешака које се јављају у корпусу и креирања метода за њихово исправљање. У циљу детекције и корекције грешака предложене су методе обраде природног језика.

## 4. ВРСТЕ ГРЕШАКА У ЕЛЕКТРОНСКИМ МЕДИЦИНСКИМ ИЗВЕШТАЈИМА

Анализом садржаја описаног скупа података дошло се до једанаест типова грешака који се јављају у електронским медицинским извештајима. У табели 1 дати су описи ових врста грешака и по пример за сваку од њих из анализираног скупа података.

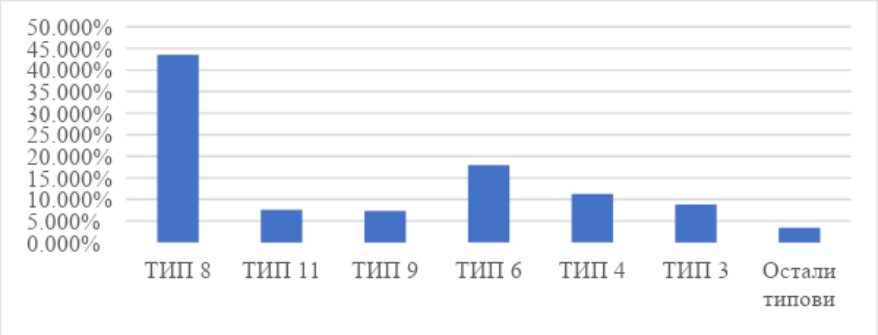
Анализом означеног скупа података дошло се до податка да је проценат грешака у електронским медицинским извештајима око 4.5%.

На основу процента појављивања грешака у обрађеном скупу података, видимо да су најзаступљенији типови грешака ТИП8, ТИП 6, ТИП 4 и ТИП 3. То значи да се у формулацији правила у алгоритму за корекцију грешака треба водити рачуна о отклањању ових типова грешки.

На слици 1. приказана је расподела типова грешака у означеном скупу.

*Табела 1 – Врсте грешака идентификовани у електронским медицинским извештајима*

| Назив грешке | Опис грешке | Пример | Исправка |
|---|---|---|---|
| ТИП 1 | изостављено двоструко слово | *поштрено* | пооштрено |
| ТИП 2 | замена места слова | *малакслаост* | малаксалост |
| ТИП 3 | додатна слова | *трупзу* | трупу |
| ТИП 4 | недостајућа слова | *темература* | температура |
| ТИП 5 | замена сличном речју | *зрело* | ждрело |
| ТИП 6 | спојене речи без размака | *тхбруфен* | тх. бруфен |
| ТИП 7 | спојене речи са случајним словом уместо размака | *помтелу* | по телу |
| ТИП 8 | изостављен (замењен) дијакритички симбол (нпр. „ц“ уместо „ч“ или „ћ“) | *кози* | кожи |
| ТИП 9 | нетачно слово | *уирус* | вирус |
| ТИП 10 | употреба слова која не припадају српској абецеди („х“ уместо „кс“) | *ехтремитетима* | екстремитетима |
| ТИП 11 | више грешака у једној речи | *маколозма* | макулозна |

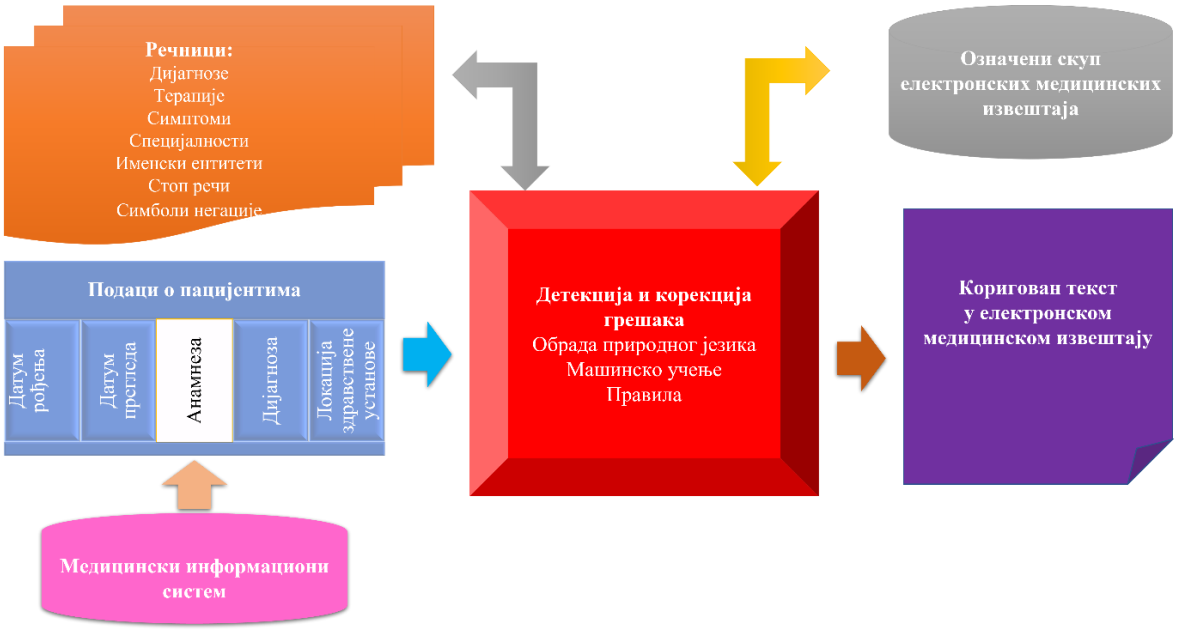



*Слика 1 - Проценат појављивања грешака у EMP-овима*

## 5. АРХИТЕКТУРА СИСТЕМА ЗА ДЕТЕКЦИЈУ И КОРЕКЦИЈУ ГРЕШАКА У МЕДИЦИНСКИМ ИЗВЕШТАЈИМА НА СРПСКОМ ЈЕЗИКУ

**Детекција грешака** се може обавити тако што ћемо пробати да речи означимо применом неког од алгоритама за означавање речи у медицинским текстовима на српском језику. За детекцију могу се користити методе засноване на речницима, или методе засноване на машинском учењу које користе означен скуп електронских медицинских извештаја [16]. Уколико реч није означена, она ће се сматрати грешком.

**Корекција грешака** се може извршити такође помоћу метода заснованих на речницима или на учењу над тренинг скупом, али је потребно да ове методе укључују у себи коришћење алгоритма за нормализацију заснован на н-грам анализи и стемеру за српски језик, и да се користе додатни кораци за исправљање грешака у виду правила за исправљање грешака [11][18]. На слици 2 је приказана архитектура система за детекцију и корекцију грешака у електронским медицинским извештајима на српском језику.




*Слика 2 - Архитектура система за детекцију и корекцију грешака у електронским медицинским извештајима на српском језику*

## 6. ЗАКЉУЧАК

Коректно написани електронски медицински извештаји могу утицати на успешност излечења пацијената, а њикова некоректност може имати кобне последице. У овом раду анализиран је скуп медицинских извештаја, и у њему су пронађене и обележене грешке с циљем да се креирају методе за њихову детекцију и корекцију. Главни допринос рада је класификација врсти грешака, приказ њихове заступљености у анализираним медицинским извештајима и предлог архитектуре система за њихову детекцију и корекцију који је базиран на специјализованим речницима, обради природног језика, тренинг скупу који се састоји из ручно означених извештаја, машинском учењу и правилима. Предмет даљег истаживања биће квантитативна анализа предложених метода и њихово експериментално извођење и компарација резултата са сличним методама за детекцију и корекцију грешака на сличним језицима.

## 7. ЛИТЕРАТУРА

[1] Henry, J. W., & Stone, R. W. (1994). A structural equation model of end-user satisfaction with a computer-based medical information system. Information Resources Management Journal (IRMJ), 7(3), 21-33.

[2] Menachemi, N., & Collum, T. H. (2011). Benefits and drawbacks of electronic health record systems. Risk management and healthcare policy, 4, 47.

[3] A. M. Milenkovic, P. Rajkovic, T. Stankovic, D. S. Janković, Application of Medical Information System MEDIS.NET in Professional Learning, 19th Telecommunications Forum (TELFOR) Proceedings of Papers, 2011, 1474-1477.

[4] S. Russell, P. Norvig, Artificial intelligence: a modern approach, 2002.

[5] D. J. Hand, N. M. Adams, Data mining. Wiley StatsRef: Statistics Reference Online, 2014, 1-7.

[6] M. W. Berry, J. Kogan, Text mining. Applications and Theory. West Sussex, PO19 8SQ, UK: John Wiley & Sons, 2010.

[7] G. G. Chowdhury, Natural language processing. Annual review of information science and technology, 2003, 37(1), 51-89.

[8] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

[9] Kukich, K. (1992). Techniques for automatically correcting words in text. Acm Computing Surveys (CSUR), 24(4), 377-439.

[10] López-Hernández, J., Almela, A., & Valencia-García, R. (2019, December). Automatic spelling detection and correction in the medical domain: A systematic literature review.

In International Conference on Technologies and Innovation (pp. 95-108). Springer, Cham.

[11] A. Avdić, U. Marovac, D. Janković, Normalization of Health Records in the Serbian Language with the Aim of Smart Health Services Realization, Facta Universitas, Series Matematics and Informatics, 2020, 35(3), 825-841.

[12] Yazdani, A., Ghazisaeedi, M., Ahmadinejad, N., Giti, M., Amjadi, H., & Nahvijou, A. (2020). Automated misspelling detection and correction in persian clinical text. Journal of digital imaging, 33(3), 555-562.

[13] Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3), 171-176.

[14] Balabaeva, K., Funkner, A. A., & Kovalchuk, S. V. (2020, June). Automated Spelling Correction for Clinical Text Mining in Russian. In MIE (pp. 43-47).

[15] Lai, K. H., Topaz, M., Goss, F. R., & Zhou, L. (2015). Automated misspelling detection and correction in clinical free-text records. Journal of biomedical informatics, 55, 188-195.

[16] A. Avdić, U. Marovac, D. Janković, Automated Labeling of Terms in Medical Reports in Serbian, Turkish Journal of Electrical Engineering & Computer Sciences, 2020, 28(6), 3285-3303.

[17] Ostrogonac, S. (2018). Modeli srpskog jezika i njihova primena u govornim i jezičkim tehnologijama (Doctoral dissertation, University of Novi Sad (Serbia)).

[18] Batanović, V., Ljubešić, N., Samardžić, T., & Petrović, M. M. (2020). Otvoreni resursi i tehnologije za obradu srpskog jezika. Primena slobodnog softvera i otvorenog hardvera.