

# РЕЦЕНЗИЯ

на дисертационен труд на тема: „*Автоматично резюмиране на правни текстове*“  
представен от **Валентин Венков Змийчаров**  
за придобиване на ОНС „доктор“ по професионално направление: 4.6. Информатика и  
компютърни науки, докторска програма Софтуерни Технологии – Откриване на знания  
Научен ръководител: проф. д-р Иван Койчев, Катедра “Софтуерни технологии”, ФМИ,  
Софийски университет „Св. Климент Охридски“  
от проф. д-мн Галя Ангелова, ИИКТ-БАН

Със заповед № РД 38-278/09.06.2025 г. на Ректора на СУ „Св. Климент Охридски“ съм определена за член на Научно жури по процедурата за защита на дисертационен труд „*Автоматично резюмиране на правни текстове*“ по професионално направление 4.6. Информатика и компютърни науки, докторска програма „Софтуерни Технологии – Откриване на знания“ с автор Валентин Венков Змийчаров. На първото заседание на Научното жури, проведено на 12.06.2025 г. съм избрана за рецензент. Настоящата рецензия е изготвена в съответствие със Закона за развитие на академичния състав в Република България (ЗРАСРБ), Правилника за неговото прилагане, както и с Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности в СУ „Св. Климент Охридски“.

## 1. Представени материали по процедурата за защита

За рецензиране е предаден пълен комплект материали, предвидени от нормативните разпоредби: (i) дисертационен труд и автореферат на български и английски език; (ii) научни публикации представящи резултатите на дисертационния труд и декларации от съавторите за равен принос, както и подробни наукометрични сведения за изданията, в които са публикувани свързаните с дисертационния труд статии; (iii) справка за покриване на минималните национални изисквания относно ОНС „доктор“; (iv) справка от система за откриване на плагиатство, протокол за проверка на оригиналност и декларация от докторанта за оригиналност на дисертационния труд; (v) материали от предзащитата (протоколи от катедрени заседания, вътрешна рецензия, доклад от научния ръководител за готовност за защита); (vi) пълен набор от административни документи свързани с протичане на докторантурата (заповеди за зачисляване, трансформиране, отчисляване, съответни доклади и становища от научния ръководител); (vii) автобиография на кандидата и негови дипломи (за ОКС „бакалавър“ и ОКС „магистър“), както и информация за дипломната му работа за получаване на магистърска степен; (viii) списък на направени промени във финалната версия на дисертационния труд след предзащитата. Предаденият комплект материали е подготвен с голямо старание и отговаря на изискванията на Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности в СУ „Св. Климент Охридски“.

Приложените наукометрични сведения за изданията „Сборник трудове на международната конференция RANLP“, CEUR-Workshop proceedings и „Сборник

трудове на международната конференция CLIB” показват, че в годините на публикуване на статиите на кандидата тези издания са били реферирани от Скопус и имат Scopus Journal Rank (SJR). Приемам статиите, представени от г-н Змийчаров, за трудове в категорията „публикации в издание със SJR без IF“. Отчитайки допълнителния коефициент 3 за умножение на точките при пресмятане на показател Г7, указан за професионално направление 4.6 в Заключителните разпоредби на Правилника за прилагане на ЗРАСРБ, приемам справката за изпълнени минимални национални критерии. (Към нея могат да се добавят още няколко точки в показател Д11 поради появили се цитирания от независими автори, които са видими в Google Scholar). По този начин се удовлетворяват количествените изисквания на Правилника за прилагане на ЗРАСРБ и Правилника за условията и реда за придобиване на научни степени и за заемане на академични длъжности в СУ „Св. Климент Охридски”.

## **2. Данни за кандидата Валентин Змийчаров**

Според представената автобиография г-н Змийчаров е роден през 1992 г. в Пазарджик. Завършва Математическа гимназия в родния си град, след което следва в СУ-ФМИ бакалавърска и магистърска програми, а впоследствие започва и докторантура. В дипломната си работа за магистърска степен „Автоматично разделяне на големи количества изображения в класове с помощта на дълбинно обучение“, защитена през 2017 г., кандидатът избира актуална тема от изкуствения интелект и демонстрира познаване и разбиране на проблемите, знания и умения за използване на невронни мрежи над големи масиви от учебни данни (770хил. изображения на 57 вида животни), и експертиза на професионален програмист при създаване на класификатор на изображения. В последните години Валентин Змийчаров е съумявал едновременно да изгражда успешна кариера на ръководител на екипи и проекти в ИТ индустрията, да съдейства на ФМИ в учебните дейности, да организира хакатони и състезания и да се ангажира с дарителски каузи за ученици от Математическата гимназия в Пазарджик. Имам и лични впечатления от него, от кратък учебен курс преди 8 години и от скорошен подкаст за Изкуствения интелект. Г-н Змийчаров е компетентен и активен експерт и успешен ръководител в ИТ-индустрията, директор на образователния отдел в СофтУни.

## **3. Съдържание и постижения на дисертационния труд**

**Актуалност.** Темата „Автоматично резюмиране“ беше едно от главните направления в компютърната лингвистика през 2017 г. при започване на докторантурата, тъй като автоматичното резюмиране е важна технология за откриване на информация във все по-нарастващия обем от текстове, които се създават ежегодно (включително в юридическата област). Прилаганите подходи бяха главно статистически техники за обработка на текста. В последните години обаче настъпиха радикални промени в компютърната лингвистика, с навлизането и усъвършенстването на подходите за дълбинно обучение: появиха се големите езикови модели и с тях се оформиха нови

понятия, очаквания и изисквания за ‘актуалност’, ‘оригиналност’, ‘резултат’ и ‘принос на дисертация’. Авторът се справя с това предизвикателство, като интегрира базови подходи и иновативни инструменти за решаване на поставената задача.

**Структура и обхват на дисертационния труд.** Представената дисертация има 114 страници, организирани в 6 глави, библиография с 58 заглавия на цитирани литературни източници, списък на фигурите, списък на таблиците и декларация за оригиналност.

Първа глава ‘*Въведение*’ съдържа кратък увод в полето, важноста на задачата и мотивация за изследването, а именно: да се извлече съществена информация от текста, като при това се гарантира, че смисълът е запазен. Представени са целите и задачите на дисертационния труд, като характерът на изследването е предимно научно-приложен.

Втора глава ‘*Обзор на предметната област*’ започва с терминологичен речник на специализирани понятия, който съдържа обяснения на 28 термина, използвани в дисертацията. След общо представяне на областта ‘обработка на естествен език’, авторът разглежда по-подробно техниките за векторизация на текста и обработки чрез невронни мрежи: вграждания на думи, рекурентни невронни мрежи, трансформатори и предварително обучени езикови модели (BERT, T5, големи езикови модели GPT). Тази част е написана компактно и интелигентно с препратки към литературни източници от последните години. Следва въвеждащ обзор на областта автоматично резюмиране на текстове, където са представени и основни предизвикателства при практически решения в системи за резюмиране. Споменати са достъпни набори от данни за резюмиране и в секция 2.6.4 са въведени подходи за оценка на автоматичното резюмиране – човешка и автоматична оценка. В секцията 2.7 ‘*Свързани разработки и съществуващи решения*’ се разглеждат 4 разработки включително многоезични над набор данни от швейцарския федерален съд. Към глава 2 бих отпредила някои бележки, например терминологичният речник да се изнесе извън основния текст на дисертацията и да се добави още литература относно най-нови резултати в използването на големи езикови модели за обработка на правни текстове (макар че е практически невъзможно да се обхване бурното развитие в тази област за последната година). В секция 2.6.4 би следвало да има по-обширна дискусия и главно, да се дадат примери за измерване на успеваемост с различни подходи, които да послужат като контекст за преценка доколко е успешен предложението от автора оригинален компонент и каква е степента на значимост на резултатите от проведените експерименти. Друг забележка е, че обзорът е ограничен до разясняване на понятия, които се използват в следващите глави на дисертацията. А би могло да се погледне по-широко към темата за резюмиране (тъй като става дума за обзор на предметната област) и да се дискутира въпросът за задаване на дължината на резюметата като параметър (отворена тема за цялата област), или да се спомене и за появата на индустриални продукти за анализ на юридически текстове с технологии на изкуствения интелект, вкл. интерфейси за резюмиране.

В трета глава ‘*Нов корпус от данни за правни текстове и техните резюмета*’ се разглежда създаването на корпус от учебни и тестови текстове, които са използвани в различни експерименти за автоматично резюмиране. Първичните документи са взети от официалния сайт на Европейския съюз, където са публикувани правни документи от

различен вид и на различни теми, преведени на всички официални езици на ЕС. Описан е процесът по събиране и почистване на текстовете, включително изхвърляне на неподходящи документи. Крайният резултат за българския език е 1776 резюмета (средна дължина 543 думи) със съответните им пълни документи (средна дължина 21152 думи), предимно текстове на регламенти, директори, решения и становища. Към корпуса са добавени многоезикови преводи на основните документи за всички 24 езика, при което е получен корпус от паралелни текстове съдържащ наличните документи. Създаденият корпус е публикуван в платформата Hugging Face и е публично достъпен.

Четвърта глава *‘Разработени и адаптирани подходи за резюмиране на текстове’* представя основния оригинален резултат на дисертационния труд – компонентите за автоматично резюмиране и експерименти с тях и със създадения корпус. Първо се избира основен подход, спрямо който ще се сравняват генерирани с различни техники резюмета; като ‘базово’ резюме се дефинират началните изречения в текста, до достигане на съотношение в дължината както в корпуса изработен по сайта на ЕС. Създадените прототипи за резюмиране, чието поведение се тества и наблюдава, са базирани върху следните техники за резюмиране: (i) екстракция на изречения с най-висок *TF-IDF*; (ii) екстракция на изречения по алгоритъма *TextRank* с използване на векторни вграждания; (iii) екстракция на изречения с помощта на генерирани от BERT вектори и *K-means* за пресмятане на центроид; (iv) екстракция на изречения с помощта на генерирани от LegalBERT вектори и *K-means* за пресмятане на центроид; (v) екстрактивен и абстрактивен подход чрез алгоритъма *PreSumm* с използване на BERT; (vi) абстрактивен подход за резюмиране с езиков модел T5 и (vii) генериране на резюмета с използване на GPT API на OpenAI. Въведен е нов подход за екстрактивно резюмиране: *StructExtSum*, който използва структурата на документа и избира съществени сегменти от текста. Освен информация за йерархичната структурата и наредбата на изреченията в текста се вземат предвид характеристиките *TF-IDF* за изреченията, както и стойностите пресметнати чрез алгоритъма *TextRank*. След пресмятане на метриката ROUGE-1 (съвпадение на думи в генерираното и оригиналното резюме) се изчислява ROUGE-1 на изреченията и с помощта на линейна регресия се пресмятат стойностите ROUGE за най-„добрите“ изречения. В тази глава се описват също особености при резюмиране за отделните езици (български и английски), които са забелязани при експериментите, както и техническата имплементация, използваните библиотеки и програмни среди, създадената инфраструктура за съхранение и обмен на данни, и възможностите за отворен достъп и повторно използване на кода и учебните данни.

Пета глава *‘Резултати от проведените експерименти’* резюмира оценки, наблюдения и изводи, направени при тестването на разнообразните техники за резюмиране над създадения корпус. Наблюденията са важни, понеже са направени над един и същи учебни данни и позволяват сравнение на успеваемостта на различните техники. Всички използвани подходи подобряват базовия модел, което не е учудващо като се има предвид, че той е възможно най-простият. Екстрактивното резюмиране с BERT вектори и *K-means* постига 36.82 ROUGE-1, което означава, че 1/3 от думите в генерираното резюме съвпадат с думи от референтното резюме, направено от човек

(златен стандарт). Съгласна съм с автора, че 39.27 ROUGE-1 за абстрактивно резюме с алгоритъма T5 е много добър резултат (Таблица 5). Подходът Bert-K-means дава най-добри резултати и при резюмиране на научни статии на английски език. Предложеният нов подход StructExtSum дава най-добри ROUGE-1 резултати при сравнение с базовия подход, TF-IDF, SlavicBERT+K-Means за български (30.04) и английски (38.99) правни текстове. Представени са и резултати от сравнение на поведението на различните подходи над корпуса с текстове на 24 езика; TextRank и GPT са най-добри за всички езици. Част 5.6 ‘Човешка оценка на качеството на резюметата’ помага за осмисляне на сложността на поставената задача и представените подходи за решаването ѝ.

Шеста глава ‘*Заключение и бъдеща работа*’ представя приносите на дисертацията и изброява насоки за бъдещи изследвания и създаване на системи за автоматично резюмиране.

#### **4. Приноси на дисертационния труд**

Авторът формулира приносите в Глава 6, стр. 96 на дисертацията. Бих нарекла изброените приноси научно-приложни, тъй като според мен научният принос се постига чрез резултати, засягащи основни и принципни постановки на научната област. Приносите според мен са в три главни направления: *(i)* създаден е нов корпус от учебни/тестови данни, който е публичен и може да се използва за експерименти и разработка на приложения на официалните европейски езици на ЕС; *(ii)* предложен е нов метод за екстрактивно резюмиране StructExtSum, който използва структурата на документите за идентифициране на ключови елементи в правни текстове и показва по-добри резултати от други техники за резюмиране; *(iii)* извършени са множество експерименти и сравнителни тестове с различни техники за резюмиране, с използване на новия корпус или други подходящи данни, и са представени обобщения и изводи за поведението на алгоритмите. Както беше казано по-горе, включването в Част 2 на обзор относно състоянието на системите за резюмиране и техните постижения като успеваемост, би позволило да се позиционира настоящата работа в контекста на областта и вероятно би помогнало да се открие оригиналността на получените резултати.

Считам, че анализът сам по себе си не е научен принос или постижение – а основа, върху която се стъпва при създаване на нови постижения, и затова не коментирам тези дейности изброени при приносите в Глава 6. Съгласна съм с ефектите и приносите в по-общ смисъл, които авторът е изброил: развиват се инструменти за обработка на български текст и те могат да се интегрират в различни приложения с цел подобряване на достъпа до информация. Създаването на ресурси е принос към дигитализацията у нас.

#### **5. Аprobация на резултатите**

Резултатите от изследванията на докторанта са представени в 4 публикации в изданията „Сборник трудове на международната конференция RANLP”, CEUR-Workshop proceedings и „Сборник трудове на международната конференция CLIB”. В годините на публикуване на статиите на кандидата тези издания са били реферирани от Скопус и

имат Scopus Journal Rank (SJR). Позволявам си да изкажа една препоръка към докторанта, и тя е да се стреми да публикува на конференции, където се събират много специалисти в областта на компютърната лингвистика и е възможно да се получат мнения от опитни изследователи и специалисти в тясната предметна област.

## **6. Автореферат**

Авторефератът съдържа 46 страници и представя адекватно обхвата и резултатите на дисертационния труд. В него са интегрирани резюмета на отделните глави на дисертацията, които представят целите и задачите на труда, обзора на предметната област, създаването на нов корпус от документи, експериментите с разнообразни техники за автоматично резюмиране, техните резултати и направените изводи, приносите и насоките за бъдеща работа. Авторефератът е оформен в съответствие с изискванията на Правилника за условията и реда за придобиване на научни степени и заемане на академични длъжности в СУ „Св. Климент Охридски“.

## **7. Заключение**

Като цяло сумарната ми оценка за дисертационния труд, автореферата и постиженията на докторанта е положителна, макар да отчитам направените съществени забележки. Получените резултати съответстват на целите и задачите, поставени в дисертационния труд. Без съмнение в процеса на работа кандидатът (както и колектива съавтори – млади учени) са придобили обширни познания в областта, натрупали са солидни програмистки умения при създаване на прототипи, както и способност за самостоятелна научна дейност. Потвърждавам, че представеният труд и публикуваните научни резултати отговарят на изискванията на ЗРАСРБ, Правилника за приложението му и съответния Правилник на СУ „Св. Климент Охридски“ за придобиване от кандидата на образователната и научна степен „доктор“ в научната област 4. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки, и че са покрити минималните национални изисквания. Не е установено плагиатство в представените по конкурса научни трудове.

Като се има предвид казаното по-горе, подкрепям присъждането на образователната и научна степен „доктор“ на Валентин Венков Змийчаров и предлагам на уважаемите членове на Научното жури да гласуват в подкрепа на такова решение.

11.08.2025 г.

Член на Научното жури:

/проф. дмн Галя Ангелова/