Рецензия

на дисертационен труд на тема „**Софтуерна система за автоматизирана обработка на неструктурирани данни от социалните мрежи**“ за присъждане на образователна и научна степен „**доктор**“ по професионално направление 4.6. „ Информатика и компютърни науки“, докторска програма „Информатика“

Автор на рецензията: доц. д-р Т.Атанасова, катедра „Информатика“, Икономически университет-Варна

Автор на дисертационния труд: ас. Борис Банков, катедра „Информатика“, Икономически университет-Варна

Основание за написване на рецензията: Заповед № 06-1177/ 2.05.2018 г. на Ректора на Икономически университет-Варна за състав на научното жури и решение на научното жури на първо заседание, проведено на 3.05.2018.

I. Данни за докторанта

Борис Банков е завършил висшето си образование в бакалавърска и магистърска степен в Икономически университет- Варна, специалност „Информатика“. Допълнителна квалификация като Front End Developer е получил в Академия Телерик. Зачислен е в докторантура на самостоятелна подготовка в докторска програма „Информатика“, считано от 21.3.2017 г. Поради успешно изпълнение на индивидуалния си план, в заседание на факултетния съвет от 2.5. 2018 г. той е отчислен с право на защита.

През периода май- септември 2012 г. Борис Банков е работил като уеб дизайнер във фирма 158ltd, Варна. От 1.09.2014 е назначен като асистент в катедра „Информатика“. Водил е занятия по дисциплините: Интернет технологии, Компютърни мрежи и комуникации, Е-бизнес 1-ва част, Компютърни технологии в рекламата и медийните комуникации и др.

Взема участие в три научно- приложни проекта. Спечелил е призово място в конкурс за най-добър млад научен работник, проведен от фирма ISIC Depot през 2016 г.

Ползва свободно английски език (ниво С2).

II. Общо представяне на дисертационния труд.

Темата на дисертацията е актуална и значима, тъй като е свързана с основни съвременни проблеми като работа с огромни обеми данни, често без семантичен модел, както и с откриване на потенциално полезни зависимости в тях.

Научната теза на докторанта е, че „обработката на неструктурирани данни от социалните мрежи дава възможност за разкриване на нова и полезна за бизнеса информация“. В тази връзка е логично поставена целта на дисертационния труд- да се създаде модел на софтуерна система за обработка на неструктурирани данни, извлечени от социалните мрежи и да се предложи концепция за реализацията й.

Дисертацията е в обем от 187 страници, разпределени във въведение, три глави, заключение, списък на информационни източници, приложения.

Във въведението е обоснована актуалността на темата. Посочени са научната теза, цел на дисертацията и задачите за постигането й, както и предмет и обект на изследване.

Първа глава започва с дефиниции за основни използвани понятия- данни, информация, знания. Сравнени са структурираните и неструктурирани данни, тъй като съществуват значителни разлики в обработката им. Разгледани са социалните мрежи като основен източник на неструктурирани данни и е направена класификация на публикациите в тях. Тъй като предмет на дисертацията е обработката на неструктурирани текстови данни, във втора точка са представени подходи за текстов анализ и по-специално класификация и клъстеризация. Внимание е отделено на метода за клъстерен анализ к-средни, който докторантът по-късно прилага. За определяне на контекста, в който се използва един термин се разглеждат възможностите на невронните мрежи и по-специално реализацията на моделите за обучение Continuous Bag-of_Words (CBOW) и Skip-gram. Разгледани са основни проблеми и проучени софтуерни приложения за работа с неструктурирани данни. Установено е, че не са открити софтуерни инструменти за анализ и извличане на знания от български текст. С

извода за необходимостта от система за обработка на данни от неструктурирани интернет източници на български език се прави преход към втора глава.

Втора глава е посветена на модела на предлаганата система. Разгледан е обхватът й чрез представяне на целта и задачите, които ще решава, както и изискванията, на които да отговаря. Конкретизирана е целта на системата, а именно- обработване на текстови масиви на **български език**, за да се открие „скрит слой информация“ в потребителските съобщения. Предложен е концептуален модел на системата за обработка на неструктурирани данни от социалните мрежи. За всеки от шестте модула са дадени функциите и конкретните особености. Данните се извличат в реално време от социалната мрежа Twitter. Приложени от докторанта са следните алгоритми:

- за първична обработка на потребителските съобщения- разпознаване на думи, изчистване на пунктуационни знаци, филтриране на хипервръзки и др.;

- за определяне честотата на срещане на термините;

- за определяне на значимостта и информационната стойност на потреби телските съобщения;

- за намиране на сходство между думите в текста;

- за определяне на активни и изключване на неактивни клъстери. Този ал горитъм е авторска разработка.

- за установяване на най-вероятни съседни думи.

Допълнително се използват и онтологии за анализ на контекста на даден термин.

Моделът може да се адаптира за извличане на данни от различни социални мрежи.

Трета глава е посветена на софтуерната реализация на предлаганата система в конкретна фирма и за избрана социална мрежа- Twitter. След анализ на използваемостта на редица софтуерни средства, е направен обоснован избор на технологични решения за разработка на системата. Предложен е софтуер с отворен код. За реализация е приложена методологията за бърза разработка на приложения. В последната точка от трета глава са представени резултатите от апробацията на системата- от извлечените данни са формирани 10 клъстерни

групи и избрани по 5 произволни думи от всяка група, което показва разпределението на дискусионните области. Докторантът е разработил и сайт, където може да се разгледат извлечените данни от Twitter и визуализират резултатите от работата на отделните модули.

Заключението представлява обобщение на съдържанието на дисертационния труд.

Списъкът с използвана литература включва 105 заглавия. Използвани са също 47 интернет източника.

В приложенията (14 бр.) е даден програмен код за реализация на основни алгоритми, използвани в системата за обработка на неструктурирани данни от социалните мрежи.

III. Преценка на структурата и съдържанието на дисертационния труд

Структурата на дисертационния труд е стандартна и балансирана по глави. Първа глава има теоретичен характер, втора- научно-приложен, а трета е с изцяло приложен характер. Разработени са в логическа последователност.

Съдържанието е информационно наситено и стегнато. Стилът на писане е научен и ясен.

Целта, която си поставя докторанта с разработката на дисертацията е постигната. Създаденият и апробиран модел на система за обработка на неструктурирани данни от социалните мрежи е приложение на най-нови информационни технологии, включително от областта на изкуствения интелект. Употребата им е подходяща особено като се има предвид големия обем на данните, извлечими от социалните мрежи и техния неструктуриран вид.

Постигнатите научни и научно- приложни резултати са убедително доказателство за възможностите на докторанта и добрата му професионална подготовка. Трудът има теоретично и практическо значение, особено с работата върху текст на кирилица.

Текстът е онагледен много добре. Дадени са пояснящи примери. Цитирането е коректно.

Авторефератът съдържа необходимите елементи и отразява коректно съдържанието на дисертацията.

IV. Идентифициране и оценяване на научните и научно – приложните приноси в дисертационния труд

Достойнствата на дисертационния труд позволяват предложените приноси да се разпределят в две групи - научни и научно-приложни. Обогатяващи теорията в изследваната научна област са следните приноси:

1. Проведено е изследване на източниците, методите, моделите и технологиите за обработка на неструктурираните данни в дигитална среда.

2. Анализирани са неструктурираните данни в социалните мрежи и е въведена класификация на различните градивни единици на потребителските съобщения в четири от най-популярните платформи за социално взаимодействие в Интернет.

С научно- приложен характер са приносите:

3. Създаден е модел на софтуерна система за автоматизирана обработка на неструктурирани данни, който е пригоден към спецификата и характера на българския език, използван в публикациите в социалните мрежи.

4. Разработен е алгоритъм за определяне на активни клъстери, при клъстеризация на постоянна емисия от данни, чрез въвеждане на променлива, която отчита времето за възстановяване на неактивни клъстери на база на тяхната значимост.

5. Разработена е концепция за реализиране и частично внедряване на предлаганата софтуерна система в организация от медийната сфера.

V. Публикации и участие в научни форуми

Публикациите на докторанта по темата на дисертационния труд са представени с 2 научни статии и 2 научни доклада. Статиите са самостоятелни, а едната е на английски език. Единият от докладите е в съавторство. Чрез посочените публикации изследванията на докторанта са получили публичност.

VI. Критични бележки и препоръки.

Сред научните интереси на докторанта, посочени в творческата биография, забелязах областта „Изкуствен интелект“ (ИИ). Препоръчвам на Борис Банков да задълбочи изследователската си работа в тази посока, тъй като научната работа по ИИ в момента е изключително интензивна и плодотворна, още повече и поради успешно използваните интелигентни методи в дисертацията.

VII. Въпроси към дисертанта

1. Как може да се модифицира предлаганата система, за да работи с данни от други социални мрежи?

VIII. Заключение

Предоставеният за рецензиране научен труд има качествата на сериозно научно изследване и отговаря на изискванията за дисертация по докторска програма „Информатика“. Докторантът демонстрира способност за самостоятелни научни изследвания с необходимата професионална квалификация. За това аз ще гласувам убедено с „да“ по процедурата и предлагам на уважаемото научно жури да присъди образователната и научна степен „доктор“ по професионално направление 4.6. „ Информатика и компютърни науки“ на Борис Банков.

10.05.2018 г. Рецензент:

Гр.Варна (доц.д-р Т.Атанасова)

РЕЦЕНЗИЯ

за **дисертация** за присъждане на образователната и научна степен „**доктор**“
на тема „Софтуерна система за автоматизирана обработка на неструктурирани данни от социалните мрежи“
с автор **ас. Борис Иванов Банков**, Икономически университет Варна
от **проф. д-р Аврам Моис Ескенази**, ИМИ БАН, професор емеритус

1.Обща информация

Основание за написване на тази рецензия е зап. РД 06-1177 от 02.05.2018 на Ректора на Икономически университет Варна, както и решението на Научното жури от неговото първо заседание. Дисертационният труд е в професионално направление 4.6 „Информатика и компютърни науки“, докторска програма „Информатика“.

2.Данни за дисертанта

Борис Банков е завършил Математическата гимназия във Варна през 2009, а след това – ИУ Варна като бакалавър по информатика през 2013 и като магистър информатик през 2014. Освен това е завършил няколко други форми на обучение – английски по програма на Кеймбриджкия университет, кратък стаж в Университет в Санкт Петербург, Академията Телерик. Участвал е в републиканска студеннска олимпиада по програмиране и в конкурс за най-добър млад лектор. За кратко работи като уеб-дизайнер, а от 2013 и досега е хоноруван и след това редовен асистент в ИУ, катедра „Информатика“. Участва в обучението на студенти с лекции и упражнения по разнообразни информатични дисциплини: „Информатика“, „Интернет технологии“, „Офис приложения“, „E-business“, „Интернет продажби“ и др. Владее английски език. Освен свързаните с дисертацията, има и други научни публикации, както и участие в 3 научноизследователски проекта, от които един международен. Членува в НТС.

По отношение на процедурата по тази защита са **спазени напълно изискванията на нормативните документи**, а именно:

1. По чл. 24 от Правилника за приложение на ЗРАСРБ – докторантът **притежава** образователно-квалификационната степен "**магистър**" – диплом **рег.№** 14487/09.09.2014 г.

2. По чл. 25, т.1 – докторантът е **отчислен с право на защита**, както личи от приложената заповед на Ректора РД 06-1161/02.05.2018.

3. По чл. 25, т.2 – докторантът (очевидно) е **представил дисертационен труд**.

4. По чл. 26 (1) – докторантът е придобил право на защита, защото е **положил (с отличен успех) 4 изпита**, определени в индивидуалния учебен план.

5. По чл. 26 (2) – налице е изискуемото **решение на факултетния съвет** (протокол 35/02.05.2018), **цитирано в заповедта на Ректора (вж. началото на рецензията)**.

6. По чл. 27 (2) – **дисертацията има изброените в тази алинея атрибути**, вкл. **Декларация за оригиналност**, представена като отделен документ.

3. Общо представяне на дисертационния труд

Представеният труд съдържа 188 стр., разпределени в Съдържание, Въведение (а не **Увод**, както е изискването на чл.27(1) на Правилника към ЗРСАРБ), 3 глави, Заключение, Библиография, 14 едностранични приложения и списък на публикациите по дисертацията. Страниците са по обем приблизително със стандартните 1800-2000 знака и като отчета, че приложенията са 14 стр, констатирам, че трудът е около горната граница на дисертациите, свързани с информатиката. Работата е добре подредена и илюстрирана с различни форми (таблици, диаграми, снимки на екрани). Езикът е професионален. Има **добър баланс** на **образователните и научни** (**приносни**) елементи. Глава 1 е 43 страници, което е около 27% от съдържателната част, което намирам за нормално.

Както вече казах в т.2 по-горе, изискуемите от чл.27(2) на ППЗРСАРБ **структурни и съдържателни атрибути** на дисертацията **са налице**. Отбелязвам още **4 характеристики на труда, които оценявам положително**:

- в работата е използван актуален софтуерен и математически инструментариум;

- тази актуалност се отнася и до темата на дисертацията, особеност, която считам, че не е необходимо да доказвам, особено в светлината на дебатите, изслушванията и мерките от буквално последните седмици, свързани със социалните мрежи;

- работата е с повишена трудност поради необходимостта от познания не само по информатика, а в решаващи моменти – от обоснован избор и квалифицирано прилагане на специфични математически инструменти (клъстеризация, невронни мрежи);

- ползваните източници са добре подбрани, номиналният им брой 152 е очакваният от една сериозна дисертация, езиковото им разпределение – също според очакванията (предимно на английски, около 20 на български и 1 на руски). Разпределението им във времето е също представително. Не ми е много ясен принципът на отделяне на някои източници в категория Интернет. Ако е спазвал стриктно правилата, авторът винаги **би следвало да е цитирал директно** (а не през друг източник). В такъв случай не си представям, че за всичките 105 (извън Интернет категорията) е минал без Интернет. Всъщност далече **по-сериозната забележка** тук е, че има източници, които фигурират в библиографията, но не можах да открия позовавания на тях (такива са например [2], [6], [8], [9], [14]).

4. Преценка на структурата и съдържанието на дисертационния труд.

Въведението на дисертацията е кратко, но съдържателно и включва всичко, което следва да се очаква – теза, основна цел, задачи, обект, предмет. Тъй като я считам за най-важна и определяща, специално оценявам **целта** – „да се създаде модел на софтуерна система за обработка на неструктурирани данни, извлечени от социалните мрежи и да се

предложи концепция за нейната реализация“ – смятам, че е ясно и (както съм отбелязвал и в други рецензии - **важно** за една работа по информатика) **конструктивно** определена. **Четирите задачи**, чрез които се постига тази цел, също оценявам като добре обмислени и водещи до постигане на целта. Тук е мястото да направя и една **забележка**. Извън всяко съмнение е, а и авторът го казва ясно, определяйки **предмета** (с.7), че изследванията и резултатите се отнасят само до **един вид неструктурирани данни – текстови**. За структурирани данни **изобщо** - става дума в обзорната част. При това положение не е ясно защо в заглавието липсва уточнението „текстови“, а и в цялата дисертация в повечето случаи се премълчава. Според мене решението е било просто – нормативната уредба позволява уточнение на темата във всеки момент, а в дисертацията е трябвало да се добави в началото, че за краткост под „неструктурирани данни“ ще се разбират „неструктурирани текстови данни“, ако изрично не е казано друго.

Глава 1, съгласно очакванията има обзорно-аналитичен характер. От нея (както и от вече спомената и оценена библиография) ясно проличава, че нещата по отношение на **образователната компонента** стоят много добре – докторантът познава състоянието на тематиката, както и на необходимите му научни резултати от по-отдалечени области ([96], [100], [104 и доста други), демонстрира способност целенасочено да ги анализира и като следствие от това - да си направи най-подходящия за своите цели избор. Подходено е комплексно – предмет на анализ са неструктурираните данни и методи за обработването им, както и наличните софтуерни инструменти за целта. Това прави извода, завършващ главата (с.50) обоснован, а с това - и **претенциите на автора за първия му принос**. Все пак бих заменил във формулировката му „доказана“ с „показана“, а и считам, че е доста неравностоен в сравнение с по-сериозните приноси 3 и 4 **Същото мисля** и за претенциите му за **втория принос**. По-точно, принос е втората част на формулировката – че е въведена класификация на различните градивни единици на потребителските съобщения, докато първата - за анализа - е просто естествена и необходима предпоставка.

Глава 2 решава задачите, които носят **най-сериозните приноси** на работата, които признавам – **3 – създаденит модел на софтуерна система за автоматизирана обработка на неструктурирани данни, който отчита спецификата на българския език от социалните мрежи**, както и 4 - разработеният **алгоритъм за определяне на активни клъстери** при клъстеризация на постоянна емисия от данни, чрез въвеждане на променлива, която отчита времето за възстановяване на неактивни клъстери на база на тяхната значимост. Преминаването към клъстеризация в реално време изисква периодично да се добавят нови обекти, което води и до промяна на центровете на клъстерите, има проблем и с неравномерното разпределение на съобщенията. Могат да възникват и нови клъстери. От друга страна може клъстер да съдържа само стари данни, т.е. физически е добре да се съхранява като история или да се деактивира. Наред с **активността** на клъстер е важно и свойството **значимост**.

За отправна точка в изследванията на дисертанта се цитира конкретен алгоритъм, който ползва променлива време за престой, както и фактор на разпад, намаляващ с времето. Така значимостта е произведение от брой вектори и фактор на разпад. Тук допълнително се пазят исторически вектори за центровете, чиято промяна води до намаляване теглата на по-старите клъстери. Други автори използват функция на затихване, която е времето за намаляване значимостта на половината от първоначалната стойност. Но тук има нещо повече, за всеки сегмент от време се натрупва сумарна статистика, която да се използва от анализатора за разбиране на хоризонта за разширяване на клъстер и броя клъстери. В трета статия, с цел да се формализира тази концепция, се използват микроклъстери и пирамидална структура на записите според времето, при която се пести място, но не се отчита значимостта.

Възможни са и офлайн алгоритми на различни етапи на онлайн емисията, които са по-удобни за изчисление. Те използват натрупаното в сегменти от времето. Чрез K-means се определят малки ядра на клъстери и след това се строи дървовидна структура чрез сегментиране на ядро. Така от класическа офлайн клъстеризация чрез сумарно представяне на елементи се преминава към coresets-клъстеризация, която в момента се счита за най-ефективната техника за онлайн клъстеризация.

На базата на този кратък обзор от 4 статии за различни техники (които техники в статиите са формализирани и анализирани с оценки и експерименти), **дисертантът прави предложение за модел - да се разделят фазите на обработка при постоянна емисия на текстове от социалната мрежа на онлайн и офлайн.**

Идеята е в зависимост от динамиката на предметната област на коментарите в социалните мрежи, при работа с постоянна емисия от данни да се поддържа списък с активни клъстери, да се следи за аномалии и поява на нови клъстери или да се терминират неактивни клъстери. Във втората част на глава 2.5 се предлага алгоритъм за определяне на активни и неактивни клъстери чрез въвеждане на променлива, която отчита времето за възстановяване на неактивни клъстери на база на тяхната значимост. Поради сложността на визираните техники и тяхната недостъпност за широка употреба, се оценява значимост на базата на максимална допустима значимост за даден клъстер. Предложението е мотивирано с конкретен прост пример (и още един подобен в приложение 1) само за $N = 3*n$ (с.92) и не е ясно как работи напр. при по-голяма стойност. Както се вижда и от глава 3 това предложение не е реализирано програмно, нито е сравнено с други методики (с.144); приносът би изглеждал още по-обоснован, ако това беше направено.

На идейно ниво (без да се формализира) е и второто предложение как да се обособява нов клъстер на базата на някое от изброените видове разстояния.

Глава 3 има за задача да покаже, че така предложените в предходната глава модели, методи и алгоритми са използваеми. Избрана е подходяща организация (медийна група Черно море), както и Twitter като социална мрежа. Добре е, че докторантът си е направил труда да проучи и стъпи на световни стандарти, когато избира своя софтуерен

инструментариум. Описал е процеса много подробно и илюстрирано. Представените експериментални данни от експлоатацията на реализираната част на софтуерната система ми дава основание да приема и претенцията за **петия принос за частично внедряване на предлаганата софтуерна система в организация от медийната сфера.**

От направения дотук анализ следват и моите изводи в съответствие с **чл.34(2) и (3)** от Правилата на ИУ, а именно:

- дисертационният труд **съдържа необходимите оригинални научни и научно-приложни резултати**; несъмнени са и задълбочените **познания** на докторанта в предметната област на дисертацията, а и в други области, необходими за получаването на резултатите;

- изхождайки от дефиницията за **монография** - научно-изследователски труд, посветен на цялостно комплексно изследване само на един значим научен проблем (тема, задача, експеримент), разработен самостоятелно или колективно, съдържащ значими оригинални теоретични изследвания на особено високо научно равнище и притежаващ разгърната структура“ (Правила за устройството и дейността на редакционните колегии, както и за реда и изискванията относно публикуването на научните издания на ИУ, чл.3, т.3), потвърждавам, че от една страна може да се говори за **съответствие на труда с тази дефиниция**, а от друга става дума за стъпка в **решаването на теоретически интересен и практически полезен проблем, активно атакуван по света.**

По **структура и съдържание оценявам автореферата положително**, смятам, че изпълнява коректно предназначението си да представи същественото от дисертациятата и съпътстваща полезна информация.

5. Идентифициране и оценяване на научните и научно-приложните приноси в дисертационния труд

В предходната точка приносите бяха идентифицирани и директно свързани с отделните части на труда. Там те бяха оценени като основателни, а някои от тях - като сериозни и значими. Това ми дава основание да заявя, че **трудът съдържа достатъчно резултати**, които са **оригинални приноси** (чл.6 (3) от ЗРСАРБ), една част от които са научни, а друга - научно-прприложни.

6. Преценка на публикациите

Общият брой публикации, представени от докторанта във връзка с дисертацията, е 7. Всички са самостоятелни, което прави отговора на въпроса за самостоятелността на приносите очевиден. От тези работи две са в национални списания, а останалите са доклади, публикувани в сборници от научни мероприятия у нас, част от тях - международни. Като отчитам, че двете списания са индексирани, констатирам, че **изискванията на чл.35(1) т.4** на Правилата на ИУ за 1 самостоятелна статия и 2 доклада, едно от които в реферирано или индексирано спнисание, **са преизпълнени.** Оценявам

като положителен, но недостатъчен факта, че едната статия е на английски, което пък прави още по-естествена препоръката ми за публикуване в авторитетни международни издания в чужбина.

7. Критични бележки и препоръки

Трудът е очевидно актуален и перспективен и затова съм малко озадачен, че не забелязах формулирани в явен и ясен вид насоки за **бъдещо развитие** както в теоретичен, така и в по-приложен аспект. Обект на по-незначително мое съмнение е използването на „единица“ във връзка с клъстерите, не съм тесен специалист, но се консултирах, за да разбера, че „елемент“ е по-подходящо. Не срещнах упоменати coresets – считаните за най-модерни напоследък клъстер техники, а и неявно ползвани от докторанта.

8. Въпроси към дисертанта

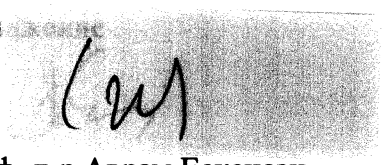
Бих искал да чуя разяснения в отговор на забележката ми относно разделянето на източниците в библиографията на две категории (краят на т.3 от тази рецензия). Също така идеи за по-нататъшно развитие на резултатите от дисертацията (т.7 по-горе).

9. Заключение

Рецензираният труд е оформен коректно и прегледно и демонстрира от една страна задълбочено познаване от страна на докторанта на предметната област, а от друга - съдържа оригинални научни и научно-приложни резултати, постигнати след сериозен обзор, анализ на релевантната научна информация, предлагане на модел и специфични алгоритми с оригинална идея и достигане до прототип на система, тестван в реални условия в определени рамки. Критичните ми бележки нямат определящ характер за крайната оценка. Следователно, рецензираният дисертационен труд има съдържание и форма, които го правят съответстващ на нормативните изисквания за дисертация за получаване на образователната и научна степен "доктор". Смятам, че са удовлетворени изискванията на ЗРСАРБ, на Правилника на МС по ЗРАСРБ и Изискванията на ИУ Варна. Това ми дава основание да му дам **положителна** оценка и да препоръчам на почитаемото научно жури **да присъди** на ас. Борис Иванов Банков образователната и научна степен „доктор“ в професионално направление 4.6 „Информатика и компютърни науки“, докторска програма „Информатика“.

София, 18.05.2018

С уе



проф. д-р Аврам Ескенази