Р Е Ц Е Н З И Я

От: проф. д.ик.н. Желю Владимиров, професионално направление 3.7 „Администрация и управление“

На: дисертационен труд от Глория Венциславова Христова, на тема „*Автоматизирана система за анализ на онлайн комуникация с клиенти чрез машинно самообучение и обработка на естествен език – структура, изграждане и бизнес приложения*“, за присъждане на образователната и научна степен “Доктор” по професионално направление 3.8 „Икономика“, научна специалност „Аналитични изследвания върху данни“.

Основание за рецензията: Заповед РД-38-326/04.07.2022 г. на Ректора на Софийски университет „Св. Климент Охридски“

1. Информация за дисертанта

Глория Венциславова Христова е завършила Софийската Математическа Гимназия (СМГ) „Паисий Хилендарски“ (2004-2012 г.). През 2016 г. се дипломира в бакалавърската програма „Стопанско управление“ на Стопански факултет към СУ „Св. Климент Охридски“. През 2018 г. защитава магистърска теза на тема „*Дизайн и разработка на автоматизирана система за определяне на потребителското отношение към софтуерни приложения чрез техники за обработка на естествен език и машинно самообучение*“ и завършва с отличие магистърската програма „Моделиране на големи данни в бизнеса и финансите“. Глория Христова продължава обучението си в Стопански факултет на СУ като редовен докторант в докторската програма „Аналитични изследвания върху данни /Data Science/“.

От септември 2021 г. Глория Христова заема длъжността „асистент“ към катедра „Статистика и иконометрия“ в Стопански факултет, където води занятия по „Количествени методи в управлението“, „Машинно самообучение за бизнес и финанси“ и „Основи на текстовия анализ и обработката на естествен език“. През последните три години тя е участвала в четири национални научни проекта, както и в няколко семинара и международни конференции. Участвала е и като ментор в състезания в сферата на аналитичните изследвания.

2. Обща характеристика на представения дисертационен труд

Представеният дисертационен труд се състои от увод, три глави, библиография и приложения с обем от 281 страници, а без приложенията и информационните източници – от 252 страници. Ползвани са 220 информационни източника на английски. Основният текст съдържа 26 таблици и 27 фигури, като още 5 таблици и 10 фигури са дадени в приложенията. Съгласно справката за изпълнение на критериите, докторант Глория Христова има 320 точки при необходим минимум от 150 точки.

В **Увода** е разкрита актуалността на проблематиката, свързана с разпространението на „големите данни“ и необходимостта от тяхното съхранение, обработка и използване от компаниите. Посочено е, че извличането, обработката и интерпретацията на данните придобива централна роля в условията на онлайн комуникация, тъй като данните от тази комуникация са отражение на *основните проблеми, които вълнуват клиентите* (с. 12).

В тази връзка, основната **цел** на изследването е дефинирана като: „Създаване на автоматизирана система за анализ на основните теми, които вълнуват клиентите, както и за анализ на удовлетвореността им от предоставените услуги в контактен център с

комуникация на български език“ (с. 16). Тази цел е конкретизирана в три големи задачи с осем под-задачи. Нейното постигане се основава на използването на различни *аналитични техники от сферата на обработката на естествен език* и *машинното самообучение.* **Обект** на дисертационния труд е онлайн чат комуникацията между клиенти и служители на голяма банка в България, а **предметът** са основните теми, които вълнуват клиентите, както и тяхната удовлетвореност от чат комуникацията с контактния център. Данните са генерирани в контактния център на тази банка.

Защитаваната **теза** е че онлайн чат комуникацията между клиенти и оператори в контактен център може да бъде ефективно извлечена, структурирана и анализирана с помощта на техники за обработка на естествен език и машинно самообучение чрез изграждането на автоматизирана система за анализ (с. 19). Проверката на тезата се реализира посредством тестване на 4 хипотези.

3. Оценка на получените научни и научно-приложни резултати

Глава първа включва подробен и критичен литературен преглед на актуални изследвания след 2016 г. относно анализа и извличането на знания от онлайн чат комуникация с фокус върху обслужването на клиенти в контактен център. В точка 1.2. акцентът е върху: описанието на тази комуникация, моделирането на теми и разрешаването на класификационни задачи. Табл. 2 (с. 58-59) представя резюме на методите за извличане на знания от онлайн чат комуникация. Направен е извода, че не са открити изследвания, анализиращи онлайн чат комуникация на български език (с. 61).

Актуална картина на обработката на естествен език и приложението на аналитични техники върху текстови данни на български език е дадена в *точка 1.3*. Анализирани са изследвания по създаването на езикови ресурси на български език, главно на групата Бултрибанк (BulTreeBank group). Показани са инструменти и системи за обработка на текст на български език (токъънизация, стеминг, лематизация и др.). Разкрити са практически приложения от анализ на текстови данни на български език за решаване на социални, икономически и бизнес проблеми (Табл. 3, с. 79-80). Очертани са етапите в обработката на естествен език – от използването на правила през статистически методи и машинно обучение до трансферното обучение.

Точка 1.4 представя методите за моделиране на теми от текстови данни в онлайн чат комуникация. Авторът е избрал алгоритъмът LDA *(Латентно разпределение на Дирихле)* за моделиране и анализ на основните теми. Основните подходи за анализ на настроението от текстови данни (използването на лексикони, машинно самообучение или комбинация от двата) са разкрити в *т. 1.5.* За прогнозиране удовлетвореността на клиенти от онлайн чат комуникацията са избрани методите за машинно самообучение и логистичната регресия.

Втора глава включва подробно изложение на методологията на изследването. Обяснено е *създаването на 4 модула*, изграждащи автоматизираната система за анализ на комуникацията на клиента с контактен център. *Модул I* съдържа алгоритъм за първоначално прочитане и структуриране на данните от онлайн чат комуникацията с помощта на програмния език Python. Показани са етапите за превръщане на данните от суров вид във вид, подходящ за обработка и количествен анализ. Общите стъпки в алгоритъма за трансформация и нормализиране на текстовите данни са представени на Фиг. 4 (с. 125).

Предложената методика в *Модул II* представлява авторска комбинация от различни подходи и техники за моделиране на теми в онлайн чат комуникация. Основните техники за *нормализация* на текста преди превръщането му в числов вид са дадени на Фиг. 5 (с. 130). След нормализацията на данните, върху тях е приложена „токъънизация“ на ниво дума, в

резултат на което всяка реплика от даден чат е представена като набор от токъни (низ от думи). Посочени са *нивата на репрезентация* на текста в зависимост от това дали се ползват цели чатове или само някои реплики. За извличането на основните теми в онлайн чат комуникация е използвана TF-IDF векторизация на данните посредством библиотеката „gensim“ в Python. Обоснован е изборът на LDA, който разглежда всеки документ като вероятностно разпределение на набора от всички теми (Фиг. 6, с. 140). Описани са две метрики за оценка на резултатите от приложението на LDA (коефициент на сложност и метрики за кохерентност – с. 145).

Модул III представя създаден от автора автоматизиран метод за прогнозиране на удовлетвореността на клиента от услугите в контактен център. Създадени са три нива на репрезентация на данните: извадка от **цели** чатове между клиент и оператор („Извадка 1“); извадка, включваща всички реплики само на клиента (“Извадка 2”); и извадка, включваща само финалните реплики на клиента в комуникацията (“Извадка 3”). Тествани са три различни алгоритъма - Бернулиевият наивен Бейсов модел, логистична регресия и класификация с опорни вектори. Използвана е $\boldsymbol{k}$-кратна крос валидация за оценка на моделите, а така също прецизност, чувствителност, F1 и F-beta спрямо класа „лош рейтинг“ с по-голяма тежест на чувствителността на модела.

Обобщение и визуализация на резултатите са представени в Модул IV, който има за цел да улесни тяхното тълкуване. Както пише авторът, докато Модулите I, II и III представляват „back-end“ на системата, Модул IV е нейният „front-end“ – т.е. това, с което крайния потребител взаимодейства директно (с. 180).

В **трета глава** автоматизираната система е апробирана чрез анализ на онлайн чат комуникацията между клиенти и оператори в контактния център на голяма финансова институция в България. Извадката от данни се състои от 38 166 чата, периодът е от 22.01.2019 г. до 01.04.2021 г., а общият брой реплики е 466 118. Финалният брой на чатовете след нормализиране на данните в Модул I е 37 529 наблюдения. Обект на анализа са чатове с клиент и един оператор, които са 29 614 или 78.9% с почти равномерно разпределение на репликите. Представени са данни за средна продължителност на чата, средна скорост на отговора и др. Анализът на основните темите, вълнуващи клиентите е осъществен с помощта на LDA алгоритъма. Използвани са двата количествени измерителя за избор на оптимален брой теми (Cv кохерентност и коефициент на сложност).

Създадени са общо 1 470 модела за оценка средната стойност на Cv кохерентността за модели с от 2 до 50 теми. Достига се до извода, че между 15 и 20 теми кохерентността на модела достига стабилна средна стойност – 16 теми след пета филтрация и с най-малък брой думи в речника (Табл. 15, с. 200). За петата филтрация е изчислена и стойността на коефициента за сложност, като отново анализът показва спад на стойността между 14-16 теми (Фиг. 23, с. 203).

В точка 3.3.1 се анализира създадения модел с оптимално представяне върху Извадка I на всяка една тема поотделно. На Фигура 24 (с. 212) са демонстрирани четири много общи теми в дискусиите с клиентите: *1. Кредитиране, 2. Дигитално банкиране, 3. Касови операции и 4. Картови продукти (кредитни и дебитни карти).* Подобен анализ е направен и върху Извадка II и III. Изведено е, че най-голямо качество на темите се постига чрез моделиране на чатовете в тяхната цялост (Извадка I), което води до *отхвърляне на Хипотеза 2*. В същото време оптималните резултати, получени върху Извадка I водят до *потвърждаване на Хипотеза 1*. Това означава, че репликите на операторите в комуникацията са необходими за извличането на важните за клиентите теми.

В т. 3.6. е показано как с помощта на предишни данни и машинно самообучение могат да се уловят характеристики на комуникацията между клиент и оператор, които с достатъчна точност да сигнализират удовлетвореност или неудовлетвореност на клиента от

услугата на контактния център (Хипотеза 3 и Хипотеза 4). Използвани са три различни репрезентации на данните (Извадка 1, 2 и 3), като са създадени са **216** модела, прогнозиращи тази удовлетвореност. Според получените резултати, чатовете сами по себе си съдържат достатъчно сигнали за създаване на модел, прогнозиращ удовлетвореността на клиента в края на комуникацията, което води до потвърждаване на Хипотеза 3. Също така, финалните реплики на клиента съдържат ценна информация относно настроението му в края на комуникацията според метриката F-beta, което води до потвърждаване на Хипотеза 4.

Заключението представлява рекапитулация на проведеното изследване, като са посочени и някои изследователски перспективи в тази област.

4. Оценка на научните и научно-приложни приноси

Формулирани са няколко приноса, които могат да бъдат обобщени по следния начин: (1) Представена е актуална картина в областта на обработката на естествен език в България, както и на възможностите за анализ на текстови данни на български език; (2) Предложена е цялостна методология за обработка и анализ на данни от онлайн чат комуникация с клиенти на български език; (3) Създадена е автоматизирана система за извличане на знания от онлайн чат комуникация с клиенти в контактен център, която би могла да се обновява в реално време при постъпване на нови данни; (4) Конструирана е методика за анализ на важните за клиентите теми, включваща основни техники за обработка и моделиране на данни с различни нива на репрезентация; (5) Създадена е методика за прогнозиране удовлетвореността на клиента от онлайн чат комуникация, която се базира само на текстови характеристики и граматическа информация; (6) Изградена е методика за интерпретация на получените резултати, която може да се използва и в други индустрии, в които се генерират подобни данни; (7) Настоящето изследване и сред първите, в които се прави анализ и извличане на знания от онлайн чат комуникация между клиенти и служители (на български език).

5. Оценка на публикациите по дисертацията

Основните компоненти на дисертационния труд са апробирани в пет публикации, от които три на международни конференции и две във списания в периода 2020-2021 г. Три от публикациите са индексирани в световно признати бази данни (Scopus и WoS).

6. Оценка на автореферата

Авторефератът е в обем от 54 страници. В него са разкрити основните моменти от дисертационния труд в синтезиран вид и като такъв отговаря на изискванията.

7. Критични бележки, препоръки и въпроси

Нямам конкретни бележки по съдържанието на текста. Единствено от редакционна гледна изглежда излишно повторението на хипотезите на с. 101, 102, 112 и 114.

8. Заключение

Дисертантът е извършил значителна работа по анализ на релевантната литература, провеждане на емпирични експерименти, проверка на хипотези посредством адекватни метрики за оценка на получените резултати и приложените методи. Проблемите са актуални и направените изводи добавят стойност към наличното знание по темата. Демонстрирано е много добро познаване на изследваната проблематика, акуратност в представянето на теоретичните подходи, коректност при позоваването на използваните публикации.

Резултатите потвърждават наличието на ефективни начини за анализ и извличане на информация от онлайн комуникация с клиенти чрез техники за обработка на естествен език и машинно самообучение. Предложена е автоматизирана система за анализ на основните теми, които вълнуват клиентите, както и за анализ на тяхната удовлетвореност от предоставените услуги в контактен център с комуникация на български език. Обосновани са практическите ползи за бизнеса от създаването на подобна система. Направените приноси са лично дело на автора. С този научен труд Глория Христова демонстрира качества на сериозен изследовател и отговорно отношение към научната дейност

Всичко това ми дава основание да предложа на уважаемото научно жури да присъди на Глория Венциславова Христова образователната и научна степен “Доктор” по професионално направление 3.8 „Икономика“, научна специалност „Аналитични изследвания върху данни“.

10.08.2022 г.
София

Рецензент:
проф. д.ик.н. Желю Владимиров