СТОПАНСКИ ФАКУЛТЕТ
SOFIA UNIVERSITY ST. KLIMENT OHRIDSKI est. 1888
FACULTY OF ECONOMICS AND BUSINESS ADMINISTRATION

Катедра „Статистика и Иконометрия“

ГЛОРИЯ ВЕНЦИСЛАВОВА ХРИСТОВА

АВТОМАТИЗИРАНА СИСТЕМА ЗА АНАЛИЗ НА ОНЛАЙН КОМУНИКАЦИЯ С КЛИЕНТИ ЧРЕЗ МАШИННО САМООБУЧЕНИЕ И ОБРАБОТКА НА ЕСТЕСТВЕН ЕЗИК – СТРУКТУРА, ИЗГРАЖДАНЕ И БИЗНЕС ПРИЛОЖЕНИЯ

АВТОРЕФЕРАТ

на дисертационен труд
за присъждане на образователна и научна степен „Доктор“ по
професионално направление: 3.8. ИКОНОМИКА,
Научна специалност: Аналитични изследвания върху данни /Data Science/

Научен ръководител: **Доц. Д-р Боряна Богданова**

София, 2022

Дисертационният труд е одобрен и насочен за публична защита от катедрен съвет на катедра „Статистика и иконометрия“ към Стопански факултет при Софийски университет „Св. Климент Охридски“, проведен на 20 юни 2022 г. (Протокол № 331).

Авторът на дисертационния труд е редовен докторант към катедра „Статистика и иконометрия“, съгласно решение на Факултетния съвет към Стопански факултет, заповед № РД 20-321/05.02.2019 г. на Ректора на СУ.

Дисертационният труд се състои от увод, изложение (в три глави), заключение, библиография и приложения с общ обем 281 стр., 37 фигури и 31 таблици.

Цитираната литература включва 220 източника на български и чуждестранни автори.

Брой публикации на автора по темата: 5.

Съдържание

I. Обща характеристика на дисертационния труд[1]

1.1. Актуалност на темата

През XXI-ви век възникна понятието **„големи данни“** (*big data*), което развълнува бизнеса, посочи нови хоризонти и отвори врати на „науката за данните“ (*data science*). Според прогноза на Statista, обемът от данни[2], генериран в световен мащаб, ще се увеличи до над 180 зетабайта до 2025 г. **Овладяването на данните и внедряването на алгоритмични решения** с помощта на машинно самообучение (*machine learning*) вече не са конкурентни предимства, а необходимост на компаниите с оглед на това да останат конкурентоспособни. Бизнесът е в **период на трансформация**, като движещата сила са науката за данните и машинното самообучение, които позволяват **автоматизиране на процеси, извличане на ценни изводи за поведението на клиентите, съкращаване на разходи, подобрение на продукти и услуги** и много други.

В условия на все по-нарастваща конкуренция, клиентът и неговата удовлетвореност са поставени на централно място [1], а данните се превръщат в ценен актив за бизнеса - източник на информация за **темите, които вълнуват клиентите**, както и техните **предпочитания**, **нужди**, **проблеми**, **настроения** и цялостна **удовлетвореност** от продуктите и услугите [2]. Според проучване, проведено от Dimension Data, включващо 1 351 организации от отрасли в 80 държави [3] - **84%** от компаниите, които работят в посока **увеличение на клиентската удовлетвореност**, отчитат увеличение на приходите си в резултат на техните усилия.

Успешен пример за **подход, в който клиентът и темите, които го вълнуват, както и нуждите и предпочитанията му са на централно място,** е този на „The New York Times“. Статия, публикувана през януари 2021 г.[3], разказва за опита на компанията в

[1] Текущият дисертационен труд е разработен в рамките на първата за Стопански факултет докторска програма, финансирана от бизнеса (*Industry-funded PhD program*). Бизнесът е достигнал фаза, в която осъзнава необходимостта от интегриране на бизнес решения, основаващи се на целево-финансирани научни изследвания. В този контекст, освен научната постановка и класическите атрибути, които следва да се съдържат в увода на дисертационния текст, е изчерпателно мотивирана и нуждата от разработване на научни изследвания, чрез които да бъдат решавани актуални бизнес проблеми.

[2] Прогнозите на Statista относно ръста в обема от данни, генериран в глобален мащаб са достъпни на следния адрес: https://www.statista.com/statistics/871513/worldwide-data-created/

[3] Статията е достъпна в дигитален формат на следния адрес - https://open.nytimes.com/we-recommend-articles-with-a-little-help-from-our-friends-machine-learning-and-reader-input-e17e85d6cf04

имплементирането на решения, базирани на машинно самообучение, с цел доближаване до нейния основен клиент - читателя. Медийният гигант прилага алгоритми за анализ на темите, които вълнуват читателите онлайн, като целта е да им бъдат препоръчвани статии, които възможно най-точно да отговарят на техните интереси.

Компании от различни индустрии опитват да достигнат до своите клиенти по всички възможни начини, с цел да разберат по-добре техните нужди и да отговорят на тях възможно най-ефективно и бързо [2]. В този контекст **данните от комуникация с клиента** под най-различна форма - било то **интервюта, отзиви, данни от социални мрежи** и други започват да имат централна роля [2], [4]. В тази комплексна картина на динамични отношения между компанията и клиента, през 2020 г. се намесва още един фактор и това е **пандемията от COVID-19**. Тя се превърна в катализатор на още по-голяма метаморфоза на бизнеса, изразяваща се в дигитализация на множество процеси и услуги. Една от най-отчетливите промени е по отношение на **комуникацията с клиента.**

Много компании, чиито бизнес е ориентиран към крайния потребител, са изправени пред предизвикателствата на изцяло **онлайн комуникация с клиента**. Възможностите за физически срещи с клиентите и проучване на тяхната удовлетвореност от услугите чрез класически подходи (задълбочени интервюта и фокус групи) са силно ограничени поради епидемичната обстановка [5]. **Възниква нуждата** от откриването на нови подходи за анализ на поведението, удовлетвореността, темите, които представляват интерес и нуждите на клиента, изцяло на база на онлайн комуникацията с него.

Сред засегнатите индустрии в световен мащаб, **огромна трансформация претърпява банковата сфера**. Доклад на KPMG [6] посочва, че в резултат на пандемията, банките са принудени да дигитализират множество процеси и услуги много по-бързо от планираното, междувременно опитвайки се да останат „близо до клиента“. Натискът на конкуренцията в лицето на финтех индустрията задълбочава проблема. Паралелно с това, обемът от клиентски данни, с които банките разполагат, се увеличава значително - прогнозите в банковия сектор за периода 2021-2026 г. са за комбиниран годишен темп на растеж на ползването на технологии за анализ на големи данни от около 22.97% [7].

Различни проучвания преди и след пандемията от COVID-19 ясно посочват **тенденциите за дигитализация на общуването** между крайния потребител и компаниите. През 2020 г., сравнено с предходната година, с цели **62%** се е увеличил броят на компаниите,

които обслужват клиенти с помощта на чат платформата WhatsApp и с **51%** се е увеличило ползването на Facebook messenger с такива цели[4]. В този бум на онлайн комуникация в разнообразни индустрии**, контактните центрове са на първа линия** в обслужването на клиента. Огромният обем от комуникация се генерира именно в тях и ролята им на **директна връзка между клиента и компанията**, им отрежда все по-важно и централно място.

Едни от най-важните аспекти за бизнеса в подобна комуникация са **темите**, които вълнуват клиента, както и **удовлетвореността** му от обслужването в контактния център [8], [9], [10], [11]. Подобни данни са **отражение на основните проблеми и теми, които вълнуват клиентите**. Това знание може да служи като индикатор за това, какво е най-важно за клиента, с какви проблеми се е сблъскал, от какво той се интересува най-вече, а оттам и какво би повлияло върху мнението му и решението му да продължи да използва продуктите и услугите. В комуникацията с контактен център, от **първостепенно значение е да се следи и удовлетвореността** - дали операторът е успял да изпълни искането и дали клиентът е останал удовлетворен [12], [9].

В този контекст, пандемията от COVID-19 постави допълнителна тежест, тъй като превърна „онлайн обслужването“ в тема с **огромно значение за бизнеса**. В период на световна пандемия, когато физическите срещи са силно ограничени, клиентът е „далеч“ и вече **не може да бъде достигнат по традиционните начини**. **Дигитализацията на общуването** доведе до ръст в обема от данни, генерирани в контактните центрове – както беше споменато по-рано, голямо **увеличение бива наблюдавано в използването на системи за чат** с цел комуникация с клиента. Подобни текстови данни могат да са с огромен обем, генерират се в реално време и са силно неструктурирани. Това прави непосилно ръчното преглеждане и обобщаване на темите, представляващи интерес за клиента, неговите проблеми, удовлетвореност или други важни индикатори за отношенията между компанията и него.

Актуалните предизвикателства пред бизнеса всъщност създават и нови възможности - дисертационният труд е фокусиран върху количествените методи за **анализ на текстови данни** (*text analytics*) и **машинното самообучение, приложени върху онлайн чат комуникация с клиента.** Подобни методи се характеризират с бързина, позволявайки

[4] Резултати от изследването са достъпни на следния електронен адрес: https://www.statista.com/statistics/1260555/top-messaging-channels-for-customer-service-by-yoy-growth/

анализ на големи данни и автоматизиране на изследването на тенденциите сред клиентското мнение, темите, които вълнуват клиентите, нагласите, нуждите и удовлетвореността им от продуктите и услугите, които една компания предоставя [9], [13], [14]. Критичният литературен преглед в рамките на дисертационния труд показва, че малко научни изследвания са фокусирани в тази област, **въпреки че в последните две години онлайн чат комуникацията с клиента се превърна в особено актуална тема за индустрията.**

Необходимостта от установяване на **надеждна методология за количествен анализ на онлайн чат комуникация с клиента** е в основата на дисертационния труд. Освен пандемията от COVID-19, фактор, който допълнително усложнява и прави този проблем още по-актуален е фактът, че в такъв тип количествени анализи голяма роля играе езика, на който бива провеждана текстовата комуникация [15]. Инструменти за обработка на естествен език, както и разнообразни автоматизации и системи за анализ съществуват **най-вече за текстови данни конкретно на английски език. Българският език далеч не е толкова застъпен** в научната литература с фокус върху количествения анализ и обработка на текстови данни. Критичният литературен преглед, анализ и синтез в Глава I на дисертационния труд показват, че **към този момент не са публикувани изследвания, посветени на анализа на онлайн чат комуникация с клиенти на български език.**

Всичко казано дотук подчертава **важността и необходимостта не само от установяването на надеждна методология за количествен анализ на онлайн чат комуникация с клиенти,** но и **актуалността на тази тема, когато става въпрос конкретно за данни на български език.** В ера на икономическа и бизнес трансформация в глобален план, когато огромен обем от разнообразни типове текстови данни се генерират всеки ден, количественият им анализ предстои да привлича все по-голям интерес, поради потенциала, който носи за бизнеса [13]. Тези тенденции неминуемо засягат и България, налагайки необходимостта от критичен литературен преглед на **възможностите и актуалното ниво на развитие в научната област, занимаваща се с анализ на текстови данни конкретно на български език.**

1.2. Обект и предмет на дисертационния труд

Обект на дисертационния труд е **онлайн чат комуникацията между клиенти и служители на голяма компания в България.** Данните от тази комуникация се генерират

в контактния център на компанията, която оперира в банковия сектор. В сфери като банковата, COVID-19 доведе до някои радикални промени в процесите, услугите и отношенията с клиента. Ако до вчера клиентите посещаваха офис на банката, за да зададат своите въпроси, днес не малка част от тях предпочитат да се свържат с нея онлайн[5]. Това доведе до огромен обем от комуникация с клиента, генерирана директно в контактните центрове на банките, а служителите в тях се оказаха на първа линия по време на кризата.

Предмет на дисертационния труд са **основните теми, които вълнуват клиентите, както и тяхната удовлетвореност от комуникацията с контактния център (посредством чат).** Темите, които клиентите засягат в комуникацията си с компанията, са ценен източник на информация за техните интереси, проблеми и нужди. Междувременно, в ера на дигитализация на комуникацията, когато клиентът и обслужването са поставени на централно място, от огромна важност е да се проследи **удовлетвореността на клиента от общуването** [4] със служителите на първа линия.

1.3. Цел и изследователски задачи на дисертационния труд

Основна цел на дисертационния труд:

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.

Набор от различни **аналитични техники от сферата на обработката на естествен език** (адаптирани към спецификите конкретно на българския език) и **машинното самообучение** е използван, за да се постигне основната цел. Много важен елемент в така дефинираната основна цел е **фокуса върху анализа на текстови данни на български език**. Литературният преглед в рамките на труда показва, че това е относително неразработена научна област. **Липсва систематизиран литературен обзор и синтез на постигнатото до**

[5] В изследване на S&P, проведено през февруари и март 2021 г., 52% от респондентите заявяват, че посещават физически офиси на банки по-рядко след началото на пандемията от COVID-19. Още интересни наблюдения върху поведението на клиентите в банковия сектор са достъпни в статия на следния адрес - https://www.spglobal.com/marketintelligence/en/news-insights/research/pandemic-pushes-customers-out-of-branches-banks-ramp-up-closures

момента в сферата. Знанията са разпръснати и фрагментирани и **в рамките на дисертационния труд е направен опит да се внесе структура в това научно поле**.

Въз основа на изводите от проведения критичен литературен преглед, анализ и синтез в Глава I на дисертационния труд са формулирани следните **изследователски задачи:**

I. Провеждане на критичен литературен преглед, анализ и синтез в следните три направления:
   - I.1. Методически литературен преглед по отношение на обекта на изследване в дисертационния труд - онлайн чат комуникация с клиенти, генерирана в контактен център.
   - I.2. Обстоен литературен преглед на научни изследвания с фокус върху анализа на текстови данни на български език - развитие в сферата и практически приложения.
   - I.3. Методически литературен преглед по отношение на предмета на изследване в дисертационния труд - моделиране на теми в текстови данни и анализ на удовлетвореността на клиента.

II. Разработване на автоматизирана система за анализ на онлайн комуникация с клиенти чрез машинно самообучение и обработка на естествен език. Фокус върху **структурата** и **изграждането** на подобна система:
   - II.1. Създаване на методика за структуриране и обработка на изследвания тип данни (**Модул I** на системата за анализ).
   - II.2. Създаване на методика за извличане на основните теми, които вълнуват клиентите (**Модул II**). Адаптация на конкретни техники в контекста на работата с данни на български език.
   - II.3. Създаване на методика за анализ на удовлетвореността на клиента от комуникацията му с контактен център посредством онлайн чат (**Модул III**). Адаптация на конкретни техники в контекста на работата с данни на български език.

III. Апробиране и приложение на практика на създадената система върху извадка от данни в рамките на реален бизнес казус. Фокус върху **бизнес приложенията** на подобна система:
   - III.1. Формиране на ясни и конкретни предложения за възможните начини, по които добитата информация от изследвания тип данни може да бъде полезна за бизнеса (**Модул IV**).

III.2. Отчитане на възможности за подобрение на създадената система и бъдещи перспективи.

1.4. Изследователска теза и хипотези

Един **основен въпрос** играе ключова роля във формулирането на тезата на дисертационния труд, а именно: **„Има ли ефективни начини за анализ и извличане на ценна информация от изследвания тип данни с помощта на методи от сферата на анализа на текст и машинното самообучение?“.** Макар и отговорът да изглежда на пръв поглед напълно ясен и праволинеен, основна роля играе въпроса, до колко въобще могат да бъдат извлечени подобни знания от този тип данни. Те се характеризират с много особености и са доста по-нестандартни за обработка, сравнено с други типове текстови данни. Откриването на отговор на този въпрос ще даде яснота до каква степен извлечената информация от подобен тип данни **добавя стойност и е ценна за бизнеса**. Така бива дефинирана и **тезата**, залегнала в основите на дисертационния труд:

Онлайн чат комуникацията между клиенти и оператори в контактен център, представлява неоползотворен богат източник на информация за отношенията на клиентите с компанията. Тази информация може да бъде ефективно извлечена, структурирана и анализирана с помощта на техники от сферата на обработката на естествен език и машинното самообучение, с цел изграждането на автоматизирана система за анализ, имаща ценни приложения в бизнеса.

Развитието на тезата се осъществява в рамките на следните изследователски въпроси, възникнали на база на литературния преглед в Глава I:

1. **Съществуват ли такива и кои са ефективните методи от сферата на машинното самообучение и анализа на текст, които могат да бъдат приложени върху обекта на дисертационния труд с цел извличане на основните теми, които вълнуват клиентите?**

Възникнали хипотези, засягащи този въпрос:

Хипотеза 1: *Традиционният метод за моделиране на теми - Латентно разпределение на Дирихле (Latent Dirichlet Allocation - LDA), би могъл да бъде приложен върху обекта на дисертационния труд, като това ще доведе до постигането на резултати, които са надеждна основа за разработването на работещо решение.*

Хипотеза 2: *При приложението на алгоритъм за моделиране на темите, които вълнуват клиента, ще бъдат постигнати оптимални резултати (по отношение на яснота и качество на получените теми), чрез ползване на извадка от реплики само на клиента, в сравнение с ползването на целия чат между него и оператора.*

Доколкото е известно на автора, **съществува само едно изследване** [16]**, което се доближава най-много до настоящото по отношение на част от поставените цели и анализиран тип данни**, а именно - прогнозиране на крайна оценка за **цялостната удовлетвореност** на клиента от онлайн чат комуникацията му с оператор в контактен център. Изследователите стигат до важно заключение - **съществуващите лексикони на настроението (*sentiment lexicons*) за анализ на текстови данни на английски език не са приложими в сферата на обслужването на клиенти и не водят до постигане на удовлетворителни резултати.** Наличието на подобни ресурси за български език е силно ограничено. Изводът на авторите в [16] води до възникването на **неизследвана ниша в научната литература** и формиране на следния изследователски въпрос:

2. **Може ли да се направи допускането, че чатовете, сами по себе си, съдържат достатъчно сигнали, така че чрез тях да се създаде модел, прогнозиращ удовлетвореността на клиента в края на комуникацията, който да бъде надеждна основа за разработването на работещо решение?**

От отговора на този въпрос зависи и възможността да се даде отговор на един доста по-генерален въпрос, а именно: **Възможно ли е да се прогнозира удовлетвореността на клиента, базирайки се основно на неговия текстови диалог (чат) с оператора?** На база на изводите на други изследователи в сферата, публикувани в [16], отговорът на този въпрос е потърсен в дисертационния труд чрез приложението на алтернативни техники, базирани изцяло на методи за машинно самообучение. Формирани са следните две хипотези:

Хипотеза 3: *Възможно е създаването на автоматизиран модел, базиран на машинно самообучение, който прогнозира удовлетвореността на клиента само на база на неговия чат с оператор. Създаденият модел ще се характеризира с оптимално представяне спрямо т.нар. „наивна“ прогноза, която не се базира на машинно самообучение.*

Хипотеза 4: *В прогнозирането на удовлетвореността на клиента могат да бъдат постигнати оптимални резултати чрез ползване на извадка, състояща се от финалните реплики на клиента в онлайн чат комуникацията (в сравнение с ползването на други алтернативни нива на репрезентация на данните).*

1.5. Изследователски метод

Изследователският метод, приложен в дисертационния труд, е изцяло ориентиран към **количествените методи за анализ на текстови данни**. Приложени са техники изцяло от сферата на обработката на естествен език и машинното самообучение с оглед постигане на основната цел на труда. Такива иновативни подходи имат потенциала **ефективно да допълнят конвенционалните техники за анализ на поведението и удовлетвореността** на клиента и да **повишат степента на рационалност** при вземането на управленски решения. Освен за бързина, тези методи биват използвани и с цел създаването на автоматизации, благодарение на които е възможно да бъдат осъществявани навременни анализи в подкрепа на вземането на по-добри управленски решения.

1.6. Обхват на изследването

Осъществената експериментална работа (в Глава III - Емпирично изследване) е ограничена до изследването на онлайн чат комуникация, генерирана в контактния център конкретно на една компания - голяма банка, оперираща в България. **Това ограничение не засяга** дисертационния труд по отношение на постигнатото в Глава I и Глава II, а именно – задълбочен преглед, анализ и синтез на напредъка в сферата, както и създаване на методология за анализ на онлайн чат комуникация в контактен център (на български език).

Макар че в емпиричното изследване е разгледан конкретен казус в банковата сфера, предложената от автора **методика за интерпретация на получените резултати** е приложима във всички останали индустрии, в които подобен тип данни биват генерирани в резултат на бизнес процесите на компанията. Именно **тази приложимост и възможност за екстраполация на стъпките в интерпретацията и анализа на подобен тип данни и в други индустрии, очертава един от приносите на дисертационния труд**.

Освен **практическите приложения** за бизнеса, **заинтересована от настоящото изследване**, може да бъде общността от изследователи в областта на анализа на текстови

данни и приложенията на подобни техники с цел оптимизация и подобрение на процеси в управлението на отношенията с клиента. Заинтересовани от дисертационния труд могат да бъдат и изследователите, занимаващи се с анализ на текстови данни на български език, тъй като е предоставен интересен поглед върху развитието в тази научна сфера, актуалните възможности и бъдещи посоки за напредък. Не на последно място, настоящото изследване би било полезно за изследователската общност, фокусирана върху изучаването на онлайн чат комуникация, както и върху анализа на подобна комуникация, проведена конкретно в контактни центрове.

1.7. Структура на дисертацията

В Глава I е осъществен подробен литературен преглед, анализ и синтез на постигнатото в научни статии, фокусирани върху основния обект и предмет на изследване в дисертационния труд. Дефинирани са някои основни понятия и хипотези, като е обоснован и избрания аналитичен подход на база на постигнатото от други изследователи в сферата. В Глава I са очертани и празнини в съществуващата литература, които настоящото изследване си поставя за цел да запълни. **Обхватът на Глава I е по-широк**, тъй като са осъществени обширен литературен преглед, анализ и синтез, целящи да представят актуална картина на развитието в сферата на анализа на текстови данни на български език. В рамките на Глава II е изложена подробно всяка стъпка в методологията, следвана с оглед изпълнение на основната цел на дисертационния труд - изграждане на автоматизирана система за анализ на онлайн комуникация с клиенти чрез машинно самообучение и обработка на естествен език. В Глава III създадената система е апробирана в рамките на реален бизнес казус, като са извлечени ценни заключения относно приложността на системата и бъдещите посоки за усъвършенстването ѝ.

II. Кратко изложение на дисертационния труд

2.1. Литературен преглед (Глава I)

В рамките на Глава I е осъществен литературен преглед в следните направления:

- ✓ Актуални изследвания, посветени на анализа и извличането на знания от онлайн чат комуникация (чрез методи за машинно самообучение и обработка на естествен език). Специално внимание е обърнато на изследванията, които са в сферата на обслужването на клиенти и разглеждат чатове, генерирани конкретно в контактни центрове.
- ✓ Научни изследвания с фокус върху анализа на текстови данни на български език - развитие в сферата и практически приложения (обстоен литературен преглед).
- ✓ Основни подходи, актуални тенденции и наложени техники за анализ на настроението (*sentiment analysis*) и моделирането на теми (*topic modeling*) в текстови данни.

Основните резултати от проведения критичен литературен преглед, анализ и синтез са документирани в [17] и [18].

2.1.1. Анализ на онлайн чат комуникация

Литературният преглед на актуални статии (публикувани след 2016 г.[6]) в областта на анализа на текст, в които основен обект на анализ е онлайн чат комуникация, се фокусира върху следните въпроси:

- Какво е постигнато от изследователите в сферата в посока на разбиране на този тип данни и какво носят те като информация за бизнеса? Какви са практическите приложения от подобни анализи?
- Какви са специфичните особености на този тип данни, които биват взети предвид по време на приложението на техники за количествен анализ върху тях?
- Поставен е фокус върху основните техники за обработка на данните, както и тествани методики в избраните публикации.

[6] В прегледа е включена и една статия публикувана през 2015 г., поради това, че притежава много сходства с някои от целите, поставени в дисертационния труд.

Всяка една изследователска статия е анализирана подробно в дисертационния труд и е създадено **структурирано резюме в табличен вид** според избрани ключови характеристики на статиите – например, източник и език на данните, приложени техники за предварителна обработка и моделиране, използвани езикови ресурси и други. В рамките на изложението по-долу са споменати само **част от най-важните заключения**, направени вследствие на литературния преглед на изследвания, фокусирани върху количествения анализ и извличането на знания от онлайн чат комуникация.

Адресираните проблеми и области на приложение на изследванията са най-разнообразни, включвайки в това число: автоматизация на процесите в контактни центрове; повишаване на ефективността на диалогови системи; подобряване на клиентското преживяване в различни онлайн комуникационни и развлекателни платформи; разкриване на онлайн измами и престъпления (например, кражба на самоличност в интернет); подобряване на процесите в сферата на образованието чрез прилагане на иновативни техники за облекчаване на процеса на обучение и други.

Не е изненадващо, че повече от половината от прегледаните научни статии анализират онлайн чат комуникация на английски език. Разбира се, причината се крие в изобилието от публично достъпни данни на английски език, както и в наличието на много езикови ресурси на разположение за този език. **В рамките на литературния преглед не са открити изследвания, осъществяващи анализ на онлайн чат комуникация на български език, което очертава и възможност за напредък в тази посока**. Най-често срещаните източници на чат комуникация, използвани в изследванията, се оказват различните социални мрежи и форуми (чат стаи). Данни, генерирани в контактен център за обслужване на клиенти, са анализирани в малка част от прегледаните статии.

По отношение на използваните аналитични техники - прегледът сочи, че след 2018 г. методите за вграждане на думи (*word embeddings*) стават все по-широко разпространени в анализа на чатове. Сред класическите методи за машинно самообучение най-популярни са Наивния Бейсов модел (*Naïve Bayes*) и SVM (*Support Vector Machines*). Сред най-скорошните изследвания биват открити и приложения на т.нар. модели тип „Трансформатор“ (*Transformer models*). В моделирането на теми в онлайн чат комуникация, алгоритъмът LDA (*Latent Dirichlet Allocation*) или негови разновидности са сред най-широко използваните.

Повече от половината от всички разгледани статии разчитат на използването на разнообразни езикови ресурси. Някои от изследванията силно зависят от използването на подобни ресурси не само във фазата на предварителна обработка на данните, но също и за последващия статистически анализ. Последното означава, че репликирането на експериментите върху данни на други езици е затруднено, ако не и невъзможно, тъй като може да не съществуват подобни езикови ресурси за тях. Нещо повече - не е изненада, че в експериментите има извършени много задачи, изискващи човешка намеса (т.е. „ръчни“ задачи, като например, проверка на правописа или анотация). Бъдещите изследвания в научната област на анализа на текст трябва да бъдат посветени на откриването на нови методи, с които да се облекчи целия аналитичен процес и да се преодолеят до някаква степен трудните и тромави „ръчни“ задачи по време на анализа на данните. Приложението на трансферно обучение е стъпка към това, но само една от статиите, включени в прегледа, използва подобни методи - [10].

Прегледът разкрива, че много малко от изследванията предлагат практическа имплементация на приложените методи и техники под формата на аналитичен инструмент (система), който да може да бъде използван за целите на автоматизиране на процеса на анализ на текст. Такъв е открит само в две от разгледаните научни статии - [12], [19]**. Може да се заключи, че повече трябва да се направи в посока на разработването на цялостни аналитични инструменти за анализ на онлайн чат комуникация и възприемането на ефективни техники за визуализация на текстовите данни.**

Вследствие на този литературен преглед, може да бъде заключено, че **има само две изследвания, които се доближават доста до някои от основните цели, залегнали в дисертационния труд и проблемите, на които той се опитва да отговори. Първото от тях засяга приложението на анализ на настроението с цел измерване на удовлетвореността на клиента от обслужването в контактен център** - [16]**. Второто изследване се фокусира върху откриването на най-често задавани въпроси/теми в онлайн чат комуникация, проведена в контактен център на банкова институция** - [8]. Фокусът в [8] е върху извличането на ключови фрази и запитвания, като са използвани разнообразни езикови ресурси за руски език. За да може обаче да се приложи подхода върху данни на български език (или друг), е необходимо да се разполага с подобни ресурси и за

него. Подробният анализ на постигнатото в тези статии спомага във формирането на методологията и някои от хипотезите, залегнали в настоящото изследване.

2.1.2. Анализ на текстови данни на български език

Една от основните цели на дисертационния труд е да представи актуална картина на развитието в сферата на анализа на текстови данни в България. Степента, до която прилагането на разнообразни техники за анализ на текстови данни е възможно и достъпно, варира за различните езици. Наличието на разнообразни езикови ресурси определя съществуващото разграничение между т.нар. „езици с малко ресурси“ (*low-resource languages*) и „езици с много ресурси“ (*high-resource languages*). От последните, не е изненадващо, че на първо място е английският език, за който съществуват голям брой корпуси, инструменти за обработка, речници, специализирани софтуерни системи за анализ на текст и т.н. Българският език се характеризира като език с малко ресурси, тъй като малък брой лингвистични ресурси са на разположение за целите на обработката и анализа му. След прегледа на актуални изследвания, посветени на анализа и извличането на знания от онлайн чат комуникация, е осъществен **преглед, анализ и синтез на постигнатото в ключови изследвания в сферата на анализа и обработката на текстови данни на български език в две основни направления**.

Първо е направен преглед на проучванията, фокусирани върху създаването на разнообразни **езикови ресурси за български език**, разработени с цел улесняване на обработката и анализа на текстови данни на този език. Примери за такива инструменти, често използвани в практиката, са различните видове езикови анализатори (*parsers*), инструменти за маркиране на части на речта (*part-of-speech tagging*) и разпознаване на именувани обекти в текст (*named entity recognition*), инструменти за стеминг (*stemming*) и други. Анализирани са приложимостта на съществуващите подобни инструменти, дали са достъпни чрез уеб интерфейс, софтуерна програма или са внедрени в установен софтуер за статистическо програмиране като R или Python. Специално внимание е обърнато на наличието на текстови корпуси на български език (и техния тип, размер, сфера на текстовите данни и ниво на анотация).

Второто основно направление на прегледа, анализа и синтеза, са **практическите приложения на анализа на текстови данни на български език за решаване на различни**

икономически или бизнес проблеми. Очертани са ключови изследователски статии главно в три практически области, които разбира се, имат много пресечни точки – клъстеризация на текстови документи, класификация на текстови документи и извличане на информация от текст. Най-голямо внимание е обърнато главно на: бизнес/икономическите проблеми, които са разгледани; сфера на текстовите данни; използвани аналитични методи; наличие на предоставени езикови ресурси като набори от данни, модели или програмен код за осъществяване на експериментите в изследването.

Изведени са заключения относно степента на развитие на научното поле, наличието и приложимостта на езикови ресурси за български език и степента, до която анализът на текстови данни на български бива използван за разрешаване на практически бизнес и икономически проблеми. Обхватът на проучването не се ограничава до преглед само на най-актуалните изследвания (например тези, публикувани през последните пет години) - вместо това на фокус са всякакви ключови изследвания в областта. Осъщественият преглед е подробен, без да се твърди, че е напълно изчерпателен. Важно да се подчертае е, че доколкото е известно на автора, това е **първият опит да се очертаят ключови изследвания в областта на анализа на текстови данни на български език** в двете направления, описани по-горе. В рамките на изложението по-долу са споменати само **част от най-важните заключения**, направени вследствие на този широк литературен преглед.

Глобалното развитие в сферата на анализа на текстови данни и по-конкретно - обработката на естествен език, преминава през три основни етапа. Първоначално биват използвани предимно подходи, базирани на правила. Такива методи, въпреки че са много по-лесни за тълкуване, отнемат много време, не обобщават добре данните и могат да бъдат лесно повлияни от човешка грешка. След този период започва прилагането на **статистически подходи към задачите за обработка на естествен език**. Методите за машинно самообучение са **много по-надеждни от тези, базирани на правила**. С помощта на статистически методи могат да бъдат извлечени и анализирани ценни зависимости в текстовите данни. Но за да бъдат приложени подобни методи, има едно важно условие, и това е наличието на анотирани данни. **Третият етап в развитието на анализа на текстови данни и обработката на естествен език настъпва едва през последните години с фокус върху трансферното обучение и използването на т.нар. модели тип „Трансформатор“** [20].

Съвсем естествено напредъкът в сферата на анализа на текстови данни на български език следва същите три етапа, макар и това развитие да се е забавило до някаква степен във времето. Прилагането на статистически подходи в експериментите с данни на български език започва относително по-късно в сравнение с глобалното развитие в сферата – това може да се обясни с липсата на анотирани данни към тогавашният момент от времето. Най-често прилагани са методики, базирани на класически подходи за машинно самообучение (различни от методите за дълбоко обучение). Това определя и една от **посоките за бъдещо развитие в сферата, а именно – фокус върху приложението на техники за дълбоко обучение**, както при разработването на инструменти за обработка на текст, така и в казуси с практически фокус. Трябва обаче да се отбележи, че това зависи до голяма степен от количеството налични данни - традиционните методи, базирани на невронни мрежи, изискват голям обем от данни на разположение, с цел постигане на висока ефективност.

В тази връзка **появата на трансферното обучение може да бъде добро алтернативно решение**. Трансферното обучение позволява използването на обучени върху огромни количества от данни модели, които в последствие могат да бъдат допълнително приспособени и прецизирани за конкретни задачи и езици. Такова прецизиране може да се извърши със значително по-малко количество данни на желания език. Без съмнение, **друга насока за бъдещо развитие е изучаването на трансферното обучение и как то може да се използва за разрешаване на проблеми в сферата на анализа на текстови данни на български език**. Проучвания, които вече са фокусирани върху тази изключително актуална тема са: [21], [22], [23], [24].

Прегледът разкрива, че **към този момент съществуващите текстови корпуси за български език са малко на брой**. Сред тях основните са - два едноезични корпуса, анотирани на различни лингвистични нива (BulTreeBank и BulNC), две лексикални бази от данни (BulNet, BTB-WordNet) и няколко двуезични/многоезични паралелни корпуса. Корпусът BulTreeBank се характеризира с висококачествена анотация на разнообразни лингвистични нива и много ключови изследователски статии го използват в създаването на разнообразни инструменти за обработка на текст. **Съществуващите инструменти и системи** за обработка на текстови данни на български език **обхващат редица задачи в тази област**, включвайки в това число: **токънизация, стеминг, лематизация (*lemmatization*), маркиране на части на речта, откриване на зависимости в текст, разпознаване на**

именувани обекти в текст, идентифициране на смисъла на думите в изречението (*word sense disambiguation*), анализ на съставните части на текста и други. Не всички от тези ресурси обаче са свободно достъпни за използване и внедряване в практически казуси/ системи в индустрията. Разработката на Popov, Osenova и Simov [25] поставя началото на изследователските усилия, насочени към интегрирането на инструменти за обработка на български език в софтуери с отворен код като Python.

Понастоящем наличните езикови ресурси за български език „живеят“ самостоятелно на различни платформи/специализирани софтуерни програми. **Бъдещите разработки в сферата зависят от, и трябва да включват, интеграция със софтуери като Python, R или технологии за обработка на естествен език като spaCy, NLTK, TextBlob** и други. Интегрирането на езикови ресурси в подобни софтуери би спомогнало да се оцени тяхната приложимост в индустрията. Съвсем естествено, ако такива интеграции съществуват, това ще улесни практиците в сферата и ще доведе до повече експерименти и изследвания с практически фокус.

Прегледът на литературата разкрива, че **основните изследователски усилия в областта на анализа на текстови данни на български език започват в началото на XXI-ви век.** Няма голям обем от изследователски статии, фокусирани конкретно върху приложния анализ на текстови данни с цел разрешаване на икономически/бизнес проблеми, а повече от половината от разгледаните статии в тази област са публикувани след 2016 г. Трябва да се отбележи обаче, че има не малко такива изследвания и напредък, конкретно в медицинската сфера. Има напредък и изявен научен интерес и в откриването на токсично и подвеждащо поведение в дискусионни форуми, както и в откриването на фалшиви новини.

Почти половината от изследванията, фокусирани върху практически бизнес/ икономически проблеми, използват като данни новинарски статии. Трябва да се направи повече в посока анализ на човешкото поведение и комуникация в социалните мрежи (например чрез данни от платформи като Twitter, Facebook, BGMamma и т.н.). Повече изследователски капацитет би могъл да бъде насочен и към анализа на настроенията, който има множество приложения в анализа на политически и социални събития [26], [27], като се характеризира с потенциал да бъде изключително ценен във финансовата [28], [29], [30], счетоводната [31] и икономическата сфера [32].

Различни социални, политически или бизнес проблеми биват адресирани с помощта на анализ на текстови данни на български език: оптимизация на уебсайтове; откриване на „манипулационни тролове“, заблуждаващи общественото мнение в интернет; анализ на настроенията на потребителски отзиви; откриване на фалшиви новини; анализ на поведението на клиенти и други. Една бъдеща посока на изследванията в сферата се състои в прилагането на по-сложни и изискващи повече ресурси задачи, като: обобщение на текстови данни, четене с разбиране (*reading comprehension*), извличане на зависимости (*relation extraction*), изследване на семантичната близост в текста (*semantic textual similarity*), създаване на системи за въпроси и отговори и диалогови системи с изкуствен интелект (*conversational artificial intelligence*).

Важно откритие, направено вследствие на широкия литературен преглед, анализ и синтез е, че към момента съществува само една статия, посветена на приложението на анализ на настроението върху данни на български език [33] **и нито една, в която се прилагат методи за моделиране на теми.** Тези две аналитични задачи са основни в настоящото изследване. В допълнение, **доколкото е известно на автора, до този момент не са публикувани научни статии, занимаващи се с анализ и извличане на знания от онлайн чат комуникация между клиенти и служители (на български език).** Настоящото изследване е първото за тази комбинация от език на изследваните данни (български), сфера на данните (онлайн чат комуникация с цел обслужване на клиенти) и аналитична задача (анализ на настроението/моделиране на теми в текст). Изследвайки въпроса, кои са възможните и най-добри подходи, които могат да се приложат, както и потенциалните проблеми в процеса, **дисертационният труд допринася към литературата, посветена на практическите приложения на анализа на текстови данни на български език.**

Прегледът на литературата разкрива липсата и съответно необходимостта от разработването на модули за автоматизиране и улеснение на обработката и анализа на текстовите данни, обект на дисертационния труд. В тази връзка, като **един от приносите на дисертационния труд, може да се посочи и възможността, конкретни процедури от създадената цялостна методология за обработка и анализ на данните (в Глава II) да бъдат репликирани от други изследователи в сферата и използвани с цел автоматизиране и улеснение на анализа на онлайн чат комуникация на български език.**

2.1.3. Моделиране на теми в текстови данни и анализ на мненията и настроенията

Осъществен е и литературен преглед на основните подходи, актуални тенденции и наложени техники за моделиране на теми в текстови данни и анализ на настроението. **Моделирането на теми** в текстови данни е много популярен метод за машинно самообучение без учител (*unsupervised learning*), който често се използва за изследване на структурата на корпуси от текстови документи, за които никаква информация не е известна предварително. Концепцията за този вид анализ е известна още от 1990 г. [34] и има приложения в множество области [35] - анализ на исторически документи, научни публикации, литературни произведения, анализ на програмен код, класификация на изображения, извличане на мнения и други. В [36] е предоставена следната дефиниция: *„Моделирането на теми е общия термин зад широка група от алгоритми, използвани за разкриване и анотиране на тематичната структура в колекция от документи“.*

Какво представляват „темите“ в контекста на този вид анализ, разяснява следващата дефиниция [34]: *„При тематичното моделиране думата „тема“ придобива специфичното значение на разпределение на вероятностите върху думите в даден текстови корпус, като същевременно се използва и в контекста на по-общото си значение, а именно - предмет на дискусия“.* Обяснено по най-опростен начин, моделирането на теми представлява групиране на думи в теми въз основа на появата им заедно в корпуса от текст, който бива анализиран. Тоест, това са колекции от думи, които могат да бъдат свързани по смисъл [34].

В контекста на единия от предметите на дисертационния труд, методите за моделиране на теми предоставят възможност за извличане на интересни изводи и забелязване на тенденции в поведението на клиентите чрез анализ на основните теми, представляващи интерес за тях, както и на възникналите проблеми с предоставяните стоки и услуги. Сред основните техники за моделиране на теми намират място LSA (*Latent semantic analysis*), NMF (*Non-negative matrix factorization*), pLSA (*probabilistic Latent Semantic Analysis*), LDA (най-широко разпространен в литературата), множество модификации на LDA (*Correlated Topic Model - CTM, Pachinko Allocation Model - PAM, Author Topic Model - ATM*) и други.

По отношение на другия предмет на дисертационния труд – в литературата анализът на удовлетвореността на клиента от обслужването в контактен център е най-често адресиран с помощта на методи за качествен анализ (*qualitative analysis*) [37], [38]. От друга страна,

количествените методи позволяват автоматизация, по-бърз анализ, и имат по-всеобхватно покритие. Последните могат да се считат като важно допълнение към качествените подходи, които пък се характеризират с по-голям детайл и точност. В дисертационния труд е поставен фокус именно върху приложението на количествени методи за измерване на удовлетвореността на клиента. **Разпознаването на това, дали клиентът оценява онлайн чат комуникацията си с контактен център като цяло положително или отрицателно, може да бъде разглеждано като задача за анализ на настроението на клиента**.

Формална дефиниция на т.нар. „анализ на настроението“ [39], [40] - *„Анализът на настроението е научна област, занимаваща се с анализ на мнението, настроението и субективизма, изразени в текст, с помощта на изчислителната мощ на компютърните технологии. Това е обширна проблемна област, като на анализ подлежат мненията, чувствата, оценките, нагласите и емоциите на хората към различни обекти – продукти, услуги, организации, лица, спорни въпроси, събития, теми и други“.*

В контекста на анализа на настроения и мнения най-често бива разглеждана задача за определяне на полярността на даден текст. Какво всъщност представлява тази “полярност” на изразеното настроение в текстови данни, Pang и Lee дефинират по следния начин [39] – *„Задачата за определяне на това, дали даден текст (изразяващ мнение) носи положителен или отрицателен емоционален заряд, се нарича задача за определяне на полярността на текста“.* **Предмет на дисертационния труд е удовлетвореността на клиента, изразена именно в нейната полярност.** Под “полярност” се има предвид конкретно, дали клиентът е оценил положително или отрицателно онлайн чат комуникацията си с контактния център. **Дисертационният труд тества емпирично дали е изпълнено допускането, че текстовите данни под формата на чатове съдържат достатъчно сигнали, чрез които да се прогнозира мнението (настроението) на клиента за предоставената услуга в края на комуникацията.**

В литературата повечето приложения на анализа на настроението са реализирани основно върху отзиви за продукти/услуги или филми, както и върху данни от социални мрежи (например, Twitter). В дисертационния труд анализът на настроението бива приложен в сфера, която е много по-малко застъпена – онлайн чат комуникация, генерирана в контактен център (сферата на обслужването на клиенти). Такива данни поставят доста предизвикателства и въпроси, свързани с обработката им, нивото на анализ, степента до

която съдържат изразени мнения/настроения, наличието на „излишна“ информация/шум (например, поздрави в началото на разговора, процес на идентификация на клиента и подобни особености в процеса на обслужване в контактен център), които биха затруднили анализа и други. **Дисертационният труд допринася към съществуваща литература, правейки изводи относно тези особености, и как те ще се отразят върху разрешаването на задачата за разпознаване на настроението в текст.**

Три основни аналитични подхода могат да бъдат използвани в задачата за анализ на настроението [41]. Първият от тях се базира на приложението на лексикони на настроенията (SentiWordNet [42], AFINN [43], VADER [44], SocialSent [45] и други) – подобни подходи не изискват наличието на тренировъчни данни. Лексиконите съдържат списъци от думи и фрази, които често биват ползвани с цел предаване на положителни или отрицателни настроения. Съществува един лексикон на настроенията [33], предназначен за данни на български език, който е свободно достъпен за изследователската общност. **Важно е да се отбележи, че този лексикон е разработен чрез данни в съществено различна сфера от тази на данните, обект на дисертационния труд**.

Вторият подход за анализ на настроенията е базиран на методи от сферата на машинното самообучение, докато третият подход е хибрид - комбинация от приложение на лексикони и машинно самообучение. Сред най-популярните класически алгоритми за класификация на текстови данни са методите на опорните вектори (*SVM - support vector machines/SVC – support vector classifier*), Наивния Бейсов Модел и логистичната регресия [46]. Постепенно навлиза и използването на модели тип „Трансформатор“ [20], комбинирани с трансферно обучение [47].

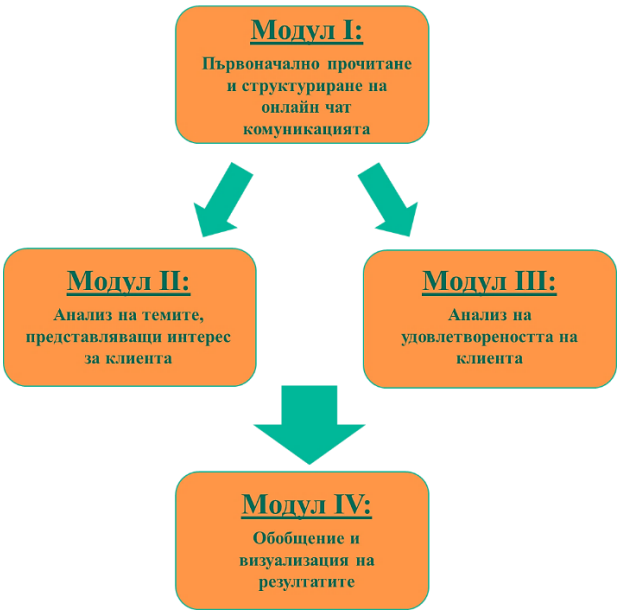
2.2. Методология (Глава II)

Стъпките в процеса по създаване на цялостната методология, с оглед постигане на основната цел на дисертационния труд, са документирани в [48], [49], [50]. Във връзка с изпълнението на основната цел са дефинирани следните четири подцели:

1. Първата междинна цел е създаването на модул, който **обработва данните и ги структурира**, така че да бъдат в подходящ за последващ анализ вид (Модул I). Целта е данните да бъдат приведени във форма, която улеснява прилагането на техники за количествен анализ.

2. Втората междинна цел е създаването на модул, който осъществява **локализиране и анализ на основните теми**, засегнати от клиентите в онлайн чат комуникацията им с контактен център (Модул II).
3. Третата междинна цел е създаването на модул, който **анализира удовлетвореността (нагласата/настроението) на клиента** от онлайн чат комуникацията му с оператор в контактен център (Модул III).
4. Четвъртата подцел акцентира върху **бизнес приложенията на създадената система** и се състои в представяне и обобщение на информацията, придобита в Модул II и III, както и на други интересни и важни индикатори, които могат да бъдат извлечени от изследвания тип данни (Модул IV).

Всяка от тези подцели засяга създаването на конкретен модул от системата за анализ на онлайн чат комуникация, който отговаря за определена аналитична задача (Фигура 1):



Фигура 1. Схема на основните модули в системата за анализ на онлайн комуникация с клиенти

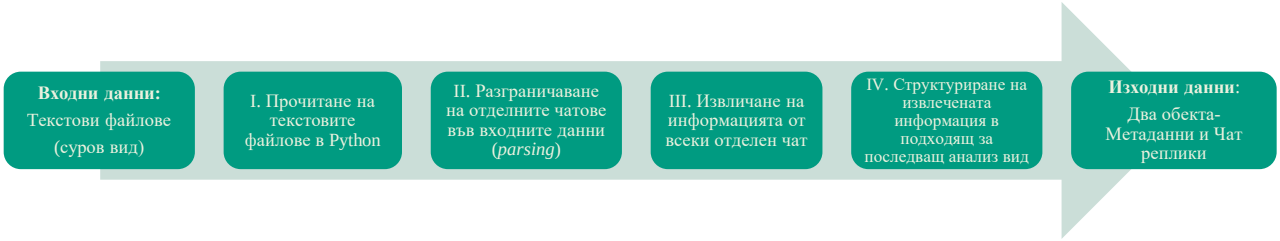
Модулите, в тяхната съвкупност, изграждат автоматизираната система за анализ на комуникацията с клиента. Тази система е създадена и **може да бъде имплементирана в реална среда чрез програмния език Python**, който бива използван във всички аналитични експерименти в рамките на дисертационния труд[7]. Сред основните библиотеки, чрез които

[7] Използвани са две интегрирани среди за разработка – Spyder и Jupyter Notebook.

са осъществени експериментите са: Gensim, scikit-learn, NumPy, pandas, Plotly, wordcloud, NLTK, treetaggerwrapper, BulStem.

2.2.1. Модул I

В Модул I бива приложен алгоритъм за първоначално прочитане и структуриране на данните (Фигура 2), тъй като техният първоначален вид (Фигура 3) не позволява приложението на каквито и да било аналитични техники. В рамките на този модул бива приложен и алгоритъм за предварителна обработка и нормализация на данните (Фигура 4). Последното е необходимо поради факта, че данните се характеризират с много особености, които ако не бъдат взети предвид, биха понижили качеството на всички последващи анализи. Изходният продукт от този модул е технически коректни и структурирани данни, които биват използвани като входни данни във всички останали модули на системата.



Фигура 2. Диаграма на алгоритъма за структуриране на информацията, извлечена от онлайн чат комуникация

2020-01-22.text - Notepad

File Edit Format View Help

```
Timestamp: 2020-01-22T15:56:19Z
Unread:false
Visitor ID: 1111111.aaaa1aaaaaa11a
Visitor Name: Visitor 11111111
Visitor Email:
Visitor Notes:
IP: 111.11.111.1
Country Code: BG
Country Name: Bulgaria
Region:
City:
User Agent: Mozilla/1.1 (Windows NT 1.1) AppleWebKit/111.11 (KHTML, like Gecko) Chrome/11.1.1111.11 Safari/111.11
Platform: Windows
Browser: Chrome

(2020-01-22 15:56:19) Visitor 11111111: Здравейте! Трябва ми помощ - забравил съм потребителското си име и паролата за онлайн банкиране.
(2020-01-22 15:56:34) Иван Иванов: Здравейте!
(2020-01-22 15:57:08) Иван Иванов: Необходимо е да посетите офис на банката, за да получите нови потребителско име и парола.
(2020-01-22 15:57:08) Visitor 11111111: Много ви благодаря!
(2020-01-22 15:57:08) Visitor 11111111: Хубава вечер!
(2020-01-22 15:57:08) Иван Иванов: Моля, хубава вечер и на вас!
================================================================================
```

Фигура 3. Структура на примерен чат в суров вид.

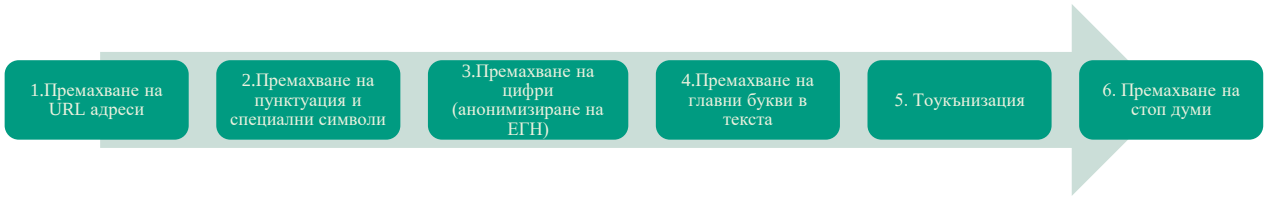
Фигура 4. Предварителна обработка на данните преди да бъдат ползвани като входни данни в Модул II/ Модул III

Приложените техники в Модул I са съобразени и със спецификите на данните, ползвани с цел апробиране на практика на създадената система за анализ на онлайн комуникация с клиента. Данните са генерирани в контактния център на голяма финансова институция в България. В случай, че е необходимо системата да бъде ползвана с цел анализ на данни, генерирани в различна чат платформа, е възможно да се наложат някои модификации на техниките, ползвани в Модул I, така че да се осигури правилно прочитане на данните и структуриране на наличната информация в тях.

2.2.2. Модул II

В Модул II бива приложена серия от аналитични техники с крайна цел – извличане и анализ на основните теми, които вълнуват клиентите и които те са засегнали в чат комуникацията си с контактен център. Чатовете са съществено различни от комуникацията с клиенти в други канали, не само по отношение на структурата и лингвистичните им особености, но и като източник на информация. Предизвикателство пред моделирането на теми е наличието на поздрави между страните в чата, шаблонните съобщения на операторите в процеса на идентификация и обслужване на клиента и други – тези части от диалога не добавят допълнителни знания, а по-скоро представляват излишна информация.

В рамките на този модул са приложени допълнителни техники за обработка на текста, които са съобразени с конкретната аналитична задача (Фигура 5). **Предложената цялостна методика в Модул II е авторско съчетание от разнообразни подходи и техники с цел постигане на поставените изследователски задачи**. Важно е да се отбележи, че някои от аналитичните техники и методи за интерпретация на резултатите, приложени с цел моделиране на теми, са усъвършенствани/допълнени, в сравнение с тези в първоначалните експерименти, описани в две от публикациите по дисертационния труд - [48], [49].



Фигура 5. Обработка на текста в Модул II

Алгоритъмът LDA е използван за целите на моделиране и анализ на основните теми, които вълнуват клиентите. Изборът не е изненадващ и се обосновава на огромния обем от съвременни статии, които прилагат този метод в разнообразни експерименти. Доколкото е известно на автора, според направения литературен преглед, **представянето и възможностите, които този алгоритъм предоставя, до момента не са били изследвани в сферата на данните, обект на анализ в настоящото изследване** (обслужване на клиенти във финансовата сфера чрез чат комуникация в контактен център). **Дисертационният труд цели да запълни тази празнина в литературата и да направи изводи относно представянето на LDA върху подобен тип данни, ограниченията и предимствата на подхода, както и спецификите на данните, които биха повлияли върху резултатите от приложението му. В тази връзка възниква и Хипотеза 1 (виж точка 1.4).**

Чат комуникацията в контактен център е смесица от реплики на оператор и клиент. Обикновено бива инициирана от клиента, който задава въпрос или излага своя проблем, търсейки помощ. Следват разясняващи въпроси, удовлетворяване на заявката и край на разговора. **В този контекст възниква въпросът: „Помагат ли изказванията на операторите в задачата за извличане на основните теми, представляващи интерес (или проблем) за клиента или по-скоро я възпрепятстват, добавяйки излишна информация в анализа?“**. Взимайки предвид и мнението на експерти в контактния център на финансовата институция, върху чиито данни е приложена на практика създадената система за анализ на онлайн чат комуникация, е формирана Хипотеза 2 (виж точка 1.4). За да се провери тази хипотеза са създадени три нива на репрезентация на данните:

1. Извадка от **цели чатове** между клиент и оператор („Извадка I“).
2. Извадка, включваща всички реплики **само на клиента** (“Извадка II”).
3. Извадка, включваща **само първите реплики** на клиента в комуникацията (“Извадка III”). Идеята е да се обхванат само репликите на клиента, в които той описва своя проблем и изпраща заявката си.

Макар и да има научни статии с подобен замисъл в други области (например, сравнение на резултатите и ефективността на анализа на темите, приложен върху пълен и резюмиран текст [51]), доколкото е известно на автора, няма други изследвания, анализиращи влиянието на нивото на репрезентация на онлайн чат комуникацията върху поведението на модели за извличане на теми като LDA или подобни на него.

Чрез LDA всеки документ в даден корпус бива представен като случайна комбинация от скрити (латентни) теми, а всяка тема бива представена като различно вероятностно разпределение на всички думи в корпуса [52]. В дисертационния труд, LDA алгоритъмът е приложен върху набора от отделни думи (т.нар. „униграми“ /*unigrams*/), съставляващи текста, **за всяко едно ниво на репрезентация на данните**. Съвсем логично, този набор от думи (наричан още „речник“) има пряко отношение към резултатите и качеството на извлечените теми. В тази връзка са извършени различни **филтрации на речника от думи** с цел постигане на оптимални резултати. Те се изразяват в промяна на параметър, който контролира включването на думи, които присъстват много често или много рядко в чатовете. Експериментира се и със създаден допълнителен списък от стоп думи (*stop words*), специфични за конкретния казус и данни, както и с филтриране на думи с помощта на инструмент за маркиране на части на речта.

Макар че моделирането на теми в дисертационния труд представлява задача в сферата на машинното самообучение без учител, в методиката са включени подходящи метрики за оценка на представянето на приложения алгоритъм, проверка на хипотезите и избор на оптимален брой теми. Това са т.нар. **коефициент на сложност** (*perplexity measure*) [52] и $\boldsymbol{C_v}$ **кохерентност на модела** (*coherency measure*) [53]. Анализирани са в детайл предимствата и ограниченията на всяка от тези две метрики. В процеса на оценка на резултатите имат влияние както **интерпретацията на автора**, така и **експертното мнение на служители** в контактния център на финансовата институция, върху чиито данни е приложена на практика създадената система за анализ на онлайн чат комуникация.

2.2.3. Модул III

В Модул III на системата, отразявайки особеностите и предизвикателствата, поставени от обекта на дисертационния труд, бива създадена **авторска методика за анализ на**

удовлетвореността на клиента от комуникацията му с контактен център посредством чат. В данните, използвани с цел апробиране на създадената система, много малка част от клиентите са поставили рейтинг на предоставената услуга. Това наблюдение е направено и в други изследвания, анализиращи онлайн чат комуникация в контактен център, и е посочено като особеност на този тип данни [16]. Именно това обуславя и необходимостта от аналитичен инструмент, който да анализира удовлетвореността на клиентите. **От най-голям интерес за бизнеса е улавянето на комуникации, от които клиентът е останал разочарован**. Възможно е обаче да съществуват чатове, в които привидно клиентът не е изразил никакво мнение/настроение и накрая все пак е поставил лоша оценка.

Вследствие на изводите от подробния литературен преглед **и взимайки предвид описаните особености на данните, в дисертационния труд са използвани методи за машинно самообучение с цел прогнозиране на удовлетвореността на клиента.** Очакването е по този начин да бъде построен модел, улавящ спецификите на изказа, използван в избраната сфера на данните – направено е допускането, че статистическият модел, учейки се на база на историческите данни, ще обхване и някои на пръв поглед „невидими“ сигнали в комуникацията, които водят до поставянето на лош рейтинг. Това предполага и постигане на по-голяма точност в сравнение с тази, която може да се постигне чрез лексикони на настроението [46]. Както беше споменато по-рано в точка 1.4, **изводите, публикувани в** [16] относно приложимостта на подобни лексикони върху данни в сферата на обслужването на клиенти, допълнително обосновават този избор.

В рамките на настоящото изследване е тествано емпирично едно много важно допускане, а именно, че **текстовият диалог (чат) между клиент и оператор, сам по себе си, съдържа достатъчно сигнали**, така че чрез него да може да се създаде модел, прогнозиращ удовлетвореността на клиента от комуникацията. Дали това допускане е реалистично или не, представлява един от изследователските въпроси в дисертационния труд, вследствие на който възниква **Хипотеза 3** (виж точка 1.4). Ако текстовият диалог наистина съдържа подобни сигнали, то тогава обучението на статистически модел чрез тези данни ще доведе до оптимални резултати, спрямо тези, получени чрез приложение на наивна прогноза (например, случайно налучкване */random guess model/*). **Ползването на исторически данни и алгоритми с цел „учене“ от тях, следва да доведе до по-добри резултати, само ако данните наистина съдържат ценни сигнали за удовлетвореността на клиента.**

Допълнително в задачата за прогнозиране на удовлетвореността възниква въпросът: **„Необходимо ли е да се използва целият текстови диалог или само част от него ще бъде достатъчна, за да се прогнозира окончателния рейтинг на клиента за услугата?“**. Анализът на изводите, направени в [16], води до възникването на Хипотеза 4 (виж точка 1.4). **В изследването става ясно, че настроенията, изразени в края на комуникацията, са с най-голяма предсказваща способност в моделирането на крайния рейтинг на клиента**. В тази връзка, анализът е приложен върху **три нива на репрезентация** на чатовете:

1. Извадка от **цели чатове** между клиент и оператор („Извадка 1“).
2. Извадка, включваща всички реплики **само на клиента** (“Извадка 2”).
3. Извадка, включваща **само финалните реплики** на клиента в комуникацията (“Извадка 3”). На това ниво е направено допускане, че именно в тези реплики клиентът или изразява благодарност, ако е останал доволен от услугата, или изразява обратното, ако услугата не е задоволителна и заявката му не е изпълнена.

Във връзка с всичко казано дотук, следва кратко описание на някои от основните стъпки в методиката в Модул III. Първо, на база на наличните данни е **формирана целева променлива** (*target variable*) – като тренировъчни данни са използвани чатове, в които **е известен** финалният рейтинг на клиента (около **13,6%** от всички налични данни). Проблемът, разглеждан в Модул III, може да бъде формално дефиниран по следния начин:

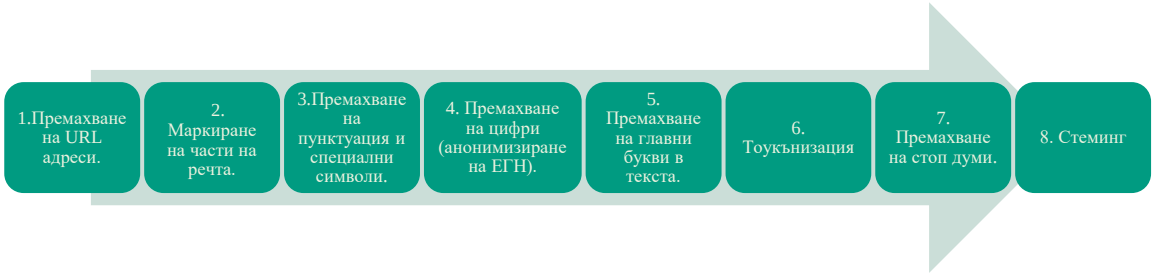
За даден чат C, задачата е той да бъде класифициран в една от следните две категории клиентски рейтинг на услугата: *Добър* (0) или *Лош* (1).

Осъществени са обработка на текста и превръщане в числов вид, съобразени както със спецификите на данните, така и с конкретно разрешавания проблем в Модул III (Фигура 6). Създадена е аналитична процедура, чрез която да се изследва ефекта от следните параметри/техники, като крайната цел е постигане на работещо решение и оптимално представяне на създадения модел за прогнозиране на удовлетвореността:

1. **Ниво на репрезентация** на онлайн чат комуникацията („Извадка 1“/“Извадка 2“/“Извадка 3“).
2. **Техники за обработка на текста** – тестване на приложението на различни форми на стеминг [54] върху данните и анализ на техния ефект върху резултатите („Стеминг

1“/„Стеминг 2“/„Стеминг 3“/„оригинална форма“). Детайли относно разликите между различните форми на стеминг са налични в основния текст на дисертационния труд.

3. Експериментиране с различни **набори от обясняващи променливи,** извлечени директно от текстовите диалози. Чатовете, сами по себе си, формират множество от потенциални **„текстови“ обясняващи променливи**. Допълнително, освен думите в чатовете, в прогнозирането биват използвани и определени граматически признаци на текста, извлечени с помощта на инструмент за маркиране на части на речта – т.нар. **„морфологични“ обясняващи променливи**. **Последните биват използвани по два начина – в комбинация с „текстовите“ обясняващи променливи или като метод за техния подбор** (селектирани са конкретни думи, които изпълняват ролята на определена част на речта). **Подборът включва следните части на речта – съществителни имена, глаголи, прилагателни имена, междуметия и наречия.** Маркирането на части на речта е приложено чрез инструмента TreeTagger [55].
4. **Алгоритъм** за машинно самообучение - изследвано е приложението на три алгоритъма - класификатор с опорни вектори (*support vector classifier - SVC*), Бернулиев Наивен Бейсов модел (Bernoulli Naïve Bayes model - BNB) и логистична регресия (избор, базиран на подробния преглед на изследвания, анализиращи онлайн чат комуникация).



Фигура 6. Обработка на текста в Модул III

С цел валидация на резултатите е приложена техниката ***k*-кратна крос валидация** (*k-fold cross validation*) [56]. За проверка на хипотезите и оценка на предсказващата способност на създадените модели са използвани няколко наложени метрики - **точност, прецизност, чувствителност, F1, F-beta**. На база на различните комбинации от изброените параметри и с помощта на метриките за оценка бива открит и модела с оптимално представяне в така формулираната задача.

Извън обхвата на емпиричните експерименти попада приложението на методи за дълбоко обучение, трансферно обучение и модели тип „Трансформатор“, които имат съществени разлики с избрания класически подход. **Целта е да се изследва какво може да се постигне в задачата за анализ на настроението в онлайн чат комуникация чрез приложението на наложени традиционни подходи от сферата на машинното самообучение** и да се направи съпоставка с постигнатото от други автори, които разчитат на ползването на лексикони на настроението в подобни задачи [16].

Значителна част от подходите, базирани на изкуствени невронни мрежи, изискват наличието на по-голям обем от данни с цел постигане на ефективни резултати. В разглеждания бизнес казус в дисертационния труд, към момента това условие не е изпълнено и натрупаният обем от данни не е в значителен размер. От друга страна, трансферното обучение адресира именно подобни случаи, като идеята е да се използват придобитите знания от подобни данни и задачи в други сфери. Към този момент, изследванията, прилагащи методи за дълбоко обучение или трансферно обучение в анализа на текстови данни на български език са много малко на брой - сферата тепърва предстои да се развива. Това определя и една от **бъдещите посоки** на настоящото изследване, която се изразява в тестване на подходи, базирани именно на методи от сферата на трансферното обучение и дълбокото обучение, както с цел репрезентация на текстовите данни, така и с цел създаването на прогнозиращ модел на удовлетвореността на клиента.

2.2.4. Модул IV

Целта на Модул IV е да се постави акцент върху интерпретацията на резултатите от приложените аналитични техники в останалите модули. Идеята е да се обърне внимание на въпроса, как могат получените резултати да са полезни в практиката и да се извлече смисъла им за бизнеса. **Модул IV улеснява тълкуването на резултатите от анализа на онлайн чат комуникация в контактен център и извличането на знания от такъв тип данни, насочвайки вниманието към бизнес приложенията на създадената система**.

Целта е да се илюстрира как резултатите биха могли да бъдат ефективно използвани и как могат да бъдат представени, така че да достигнат по-лесно до хората от страна на бизнеса, които са и най-заинтересовани от тях. В Модул IV са използвани прости

аналитични техники в комбинация с техники за визуализация (например, облаци от думи /*word clouds*/ или графики, които проследяват промените в дадена характеристика с течение на времето). Обединени са резултатите от Модул II и Модул III, като по този начин се постига синергичен ефект в анализа и интерпретацията им. Последното позволява да се открият зависимости между нивото на удовлетвореност на клиента от обслужването и засегнатите теми в комуникацията му с контактен център.

По отношение на конкретната банкова финансова институция, чиито данни са използвани за апробиране на създадената система, заинтересовани от последната биха били поне няколко отдела - например, отдел „Контактен център“, отдел „Качество на обслужването“, отдел „Управление на взаимоотношенията с клиенти“. Съвкупността от аналитични техники и визуализации в Модул IV илюстрира как би изглеждала една подобна система за нейните крайни потребители от страна на бизнеса. Докато Модул I, Модул II и Модул III представляват т.нар. „back-end“ на системата, Модул IV е нейният „front-end“ - тоест това, с което крайният потребител взаимодейства директно и чието предназначение е да му осигури възможност сам да намери отговор на въпросите:

- Какви знания биват извлечени от данните?
- С какво тези знания са ценни за бизнеса?
- Какви бизнес решения могат да бъдат взети (или да бъдат подпомогнати) на базата на извлечените от данните знания?

2.3. Емпирично изследване (Глава III)

В рамките на Глава III от дисертационния труд, системата за анализ на онлайн чат комуникация с клиенти е **апробирана в рамките на конкретен реален бизнес казус**. Анализирана е онлайн чат комуникацията между клиенти и оператори в контактния център на голяма финансова институция, оперираща в България. Голяма част от основните резултати и изводи, направени вследствие на емпиричните експерименти са документирани в [48], [49], [50]. **Както беше споменато и по-рано, някои от аналитичните техники са усъвършенствани/допълнени, в сравнение с тези в първоначалните експерименти**. Извадката от данни, ползвана с цел апробиране на системата, се състои от **37 529** чата, проведени в периода от **22.01.2019 г.** до **01.04.2021 г**. Тъй като в анализа биват използвани

различни нива на представяне на тази извадка (т.нар. „репрезентации“ на онлайн чат комуникацията), за улеснение, в нейния пълен вид тя бива означена като „Извадка I“.

2.3.1. Анализ на темите (Модул II)

Първоначално, анализът на основните теми, които вълнуват клиентите, е осъществен върху Извадка I (тя съдържа всички реплики на клиент и оператор). Следващ експеримент цели да тества коя репрезентация на данните (виж точка 2.2.2) води до оптимални резултати от моделирането на теми. Анализът е осъществен чрез приложение на алгоритъма LDA в условията на машинно самообучение без учител - реалният брой теми в извадката не е известен предварително. С помощта на емпирични методи е създадена процедура, по която да се определи оптималният им брой – използвани са метриките C_v кохерентност и коефициент на сложност. Последният е насочен към това, да оцени как моделът ще се представи върху данни, които “не е виждал”, докато метриката за кохерентност оценява качеството на получените теми от гледна точка на разбираемост и смисъл. В този смисъл, кохерентността се приближава до т.нар. „човешко“ възприятие за качеството на модела. C_v кохерентността заема стойности между 0 и 1. По-високата стойност на C_v говори за по-смислени и разбираеми теми и съответно – модел с по-ефективно представяне. Резултатите в утвърдени статии в сферата сочат, че в практиката се е наложило следното - стойности под 0.4 говорят за ниско към лошо качество на създадения модел, докато стойности между 0.6 и 0.8 се считат за оптимални [51], [57], [58]. Стойности над 0.9 е малко вероятно да се получат и се разглеждат по-скоро като подозрителни[8].

Тънкост и много важен момент в моделирането на теми е изборът на речника от думи, който ще бъде използван. В експериментите са тествани **пет различни филтрации на речника от думи в Извадка I**. Тези филтрации представляват комбинация от промени в следните параметри – **премахване на най-често/най-рядко срещани думи, използване на допълнителен ръчно създаден списък от стоп думи и подбор на думи чрез маркиране на частите на речта**[9]. На Таблица 1 е представено сравнение на резултатите от приложението на LDA върху различните филтрации на Извадка I.

[8] Представените референтни стойности се отнасят за цялостната кохерентност на даден модел (т.е. средната стойност на сумата от индивидуалните кохерентности на всяка тема в модела).

[9] По-конкретни детайли относно разликите между различните филтрации са налични в основния текст на дисертационния труд.

На база на всички изводи от проведения обширен анализ на Извадка I може да бъде заключено, че модели, съдържащи от 15 до 20 теми, се характеризират с оптимално представяне, както по отношение на кохерентност на получените теми, така и по отношение на поведение на модела върху данни, които не са участвали в тренирането му (според коефициента на сложност). Важно е да се уточни, че за избор на оптимален брой теми според C_v е приложен подход, предложен в някои наложени изследвания в сферата [51], [59]. В анализа за всяка една филтрация е създадена графика, която проследява нивото на кохерентност на моделите за всеки брой теми (изследвани са модели с от 2 до 50 теми) и позволява лесно да бъде отличен моментът, в който кохерентността започва да варира около една и съща стойност.

Таблица 1. Извадка I – сравнение на резултатите от приложението на LDA

Вид Филтрация	*Брой думи в речника*	*Избран оптимален брой теми*	*Средна стойност на C_v за избрания брой теми*	*Средна стойност на C_v след избрания брой теми*
Филтрация 1	13 396	15	0.6127	0.5756
Филтрация 2	13 177	20	0.5844	0.5617
Филтрация 3	8 556	15	0.6233	0.5716
Филтрация 4	8 497	20	0.5308	0.5247
Филтрация 5	**4 592**	**16**	**0.6655**	**0.6163**

Най-голяма стойност на C_v е постигната след приложението на Филтрация 5 върху речника от думи, ползван в моделирането на Извадка I – селектирани са само съществителни имена с помощта на инструмент за маркиране на части на речта в текстове на български език [55]. И коефициентът на сложност, и метриката C_v показват много добро представяне на модел с 16 теми - това е финалният избор за оптимален брой теми в Извадка I. Изборът е в унисон и с бизнес очакванията по отношение на броя генерални теми, представляващи интерес за клиентите, които би трябвало да присъстват в данните.

Чрез приложението на алгоритъма LDA, всеки един чат може да бъде описан чрез разпределението на темите в него. **Тези думи, които се характеризират с най-голяма вероятност в дадена тема, определят това за какво най-общо е дадената тема – тоест те изразяват основната концепция на темата.** Работа на анализатора е да извлече смисъла, логиките между думите и да определи темата. В основния текст на дисертационния труд подробно са анализирани откритите теми, представляващи интерес за клиентите, тяхното

качество, ключови думи, чатове, в които доминират и ниво на асоциация и зависимости помежду им. Използвана е библиотека за специализирани интерактивни визуализации [60], която улеснява интерпретацията на темите, откриването на интересни зависимости между тях, групирането им в по-общи направления („клъстери“) и позволява анализ на това, доколко доминират в извадката от данни. Анализът на темите води до разграничение на четири основни направления в дискусиите с клиентите, а именно:

1. **Кредитиране и различни етапи в този процес** - изисквания при отдаване на кредит, процес на кредитиране, становища по кредити/рефинансиране и други.
2. **Дигитално банкиране и плащане в интернет** - регистрация в онлайн банкиране, сигнали за проблеми в онлайн банкирането, получаване на СМС с код при плащане в интернет и други.
3. **Касови операции** - теглене/внасяне на парични наличности в офис, въпроси относно обмяна на чужди валути (актуални курсове на валути), търсене на информация за локация/работно време на офиси и други.
4. **Картови продукти (кредитни и дебитни карти) –** откриване/закриване на сметки (условия), такси, постъпления по сметки, осъществяване на преводи, запори и други.

Не се забелязва силно доминиране на някоя от темите, но графичният анализ показва, че в данните малко повече се появяват заявки и въпроси относно процеса по кредитиране и таксите по картови продукти. След обширния анализ на Извадка I, следваща част от емпиричните експерименти цели да изследва коя репрезентация на данните води до оптимални резултати в процеса на извличане на темите, които вълнуват клиентите (виж Хипотеза 2, точка 1.4). В точка 2.2.2 са описани изследваните три репрезентации на данните (Извадка I, II и III). Извадките са създадени като се следват сходни процедури и биват приложени едни и същи техники за обработка на текста.

В анализа на Извадка I са тествани пет различни филтрации на речника от думи. Макар че Филтрация 5 се оказва оптимална за Извадка I, с цел пълнота на анализа, за Извадка II и Извадка III са анализирани ефектите и на част от другите филтрации (тези, които биха били приложени по-често в практиката). Стойностите на параметрите във всички филтрации са същите като тези, приложени в анализа на Извадка I. Резултатите отново са анализирани графично, а в Таблица 2 е осъществен обобщаващ сравнителен анализ между полученото чрез Извадка II и Извадка III и най-добрия модел, създаден върху Извадка I.

Таблица 2. Извадка II и Извадка III – сравнение на резултатите от приложението на LDA

Извадка	*Вид филтрация*	*Брой думи в речник*	*Избран оптимален брой теми*	*Средна стойност на C_v за избрания брой теми*
Извадка I	**Филтрация 5**	**4 295**	**16**	**0.6655**
Извадка II	Филтрация 2	10 221	21	0.5268
	Филтрация 3	6 278	17	0.5363
	Филтрация 5	**3 725**	**15**	**0.5423**
Извадка III	**Филтрация 2**	**5 191**	**15**	**0.4408**
	Филтрация 3	3 153	16	0.4393
	Филтрация 5	2 007	15	0.3987

Следва обобщение и дискусия на част от най-важните изводи на база на целия емпиричен анализ в посока моделирането на теми, представляващи интерес за клиента:

- ✓ **Речникът от думи, ползван по време на моделирането на теми, има голямо влияние** върху резултатите. Различните филтрации, приложени върху него, могат да доведат до немалки разлики в качеството на получените теми, но **не повлияват избора на оптимален брой теми** – този брой остава сходен. **Филтриране на речника от думи на база на части на речта** и селектиране само на съществителни имена (Филтрация 5) води до **оптимални резултати** в моделирането на теми чрез Извадка I и Извадка II. Може да се направи извода, че съществителните имена предоставят най-важен контекст по отношение на засегнатите от клиентите теми.
- ✓ **Моделът с най-високо качество на темите** е създаден върху **Извадка I** след приложение на **Филтрация 5**. Сравнено с останалите филтрации, тази води до най-голямо съкращение в речника от думи, което се оказва оптимално. Резултатите са пример за това, как **не е важно количеството на думите в речника, а тяхното качество** – чрез подбора се селектират по-важните думи и се филтрира излишния шум, като това води до по-добро разграничение между темите. Така се **оптимизира времето за трениране на модела** и междувременно се **намалява вероятността от прекомерно нагаждане към данните**.
- ✓ Все пак, **типът филтрация е важно да бъде съобразен с големината на извадката (общия брой думи в нея)**. Приложението на Филтрация 5 води до неоптимални резултати върху Извадка III, като най-вероятната причина за това се крие в драстичното намаляване на броя думи, включени в моделирането на теми. Причината е, че тази извадка е значително по-малка, тъй като се състои само от първите реплики на клиента. В тази връзка, съществуват специални методи, които са насочени към **моделирането на**

теми в кратки текстове (такива, например, са статусите в социалната мрежа Twitter) [61]. В случая, такива методи биха били по-подходящи за приложение върху Извадка III, но продължаването на експериментите в такава посока излиза извън обхвата на дисертационния труд и се счита като една **бъдеща перспектива**.

- ✓ На база на резултатите, получени върху Извадка I, **Хипотеза 1 (виж точка 1.4) може да бъде потвърдена**. Макар че онлайн чат комуникацията представлява нестандартен тип данни и се характеризира с много особености, традиционният метод за моделиране на теми **LDA води до създаването на модел, характеризиращ се с висока кохерентност**. Измерената кохерентност на модела, трениран върху Извадка I (Филтрация 5), както и детайлният анализ на темите в него, потвърждават, че LDA се представя на оптимално ниво в задачата за моделиране на теми, водейки до постигането на **работещо решение**.
- ✓ Сравнителният анализ между трите репрезентации на чат комуникацията сочи, че ползването на **Извадка I, която съдържа чатовете в тяхната цялост, води до постигане на оптимални резултати** (независимо от ползваната филтрация на речника от думи). Кохерентността на темите, постигната чрез Извадка II и Извадка III, е значително по-ниска от тази, постигната чрез Извадка I. Това свидетелства за неоптимално общо представяне на моделите и ниско ниво на интерпретация и качество на получените теми. Резултатите върху тези две извадки се характеризират и с **по-голяма вариабилност**, което говори за **по-ниска степен на увереност и надеждност на моделите**.
- ✓ На база на тези резултати, **Хипотеза 2 следва да бъде отхвърлена** – използването само на клиентските реплики в чатовете **не** води до създаване на модел с оптимално представяне, спрямо това, постигнато чрез включването и на операторските реплики в моделирането. Оттук следва, **че репликите на операторите в комуникацията носят важен контекст**, който спомага за извличането на темите, представляващи интерес за клиентите и създаването на работещо решение.
- ✓ Важно е да се отбележи, че макар и качеството на извлечените теми да намалява при моделирането на Извадка II и Извадка III, **оптималният брой теми се запазва** с някои малки разлики (отново е около 15-20 теми). Този резултат е логичен, тъй като очакването би било, независимо от репрезентацията, извлечените теми, представляващи интерес за клиента да имат сходен брой и да се характеризират с доста прилики.

- ✓ Една **бъдеща перспектива** е да се изследва приложението на техники като стеминг и лематизация и техния ефект върху моделирането на теми в данни на български език. Друга **бъдеща перспектива** се състои в изследване на възможностите, които предоставят дълбокото обучение и трансферното обучение в посока на моделирането на теми на български език, и приложимостта на подобни техники върху обекта на дисертационния труд.

В заключение, емпиричното изследване постига всички поставени цели по отношение на моделирането на теми в онлайн чат комуникация. В рамките на емпиричния анализ са изследвани Хипотеза 1 и Хипотеза 2 (виж точка 1.4) и е формиран отговор на един от поставените изследователски въпроси в дисертационния труд. **Комбинацията от традиционния алгоритъм LDA, тестването на различни репрезентации на данните, както и различни филтрации на речника от думи, формират един цялостен ефективен подход за моделиране на теми в онлайн чат комуникация.**

Извлечени са ценни изводи относно различните нива на репрезентация на онлайн чат комуникацията в контактен център - тема, която доколкото е известно на автора, до този момент не е била засягана от други изследователи в сферата. В дисертационния труд е демонстрирано как тези нива на репрезентация на данните влияят върху резултатите и най-вече върху качеството на извлечените теми чрез LDA. В допълнение е предложена и надеждна **методика за обработка и нормализация** на изследвания тип данни преди пристъпване към моделиране на теми.

Доколкото е известно на автора, **представянето, възможностите и ограниченията на алгоритъма LDA, до този момент не са били изследвани в сферата на данните, обект на анализ в дисертационния труд (онлайн чат комуникация в контактен център на банкова институция).** Въпреки очевидните практически ползи от изводите в изследването, конкретно по отношение на комуникация с клиента в банковата сфера, е много важно да се подчертае **приложимостта на създадената система и в други индустрии**, в които подобен тип данни биват генерирани в резултат на бизнес процесите на компанията.

Може да се обобщи, че в дисертационния труд е предложена **цялостна методика** за извършване на анализ на темите, представляващи интерес за клиента, която преодолява някои трудности, възникващи поради характерните особености на чат комуникацията. За разлика от други изследвания, имащи подобни цели, в настоящото е демонстриран **изцяло**

статистически подход. Единственият езиков ресурс, необходим с цел анализът да се приложи върху подобни данни на друг език, е инструмент за маркиране на части на речта. **Това увеличава и възможността подхода, предложен в дисертационния труд, да бъде репликиран и от други изследователи в сферата.**

2.3.2. Анализ на настроението на клиента (Модул III)

Модул III на автоматизираната система за анализ цели да установи надеждна методика за прогнозиране на удовлетвореността на клиента от онлайн чат комуникацията му с контактен център. От Таблица 3 става ясно, че само **13,66%** от всички чатове в Извадка I (тази извадка представлява всички налични данни) се характеризират с рейтинг, поставен от страна на клиента – за останалите **86,34%** няма яснота до колко качествено е била изпълнена услугата. **Целта в Модул III е с помощта на историческа информация и машинно самообучение да се обхванат белези и зависимости в комуникацията между клиент и оператор, които с достатъчна точност да сигнализират, дали клиентът е останал доволен или по-скоро недоволен от предоставената му услуга**. За целите на създаването на модел, прогнозиращ удовлетвореността на клиента, е използвана извадката от чатове, в които рейтингът на клиента е известен – тази извадка се състои от **5 125 чата**, проведени в периода от **28.01.2019 г.** до **01.04.2021 г**. Тази извадка, съдържаща чатовете в техния пълен вид (тоест включвайки всички реплики в тях), бива означена като „Извадка 1“.

Таблица 3. Разпределение на чатовете в Извадка I спрямо наличието на рейтинг

	Брой наблюдения	*В процент*
Чатове с рейтинг	5 125	13,66%
Чатове без рейтинг	32 404	86,34%
Общ брой	37 529	100%

Клиентите могат да поставят само два вида рейтинг на услугата - **„добър“ или „лош“.** Разпределението на целевата променлива е силно небалансирано - от Таблица 4 става ясно, че доминантната категория рейтинг е „добър“ (разбира се, това наблюдение е и до известна степен очаквано). Услуги, оценени като „лоши“, представляват около **10,4%** от извадката. Силното ниво на небалансираност на целевата променлива в извадката със сигурност има влияние върху резултатите и представянето на класификационния модел, но в експериментите е запазено това съотношение между класовете на целевата променлива.

Таблица 4. Разпределение на целевата променлива (рейтинг на услугата)

Рейтинг	*Брой наблюдения*	*В процент*
Лош	533	10.4%
Добър	4 592	89.6%
Общ брой чатове	5 125	100%

В рамките на експеримента са изследвани Хипотеза 3 и Хипотеза 4 (виж точка 1.4). Във връзка с Хипотеза 3 е въведено понятието „наивна прогноза“ - това е прогноза, която **не се базира на машинно самообучение.** В същността на „наивната прогноза“ стои идеята да бъде сравнено представянето на модел, който не знае нищо за данните (не извлича знания от тях, не анализира зависимостите между отделните наблюдения) и модел, който се обучава от историческата информация, която предоставят данните, с цел да направи прогноза за стойността на целевата променлива. Подобно сравнение директно би дало отговор на въпроса, дали въобще има смисъл и полза от това да бъдат ползвани методи за машинно самообучение с цел разрешаване на разглеждания проблем. **Като най-добра отправна точка за сравнение с представянето на създадения прогнозиращ модел е избрана наивна прогноза, която взима предвид разпределението на класовете в данните (това знание е известно предварително).** В случая при този тип „наивна прогноза“ на случаен принцип 10,4% от чатовете се класифицират като „лоши“ и 89.6% като „добри“ (този вариант е познат под името „*weighted guess classifier*“). Стойностите на метриките в Таблица 5 са използвани като отправна точка в оценката на това, до колко приложението на машинно самообучение върху обекта на дисертационния труд помага в задачата за прогнозиране на удовлетвореността на клиента.

Таблица 5. Стойност на метриките за оценка при „наивна прогноза“ в разглеждания казус

	Точност	*Чувствителност*	*Прецизност*	*F1*	*F-beta*
Наивна прогноза	~0.81	~0.104	~0.104	~0.1	~0.1

Цел на анализа е да бъдат тествани различни параметри/техники, представляващи интерес, така че да се получи оптимално представяне на прогнозиращия модел и да се постигне във възможно най-голяма степен работещо решение. Ефектът от следните параметри/техники е изследван по време на експериментирането с данните и търсенето на оптимално решение (в точка 2.2.3 са предоставени повече детайли) – 1. **Ниво на**

репрезентация на онлайн чат комуникацията (свързано с Хипотеза 4 - виж точка 1.4) - създадени са Извадка 1, Извадка 2 и Извадка 3; 2. Приложени **техники за обработка на текста** (стеминг); 3. Експериментиране с различни **набори от обясняващи променливи** („текстови“ и „морфологични“); 4. Приложен **алгоритъм за машинно самообучение**.

Създадена е аналитична процедура, чрез която да бъдат комбинирани и тествани всички възможни комбинации измежду избраните четири параметъра. Важно е да се отбележи, че върху всяко ниво на репрезентация (Извадка 1, 2 или 3) са приложени едни и същи техники за обработка на текста преди моделирането на данните. Създадени са общо **216** модела, прогнозиращи удовлетвореността на клиента (**3** нива на репрезентация на онлайн чат комуникацията Х **4** различни форми на стеминг, приложени върху текста Х **6** различни набора от обясняващи променливи Х **3** различни прогнозиращи алгоритъма). Необходимо е да се отбележи, че анализът е усъвършенстван в някои отношения, сравнено с първоначалните експерименти, публикувани в [50] – 1. подобрена е процедурата по валидация на моделите; 2. всяка форма на стеминг, приложена върху данните (включително и оригиналният вид на текста), е включена като отделен параметър в задачата за прогнозиране; 3. Експериментът е разширен – включени са още два метода за машинно самообучение (Бернулиев Наивен Бейсов модел и класификатор с опорни вектори).

За всяко ниво на репрезентация на данните са създадени общо 72 модела. Моделирана е **вероятността клиентът да постави „лош“ рейтинг на услугата, предоставена му посредством онлайн чат**. В оценката на представянето на моделите, с най-голяма тежест биват взети предвид метриките чувствителност и F-beta, тъй като от основно значение за бизнеса е създаденият прогнозиращ модел точно да разпознава онлайн чат комуникация от която клиентът **не е останал доволен.** Акцентира се върху способността на модела да идентифицира правилно всички наблюдения, попадащи в класа с рейтинг „лош“, дори това да е до някаква степен за сметка на прецизността му. За улеснение на сравнителния анализ, Таблица 6 представя трите модела с оптимално представяне (според F-beta), съответно за всяко едно ниво на репрезентация на данните.

Таблица 6. Модели с оптимално представяне според F-beta (за всяка извадка)

Извадка	*Алгоритъм*	*Обработка на текста*	*Набор от обясняващи променливи*[10]	*F1*	*Точност*	*Чувстви-телност*	*Прециз-ност*	*F-beta*
Наивна прогноза				~ 0.1	~ 0.81	~ 0.104	~ 0.104	~ 0.1
Извадка 1	SVC	Оригинална форма	Модел с n-grams + части на речта	0.8921	0.8972	0.4089	0.5059	0.4250
Извадка 2	SVC	Стеминг 1	Модел с n-grams + части на речта	0.884	0.889	0.383	0.457	0.395
Извадка 3	**BNB**	**Стеминг 3**	**Модел с n-grams**	**0.8630**	**0.8533**	**0.4784**	**0.3494**	**0.4450**

Следва кратко резюме и дискусия на част от по-важните изводи, направени по време на емпиричния анализ, посветен на моделирането на удовлетвореността на клиента:

- ✓ **<u>Потвърдена е Хипотеза 3</u>** - приложението на методи за машинно самообучение с цел разрешаване на задачата за прогнозиране на удовлетвореността на клиента, води до постигането на оптимални резултати, спрямо тези получени чрез „наивни“ подходи, които не се базират на техники за извличане на знания от данни. Всеки от моделите в Таблица 6 се представя по-ефективно в задачата за прогнозиране на удовлетвореността на клиента, в сравнение с това, което може да се постигне чрез „наивна“ прогноза.
- ✓ Приложените емпирични техники позволяват формирането на отговор на втория изследователски въпрос, залегнал в основите на дисертационния труд (виж точка 1.4). **Чатовете, сами по себе си, съдържат достатъчно сигнали за удовлетвореността на клиента, така че чрез тях да се създаде прогнозиращ модел, който представлява надеждна основа за разработването на работещо решение в бизнеса.**
- ✓ Избраният модел, характеризиращ се с **оптимално представяне** спрямо всички останали (според F-beta), е създаден само с помощта на финалните реплики на клиента (Извадка 3), върху които е приложен „Стеминг 3“. Този модел е създаден с помощта на Наивния Бейсов алгоритъм, приложен върху набор от униграми и биграми (*bigrams*), извлечени от текста. В тази връзка е **<u>потвърдена Хипотеза 4</u>** - ползването на извадка, състояща се от финалните реплики на клиента в комуникацията, води до постигането на оптимални резултати в прогнозирането (според метриката F-beta, която има най-голяма тежест в процеса на оценка на резултатите).

[10] N-grams моделът включва използването на т.нар. униграми (единични думи) и биграми (две последователни думи), извлечени от текста. Към тях може да бъде добавена и информация за частите на речта. Повече детайли относно наборите от обясняващи променливи са налични в основния текст на дисертационния труд.

- ✓ От последното може да се направи изводът, че **финалните клиентски реплики носят най-ценна информация и важни сигнали по отношение на удовлетвореността на клиента.** В подкрепа на този извод е и фактът, че според конкретни метрики за оценка, ефективността на моделите, тренирани върху Извадка 3, е на същото ниво, ако не и по-добро спрямо това, постигнато чрез останалите нива на репрезентация. Едно предимство на Извадка 3 е, че в нея е осъществен по-голям подбор на променливи, тъй като тя включва само финалните реплики на клиентите. Това води до намаляване на вероятността за прекомерно нагаждане към данните, както и до съкращаване на времето за трениране на модела. Очевидно, голяма част от излишната информация (която не добавя ценни знания), бива изключена на това ниво на репрезентация на данните.
- ✓ **Потвърждението на Хипотеза 4 и направените изводи са в унисон с резултатите на други изследователи по отношение на конкретно разглеждания проблем**. Park и съавтори [16] анализират кои са обясняващите променливи, които се характеризират с най-голяма предсказваща способност в задачата за прогнозиране на удовлетвореността на клиента. Техните резултати показват, че настроението на клиента във финалните му реплики в комуникацията е сред най-важните променливи в прогнозирането.

В заключение, предложен е **изцяло статистически подход** за анализ на удовлетвореността на клиента от услугите, предоставени в контактен център чрез онлайн чат. Създадена е методика, която се базира изцяло на информацията, която предоставят текстовите диалози между клиенти и оператори. Разглежданият проблем е анализиран от различни гледни точки, като са изследвани три нива на репрезентация на данните и са извлечени изводи относно техния ефект и полезността им в задачата за прогнозиране на удовлетвореността. **Доколкото е известно на автора, настоящото изследване е първото такова, изрично насочено към прогнозирането на удовлетвореността на клиента от чат комуникация, базирайки се единствено на текстови характеристики и граматическа информация, извлечена от текста.** Изследването допринася и за развитието на науката в областта на приложния анализ на текстови данни на български език, с цел разрешаване на актуални икономически и бизнес проблеми.

Важно е да се отбележи, че предложената методика не разчита на лексикони на настроението, което я прави подходяща за езици с малко ресурси, какъвто е и българският език. Настоящото изследване е първото, което засяга тази **комбинация от**

изследван проблем и език, на който са текстовите данни (доколкото е известно на автора). Изследването намира място сред малкото такива, посветени на анализ на настроението в онлайн чат комуникация, проведена с цел обслужване на клиенти. Данните, използвани с цел апробиране на създадената система за анализ, са в банковата сфера, но представените **методика, аналитични техники и подход за интерпретация** могат да бъдат приложени и върху други подобни данни, генерирани в разнообразни индустрии. В допълнение, резултатите и изводите могат да са полезни и на други изследователи в сферата, които анализират подобен тип данни.

Една от **бъдещите възможности за развитие** е в посока на приложението на трансферно обучение и модели тип „Трансформатор“. Бъдеща перспектива представлява изследването на това, дали съществуват предварително обучени модели (*pre-trained models*), които могат да бъдат приложени върху данни на български език и ако това е така - да се потърси отговор на въпроса, до каква степен тези модели са приложими върху обекта на дисертационния труд, както и в задачата, която бива разрешавана. Интересна посока за бъдеща разработка се състои в изследването, дали подобни подходи биха могли да се комбинират с техниките, илюстрирани в дисертационния труд, така че да се получи още по-точна оценка за удовлетвореността на клиента. Интерес представляват и възможностите за приложение на методи за дълбоко обучение и съпоставката им с класическия подход за машинно самообучение, използван в дисертационния труд. Разбира се, ефективни резултати от приложението на методи, базирани на изкуствени невронни мрежи, биха се очаквали в случай, че бъде натрупан по-голям обем от налични данни в настоящия казус.

Друга посока за развитие се базира на част от резултатите, публикувани в [16]. Авторите стигат до извода, че метаданните към подобни текстови диалози между клиенти и оператори (например, продължителност на чата, реакция на оператора, време от деня, брой реплики и други) имат ниска предсказваща способност в задачата за прогнозиране на удовлетвореността. Една бъдеща посока за развитие е да се изследва какъв ще е ефектът от добавянето на метаданни като обясняващи променливи в задачата за прогнозиране на удовлетвореността и дали те ще подобрят представянето на модел, създаден само чрез „текстови“ променливи. Например, интересно би било да се изследва, дали по-голямата средна скорост на отговор на оператора, както и максималното му забавяне са фактори в поставянето на „лош“ рейтинг. Интересна идея е да се съпоставят резултатите, получени

чрез използването само на структурирани данни (метаданни) с цел прогнозиране на удовлетвореността на клиента и тези, получени чрез ползването само на неструктурирани данни (текста на чатовете), по подобие на идеята, илюстрирана в [21].

2.3.3. Обобщение и визуализация на резултатите (Модул IV)

Модул IV обединява допълнителни аналитични техники за извличане на знания от данните, чиято основна цел е да:

- ✓ Подпомогнат в интерпретацията и придобиването на повече контекст относно извлечените теми, представляващи интерес за клиента.
- ✓ Да се проследи развитието на темите във времето.
- ✓ Да се състави оценка за нивото на комплексност на засегнатите теми/заявки.
- ✓ Да се обединят знанията, извлечени в Модул II и Модул III, като по този начин се получи синергичен ефект от анализа на темите, представляващи интерес за клиента и анализа на удовлетвореността му от обслужването.

Сред приложените аналитични техники в Модул IV намират място: 1. Графичен анализ и опознаване на темите, представляващи интерес за клиентите с помощта на т.нар. „облаци от думи“; 2. Създаване на индикатори за **сложността** на клиентските заявки, попадащи в дадена тема (през призмата на метаданните, налични към всеки един чат в извадката – например, продължителност на чата, брой изказвания в него, скорост на отговор на оператора); 3. Анализ на удовлетвореността на клиента в дискусии, попадащи в различни теми; 4. Извличане на ключови фрази от дискусиите, попадащи в определени теми; 5. Графичен анализ на разпределението на темите във времето (откриване на пикове в конкретен тип заявки и проследяване на тенденции сред интересите на клиентите).

В Модул IV са илюстрирани само част от възможните приложения на системата за анализ на онлайн комуникация, която се характеризира с множество посоки за усъвършенстване и добавяне на разнообразни функционалности в полза на крайния потребител в лицето на бизнеса. Потенциално системата би могла да се разшири и да се включат още данни от други канали за комуникация с клиента. На първо място, от подобни анализи в дадена компания най-вече биха се възползвали мениджърите в отдел „Контактен център“, тъй като са предоставени директни измерители за представянето на операторите и покритието на различни категории заявки. Подобни анализи биха довели до редица

оптимизации на процеси в контактния център. Отделно от това, последните години се наблюдава тенденцията да се имплементират автоматизирани диалогови системи с цел оптимизация на процеса на обслужване. В тази връзка, моделирането на теми и извличането на ключови фрази от диалозите могат да бъдат полезни в поне две направления.

Първо, предоставена е оценка за сложността на клиентските заявки, засягащи определени теми. Последното е в пряка връзка с това, кой тип заявки се характеризират с най-голям потенциал да бъдат автоматизирани лесно, и обратното - кои заявки ще е най-трудно да се поемат от подобни системи и биха изисквали надзор от страна на операторите. Второ, извличането на ключови фрази, както и подобни анализи в тази посока (например, извличане на синонимни фрази) биха били от голяма полза в процеса на обучение на чатбот (често срещаните клиентски фрази могат да играят ролята на тренировъчни данни).

Друго звено в компанията, заинтересовано от системата за анализ на онлайн комуникацията с клиента е отдел „Качество на обслужването“. Интерес би представлявало да се анализира в кои случаи е по-вероятно клиентът да остане разочарован и какви са причините за това (в детайл), като се разгледат подробно такива дискусии. Така биха се идентифицирали „проблемни“ теми за клиента и посоки, в които услугите е необходимо да се подобрят. Компаниите често използват анкети с цел анализ на мненията сред клиентите. Този подход обаче се базира на това, какво компанията смята за важно и интересно. Чрез използването само на такива методи, някои важни теми могат да останат в „сляпото петно“ на бизнеса. Именно поради това, анализи на непринудената комуникация между клиент и оператор са много ценни, тъй като могат да бъдат открити важни теми и въпроси, за които хората от бизнеса въобще не са се досетили.

Друго звено, което също би се възползвало от създадената система, е отдел „Управление на взаимоотношенията с клиенти“. С помощта на търсене по ключови думи, в комуникацията биха могли да се идентифицират много конкретни въпроси/проблеми на клиенти във връзка с определени продукти, услуги и теми. Системата би могла да подпомогне и проследяването на успеха на маркетингови кампании и различни инициативи на компанията. В допълнение, знанията, извлечени от комуникацията с клиенти, биха могли да спомогнат в профилирането на клиенти и изготвянето на специални предложения.

В заключение, анализът на темите, които вълнуват клиентите, както и на удовлетвореността им от предоставените услуги в контактния център, в комбинация с

анализ на метаданните, налични към всеки един чат, водят до постигането на множество синергични ефекти и извличане на допълнителни знания от данните. Илюстрирани са само част от ползите за бизнеса с помощта на конкретни примери и идеи за отдели, които биха се възползвали от функционалностите на системата за анализ на онлайн комуникацията с клиента. Така, резултатите в дисертационния труд са представени не само от аналитична гледна точка, но и от гледна точка на тяхното значение и интерпретация за хората от бизнеса - **подчертани са приложността и смисъла на резултатите в практиката**.

III. Заключение

В заключение, с помощта на приложения изследователски метод, базиращ се изцяло на количествени методи за анализ на текстови данни, основната цел на дисертационния труд е постигната. Създадена е 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. Илюстрирана е структурата на една подобна система, изложени са подробно методите за изграждането ѝ, както и нейни конкретни приложения в бизнеса. Предоставена е и актуална картина и детайлен анализ на постигнатото, и бъдещите посоки за развитие в сферата на анализа на текстови данни на български език. Чрез проведените емпирични експерименти е предоставен отговор на двата изследователски въпроса, залегнали в основите на дисертационния труд. Всяка една от хипотезите, дефинирани във връзка с тези два изследователски въпроса, е проверена с помощта на специално избрани подходящи метрики за оценка на получените резултати и ефективността на приложените методи.

Установена е методология, базираща се на комбинация от аналитични процедури и техники, чрез които да бъде открит отговор на ключовия въпрос, залегнал в тезата на дисертационния труд. Резултатите, документирани в рамките на труда и свързаните с него публикации, потвърждават съществуването на ефективни начини за количествен анализ и извличане на ценна информация от изследвания тип данни. Оттук следва, че онлайн чат комуникацията между клиенти и оператори в контактен център, представлява неоползотворен богат източник на информация за отношенията на клиентите с компанията. Демонстрирано е как тази информация може да бъде ефективно извлечена, структурирана и анализирана с помощта на техники от сферата на обработката на естествен език и

машинното самообучение, с цел изграждането на автоматизирана система за анализ, имаща ценни приложения в бизнеса.

Дисертационният труд насочва изследователския интерес, както и потенциално този на бизнеса, към възможните приложения на анализа на онлайн чат комуникация между клиенти и служители - с какво такъв тип данни могат да бъдат полезни за бизнеса, каква информация може да бъде извлечена от тях, какви техники за обработка и анализ могат да бъдат приложени с цел извличане на ценни знания. Поставен е фокус върху най-важните аспекти за бизнеса в една такава комуникация - удовлетвореността на клиентите и основните теми, които присъстват в дискусиите им със служителите на компанията. Последното придобива още по-голяма важност в контекста на пандемията от COVID-19 и огромния ръст в онлайн комуникацията в глобален мащаб - това води и до прекомерна натовареност на служителите в контактни центрове, които са на първа линия в тази ситуация.

Практическите ползи за бизнеса от създаването на една подобна система за анализ на онлайн комуникацията с клиента са много - подобряване на обслужването на клиентите, по-добро разбиране на клиентските нужди и проблеми, усъвършенстване на продукти и услуги, откриване на тенденции и теми, които вълнуват най-много клиентите, проследяване на ефективността/успеха на различни кампании, подпомагане на процеса на създаване на чатбот или друг тип автоматизирана диалогова система и много други. Макар че системата за анализ е апробирана в рамките на конкретен казус в банковата сфера, предложената от автора **методика за интерпретация на получените резултати, както и използваните аналитични техники са приложими** във всички останали индустрии, в които подобен тип данни биват генерирани в резултат на бизнес процесите на компанията. Направените изводи могат да бъдат ценни във всеки един бизнес контекст, в който ориентираността към клиента е от най-голямо значение и представлява ключ към развитие на бизнеса и бъдещ успех.

IV. Използвана литература[11]

[1] S. Gupta and D. Ramachandran, "Emerging market retail: transitioning from a product-centric to a customer-centric approach," Journal of Retailing, vol. 97, no. 4, pp. 597-620, 2021.

[2] M. A. Camilleri, "The use of data-driven technologies for customer-centric marketing," International Journal of Big Data Management, vol. 1, no. 1, pp. 50-63, 2020.

[3] Dimension Data, "2017 Global Customer Experience Benchmarking Report. Digital crisis or redemption. The uncomfortable truth," 2017.

[4] A. Qasem and W. Alhakimi, "The Impact of Service Quality and Communication in Developing Customer Loyalty: The Mediating Effect of Customer Satisfaction," Journal of Social Studies, vol. 25, no. 4, 2019.

[5] B. Lobe, D. Morgan and K. A. Hoffman, "Qualitative data collection in an era of social distancing," International Journal of Qualitative Methods, vol. 19, no. 1-8, Art. no. 1609406920937875, 2020.

[6] KPMG, "Standing firm on shifting sands. Global banking M&A outlook H2 2020," 2020.

[7] Mordor Intelligence, "Big Data Analytics In Banking Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026)," 2021.

[8] E. Pronoza, A. Pronoza and E. Yagunova, "Extraction of Typical Client Requests from Bank Chat Logs," in Mexican International Conference on Artificial Intelligence, 2018.

[9] S. Kumar, A. K. Kar and P. V. Ilavarasan, "Applications of text mining in services management: A systematic literature review," International Journal of Information Management Data Insights, vol. 1, no. 1, Art. No. 100008, 2021.

[10] J. Wang, J. Wang, C. Sun, S. Li, X. Liu, L. Si, Z. M. and G. Zhou, "Sentiment Classification in Customer Service Dialogue with Topic-Aware Multi-Task Learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[11] M. Zaki and A. Neely, "Customer experience analytics: dynamic customer-centric model," Handbook of Service Science, Volume II, pp. 207-233, 2019.

[12] S. Roy, R. Mariappan, S. Dandapat, S. Srivastava, S. Galhotra and B. Peddamuthu, "QART: A System for Real-Time Holistic Quality Assurance for Contact Center Dialogues," in Thirtieth AAAI conference on artificial intelligence, 2016.

[13] M. Anandarajan, C. Hill and T. Nolan, Practical text analytics. Maximizing the Value of Text Data, Springer, 2019.

[14] G. Miner, D. Delen, J. Elder, A. Fast, T. Hill and B. Nisbet, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[15] A. Magueresse, V. Carles and E. Heetderks, "Low-resource languages: A review of past work and future challenges," arXiv:2006.07264 [cs.CL], 2020.

[16] K. Park, J. Kim, J. Park, M. Cha, J. Nam, S. Yoon and E. Rhim, "Mining the minds of customers from online chat logs," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 2015.

[17] G. Hristova, "Text Analytics in Bulgarian: An Overview and Future Directions," Cybernetics and Information Technologies, vol. 21, no. 3, pp. 3-23, 2021.

[18] G. Hristova, "A SURVEY OF TEXT MINING METHODS APPLIED ON CONVERSATIONAL DATA," Scientific Research of the Union of Scientists in Bulgaria – Plovdiv, series B. Natural Sciences and Humanities, Vol XX. VIIIth International Conference of Young Scientists, 2020.

[19] C. H. Chen, W. P. Lee and J. Y. Huang, "Tracking and recognizing emotions in short text messages from online chatting services," Information Processing & Management, vol. 54, no. 6, pp. 1325-1344, 2018.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, pp. 5998-6008, 2017.

[21] B. Velichkov, I. Koychev and S. Boytcheva, "Deep learning contextual models for prediction of sport event outcome from sportsman's interviews," in Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), 2019.

[22] Y. Dinkov, I. Koychev and P. Nakov, "Detecting Toxicity in News Articles: Application to Bulgarian," arXiv:1908.09785 [cs.CL], 2019.

[23] M. Hardalov, I. Koychev and P. Nakov, "Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian," arXiv:1908.01519 [cs.CL], 2019.

[24] B. Velichkov, S. Gerginov, P. Panayotov, S. Vassileva, G. Velchev, I. Koychev and S. Boytcheva, "Automatic ICD-10 codes association to diagnosis: Bulgarian case," in CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, 2020.

[25] A. Popov, P. Osenova and K. Simov, "Implementing an End-to-End Treebank-Informed Pipeline for Bulgarian," in Proceedings of the 19th Workshop on Treebanks and Linguistic Theories, 2020.

[26] G. Hristova, B. Bogdanova and N. Netov, "Design of ML-based AI System for Mining Public Opinion on E-government Services in Bulgaria," in AIP Conference Proceedings, (accepted for publication).

[27] G. Hristova, B. Bogdanova and N. Netov, "Data Mining of Public Opinion: An Overview," in AIP Conference Proceedings, (accepted for publication).

[28] B. Bogdanova and E. Stancheva-Todorova, "ML-based predictive modelling of stock market returns," in AIP Conference Proceedings, 2021.

[29] G. Mengov, I. Nenov and I. Zinovieva, "A model for collective emotion forecasts financial data," IFAC-PapersOnLine, vol. 52, no. 25, pp. 208-213, 2019.

[30] I. Ivanov, S. Kabaivanov and B. Bogdanova, "Stock market recovery from the 2008 financial crisis: The differences across Europe," Research in International Business and Finance, vol. 37, pp. 360-374, 2016.

[11] Пълният списък на използваната литература може да бъде открит в дисертационния труд.

[31] E. Stancheva-Todorova and B. Bogdanova, "Enhancing investors' decision-making–An interdisciplinary AI-based case study for accounting students," in AIP Conference Proceedings, 2021.

[32] I. Nenov, G. Mengov, K. Ganev and R. Simeonova–Ganeva, "Neurocomputational economic forecasting with a handful of data," Comptes rendus de l'Académie bulgare des Sciences, vol. 74, no. 10, 2021.

[33] B. Kapukaranov and P. Nakov, "Fine-grained sentiment analysis for movie reviews in Bulgarian," in Proceedings of the International Conference Recent Advances in Natural Language Processing, 2015.

[34] J. Boyd-Graber, Y. Hu and D. Mimno, "Applications of Topic Models," Foundations and Trends in Information Retrieval, vol. 11, no. 2-3, p. 143–296, 2017.

[35] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," Multimedia Tools and Applications, vol. 78, no. 11, pp. 15169-15211, 2019.

[36] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," EAI Endorsed transactions on scalable information systems, vol. 7, no. 24, 2020.

[37] A. M. Dean, "The Impact of the Customer Orientation of Call Center Employees on Customers' Affective Commitment and Loyalty," Journal of Service Research, vol. 10, no. 2, pp. 161-173, 2007.

[38] A. Rafaeli, L. Ziklik and L. Doucet, "The Impact of Call Center Employees' Customer Orientation Behaviors on Service Quality," Journal of Service Research, vol. 10, no. 3, pp. 239-255, 2008.

[39] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.

[40] B. Liu, "Sentiment analysis and opinion mining," Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1-167, 2012.

[41] W. Medhat, A. Hassan and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5, no. 4, pp. 1093-1113, 2014.

[42] S. Baccianella, A. Esuli and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), 2010.

[43] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," arXiv:1103.2903 [cs.IR], 2011.

[44] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," in Proceedings of the International AAAI Conference on Web and Social Media, 2014.

[45] W. L. Hamilton, K. Clark, J. Leskovec and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in Proceedings of the conference on empirical methods in natural language processing, 2016.

[46] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowledge-Based Systems, vol. 89, pp. 14-46, 2015.

[47] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang and C. Liu, "A survey on deep transfer learning," in International conference on artificial neural networks, 2018.

[48] G. Hristova, "Topic modeling of chat data: A case study in the banking domain," in AIP Conference Proceedings, 2021.

[49] G. Hristova, "Topic Modeling of Chat Data: Experimenting with Different Levels of Chat Data Representation by Utilizing a Latent Dirichlet Allocation Model," Journal of Economic Boundaries And Transformation, (accepted for publication).

[50] G. Hristova, "Text Analytics for Customer Satisfaction Prediction: A Case Study in the Banking Domain," in AIP Conference Proceedings, (accepted for publication).

[51] S. Syed and M. Spruit, "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation," in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2017.

[52] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[53] M. Röder, A. Both and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 2015.

[54] P. Nakov, "BulStem: Design and evaluation of inflectional stemmer for Bulgarian," in Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), 2003.

[55] K. Simov, P. Osenova and M. Slavcheva, "BTB-TR03: BulTree-Bank Morphosyntactic Tagset. BTB-TS version 2.0.," Technical report, Bulgarian Academy of Sciences, Sofia, Bulgaria, 2004.

[56] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, New York: Springer, 2013.

[57] H. Lei and Y. Chen, "Concentrated Document Topic Model," arXiv:2102.04449 [stat.ML], 2021.

[58] R. Taylor and J. A. D. Preez, "ALBU: An approximate Loopy Belief message passing algorithm for LDA to improve performance on small data sets," arXiv:2110.00635 [cs.LG].

[59] K. Stevens, P. Kegelmeyer, D. Andrzejewski and D. Buttler, "Exploring topic coherence over many models and many topics," in Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012.

[60] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 2014.

[61] J. Qiang, Z. Qian, Y. Li, Y. Yuan and X. Wu, "Short text topic modeling techniques, applications, and performance: a survey," IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 8, 2020.

V. Приноси на дисертационния труд

1. Създаване на широкообхватен литературен източник, представящ актуална картина на развитието в сферата на обработката на естествен език в България, както и на възможностите за анализ на текстови данни на български език.
2. Създаване на методика за интерпретация на получените резултати, характеризираща се с приложна значимост и възможност за екстраполация на стъпките в нея. Методиката е приложима не само в банковата индустрия, но и във всички останали индустрии, в които подобен тип данни биват генерирани в резултат на бизнес процесите на компанията.
3. Създаване на автоматизация на процеса на извличане на знания от онлайн чат комуникация с клиента, проведена в контактен център. Илюстрираните количествени методи за анализ могат да бъдат комбинирани с установени подходи за качествен анализ, така че бизнесът да може да взима по-добри решения, характеризиращи се с по-висока степен на рационалност.
4. Като допълнителен принос може да се разглежда и възможността, създадената автоматизирана система да се адаптира при постъпването на нови данни и информацията да се обновява в реално време. Важно е да се отбележи, че макар и системата да е адаптивна, съществува необходимост от провеждане на периодичен мониторинг и актуализация на прогнозиращите модели, спрямо променящите се характеристики на данните във времето.
5. Създаване на методология за анализ на онлайн чат комуникация с клиенти конкретно на български език. Конкретни алгоритми и процедури за обработка и анализ могат да бъдат репликирани от други изследователи в сферата и използвани с цел автоматизиране и улеснение на анализа на подобен тип данни (например, създаване на библиотеки, модули и други).
6. Създадена е методика за прогнозиране на удовлетвореността на клиента от онлайн чат комуникация, която се базира единствено на текстови характеристики и граматическа информация, извлечена от текста.
7. Създадена е и методика за извършване на анализ на темите, представляващи интерес за клиента. Посочени са важни техники за обработка и моделиране, необходими с цел извличането на знания от изследвания тип данни. Направени са ценни изводи относно

различните нива на репрезентация на онлайн чат комуникацията в контактен център - тема, която доколкото е известно на автора, до този момент не е била засягана от други изследователи в сферата.

VI. Списък с публикации по дисертацията

1. G. Hristova, "A SURVEY OF TEXT MINING METHODS APPLIED ON CONVERSATIONAL DATA," Scientific Research of the Union of Scientists in Bulgaria – Plovdiv, series B. Natural Sciences and Humanities, Vol XX. VIIIth International Conference of Young Scientists, 2020.
2. G. Hristova, "Text Analytics in Bulgarian: An Overview and Future Directions," Cybernetics and Information Technologies, vol. 21, no. 3, pp. 3-23, 2021.
3. G. Hristova, "Topic modeling of chat data: A case study in the banking domain," in AIP Conference Proceedings, 2021.
4. G. Hristova, "Topic Modeling of Chat Data: Experimenting with Different Levels of Chat Data Representation by Utilizing a Latent Dirichlet Allocation Model," Journal of Economic Boundaries and Transformation, (accepted for publication).
5. G. Hristova, "Text Analytics for Customer Satisfaction Prediction: A Case Study in the Banking Domain," in AIP Conference Proceedings, (accepted for publication).