

ТЕОРЕТИЧНИ АСПЕКТИ НА СТАТИСТИЧЕСКИЯ АНАЛИЗ НА НЕСТРУКТУРИРАНИ ДАННИ ОТ ЧАТБОТОВЕ ЗА УСЪВЪРШЕНСТВАНЕ НА КЛИЕНТСКОТО ПРЕЖИВЯВАНЕ

Theoretical aspects of statistical analysis of unstructured data from chatbots for
improving customer experience

Биляна Голешова¹

e-mail: bilyana.goleshova@unwe.bg¹

Абстракт

В днешния дигитализиран свят една от компютърните програми, която спестява много време и усилия както на потребителите, така и на компаниите е чатбота. Той улеснява обслужването на клиенти чрез автоматизация и, едновременно с това, генерира голям обем високоразмерни, неструктурирани текстови данни. Настоящият доклад има за цел да създаде теоретична и методологическа рамка за статистически анализ на разговорите в чатбот чатове, като преобразува неструктурирания диалог в статистически измерими величини за нуждите на икономическия анализ. Тъй като традиционните статистически методи не са достатъчни за извличане на икономическа оценка от такъв тип текст за подобряване на начина, по който се обслужват клиентите, се приложи подход, който се фокусира върху усъвършенствани методи за обработка на естествен език (NLP) като: токенизация, лематизация, стоп думи, векторизация, анализ на настроеността и тематично моделиране. Приложи се и статистика за извличане на изводи и заключения и прогнозни модели, за да се разбере дали промяната в чатбота наистина води до удовлетворяване на потребителите и да се прогнозира как ще бъде в бъдеще. Даде се светлина и върху това как анализа на такива данни подобрява клиентското изживяване.

Накрая се достигна до заключението, че целта на доклада е напълно изпълнена. Той представи нужният теоретичен фундамент като показва връзката между статистическия анализ на езикови данни и възможността за вземане на информирани решения, които усъвършенстват потребителското преживяване, увеличават клиентската лоялност и подобряват процесите.

Abstract

In today's digitalized world, one of the computer systems that saves a lot of time and effort for both users and the company is the chatbot. It facilitates customer service through automation and at the same time creates a large volume of high-dimensional, unstructured text data. This report aims to create a theoretical and methodological framework for statistical analysis of chatbot conversations by transforming unstructured dialogue into statistically measurable quantities for the needs of economic analysis. Since traditional statistical methods are not sufficient to extract economic value from this type of text to improve the way customers are served, an approach was applied that focuses on advanced natural language processing (NLP) methods such as: tokenization, lemmatization, stop words, vectorization, sentiment analysis, and thematic modeling. Inferential statistics and predictive models were also applied to understand whether the change in the chatbot really leads to user satisfaction and to predict how it will be in the future. It also sheds light on how analyzing such data improves the customer experience.

¹ Докторант, катедра „Статистика и иконометрия“, факултет „Приложна информатика и статистика“, Университет за национално и световно стопанство

PhD student, Department of Statistics and Econometrics, Faculty of Applied Informatics and Statistics, University of National and World Economy

Finally, it was concluded that the purpose of the report was fully fulfilled. It presented the necessary theoretical foundation by showing the connection between the statistical analysis of linguistic data and the ability to make informed decisions that enhance the user experience, increase customer loyalty and improve processes.

Ключови думи: неструктурирани данни, статистически анализ, чатбот, клиентско преживяване
JEL: C45, C55, D91, M31

Въведение

С напредването на технологиите се улесни и начинът, по който организациите в сферата на обслужване на клиенти взаимодействат със своите потребители. Обработвайки милиони разговори ежедневно, чатботовете генерират огромни количества неструктурирани текстови данни под формата на диалог. Този вид данни са безценни, но скрити. Безценни са, защото издават емоциите, желанията и болките на потребителите, благодарение на което компаниите имат възможност да разберат нуждите им и да усъвършенстват клиентското преживяване. Същевременно са известни като скрити, защото теоретичните аспекти на ефективното извличане на знание от тях остава предизвикателство. Традиционните статистически методи са създадени за работа със структурирани количествените данни и изпитват затруднения при срещата със сложния и неструктуриран естествен език. Поради това настоящият доклад има за цел да се представят теоретичните основи и методологии за статистически анализ на неструктурирани чатбот данни с фокус върху усъвършенстването на клиентското преживяване.

Теоретични основи на изследването

1. Преглед на литературата по темата

Значително нарастващата употреба на чатботове и ключовата роля на анализа на неструктурирани данни, генерирани от тези системи, за оптимизиране на клиентското преживяване, са водеща тема, която намира силен отзвук в съвременната научна литература. Доказателство за мащабността и важността на чатбота в сферата на обслужване на клиенти са докладите на Juniper Research, които посочват, че чатботовете ще спестят милиарди долари на индустрията и ще обработват над 50% от всички комуникации в електронната търговия до средата на десетилетието.

В едно от своите изследвания свързано с големите данни българският автор Dimitrova V. (2025) изразява мнение, че това, което е измеримо и класифицируемо може да се преобразува в данни и че в повечето случаи данните се събират, обработват и анализират с идеята да се подобри разбирането на определена система. Тези нейни твърдения важат с пълна сила относно извличането, обработването и анализа на неструктурирани данни от чатботове, с цел да се подобри системата, а това да доведе и до усъвършенстване на клиентското изживяване.

Настоящият преглед на литературата разглежда темата през погледа на изследователите в три основни направления:

➤ Клиентското преживяване и чатбот взаимодействия

Основополагащи са твърденията за потребителското преживяване на Verhoef, P. C. (2009) и Lemon, K. N. & Verhoef (2016). Според Verhoef, P. C. (2009) то е цялостен, взаимосвързан подход, който се формира не само от директните взаимодействия по време на покупката, но и от индиректните взаимодействия, като включва когнитивни, емоционални, поведенчески, сетивни и социални реакции на клиента. По-късно Lemon, K. N. & Verhoef, P. C. (2016) стигат до заключението, че чатботът е само една точка на контакт, която трябва да бъде безпроблемно управлявана и интегрирана в цялото пътуване на клиента, за да се постигне положително преживяване.

Основен извод на Gnewuch, U., Morana, S., & Maedche, A. (2017) е, че успехът на чатботовете в обслужването на клиенти зависи от точния баланс между ефективност и социални умения, както и от точното им позициониране спрямо типа заявка.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020) също взимат отношение по въпроса смятайки, че дизайнът и характеристиките на чатбота оказват влияние върху удовлетвореността и намеренията за бъдещо поведение на потребителите.

➤ Обработка на естествен език в анализа на чатбот данни

Позицията на Jurafsky, D., & Martin, J. H. (2023) относно обработката на естествен език в анализа на чатбот данни е следната: Естественият език предоставя основните методи и инструменти за превръщане на неструктурираните текстови данни, генерирани от чатбот взаимодействията, в структурирана, приложима информация, като позволява на системите да разбират, интерпретират и класифицират намеренията на клиента, чувствата и темите на дискусията, с цел оптимизиране на клиентското преживяване. Това, което позволява на чатбот данните да бъдат анализирани и превърнати в стратегическа информация е, че благодарение на естествения език се разбира целта на клиента, намеренията и чувствата му.

Два аргумента на Pang, B., & Lee, L. (2008) и Liu, B. (2020) към анализа на чатбот данни се фокусират върху анализа на настроеността и фактологичното извличане на знания от текст. Основното мнение на Pang, B., & Lee, L. (2008) е, че NLP може систематично да класифицира и извлича субективна информация от текст, като определя дали изразеното мнение в чатбот разговор е положително, отрицателно или неутрално, както и каква е силата на това определеното чувство. Liu, B. (2020) констатира, че обработката на естествен език е незаменима за извличането на детайлно знание от чатбот разговорите, чрез техники като анализ на настроеността и извличане на мнения, което позволява на бизнеса да идентифицира и количествено да измери не само общото настроение на клиента, но и кои конкретни аспекти на услугите или продуктите предизвикват тези реакции.

Последната теза, която се разгледа по въпроса е на Adamopoulou, E., & Moussiades, L. (2020). Твърдят, че чрез компонентите си за разбиране и генериране естественият език е технологичният стълб, който определя типа, функционалността и сложността на всеки чатбот, като по този начин влияе върху качеството и количеството на данните, които могат да бъдат анализирани за подобряване на клиентското преживяване.

➤ Статистически и машиннообучаващи подходи за анализ на неструктурирани данни

Превръщането на свободен, словесен текст в структурирана форма позволява използването на статистически методи и алгоритми за машинно обучение

Твърдението на Creswell, J. W., & Plano Clark, V. L. (2017), отнасящо се до начинът, по който машиннообучаващите подходи трябва да бъдат комбинирани в изследване за постигане на по-ефективни резултати е следното: Най-силното и изчерпателно разбиране на смесени явления, като клиентското преживяване с чатботове, се постига чрез прилагане на смесени методи, при които количествените и качествените подходи се обединяват, анализират и интегрират в рамките на едно проучване.

Наблюдава се сходство в твърденията на разгледаната литература, което дава стабилно основание за извършване на теоретично изследване по темата в настоящия доклад. Стана ясно, че успешното функциониране на чатботовете зависи от баланса между ефективност и социална комуникация, както и от техния дизайн и въздействие върху удовлетвореността на клиента.

Единодушно е мнението, че обработката на естествен език е ключова технология, която позволява неструктурираните текстови данни от чатбот взаимодействия да бъдат трансформирани в структурирани (със стратегически приложима информация), което води до прилагането на статистически и машиннообучаващи методи. Технологията дава възможност системите да интерпретират намеренията, емоциите и темите на клиентските съобщения, което е от съществено значение за оптимизиране на клиентското преживяване. Най-пълното и ефективно разбиране на сложни явления като клиентското преживяване се постига чрез интегрирането на количествени и качествени подходи в рамките на едно цялостно изследване, като за целта могат да се използват смесените методи.

2. Теоретична рамка

Теоретичната рамка в настоящото изследване обхваща важни и основополагащи за разработката понятия, които следва да бъдат разяснени:

- Чатботът е софтуерно приложение, базирано на изкуствен интелект, което симулира човешки разговор. Той комуникира с потребителите чрез текстови или гласови методи, използвайки обработка на естествен език, с цел разбиране намеренията на потребителя и предоставяне на автоматизирани отговори или изпълняване на задачи.
- Клиентското преживяване представлява съвкупността от всички възприятия, емоции и оценки, които един клиент формира с дадена марка или компания от първоначалния им контакт до следпродажбеното обслужване.
- Неструктурираните данни са данни, които нямат предварително дефинирана структура, което ги прави трудни за съхранение, обработка и анализ.

След като се изясниха основните понятия - чатбот, клиентско преживяване и неструктурирани данни, вече е възможно да се представи начинът, по който настоящата теоретична рамка съчетава тези елементи, структурирайки анализа в три основни, взаимосвързани области, а именно: статистически анализ на неструктурирани данни, обработка на естествен език и управление на клиентското преживяване.

Теоретичната основа за статистическия анализ на неструктурирани данни се базира на текстов анализ. Текстовият анализ е процес на извличане на скрита информация от текст, което позволява последваща статистическа обработка (регресионен анализ).

Основната цел на обработката на естествен език е да даде възможност на компютрите да разбират, интерпретират и генерират човешки език по смислен начин. NLP е много важна за анализа на неструктурирани данни. В този доклад той включва предварителна обработка, количествено представяне, анализ на настроенията и тематично моделиране.

Управлението на клиентското преживяване обхваща емоционалните, когнитивните и поведенческите реакции на клиента към взаимодействието с дадена марка. Оценява се чрез метрики като удовлетвореност (CSAT), лоялност (NPS) и усилия на клиента (CES) и как те се влияят от качеството на обслужване. От поведенческите аспекти се включва това как емоциите, несигурността и очакванията на клиента, изразени в чата, влияят на бъдещото му икономическо поведение. Много е важна и концептуализацията на успешен разговор (решен проблем, удовлетворение).

Методология за статистически анализ на чатбот данни

Методологията за статистически анализ на чатбот данните е систематизиран процес, който превръща необработените потребителски диалози в полезни заключения, необходими за подобряване на потребителското изживяване в чатбота. Достигането до целта изисква преминаването през следните етапи:

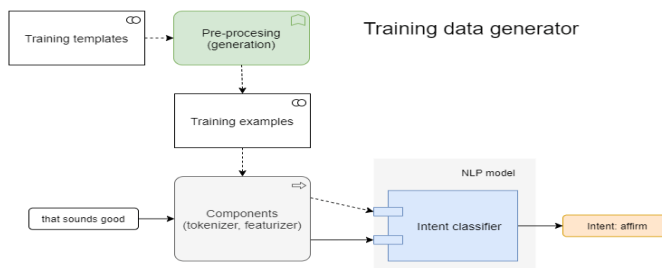
1. Етапи на подготовка на данните

- Предварителна обработка на чатбот данни

Това е първият етап при анализ на данни на разговори събирани от чатбот. Идеята е суровите чатбот разговори да се трансформират в стандартизиран масив от данни, готов за статистически анализ (напр. за изчисляване на честоти, анализ на тематиката или оценка на удовлетвореността). За целта:

- 1) Прави се токенизация. Токенизацията е разделяне на текста от диалога на по-малки смислови единици по известни като токени, т.е. целия разговор се разбива на отделни думи или фрази. Това е необходимо, защото статистическият анализ се нуждае от брой срещания на отделни думи или фрази и не може да работи с цели изречения.
- 2) Премахват се стоп думи. Стоп думите са често срещани, но неинформативни думи, които често потребителите използват в чата като: „с“, „на“, „в“, „и“ и др. Те с нищо не допринасят за смисъла на съобщението или за статистическия анализ. С тяхното премахване се предотвратява доминирането на тези думи в честотния анализ и се намалява обема на данните.
- 3) Извършва се лематизация. Това е процес на превръщане на думите в тяхната основна форма, по известна като лема. Идеята е думи с еднакъв смисъл да се броят заедно (Например говоря, говорих, говорил).

Тези етапи по предварителна обработка на данните е представен визуално с помощта на следната диаграма:



Източници: „ Intersog (2023-2024)“

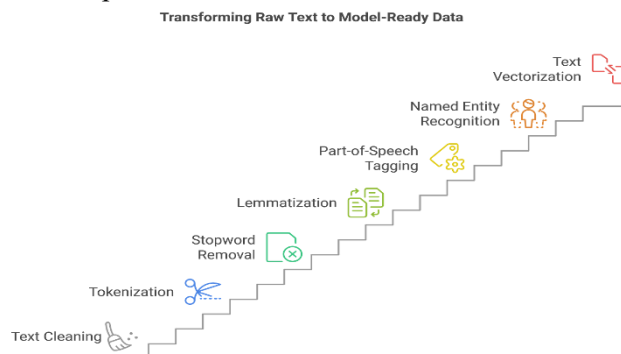
Фигура 1: „ Предварителна обработка на чатбот данни“

➤ Количествено представяне (векторизация) на чатбот данни

След предварителната обработка е необходимо думите да се превърнат в числови стойности, с цел улесняване на по-нататъшния анализ. Известни техники за това са:

- 1) Bag-of-Words изветна на български като торба с думи създава речник от всички уникални думи в данните, като брои колко пъти се среща определена дума в дадено съобщение
- 2) TF-IDF (Term Frequency-Inverse Document Frequency) на български честота на термина - обратна честота на документа освен честотата на думата в дадено съобщение, отчита и колко рядка е тази дума в целия набор от данни. С което дава по-голяма тежест на специфичните, важните думи в текст.
- 3) Word2Vec (Word to Vector) или вектор на думи улавя семантичните връзки и позволява количествени операции. Например кралица - жена, принц - мъж.
- 4) BERT Embeddings (Bidirectional Encoder Representations from Transformers), в превод двупосочни представяния на енкодера от трансформъри най-точно прихваща нюансите и контекста в сложните чатбот разговори. Една и съща дума има различен вектор в зависимост от нейното значение в конкретното изречение.

За по-ясна представа на читателя подготвителният етап е представен в обобщен вид, визуално на следващата диаграма:



Източници: „ Chaitanya Sawant, (2025),“

Фигура 2: „Етапи на подготовка на данните“

2. Описателна статистика на чатбот данни

Описателната статистика на чатбот данни е процес по събиране, обобщаване и представяне на ключови показатели от тези данни, с цел да се разбере какво се случва с потребителите. Без да представят сложни изводи и прогнози дава обобщена картина за ефективност, за поведението на потребителите и къде има нужда от подобрене.

➤ Честотен анализ на чатбот данни

Процес на изброяване, на това колко често се срещат конкретни елементи в даден набор от данни. При чатботовете честотният анализ се извършва върху почистени и токенизирани текстове, които обменят информация между потребителя и системата, както и върху метаданните (времето нужно за отговор). Идеята е да се открият често срещани думи или фрази, които да разкрият предпочитанията на потребителите, с цел да се оптимизира самия чатбот, за да се усъвършенства клиентското преживяване.

1) Честота на потребителските намерения

Това е анализ на най-често срещаните нужди на потребителя, известна метрика в статистиката като мода (най-често срещата тема). Обикновено измерва защо потребителя започва разговор, което позволява на екипа да приоритизира подобренията. Честотата на потребителски намерение се представя таблично. Таблицата е с категории и проценти на съответни разговори попадащи в отговарящата им категория.

2) Честота на ключови думи

Определя кои смислови думи се срещат най-често в потребителските чат разговори след като са премахнати стоп думите, за които се спомена по-горе в разработката. Освен това с метода TF-IDF се дава по-голяма тежест на наистина важните думи от разговора. Целта на този анализ е да идентифицират неочаквани теми или някакви пропуски в обучението на чатбот. Честотата на тези думи обикновено се представя списъчно (първите 15 думи например) или с облаци от думи.

3) Честота на пътищата на разговора

Изследва кои са най-често срещаните пътища и последователност от стъпки преди разговорът да бъде завършен или прехвърлен към човек (въпрос от клиента - отговор от потребителя). По този начин разкрива неефективни моменти или бъгове в логиката на чатбота. Резултатите се представят с помощта на Sankey диаграма или матрица. Sankey диаграмата проследява как потребителите преминават през различни категории въпроси и действия в разговорите с чатбота.

➤ Разпределение на дължината на разговора и времето за реакция

Извършва се, за да оцени ефективността на системата и качеството на потребителското изживяване. За целта в статистиката се използват централната тенденция и вариацията.

1) Разпределение на дължината на разговора

Дължината на разговора се измерва обикновено като общия брой съобщения - реплики, разменени между потребителя и робота в рамките на един чат разговор. Централната тенденция показва колко съобщения средно са необходими за приключване на един разговор, като се предпочита ниската средна стойност, защото показва бързо и директно разрешаване на проблема. Високата може да означава, че чатботът не разбира веднага въпроса. Обаче в случай, че средната стойност е много по-висока от медианната, става ясно, че има малък брой много дълги и сложни разговори, които изкривяват средното. Стандартното отклонение, от своя страна, показва колко силно се различават отделните разговори от средния брой реплики. Високото стандартно отклонение показва непоследователност, чатботът може да е много ефективен за рутинни въпроси, но много неефективен за сложни или нови теми и тази неефективност трябва да се поправи.

2) Разпределение на времето за реакция

Времето за реакция измерва закъснението между изпращането на потребителското съобщение и генерирането на отговор от чатбота. То визуализира нивото на техническа производителност и потребителско изживяване. Централната тенденция измерва времето в милисекунди или секунди, което отнема на чатбота да отговори. То трябва да бъде ниско, за да е ефективно. Чрез стандартното отклонение се разкрива нестабилност в система на чатбот. При положение, в което средното време е ниско, голямото стандартно отклонение показва, че част от потребителите получават много бавен отговор. Това може да се случи при силно натоварване.

3. Методи за анализ на съдържанието на данни от чатбота

Този тип модели имат задачата да извлекат модели, тенденции и съществено важна информация от текстовите диалози между потребителите и робота, с идеята да се подобри

клиентското преживяване. Основните методи за анализ на съдържанието на чатбот данни се основават на обработката на естествен език и следва да бъдат разгледани.

➤ Тематично моделиране на чатбот данни

Тематичното моделиране се фокусира върху намирането на латентни теми в голям набор от текстови данни от чатбот, без предварително дефинирани категории. Това помага за приоритизиране актуализациите на съдържанието и подобряване на конкретни отговори на робота.

Известни подходи за осъществяване на този метод са:

1) LDA (Latent Dirichlet Allocation) или Латентно разпределение на Дирихле

Това е алгоритъм, който приема, че всеки чат-разговор е комбинация от няколко теми, а всяка тема е комбинация от токени. И настоящият въпрос е какво всъщност прави това разпределение в чатбота? То преминава през всички думи в разговора и вкарва всяка дума към случайно избрана от него тема. След това започва да преразглежда вкараните думи, като търси по-добро разпределение, с цел да напасне така думите по темите, че да стане ясно защо се виждат точно определени думи, групирани в точно определени разговори. Например диалог в чатбот съдържаща топка, волейбол и кош е спортна, затова местенето на този разговор в раздел „спорт“ увеличава логиката на цялата система за разпределение. Идеята е така получените теми да са ясни и смислени за анализатора.

2) NMF (Non-Negative Matrix Factorization), в превод Неотрицателна матрична факторизация

Този подход разлага голямата матрица на думите на две по-малки матрици, за по-лесно интерпретиране. Получават се две матрици: едната показва важността на всяка тема в съответен разговор, а другата - важността на всяка дума в дадена тема. Важно е да се отбележи, че този алгоритъм изисква всички числа да са неотрицателни, за да може получените теми да са адитивни и лесни за разбиране.

➤ Анализ на настроенията (Sentiment Analysis)

Прави се, за да се определи емоционалната нагласа на потребителя в определен чатбот - разговор (позитивна, неутрална, негативна). Целта е бързо да се идентифицират нивата на удовлетвореност и да даде приоритет на проблемите, които роботът не е успял да разреши.

Два са подходите за sentiment анализ, а именно:

1) Лексикални подходи (Lexicon-Based)

Този метод разчита на предварително създадени речници от думи, които са разделени в две категории - позитивни и негативни. Всеки речник съдържа списък от думи и асоциации с оценки за настроение. Например „гримьор“ - 0, „опитен“ - (+2), „неопитен“ - (-2). Алгоритъмът сканира съобщението на клиента, след това събира оценките от настроенията на всяка дума и ги сумира или осреднява, с цел да изчисли общото настроение на целия разговор.

2) Подходи базирани на машинно обучение (Machine Learning)

При чатботовете тези подходи се използват за класифициране, регресия и прогнозиране. Този модел използва обучени техники (като Наивен Бейс, SVM, Deep Learning), които се прилагат за разбиране на по-сложни езикови структури. Работи като първо се обучава върху голям набор от данни от разговори, които вече са отбелязани от хора като „позитивни“, „неутрални“ и „негативни“. Той се научава не само от отделните думи, а от последователностите от думи и цялостния контекст на изречението. Така, когато моделът види ново съобщение от чатбота, прилага вече научените модели, за да предскаже с максимална точност какво е общото настроение. Важно е да се отбележи, че за разлика от лексикалните подходи, тези разбират сарказми.

4. Статистика за изводи и заключения и прогностични модели за чатбот данни

➤ Статистика за изводи и заключения

Статистиката за изводи и заключения е много важна при анализа на чатбот данни, защото позволява да се правят достоверни изводи и заключения за цялата потребителска база, основавайки се на данни от извадка (обикновено това са обикновен период, група разговори). Тя

нахвърля обобщената картина, която представя описателната статистика и дава възможност да се разбере, дали промяната в чатбота наистина води до удовлетворяване на потребителите. Необходимо е да се направи следното:

1) Тестване на ефективността

Целта е да се провери дали иновациите в дизайна на диалога, във функционалността или алгоритъма за разбиране на естествен език водят до статистическо значимо подобрене. За да се достигне до целта се прави A/B тестване, което включва хи-квадрат тест и t - тест. Чрез сравнение между старата и новата версия на чатбот група с хи-квадрат теста се взема решение, дали са успешно разрешени запитванията в чата, а с помощта на t - теста средният рейтинг на удовлетвореност и времето за разрешаване на проблема на клиента. Удовлетвореността на клиента се представя с помощта на доверителни интервали, за да се направи надеждно заключение за цялата потребителска база.

2) Оценка на обръкването

Оценява се какво въздействие има върху клиента обръкването на чатбота. Когато не разбира и прехвърля запитванията към агент. Това се прави с помощта на корелационен и регресионен анализ.

➤ Прогностични модели

Задачата на прогностичните модели за чатбот данни е да предскаже бъдещото клиентско удовлетворение или риска от неуспешен разговор, въз основа на текущи или исторически данни от диалози. Това помага на системата да вземе предварителни мерки в реално време, преди клиента да остане недоволен.

Прилагат се различни техники от машинното обучение и статистиката за предвиждане на резултата от чатбот - разговора.

1) Регресионен анализ (Regression Analysis)

Регресионния анализ помага за установяване на количествена връзка между характеристиките на разговора и степента на удовлетвореност. За анализа се използват променливи като брой съобщения преди разрешаване на проблема, време за реакция на робота, брой обръквания.

2) Анализ на тоналността (Sentiment Analysis)

Използва се за непрекъснато оценяване на емоционалното състояние на клиента като класифицира всяко съобщение като положително, неутрално, отрицателно. В случай, че се влоши тоналността, рискът от неуспешен разговор се повишава и може да се достигне до неудовлетвореност на клиента. Подходите за този тип анализ се разгледаха в подточка „Методи за анализ на съдържанието на данни от чатбота“ и няма да се обръща отново внимание върху тях.

3) Дървета на решенията и случайни гори (Decision Trees & Random Forests)

Използват се при нелинейни данни, като създават поредица от правила, с дървовидна структура, за достигането на достоверни прогнози. Например при положителен тон, време за реакция на робота от 1 секунда, брой съобщения 3 се прогнозира за висока клиентска удовлетвореност.

4) Анализ на времеви редове (Time Series Analysis)

Прилага се за прогнозиране на количествен обем от данни във времето. Анализирайки исторически времеви ред прогнозира за тенденцията на развитие, сезонност, цикличност. Така предвижда броя на разговорите, които ще пристигнат в бъдещето, което помага за оптимизиране на разпределението на ресурсите.

Приложение на анализа на чатбот данни за усъвършенстване на клиентското преживяване

1. Клиентски трудности

Тематичното моделиране и анализа на настроението имат задачата да разкрият съдържанието и емоционалния заряд на потребителското запитване, с което се идентифицират повтарящи се проблеми и области за усъвършенстване. Например ако по-голям брой от чат

разговорите съдържат тема „грешка“, „не мога да вляза“, „проблем с паролата“, с тематичното моделиране системата автоматично открива, че има проблем с достъпа и го определя като клиентска трудност. Ако пък има връщане на продукт и тематичното моделиране в комбинация със sentiment анализа отчетат висок процент отрицателен емоционален тон на потребителя, се констатира, че процесът по връщане е сложен и бавен - отново клиентска трудност.

2. Подобряване на диалоговите взаимодействия

С помощта на статистическите методи се измерва ефективността на чатбота и се идентифицират местата където той се проваля. По-високият процент на неуспех на работата да разбере намерението на потребителя, сигнализира, че модела не е достатъчно обучен да разбира естественият език. Тези провали са причина за обогатяването на тренировъчните данни на робота и добавяне на нови синоними, с цел да се увеличи точността на отговорите. Ако пък определена тема води до постоянно прехвърляне към агент, означава че логиката на робота или отговорът по тази тема не са достатъчни и е необходимо да се интегрира API (Application Programming Interface, в превод програмен интерфейс за приложения) за директно извличане на информация и прецизиране на отговора на робота.

3. Персонализирано и предсказуемо обслужване

С помощта на клъстерният анализ може да се сегментират клиентите според тяхното поведение и нужди разкрити в чат разговора. Например в групи като „Лоялни клиенти“, „Проблемни клиенти“, или пък „Информационни клиенти“. След като се идентифицира групата, роботът може да предостави персонализирани отговори още в началото на разговора. Например, ако клиентите от групата на „информационните“, често питат за промоции, робота може да изпревари нуждите им като предложи промоция преди потребителя да е попитал, с което създава положително клиентско преживяване.

Предизвикателства и бъдещи насоки при анализа на чатбот данни

1. Предизвикателства

➤ Методологични предизвикателства

Този тип предизвикателства са насочени към разбирането на естествения език и работата с големи и неструктурирани данни.

1) Обработката на сарказъм, жаргон и неясноти в контекста

На системите за анализ на настроението им е трудно да разпознават, когато положителни думи се използват в негативен смисъл. Затрудняват се и когато клиентите използват съкращения, емотикони или жаргон. За да могат да анализират такива е необходимо да бъдат обучени с реалистични данни (с включен шум). Още едно предизвикателство е разбирането на местоимения и препратки. Липсата на това разбиране води до тематични грешки.

2) Мащабируемост при обработката на големи обеми от данни

Сложните методи като дълбочинно обучение, обработка на естествен език, клъстерен анализ изискват силна разпределителна изчислителна среда, за да се приложат. Процесите трябва да са оптимизирани за скорост и автоматизирани, за да се получи анализ в реално време, който е важен за своевременна реакция при проблеми.

3) Проблем с качеството на данните

Необходими са строги процеси за предварителна обработка на данните, тъй като те са нестандартизирани и шумни (премахване на символи, дублиране, попълване на липсващи стойности).

➤ Законодателни аспекти

Анализът на чатбот данни включва лична, чувствителна и емоционално натоварена информация за клиентите, за това трябва да се спазва общият регламент за защита на данните на ЕС. Така няма да се нарушава неприкосновеността и личния живот на потребителите.

2. Бъдещи насоки

➤ Необходимо е да се интегрира генеративен изкуствен интелект (Големи езикови модели) в анализа на чатбот разговори и усъвършенстването на клиентското преживяване.

Големите езикови модели могат да извършват много по-точен и ясен анализ на настроението, както и тематично моделиране от обичайните алгоритми. Те могат да обобщават дълги разговори, които след това да бъдат статистически анализирани. Освен това имат възможности

да констатираят набор от провали на чатбота и да предложат по добри и точни контекстуални отговори за обучението на робота. Големите езикови модели умеят да идентифицират пропуски, анализирайки хиляди негативни разговора и да предложат промени в езика и правилата.

- В бъдеще чатбот данните да се комбинират със структурирани данни от други източници (история на покупките, посетени страници на уебсайта, честота на използване на приложението).

Това позволява да се направи клъстерен анализ като се идентифицират проблемите и корелациите между поведението на потребителите онлайн и нуждата от поддръжка на системата.

Заклучение

Настоящият доклад обстойно изследва теоретичните аспекти на статистическият анализ на неструктурирани данни от чатбот разговори, с идеята да се направи подобрене на клиентското преживяване и да се избегне достигането до фрустрация на потребителя.

При прегледа на литературата на изследването се разгледа темата през погледа на различни изследователи и се констатира тяхното мнение по въпроса. В теоретичната рамка се обясни значението на базовите понятия по темата (неструктурирани данни, клиентско преживяване и чатбот) и се хвърли светлина върху нейния обхват, след което се изложиха основните теоретични подходи използвани в анализа. Обърна се внимание на предварителната обработка на данните (токенизация, стоп думи, лематизация), която е важна за превръщането на неструктурирания текст във вид подходящ за анализ. Към подготовката, за превръщането на думите в числови стойности се включи и векторизацията, която обхваща методи като: Bag-of-Words, TF-IDF, BERT Embeddings и Word2Vec. Впоследствие се представиха техники за честотен анализ на чатбот данни и се хвърли светлина върху методите за анализ на съдържанието на такива данните. Както стана ясно, методите за анализ на съдържанието на чатбот данни са изключително важни, за извличането на модели, тенденции и съществено важна информация от диалозите между потребителите и робота, с идеята да се подобри клиентското преживяване. Този анализ се фокусира върху техники като тематично моделиране, благодарение на което се намират латентни теми в голям набор от текстови данни от чатбот, без предварително дефинирани категории и анализ на настроението, с който се определя емоционалната нагласа на потребителя и се идентифицират нивата на удовлетвореност. На последно място се разгледаха статистически методи за анализ на изводи и заключения и методи за извършването на прогнози. Статистическите методи за изводи и заключения са от особено значение за анализа, защото дават възможност да се дефинира краен и достоверен извод за това, дали промяната в чатбота наистина води до удовлетворяване на потребителите и въз основа на това се правят бъдещи прогнози. Накрая в една точна се обясни как всички тези техники за анализ на данни от чатбот допринасят за подобряването на клиентското преживяване.

След като се представи нужният теоретичен фундамент, докладът показва връзката между статистическият анализ на езикови данни и възможността за вземане на информирани решения, които подобряват потребителското преживяване, увеличават клиентската лоялност и подобряват процесите. От това следва, че поставената цел, да се представят теоретичните основи и методологии за статистически анализ на неструктурирани чатбот данни с фокус върху усъвършенстването на клиентското преживяване, е напълно постигната.

Основен принос на доклада към теоретичната база е, че представя цялостен теоретико-методологичен модел, който систематизирано свързва статистическите методи и обработката на естествен език с управлението на клиентското изживяване.

Източници

1. Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology, AI & Society, 35(3), pp. 535 - 547.

2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, pp. 993 - 1022.
4. Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction, *Journal of Business Research*, 117, pp. 587 - 595.
5. Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and Conducting Mixed Methods Research* (3rd ed.), SAGE publications, Chapter 1 - 4.
6. Dimitrova, V., (2025). Modeling Big Data Using Python, Challenges to the Statistical Science and the Business Analysis Profession in the Era of Big Data and Artificial Intelligence: Conference Proceedings, Sofia, pp. 30 - 40. ISBN 978-619-232-923-5.
7. Gatzioufa, P., & Saprikis, V. (2022). A literature review on users' behavioral intention toward chatbots' adoption, *Applied Computing and Informatics*, 18(1), 1 - 17.
8. Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *Proceedings of the 38th International Conference on Information Systems (ICIS 2017)*, Seoul, South Korea.
9. <https://intersog.com/blog/strategy/three-methods-of-pre-processing-data-in-chatbot-development/>
10. Jurafsky, D., & Martin, J. H. (2023). Chatbots and dialogue systems. In *Speech and Language Processing* (3rd ed.), Draft, Chapter 15, pp. 309 - 330
11. Kaczorowska-Spychalska, D. (2019). How chatbots influence marketing. *Management*, 23(1), pp. 251 - 270.
12. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization, *Nature*, 401(6755), pp. 788 - 791.
13. Lemon, K. N., & Verhoef, P. C. (2016). Understanding customer experience throughout the customer journey, *Journal of Marketing*, 80(6), pp. 69 - 96.
14. Liu, B. (2012). *Sentiment analysis and opinion mining*, Morgan & Claypool Publishers, ISBN 9781608458851
15. Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2nd ed.), Cambridge University Press, Chapter 1 - 4, pp. 1 - 113.
16. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1 - 2), pp. 1 - 135.
17. Sawant, C. (2025, January 3). From messy text to model-ready data: A guide to NLP preprocessing, Medium.
18. Schöbel, S., Schmitt, A., Benner, D., & Janson, A. (2021). Future directions for chatbot research: An interdisciplinary research agenda, *Computing*, 103(10), 2915 - 2942
19. Verhoef, P. C., et al. (2009). Customer experience creation: Determinants, dynamics and management strategies, *Journal of Retailing*, 85(1), 31 - 41.