Извличане на знания от неструктурирани данни чрез анализ на мнението на потребители

Станимира Йорданова*,
Камелия Стефанова**

Резюме: Информацията в бизнес организацията постоянно расте под формата на структурирани и неструктурирани данни. До 2018 г. повече от половината от големите организации в световен мащаб ще се конкурират, използвайки интелигентни аналитични средства за анализ на всякакви видове данни с цел извличане на ново знание. Една от тенденциите, които влияят върху бързото развитие на този пазар, е усъвършенстването на бизнес аналитичността чрез обогатяване с нови методи и алгоритми, способни да извличат и обработват данни от нови източници, предоставящи неструктурираните данни, резултат от взаимодействието с клиентите (Gartner Inc, 2016). Такива източници могат да бъдат социалните онлайн медии и потребителски сайтове, където потребители на стоки и услуги изразяват своето мнение. Използването на методи за анализ на мненията на потребителите позволява на бизнес организациите да се запознаят с мнението на своите клиенти чрез идентифициране на положителни или отрицателни мнения за техните продукти или услуги в интернет. Извличането на знания от коментари на потребители изисква структуриране на данните, след което с помощта на методи и средства на извличане на закономерности от данни и текст, данните се анализират. Получените резултати се визуализират с подходящи средства за бизнес интелигентност, за да допринесат за получаването на нови знания в процеса на подпомагане вземането на управленски решения.

Настоящата статия представя основни понятия, методологии и средства, свързани с изследването, и предлага методология за извличане на знания от неструктурирани данни чрез анализ на мнението на потребители.

Ключови думи: извличане на знания от данни и текст, анализ на мнение на потребители, бизнес интелигентност.

JEL: C63.

1. Увод

В динамично развиващата се бизнес среда правилното и навременното използване на информацията, получена в организацията и от нейните взаимоотношения, днес дава важно конкурентно предимство. Организациите повече от всяко-

* Станимира Йорданова е докторант в катедра „Информационни технологии и комуникации“ на УНСС, e-mail: smira.yordanova@gmail.com

** Камелия Стефанова е доктор, професор в катедра „Информационни технологии и комуникации“ на УНСС, e-mail: kstefanova@unwe.bg

га се нуждаят от събиране и обработване на необходимите данни, за да разбират по-добре случващите се процеси и вземат аргументирани решения. Стремейки се да отговорят на жестоката конкуренция, те непрекъснато се фокусират върху възможностите за постигане на по-висока удовлетвореност на клиентите чрез подобряване и разширяване на своите продуктови предложения и повишаване на качеството на услугите. С цел установяване на клиентските предпочитания и получаване на опит от предоставянето на продукти и услуги, бизнесът обикновено организира специални маркетингови проучвания за изучаване на мнението на клиентите. Резултатите от тях се обработват чрез различни методи за анализ. Този подход на събиране на мнението и нуждите на клиентите е доказано скъп, отнема значително време, включва участие на специализирани консултантски фирми за организиране и провеждане на проучвания и допълнително, едни от големите недостатъци са, че обхваща ограничена извадка от потребители, а резултатите са валидни за твърде кратък период от време, тъй като мнението на потребителите се променя бързо.

Бурното развиване на интернет технологиите и масираното използване на социалните онлайн медии предостави средства на потребителите на стоки и услуги да споделят без ограничения своите преживявания и опит от използването им в интернет пространството чрез публикуването на коментари, мнения, статии, оценки. Търсейки възможности как да подобри предлаганите продукти или услуги, как да разбира по-добре поведението на клиентите, както и да взема информирани управленски решения, бизнесът трябва да започне да използва съдържащата се ценна информация в публикуваните коментари. Постоянно нарастващото съдържание в различните коментари на потребителите в интернет прави невъзможно тяхното ръчно обработване. Успешното използване на информацията, съдържаща се в мнението на потребителите, може да се постигне чрез извличане, обработка и анализ на големи обеми от неструктурирани данни, получени от различни публикации в интернет. Тези данни нямат определена структура или модел и затова не могат да бъдат директно обработени и разбрани. Те първо трябва да преминат през процеси на структуриране и извличане на полезна информация, след това да бъдат анализирани с помощта на методи и средства на извличане на закономерности от данни (Data Mining), в последствие получените резултати да бъдат визуализирани с подходящи средства за бизнес интелигентност и чак тогава биха допринесли за получаването на нови знания за подпомагане вземането на управленски решения.

Настоящата статия има за цел да представи теоретичен преглед на свързаните с изследването основни понятия, методологии и средства и да предложи методология за извличане на знания от неструктурирани данни чрез анализ на мнението на потребители, съчетавайки съществуващи методи и средства за извличане на знания от текстово съдържание, извличане на закономерности от данни, машинно обучение и обработка на естествен език.

2. Извличане на знания от неструктурирани текстови данни

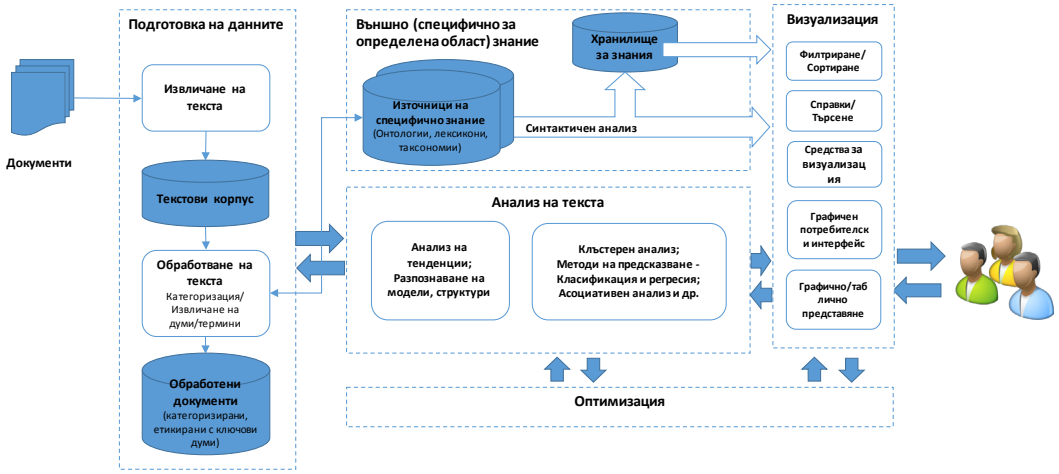
Извличането на знания и закономерности от данни (Data Mining) представлява процес на откриване на връзки, зависимости, повтарящи се модели, тенденции и аномалии в големи масиви от структурирани данни, съхранявани в складове чрез

използване на алгоритми от областта на машинното обучение, разпознаването на образи, статистиката и визуализацията на данни. Целта е да се разкрие скрита предварително неизвестна информация, полезна за бизнес потребителя от данни, които са нормализирани и съхранени в определена структура и логика. Освен от структурирани данни, ново знание може да се извлече и от неструктурирани данни, които все повече нарастват в информационната база на бизнес организацията. Неструктурираните и полуструктурираните данни са обект на анализ от процеси на извличане на знания от текст (Text Mining) под формата на документи, съдържащи текст или големи колекции от документи, наречени текстови корпус. Документи, които нямат типографски елементи като препинателни знаци, големи букви и маркери в текста, които да спомагат идентифицирането на важни в документа компоненти (параграфи, таблици и др.) често се обозначават като документи в свободен формат или неструктурирани. Документи, които имат такива елементи като съобщения от електронната поща, pdf или word документи, се приемат за полуструктурирани.

Извличането на знания от текст обхваща методи и средства за структуриране на текстова информация, извличане на модели от структурирани текстови данни, оценка на моделите, интерпретация и визуализация на получените резултати. Процесът на обработка и извличане на знание от текстово съдържание обхваща три етапа. Първият етап е събиране и организиране на текстови документи със специфична област на приложение и използване, в резултат на което се генерира корпус или колекция от документи, която ще се обработва. Тези документи съдържат текст, който не може да се обработи от алгоритмите на извличане на знания от данни и трябва да премине процес на предварителна обработка. Вторият етап обхваща техники за обработване на текстовото съдържание в корпуса с цел структурирането му в подходящ вид за анализ от алгоритмите за извличане на знания от данни. Третият етап е извличане на полезно знание от текстовото съдържание чрез използване на алгоритми за класификация, клъстеризация, асоциация и др.

Структурирането на текстовото съдържание в данни чрез идентифициране и извличане на отличителни елементи от текста, представящи съдържанието и подходящи за използване от алгоритмите на извличане на знания, е съществена особеност в текстообработката, която отличава системите за извличане на знание от текст от системите на извличане на закономерности от структурирани данни. Съвкупността от отличителни елементи на текста оформя представителен модел на документа. Разработването на ефективен представителен модел на документа е важен етап в структурирането на текста, изискващ прилагане на методи от други компютърни дисциплини, обработващи текстово съдържание на естествен език. Постигане на правилното съотношение между обем и семантично ниво на елементите в модела, които най-точно да представят текстовото съдържание в съчетание с идентифициране на тези елементи, които ще бъдат подходящи като най-ефективни за разкриване на нови модели и тенденции, е предизвикателство при проектирането на системи, обработващи текстово съдържание.

Архитектурата на системите за извличане на знания от текст на функционално ниво (фигура 1) обхваща следните задачи: (1) Подготовка на текста; (2) Анализ на текста; (3) Визуализация на резултатите и (4) Оптимизация.



Фигура 1. *Архитектура на система за извличане на знания от текст (адаптирана от Feldman R., Sanger J. P 2007)*

Подготовката на текста обхваща извличане на документи от източници и прилагане на методи за обработване на текстовото съдържание. Извлечените документи оформят корпус (колекция) от документи. След това се прилагат методи за обработване на текстовото съдържание в корпуса с цел структуриране в подходящ вид за прилагане на алгоритми за извличане на знания от данни. Сърцето на всяка система за извличане на знание от текст е аналитичната част, в която може да се използват различни методи като анализ на тенденции, разпознаване на модели в данни и други методи, най-вече, използвани в извличането на закономерности от данни. Анализът на тенденциите използва датиране на документите в корпуса, така че да могат да се правят различни сравнения между подмножества в корпуса, отнасящи се до определен времеви период. Методите за анализ на модели целят да открият наличие на връзки между понятия в корпуса от документи. Методите за извличане на закономерности от данни, които се използват и при извличане на знания от текст, са: клъстерен анализ, който най-често се използва за създаване на таксономии от понятия в дадена област; класификация като метод за предсказване на стойности на променлива на базата на известни стойности на останалите (независими) променливи; асоциативен анализ – за откриване на интересни връзки между променливи в големи бази данни. Подготовката на текста и анализът са най-критичните задачи за проектиране и изпълнение, без които системата за извличане на знание от текст не може да съществува.

Някои системи за извличане на знание от текст обработват и анализират документи, които принадлежат към определена област/домейн като например биология, финансови услуги, международно право и други. Тези области включват специфично знание, характерно само за конкретната област, организирано под формата на онтологии, речници с понятия или таксономии. Това знание се използва в процесите на подготовка на текста, анализ и извличане на знание от текста и визуализация на резултатите в системата, като често се съхранява в Хранилище за знания

(Knowledge Base) и е достъпно за различни елементи от системата. Визуализацията цели да представи на потребителите на системата за извличане на знания от текст резултатите във вид, подходящ за решаването на различни проблеми, където навременното и правилно им представяне и обобщаване подпомага процесите на вземане на правилни и бързи решения.

Източниците, от които се извличат знания от текстово съдържание, са разнообразни. От гледна точка на бизнес организацията могат да се разделят на вътрешни и външни. Вътрешните източници съдържат информация, която се създава в резултат на дейността на организацията (доклади, отчети, имейли и др.), а външните източници се създават извън нея, но съдържат важна и значима информация за дейността ѝ (социални медии, мобилни данни, текстови съобщения, информация за местоположение, съдържание на уеб сайт, форуми, блогове, коментари на потребители, сайтове и други инструменти и технологии за публикуване на съдържание в интернет). Коментарите на потребители, публикувани в интернет, в които те споделят своя опит от използването на стоки и услуги, са в неструктуриран вид от външен източник с текстово съдържание. Извличане на полезно знание за организацията от коментара може да стане чрез прилагането на анализа на мнението на потребителите.

3. Извличане и анализ на мнения от потребителски коментари

В англоезичната научна литература понятията „sentiment analysis“ и „opinion mining“ най-често се използват като синоними. За първи път и двете понятия се срещат през 2003 г., но за начало на научните изследвания в тази област се смятат 2000 г. и 2001 г. като изследването на отделни задачи, влизащи в обхвата на понятието, е предмет на научна дейност и преди 2000 г. В днешно време, изследователската дейност в тази научна област се развива с бързи темпове. Основни причини за това са бързото развитие на методите, прилагани в машинното обучение (Machine Learning), в обработването на естествения език (Natural Language Processing) и в извличането на информация (Information Retrieval); нарастването на данните в интернет, които се използват за изграждане и тестване на модели, като резултат от бурното развитие на социалните мрежи, блогове, сайтове за коментари на стоки и услуги и др. и не на последно място – реализирането и комерсиализацията на интелигентни приложения за анализ на мнението на потребители.

Понятието се дефинира както от учените, работещи в областта, така и от разработчиците на софтуерни решения за анализ на мнението на потребителите. Например W. Kurtz (2013) дава дефиниция на sentiment analysis: прилагане на обработка на естествения език, компютърна лингвистика и анализиране на текст с цел идентифициране и извличане на субективна информация от източници. Kumar и Sebastian (2012) приемат sentiment analysis като „автоматизиран анализ на субективността, подобен на извличане на мнение, който се фокусира върху извличане и класифициране на текстове с машинен език и компютърно програмиране“. За Attensity (2014), доставчик на решения в областта на sentiment analysis, това е „процес на определяне на тона или отношението в един коментар“, като акцентира върху бизнес значението му: за всеки анализатор на социалните медии, потребителските мнения, емоции, нагласи и желания водят до реално приложими прозрения, които могат директно да се реализират в нови продукти, в маркетинг стратегии или в търговията и отношенията с клиентите.

Общото при всички определения е, че sentiment analysis прилага средствата за обработка на текст и естествен език с цел идентифициране и извличане на субективна информация от текстови материали. Основна задача на анализа е да класифицира даден текст или елемент от текста като положителен, отрицателен или неутрален в зависимост от изразеното чувство от страна на субекта към обекта.

В българската литература трудно се намира превод на двете понятия. Разграничение между тях може да се направи с оглед дефинициите на чувство (sentiment) и мнение (opinion). Според тълковен онлайн речник чувство е „психическо състояние, предизвикано от външни въздействия; душевно преживяване, емоция“, а мнение е „възглед, схващане, отношение по някакъв въпрос, преценка, оценка“. В този смисъл sentiment analysis може се преведе като анализ на чувства, а opinion mining – като извличане на мнения. Извличането на мнения може да се приеме като по-обширното понятие, което включва анализа на чувства и може да се преведе като извличане и анализ на мнения, с които се изразяват или предполагат позитивни или негативни чувства. В настоящата статия се използва ***анализ на мнения на потребители.***

4. Същност на процеса, видове методи и средства за анализ на мнения на потребители

Основната задача в процеса на извличане и анализ на мнения може да се дефинира по следния начин:

В едно множество D от текстови документи, които съдържат мнения или чувства на потребители за обекти, да се извлекат всички атрибути и компоненти на обекта във всеки документ d ∈ D и да се определи дали коментарите са позитивни, негативни или неутрални.

Обект на мнението може да бъде продукт, услуга, човек, събитие, организация или тема. Всеки обект има характеристики, наречени атрибути или компоненти, които се коментират и за които се изразява мнение. Полярността на мнението (емоционалният заряд) е ориентацията, която показва дали мнението е положително, отрицателно или неутрално. Повечето научни разработки са фокусирани върху бинарната класификация на позитивен и негативен клас. Автор на мнението е човекът, който е изразил своите чувства в коментара. Извличането на обектите на мнението и класифицирането на мнението чрез определяне на полярността на изразеното мнение са едни от основните задачи в анализа на мнения.

Процесът на извличане и анализ на мнения като част от обхвата на извличане на знания от текстово съдържание в най-общ вид обхваща три етапа:

1. Предварителна обработка на текста
2. Класификация на текста и
3. Обобщаване и представяне на резултатите.

Коментарите, които потребителите публикуват в интернет, са в текстови вид и не могат директно да се обработват с техниките и методите за извличане на закономерности от данни. Преди да се приложат тези методи, е необходимо текстовият документ да се подготви чрез обработване на синтактичната и семантичната структура на текстовите документи и представяне на документа като съвкупност от думи. В предварителната обработка на текста се използват различни техники от извличането на знания от текст (TM) и от лингвистична обработка на текста (NLP) с цел да се обогати наличната информация за думите, като изборът им зависи от конкретната задача и поставената цел. Краткото описание на тези техники е представено на фигура 2.

Техника от ТМ	Описание
Разделяне на текста на графични думи и пунктуационни знаци (Tokenisation)	Разделяне на текста на съставните му думи и пунктуационни знаци.
Лематизация (Lemmatisation)	Групиране на различните форми на една дума в лема.
Стоп думи или „шумови“ думи (Stopwords)	Премахване на думи, които не носят важно значение (цифри, знаци, местоимения, предлози и др.)
Стеминг (Stemming)	Съкращаване на думата до нейния корен, което намалява размера на речника от думи в документа и подобрява резултатите, особено при по-малки множества от данни.
Филтриране на думи (Filtering Tokens)	Филтриране на думите по различни критерии (напр. дължина на думата).
Трансформиране на думи (Transforming Tokens)	Трансформиране на думите (напр. превръщане на главните букви в малки).
Техника от NLP	**Описание**
Автоматичен морфологичен анализ (Part-of-Speech Tagging)	Маркиране на частите на речта в изречението – съществителни, глаголи, наречия, съюзи и т.н.
Автоматичен синтактичен анализ (Parsing)	Генерира дърво на връзките между думите в изречението, синтактично дърво.
Частичен автоматичен синтактичен анализ (Text Chunking)	Разделяне на текста на синтактично свързани думи. Предшества генерирането на синтактичното дърво.
Многозначност (Word Sense Disambiguation)	Техники за отстраняване на многозначността на думите или фразите.

Фигура 2. *Техники за предварителна обработка на текста*

Класификация на текста може да се извърши на три нива: класификация на целия документ, на ниво изречение, като определи полярността на всяко изречение и на ниво конкретен аспект (характеристика) на обекта.

Съществуват различни предизвикателства и трудности, които оказват влияние върху процеса на класификация на коментарите. На първо място, съществуват два типа мнения: обикновени мнения и сравнителни мнения. Обикновените мнения изразяват чувства по отношение на определен един обект. За разлика от тях, сравнителните мнения изразяват чувства по отношение на няколко обекта, като ги сравняват. Как да се идентифицират такива изречения е основна трудност в анализа на мнението в сравнителни изречения. Анализът на сравнителните мнения не е изучаван задълбочено в научната литература.

Друг проблем е съставяне на речници, които се използват в класификацията на мнения. Най-важните индикатори на чувства са думите и фразите, изразяващи позитивни или негативни заряди (например прилагателни, наречия и т.н). Съвкупността от тези думи и фрази оформят речник или лексикон, който се използва в анализа на мнения. Съставянето на речниците има своите предизвикателства:

- в различните области на използване една дума може да носи различен заряд;
- изречение, в което има дума, изразяваща чувство, да не изразява мнение;
- саркастични изречения и изречения, които не съдържат дума, изразяваща мнение, но изразяват мнение.

От една страна, предизвикателство пред учените в тази област е да проектират алгоритми, които да компилират такива лексикони, от друга страна, тези реч-

ници от думи са зависими от предметната област, в която се прилагат.

Други научни предизвикателства, които съществуват пред анализа на мнения и са свързани с обработката на естествения език, са прилагане на процесите на намиране на всички изрази, които се отнасят до един и същ обект в текста (co-reference resolution); справяне с думите в изречението, които променят заряда (negation); и многозначност на думите (word sense disambiguation).

В процеса на класификация на мнението се използват различни подходи, които могат да се разделят на машинно обучение (МО), подход, базиран на лексикон/речник и хибридни подходи. МО обхваща *машинно обучение с учител* (supervised machine learning) и *машинно обучение без учител* (unsupervised machine learning). МО с учител използва две множества от данни – обучителни данни и тестови данни. Обучителните данни включват входни данни и очаквани резултати и служат за „учител“ по отношение на тестовите данни. Алгоритъмът на МО изгражда модел за предсказване на нови данни, използвайки обучаващите данни и тествайки модела върху тестовите данни. При МО без учител алгоритъмът трябва да открие моделите и структурите в данните, без да разполага с обучително множество. Някои от методите, които най-често се използват в машинното обучение с учител за извличане на ключови думи и фрази (атрибути), където съставянето на множество от ефективни атрибути оказва влияние върху класификацията на мнения, са представени на фигура 3.

Техники за избор на атрибути	**Описание**
Двоично представяне на термините в текста (Binary Presence)	Наличие на термин (дума или фраза), с двоично представена стойност на атрибутите, в който записите показват дали един термин се появява (заема стойност 1) или не (заема стойност 0)
Абсолютна честота на термин (Term Occurrences)	Колко пъти терминът се появява в съответния документ nw като абсолютна стойност.
Относителна честота на термин (Term Frequency, TF)	Относителна честота на термин в документ като отношение между колко пъти думата се появява в съответния документ nw (където w е думата) и общия брой на думите в документа, n: $TF=\frac{n_w}{n}$
TF–IDF претегляне (Term Frequency–Inverse Document Frequency, TF–IDF)	Техника за калкулиране на теглото на думите или фразите в текста. $TF–IDF=\frac{n_w}{n} * \log_2 \frac{N}{N_w}$, където N е общият брой на документите, а Nw е броят на документите, които съдържат думата w.
Автоматичен морфологичен анализ (Part-of-Speech Tagging)	Прилагането на автоматичен морфологичен анализ обикновено е свързано с наблюдението, че най-често характеристиките на продуктите и услугите се изразяват със съществителни думи и фрази, а прилагателните и наречията са носители на емоционален заряд и като такива участват в определянето на полярността. Използвайки тази техника, всяка дума в текста се идентифицира като част на речта и се извличат само съществителни (в анализа на ниво аспект) и/или прилагателни и наречия, в зависимост от поставената задача.

Фигура 3. *Методи за извличане на ключови думи*

Под понятието Term (термин) се разбира една дума или фраза, състояща се от няколко думи, които са извлечени директно от текстовия корпус в определена предметна област чрез средства и методи на обработване на естествения език. TF-IDF стойностите се използват за създаване на векторно представяне на документите. Всеки компонент на вектора съответства на TF-IDF стойността на определена дума в корпуса. Термини, които не се срещат в даден документ, приемат стойност нула. Този вид представяне на текста се нарича векторен пространствен модел, който се използва в процеса на извличане на информация и класификацията. Всяко измерение на пространството съответства на термин от речника на текстовия корпус.

В прилагането на МО с учител се използват различни видове класификатори: дърво на решенията, невронни мрежи, метод на поддържащи вектори, класификатори, базирани на правила, Наивен Бейсов класификатор и максимална ентропия. В научната литература методът на поддържащи вектори, Наивният Бейсов класификатор и максималната ентропия са най-използвани и сравнявани методи за класификация. Според проучване на литературата в областта на анализа на мнения, направено от Tang, Tan, Cheng, 2009, методът на поддържащи вектори и Наивният Бейсов класификатор са най-добрите модели за класификация на ниво документ в един домейн. Дървото на решенията се използва в твърде малко научни разработки, основно за класификация на филмови и продуктови коментари.

При други подходи се използват лексикони/речници от негативни и позитивни думи в процеса на класификация. Процесът на събиране на думи, които оформят речника започва с ръчно събиране на малък набор от думи, съдържащи положителен и отрицателен заряд, който се увеличава чрез търсене на техните синоними и антоними в големи речници като WordNet®. Основен недостатък на този подход е невъзможността да се намерят думи, характерни за определен домейн или контекст. Използването на думи от текстовия корпус решава този проблем, защото се съставя списък с думи, изразяващи чувство, като се търсят други думи в текстовия корпус със специфичен контекст. Хибридните подходи са най-прилагани в научните разработки, като съчетават машинното обучение с учител и без учител и използването на лексикони/речници за класификацията на текста. Най-често се прилагат в анализа на мнението на ниво аспект.

Анализът на мнението на потребителите е специфична задача от извличането на знания от текст, в която се използват техники и методи за извличане на закономерности от данни и обработване на естествения език. Затова повечето средства за анализ на текст и данни предоставят възможности и за анализ на мнение на потребителите. Те се делят основно на комерсиални и безплатни или с безплатна версия. Повечето комерсиални продукти са целево предназначени за анализ на текстово съдържание (като Attensity, NetOwl, Lexalytics Text Analytics и др.) и предлагат различни методи за обработка на текстово съдържание и анализ на мнения или са част от серия аналитични продукти (като IBM Text Analytics, SAS Text Analytics и др.). Някои от продуктите предоставят функционалности в облак (като AlchemiAPI, MeaningCloud, DiscoverText). Сред софтуерите с отворен код са GATE: General Architecture for Text Engineering, Text Mining Infrastructure in R, NLTK: Natural Language Toolkit, RapidMiner, WEKA, OpenNLP, KNIME Text Processing, STANFORD CORENLP.

5. Методология за извличане на знания от неструктурирана информация чрез анализ на мнението на потребители

В изследването се предлага методология за извличане на знания от неструктури-

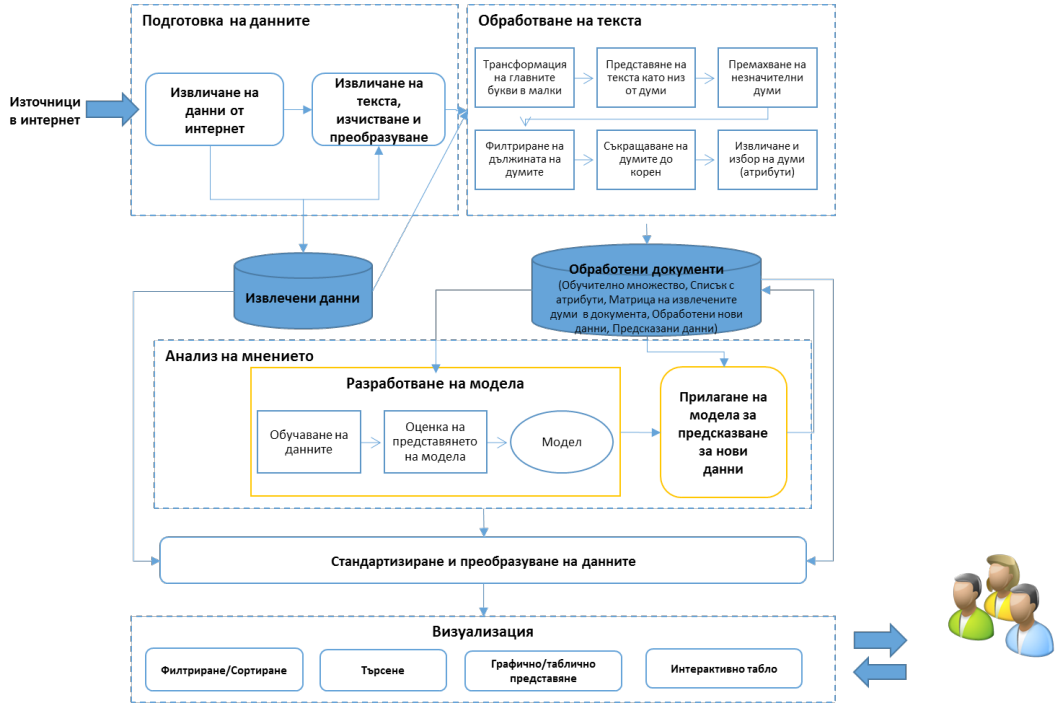
рана информация чрез анализ на мнението на потребители, съчетавайки съществуващи методи и техники в текстообработката и класификацията на мнения. Тя обхваща процеси, методи и средства за извличане на данни и текстово съдържание от интернет (съдържащо мнение на потребители), подготвяне на текстовото съдържание чрез обработване и структуриране, изграждане на модел за предсказване на мнението чрез машинно обучение с учител, оценяване на представянето на модела, прилагане на модела за нови данни и визуализация на извлечените данни от интернет и предсказаните резултати чрез средствата на Бизнес интелигентността.

тези знания могат да подпомогнат решаването на бизнес проблеми в организацията.

Методологията за извличане на знания от неструктурирана информация чрез анализ на мнението на потребители е представена на фигура 4 и обхваща четири етапа:

(1) Подготовка на данните;
(2) Обработване на текстовите данни;
(3) Анализ на мнението на потребителите – разработване и прилагане на модела за предсказване;
(4) Визуализация.

Подготовката на данните обхваща извличането на данните от източници и извличане на текстовото съдържание, необходимо за изграждане на модела и за пред-



Фигура 4. *Методология за извличане на знания от неструктурирана информация чрез анализ на мнението на потребителите*

Основната задача на изследването е да се провери какви знания могат да се извлекат от мнението на потребителите и как

сказване на мнението. Данните се извличат от интернет сайтове чрез използване на import.oi и се съхраняват в excel формат.

Това е уеб-базирана платформа за извличане на данни от интернет без кодиране чрез създаване на модел, на основата на който се извлича необходимата информация от посочени страници. Данните се съхраняват в облак и могат да бъдат изтеглени като CSV, Excel, Google таблици или JSON или споделени. След това данните се изчистват от грешки, повторения, преобразуват се за целите на следващия етап и се съхраняват.

При предварителното обработване на текстовото съдържание текстът се структурира чрез последователно приложение на няколко техники. Първо, чрез алгоритмите на трансформация, главните букви се превръщат в малки. Разделянето на текста на графични думи и пунктуационни знаци е задължителна стъпка в процеса на обработката и представя текста като последователност от думи, в резултат на което се създава списък с думи. С цел да се постигнат по-добри резултати при класификацията, се премахват думи без особено значение, филтрират се думи с определена дължина и се съкращават думи до корен. Извличането на думите, наречени още атрибути, които се използват в машинното обучение, се извършва чрез изчисляването на тяхната честота. С изследователска цел ще се тестват следните техники: двоичното представяне на атрибутите в документите, колко пъти една дума се среща в даден документ и TF–IDF претегляне. В процеса на трансформация се генерира Матрица на извлечените думи от текста (Term Document Matrix, TDM). Това е математическа матрица, която описва честотата на думите, които се намират в корпуса от документи. Тази матрица се разпознава от всеки алгоритъм на машинно обучение. В нея редовете съответстват на документите, а колоните са думите от речника на корпуса, а стойностите в клетките отговарят на калкулираната честота на съответната дума. Списъкът с атрибутите, матрица на извлечените думи от текста и обучителното множество от текстови данни се съхранява и участва в разработването на модела за предсказване на мнението на потребителите.

Третият етап обхваща процес на анализ на мнението на потребителите на ниво документ, като задачата е да се определи дали целият текст е положителен или отрицателен. Това е бинарна класификация, при която се прилага алгоритъм на машинно обучение с учител, като се изгражда **модел за предсказване на мнението на потребители** чрез обучение на данните, тестване на модела за класификация и оценка на неговото представяне.

Обучителният процес обхваща две фази: обучение и тестване, при които се използват две множества от данни – обучаващо множество и тестово множество. За оценката на представянето на модела се използва кръс проверка, която се повтаря 10 пъти (10-fold Validation). Данните се разделят на 10 части/извадки с еднакъв брой примери. От десетте извадки, една се използва само в тестовата фаза и не участва в обучението. Останалите девет се използват в обучението като обучително множество. Процесът на валидация се повтаря 10 пъти, като всеки път в тестовия процес участва различна извадка от цялото множество данни. Методът на формиране на извадката е стратифициран с цел изграждане на случайни подмножества, за да гарантира, че разпределението по клас в подмножествата е същото, както в цялото множество. Изследвания в областта на оценката на моделите са показали, че най-добра оценка на грешката се получава, когато валидацията се повтори десет пъти. Стратифицираното кръс валидиране, повторено 10 пъти, се е превърнало в стандартна техника за оценка на модели при малки обеми от данни.

В процеса на класификация се извършва разделяне на множеството текстови обекти на два класа – положителен и отрицате-

лен, като за метод на класификация се използва дървото на решенията. Дървото на решенията (ДР) е структура, която включва корен, клони и листа. Всеки вътрешен възел обозначава тестване на атрибут, всеки клон е резултат от теста и всяко листо притежава етикет, обозначаващ клас. Най-горният възел в дървото е коренът. Най-важно при изграждането на дървото е изборът на атрибут, по който да стане разделяне на множеството. За построяването на дърво на всеки вътрешен възел е необходимо да се намери такова условие, което би разделило множеството, асоциирано с този възел на подмножества. За такава проверка се избира един от атрибутите. Общото правило за избор на атрибут е избраният атрибут да разделя множеството така, че получените подмножества да се състоят от обекти, принадлежащи към един клас или количеството обекти от други класове да са максимално малко. Критерии за избор на атрибут, който да разделя множеството на подмножества е ентропията, Information Gain, Gain Ratio.

Ентропията е понятие от информационната теория и мярка, характеризираща количеството информация. При множество S, съдържащо позитивни и негативни примери, Ентропията на S е:

$$H(S) = -p_+\log_2 p_+ - p_-\log_2 p_- \quad (1)$$

p_+ – пропорцията на позитивните примери в S

p_- – пропорцията на негативните примери в S

Ентропията е 0, когато всички членове на S принадлежат към един клас. Ако всички са позитивни ($p_+ = 1$), ентропията е 0 ($-1.\log_2 1 - 0.\log_2 0 = -1.0 - 0.0 = 0$). Ентропията е 1, когато S съдържа равен брой примери, които принадлежат към двата класа. Ентропията е между 0 и 1, когато S съдържа неравен брой примери, принадлежащи към двата класа.

Information gain (IG) се измерва чрез очакваното намаляване на ентропията, причинено от разделянето на примерите според определен атрибут и се изчислява по следната формула:

$$Gain\ (S,A) = H(S) - \sum_{v \in Values\ (A)} \frac{|S_v|}{|S|} H(S_v) \quad (2)$$

където:

Gain (S, A) – очакваното намаляване на ентропията, причинено от това, че стойността на атрибут A е известна.

S е цялото множество от примери, описано чрез атрибути, един от които е атрибут A.

S_v е подмножество на S, за което атрибут A заема стойност v ($S_v = \{s \in S | A(s) = v\}$)

Values(A) са всички възможни стойности на атрибут A.

$\frac{S_v}{S}$ е отношението между броя на примерите в подмножеството S_v, за което атрибут A заема стойност v и общия брой на примерите в множеството S.

H(S) е ентропията на множеството S, която се изчислява по формула (1).

$H(S_v)$ е ентропията на подмножеството S_v, за което атрибут A заема стойност v. Изчислява се чрез формула (1) за подмножество S_v.

Най-добрият атрибут за разделяне е този с най-висок Gain.

Gain ratio (GR) е модификация на IG, при която се вземат предвид броят и размерът на клоните на дървото чрез разделяне (4) на IG на стойността на потенциалната информация (3), генерирана от разделянето на обучаващото множество S на подмножества S_v, използвайки атрибут A (Intrinsic Information, IntI):

$$IntI\ (S,A) = -\sum_v \frac{|S_v|}{|S|} \log\left(\frac{|S_v|}{|S|}\right) \quad (3)$$

$$GR\ (S,A) = \frac{Gain\ (S,A)}{IntI\ (S,A)} \quad (4)$$

Алгоритъмът на ДР разраства дървото толкова дълбоко, колкото е нужно, за да се класифицират примерите от обучаващо-

то множество перфектно. Това всъщност може да доведе до преспециализация на дървото (overfitting), или пълно съответствие на модела с обучаващите данни (Кабакчиева, 2012), когато в данните съществуват шумове или когато броят на обучаващите примери е твърде малък, за да произведе представителна извадка на целевата променлива. Преодоляването на преспециализацията на дървото се постига чрез спиране на разрастването на дървото или спиране на обучението (Tree Pruning). Подрязаните дървета са по-малки и по-малко сложни. Подходите при подрязването на дървото са два: предварително подрязване (Pre-pruning) и последващо подрязване (Post-pruning). При първия подход дървото се подрязва предварително, като се спира неговото разрастване чрез използването на правила за спиране (например прагови стойности за брой листа, брой стъпки в процеса на разделяне, степен на чистота на възлите и др.). При втория подход дървото се подрязва след като е изградено напълно, като се премахват поддървета, съгласно избрани критерии.

Оценката на точността на модела се прави чрез използването на кръс проверка. При кръс проверката точността на класификацията на тестовото множество се сравнява с точността на класификация на обучаващото множество. В резултат на кръс проверката се генерира матрица, представена на фигура 5.

В матрицата True Positive (TP) са правилно класифицирани позитивни примери, а True Negative (TN) са правилно класифицирани негативни примери. False Positive (FP) са негативните коментари, неправилно предсказани като позитивни, а False Negative (FN) са позитивните коментари, неправилно предсказани като негативни. Изчисляването на стойностите на показателите за оценка са представени на фигура 6.

		Class = pos ***true pos***	**Class = neg** ***true neg***
Предсказан клас	Class = pos *predicted pos*	TP	FP
	Class = neg *predicted neg*	FN	TN

Фигура 5. *Матрица при кръс проверка*

Стойност	**Калкулации**
Correct predictions	TP + TN
Incorrect predictions	FP + FN
Total scored cases	TP + FP + TN + FN
Error rate	(FP + FN)/(TP + FP + TN + FN)
Accuracy rate	(TP + TN)/(TP + FP + TN + FN)
Precision rate	TP/(TP + FP)
Recall rate	TP/(TP + FN)
Pos Class Recall	TP/(TP + FN)
Neg Class Recall	TN/(TN + FP)
Pos Class Precision	TP/(TP + FP)
Neg Class Precision	TN/(TN + FN)

Фигура 6. *Изчисляване на показателите за оценка*

При крос проверката се изчисляват точност (Accuracy), прецизност (Precision), (Recall). Точността (Accuracy) е процентът на правилно класифицираните примери от множеството. Прецизността (Precision) е процентът на правилно класифицираните примери от съответния клас – позитивен и негативен. Recall показва от всички позитивни коментари колко са правилно класифицираните примери като позитивни и колко са реално позитивните коментари, неправилно предсказани като негативни.

Целта при изграждането на модела за предсказване на мнението на потребителите в интернет е използваният алгоритъм за класификация на коментарите да постигне точност на класификацията над 70%, като се експериментира с параметрите на основния модел с цел постигане на по-добра класификация на негативните коментари, като се запазва точността на класификация на позитивните коментари.

Изграденият модел за предсказване на мнението на потребителите се прилага за нови данни с цел предсказването им като позитивни и негативни. Новите данни се извличат от интернет и преминават през процес на обработка с цел структуриране на текстовото съдържание.

RapidMiner Studio се използва за реализацията на модела и експериментирането с цел постигане на по-добри резултати, тъй като този софтуерен продукт е безплатен и предоставя възможности за прилагане на методи за обучение с учител, методи за валидиране на резултатите и чрез Text Processing Extension могат да се използват необходимите техники и методи за обработка на текст.

В последния етап на Визуализация се използват Бизнес интелигентни средства за представяне на извлечените данни от интернет и резултатите от анализите на крайните бизнес потребители. Основната цел при визуализацията е информацията да подпомогне потребителите при решаване на бизнес проблеми и организациите в процеса на проследяване на развитието и оценяване на ефективността на предоставените услуги и за коригиране на техните стратегии и планове. Преди да се проектира визуализацията, е необходимо данните да бъдат стандартизирани и съхранени. Чрез Бизнес интелигентните средства се проектират и реализират необходимите справки и интерактивни бизнес табла. Qlik Sense Desktop се използва за проектирането и реализирането на Бизнес интелигентно приложение, представящо извлечените данни от сайтовете за коментари и резултатите от анализа на мнението на потребителите в интернет.

6. Заключение

Предложената методология се използва за прилагане на анализа на мнението на потребителите с цел извличане на ново знание от коментарите на потребителите на стоки и услуги в полза на бизнеса и за решаване на бизнес проблеми в организациите. Експериментално методологията се реализира в областта на предоставянето на хотелски услуги, като текстовите данни са коментарите, публикувани от потребителите (гости) на хотели в интернет. Източници, от които се извличат данните, необходими за изграждане и прилагане на модела за предсказване на мнението на потребителите и за изграждане на Бизнес интелигентно приложение, са едни от най-големите сайтове за предоставяне на хотелски услуги – Booking.com, TripAdvisor.com и Hotel.com. Като част от научната работа се разработва модел за класифициране и предсказване на коментарите на потребителите в интернет, използващ дървото на решенията. Моделът ще се изследва експериментално, като се променят параметрите на алгоритъма на класификация с цел постигане на точност над 70%. Моделът ще се приложи по отношение на

нови неизвестни данни с цел предсказване на мнението на потребителите на хотелски услуги. Резултатите от анализа ще се използват в процеса на проектиране и разработване на Бизнес интелигентно приложение с цел ясна и полезна визуализация на информацията от коментарите и извличане на ново знание за бизнес потребителите, което би подпомогнало решаването на бизнес проблеми в организацията.

Цитирани източници:

Кабакчиева, Д., 2012. Технологии за извличане на знания, книга, Управление на знания, Авторски колектив, гл. 14, Софийски университет, ISBN 978-954-18-0839-9.

(Kabakchieva, D., 2012. Tehnologii za izvlichane na znania, Upravlenie na znania, Sofiyski Universitet, gl. 14, ISBN 978-954-18-0839-9)

Тълковен онлайн речник, 2016, [онлайн], наличен на http://talkoven.onlinerechnik.com/, [последно влизане 13.06.2016].

(Talkoven rechnik, 2016, [onlayn], nalichen na http://talkoven.onlinerechnik.com/, [Posledno vlizane 13.06.2016])

Bing Liu, 2012. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers

Bing Liu, Lei Zhang, 2012. A survey of opinion mining and sentiment analysis, Mining Text Data.

Feldman R., J. Sanger, 2007. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.

Huifeng Tang, Songbo Tan, Xueqi Cheng, 2009. A survey on sentiment detection of reviews, Expert System with Applications 36 10760-10773.

Import.io, https://www.import.io/, [Accessed 10.06.2016])

Kumar, Akshi, and Sebastian, Teja, Mary, 2012. Sentiment analysis. A perspective on its past present and future. *International Journal of Intelligent Systems and Applications*, 4 (10): 1-14.

Kurtz W., 2013. Sentiment, Sentiment Analysis and Sentiment Intelligence, Available at: https://waynekurtz.wordpress.com/6bi-home/sentiment-sentiment-analysis-and-sentiment-intelligence/, [Accessed 10.06.2016])

L. Kart, G. Herschel, A. Linden, J. Hare, 2016. Magic Quadrant for Advanced Analytics Platforms, Gartner, Inc., 09 February 2016, Available at: https://www.gartner.com/doc/reprints?id=1-2WQY2ZJ&ct=160121&st=sb, [Accessed 10.06.2016])

Pang B, Lee L., S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

Princeton University „About WordNet.“ WordNet. Princeton University. 2010, http://wordnet.princeton.edu

Qlik Sense Desktop 2.2, http://www.qlik.com/products/qlik-sense, [Accessed 10.06.2016])

RapidMiner Studio 7.1, https://rapidminer.com/, [Accessed 10.06.2016])

Salton G., Wong A., Yang C. S., A vector space model for automatic indexing, Communications of the ACM, Vol. 18 No. 11.

Shinde P., Govilkar S., 2015. A Systematic study of Text Mining Techniques, *Internationalm Journal on Natural Language Computing* (IJNLC) Vol. 4, No.4.

The Artistry Behind a Reliable Sentiments Engine, 2014, Attensity White paper.