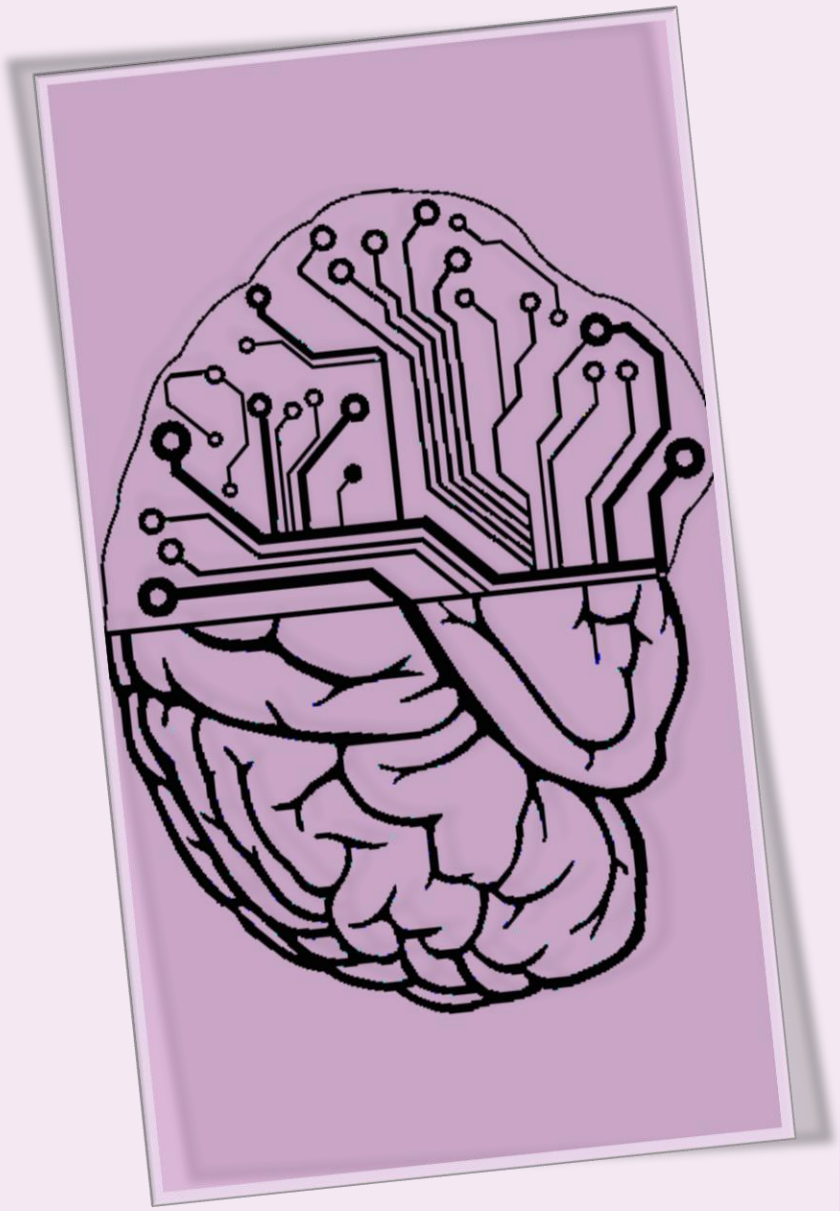
СТОПАНСКИ ФАКУЛТЕТ

Катедра "Статистика и иконометрия"

Специалност "Моделиране на големи данни в бизнеса и финансите"



МАГИСТЪРСКА ТЕЗА

Дизайн и разработка на автоматизирана система за определяне на потребителското отношение към софтуерни приложения чрез техники за обработка на естествен език и машинно самообучение

Автор: Глория Христова

Научен ръководител: Доц. Д-р Боряна Богданова

Март 2018

София

Резюме

Анализът на настроението е актуална тема, приковаваща интереса в много сфери на нашето общество, поради безбройните ѝ приложения и ползи. Бизнесът, както и социалният и политически свят, все повече осъзнават нуждата си и се възползват от подобен род изследвания, като това неминуемо довежда до засилен интерес към тази научна област и в академичният свят. Настоящата разработка си поставя за цел, именно създаването на автоматизирана система за разпознаване на настроението в текст. Обект на нашето изследване е потребителското отношение, изразено в отзиви за мобилни приложения. За да разрешим основната задача в нашата работа, ние използваме и изследваме представянето на основни алгоритми на машинното самообучение. Прилагаме техники от обработката на естествен език и извличаме различни описващи текста променливи. Резултатите ни показват, че така създадената система за разпознаване на настроението, се характеризира с голяма предвиждаща способност (по време на валидацията на системата е постигната точност от 85%). Един от приносите на настоящата работа е предложената методология за обработка и привеждане на текстови данни, част от домейна на мобилните приложения, в подходящ за количествен анализ вид. Методологията е авторско съчетание на различни техники от обработката на естествен език и е приложима и в други изследвания, в които се борави с данни в разглеждания домейн. Наред с това, нашите изводи и резултати могат да послужат за сравнителен анализ на представянето на основни алгоритми на машинното самообучение, използвани за разпознаване на емоцията в потребителски отзиви - както в домейна на мобилните приложения, така и като цяло в сферата.

Ключови думи:

Наука за данните, Извличане на знания от текст, Обработка на естествен език, Анализ на настроението и емоцията, Машинно самообучение, Извличане на знания от данни, Класификация, Потребителски коментари, Софтуерни приложения, Google Play Store, Python

Съдържание

Списък на включените фигури

Списък на включените таблици

Речник на термините

Термин (бълг.)	Термин (англ.)
F1-мярка	F1-measure
n-gram модел	n-gram model
Tf-idf репрезентация	Term frequency-inverse document frequency
Анализ на настроения и чувства	Sentiment analysis
Априорната вероятност	Prior probability
Аспектно-базиран синтез на мнението	Aspect-based opinion mining
Балансирана извадка	Balanced dataset
Бернулиев Наивен Бейсов Модел	Bernoulli Naïve Bayes
Биграми/ Две последователни думи	Bigrams
Бизнес разузнаване	Business intelligence
Бинарна класификация	Binary classification
Булев модел	Boolean model
Бъг/Програмна грешка	Bug
Валидационна извадка	Validation set
Векторизация	Vectorization
Векторно-пространствен модел	Vector-space model
Вероятностна стойност	P-value
Вероятностни модели за моделиране на език	Probabilistic language models
Големи данни	Big data
Дискурсивен анализ	Discourse analysis
Зависима променлива/ Етикет/ Целева променлива	Target variable
Извличане на знания от данни	Data mining
Извличане на знания от текст	Text mining
Извличане на знания от уеб източници	Web mining
Търсене на информация	Search and Information retrieval
Извличане на информация	Information extraction
Извличане на концепции	Concept extraction
Изкуствен интелект	Artificial intelligence
Изходен програмен код	Source code
Интегрирана среда за разработка	Integrated development environment, IDE
Интернет на нещата	Internet of Things
Информационната наука	Information Science
Класификация на документи	Document classification
Класификация с повече групи	Multiclass classification
Клъстеризация на документи	Document clustering
Компютърна лингвистика	Computational Linguistics
Лапласово изглаждане на данните	Laplace smoothing
Лексикон на настроението от емотикони	Emoticon sentiment lexicon
Лексикон с думи, носещи дадено настроение	Sentiment lexicon
Лематизация	Lemmatization
Линейна класификация чрез опорни вектори	Linear Support Vector Classifier
Логистична регресия	Logistic Regression
Макро F1-мярка	Macro F-measure
Максимална ентропия	Maximum entropy
Матрица на грешките	Confusion matrix/error matrix
Машинно самообучение	Machine Learning
Машинното самообучение без учител	Unsupervised learning

Машинното самообучение с учител	Supervised learning
Метод на k най-близки съседи	k-nearest neighbours algorithm
Метод на опорните вектори	Support vector machines
Метода на случайната гора	Random forest
Методи базирани на лексикон	Lexicon-based approach
Минал опит	Past Experience
Моделиране на теми в текст	Topic modeling
Морфологичен разбор на изречението	Part-of-speech tagging (POS tagging)
Мярка за взаимна информация	Mutual information
Наивен Бейсов Модел	Naïve Bayes
Намаляване на размерността на данните	Dimensionality reduction
Настройване на статистически модели	Model tuning
Наука за данните	Data Science
Нормализация на буквите в текста	Case normalization
Облаци от думи	Word clouds
Обработка на естествен език	Natural Language Processing
Обясняващи/Независими променливи	Explanatory variables/ Features
Отдалечено наблюдение	Distant supervision
Отрицание	Negated expressions
Подход базиращ се на корпус от думи	Corpus-based approach
Подход базиращ се на речници	Dictionary-based approach
Постериорната вероятност	Posterior probability
Потребителски изисквания към приложение	Feature requests
Предвиждащи алгоритми	Predictive analytics
Прекомерно нагаждане към извадката	Overfitting
Препоръчващи системи	Recommender systems
Прецизност	Precision
Регуляризация на логистичната регресия	Regularized logistic regression
Релационни структури от данни	Relational databases
Ръчното поставяне на етикет	Manual labeling
Синонимни емотикони	Emoticon synsets
Система за извличане на данни от интернет	Web crawler
Стеминг	Stemming
Стоп думи	Stop words
Субективни фрази	Subjectivity phrases
Суров вид на данните	Raw data
Тестова извадка	Test set
Торба с думи	Bag of words
Токъни/ Низ от думи	Tokens
Токъницазия	Tokenization
Точност	Accuracy
Тренировъчна извадка	Training set
Търсене в решетка от предефинирани стойности	Grid search
Униграми/ Единични думи	Unigrams
Унифицирано разпределение на класа	Uniform class distribution
Хаштаг	Hashtag
Хибридни методи	Hybrid approach
Хи-квадрат тест	Chi-square test
Хиперпараметър	Hypeparameter
Чувствителност	Recall

Увод

Науката за данните (*Data Science*) завладя 21-ви век. Бизнесът започна да осъзнава все повече нуждата си от аналитични инструменти, с които да разбира по-добре, както вътрешните си процеси, така и външната среда, която го обгражда – клиенти, конкуренти, потенциални партньори и още много други аспекти от този необятен свят. Междувременно с възможностите, които предоставя интернет пространството, масовото използване на смарт устройства, разрастването на социалните мрежи, подема на интернет на нещата (*Internet of Things- IoT*) и като цяло забележителното развитие на технологичния свят, обемът на данните с които разполагаме се увеличава многократно с всеки изминал ден. Така постепенно възникна и понятието "големи данни" (*Big data*), за което всички днес говорим и се интересуваме, независимо от това дали сме част от академичния, политическия или пък света на бизнеса.

В глобален план огромна част от тези големи данни е в неструктуриран вид[1]. Това означава, че тях не можем да открием в добре познатите ни релационни структури от данни (*relational databases*), които сме свикнали да обработваме лесно. Съществена част от тази неструктурирана информация произлиза от дигиталният свят и приема основно формата на изображения, видео и текст (журнали, списания, книги, документи, имейли и всякакъв друг възможен вид кореспонденция). Това пък от своя страна дава ход на прогреса в централни сфери, част от науката за данните и изкуственият интелект (*artificial intelligence*). Техники в областта на извличането на знания от данни (*data mining*), машинното самообучение (*machine learning*), извличането на знания от текст (*text mining*), обработката на естествен език (*natural language processing*) ни предоставят разнообразни начини за опознаване, обработка, моделиране, откриване на нови зависимости и създаване на прогнози. Те ни помагат в анализирането не само на данни, част от релационни структури, но могат да бъдат прилагани и върху неструктурирана информация, произлизаща от разнообразни източници.

Анализът на текст и обработката на естествен език са научни области известни ни още през 20-ти век, но едва в началото на 21-ви век започват да бъдат по-широко

[1]Това твърдение е подкрепено от американския гигант в областта на информационните технологии IBM (International Business Machines Corporation) и други компании публикуващи свои изследвания и прогнози относно развитието на дигиталния и аналитичния свят, като Gartner, Merrill Lynch и други.

употребявани и получават значително по-голяма гласност. Причина за това са както развитието на технологиите и компютърната мощ, позволяващи извършването на много по-сложни математически изчисления, така и купищата източници на информация под формата на текст в интернет пространството, която в много случаи е напълно достъпна за всеки. Интересът към тези две научни области продължава да се засилва и към днешна дата.

Анализът на текст е обширно научно поле, в което се преплитат различни сфери – компютърна лингвистика (*Computational Linguistics*), извличане на знания от данни, машинно самообучение, бази от данни, статистика, изкуствен интелект и библиотечно-информационни науки (Miner, 2012). Настоящото изследване засяга анализа на текст, използвайки основно методи на извличането на знания от данни и машинното самообучение, като разбира се донякъде има досег и с останалата част от изброените научни сфери, тъй като те притежават множество допирни точки и често се застъпват.

Така по естествен път стигаме и до основната тема в настоящата работа, а именно анализът на настроения и чувства (*sentiment analysis*). Този вид анализ се занимава с извличане на субективна информация от текст (чувства, мнения и емоционални състояния на човека, изразени в текста) с помощта на методи за обработката на естествен език и извличането на знания от данни (Miner, 2012).

През август 2017г. списание Forbes публикува статия, озаглавена “Анализ на настроението: от критична важност в бизнеса за подобряване на клиентското преживяване”.[2] В нея Pradeep Govindasamy, главен технически директор на Cigniti[3], подчертава ползите и дори необходимостта на всеки един бизнес от използването на такъв вид анализ, ако иска да бъде успешен и да подобри своите решения в сферата на управлението на взаимоотношенията с клиенти (*Customer Relationship Management*). Според Pradeep Govindasamy, в днешно време потребителите са станали силно емоционални, като обичат да споделят своите опит, преживявания и взаимодействие с компаниите и техните стоки и услуги. Това става най-често посредством различни платформи като Twitter, Facebook, сайтове за потребителски отзиви, блогове и други.

[2]Govindasamy, Pradeep. Sentiment Analysis: A Business-Critical Need To Improve Customer Experience // Forbes. Jersey City. 03.08.2017. [cited 03.08.2018] Available from: https://www.forbes.com/sites/forbestechcouncil/2017/08/03/sentiment-analysis-a-business-critical-need-to-improve-customer-experience/#251c7845352b

[3]Cigniti is Global Leader in Independent Software Testing Services // Cigniti. Irving TX. 2004. [cited 03.01.2018] Available from: http://www.cigniti.com/

Същата статия споменава и резултатите от глобално изследване на Nielsen[4] през 2015 г., според което 83% от хората се доверяват напълно на препоръки от семейство и приятели и 66% се доверяват на потребителски мнения, публикувани в онлайн пространството. Pradeep Govindasamy споменава и важната роля на мобилното приложение, в случаите когато компанията притежава такова. Позовавайки се на още проучвания на потребителското поведение, той подчертава факта, че хората внимателно прочитат отзивите преди да свалят и използват дадено приложение. Последното означава, че бизнеса трябва да полага постоянни грижи за това да осигури отлично потребителско изживяване на своите клиенти и да поддържа висок рейтинга на своето мобилно приложение. Така анализът на настроението се превръща в основен инструмент за подобряване на стоки, услуги, фирмена репутация и път към превръщането в пазарен лидер.

Анализът на чувства и настроения в текст е изключително повлиян от сферата, в която се прилага. Изразите и думите, които хората употребяват са различни в зависимост от това какъв е предмета (например, стоката или услугата), за която те изразяват своето мнение. Дадени думи биха могли да значат нещо положително в даден контекст и точно обратното в друг. Това е показано и в редица изследвания на изявени автори в тази област - (Liu, 2012), (Pang & Lee, 2008) и други.

В повечето анализи на чувства и настроения в потребителски мнения са изследвани основно отзиви за филми (в сайтове като Rotten Tomatoes и IMDb), за потребителски стоки (например, отзиви в Amazon) и хотели и ресторанти (в сайтове като TripAdvisor и Yelp) - (Pang, Lee, & Vaithyanathan, 2002, July), (Narayanan, Arora, & Bhatia, 2013, October), (Fang & Zhan, 2015), (Gezici, Yanıkoğlu, Tapucu, & Saygın, 2012). По-малко изследвания са посветени на потребителски отзиви за мобилни приложения, въпреки наличието на източници на такъв тип данни. Разработките, изследващи съдържанието на отзиви за софтуерни приложения в голяма част от случаите са фокусирани основно не върху анализа на настроението на потребителя и езиковите средства, с които той го изразява в тази сфера, ами върху извличането на информация (например, конкретни оплаквания, проблеми или изисквания на потребителите) от отзивите, агрегирането ѝ, откриване на основните дискутирани теми в отзивите и други - (Fu, et al., 2013, August), (Panichella, et

[4]Global trust in advertising // Nielsen. Mumbai. 28.09.2015. [cited 03.01.2018] Available from: http://www.nielsen.com/in/en/insights/reports/2015/global-trust-in-advertising-2015.html

al., 2015, September). На база на направеният литературен преглед в настоящото изследване, можем да кажем, че значително по-малко са авторите, които се задълбочават главно в изучаването на извличането на настроението от текст в сферата на мобилните приложения, както и свързаните с тази сфера специфики в езика и изразяването на потребителя – (Guzman & Maalej, 2014, August), (Hoon, Vasa, Martino, Schneider, & Mouzakis, 2013, November). Това би могло да се окачествти като известен пропуск, тъй като такъв вид анализ носи своите позитиви и в тази област не по-малко отколкото в горепосочените (в глава I - „Литературен преглед“ - е представен по-детайлен обзор на някои от споменатите разработки).

Един от тези позитиви, който трябва да бъде споменат е, че разработването на система за разпознаване на чувства и настроения в потребителски коментари за софтуерни приложения, би спомогнала разработчиците им да бъдат постоянно в течение с това, дали тяхното приложение се харесва от потребителите. Подобна система би могла да бъде изключително полезна за анализ на отзиви, публикувани и в други канали за комуникация с потребителя - например, сайтове без рейтингова система[5], която да помага при разграничаването на публикуваните положителни и отрицателни мнения. Това могат да бъдат вътрешни системи за комуникация с потребителя, социални мрежи, специализирани сайтове, анкети и други. Цялата тази неструктурирана информация е важна за разработчиците на приложения. Следователно, обединението ѝ и способността да бъде анализирана от единна система за разпознаване на настроението на потребителя биха улеснили тяхната работа. Прост пример, който можем да дадем, е как чрез подобна система софтуерните разработчици могат да следят в реално време отношението на техните потребители след пускането на най-актуалната версия на мобилното приложение, което разработват. Martin et al. (Martin, Sarro, Jia, Zhang, & Harman, 2017) подчертават, че отзивите и рейтинга на дадено мобилно приложение са силно обвързани с неговият брой сваляния и продажби.

Извличането на информация за настроението на потребителя има добавена стойност не само поради горепосочените позитиви. Това знание е свързано и би могло в бъдеще да бъде използвано като основа за други важни аналитични задачи, които биха били интересни за бизнеса в софтуерната индустрия (а и като цяло), като например:

[5]Под “рейтингова система” се имат предвид сайтовете, в които освен че потребителя може да остави коментар, също така може да приложи и оценка за даденият продукт/услуга в добавка към своя коментар.

- Извличане на препоръки на потребителите, чести проблеми (бъгове), въпроси по използването на приложението, искания за различни функционалности и други – цялото това знание може да бъде използвано за усъвършенстване на мобилното приложение и удовлетворяване на нуждите на потребителя;
- Създаване на по-задълбочен анализ, който да разпознава не само емоцията на даденият потребител, но и да разпознава към кой аспект на мобилното приложение е насочена тази емоция – този анализ спомага за дълбоко опознаване на клиента, неговите изисквания, силните и слаби страни на даденото мобилно приложение и т.н;
- Изследване на влиянието на потребителските мнения върху продажбите;
- Анализ на задържането и удовлетвореността на клиентите и други.

Всичко казано до момента ни насочва към идеята за централното място на анализа на чувства и настроения на потребителя и основополагащата му роля в множество други аналитични задачи, чието разрешаване би спомогнало значително за вземането на бизнес решения. В глава „Литературен преглед“ е направен обзор на по-важните изследвания в тази обширна сфера, от които става ясно, че това е така.

Предмет на настоящото изследване е емоцията на потребителя. Тъй като “емоция” е обширно понятие, следва да поясним, че става въпрос конкретно за отношението на потребителя към даденият продукт – дали е положително или отрицателно настроен към него. Това означава, че настоящата работа засяга потребителските емоции конкретно в тяхната полярност. Важността на това, даден бизнес да е наясно с това как се чувстват потребителите на неговите продукти е безспорна, тъй като това е от първостепенно значение за бъдещото му развитие.

Liang et al. (Liang, Li, Yang, & Wang, 2015) демонстрират, че информацията, която потребителите разпространяват онлайн от уста на уста (*online word of mouth - eWOM*) има значим ефект върху продажбите на мобилни приложения, като изследват в дълбочина качествените характеристики на тази информация, използвайки потребителски отзиви от iOS App Store[6]. Има още много изследвания, които изучават и доказват значимият ефект

[6]iOS App Store представлява платформа за цифрова дистрибуция на мобилни приложения на базата на iOS, разработена и поддържана от компанията Apple.

на потребителските коментари в интернет пространството върху продажбите на даден продукт - (Chevalier & Mayzlin, 2006), (Duan, Gu, & Whinston, 2008) и други.

Обект на настоящото изследване са отзиви на потребители за софтуерни приложения в Google Play[7]. Считаме, че това е извадка, чиито характеристики са подходящи за изпълнението на целта на тезата и ще обхване спецификите на изказа, които хората използват в тази сфера, за да изразят своето мнение.

По данни от проучвания на пазара на мобилни приложения, Google Play е на първо място по брой приложения, които потребителите могат да свалят и използват на своите мобилни устройства (по данни от март 2017 - 2,8 млн.)[8]. Това прави Google Play най-водещата фигура на този пазар, следвана от основният си конкурент – iOS App Store, чийто потребители по същото време са можели да избират между 2,2 млн. приложения.

Така определените предмет и обект на настоящата теза ни водят към дефинирането на нейната основна цел, а именно:

Създаването на автоматизирана система за разпознаване на емоцията на потребителя, изразена в отзиви за мобилни приложения.

В така поставената цел, с понятието "емоция" означаваме полярността на потребителското настроение, така както вече беше дефинирано и в предмета на настоящото изследване.

Извличането на тази информация чрез аналитични техники би довело до общо подобряване на обслужването на клиентите и съответно - развитие на бизнеса в положителна посока. Подобна система би спомогнала неимоверно за автоматизацията на този процес по опознаване на клиента – би било непосилно за разработчиците на приложения да прочитат хилядите отзиви на потребители, които е възможно да идват от различни канали за комуникация с клиента, както беше споменато по-рано. Трябва да се отбележи, че се появяват и компании, чиято основна дейност се състои именно в осъществяването на този вид анализ – Appbot[9] е пример за такава фирма. Нейната основна цел е подпомагането на работата на екипи от софтуерни разработчици на мобилни приложения чрез предоставянето на текуща информация относно емоцията,

[7]Google play представлява официален онлайн магазин за мобилни приложения, поддържани от операционната система Android.
[8]Number of apps available in leading app stores as of March 2017 // Statista. New York. 03.2017. [cited 03.01.2018] Available from: https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/
[9]We know how important app reviews are to the success of your app // Appbot. 2011. [cited 03.01.2018] Available from: https://appbot.co/about

преживяванията и затрудненията, които споделят потребителите им в своите отзиви. Фирмата е създадена от своя основател именно поради нуждата му от система, която да анализира хилядите отзиви за неговото приложение, осъзнавайки че това ще доведе до създаването на още по-добър продукт.

Така формулираната цел ще постигнем чрез решаването на следните **изследователски задачи**:

1. Преглед на литературата в сферата на:

1.1. анализа на чувства и настроения и конкретно емоциите, които са предмет на настоящото изследване

1.2. обработката на естествен език и извличането на знания от текст.

2. Избор на методология за превръщане на текста в подходящ за количествен анализ вид чрез техники за обработката на естествен език.

3. Формиране на методология с цел създаването на автоматизирана система за разпознаване на емоцията в отзиви за мобилни приложения чрез методи на машинното самообучение. Тази задача се състои от решаването на следните по-малки задачи:

3.1. Дефиниране на зависима променлива с цел превръщане на проблема в такъв на машинното самообучение с учител (*Supervised machine learning*).

3.2. Дефиниране на множество от обясняващи променливи от текста – използване на думите в текста, но и тестване на други променливи, извлечени от него – емотикони, специални препинателни знаци и други характеристики, описващи текста и носещи информация за емоцията, която носи той.

3.3. Избор на техника за справяне с размерността на данните и подбор на обясняващите променливи.

3.4. Тестване на най-често използваните алгоритми на машинното самообучение за анализа на текст и емоцията в него – обучение на модели, използване на техники за настройването им с цел подобряване на представянето.

3.5. Избор на най-добре представящия се алгоритъм според валидационните резултати и избрани статистики за съпоставка на моделите.

4. Отчетлива интерпретация на получените резултати, сравнение с утвърдени изследвания в сферата, отчитане на възможности за подобрение на създадената система и бъдещи перспективи.

Въз основа на всичко казано до момента и така дефинираните предмет, обект и цел на настоящото изследване, можем да формулираме тезата, залегнала в неговата основа по следният начин:

Система за разпознаване на полярността на емоцията, изразена в потребителски отзиви за мобилни приложения, би могла да бъде успешно създадена чрез подходящи методи на машинното самообучение, традиционно използвани в подобни изследвания, и комбинация от добре подбрани обясняващи променливи.

Така формулираната теза можем да развием в следните две хипотези, които ще бъдат изследвани емпирично в рамките на настоящата работа:

- Добавянето на обясняващи променливи, извлечени от текста и различни от думите в него (например, използване на емотикони), ще подобри общото представяне на системата в сравнение с представянето ѝ, когато използваме само думите в него.
- Най-добро общо представяне на системата очакваме да получим използвайки класификация чрез опорни вектори (*Support Vector Classifier*).

Основен принос на настоящото изследване е в създаването на методология за обработка и привеждане на текстови данни, част от домейна на мобилните приложения, в подходящ за количествен анализ вид. Наред с това, резултатите от изследването могат да бъдат използвани с цел сравнителен анализ на представянето на основни алгоритми на машинното самообучение, използвани за разпознаване на емоцията в потребителски отзиви.

Останалата част от настоящата работа е организирана, както следва. Глава I предоставя литературен обзор на изследователската дейност в обширното поле на анализа на настроението, като ще обърнем специално внимание и на изследвания конкретно в домейна на мобилните приложения. Глава II представя детайлно разработената методология в настоящата работа за създаване на система за разпознаване на настроението. В Глава III тази методология ще приложим върху обекта на настоящата разработка – отзиви за мобилни приложения. Там ще се запознаем с характеристиките на данните и ще тълкуваме подробно резултатите, получени в емпиричното изследване. В заключение ще направим обзор на всичко постигнато в настоящият труд, като ще очертаем детайлно и неговите приноси.

I. Литературен преглед

В настоящата глава ще си отговорим на въпроса какво всъщност представлява анализа на текст, ще открием къде е мястото на анализа на настроението в това обширно научно поле, като предоставим и формални дефиниции на тези понятия. Ще се запознаем с важни задачи, които засяга анализа на настроението, както и с основните методи, които помагат за разрешаването им. Ще направим литературен обзор на значими разработки в сферата и ще проследим къде намира приложение този вид анализ. Ще мотивираме и уточним обекта и предмета на настоящото изследване, както и хипотезите, залегнали в него. Така по естествен начин ще позиционираме нашата разработка в рамките на съществуващата литература в областта на анализа на настроението в текст.

1.1. Анализ на текст

Както вече споменахме в увода на настоящото изследване, дигитализацията и подема на технологиите, промяната в начините на общуване между хората и бума на социалните мрежи доведоха до възникването на огромни количества от неструктурирани данни, които могат да съдържат ценна информация, засягаща безброй области в нашият свят. Тъй като голяма част от тази информация можем да открием под формата на текст, науките занимаващи се с анализа на текст и обработката на естествен език започват да привличат все повече интерес.

Науката за извличане на знания от текст се занимава с анализирането на писмен език, което означава, че тя е насочена към разбирането и опознаването на най-сложната аналитична система, позната ни досега – човешкият мозък. Това определя науката като комплексна и разностранна. Следва дефиниция на тази наука, предоставена от (Miner, 2012):

Дефиниция № 1: Науката за извличане на знания от текст се занимава с редица технологии за анализ и обработка на полуструктурирани и неструктурирани текстови данни. Всички тези технологии биват обединени от необходимостта от "превръщането на текста в числа" (конвертиране на текста в структуриран цифров формат), така че мощни алгоритми да могат да бъдат приложени върху големи бази от документи.

Модерните системи за анализ на текст възникват основно след огромният подем на технологичният свят. Разбира се, в основата на това развитие стои именно нуждата от това да можем лесно да достъпваме, обработваме и анализираме нарастващото количество на текстова информация, която се генерира в световен мащаб. Добре е да споменем, че науката за извличане на знания от текст води началото си от библиотечната наука (*Library science*) и нуждата от това да можем да класифицираме, обединяваме в групи текстовите документи (например, книги в библиотека) и да ги обобщаваме. По-нататък, възникването на информационната наука (*Information Science*) позволява електронното съхраняване на текст и предизвиква революция в начина, по който биват свързвани авторите и ползвателите на дадена текстова информация. С течение на времето, започва съсредоточаване на изследователския интерес върху извличането на информация и знание от текст чрез методи за обработка на естествен език. Тази практическа нужда стимулира опитите на изследователи в сферата на изкуствения интелект за разработването на интелигентна машина, която да симулира сравнително точно човешкия мозък, представляващ най-сложната мисловна машина. Така тези три научни дисциплини и тяхното развитие стоят в основата на модерната наука за извличане на знания от текст. Miner (Miner, 2012) предоставя доста изчерпателен преглед на това историческо развитие на сферата, проследява отделните му фази и накрая предоставя и поглед към бъдещето, прогнозирайки, че то се крие в "дълбокото" откриване на смисъла в текст (*deep discovery*) – обучение на машината на "четене между редовете", идентифициране на идиоми, сарказъм и инсинуация, изразени в текст.

В увода на настоящата работа споменахме, че в науката за анализ на текст се преплитат различни сфери. Miner (Miner, 2012) внася яснота в това обширно научно поле, като изброява основните седем практически сфери на анализа на текст. Често типична задача за анализ на текст ще включва техники от няколко от тези сфери, тъй като те се застъпват в голяма степен. Тези седем сфери са (Miner, 2012): търсене и извличане на информация (*Search and information retrieval*), клъстеризация на документи (*Document clustering*), класификация на документи (*Document classification*), извличане на знания от уеб източници (*Web mining*), извличане на информация (*Information extraction*), обработка на естествен език и извличане на концепции (*Concept extraction*). Няма да се спираме на подробни описания на всяка една от тези сфери, а ще насочим вниманието си към анализа

на настроението, върху който основно е фокусирана настоящата работа и ще обособим неговото място в тези преплитащи се обширни научни полета.

Анализът на настроението намира своето място в пресечната област между класификацията на документи, извличане на концепции от текст и обработката на естествен език. Това е така, тъй като задачите за разпознаване на настроението включват: идентифицирането и групирането на думи, имащи подобно семантично значение (с това се занимава основно сферата на извличането на концепции от текст); класифицирането на документа в дадена група, носеща определено настроение (например, изразяващ положително или отрицателно настроение на автора); обработка на текста чрез различни техники, така че човешкият език да бъде "разбран" от компютъра, така както бива разбиран от човека.

1.2. Анализ на настроението в текст

Анализът на настроението е част от голямото научно поле на анализа на текст, както вече стана ясно в предната точка. Той е залегнал и в основата на настоящата работа, чиято цел е именно създаването на автоматизирана система за разпознаване на настроението, изразено в текст. Още в уводната част стана ясно какво представлява този вид анализ, но тук ще приложим неговата формална дефиниция, предоставена от някои от най-видните изследователи в сферата - (Pang & Lee, 2008), (Liu, 2012):

Дефиниция № 2: Анализът на настроението е научна област, занимаваща се с анализ на мнението, настроението и субективизма, изразени в текст, с помощта на изчислителната мощ на компютъра. 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.

Може би предшественик на този вид изследвания е това на Stone et al. (Stone, Dunphy, & Smith, 1966), което засяга количествените методи за моделиране и анализ на текст и отделно от това провежда изследване на психологическото състояние на хора на база на анализ на вербалното им поведение. След 2001 г. започва все по-усилен изследователски интерес към това научно поле, започвайки с разработките на Pang et al. (Pang, Lee, & Vaithyanathan, 2002, July) и Turney (Turney, 2002, July).

В контекста на анализа на чувства и настроения най-често решаваме задача за определяне на полярността на даден текст. Следва да дефинираме какво всъщност представлява тази "полярност" на настроението в текста (Pang & Lee, Opinion mining and sentiment analysis, 2008):

Дефиниция № 3: Задачата за определяне на това, дали даден текст (изразяващ мнение) носи положителен или отрицателен емоционален заряд, наричаме задача за определяне на полярността на текста.

Пионерските изследвания в сферата на анализа на настроението разрешават именно такава задача - (Pang, Lee, & Vaithyanathan, 2002, July), (Turney, 2002, July), (Dave, Lawrence, & Pennock, 2003, May). Понякога в тази задача може да бъде включен и "неутрален клас", който обикновено означава "липса на мнение" (Pang & Lee, 2008).

Ясно е, че човешките емоционални състояния се характеризират с доста по-богата палитра. Това означава, че могат да бъдат използвани много повече измерения за окачествяване на настроението в даден текст, като например "разочарование", "гняв", "радост", "тъга", "въодушевление" и други (Cambria, Das, Bandyopadhyay, & Feraco, 2017). Пример за такова изследване е това на Mohammad (Mohammad, 2012, June), където бива направен опит за създаването на корпус от постове в социалната мрежа Туитър, всеки от които бива разпределен в група, носеща определена емоция – гняв, отвращение, страх, радост, тъга и изненада. За целта авторите използват т.нар. "хаштагове" (*hashtags*) във всеки туит, които подсказват емоцията, изразена от автора (например, хаштаг "#щастлив"). Трябва да се има предвид, че подобен вид анализ на емоционалното състояние на автора на даден текст е по-сложен за изпълнение най-малкото поради донякъде субективният характер на задачата.

Всичко казано до момента е пряко свързано с предмета на настоящата работа, така както вече беше дефиниран в увода ѝ. Предмет в тази разработка е човешката емоция, изразена в текст, конкретно в нейната полярност. От *Дефиниция № 3* следва, че под "полярност" се има предвид конкретно дали текста изразява положителната или отрицателна емоция на автора.

Избираме този вариант, основно защото разпознаването на по-задълбочени емоционални и психически състояния на човека ще изисква по-сложни техники за

обработка на естествен език, ще има субективен характер и по-трудно ще се оцени представянето на създадената система. Във връзка с последното е много вероятно да се наложи и по-значителна човешка намеса, което по-скоро бихме искали да избегнем в настоящата работа (Cambria, Das, Bandyopadhyay, & Feraco, 2017).

1.3. Методи за анализ на настроението в текст

Начините за разрешаване на задачата за разпознаване на настроението в текст можем да разделим общо на три основни вида (Medhat, Hassan, & Korashy, 2014) – методи на машинното самообучение (*machine learning approaches*), методи базирани на лексикон (*lexicon-based approach*) и хибридни методи (*hybrid approach*). Първият вид използва различни предвиждащи алгоритми. Вторият използва лексикон с подбрани и установени думи, означаващи конкретно настроение (*sentiment lexicon*) – може да бъде разделен на две категории – базиращ се на речници (*dictionary-based approach*) и базиращ се на корпус от думи (*corpus-based approach*). Последните две използват статистически и семантични методи за разпознаване на полярността на настроението. Хибридните методи комбинират изброените. В своят обзор на редица актуални изследвания в сферата на анализа на настроението, Medhat et al. (Medhat, Hassan, & Korashy, 2014) показват, че броят научни изследвания, използващи метода на машинното самообучение или метода базиран на лексикон се мени през годините. Изводът на авторите е, че най-често от изследователите в областта са били използвани лексикони за разпознаване на настроението поради причината, че генерализират по-добре и могат да бъдат използвани за редица проблеми в този вид анализ. Според тях, методите на машинното самообучение за анализ на настроението продължават да бъдат отворено поле за научни изследвания. Те носят своите позитиви поради относителната простота в сравнение с другите методи и факта, че използват тренировъчни данни (*training data*), което ги прави много по-адаптивни към домейна на текста, който бива анализиран (следващата точка в тази глава хвърля светлина върху огромното значение на домейна в анализа на настроението). Използването на семантичната ориентация посредством речници може да обобщава добре, но за сметка на това методите на машинното самообучение имат доста по-голяма точност твърди Ravi в своят преглед на сферата (Ravi & Ravi, 2015).

В настоящата работа ще бъдат използвани методи на машинното самообучение за разрешаване на задачата за разпознаване на настроението. Очакваме по този начин да бъде построена система, улавяща напълно спецификите на изказа използван в домейна, който е избран (отзиви за мобилни приложения). Това предполага и постигането на по-голяма точност в сравнение с тази, която бихме получили с методи, базирани на лексикони.

Следва да добавим и малко повече детайли в тази точка относно основните методи на машинно самообучение и как биват използвани за разпознаване на настроението.

В обширната сфера на науката за данните, когато говорим за машинно самообучение, имаме предвид дял от компютърните науки, който се състои основно от създаването на ефективни и точни предвиждащи алгоритми (*predictive analytics*). Прилагаме и формална дефиниция, предоставена от Mohri et al. (Mohri, Rostamizadeh, & Talwalkar, 2012):

Дефиниция №4: Машинното самообучение представлява съвкупност от изчислителни методи, използващи минал опит с цел създаване на по-точни прогнози. Под "минал опит" (*past experience*) се имат предвид данните, свързани с даденият проблем, с които вече разполагаме и можем да използваме, за да го разрешим.

Трябва да се има предвид, че качеството и големината на тези данни са от основно значение за успеха на изследването (Mohri, Rostamizadeh, & Talwalkar, 2012).

Машинното самообучение е обширно понятие, в което могат да бъдат разграничени различни (подсектори) групи от методи. В основата си то представлява едно взаимодействие между "обучаващия се" (*the learner*) и външната среда, като в ролята на този обучаващ се стои компютъра (Shai & Shai, 2014). Това обуславя възможността да разграничим задачите за машинно обучение в две основни категории именно в зависимост от особеностите на тази интеракция, като трябва да се има предвид, че с това не се изчерпват възможните разграничавания. Тези две основни категории са:

- Надзиравано машинно самообучение (с "учител" – *supervised learning*)
- Ненадзиравано машинно самообучение (без "учител" – *unsupervised learning*)

В случая на надзиравано самообучение, тренировъчният пример (миналия опит/данни), който ще бъде използван за обучение от компютъра съдържа значима информация, която липсва в тестовите данни, върху които ще бъде приложено наученото от алгоритъма. Опита, който е придобил обучаващия се компютър в процеса на самообучение бива използван, за да се прогнозира/предвиди липсващата информация в тестовите данни. В този сценарий можем да кажем, че външната среда влиза в ролята на "надзирател" или "учител" на обучаващия се компютър чрез предоставянето на тази допълнителна информация (Shai & Shai, 2014). Тази допълнителна информация всъщност е зависимата променлива в анализа (наричана още етикет/целева променлива – *target variable*). Входните данни които използваме, за да предвидим стойността на зависимата променлива, наричаме обясняващи/независими променливи (*explanatory variables*).

В случая на ненадзиравано машинно самообучение нямаме "учител", който да предостави допълнителна информация към тренировъчните данни – в този сценарий стойността на целевата променлива не е известна по време на обучението. В това именно се състои и основната разлика между надзиравано и ненадзиравано самообучение. Очевидно е, че в този случай е много по-трудно да се оцени представянето на обучения алгоритъм за предвиждане.

В прегледа на научни статии, осъществяващи анализ на настроението (в точка 1.5 от тази глава) ще стане ясно и кои от алгоритмите на машинното самообучение са били най-често използвани за анализ на настроението, като това наше знание е пряко обвързано и със създадените хипотези, залегнали в настоящата теза.

1.4. Приложения в различни домейни

Както вече споменахме и в увода на настоящата работа, когато говорим за анализ на настроението в текст, от огромно значение е т.нар. домейн (сфера), в който той ще бъде използван. Приложения на този вид анализ можем да открием във всяка една сфера в нашият свят, когато се интересуваме от това какво мислят другите – от потребителски стоки, услуги, здравеопазване и финансови услуги до различни обществени събития и политически избори. Безспорно обаче, изразните средства, с които хората ще изкажат своето мнение ще се различават значително в зависимост от контекста и домейна на текста (Pang & Lee, 2008). Това показва и Read (Read, 2005, June) в своето значимо

изследване на представянето на различни алгоритми на машинното самообучение за анализ на настроението в разнообразни домейни – от анализ на отзиви за филми до новинарски статии. Най-прост пример защо домейна е важен и влияе пряко върху анализа на настроението в текст е думата “freezing” – ако разглеждаме отзиви за хладилници думата е много вероятно да означава положително мнение, но същата дума използвана в отзиви за софтуерни приложения ще има отрицателно значение. Можем да дадем подобен пример и с думата “predictable” – ако говорим за кола, тази дума ще има положително значение, но ако даваме мнението си за филма, който сме гледали – това по-скоро предполага отрицателно мнение. Има случаи, когато дори различните форми на една и съща дума могат да бъдат асоциирани с противоположни емоции (например, “improved” и “improve” в отзиви за софтуер). Така става ясно, че домейна има централна роля в представянето на системи за разпознаване на настроението.

Преди да продължим разискването на домейна и неговото важно значение е добре да направим и една вметка относно източниците на текст, който може да подлежи на подобен анализ. Liu (Liu, 2012) посочва, че те могат да бъдат най-разнообразни – новини, туитове, дискусии във форуми, блогове, постове във Фейсбук, отзиви за продукти и услуги и още много други. Всъщност в това число можем да включим абсолютно всякакъв текст, независимо от източника, като единственото необходимо и достатъчно условие е той да изразява мнението на автора относно дадена тема/обект. Все пак, Liu (Liu, 2012) подчертава, че отзивите са изключително фокусирани и богати на изразни средства – това ги прави много по-лесни за анализ, тъй като съдържат малко излишна информация. Поради тази причина те са предпочитаният източник на данни в повечето изследвания, засягащи анализа на настроението. Според Liu дискусиите във форуми са най-трудни за такъв вид анализ, тъй като там потребителите могат да засягат много теми, докато междувременно и комуникират помежду си. Liu също подчертава важността на домейна и изразява своето мнение, че отзиви за стоки и услуги обикновено са по-лесни за анализ, доколкото обществени и политически дискусии предоставят много повече предизвикателства, тъй като засягат комплексни теми и се характеризират с висока концентрация на сарказъм и ирония.

Съществуват проучвания изцяло посветени на това да осъществят цялостен литературен преглед на научното поле, занимаващо се с анализа на настроението в текст -

(Vinodhini & Chandrasekaran, 2012), (Medhat, Hassan, & Korashy, 2014), (Ravi & Ravi, 2015) и други. Поглеждайки прегледа на изследвания в сферата, осъществен от Vinodhini et al. виждаме, че те са реализирани основно в домейна на отзивите за филми, а почти цялата останала част от прегледани изследвания са в домейна на отзивите за продукти в Amazon. Според авторите, анализът в тези две сфери се различава значително, тъй като в отзивите за филми има висока концентрация на ирония, доколкото в отзивите за продукти често се наблюдават много специфики на изказа, свързани с конкретни характеристики на продуктите, както и наличието на смесица от положителни и отрицателни изказвания в един отзив към тези отделни характеристики на продукта. В своето изследване Medhat et al. (Medhat, Hassan, & Korashy, 2014) правят преглед на 54 актуални научни статии на тема анализ на настроението и също констатират, че в огромна част от случаите за анализ са били използвани отзиви за продукти (отново източника е Amazon). В техният преглед намират място и доста изследвания, които използват данни от социални мрежи (най-често Twitter). В изследването на Ravi (Ravi & Ravi, 2015) също могат да бъдат открити подобни наблюдения. Може би една от основните причини отзивите за продукти и филми да бъдат най-често използвани в изследвания на анализа на настроението е, че сайтовете в които биват поместени подобни отзиви разполагат и с рейтингова система, което спомага за по-лесното обособяване на положителните и отрицателните коментари (за тази цел се използват различни техники, които ще изясним по-късно в настоящата работа). Последно е важно да се отбележи, че съществуват и опити за създаването на системи за разпознаване на настроението, пригодни за използване в различни домейни, но това продължава да бъде предизвикателство в сферата именно поради споменатите по-рано трудности.

На база на направеният литературен преглед става ясно, че отзиви за софтуерни приложения биват използвани, но много по-рядко в научните разработки в сферата на анализа на настроението. Източници на подобен вид данни обаче, съществуват, като дори могат да бъдат открити в платформи за споделяне на мнение с рейтингова система. Мобилните приложения са една отделна и обширна категория продукти, в която също съществуват конкретни специфики на изказа на потребителите (да вземем за пример думи като “bug”, ”update”, ”release”, “download”, “fix” и т.н). Последното означава, че ако искаме да постигнем висока предвиждаща точност с подобен вид данни ще се нуждаем от

специално създадена система, работеща конкретно за този домейн. Анализът на настроението в сферата на мобилните приложения също носи своите позитиви (в увода по-рано споменахме дори за наличието на фирми, занимаващи се основно с такава дейност), които отново ще подчертаем малко по-нататък, правейки преглед на вече съществуващи изследвания в този домейн.

Поради изложените наблюдения, за обект на настоящото изследване са избрани именно отзиви за мобилни приложения. Този домейн представлява интерес основно поради причината, че е по-слабо застъпен в съществуващата литература в сферата. Малко са изследванията на спецификите на изразните средства, начините на обработката на този вид данни, както и представянето на алгоритми на машинното самообучение конкретно в този домейн.

1.5. Обзор на научни статии с основен фокус върху анализа на настроението

Следва литературен обзор на научни изследвания, в които се ползват методи на машинното самообучение за анализ на настроението. В тази точка ще разгледаме домейни, различни от този, който е обект на настоящата работа, а в следващата ще разгледаме и разработки конкретно в него. Ще обърнем внимание на поставените цели в тези изследвания, използваните техники за обработка на естествен език, приложените алгоритми и получените резултати. Някои от техниките за обработка на естествен език, споменати тук и използвани и в настоящото изследване са дефинирани в глава II – "Методология". За останалите посочени техники, за които не сме дали дефиниция в глава II, такава може да бъде открита в съответния литературен източник на който сме се позовали. В учебника на Manning и Schütze (Manning & Schütze, 1999) също могат да бъдат открити много от основните концепции и техники в сферата на обработката на естествен език.

В своето пионерско изследване в научното поле на анализа на настроението, Pang et al. (Pang, Lee, & Vaithyanathan, 2002, July) създават класификационни модели за разпознаване на полярността на настроението в отзиви за филми – т.е. отново се разглеждат положителни и отрицателни емоции. Авторите имат за цел да изследват представянето на три алгоритъма на машинното самообучение, използвани за

класификация на текста – Наивен Бейсов модел (Naïve Bayes), Метод на опорните вектори (*Support vector machines - SVMs*) и Максимална ентропия (*Maximum entropy*)[10]. Авторите използват данни от IMDB[11], като рейтинговата система в този сайт им помага за създаване на зависимата променлива (настроението) без да има нужда от човешка намеса за това (повече детайли относно подобни техники има в глава II – "Методология"). В обработката на текста не са премахнати стоп думи (*stop words*), не е приложен стеминг (*stemming*) и е взето предвид отрицанието в текста (*negated expressions*). Като обясняващи променливи са използвани унограми и биграми (*unigrams* и *bigrams*), като авторите изпробват различни схеми. Резултатите от изследването показват, че алгоритмите се представят по-добре в сравнение с постигнатото от хора в тази задача за разпознаване на настроението. Метода на опорните вектори се характеризира с най-добро представяне, като това е постигнато чрез използването само на унограми. Авторите заключават, че някаква форма на дискурсивен анализ (*Discourse analysis*) би подобрила представянето на алгоритмите с този вид данни.

Gezici et al. (Gezici, Yanıkoğlu, Tapucu, & Saygın, 2012) също се занимават с класификация на полярността в текст, но в тяхното изследване домейна е различен – анализират се отзиви за хотели в TripAdvisor[12]. Основната цел на авторите, е да изследват дали представянето на алгоритми на машинното самообучение (конкретно метода на опорните вектори и логистичната регресия) ще се подобри след въвеждането на специално създадени обясняващи променливи от изреченията в текста. Изводите са, че макар и с малко, все пак създаденият набор от описващи текста променливи води до увеличение в точността на класификационните модели, като в бъдеще авторите искат да задълбочат това изследване чрез тестване върху разнообразни по вид и големина данни, като например блогове.

Narayanan et al. (Narayanan, Arora, & Bhatia, 2013, October) използват популярните данни с отзиви за филми от IMDB, за да изследват различни начини за подобрение на точността на класификацията чрез Наивен Бейсов модел. Настроението в текста отново е

[10] Логистичната регресия разглежда проблеми с два класа на зависимата променлива, а максималната ентропия представлява генерализация на този случай за *N* на брой класа на зависимата променлива. Оттук нататък в настоящата работа, тези два термина ще бъдат използвани взаимнозаменяемо.
[11] Watchlist // IMDb. United Kingdom. 1993. [cited 11.12.2017] Available from: http://www.imdb.com/
[12] Latest reviews. Lowest prices // Trip Advisor. Massachusetts. 2000. [cited 11.12.2017] Available from: https://www.tripadvisor.com/

разглеждано в неговата полярност. В обработката на текста е взето предвид използването на отрицание в текста. В обясняващите променливи са включени биграми и триграми, като е използван метод за селектирането на тези с по-добра предвиждаща способност чрез мярката за взаимна информация (*Mutual information*). Изводите са, че след избора на подходяща обработка на текста и добрият подбор на обясняващи променливи, Наивният Бейсов модел може да постигне много висока точност. Така той догонва представянето на метода на опорните вектори и логистичната регресия, които са най-често използвани за същата задача, но обикновено се характеризират с по-висока точност.

Fang и Zhan (Fang & Zhan, 2015) предлагат генерален подход за разрешаване на задачата за разпознаване на полярността на настроението в отзиви за продукти в Amazon. В своето изследване, те включват данни за четири категории продукти – за красота, книги, електронни и стоки за дома. По време на обработката на естествен език са оставени само изречения, съдържащи думи, които носят информация за настроението (т.нар. субективни фрази – *subjectivity phrases*). С помощта на морфологичен разбор на изречението (*part-of-speech tagging- POS tagging*) за обясняващи променливи в анализа са селектирани само прилагателни имена, съществителни имена и глаголи, тъй като те са основните думи, носещи информация за настроението. И в това изследване, авторите вземат предвид използването на отрицание в текста чрез конкретни техники от обработката на естествен език. Изследва се представянето на Наивен Бейсов модел, Метод на опорните вектори и Метода на случайната гора (*Random forest*). В бъдеще авторите искат да изпробват предложената методология и върху други видове текстови данни.

Tripathy et al. (Tripathy, Agrawal, & Rath, 2015) провеждат сравнително изследване на представянето на Наивния Бейсов модел и метода на опорните вектори за разрешаване на задачата за разпознаване на полярността на настроението. Отново са използвани данни с отзиви за филми. Като част от обработката на текста, биват премахнати стоп думи и специални символи в текста. Също така целенасочените повторения на букви в някои думи също са премахнати. Въпреки, че авторите не навлизат в детайли относно обработката и създаването на класификационните модели, тяхното изследване очертава добре цялостната картина на методологията и основните стъпки, през които всеки изследовател преминава в един подобен анализ.

Vohra и Teraiya (Vohra & Teraiya, 2013) представят изследване на основните техники, използвани в сферата на анализа на настроението. Те констатират, че голяма част от изследователите в сферата са посочили метода на опорните вектори, като най-добре представящия се алгоритъм от машинното самообучение в разрешаването на задачата за разпознаване на полярността на настроението. До същият извод са достигнали и Ravi и Ravi (Ravi & Ravi, 2015), които правят преглед на 161 научни статии, публикувани в периода 2002-2015 г. В тяхното изследване се вижда, че в 55 статии е бил използван този метод, като в по-голямата част от тях, именно с него е била постигната най-голяма предвиждаща способност на създадената система.

От всички разгледани литературни прегледи на тази научна сфера (както и в настоящия) става видно, че когато биват използвани методи на машинното самообучение за класификация на настроението в текст, това са най-често три конкретни алгоритъма – Наивен Бейсов модел, Метод на опорните вектори и Логистична регресия. Бивайки очевидно утвърдени техники за анализ на настроението, ние също ще се спрем на тях, за да постигнем целта на настоящата работа. Важно е да се отбележи, че това са методи на надзираваното машинно самообучение.

Така литературният обзор в настоящата работа ни води и до възникването на едната от хипотезите, които ще бъдат тествани емпирично в рамките на нашето изследване. Тя засяга конкретно техниките, с които ще бъде създадена системата за разпознаване на настроението и вече беше спомената и в увода на работата. **На база на резултатите от анализи на настроението в отзиви от различни домейни, в настоящата разработка допускаме, че най-добро общо представяне на създадената система ще получим използвайки класификация чрез опорни вектори. Нашето очакване е, че спрямо обекта на това изследване (отзиви в домейна на мобилните приложения), този класификатор отново ще успее да се справи най-добре със задачата да разграничи положителните от отрицателните коментари.**

1.6. Обзор на научни статии в софтуерния домейн

Следва да направим литературен обзор на научни статии, занимаващи се с анализа на настроението конкретно в домейна на мобилните приложения, както и да посочим

изследвания, фокусирани не само върху този вид анализ, но намиращи място в същия домейн.

Zhang et al. (Zhang, Hua, Wang, Qian, & Zhang, 2014) подчертават, че последните години има драстично увеличаване на броят софтуерни приложения за мобилни устройства и таблети, а потребителите могат да коментират и дават своята обратна връзка по всяко време чрез различни платформи за комуникация. Въпреки това, авторите потвърждават, че много малко са изследванията на различни класификационни алгоритми за определяне на настроението в текст от тази сфера. Поради тази причина те осъществяват именно такъв анализ, използвайки данни на китайски език за мобилно приложение за комуникация в iOS App Store. Тяхното изследване показва, че характеристиките (пр. средна и максимална дължина) на отзивите за мобилни приложения се различават значително от отзивите за софтуерни приложения за настолен компютър. За разпознаване на полярността на настроението, авторите използват метод на опорните вектори и Наивен Бейсов модел, като последният дава по-добри резултати върху техните данни.

Hogenboom et al. (Hogenboom, et al., 2015) провеждат обстойно изследване на използването на емотикони в текст и акцентират върху това, колко важна информация относно настроението на автора се съдържа в тях. Емотиконите биват разглеждани като способ за подобряване на резултатите на система за разпознаване на настроението в текст. Авторите осъществяват класификация на настроението, чрез методи базирани на лексикон. За да оценят значението на емотиконите в текста, те провеждат експеримента първо без включване на емотиконите – това е индикатор за базовото представяне на класификацията. След това анализа се провежда отново, но този път с включената информация за настроението, което носят емотиконите. Използвани са данни от туитове и съобщения в блогове (на холандски език) за разработване на системата и данни за мобилни приложения в iOS App Store (на английски език) за валидиране на нейната работа. Резултатите показват, че има статистическо значимо увеличение на точността на системата за разпознаване на настроението след добавянето на емотиконите. Това увеличение се наблюдава по-драстично за холандските съобщения, но макар и доста по-малко, такова повишение на точността е наблюдавано и в отзивите за мобилни приложения. Освен използването на предефиниран лексикон с думи, носещи определено

настроение, за анализа авторите използват помощта на лексикон на настроението от емотикони (*emoticon sentiment lexicon*). Всяка емотикона в него представлява поредица от символи, заедно с асоциираната с нея оценка на настроението, което тя носи (*sentiment score*). В този лексикон емотиконите биват организирани в по-големи групи – синонимни емотикони (*emoticon synsets*) – които носят една и съща емоция (например, групи от емотикони означаващи щастие, тъга, любов, скука и други).

На база на изследването на Hogenboom et al. (Hogenboom, et al., 2015) възниква и другата от хипотезите, залегнала в настоящият експеримент. Следвайки примера в това изследване, ние допускаме, че въвеждането на емотикони в класификацията на настроението ще доведе до увеличение на точността на създадената система. Нашият подход се различава от този на Hogenboom et al. в това, че за класификация на полярността са използвани методи на машинното самообучение (а не лексикони). По този начин ние не разчитаме на предефинирани думи, емотикони и оценки на настроението, които те носят, а оставяме алгоритмите сами да научат тази информация от данните използвани в експеримента. Също така, в нашият експеримент извличаме и още няколко характеристики, описващи текста и различни от думите в него. **Хипотезата, която ще бъде изследвана емпирично в рамките на настоящият труд гласи, че добавянето на тези допълнителни обясняващи променливи (в лицето на групи от емотикони, носещи конкретно настроение и други описващи текста характеристики) ще подобри общото представяне на модела в сравнение с представянето му, когато използваме само думите в текста за обучение на алгоритмите.**

Guzman и Maalej (Guzman & Maalej, 2014, August) създават метод за извличане на отделните аспекти на дадено мобилно приложение, които потребителите са коментирали и настроението, изразено към тях. В литературата това може да бъде открито под наименованието аспектно-базиран синтез на мнението (*Aspect-based opinion mining*) - (Liu, 2012). Авторите първо идентифицират конкретни аспекти на дадено мобилно приложение, извличат отношението към тези аспекти чрез анализ на настроението, след което използват техники от анализа за моделиране на теми в текст (*Topic modeling*) – по този начин те обединяват сходните дискутирани аспекти в по-общи групи. Този вид анализ изисква по-задълбочени техники от обработката на естествен език, но резултатите могат да бъдат от огромно значение и стойност. В изследването са използвани отзиви за

приложения в Google Play Store и iOS App Store. Подходът на авторите и резултатите помагат на разработчиците на приложения да филтрират, агрегират и анализират по-лесно богатата информация, съдържаща се в отзивите на потребителите. Авторите подчертават, че тази информация може да бъде използвана за идентифициране на нови изисквания на потребителите или планиране на бъдещи пускания на нови версии. Авторите имат намерение да тестват системата в реална среда, за да може да бъдат установени евентуални нейни недостатъци.

Gu и Kim (Gu & Kim, 2015, November) си поставят за цел да отговорят на въпроса кои елементи на мобилното приложение са обичани от потребителите му. Те подобряват описаната в предният параграф система на Guzman и Maalej (Guzman & Maalej, 2014, August) за извличане на аспекти и потребителското настроение към тях. За да оценят представянето на създадената система в реална среда, авторите я прилагат върху отзивите за 17 популярни приложения за Android, след което представят резултатите на разработчиците на тези приложения. Обратната връзка от тях е изключително положителна. Те потвърждават, че системата им е помогнала да уловят бързо и лесно настроенията на потребителите към отделните елементи на приложенията и така да могат да идентифицират предимствата и недостатъците на продукта.

Fu et al. (Fu, et al., 2013, August) също разработват система за обобщаване на потребителските отзиви за мобилни приложения. Използваните данни са от Google Play. Предложеният от тях способ открива несъответствия между посоченият рейтинг към отзива и коментара на потребителя чрез помощта на регресионен модел. Също така авторите използват техника за моделиране на теми в текст, за да могат да идентифицират основните причинители на недоволство у потребителите на дадено приложение. Разработената система дори предлага възможността да се проследи как оплакванията на потребителите са се променяли във времето. В своята бъдеща работа авторите възнамеряват да адаптират системата и към данни от други подобни онлайн платформи за мобилни приложения.

В своето изследване Panichella et al. (Panichella, et al., 2015, September) използват данните от анализа на Guzman и Maalej (Guzman & Maalej, 2014, August) като комбинират техники от естествената обработка на език и анализа на настроението, за да класифицират отзиви за мобилни приложения в предефинирани категории, отнасящи се за поддръжката

и развитието на софтуерните приложения (отзиви, информиращи за даден аспект на приложението; отзиви, търсещи информация/помощ; отзиви, предлагащи идеи за подобряване на приложението; отзиви, описващи проблеми с приложението). За анализа на настроението авторите използват Наивен Бейсов модел. В бъдещата си работа авторите искат да подобрят анализа, използвайки техники за моделиране на теми в текст и да разширят изследването чрез данни за разнообразни други приложения.

През 2017 г. разработка на Ciurumelea et al. (Ciurumelea, Schaufelbühl, Panichella, & Gall, 2017, February) представлява подобрение на споменатото изследване в предния параграф. Ciurumelea et al. твърдят, че предефинираните категории, използвани от Panichella et al. (Panichella, et al., 2015, September) са полезни, но прекалено общи. За да може да отговарят на конкретни нужди на потребителите, отзивите в изследването на Ciurumelea биват класифицирани в повече категории (използвани са две равнища на категориите), отнасящи се за развитието и поддръжката на едно мобилно приложение. След "прочитане" на даден отзив, системата, която те разработват е способна да препоръча кои файлове с изходния програмен код (*source code*) трябва да бъдат променени, така че да може да се отговори на проблема, описан от потребителя в този отзив. Външни оценители на системата потвърждават, че тя спестява около 75% от времето на разработчиците, необходимо за ръчна проверка на отзивите.

Hoon et al. (Hoon, Vasa, Schneider, & Mouzakis, 2012, November) като други автори в сферата също подчертават своето наблюдение, че въпреки наличието на много разработки засягащи потребителски отзиви в различни домейни, продължаваме да знаем малко за този, свързан с мобилни приложения. Те осъществяват обширен анализ на 8.7 милиона потребителски отзиви в iOS App Store. Основната им цел е да се запознаят по-подробно с използваната лексика от потребителите в тази сфера. Авторите откриват, че най-често употребяваните думи са такива, изразяващи емоция. Също така набора от думи, чрез който бива предадено отрицателно мнение се оказва по-разнообразен в сравнение с изказа на потребителите, когато мнението е положително. Някои от откритията на авторите съвпадат с тези в настоящата работа (това става ясно в глава III), въпреки че източника на данни е различна платформа за пазаруване на приложения (все пак, продукта е един и същ – мобилни приложения).

В следващо свое изследване, същите автори (Hoon, Vasa, Martino, Schneider, & Mouzakis, 2013, November) акцентират върху важността на рейтинговата система в платформи като iOS App Store, тъй като тя дава директна оценка за качеството на дадено мобилно приложение, а този пазар представлява силно конкурентна среда. Освен че отзивите дават обратна връзка на разработчиците за приложенията, те предоставят и данни за анализ на конкуренцията. Поради тази причина авторите провеждат задълбочено изследване на съдържанието на тези отзиви. Те представят анализ на това, до колко рейтинга (под формата на брой звезди), поставен от потребителя, кореспондира на настроението, изразено в коментара му за приложението. Изводите са, че при по-кратки отзиви (до пет думи) има съвпадение между вербалната и числова оценка на потребителите и разработчиците биха могли да използват само последната (когато отзивите са по-кратки).

Изследването на Grano et al. (Grano, et al., 2017, September) отново е свързано с анализ на отзиви за мобилни приложения с цел поддръжка и усъвършенстване на софтуера. Те използват данни от Google Play Store за 395 различни приложения. Основната цел на тяхната работа е да създадат такъв набор от данни, който да бъде използван от изследователите в сферата за провеждане на експерименти с цел по-добро разбиране на това как отделни аспекти, свързани с качеството на програмният код могат да рефлектират върху отзивите за приложението и неговия рейтинг. Също така авторите искат да спомогнат за разбирането на това как разработчиците реагират на конкретни потребителски нужди, изразени в отзивите и ги вземат предвид по време на разработката на приложението. В своята работа, подобно на описаните по-рано изследвания, Grano et al. разпределят отзивите в предефинирани категории, свързани с намерението на автора (например, съобщение за проблем или предложение за подобрение на софтуера). Авторите предоставят публично достъпен набор от данни, в който отзивите са свързани с конкретни версии на приложенията и са класифицирани според вида обратна връзка от потребителя (чрез предефинираните категории). Също така мобилните приложения, част от тези данни, са анализирани от авторите и притежават оценка на качеството на програмният код (за тази цел са използвани различни метрики за качество).

Настоящата работа е с основен фокус върху анализа на настроението. Grano et al. предоставят данни, които могат да послужат за най-различни разработки в домейна на

мобилните приложения, включително и за настоящата. Предоставената извадка от тях е с голям размер на броят отзиви и засяга разнообразни видове приложения (това ще покажем по-късно в глава III), като междувременно е и изключително скорошна към датата на разработка на настоящата работа.

Поради изложените причини, в това изследване с цел създаване на автоматизирана система за разпознаване на настроението, са използвани данните за потребителски отзиви в Google Play от разработката на Grano et al. (Grano, et al., 2017, September).

1.7. Приложения на анализа на настроението

Направеният литературен преглед до този момент вече успя да хвърли светлина върху актуалността на разглежданият проблем в настоящата работа. Все пак в тази последна точка бихме искали да обобщим накратко разнообразните приложения на една система за разпознаване на настроението.

Liu (Liu, 2012) подчертава, че много от големите световни корпорации (Microsoft, Google, Hewlett-Packard, SAP, SAS) разработват вътрешно подобни системи за своите нужди, като именно този интерес от страна на бизнеса мотивира и огромният такъв от страна на академичният свят към анализа на настроението.

Според Ravi (Ravi & Ravi, 2015) анализа на настроението все още се намира в своя най-ранен стадий от гледна точка на разпространението му в реалният свят. Той бива прилаган в разнообразни сфери, като бъдещето предполага откритие на още области, в които той е необходим, а резултатите от него – ценни за бизнеса. **Някои от настоящите сфери, в които прилагаме този вид анализ са:** предвиждане на цените на пазара за различни финансови инструменти; изследвания на влиянието на социалните мрежи върху цените на акции, както и върху фирменото развитие; измерване на интереса на инвеститори; предвиждане на боксофис в сферата на кино-индустрията; предвиждане на нивото на полезност на коментари относно стоки и услуги; създаване на предвиждащи модели за напускащи клиенти/служители (*churn prediction*); измерване на удовлетвореността на клиента; създаване на препоръчващи системи (*recommender systems*) за реклами в социалните мрежи, филми, ресторанти, хотели, туристически дестинации и други; безброй приложения в маркетинга – опознаване на конкуренцията, измерване на

успеха на маркетингови кампании, измерване на репутацията на бранд/компания, изследване на обратната връзка от клиента, откриване на проблеми в продукти; различни проучвания в социалната и политическа сфера.

Някои от приложенията в софтуерният домейн са (Martin, Sarro, Jia, Zhang, & Harman, 2017): анализ на потенциални проблеми (бъгове – *bug reports*) с мобилното приложение; препоръчващи системи, спомагащи на разработчиците да планират промените и обновленията на софтуера; изследвания на това как промените в софтуера влияят на потребителските рейтинги; анализ на конкуренцията; анализ на реакциите на потребителя спрямо промени в цената и рейтинга на приложението; запознаване с потребителските изисквания към дадено приложение (*feature requests*); анализ на най-честите оплаквания на потребителите; откриване и приоритизиране на информативни потребителски отзиви; препоръчващи системи на база на преференциите на потребителя; подобряване на системите за бизнес разузнаване (*Business intelligence*) на софтуерните компании и други.

В заключение на тази глава, можем да кажем, че поставяйки си целта за създаването на система за разпознаване на настроението в отзиви за мобилни приложения, ние се надяваме да допълним съществуващата литература в този домейн. От литературният обзор става ясно, че всяко едно подобно изследване в тази научна област носи своите специфики и е уникално само по себе си поради няколко причини.

Това се дължи първо на факта, че в анализа на текст има безброй техники за обработка на естествен език. Избора на набор от такива, които да бъдат приложени, е строго специфичен в зависимост от особеностите на конкретния текст, обект на изследването. Настоящата работа представя методология за осъществяването на тази обработка и привеждане на данните, част от този домейн, в подходящ за количествен анализ вид. Методологията е авторско съчетание на различни техники от обработката на естествен език и е пригодена напълно за задачата, която разрешаваме в настоящият труд.

Второ, подбора и избора на характеристики (обясняващи променливи), описващи текста също може да доведе до различни резултати и е абсолютно индивидуален (отново зависи от спецификите на текста и това се вижда в резултатите на изследвания в различни домейни). Трето, резултатите от прилагането на алгоритми на машинното самообучение могат да бъдат предвидени в много малка степен. Освен, че отново зависят от домейна

(например текст, богат на ирония или сарказъм (отзиви за филми и политически събития) - ще затрудни и влоши представянето на класификатора), те зависят изцяло и от решенията взети на предните две посочени стъпки (обработка на текста и подбор на обясняващи променливи). Могат да зависят и от много други фактори, като качество и големина на данните и подходящо настройване на класификационните модели (*model tuning*).

В настоящата работа адресираме и проблема, свързан с използването на емотикони и други обясняващи променливи (различни от думите в текста), като способ да се увеличи предвиждащата способност на системата за разпознаване на настроението.

II. Методология

Методологията на настоящият анализ е внимателно съобразена и избрана, така че да успее да отговори на поставените цел на изследването и изследователски задачи.

Анализите в областта на науката на данните или по-конкретно в областта на извличането на знания от данни и машинното самообучение споделят един и същ жизнен цикъл. Под жизнен цикъл се имат предвид отделните главни стъпки в методологията по създаването на едно изследване, които обикновено са едни и същи без значение от спецификите на конкретния проект.

В научната литература много автори дават описание на процеса на моделиране на данни и основните етапи, през които преминава изследователя в подобен род анализи, но макар и с малки разлики понякога, концепцията си остава една и съща.

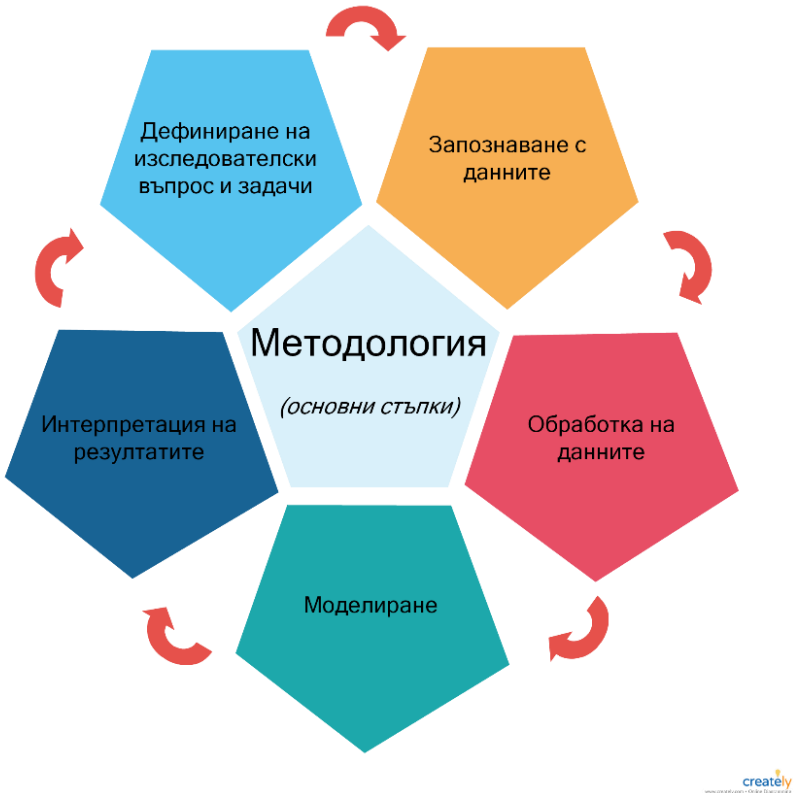
Фигура 1[13] представя основните етапи в методологията на настоящото изследване. Схемата на всяка една от стъпките не бива да бъде разглеждана като линеен процес. Както във всеки един проект в сферата на науката на данните, това е цикличен процес – на всяка стъпка оценяваме получените резултати, преценяваме дали всички стъпки преди това са изпълнени както трябва, дали трябва да бъде променено нещо и след това определяме как да се продължи напред (Peng & Matsui, 2016).

Всеки един проект в сферата започва с дефиниране на проблема, който трябва да бъде разрешен. В настоящата работа на този етап изясняваме някои основни моменти в анализа – дефинираме проблема, който ще бъде изследван; посочваме техники, с които би могъл да бъде разрешен, имайки предвид спецификите му; формираме хипотези относно резултатите от емпиричното изследване.

Вторият етап в процеса е запознаването с данните, които ще бъдат използвани в емпиричното изследване. На тази стъпка наблюдаваме данните от различни перспективи, създаваме целевата променлива (подробно описание на целият процес, както и формални дефиниции на използваните термини могат да бъдат открити по-долу в тази глава) и взимаме предвид някои от спецификите на данните, за които смятаме, че ще бъдат от значение в следващите етапи от изследването.

[13] Схемата е създадена с помощта на Creatly // Creately. Australia. 08.11.2011. [cited 15.01.2018] Available from: https://creately.com/

Следва обработката на данните. Този етап е от решаващо значение за успеха на цялото изследване. Тъй като работим с текстови данни, на тази стъпка нашата основна задача е, те да бъдат приведени в подходящ за количествен анализ вид (т.е. да бъдат представени в числов вид). По-важни фази през които преминаваме на този етап, са изчистването на данните, токъницазия (*tokenization*), лематизация (*lemmatization*),



Фигура 1. Основни етапи в методологията на настоящото изследване

извличане на обясняващи променливи. Колкото по-качествено обработим данните на този етап, толкова по-добри резултати можем да очакваме от изследването.

След приготвянето на данните стигаме до етапа на тяхното моделиране – прилагаме избрани алгоритми; валидираме и оценяваме представянето на моделите на база на конкретни метрики.

На последната фаза от изследването следва да интерпретираме резултатите, да си отговорим на създадените хипотези и да открием потенциални слабости в анализа, чрез което да идентифицираме възможности за подобряване на резултатите. Както споменахме по-рано, целият процес е цикличен, тъй като именно с разкриването на тези потенциални

възможности следва да предефинираме съществуващият или направо да създадем нов изследователски въпрос и задачи, с което да започнем нов цикъл на анализа.

Преди преминаване към подробно описание на методологията по създаване на система за разпознаване на полярността на емоцията в текст, стъпка по стъпка е важно да спомена, че софтуерът, използван за създаването на целия експеримент е програмният език Python. Използваната интегрирана среда за разработка (*integrated development environment, IDE*) чрез програмния език Python е Spyder. Основните библиотеки, които са ползвани за анализа са: Natural Language Toolkit (nltk), scikit-learn, NumPy, pandas. За визуализациите е използвана средата Jupyter Notebook с помощта на библиотеките plotly[14] и wordcloud.

Анализиран е само текст на английски език, тъй като някои от използваните библиотеки и техники са разработени конкретно според спецификите и граматичните правила на този език, а това е от кардинално значение за анализа. Това съображение е взето предвид и при избора на данни за осъществяване на целия експеримент.

2.1. Основни концепции

Както вече стана ясно в глава I, анализът на чувства и настроения в настоящата теза е осъществен основно с методи на машинното самообучение.

Ще разглеждаме проблем от машинното обучение с “учител”. Това ни позволява освен използването на конкретни алгоритми, така и извеждането на относително точна оценка на представянето и резултатите от обучението.

Както по-рано беше изяснено в предмета на настоящото изследване – разрешаваме проблем, свързан с разпознаването на полярността на потребителската емоция. Според ***Дефиниция № 3,*** това означава, че се ограничаваме в определянето на това, дали даден текст (потребителски отзив) отразява положително или отрицателно настроение. От това следва, че целевата променлива заема точно две категории (положителна и отрицателна), което обуславя проблема като такъв на класификацията в машинното самообучение. Нещо повече - класификацията в този случай е бинарна (*binary classification*), тъй като

[14]Plotly предоставя онлайн графични и аналитични инструменти за създаване на визуализации на данни, като съществуват библиотеки за ползването му в Python, R, MATLAB и други софтуери.

съществуват точно две категории (при повече от две категории, говорим за класификация с повече групи – *multiclass classification*).

Входните данни за анализа представляват отзиви на потребители за мобилни приложения, като всяко от тях има и оценка от автора – между 1 и 5 звезди. За целите на анализа са оставени само отзиви, в които авторът е приложил и своето мнение – т.е. задължително съществува и текст от всеки потребител, а не само оценка между 1 и 5. Все пак продължава да стои въпросът, как именно ще създадем зависимата променлива, така че тя да се състои само от два класа – положителен и отрицателен (при положение, че разполагаме с текста и пет степени в оценъчната скала на рейтинга). Конкретно чрез създаването на тази целева променлива ще можем да разглеждаме проблема като такъв на надзираваното машинно самообучение.

Отговорът е, че именно наличието на споменатия рейтинг към всеки отзив на потребител ни позволява да създадем нашата зависима променлива. За тази цел е използвана методиката, спомената от (Cambria, Das, Bandyopadhyay, & Feraco, 2017) и наричаща се "отдалечено наблюдение" (*distant supervision*).

Преди да опишем тази методика е добре да отбележим, че друга техника за откриване на зависимата променлива, която Cambria et el. (Cambria, Das, Bandyopadhyay, & Feraco, 2017) споменават, е ръчното поставяне на етикети (стойности) на зависимата променлива (*manual labeling*). Тази задача може да бъде изпълнена чрез делегиране на работата на множество непознати онлайн потребители (*crowdsourcing*). Пример за подобна онлайн платформа, която позволява това е Amazon Mechanical Turk[15]. Принципът на работа и потенциалните проблеми, които биха могли да възникнат с тази платформа са разгледани подробно от Paolacci et al. (Paolacci, Chandler, & Ipeirotis, 2010). Основното предимство на ръчната пред техниката за отдалечено наблюдение е, че прилежащият етикет към всеки отзив бива верифициран от човек. Недостатък обаче е, че не можем да бъдем напълно сигурни в качеството на изпълнената работа, както и това, че със сигурност е доста по-времеемка в сравнение с другият предложен метод.

Името на техниката за отдалечено наблюдение произлиза от това, че няма ръчно поставяне на етикет (стойност) на зависимата променлива – т.е. не се изисква човек да

[15] Human intelligence through an API. Access a global, on demand, 24x7 workforce // Mechanical turk. United States. 2005. [cited 14.12.2017] Available from: https://www.mturk.com/

прочита всеки един отзив и да определя според прочетеното, дали отзива е с категория “положителен” или категория “отрицателен” (както е в другият вид предложена методика). В този смисъл се използва определението “отдалечено” в наименованието на метода. Отново можем да дадем пример с използването на т.нар. “хаштагове” в туитове с цел откриване на настроението на автора. Както споменахме в Глава I, тази техника е използвана в изследването на Mohammad (Mohammad, 2012, June). Например, туитове с хаштаг “#щастлив” могат да бъдат отнесени към категория “положителен”. Това именно, е пример за използване на отдалеченото наблюдение с цел поставяне на етикет. Cambria et al. (Cambria, Das, Bandyopadhyay, & Feraco, 2017) дават пример и с отзиви за продукти, продавани от американската компания Amazon.com, Inc., където пък рейтинговата система със звездички може да бъде използвана за определяне на целевата променлива. Използва се логиката, че отзив, който е с 1 звезда отразява отрицателно мнение на потребителя, а отзив от 5 звезди е силна индикация за положително мнение.

В настоящата работа е използван именно този метод за създаване на зависимата променлива. Още много са изследванията, в които е използвана техниката за отдалечено наблюдение с цел поставяне на етикет - (Pang, Lee, & Vaithyanathan, 2002, July), (Cui, Mittal, & Datar, 2006, July), (Hogenboom, и др., 2015) и други.

В литературата се наблюдават различни способи за използване на рейтинга към потребителски коментари, като най-често авторите обединяват отзиви с една и две звездички в отрицателният клас и отзиви с четири и пет звездички в положителният клас. Мнения на потребители с три звезди се считат за неутрални (Liu, 2012).

Тъй като целта на настоящата работа е да се обучи модел, разграничаващ положителни от отрицателни коментари - неутралният клас не участва в изследването, което означава, че отзиви с три звездички няма да бъдат анализирани. Това предполага и улесняване на класификационната задача, както споменава Liu (Liu, 2012).

Отрицателният клас в настоящото изследване е съставен от коментарите на потребители, които са с оценка само една или две звезди. Положителният клас е създаден от отзиви с пет звезди. Причината да не се включат мненията с четири звезди е, че положителните коментари като цяло са много повече, отколкото тези с отрицателна оценка – т.е. разполагаме с предостатъчно данни за положителният клас. Изключвайки тези мнения и използвайки най-безспорно положителните (т.е. с максимум звезди)

постигаме по-голямо разграничение между двата класа в разглежданият проблем. Този способ е използван от Rain (Rain, 2013), който предполага влошаване на резултатите от класификацията, когато се ползват отзиви от всеки вид рейтинг – според него така различията между тях намаляват, а това затруднява разграничаващата способност на модела.

Последният важен елемент, който ще засегнем в тази точка, изясняваща основната постановка в настоящият анализ, е разпределението на така създадените класове на целевата променлива. Ще работим с унифицирано разпределение на класа (*uniform class distribution*) или т.нар. "балансирана извадка" (*balanced dataset*). Това означава, че ще работим с еднакъв брой наблюдения в положителния и отрицателния клас. Тази мярка се налага поради факта, че някои от алгоритмите на машинното самообучение са чувствителни към съотношението на класовете на целевата променлива и това е в полза на класа, който се среща по-често. Това може да доведе до проблеми в оценката на представянето и подвеждащи резултати. Тези проблеми са разисквани в научната статия на Chawla et al. - (Chawla, Japkowicz, & Kotcz, 2004), която предоставя и детайлен преглед на значимите изследвания в този контекст.

Поради горепосочените причини, на случаен принцип от цялата извадка с данни са изтеглени еднакъв брой наблюдения от положителният и отрицателният клас на целевата променлива. Голяма част от изследванията, в които се прилага класификация на настроението, използват именно този начин за създаване на извадка, чрез която да бъде обучен класификационният алгоритъм - (Pang, Lee, & Vaithyanathan, 2002, July), (Gräbner, Zanker, Fliedl, & Fuchs, 2012), (Narayanan, Arora, & Bhatia, 2013, October) и други.

2.2. Обработка на данните

Текстовите данни могат да бъдат определени като силно неструктуриран вид данни. За да може да бъде извлечена информация от тях с помощта на математически и статистически подходи, те трябва да бъдат приведени в подходящ за анализ вид. Това включва серия от трансформации, като крайната цел е данните да бъдат представени в числов вид. Накратко, основният въпрос, който си поставяме на този етап е с какви методи да преобразуваме данните в числов вид, така че да можем да приложим върху тях някои от алгоритмите за извличане на знания от данни.

2.2.1. Ниво на репрезентация на текста

Първо, трябва да бъде изяснено какво ще бъде нивото на репрезентация на текста. Liu (Liu, 2012) посочва три основни нива на репрезентация на текста, които биват ползвани в анализа на чувства и настроения, а именно – на ниво документ (*document level*), на ниво изречение (*sentence level*) и на ниво аспект (*aspect level*).

Когато анализираме на ниво "документ", задачата на анализа е да класифицира правилно, дали целият документ изразява положително или отрицателно мнение. В контекста на отзивите за продукти, това ниво на анализ дава отговор на въпроса, дали отзива на даден потребител е като цяло положителен или отрицателен за дадения продукт. Този метод почива на хипотезата, че един отзив се отнася точно за един продукт (не повече).

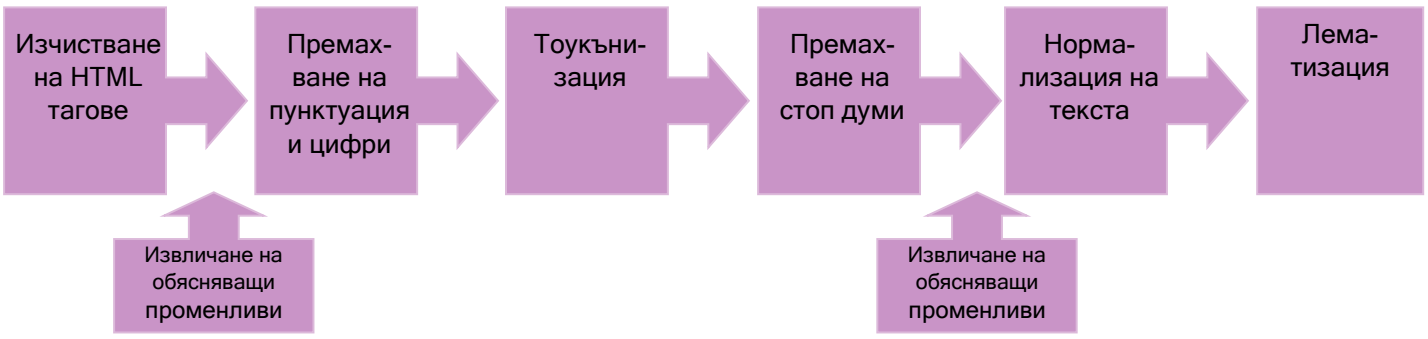
Когато анализираме на ниво "изречение", тогава задачата е същата, но този път я отнасяме към всяко едно изречение в целият набор от отзиви. Последното ниво (аспектно) е най-фино, тъй като там задачата е да се открие отношението на автора към всеки един аспект на продукта/услугата/темата, който е засегнал в неговото мнение. Прост пример, който можем да дадем, е отзив за мобилен телефон, в който потребителят изразява мнение, че е много доволен от камерата на телефона, но батерията е слаба. Анализът на ниво "аспект" ще открие, че относно камерата в даденият отзив е изразено положително мнение, но относно батерията – отрицателно. Също така това ниво изисква идентифициране на аспектите в целия корпус от документи преди прилагането на анализа на настроението (аспектите в отзиви за мобилни телефони могат да бъдат – "батерия", "дисплей", "камера" и др.). Този пример показва, че аспектната репрезентация е най-сложна и на най-дълбоко ниво, тъй като разбива напълно документа на съставните му части.

В рамките на настоящата работа ще се съсредоточим върху анализа на настроението на ниво "документ". Това означава, че всяко едно мнение в извадката ще представлява един "документ", като нашата задача ще бъде да се класифицира правилно настроението на потребителя на дадено мобилно приложение, изразено в този "документ".

2.2.2. Обработка на текста преди преобразуването му в числов вид

Преди преобразуването на данните в числов вид, е добре да се приложат някои предварителни общи стъпки, използвани при обработката на естествен език. За улеснение, **Фигура 2** представя приложените трансформации върху текста, като следва подробно описание на всяка от тях. Важно е да се отбележи, че предложената методология за обработка на текстовите данни е авторско съчетание от разнообразни подходи, прилагани от различни автори в техните изследвания в областта на анализа на текст - (Miner, 2012), (Weiss, Indurkhya, & Zhang, 2015), (Hofmann & Chisholm, 2016). Тя е съобразена изцяло със спецификите на конкретния проблем, който разрешаваме и с особеностите на данните, обект на настоящото изследване.

Трябва да се има предвид, че отзивите, писани от потребители в интернет се характеризират с много шум – правописни грешки, съкращения на думи, жаргон и др. Поради тази причина всяка стъпка в последващата методология за обработка на данните е съобразена с тези специфики на данните.



Фигура 2. Схема на приложените техники за обработка на естествен език върху текста

Тъй като текстовите данни много често биват извлечени директно от интернет страници, задължителна стъпка при обработката, е да се провери и да се премахне евентуалното наличие на HTML тагове и коментари, които е възможно да попаднат в данните по време на извличането им от тези страници. HTML таговете представляват структурна единица на един уеб документ, като са съставени от символ за отварящ таг "<", служебна дума или абревиатура дефинираща структурата и затварящ таг ">".

Следващата стъпка е премахването на пунктуационните знаци, специалните знаци и цифрите. Тъй като по-рано уточнихме, че анализът ще бъде на ниво цял документ – отделните изречения в даден отзив се считат за едно цяло. Важно е да се уточни, че преди

тази стъпка са извлечени всички обясняващи променливи, които възнамеряваме да използваме по време на моделирането и са пряко обвързани с пунктуационните или специалните знаци, използвани в отзивите. Подробно описание на тези променливи е приложено по-нататък в настоящата глава.

Следваща стъпка по време на трансформацията на данните е т.нар. "тоукънизация" (*tokenization*). Тази техника дефинираме по следният начин (Miner, 2012):

Дефиниция № 5: Тоукънизацията представлява разбиване на даден свързан текст на съставните му (най-малки) части, а именно – думи, символи или други елементи, част от него. Тези елементи наричаме "тоукъни".

Понякога тоукънизацията може да бъде и на друго ниво, но за целите на настоящото изследване е необходима тоукънизация на ниво дума. За извършването на тоукънизация в английския език, ролята на разделители на отделните елементи в текста играят интервалите и точките. С помощта на интервалите, текста е разделен на тоукъни, за да може да се продължи към следващите стъпки по обработката на данните. На този етап всеки отзив представлява набор от тоукъни – низ от думи (важно е да се отбележи, че в това число включваме и добавените обясняващи променливи, създадени от пунктуационните знаци - ако в даденият отзив съществува такава променлива, то тя също ще представлява отделен тоукън след трансформацията).

След тоукънизацията идва ред на премахването на т.нар. стоп думи. Тях дефинираме по следния начин (Weiss, Indurkhya, & Zhang, 2015):

Дефиниция № 6: Стоп думите са най-често използваните думи в даден език, които нямат особено значение, нито притежават предвиждаща способност.

Според тази дефиниция, в българският език това могат да бъдат местоимения и предлози, като "аз", "на", "с" и т.н. Настоящият анализ е на английски език и пример за такива думи са - "the", "a", "an", "I" и т.н. Очевидно е, че тези думи не ни дават информация относно настроението в текста, но междувременно са много често разпространени и внасят шум в данните.

Следва нормализация на буквите в текста – премахване на главните букви (*case normalization*). Така всички думи се превръщат в изписани с малка буква. Това е от голямо значение когато софтуерът, с който биват обработени данните, различава главните от малки букви (*case sensitive*). Python се причислява към тези софтуери и думите “amazing” и “Amazing” ще бъдат разпознати като две различни думи в последващия анализ. Поради тази причина се извършва нормализацията. И тук трябва да се отбележи, че ако има обясняващи променливи, свързани с големината на буквите в текста, те трябва да бъдат извлечени преди тази стъпка.

Като част от нормализацията на данните е приложена и една техника, която е особено подходяща в анализа на кратки експресивни текстове, които са съставени основно от изречения, предаващи емоцията на автора - това е боравенето с думи, изкуствено удължени чрез повторение на букви. Авторите на отзиви много често целенасочено повтарят буквите в дадена дума, която носи емоционален заряд, за да акцентират върху нея и да предадат по-изразително своята емоция. Прост пример, който може да се даде, е изписването на думата “good” като “gooooooooood”, за да се засили нейното значение. Broody et al. (Brody & Diakopoulos, 2011, July) показват колко разпространено е целенасоченото удължаване на думите в микроблоговете, както и колко важно е то да бъде взето предвид, когато става въпрос за анализ на настроението. В настоящата работа, всички думи, които имат повече от две повтарящи се букви една след друга, са трансформирани, така че този брой да не надвишава две. След тази трансформация, думи като “goodddddd” отново няма да бъдат с правилен правопис, но ще бъде постигната съществена нормализация на текста, тъй като различните изписвания на думата ще бъдат сведени до значително по-малък брой. Разбира се, трябва да се има предвид, че съществуват техники (Brody & Diakopoulos, 2011, July) за по-прецизно обработване на такива думи, но те попадат извън обхвата на настоящото изследване.

Следващата стъпка в първоначалната обработка на данните е т.нар. “лематизация” (*lemmatization*). Дефинираме я по следният начин (Miner, 2012):

Дефиниция № 7: Лематизацията представлява техника за групиране на различните форми на една и съща дума в една обща “лема”, на база на тяхната основа, контекста, в който са използвани и каква част на речта са.

Тази техника е по-сложна и изисква повече информация, в сравнение с по-често използваната – "стеминг" (*stemming*). Стеминга представлява свеждане на думите до техния корен, като не се взема предвид нито контекста, нито каква част на речта е думата. Очевидно, тази трансформация е по-проста, но води до загуба на информация. Хубав пример, който може да се даде в тази връзка е с думата "parking", която може да бъде използвана и като съществително, обозначаващо места за паркиране, и като глагол в съответното време, изискващо продължителната му форма (частицата -ing). Стеминга ще трансформира и двете до думата "park", а лематизацията ще трансформира само глагола в "park"- съществителното "parking" ще бъде запазено в тази си форма.

Последната стъпка, част от обработката на данните, е премахването на думи, които се състоят от една буква. В английският език има две такива думи – "a" и "I". Тези две думи би трябвало да са част от списъка със стоп думи и на този етап да са вече премахнати от данните. Въпреки това, не трябва да се забравя, че данните с които боравим са далеч от чисти и често може да се натъкнем на отзиви, в които има както грешни изписвания на цели думи, така и просто написани символи и знаци, които не носят никакъв конкретен смисъл. Поради тази причина това е превантивна мярка, така че да бъдем сигурни, че шума в данните е сведен до някаква степен.

Така, предварителната обработка на текстовите данни приключва и вече може да пристъпим към превръщането им в числов вид.

2.2.3. Превръщане на текстовите данни в числов вид

Както по-рано беше споменато, нашата цел е да превърнем текста в цифров вид или казано по по-прост начин - думите да бъдат описани с помощта на числа, за да могат да подлежат на количествен анализ впоследствие. За тази цел ще използваме основният и един от най-популярните похвати за представяне на текст в сферата на търсенето и извличането на информация и обработката на естествен език, а именно векторно-пространственият модел (*the vector-space model*).

Преди да преминем към формалната дефиниция на този модел и да поясним неговото предназначение, би било добре да споменем и някои от алтернативите за репрезентация на текст отново произлизащи от сферата на извличането на информация. Това са Булевият модел (*Boolean model*), както и различни вероятностни модели за

моделиране на език (*probabilistic language models*). Въпреки съществуването на други методи, когато говорим за класификация на текст, най-често използваният начин за репрезентацията му, е именно векторно-пространственият модел, както поради неговата простота, така и поради факта, че позволява използване на пространствената близост за изчисляване на семантичната близост между думите. Поради тази причина, другите споменати методи за репрезентация на текст няма да бъдат разисквани в настоящата работа, като предоставяме референции към разработки, където могат да бъдат открити обстойни разяснения около репрезентацията на текст, както и формални дефиниции на съществуващите методи за това. Manning et al. (Manning, Raghavan, & Schutze, 2008) предоставят детайлно обяснение на някои от най-важните и основополагащи концепции в сферата на извличането на информация. Zhai (Zhai, Statistical language models for information retrieval, 2008) прави критичен преглед на съществуващата литература и постигнатото при използването на статистически езикови модели (*statistical language models*) за извличане на информация.

Векторно-пространственият модел е за пръв път дефиниран от Salton et al. през 1975 г. (Salton, Wong, & Yang, 1975).

Дефиниция на модела:

Нека съществува пространство от документи, които ще означаваме с D_i, всеки от които би могъл да се идентифицира с един или повече индексни термини (*index terms*), които ще означаваме с T_j. Тези термини могат да имат тежести, отразяващи тяхната важност, а могат и да нямат такива и да бъде отбелязано само присъствието или липсата им с 1 и 0. В случай, че имаме t различни индексни термина, а това означава t дименсии, всеки документ D_i може да бъде представен чрез t-дименсионен вектор по следния начин:

$$\mathrm{D_i} = \left(\mathrm{d_{i1}}, \mathrm{d_{i2}} \ldots, \mathrm{d_{ij}}\right), \qquad (1)$$

където с d_{ij} означаваме тежестта на j-тият термин в документа D_i.

По-неформално определение е, че в този модел, документа (в нашият случай - отзива) бива представен като вектор, като елементите на този вектор индикират

присъствието или липсата на даден набор от думи във въпросния документ (Miner, 2012). Когато този набор от думи включва абсолютно всички уникални думи, присъстващи в корпуса от документи, обикновено резултата е пространство с огромен брой дименсии (броят дименсии всъщност е равен точно на броя уникални думи в целият корпус от документи). Още по-просто обяснение е, че всеки вектор, изобразяващ даден документ от нашият корпус с документи, ще има толкова на брой елементи, колкото уникални думи има сред всички документи, с които разполагаме в корпуса. Това означава, че в настоящата работа, всеки отзив ще бъде представен под формата на вектор в *n*-мерното пространство, където с *n* ще означаваме броят думи, които ще бъдат използвани в анализа.

Както пояснихме малко по-рано, този модел е споменат и използван в много от значимите изследвания, засягащи анализа и класификацията на текст, като - (Pang, Lee, & Vaithyanathan, 2002, July), (Yang & Liu, 1999, August), (Turney & Pantel, 2010) и още много други.

В основата на векторно-пространственият модел е залегнало косвено едно голямо допускане, което наричаме "торба с думи" *(bag of words*) и можем да дефинираме по следния начин (Manning & Schütze, 1999):

Дефиниция № 8: "Торба с думи" представлява такава репрезентация на текста чрез думите в него, според която техният ред няма значение. Все пак е интуитивно документи със сходства в тази репрезентацията да имат и сходно съдържание.

Името на това допускане произлиза именно от факта, че документа бива описан от думите, които го съставят, като техният ред не се взима предвид. Разбира се, за всеки е ясно, че този ред не е без значение, но според Miner (Miner, 2012) това допускане не е такъв проблем, когато се прилага за класификация или клъстеринг на текст, където отделните думи са достатъчни, за да може да се диференцира между семантичните концепции в текста.

Процесът, по който превръщаме текста във вектори (или т.нар. "векторизация" - *vectorization*), заема основно три форми (Miner, 2012). Най-често тези форми биват бинарна, честотна (в абсолютна стойност или относителна) и tf-idf (*term frequency–inverse document frequency*). Бинарната форма представя текста като нули и единици, където единицата индикира присъствие на дадена дума, а нулата – липса. Честотната форма

представя текста чрез броят присъствия на дадената дума в текста, а tf-idf репрезентацията представлява метод, който отразява колко важна е дадена дума в корпуса от думи - думи, които често се повтарят в корпуса от всички документи ще имат по-малка тежест и обратното (пример за такава дума в английският език е “the”).

Няма да навлизаме в по-подробно описание на всяка репрезентация, тъй като в настоящата работа, избраната форма на векторизация е бинарна поради две основни причини. Първо, тази форма очевидно е по-опростена и ясна в сравнение с другите и второ – Pang et al. (Pang & Lee, 2008) споменават своето откритие, че в сферата на анализа на настроения и чувства те получават по-добри резултати с бинарната репрезентация, отколкото използвайки честотата на думите (честотата заляга и в популярния в сферата на извличането на текст tf-idf метод). Според тях, това откритие е показателно за разликите между анализа на настроения и чувства и анализа за моделиране на теми в текст. То показва, че първият не се повлиява толкова от това, колко пъти присъства дадена дума в текста, доколкото в анализа за извличане на теми от текст – честотата на думите е от централно значение.

2.3. Извличане на обясняващи променливи от текста

2.3.1. Използване на думи като обясняващи променливи

Всеки един проблем на машинното самообучение, който е разрешаван чрез методи на надзираваното обучение, се нуждае от обясняващи променливи, които да бъдат използвани за предвиждане на зависимата променлива. Когато говорим за анализ на текст, обикновено основно думите играят ролята на тези променливи. Вече споменахме концепцията “торба с думи”, като в ***Дефиниция № 8*** интуитивно асоциирахме тази техника с описанието на текста само чрез отделните (единични) думи, от които е съставен – това са т.нар. “униграми” (*unigrams*). Нищо не пречи обаче, в тази концепция да бъдат включени и две последователни думи (“биграми” - *bigrams*), три (*trigrams*) или повече, в зависимост от целите и спецификите на изследването. Това представлява *n-gram* модел, който дефинираме по следният начин (Miner, 2012):

Дефиниция № 9: *N-gram* моделът представлява репрезентация на текста чрез торба с думи, при която биват използвани не само отделните думи, но и съчетанията от *n* на брой последователни думи.

Разбира се, колкото по-голямо е числото *n*, толкова по-голяма е изчислителната тежест и време за извличането на тези променливи. За да стане напълно ясно, следва прост пример за това, какви характеристики ще извлече модел от биграми (*2-gram model*) в дадено изречение.

Пример:

Нека имаме изречението: “This app is amazing! Keep up with the good work!”.

След предварителната обработка на данните, то ще изглежда по следният начин[16]:

[“this”,”app”,”is”,”amazing”,”keep”,”up”,”with”,”the”,”good”,”work”]

Това всъщност е модел от униграми, тъй като даденият отзив е представен само чрез отделните думи, които го съставляват. Ето как ще изглежда модел от биграми, представящ дадения отзив:

[“this app”, “app is”, “is amazing”, ”amazing keep”, “keep up”, “up with”, “with the”, “the good”, “good work”]

От примера става ясно, че биграмите ще са по-рядко срещани в даден корпус от документи, отколкото униграмите, тъй като са комбинации от две думи. Трябва да се има предвид, че с увеличаване на стойността на *n* в n-gram модела, може да се доведе до ситуация на прекомерно нагаждане към тренировъчната извадка (*overfitting*). Именно поради тази причина, в настоящото изследване като обясняващи променливи са извлечени само униграми и биграми.

2.3.2. Използване на обясняващи променливи, различни от думи

Въпреки, че най-често в анализа на текст, използваните обясняващи променливи са именно думите, съществуват и други характеристики на текста, които носят информация, могат да бъдат извлечени от него и използвани за предвиждане. Поради тази причина, в настоящата работа не сме се ограничили само до използването на униграми и биграми, но

[16] Представеният вид на изречението е без премахнати стоп думи.

сме добавили и още няколко обясняващи променливи, които са директно извлечени от текста.

Емотиконите представляват лесен, бърз и ефективен начин за изразяване на настроения в интернет пространството. В тях е заложен емоционалният заряд на потребителя и те могат да бъдат изключително ценни, когато става въпрос за анализ на чувства и настроения. В допълнение, има логика емотиконите да са широко употребявани именно във всякакъв вид съобщения, в които потребителя трябва да се изрази кратко – туитове, кратки отзиви, чат съобщения, статуси в социални мрежи и др. Както вече споменахме в глава I - това изтъкват и Hogenboom et al. (Hogenboom, et al., 2015) в своето изследване, където показват, че представянето на модел за разпознаване на настроението в данни от Twitter и форуми се подобрява значително след използването на емотикони.

Преди да продължим с методологията по създаването на променливите, трябва да направим едно важно пояснение. В интернет пространството съществуват емотикони и емоджи, като двете не бива да бъдат бъркани. Емотиконите представляват сбор от препинателни знаци, букви и цифри, използвани за създаване на икони на изображения, които обикновено показват емоция или чувство. Емоджи се появяват по-късно, като името произлиза от японски език. Те представляват пиктограми на лица, обекти и символи.[17] Ако дадена икона на лице, може да бъде създадена само със символи на клавиатурата – това е емотикона. Ако не може да бъде пресъздадена чрез препинателни знаци – това е емоджи.

В настоящата работа са извлечени само емотикони с помощта на специална библиотека в Python, разработена именно за тази цел (библиотеката "*emot"*). Първата стъпка е да извлечем всички използвани емотикони в данните, след което да създадем групи от тях, в зависимост от смисъла и емоцията, която предават. За осъществяването на тази цел помага доста и подробният списък в уикипедия, представящ множество групи от емотикони, заедно с техните значения.[18] Списъка е изключително подробен и е създаден на база на голям брой интернет източници, в които официално са изброени емотиконите, използвани в различни платформи за комуникация (например, Yahoo Messenger, MSN

[17] Grannan Cydney. What`s the difference between emoji and emoticons // Encyclopaedia Britannica. Chicago. 2015. [cited 20.10.2017] Available from: https://www.britannica.com/story/whats-the-difference-between-emoji-and-emoticons

[18] List of emoticons // Wikipedia Commons. 05.02.2008. [cited 20.10.2017] Available from: https://en.wikipedia.org/wiki/List_of_emoticons

Messenger), както и с помощта на официалната страница на Уникод Консорциумът[19]. Значимо изследване относно използването на емотикони за комуникация във виртуалното пространство и анализ на техният ефект върху интерпретацията на съобщенията, разменяни между хората по този начин представят Walther et al. (Walther & D'Addario, 2001).

Именно създадените групи от емотикони с подобна емоция, представляват обясняващите променливи, които ще участват в анализа. Типа им отново е бинарен - ако в даден отзив съществува емотикона от съответната група – това бива означено с единица (в противен случай променливата заема стойност 0).

Освен изброените до този момент, от текста са създадени още три обясняващи променливи. Всички са от бинарен тип, като две от тях отново използват наличието на определени препинателни знаци в отзивите.

Първата от тях издава наличието на удивителен знак в отзива. При работа с данните се забелязва, че в повечето случаи на положителни отзиви се наблюдава и използването на удивителни знаци в текста, чрез които потребителят сякаш придава още по-голяма тежест на положителните думи и възклицанието си от съответното приложение.

Втората издава наличието на въпросителни знаци в отзива. Тук наблюдаваме обратната тенденция – често потребители, които са имали проблем или трудност при употребата на дадено приложение задават въпрос. Поради тази причина въпросителният знак в отзива може да бъде индикация за отрицателно мнение на потребителя.

И последната обясняваща променлива е свързана със стила на писане на потребителя и по-конкретно – индикира дали потребителя е използвал само главни букви при изразяването на своето мнение. За всички потребители на интернет или като цяло на електронни устройства е ясно, че в тази дигитална комуникация много често се използват само главни букви, когато човек иска да изрази ярост, гняв, раздразнение и други подобни състояния на фрустрация. Разбира се, съществуват и примери за обратното (изразяване на силно положителна емоция). Все пак, тази особеност определено носи информация за

[19] Това е организация, посветена на разработването, поддръжката и популяризирането на интернационалният стандарт Unicode, който засяга репрезентацията на текста във всички съвременни софтуерни продукти. Повече информация може да бъде открита на техният официален уебсайт - http://www.unicode.org/

емоционалния интензитет на текста и поради тази причина е включена в анализа като отделна обясняваща променлива.

2.3.3. Подбор на обясняващите променливи

След извличането и създаването на обясняващи променливи, идва ред и на техният подбор. Когато работим с текстови данни, тези променливи могат да достигнат огромно число (поне колкото уникални думи има в целият корпус от документи, който анализираме). Това налага ползването на техники за справяне с размерността на данните (*dimensionality reduction*) или прилагането на алгоритми и статистики за избиране на обясняващите променливи, които имат по-добра предвиждаща способност.

В настоящото изследване е използван хи-квадрат теста с цел избиране на променливите, които да бъдат включени в моделирането на данните. Следва формална дефиниция на този тест, за пръв път изследван от Karl Pearson през 1900г. (Pearson, 1900).

Формална дефиниция на хи-квадрат тест:

Хи-квадрат статистиката може да бъде изчислена по следният начин:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \text{ където} \qquad (2)$$

χ^2 – кумулативна тест статистика на Pearson, която асимптотично клони към хи-квадрат разпределение.

O_i - наблюдавана честота на наблюденията от тип i.

E_i - очаквана честота (в случай на липса на зависимост между променливите) на наблюденията от тип i.

В нашият случай този тест е осъществен за всяка една обясняваща променлива и тъй като съществуват само два класа в настоящата класификационна задача, това означава че i ще заема точно две стойности (честота в положителни и честота в негативни отзиви).

По-неформална дефиниция, която можем да предоставим е, че хи-квадрат теста е статистически непараметричен тест, който може да бъде ползван за откриването на зависимост между две категорийни променливи. Хи-квадрат статистиката ни показва разликата между наблюдаваните стойности и тези, които бихме очаквали в случай на липса на зависимост между изследваните две величини (нулевата хипотеза на този тест е, че липсва зависимост между двете изследвани величини). Използвайки вероятностната стойност на хи-квадрат статистиката (*p-value*), можем да определим дали съществува статистически значима зависимост между изследваните две променливи (Gibbons & Chakraborti, 2011).

В настоящата работа, чрез хи-квадрат теста селектираме променливите, които се характеризират с по-силна връзка със зависимата променлива. Тази процедура за подбор е посочена както в учебници, посветени на анализа на текст, така и е използвана в изследванията на редица автори при избора на обясняващи променливи, които да бъдат включени в класификация на настроението или други предвиждащи модели в сферата на анализа на текст – Hofmann (Hofmann & Chisholm, 2016), Miner (Miner, 2012), Cui et al. (Cui, Mittal, & Datar, 2006, July), Feldman и Sanger (Feldman & Sanger, 2007) и други.

Преди пристъпване към следващата част от методологията, е добре да разясним какъв е финалният вид на данните, непосредствено преди използването им от алгоритъм за моделиране.

След приложените трансформации, всеки отзив бива превърнат във вектор от бинарни стойности, индикиращи присъствието или липсата на всяка една от избраните обясняващи променливи (единични думи, две последователни думи, емотикони и други), използвани в анализа. За да стане по-ясно, текстовите данни биват превърнати в бинарна матрица с размери m X n, където m е броя отзиви, използвани в изследването, а n е броя обясняващи променливи, които ще бъдат използвани за обучение на модела за разпознаване на настроението. Това именно е представяне на данните чрез векторно-пространственият модел, който дефинирахме по-рано в тази глава.

2.4. Методи за моделиране

При избора на алгоритми за моделиране, трябва да се има предвид основната особеност на текстовите данни и векторно-пространственият модел, чрез който те са

представени, а именно – размерността. Това означава, че трябва да се насочим към алгоритми, които се справят ефективно с данни с голяма размерност.

По време на литературният преглед стана ясно, че Наивният Бейс, Логистичната регресия и Метода на опорните вектори са най-често използваните и популярни алгоритми за класификация на текст. Именно поради тази причина, това са и избраните методи, които ще бъдат тествани за решението на настоящата класификационна задача.

2.4.1. Наивен Бейсов Модел

Наивният Бейсов модел представлява вероятностен генеративен класификатор, в чиято основа е залегнало правилото на Бейс (McCallum & Nigam, 1998, July). Според това правило можем да изчислим вероятността на дадено събитие, базирайки се на нашата предварителна информация относно условията, които биха могли да бъдат свързани с това събитие.

Теоремата на Бейс гласи:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \text{където} \qquad (3)$$

A и B са събития, като $P(B) \neq 0$.

$P(A|B)$ - условната вероятност събитието A да настъпи, в случай че B е настъпило.

$P(B|A)$ - условната вероятност събитието B да настъпи, в случай че A е настъпило.

$P(A)$ и $P(B)$ - вероятностите за настъпване на събитията, независимо едно от друго.

Теоремата на Бейс се опитва да намери вероятността за настъпването на събитието A, в случай, че B е настъпило.

В уравнение *(3)*, $P(A)$ дефинираме още като **априорната вероятност (*prior probability*) на събитието A**. Тя отразява нашето предварително очакване за вероятността на събитието A преди да разполагаме с допълнителна информация.

В уравнение *(3)*, $P(A|B)$ дефинираме още като **постериорната вероятност (*posterior probability*) на събитието A.** Тя представлява вероятността на събитието A след

като сме взели предвид цялата допълнителна информация, с която разполагаме – т.е. когато сме взели предвид събитието B.

Прилагайки теоремата на Бейс, вероятността даден отзив да принадлежи на даден клас можем да дефинираме по следният начин:

$$P(y_k|W) = \frac{P(W|y_k)P(y_k)}{P(W)}, \text{където} \quad (4)$$

y_k - класът на документа. В нашият случай k ще заема две стойности – положителен или отрицателен клас - $k = \{1,2\}$.

W- вектор на обясняващите променливи (в нашият случай това са думите съставляващи документа). $W = (w_1, w_2 \dots w_v)$, където с *v* означаваме броя обясняващи променливи.

В Наивният Бейсов модел има едно огромно допускане, което гласи, че обясняващите променливи са абсолютно независими една от друга или по-просто казано – присъствието или липсата на дадена дума не е свързано с присъствието или липсата на друга. Това на практика не е вярно, особено когато говорим за текст. Оттам произлиза и името на модела. Това допускане за независимост можем да дефинираме по следният начин:

$$P(A \cap B) = P(A) \times P(B), \quad (5)$$

Уравнение *(5)* представлява общата вероятност за настъпване на събитията A и B, при положение, че допускаме независимост между тези две събития – в такъв случай тя е равна на тяхното произведение.

Така, използвайки това допускане, можем да представим уравнение *(4)* по следния начин:

$$P(y_k|w_1, w_2, \dots w_v) = \frac{P(w_1|y_k)P(w_2|y_k) \dots P(w_v|y_k)P(y_k)}{P(w_1)P(w_2) \dots P(w_v)} \quad (6)$$

В знаменателят сме заместили вероятността на документа с произведението от вероятностите на обясняващите променливи. Това правим именно заради допускането за независимост по време на генерирането на документа. Произведението включва вероятностите за настъпването на всички елементи (основно думи), които сме избрали да използваме за репрезентиране на отзива (това са отделните елементи на отзива). В числителят на формула *(6)* вероятността документа да е W при положение, че знаем стойността на зависимата променлива y също е заместена от произведението на отделните вероятности на всеки един елемент, съставляващ отзива, при положение, че класа y е известен.

Уравнение *(6)* може да представим и така:

$$P\left(y_k \middle| w_{1,} w_2, \dots w_v\right) = \frac{P(y_k)\prod_{i=1}^{v} P(w_i|y_k)}{P(w_1)P(w_2)\dots P(w_v)}, \qquad (7)$$

Създаването на Бейсовият класификатор се състои в превръщането на последното уравнение *(7)* в оптимизационна задача. В нея изчисляваме вероятностите на даден набор от обясняващи променливи W за всички възможни стойности на класа y и избираме варианта, при който вероятността за принадлежност на отзива към даден клас е максимална.

Тъй като знаменателят в уравнение *(7)* е константна величина, вероятността, която ще изчислим за който и да е документ, няма да се повлияе от него. Това прави израза в знаменателя излишен по време на изчислението на вероятността даден документ да принадлежи на някой от двата класа и поради тази причина не го включваме в оптимизационната задача. Така тя придобива следният вид:

$$\hat{y} = \underset{y}{\operatorname{argmax}} P(y) \prod \prod_{i=1}^{v} P(w_i|y), \qquad (8)$$

Уравнение *(8)* е общата форма на Бейсовият модел. В зависимост от допускането, което правим за разпределението на обясняващите променливи, съществуват три разновидности на модела – Полиномен, Бернулиев и Гаусов. Последният предполага

нормално разпределение на променливите, Полиномният модел е подходящ за променливи, които са представени чрез тяхната честота в извадката, а Бернулиевият модел се прилага, когато променливите следват биномно разпределение (каквото е и в нашият случай, тъй като обясняващите променливи са бинарни). Това означава, че при класификацията на текст, Полиномният Бейсов модел използва честотата на думите, доколкото Бернулиевият модел използва информация само относно тяхното присъствие или липса в текста. Това прави Полиномният модел по-малко подходящ за използване върху кратки документи.

Поради всички изброени специфики най-подходящ за конкретното изследване е Бернулиевият Наивен Бейсов модел и в настоящата задача, ще бъде използвано разпределението на Бернули, за да се апроксимира вероятностното разпределение на W по отношение на всеки клас y - $P(w_i|y_k)$.

Следващото уравнение е пример за това как ще изчислим $P(W|y_k)$ за един документ:

$$P(W|y_k) = P(w_1|y_k) \times P(w_2|y_k) \dots P(w_v|y_k), \quad (9)$$

Това е произведението от вероятностите на отделните елементи на W.

Правейки допускането за вероятностната плътност на величината W, следваща разпределение на Бернули, всяка вероятност $P(w_i|y_k)$ в това уравнение всъщност може да бъде представена като сбора от произведението на вероятността на думата спрямо класа на y_k (k заема точно две стойности) и стойността на вектора b_i, индикиращ, дали тази дума присъства или не в настоящият документ, който разглеждаме:

$$P(w_i|y_k) = b_i P(w_i|y_k) + (1 - b_i)\left(1 - P(w_i|y_k)\right) \quad (10)$$

В случай, че думата не присъства в даденият документ, горната сума ще вземе предвид това, като първият елемент в нея ще бъде равен на 0, а вторият ще бъде равен на вероятността думата да не присъства в документа (според теория на вероятностите, това получаваме като от 1 извадим вероятността елемента да присъства).

Така използвайки горното представяне, можем да изведем и формулата на Бернулиевият модел, по която изчисляваме вероятността за това даден документ да принадлежи към даден клас:

$$P(y_k|W) \propto P(W|y_k)P(y_k) \propto P(y_k)\prod_{i=1}^{v}\left[b_i P(w_i|y_k) + (1-b_i)\left(1-P(w_i|y_k)\right)\right] \quad (11)$$

y_k - класът на зависимата променлива. При бинарна класификация на настроението, заема две стойности – положителен и отрицателен.

$p(W|y_k)$ е вероятността на документът W (представен от вектор от бинарни стойности на обясняващите променливи), при условие, че знаем стойността на y_k. Тази вероятност пресмятаме чрез вероятностите на думите, съставляващи документа, както беше обяснено по-горе.

С $i = 1 \dots v$, означаваме броят обясняващи променливи, които ще използваме за анализа.

$b_i = \{0,1\}$ и означава дали дадената обясняваща променлива присъства или отсъства в дадения документ.

Въпреки слабостта на основното допускане на модела (за независимост между отделните думи), Наивният Бейсов модел класифицира добре, като това е подкрепено от факта, че се нарежда сред най-използваните алгоритми за класификация на текст. Класът на зависимата променлива, който бива избран след класификацията, обикновено се характеризира с много по-голяма вероятност от тази, която са получили другите класове. Тази вероятност може да се различава значително от истинската, но класификатора взима решение на база на това, за кой клас е изчислил най-голяма вероятност – в този смисъл по-голямо значение отдаваме на точността, с която се класифицират анализираните наблюдения, отколкото точността на оценената за всеки клас вероятност. Manning et al. пояснява много добре с думите, че правилната оценка предполага точна прогноза, но правилната прогноза не предполага точна оценка – Наивният Бейс прави не до там точни прогнозни стойности на вероятността, но класифицира добре (за още по-подробно описание и примери вижте - (Manning, Raghavan, & Schutze, 2008)).

В Наивният Бейсов модел съществува и хиперпараметър α, чиято стойност може да бъде контролирана с цел подобряване на представянето на модела. Кратко уточнение, което трябва да направим е, че хиперпараметрите се определят извън модела и не могат да бъдат изчислени от данните, а параметрите се изчисляват чрез данните.

Чрез този хиперпараметър в настоящата работа е осъществено Лапласово изглаждане на данните (*Laplace smoothing*). То е необходимо поради проблема, който възниква, когато класифицираме нов документ (който не е част от тренировъчната извадка), където има вероятност да не присъства нито една от обясняващите променливи, които сме ползвали за построяване на модела. В този случай, новият документ няма да може да бъде класифициран. Чрез изглаждането целим да избегнем такива ситуации. В своето изследване на този проблем и как той се отразява върху резултатите от класификацията Indriani et al. (Indriani & Nugrahadi, 2016, October) показват, че сред изпробваните техники за изглаждане, най-добри резултати са постигнати с Лапласово изглаждане на данните. Алтернативни методи са: изглаждане на Дирихле (*Dirichlet smoothing*), Абсолютно намаляване (*Absolute Discounting*), метода на Jelinek-Mercer (*Jelinek-Mercer smoothing*) - (Indriani & Nugrahadi, 2016, October), (Zhai & Lafferty, 2004).

Изпробвани са различни стойности за α и е избрана тази, при която има най-голямо подобрение в представянето на модела.

Използвана е решетка от предефинирани стойности за α параметъра (*grid search*). Стойностите, които ще бъдат тествани, са избрани на база на това, какво се препоръчва в литературата - (Indriani & Nugrahadi, 2016, October), (Zhai & Lafferty, 2004). За да бъде избрана стойността на хиперпараметъра, с която модела се представя най-добре, е използвана валидационната извадка в изследването, като процеса по осъществяването на тази валидация е представен детайлно в точка 2.5. “Валидация и оценка на представянето на моделите” в настоящата работа.

2.4.2. Логистична регресия

Логистичната регресия представлява дискриминативен класификатор, в който зависимата променлива заема точно две стойности. В нашият случай – положителният и отрицателният клас.

Функционална форма на модела (Friedman, Hastie, & Tibshirani, 2010):

$$\log\frac{P(Y=1|w)}{P(Y=0|w)} = \log\frac{P(Y=1|w)}{1-P(Y=1|w)} = \beta_0 + \sum_{i=1}^{V} w_i\beta_i\,,\ \beta_i \in R, \text{където} \qquad (12)$$

w_i – обясняващи променливи в модела, $w_i \in R$, $i = 1 \dots V$

Y – зависима променлива в модела, $Y = \{0,1\}$

β_i – параметри на модела, които търсим, така че полученият модел възможно най-добре да описва данните, $i = 1 \dots V$

Последната формула *(12)* може да бъде трансформирана така, че директно да изразява вероятността даден документ да принадлежи към клас 1:

$$p(Y=1|W) = \frac{e^{\beta_0+\sum_{i=1}^{V} w_i\beta_i}}{1+e^{\beta_0+\sum_{i=1}^{V} w_i\beta_i}} \qquad (13)$$

Логистичната регресия също притежава хиперпараметър, който може да бъде използван с цел подобряване на представянето на модела. Чрез него можем да контролираме силата на регуляризацията на логистичната регресия (*regularized logistic regression*). Просто обяснение на регуляризацията в логистичната регресия е, че тя представлява метод за намаляване на сложността на модела (Shai & Shai, 2014). Използвайки тази техника, ние се предпазваме от ситуация на прекомерно нагаждане към тренировъчната извадка.

Изпробвани са различни стойности на регуляризиращият хиперпараметър (отново чрез решетка от предефинирани стойности, избрани на база на това, какво се препоръчва в научната литература, засягаща тази тематика - (Zhu & Hastie, 2004)) на логистичната регресия и е избрана тази, при която има най-голямо подобрение в представянето на модела върху валидационната извадка - (Pedregosa, et al., 2011), (Friedman, Hastie, & Tibshirani, 2010).

2.4.3. Метод на опорните вектори

Методът на опорните вектори представлява група от алгоритми на машинното самообучение с учител, които могат да бъдат използвани, както за класификация, така и за регресионен анализ.

Проблемът в настоящото изследване е класификационен, и поради това, е използван класификатор с опорни вектори (*support vector classifier*). По-конкретно - приложен е линеен класификатор, тъй като при наличието на много обясняващи променливи, неговото използване се препоръчва (Hsu, Chang, & Lin, 2003). Други алтернативи на метода на опорните вектори за класификация са – метод с радиално-базисна функция (*radial basis function*), метод с полиномна функция (*polynomial function*) и метод със сигмоидна функция (*sigmoid function*).

Линейният класификатор с опорни вектори решава следната оптимизационна задача (Hsu, Chang, & Lin, 2003):

$$\min_{x,b,\xi} \frac{1}{2} x^T x + C \sum_{i=[1,|V|]} \xi_i \qquad (14)$$

При условие, че:

$$y_i(x^T w_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \text{ и } \forall i \in [1, |V|] \qquad (15)$$

Означение:

x – вектор от тежести (коефициенти) на обясняващите променливи;

w_i – вектор от обясняващите променливи (в случая на бинарна репрезентация на текста този вектор е съставен от 0 и 1);

y – зависима променлива; в случай на бинарна класификация - $y_i = \{-1,1\}$

V – тренировъчните данни (вектор от обясняващи променливи + етикет); $V = \{(w_i, y_i)\}$, $i = 1 \dots |V|$

b – отклонение (константна величина)

C – регуляризиращ хиперпараметър на грешката (*penalty parameter*)

ξ_i – параметър, означаващ допустимата грешка на всяко наблюдение

Както става ясно от определението на модела, този алгоритъм също притежава хиперпараметри, които могат да бъдат контролирани с цел подобряване на представянето му. В настоящата работа са изпробвани различни стойности на хиперпараметъра C, който има регуляризираща функция (както и в случаят на логистичната регресия) - (Hastie, Rosset, Tibshirani, & Zhu, 2004). Колкото по-голяма е стойността му, толкова по-голяма е регуляризацията на грешката и обратното. Чрез този хиперпараметър контролираме степента на нагаждане на модела към данните. Отново използваме решетка от предефинирани стойности за хиперпараметъра C, за да открием тази, за която получаваме най-добро представяне. И в този случай стойностите са избрани на база на това, какво се препоръчва в научната литература - (Hsu, Chang, & Lin, 2003), (Hastie, Rosset, Tibshirani, & Zhu, 2004).

В края на тази секция е важно да спомена, че за имплементирането на всеки един от алгоритмите, търсенето на хиперпараметри и осъществяването на регуляризация, е използвана библиотеката “Scikit-learn” в Python (Pedregosa, et al., 2011). В ръководството за използване на библиотеката могат да бъдат открити подробни пояснения и формални дефиниции на всички техники, залегнали в основата на функциите, използвани за създаване на моделите и тяхното настройване.[20]

2.5. Валидация и оценка на представянето на моделите

Основната идея на валидацията, когато моделираме данни, е да гарантираме устойчивостта на създадения модел и да оценим неговото истинско представяне, когато бъде използван в реална среда. За тази цел е най-добре той да бъде валидиран върху данни, различни от тези, с които е създаден. Това можем да направим, предварително разделяйки извадката на три части – тренировъчна, валидационна и тестова извадка (Shai & Shai, 2014). Тренировъчната бива използвана за обучение на модела. Валидационната бива използвана за настройване на модела и селектиране на най-добрият от всички тествани. Тестовата бива използвана за валидация и оценка на представянето в реална

[20] User Guide // Scikit-learn. 2007. [cited 20.10.2017] Available from: http://scikit-learn.org/stable/user_guide.html

среда – получаваме оценка за действителната грешка на модела (*true error*). Разбира се, когато големината на цялата извадка, с която разполагаме, не позволява отделянето ѝ на няколко части, може да се прибегне към алтернативни методи, като например крос-валидация (*cross validation*) - (Shai & Shai, 2014).

В настоящата работа цялата извадка с отзиви е разделена на три части, според описаният по-горе подход, тъй като разполагаме с достатъчно данни за това.

Тренировъчната извадка е използвана само по време на обучението на моделите. Валидационната извадка е използвана, за да може да бъдат избрани стойностите на хиперпараметрите и броят обясняващи променливи, с които различните класификационни алгоритми постигат най-добро представяне. Тестовата извадка е използвана само при валидация на крайните модели, които са били селектирани чрез валидационната. Именно, при тази последна валидация, получаваме най-реална оценка на представянето на обучените модели, тъй като те биват прилагани върху данни, които никога преди това не са виждали (*unseen data*) – тези данни не са участвали по никакъв начин в създаването на моделите.

За валидация и оценка на представянето на моделите са използвани подходящи за това метрики, а именно:

⇒ Точност на модела (*accuracy*);
⇒ Прецизност на модела (*precision*);
⇒ Чувствителност на модела (*recall*);
⇒ F1-мярка (*f1-measure*).

Следва формална дефиниция на всяка една от тези метрики, като преди това ще въведем и още едно понятие, което ще ни помогне за дефинирането на тези оценителни статистики, а именно “матрица на грешките” (*confusion matrix/error matrix*).

Матрицата на грешките представя резултатите от бинарна класификация по следният начин (Sammut & Webb, 2011):

Таблица 1. Матрица на грешките – дефиниция

	Прогнозна стойност на зависимата променлива		
	Общ брой примери в извадката	**Класифицирани като клас 1 (*positive class*)**	**Класифицирани като клас 2 (*negative class*)**
Истинска стойност на зависимата променлива	**Примери от клас 1**	Предсказани правилно от клас 1 (true positive)	Предсказани грешно от клас 1 (false negative)
	Примери от клас 2	Предсказани грешно от клас 2 (false positive)	Предсказани правилно от клас 2 (true negative)

Както става ясно, ***Таблица 1*** ни предоставя детайлна разбивка на резултатите от процеса на класификация на данните. Виждаме четири измерителя (true positive, true negative, false positive, false negative), които ни дават информация за броя предсказани правилно и броя предсказани грешно случаи от всеки клас на зависимата променлива.

Точността на модела (Sammut & Webb, 2011) се изчислява най-просто и представлява отношението между правилно класифицираните примери и общият брой примери в извадката, използвана за валидация (валидационна или тестова):

$$\text{Точност} = \frac{\text{Предсказани правилно от клас 1} \ + \ \text{Предсказани правилно от клас 2}}{\text{Общ брой примери в извадката}}, \qquad (16)$$

Точността не ни дава информация относно представянето на даденият алгоритъм поотделно за всеки от двата класа, а ни дава обща оценка. Поради тази причина в процеса по оценяване, взимаме предвид и останалите оценки – прецизност, чувствителност и F1-мярката. Те биват изчислявани поотделно за всеки един от класовете на зависимата променлива.

Прецизността на класификацията (Sammut & Webb, 2011) е оценка на това до колко примерите, класифицирани като част от даден клас, наистина принадлежат на него. Изчисляваме я по следната формула[21]:

[21] Показаният пример е за изчисляване на прецизността на класификатора за клас 1. Същата формула прилагаме и за другият клас на зависимата променлива.

$$\text{Прецизност} = \frac{\text{Предсказани правилно от клас 1}}{\text{Предсказани правилно от клас 1} + \text{Предсказани грешно от клас 2}} \quad (17)$$

Чувствителността на класификатора (Sammut & Webb, 2011) е оценка на това, колко от примерите в даден клас са били класифицирани правилно:

$$\text{Чувствителност} = \frac{\text{Предсказани правилно от клас 1}}{\text{Предсказани правилно от клас 1} + \text{Предсказани грешно от клас 1}} \quad (18)$$

Тези две мерки са така дефинирани, че когато опитваме да увеличим стойността на едната, другата намалява и обратното. В зависимост от класификационната задача, която разрешаваме, е възможно да искаме да избегнем допускането на единият тип грешка, доколкото другият тип грешка да не е с особена важност при разрешаването на конкретната задача. Има и случаи, в които по-скоро бихме искали да вземем предвид и двете мерки, какъвто е и случаят в настоящата работа. Поради тази причина използваме и още една оценителна мярка, която комбинира прецизността и чувствителността. Това е F1-мярката.

F1-мярката представлява средната хармонична стойност на прецизността и чувствителността на модела (Sammut & Webb, 2011). Тя поставя една и съща тежест и върху двете, и се изчислява по следният начин:

$$\text{F1 мярка} = 2\frac{\text{Прецизност} \times \text{Чувствителност}}{\text{Прецизност} + \text{Чувствителност}} \quad (19)$$

Метриката е подходяща именно в случаи, когато искаме да вземем предвид и двете статистики, измерващи грешката на модела.

Последно ще дефинираме и една разновидност на F1-мярката, която е използвана в анализа, а именно макро F1-мярката. Тя представлява средната стойност на F1-мерките, изчислени върху положителния и отрицателния клас поотделно. Ето и простата формула, по която тя бива изчислена:

$$\text{Макро F1 мярка} = \frac{\text{F1 мярка}_{pos} + \text{F1 мярка}_{neg}}{2} \quad (20)$$

Този макро-подход може да бъде приложен и върху останалите по-рано дефинирани метрики и е полезен, тъй като ни позволява да обединим информацията и оценим цялостното представяне на всички класове чрез различните статистики.

III. Емпирично изследване[22]

До този момент в настоящата работа беше направен литературен преглед на основни изследвания в сферата на анализа на чувства и настроения и извличането на знания от текст. Също така беше предложена методология за създаването на система за разпознаване на полярността на настроението в потребителски коментари. В настоящата глава следва да приложим тази методология и на база на емпиричен анализ да проверим хипотезите, залегнали в основата на нашето изследване.

3.1. Данни

Както по-рано в настоящата работа споменахме, данните използвани за емпиричното изследване са потребителски отзиви за мобилни приложения, извлечени от Google Play[23]. Те са публично достъпни и са предоставени от Grano et al. в тяхното изследване (по-рано споменато в глава I -"Литературен преглед") на обратната връзка, която потребителите на мобилни приложения предоставят в своите отзиви и как тя може да бъде използвана за подобряване и усъвършенстване на разработката на софтуер (Grano, et al., 2017, September). Данните са предоставени в табличен вид (в csv формат), което улеснява значително последващата им обработка в настоящата работа.

Авторите на това изследване създават система за извличане на данни от интернет (*web crawler*), чрез която изтеглят от Google Play Store потребителските отзиви за 395 различни мобилни приложения за операционната система Android. Чрез този инструмент за извличане на данни е изтеглена следната информация от Google Play Store:

⇒ Текста с потребителският отзив;

⇒ Името на приложението, за което се отнася даденият отзив (*Android package name*);

[22] При желание от страна на рецензента на настоящата разработка, всички скриптове, чрез които е създаден експеримента могат да бъдат предоставени. Те са достъпни в облачното пространство Google Drive, като при поискване от негова страна на следният имейл адрес: Gloria_hristova@yahoo.com , веднага ще му бъде изпратен линк за достъп до тях.

[23] Google play // Play google. 06.03.2012. [cited 22.09.2017] Available from: https://play.google.com/store/apps

⇒ Рейтинга, който е поставил потребителя на приложението (по 5-звездна рейтингова система);

⇒ Датата, на която е публикуван отзива.

На база на това, кога е датата на излизане на всяка една версия на приложенията и датата на публикация на потребителските отзиви, авторите създават и колона с информация за това, коя е версията на мобилното приложение, за която се отнася даденият отзив.

Данните, предоставени от Grano et al. (Grano, et al., 2017, September) се състоят от **288 065** потребителски коментара за **629** версии на **395** мобилни приложения за Android. Публикувани са в Google Play в периода от **01.01.2014 до 02.05.2017 г.**

Таблица 2 показва първоначалният вид на данните (в csv формат). Важно е да се отбележи, че за целите на настоящата работа и анализ, са използвани само данните за текста на потребителският отзив, оценката на потребителя и датата на която е публикуван отзива.

Таблица 2. Първоначален вид на данните

id	package_name	review	date	star	version_id
7bd227d9-afc9-11e6-aba1-c4b301cdf627	com.mantz_it.rfanalyzer	Great app! The new version now works on my Bravia Android TV which is great as it's right by my rooftop aerial cable. The scan feature would be useful...any ETA on when this will be available? Also the option to import a list of bookmarks e.g. from a simple properties file would be useful.	October 12 2016	4	1487
7bd22905-afc9-11e6-a5dc-c4b301cdf627	com.mantz_it.rfanalyzer	Great It's not fully optimised and has some issues with crashing but still a nice app especially considering the price and it's open source.	August 23 2016	4	1487
7bd2299c-afc9-11e6-85d6-c4b301cdf627	com.mantz_it.rfanalyzer	Works on a Nexus 6p I'm still messing around with my hackrf but it works with my Nexus 6p Trond usb-c to usb host adapter. Thanks!	August 04 2016	5	1487
7bd22a26-afc9-11e6-9309-c4b301cdf627	com.mantz_it.rfanalyzer	The bandwidth seemed to be limited to maximum 2 MHz or so. I tried to increase the bandwidth but not possible. I purchased this is because one of the pictures in the advertisement showed the 2.4GHz band with around 10MHz or more bandwidth. Is it not possible to increase the bandwidth? If not it is just the same performance as other free APPs.	July 25 2016	3	1487
7bd22aba-afc9-11e6-8293-c4b301cdf627	com.mantz_it.rfanalyzer	Works well with my Hackrf Hopefully new updates will arrive for extra functions	July 22 2016	5	1487

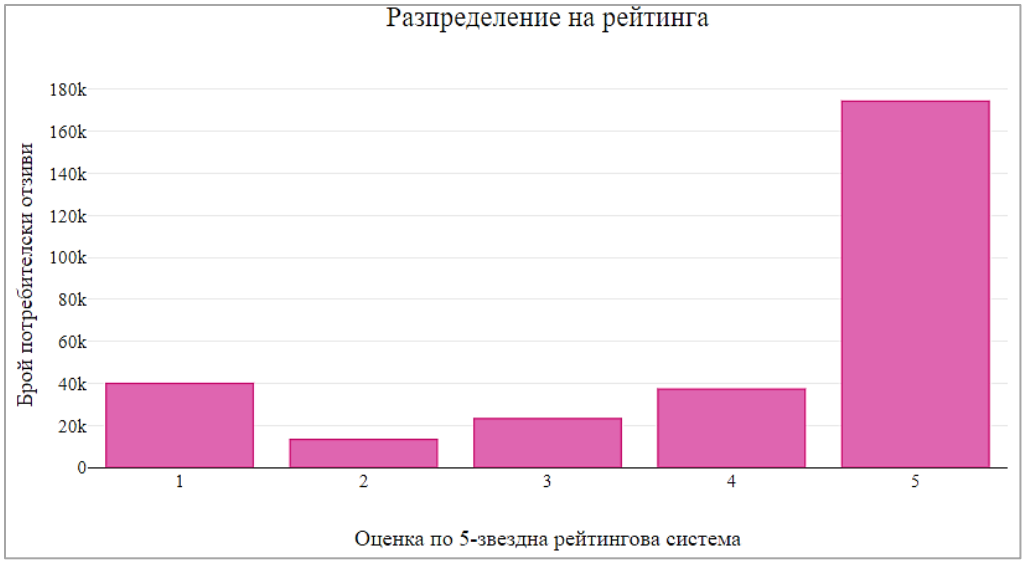
Авторите на изследването предоставят информация и относно това, за какви категории приложения са извлечените потребителски отзиви. **Таблица 9** представя тази информация в Приложенията към настоящата работа. В данните присъстват отзиви за 24 различни категории мобилни приложения, като най-много коментари има за категория “Инструменти”, в която попадат разнообразни софтуери, улесняващи работата на потребителите на смарт устройства – антивирусни програми, програми за създаване и обработка на документи, фенерчета, емоджи клавиатури, инструменти за дистанционно

управление, четци на баркодове и други. От **Таблица 9** става ясно, че данните покриват широк набор от мобилни приложения, притежаващи различни предназначения и функционалности. Това може да се счете като положително за настоящият анализ, тъй като означава, че данните се характеризират с по-разнообразна лексика и ще ни предоставят по-просторен поглед в разглежданият домейн – потребителите може да срещат различни проблеми или предимства на дадени приложения и да менят изказа и думите, които използват в своите отзиви за тях.

3.2. Първоначален вид на данните

След основното описание на данните в предната точка, следва по-подробно запознаване с тях.

В глава II вече пояснихме как ще бъде създадена зависимата променлива в анализа, така че той да може да бъде разрешен чрез методи на машинното самообучение с учител. За целта използваме рейтинга, който потребителите посочват към своят коментар за дадено мобилно приложение. На **Фигура 3** виждаме хистограма на разпределението на този рейтинг в цялата извадка. Става ясно, че най-често потребителите дават най-силно положителната оценка (5 звезди), а най-рядко е била поставяна оценка от 2 звезди. Коментарите с пет звезди представляват около 60% от данните – очевидно потребителите са по-склонни да дават положителни оценки. На база на литературният преглед, направен в настоящото изследване, можем да кажем, че това наблюдение се среща в изследванията на много автори, занимаващи се с анализ на потребителски отзиви в сайтове с рейтингови системи, като това е без значение от конкретната сфера, за която се отнасят отзивите - (Fang & Zhan, 2015), (Cui, Mittal, & Datar, 2006, July) и други.



Фигура 3. Хистограма на разпределението на рейтинга в цялата извадка от отзиви

На **Фигура 10** в Приложенията можем да разгледаме и разпределението на средната стойност на рейтинга, присъден от потребителите по време на целият времеви период, който обхваща извадката (Януари 2014 – Май 2015). На графиката се вижда, че на месечна база този рейтинг е около 4 звезди, като няма значителни промени в стойността му.

3.3. Характеристики на извадката

В тази точка ще представим основните особености на извадката, използвана за разработването на система за разпознаване на настроението на потребителя. Още в глава II става ясно, че в този процес няма да участват всички данни с които разполагаме, а част от тях – в тази точка ще разгледаме по-подробно конкретно техните специфики.

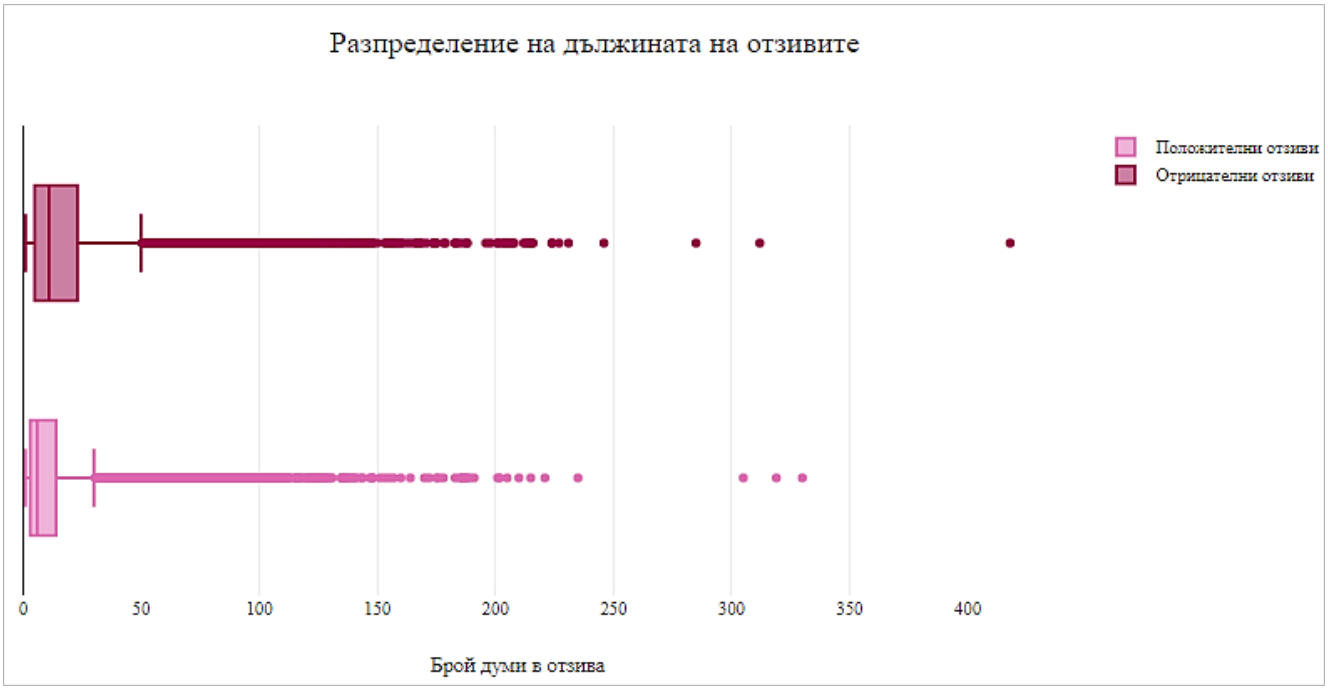
Разглежданият проблем в нашето изследване се състои основно в това, да се определи настроението, което носи даден отзив, като положително или отрицателно. В настоящият анализ, отрицателния клас на зависимата променлива е създаден от отзиви с една и две звезди, а положителния – само от отзиви с пет звезди (те са достатъчно на брой, така че да не се налага включването на такива с четири звезди). Както беше споменато по-рано в глава II, целта е по този начин да има по-голямо разграничение между двата класа, за да може да се увеличи дискриминиращата способност на алгоритмите, с които ще обучим модела за разпознаване на настроението. Чрез отзивите с тези видове оценка, на случаен принцип създаваме “балансирана“ извадка (Chawla, Japkowicz, & Kotcz, 2004) –

положителният и отрицателният клас са с еднакъв брой наблюдения. Така извадката в анализа се състои от общо **96 000** потребителски отзива – **48 000** положителни и **48 000** отрицателни.

Една от интересните особености на данните, която забелязваме е, че отрицателните отзиви се характеризират с по-голяма дължина (измерена в брой думи - без да включваме символи и цифри в това число) в сравнение с положителните. Това означава, че потребителите на дадено приложение са склонни да бъдат по-описателни в своите отзиви, когато изразяват недоволство (макар и изразяването на недоволство да се случва значително по-рядко, както забелязахме по-рано). Също така, това наблюдение означава, че може би набора от думи, които индикират отрицателно мнение е по-богат, в сравнение с този за положителни мнения. Последното ни допускане всъщност е установено и подкрепено с доказателства в изследването на Hoon et al. (Hoon, Vasa, Schneider, & Mouzakis, 2012, November).

Фигура 4 представя разпределението на броят думи в отзив, в зависимост от класа, който разглеждаме – от нея ясно си личи, че отрицателните отзиви се характеризират с по-голяма дължина. **Таблица 10** в Приложенията ни предоставя по-детайлна информация за описателните статистики, свързани с дължината на отзивите, измерена в брой думи (информацията е представена поотделно за всеки клас).

В Приложенията, **Фигура 11** ни показва разпределението на дължината на всички отзиви (без да ги разделяме спрямо емоцията). От нея става ясно, че над 75% от цялата извадка се характеризира с отзиви, не по-дълги от 25 думи. От **Таблица 10** пък, директно откриваме, че до 75-ти персентил, дължината на отзивите в положителният клас е в диапазона между 1 и 14 думи, а в отрицателният – между 1 и 23. Тази краткост на изказа не би трябвало да ни изненадва, тъй като е напълно типична за данни, част от подобна онлайн платформа за споделяне на мнение.



Фигура 4. Разпределение на броят думи в отзив, в зависимост от класа, който разглеждаме

3.4. Обработка на текста

Преди пристъпването към по-задълбочен анализ и прилагането на алгоритми от машинното самообучение върху данните, е необходимо те да бъдат приведени в числов вид. Процеса по превръщането им в такъв е подробно описан в глава II (“Методология”), поради което в тази част няма да пояснявам отново всяка стъпка, а само ще демонстрираме как от суровия вид на един отзив в нашата извадка се получава низ от думи, които след това ще бъдат използвани за избор на обясняващи променливи и спрямо това отзивите ще бъдат превърнати в бинарни вектори.

Пример:

Това е първоначалният (суров) вид на един от положителните отзиви в нашата извадка:

“Best Gallery in the Android Universe! AWESOME Improvements VERY NICE!Fast light-weight pack with the right amount of features for a great gallery.No ads and pop ups!1.06 mb gallery is real! ☺”

Следвайки описаната методология в глава II, първата стъпка в неговата обработка е да премахнем html таговете, ако такива са попаднали в текста. Тъй като в този отзив не съществуват подобни следи, останали след извличането на данните от интернет, той няма да промени вида си на този етап. Следващата стъпка ще бъде да извлечем част от обясняващите променливи, описващи текста и различни от думите в него. Това правим преди изчистването на препинателните знаци, тъй като тези допълнителни променливи са свързани именно с тях и ще изгубим информация, ако ги премахнем на този етап.

В този случай в настоящият отзив съществуват две от допълнително създадените променливи в анализа – емотикона от групата на изразяващите щастие и наличие на възклицание. Тези символи в отзива ще бъдат заменени със съответните тоукъни, индикиращи наличието на тези две променливи. Ето как ще изглежда отзива след промяната:

“Best Gallery in the Android Universe! AWESOME Improvements VERY NICE!Fast light-weight pack with the right amount of features for a great gallery.No ads and pop ups!1.06 mb gallery is real! smileyhappy hasexclamation”

От примера става ясно, че емотиконата е заменена с текст, който обозначава наличието ѝ, както и дава информация за това към коя група от емотикони принадлежи (изразяващи щастие). Със следващият тоукън е обозначено присъствието на удивителен знак в отзива (броят удивителни знаци не се взима предвид). Тази стъпка в обработката може да бъде пропусната в случаят, когато работим само с думите в текста и не извличаме друга информация от него.

На този етап вече можем да изчистим отзива от всички пунктуационни знаци, специални знаци и цифри. Ето неговият вид след тази стъпка:

“Best Gallery in the Android Universe AWESOME Improvements VERY NICE Fast light weight pack with the right amount of features for a great gallery No ads and pop ups mb gallery is real smileyhappy hasexclamation”

Както се вижда, всички тези знаци са заместени с интервал в данните. До този момент отзивът все още е под формата на цял текст. Следващата стъпка е неговата тоукънизация, чрез която ще го превърнем в низ от думи (тоукъни). Ето неговият вид в Python след тази стъпка:

['Best', 'Gallery', 'in', 'the', 'Android', 'Universe', 'AWESOME', 'Improvements', 'VERY', 'NICE', 'Fast', 'light', 'weight', 'pack', 'with', 'the', 'right', 'amount', 'of', 'features', 'for', 'a', 'great', 'gallery', 'No', 'ads', 'and', 'pop', 'ups', 'mb', 'gallery', 'is', 'real', 'smileyhappy', 'hasexclamation']"

Следва да бъдат премахнати стоп думите в отзива:

['Best', 'Gallery', 'Android', 'Universe', 'AWESOME', 'Improvements', 'NICE', 'Fast', 'light', 'weight', 'pack', 'right', 'amount', 'features', 'great', 'gallery', 'No', 'ads', 'pop', 'ups', 'mb', 'gallery', 'real', 'smileyhappy', 'hasexclamation']"

Виждаме, че последната стъпка премахва успешно думите, които не носят никакво значение и информация относно настроението в текста.

Следва нормализация на текста – всички тоукъни трябва да са изписани с малки букви:

['best', 'gallery', 'android', 'universe', 'awesome', 'improvements', 'nice, 'fast', 'light', 'weight', 'pack', 'right', 'amount', 'features', 'great', 'gallery', 'no', 'ads', 'pop', 'ups', 'mb', 'gallery', 'real', 'smileyhappy', 'hasexclamation']

Следваща стъпка е да се обработят думите, изкуствено удължени чрез повторение на букви. В настоящият пример няма такива и поради това видът му се запазва на този етап.

Следва да се осъществи лематизация на текста. Ето видът му след прилагането на тази техника:

['best', 'gallery', 'android', 'universe', 'awesome', 'improvement', 'nice, 'fast', 'light', 'weight', 'pack', 'right', 'amount', 'feature', 'great', 'gallery', 'no', 'ad', 'pop', 'ups', 'mb', 'gallery', 'real', 'smileyhappy', 'hasexclamation']"

Примерът ясно показва как лематизацията успешно се е справила в премахването на множественото число на съществителните имена в отзива.

Последно, отзивите биват изчистени от думи, състоящи се от една буква (възможно е наличието на единични букви в отзивите без никакъв смисъл, тъй като данните са далеч от чисти и често може да се срещнат такива неточности и проблеми в тях). Тъй като в настоящият пример не съществуват такива, той запазва своя вид.

По този начин бива осъществена предварителната обработка на данните, така че те да бъдат в по-подходящ вид за подбора на обясняващи променливи, прилагането на векторно-пространственият модел и превръщането на отзивите в бинарни вектори, индикиращи присъствието или липсата на избраните обясняващи променливи.

Последно в тази точка, разясняваща обработката на текста, би било добре да поясним кои са групите от емотикони, създадени чрез данните. Както беше описано по-рано в глава "Методология", от отзивите първо биват извлечени всички използвани емотикони с помощта на специална библиотека в Python, помагаща в тази задача. След това чрез подробният списък на значението им, с който разполагаме, те биват разпределени в няколко групи, като всяка от тях носи определена емоция. В нашите данни открихме **95** уникални емотикони, като ги разпределихме в **10** различни групи, а именно изразяващи чувство на: щастие, смях, любов, изненада, нещастие, гняв, скептичност, намигване, притеснение и закачка. Те образуват **10 допълнителни обясняващи променливи** в анализа, които са пряко обвързани с емоционалната натовареност, която той носи. Поради това, очакването ни е, че те ще бъдат силно дискриминативни.

3.5. Визуализации на най-често използваните думи

След обработката на данните и привеждането им във вид подходящ за по-нататъшен анализ, е добре да се запознаем с това кои са най-често използваните думи в положителните и отрицателните отзиви. Това знание ще ни помогне за разрешаването на поне три задачи:

- ✓ Проверка на резултата след всички приложени трансформации - ако има съществени проблеми в обработката и грешки, незабелязани до момента, на тази стъпка е възможно да бъдат уловени;
- ✓ По-дълбоко опознаване на данните и проверка дали това, което наблюдаваме съвпада с нашите очаквания;
- ✓ Създаване на предположение относно това кои обясняващи променливи ще имат по-голяма предвиждаща способност и съответно ще бъдат допуснати в разработката на модела след подбора им, който е следващата стъпка от анализа.

Тази част на изследването е осъществена основно чрез помощта на визуализации под формата на облаци от думи (*word clouds*), създадени чрез програмният език Python и

библиотеката “wordcloud”. Това е често срещана техника в подобен род изследвания и е доста подходяща, когато анализираните данни са под формата на текст - (Hofmann & Chisholm, 2016), (Coelho & Richert, 2015). Големината на думите в облака отразява честотата на думите в анализираният текст.[24]

На **Фигура 5** можем да видим облак, построен от най-често срещаните думи в положителни отзиви. Важно е да отбележим, че наблюденията ни са сходни с тези на Hoon et al. (Hoon, Vasa, Schneider, & Mouzakis, 2012, November), които също предоставят детайлен поглед върху най-често ползваните думи в отзиви за мобилни приложения с 5 звезди (с разликата, че данните в тяхното изследване са извлечени от iOS App Store). Съществена част от топ думите, които те посочват, съвпада с тези, открити в нашите данни – “awesome”,”best”,”easy”,”amazing”, “thanks”, "useful”, “simple”.

Фигура 5. Най-често срещани униграми в положителни отзиви

В Приложенията може да бъде открита по-подробна графика и на 20-те най-често срещани униграми в положителни отзиви, даваща информация за техният брой - **Фигура 12**.

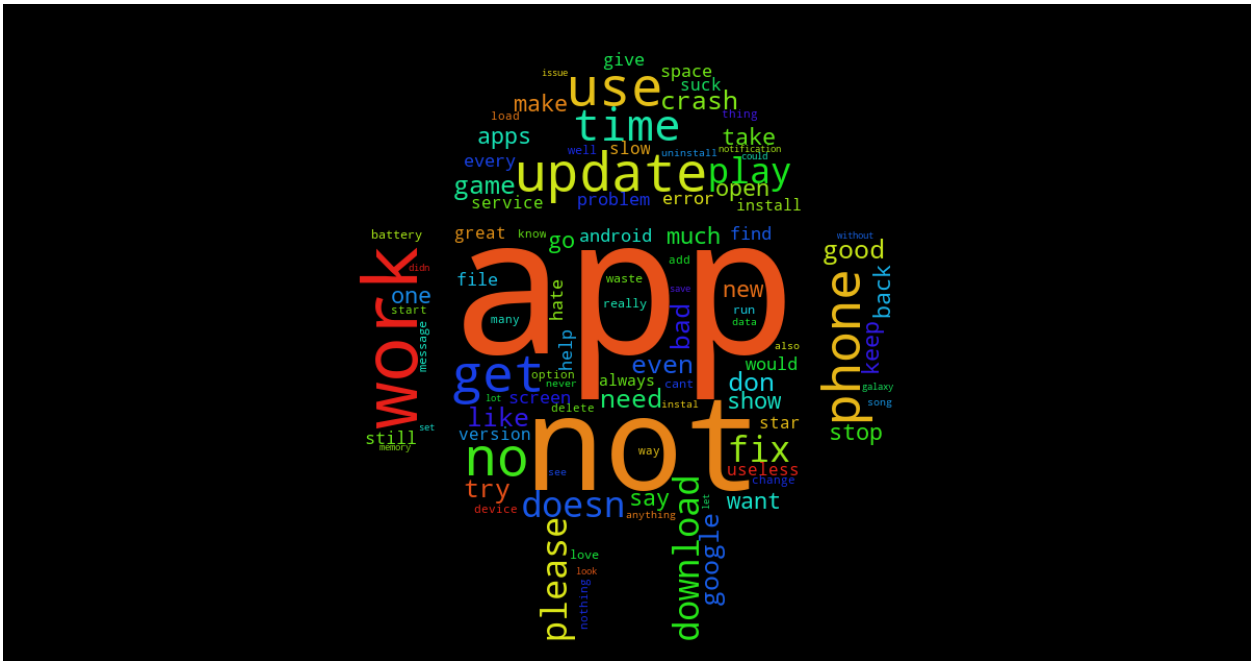
На **Фигура 5** ясно се вижда, че преобладаващата част от униграмите в положителни отзиви са прилагателни или глаголи, изразяващи удовлетворението на

[24]В зависимост от изследването е възможно вместо честота да бъде използван и друг измерител за важността на думите.

потребителя – “good”, “love”, ”great”, “best”, “simple”, “work”, “thank”, ”use” и други. Също така се появяват и думи, които са по-скоро свързани с контекста и средата, поради което са много често срещани в целия набор от данни, но не носят емоция. Такива са – “app”, ”google”, “play“. Тези думи се очаква да бъдат често срещани и в отрицателният клас отзиви и съответно - да имат по-малка предвиждаща способност, тъй като чрез тях трудно би могло да се разграничи настроението в текста (имайки предвид факта, че са характерни и за двата типа отзиви). Разбира се, срещаме и думи, чието наличие в положителни отзиви на пръв поглед изглежда противоречиво – “help”, “problem”, “please”. Има множество варианти за тълкуване на присъствието им. Без да четем отзивите, бихме могли да предположим, че става въпрос за разрешени проблеми на потребителите чрез нова версия на дадено приложение. Също така това може да е сигнал за потребители, които като цяло харесват даденото мобилно приложение, но съобщават, че има неща, които могат да се подобрят в него. Възможно е и да съществува несъответствие между рейтинга и текста, което би било пречка по време на процеса на моделиране.

В Приложенията може да бъде открит облак от думи, създаден чрез най-често срещаните биграми в позитивните отзиви - **Фигура 13**. Той ни помага да отговорим на някои от възникналите въпроси след разглеждането на облака с позитивни униграми по-горе. Вижда се, че негативната дума “problem” всъщност доста често се е срещала в съчетание с думата “no”. Това е пример за използване на отрицателни частици (“no”), сменящи настроението в текста – целият израз всъщност е положителен. Откриваме и думата “please” в съчетание с “add”, което показва, че допускането ни за това, че потребителя харесва приложението и дава висока оценка, но моли за добавянето на подобрения, е вярно.

Следва да разгледаме по същият начин и отрицателният клас от отзиви. **Фигура 6** представлява облак, показващ ни най-често срещаните думи в отрицателни отзиви. Отново забелязваме съвпадение с резултатите на Hoon et al. (Hoon, Vasa, Schneider, & Mouzakis, 2012, November). Топ негативните думи, които те посочват съвпадат с част от тези, които откриваме в нашето изследване – “crashes”, “fix”, “bad”, “doesn`t”, ”download”, “useless”.



Фигура 6. Най-често срещани униграми в отрицателни отзиви

Веднага можем да забележим, че в облака на **Фигура 6** основно се срещат думи, изразяващи някакво недоволство или проблем – “crash”, “problem”, “error”, “useless”, “uninstall”, “hate” и т.н. Доста използвани са отрицателни частички и изрази, като – “not”, “no”, ”doesn”. Тук отново можем да констатираме наличието на същите често срещани думи, които са обвързани с целият контекст, не носят емоция и бяха открити и в положителния облак от униграми. Също така забелязваме и присъствието на няколко думи, говорещи за положителна емоция – “great”, “love”, “good”. Това явление можем да си обясним с няколко причини. Първата и най-вероятна причина е, че тези думи просто са част от изрази, които изказват недоволство чрез отрицателни частици – “not good”, “not great”, “no love”, “doesn work” и т.н. Това обаче не е уловено в тези визуализации, тъй като те са създадени само от най-често срещаните униграми. Друга възможна причина е, че потребителите може да използват ирония в своите коментари, изразявайки недоволството си чрез използването на положителни думи. Друга причина би могло да бъде поставянето на ниска оценка поради конкретен проблем, въпреки че като цяло приложението се харесва от потребителя. Всички изказани предположения със сигурност могат да доведат до затрудняване на построяването на модел за разпознаване на емоцията, тъй като донякъде ще се загуби дискриминиращата способност на някои думи.

Тук ще спрем, за да обърнем малко повече внимание на първата от горепосочените причини – използването на отрицание, тъй като тя представлява важен елемент, когато говорим за анализ на текст, и по-специално – анализ на настроението в него. За да стане напълно ясно защо, можем да дадем малък пример с думата "good". Ако тя е част от избраните обясняващи променливи в текста, то нейното присъствие ще бъде индикирано във всеки един отзив, в който тя участва. Разбира се, ние асоциираме тази дума с положителна емоция и най-вероятно тя наистина ще преобладава в положителни отзиви. За жалост обаче, ако в дадената извадка потребителите, изразяващи негативно мнение, често са използвали отрицателни частици в своите изказвания в комбинация с тази дума, като например израза "not good" – унограмата "good" ще се среща често, както в единият клас, така и в другият клас на класификационната задача. Това автоматично довежда до загуба на нейната дискриминативна способност.

"Отрицанието" (*negated expressions*) може да се определи, като граматическа категория, която позволява промяната на истинския смисъл на дадено твърдение. Отрицанието обикновено бива изразено чрез думи, като "not" и "never" и може значително да повлияе върху настроението, което носи текста (Cambria, Das, Bandyopadhyay, & Feraco, 2017). Съществуват техники за справянето с този проблем, който (както споменахме вече) може да доведе до загуба на дискриминираща способност у някои думи. Използването на такива методи би било една възможност за подобрение на настоящия анализ (Pang & Lee, Opinion mining and sentiment analysis, 2008).

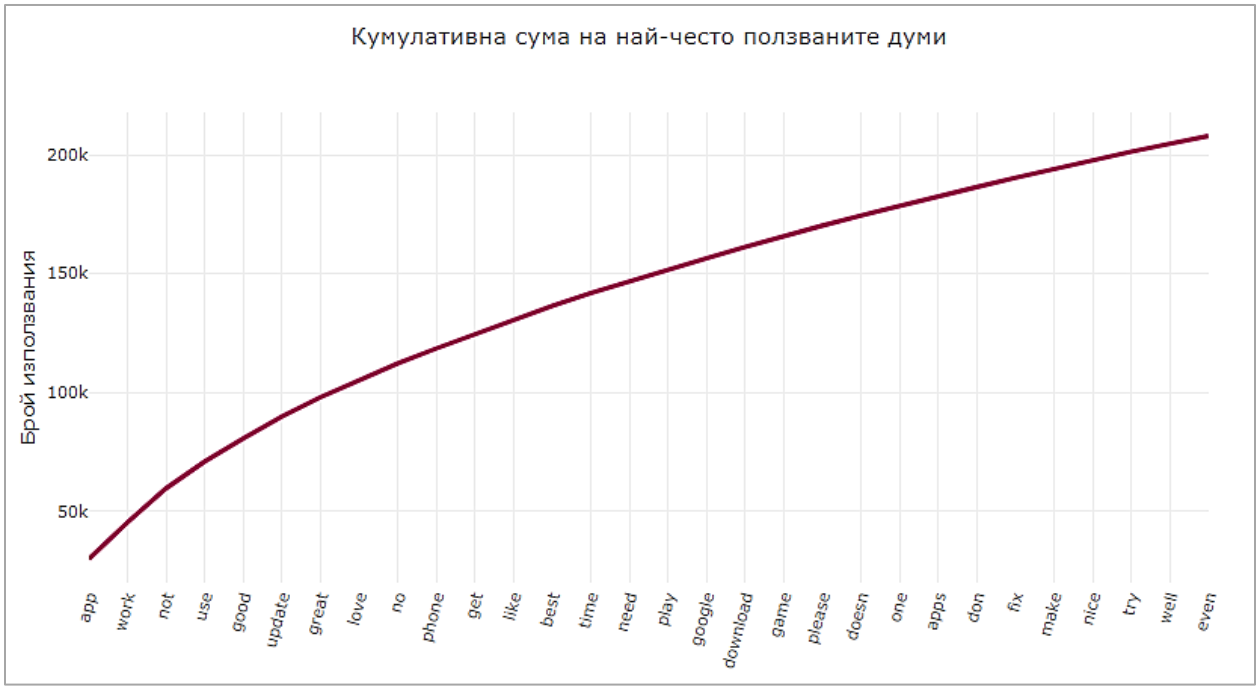
В Приложенията (**Фигура 15**) може да бъде открит облак от най-често срещаните биграми в негативни отзиви. На него ясно виждаме, че предположението ни относно наличието на много отрицателни изрази е напълно вярно – откриваме "doesn work", "not work", "not good", "not able" и т.н.

В Приложенията (**Фигура 14**) отново може да бъде открита и по-подробна графика на 20-те най-често срещани думи в отрицателни отзиви, която показва техният брой. На нея се вижда ясно, как думата "not" е била използвана над 12 000 пъти в отрицателни отзиви. За сравнение, в положителните тя бива открита само около 2 000 пъти или шест пъти по-рядко. Думата "app" е свързана с контекста и средата и напълно очаквано представлява най-често използваната дума в целият набор от данни. В графиката с топ думи в негативни мнения се забелязва много по-широка употреба на глаголи, а в

позитивните отзиви – на прилагателни. Можем да направим предположение, че когато потребителите са недоволни, те прилагат и описание на своят проблем в отзива, използвайки думи, означаващи действие (което е обяснение и на това защо негативните отзиви са по-дълги), а когато изразяват задоволство, най-често просто употребяват положителни определения за даденото приложение като цяло, без навлизане в детайли относно опита им с него.

От създадените визуализации можем да направим извод, че няма съществени проблеми, възникнали по време на обработката на текста и превръщането на всеки отзив чрез токънизация в низ от думи. Също така, можем да направим предположение, че някои от по-често срещаните думи за единия или другия клас на зависимата променлива, които наблюдавахме в създадените облаци, е много вероятно да бъдат част от думите с по-голяма предвиждаща способност, които ще бъдат избрани в следващият етап от настоящото изследване.

След обработката на целият набор от данни и неговата токънизация, ние разполагаме с общо **795 019** (с повторенията) токъна от униграми. Този брой съставлява цялата извадка от отзиви. От тях **37 132** са уникални. **Фигура 7** представлява графика на кумулативната сума на броят на първите 30 най-често използвани думи, както в положителни, така и в отрицателни мнения. От нея разбираме, че първите 30 от тях отговарят за около 200 000 токъна или **25%** от извадката. Това би могло да се тълкува като сигнал за липса на достатъчно разнообразие в набора от думи, с които разполагаме. Този извод трябва да се има предвид, тъй като такава особеност би могла да се отрази на представянето на модела, когато бъде използван върху нови данни (различни от тези, върху които е обучен и валидиран).



Фигура 7. Графика на кумулативната сума на броят на първите 30 най-често използвани думи в цялата извадка (положителни + отрицателни отзиви), използвана за анализа

3.6. Резултати

Преди да представим резултатите от емпиричното изследване, нека си припомним отново целта на настоящата работа, а именно - създаването на автоматизирана система за разпознаване на полярността на емоцията в отзивите на потребители на мобилни приложения. Допускането, че това е възможно чрез използването на методи на машинното самообучение, ни доведе до възникването на две хипотези, които ще бъдат проверени емпирично в следващата част от настоящата работа.

Нека припомним, че първата от тези хипотези засяга конкретно използването на обясняващи променливи в експеримента. Нашето предположение е, че добавянето на допълнителни променливи, извлечени от текста и различни от думите в него, ще подобри общото му представяне в сравнение със случая, в който използваме само думите в текста. Второто ни предположение е, че от моделите, които ще бъдат тествани (описани подробно в глава II – "Методология"), най-добро общо представяне ще получим използвайки линейна класификация чрез опорни вектори. Това как моделите се справят, бива оценено на база на избрани метрики (по-рано дефинирани в глава II).

Преди процеса по подбор на обясняващи променливи е важно да отбележим, че извадката е разделена на три части – тренировъчна, валидационна и тестова. **Таблица 3** разкрива броя наблюдения във всяка една от тях. Двата класа на зависимата променлива в тези три извадки са с равномерно разпределение (в съотношение 1:1).

Таблица 3. Размер на извадките, с помощта на които ще бъде създадена системата за разпознаване на настроението

Брой наблюдения (отзиви)	Тренировъчна извадка	Валидационна извадка	Тестова извадка
	60 000	20 000	16 000

3.6.1. Подбор на обясняващи променливи

Чрез подбора на обясняващи променливи се справяме с размерността на данните, като избираме тези, които се характеризират с по-добра предвиждаща способност, за да включим само тях в процеса на моделиране.

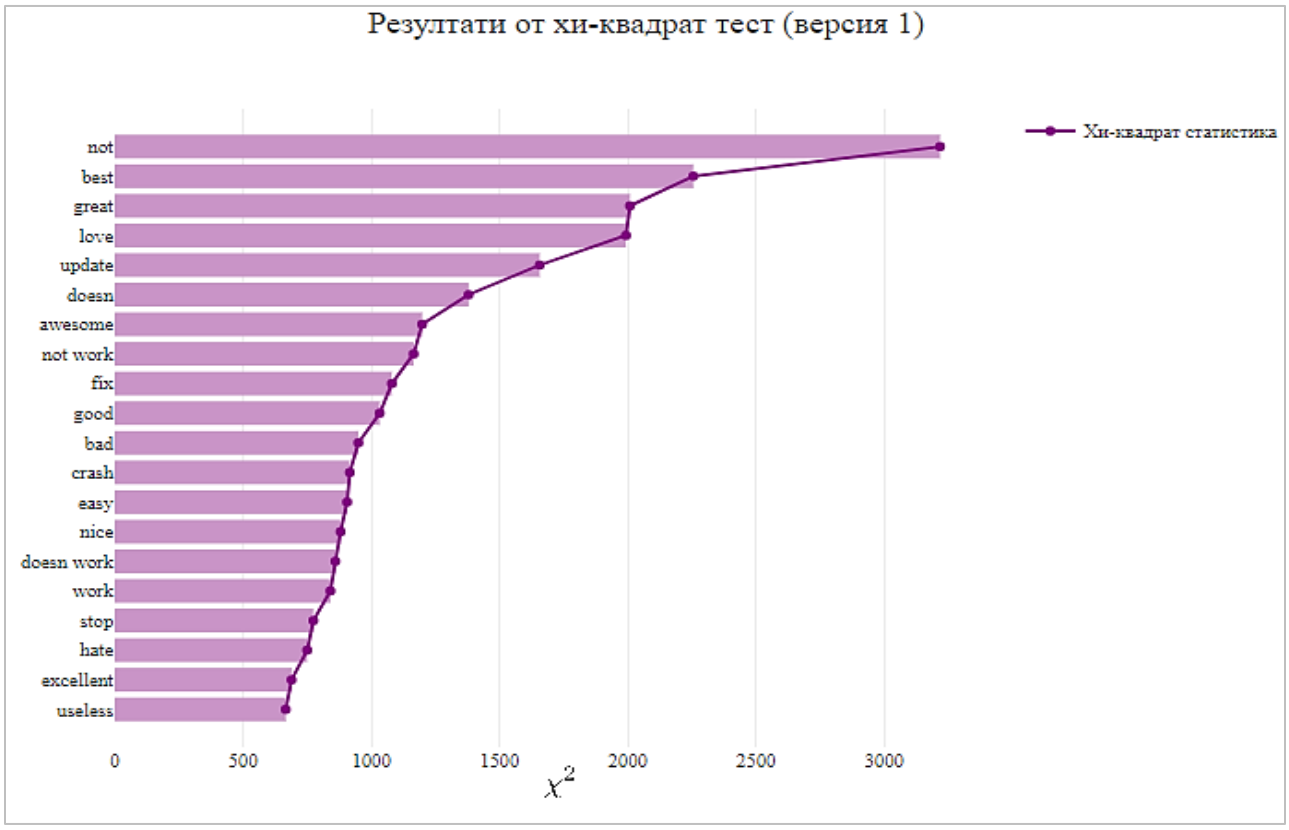
Подборът е осъществен само върху тренировъчните данни чрез използването на хи-квадрат тест, дефиниран в глава II. Съгласно избраната методология, моделите са тествани с три различни стойности на броя обясняващи променливи, включени в тях – 5 000, 7 000 и 10 000. Тези стойности са избрани на база на опита с дадената извадка. По време на провеждане на експериментите стана ясно, че при използване на над 10 000 обясняващи променливи, времето за изчисление нараства значително на машината, използвана за изследването, а подобрението в представянето е незначително. [25] Поради тази причина не са тествани модели с над 10 000 променливи. Възможно подобрение на експеримента в тази част, е все пак да се изследва по-обстойно оптималният брой променливи за всеки алгоритъм и чрез валидация да се забележи в кой момент има влошаване на представянето на алгоритъма, както са направили Narayanan et al. (Narayanan, Arora, & Bhatia, 2013, October) и Fang et al. (Fang & Zhan, 2015).

Независимо от това колко ще бъде броят обясняващи променливи във финалният модел, както и това кой от тях ще се представи най-добре в решението на класификационната задача, на този етап вече можем да определим кои са обясняващите

[25] Целият експеримент е проведен на машина с 8GB RAM и процесор - Intel(R) Core(TM) i7- 5500U CPU @ 2.40GHz, 64-bit операционна система.

променливи в нашата извадка, характеризиращи се с най-силна връзка със зависимата променлива. Това разбираме именно чрез резултатите от хи-квадрат теста. Променливите, за които получаваме най-голямо число на хи-квадрат статистиката са тези, които според теста би трябвало да имат най-голяма дискриминираща способност. Те със сигурност ще участват в създаването на всеки един от класификаторите в настоящата работа.

Фигура 8 показва първите двадесет най-силни променливи според хи-квадрат теста в случай, че използваме само униграми и биграми (за удобство, ще наричаме този вариант на експеримента "версия 1"), а **Фигура 9** показва резултата, след добавянето на допълнително извлечените от текста емотикони и други обясняващи променливи ("версия 2" на експеримента), вече описани подробно в глава II - "Методология".



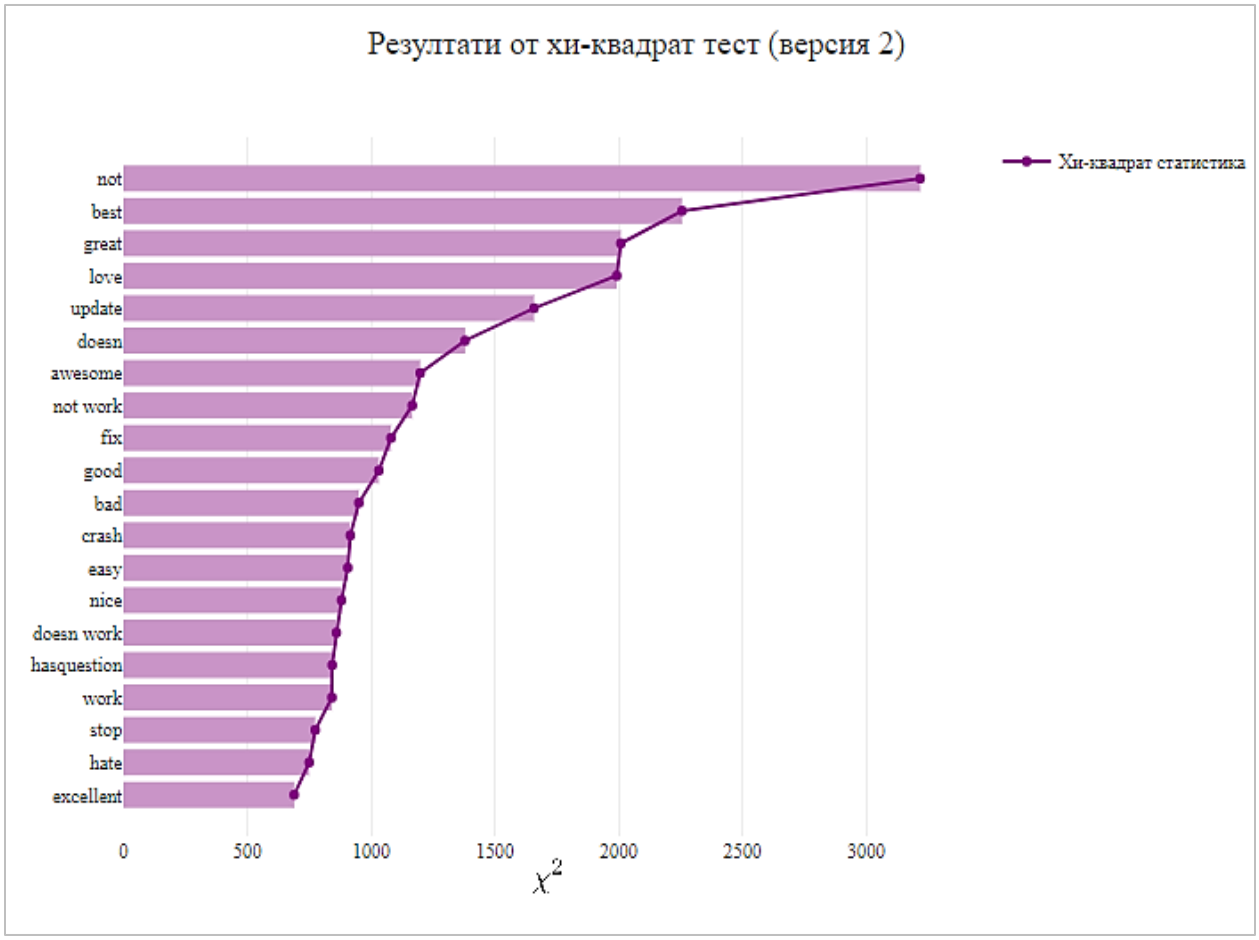
Фигура 8. Най-силни променливи според хи-квадрат теста (версия 1 – включени са само униграми и биграми)

От **Фигура 8** става ясно, че сред първите двадесет най-силно дискриминативни променливи участват само две биграми, останалата част са единични думи. Разглеждайки ги, става ясно и е напълно логично, кои от тях са определящи при предвиждане на положителния клас и кои от тях на отрицателния. Думи като "best", "great", "love" носят

силно положителен заряд, като абсолютно противоположното важи за “bad”, “crash”, “hate”, “useless”. Те са ни познати и от визуализациите на най-често използваните думи в предната точка на настоящата работа. Частичката за отрицание “not” има най-силна предвиждаща способност според резултатите от теста.

Биграмите “doesn work” и “not work” (виж **Фигура 8**) носят един и същ смисъл, но участват като отделни променливи. Може би е по-подходящо подобни изрази да бъдат обединени в един токън. Проста идея за това как можем да получим такова представяне, е да заместим всички изрази “doesn” с “not”. Дали подобни допълнителни обработки на данните ще доведат до по-добри резултати не е проверено в това изследване, но това е възможност, която би могла да бъде изследвана в бъдещи подобрения на настоящата работа.

От **Фигура 9** става ясно, че само една от допълнително създадените променливи се нарежда сред 20-те най-силно дискриминативни във версия 2 на експеримента. Това е обясняващата променлива, сигнализираща за наличието на въпрос в даденият отзив (останалите думи са същите, тъй като няма други промени, нанесени върху набора от данни, освен добавката на допълнително извлечените от текста обясняващи променливи).



Фигура 9. Най-силни променливи според хи-квадрат теста (версия 2 – включва допълнително добавените обясняващи променливи)

В Приложенията може да бъде открита и таблица с първите 100 най-силни обясняващи променливи, даваща информация за стойността на хи-квадрат статистиката и прилежащата ѝ вероятностна стойност за всяка от тези променливи - **Таблица 11**. Таблицата е разделена на две части - съответно за версия 1 и версия 2 на набора от обясняващи променливи с които се работи. От **Таблица 11** става ясно, че във версия 2 следните допълнително създадени обясняващи променливи са станали част от първите 100 най-дискриминативни такива:

- ✓ наличие на въпрос в отзива;
- ✓ наличие на възклицание в отзива;
- ✓ наличие на емотикона, изразяваща щастие;
- ✓ наличие на емотикона, изразяваща тъга.

Това наблюдение донякъде може да се тълкува като положителен сигнал за уместността на така извлечените допълнителни характеристики от данните и в известна степен като подкрепа в полза на първата хипотеза в настоящата работа.

3.6.2. Резултати от процеса по моделиране на данни

В тази точка ще представим подробно процеса и отделните стъпки в моделирането на данните и избирането на най-добре представящ се алгоритъм в решаването на основната задача в рамките на настоящата разработка. Ще интерпретираме резултатите от всеки един етап в този процес и ще обсъдим до колко те съвпадат с нашите очаквания и хипотези.

3.6.2.1. Резултати от първи етап в процеса по моделиране на данните

На първият етап в процеса бяха обучени общо 204 модела. Това се дължи на факта, че използваме търсене на хиперпараметри в решетка от предефинирани стойности (*grid search*) за всеки един от трите алгоритъма (подробно представени в глава II - “Методология”), като междувременно изпробваме и различен брой обясняващи променливи. Комбинацията от всички възможни варианти за тези величини и това, че тестваме два пъти – един път използвайки само думите в текста и един път добавяйки допълнителните променливи – доведе до този общ брой на обучените модели в първи етап от процеса.

Таблица 12[26] в Приложенията показва резултатите от този начален етап. Всеки един от моделите е обучен, използвайки данните само в тренировъчната извадка и е валидиран върху валидационната извадка. За момента не разглеждаме подробно всички характеристики на създадените модели (чрез мерките за прецизност, чувствителност и точност). Това ще бъде направено на следващ етап, тъй като първо, класификаторите трябва да бъдат сведени до по-обозрим брой, за да може да бъдат сравнени. Именно поради това на този етап, моделите са отсяти само с помощта на макро F1-мярката, която както беше обяснено по-рано в глава “Методология”, комбинира мерките за прецизност и чувствителност и за двата класа, като не поставя тежест върху нито една от двете. Тази статистика е изчислена само върху валидационната извадка, като **Таблица 12** показва

[26] За улеснение, имената на моделите в тази таблица, както и в останалите, са представени чрез английските им съкращения.

нейните стойности за всички различни постановки на трите изследвани алгоритъма. Макар и малки, разликите са отчетени (разлики в макро F1-мярката има чак след третият или четвъртият знак след десетичната запетая), следвайки примера на Cui et al. (Cui, Mittal, & Datar, 2006, July). Те също тестват представянето на няколко алгоритъма, променяйки различни параметри в своят експеримент, като получените резултати за някои от моделите са изключително близки.

Преди отсяването на база на макро F1-мярката разполагаме с общо 102 класификатора, създадени само чрез униграми и биграми и 102 класификатора, създадени чрез същите думи, но в добавка с допълнителните обясняващи променливи.

На този етап вече бихме могли да проверим първата хипотеза в настоящото изследване, а именно, дали представянето на класификаторите бива подобрено след добавянето на допълнителните характеристики, извлечени от текста. За целта избираме най-добре представилите се алгоритми на база на макро F1-мярката от всеки вид, като запазваме и разбивката по брой променливи, които включва класификатора, за да има съпоставимост на резултатите.[27] По-опростено обяснение е, че за всеки вид алгоритъм и брой включени обясняващи променливи в него – избираме модела с тази стойност на хиперпараметъра, за която сме получили най-висока стойност на макро F1-мярката. Това правим два пъти – един път за моделите, създадени само с униграми и биграми и един път за тези, създадени с допълнително добавените към тях характеристики на отзивите. На **Таблица 12** избраните модели са отбелязани в лилаво.

По този начин на първи етап от експеримента, свеждаме класификаторите до 18 на брой, като **Таблица 4** ни показва резултата и стойността на макро F1-мярката за всеки един от избраните модели.

[27] Няма да е съпоставимо да сравним Бейсов модел, създаден с 5000 променливи (без допълнителните) и същият алгоритъм, но с 10000 променливи (с допълнителните), тъй като ако има подобряване на представянето във вторият вариант, то това може да се дължи просто на факта, че модела е обучен с повече обясняващи променливи (което неминуемо увеличава степента на нагаждане към извадката), а не поради по-добрата дискриминативна способност на създадените от нас допълнителни обясняващи променливи.

Таблица 4. Резултати от първи етап на моделиране - макро F1-мярка на най-добре представилите се алгоритми в двете версии на експеримента (разбивката по обясняващи променливи е запазена)

Брой променливи	NB - v1	NB - v2	Linear SVC - v1	Linear SVC - v2	LogReg - v1	LogReg - v2
5 000	0.8072	0.8095	0.8430	0.8440	0.8426	0.8438
7 000	0.8069	0.8095	0.8439	0.8449	0.8444	0.8445
10 000	0.8067	0.8097	0.8455	0.8464	0.8443	0.8453

От **Таблица 4** става ясно, че подобрението след добавянето на новите променливи в класификацията е малко – разлика в стойността на макро F1-мярката се появява чак в третия знак след десетичната запетая. На таблицата в лилаво са отбелязани моделите, в които е констатирано увеличение.

В края на този етап, трябва да отбележим, че не само не се наблюдава влошаване в моделите след добавянето на допълнителните обясняващи променливи, но и съществува известно подобрение в представянето. Поради тази причина можем да приемем за вярна първата хипотеза в настоящото изследване. Трябва да се обърне внимание и на факта, че въпреки малкото подобрение (чак след третият десетичен знак) моделите се характеризират с доста добро представяне (на този етап оценено само на база на макро F1-мярката), сравнено с други утвърдени изследвания в областта на анализа на потребителски отзиви - (Chen, Liu, & Chiu, 2011), (Vinodhini & Chandrasekaran, 2012).

3.6.2.2. Резултати от втори етап в процеса по моделиране на данните

В първи етап от процеса на моделиране, успяхме да установим, че макар и минимално, все пак съществува някакво подобрение в класификацията, след добавянето на допълнителни променливи. Поради тази причина в следващите етапи в процеса по създаване на система за разпознаване на полярността на емоцията в текст ще продължим работата си, изследвайки по-подробно моделите, създадени с тези допълнителни обясняващи променливи.

На този втори етап целим да установим кои от класификаторите, включващи допълнително създадените променливи се представят най-добре. На първи етап, вече беше осъществен подбор на база на стойността на хиперпараметрите. В настоящата фаза критерият, по който подбираме моделите е броя обясняващи променливи в тях – при каква

стойност на тази величина постигаме най-добро представяне за всеки вид алгоритъм. На този етап отново оценяваме моделите на база на макро F1-мярката и свеждаме броят им до три. От **Таблица 5** става ясно, кои са тези модели. Вижда се, че и трите вида алгоритъм са се представили най-добре, когато използваме максималният брой обясняващи променливи, с които сме тествали (10 000).

Таблица 5. Резултати от втори етап на моделиране - макро F1-мярка на най-добре представилите се алгоритми само във втора версия на експеримента (разбивката по обясняващи променливи е запазена)

Версия 2			
Брой променливи	**NB**	**Linear SVC**	**LogReg**
5 000	0.8095	0.8440	0.8438
7 000	0.8095	0.8449	0.8445
10 000	0.8097	0.8464	0.8453

След тези два етапа бихме могли да кажем, че успешно сме селектирали най-добре представящите се три класификатора от общо 204, на база на стойността на макро F1-мярката (изчислена върху валидационната извадка) и според следните критерии:

- ✓ Дали да бъдат използвани или не допълнително извлечените обясняващи променливи от текста;
- ✓ Стойност на хиперпараметъра, избрана чрез техника за търсене на хиперпараметри в решетка от предефинирани стойности;
- ✓ Брой включени обясняващи променливи в модела.

3.6.2.3. Резултати от трети етап в процеса по моделиране на данните

На трети етап в процеса по създаване на система за разпознаване на полярността на емоцията в текст, вече можем да анализираме по-детайлно най-добре представилите се три класификатора. В този последен етап на подбор, следва да проверим и втората хипотеза в настоящото изследване, а именно кой алгоритъм се представя най-добре в класификационната задача, която разрешаваме в рамките на този труд.

Подборът до този момент е осъществен така, че да се изберат оптимално параметрите, с които трите класификационни алгоритъма постигат най-доброто си представяне върху валидационната извадка.

Таблица 6 ни показва по-подробно характеристиките на тези модели[28].

Таблица 6. Резултати от трети етап на моделиране - най-добре представили се модели от всеки вид алгоритъм- детайлна разбивка на резултатите върху валидационната извадка

Резултати върху валидационна извадка								
	Брой променливи	Прецизност$_{pos}$	Прецизност$_{neg}$	Чувствителност$_{pos}$	Чувствителност$_{neg}$	F1-мярка$_{pos}$	F1-мярка$_{neg}$	Точност
NB	10 000	0.7619	0.8840	0.9059	0.7169	0.8277	0.7917	0.8114
Linear SVC	10 000	0.8169	0.8828	0.8938	0.7996	0.8536	0.8391	0.8467
LogReg	10 000	0.8160	0.8814	0.8925	0.7987	0.8525	0.8380	0.8456

На база на резултатите до този момент, вече бихме могли да си изградим представа за това, кой класификатор се справя по-добре и кой по-зле в настоящата задача.

Става ясно, че на база на точността, както и на почти всички останали показатели Наивният Бейсов модел има по-лошо представяне от това на останалите два алгоритъма. Виждаме, че всички модели се характеризират с по-висока чувствителност за сметка на прецизността им относно положителния клас в изследването. Това означава, че алгоритъмът е разпознал правилно голям процент от положителните отзиви (чувствителност на модела). Последното е за сметка на това, че по-малък процент от примерите, които е определил като положителни – наистина са били такива (прецизност на модела). Това означава, че класификаторите имат тенденция неправилно да класифицират негативни отзиви като позитивни.

Разлика в представянето на логистичната регресия и метода на опорните вектори има чак в третия знак след десетичната запетая, като това важи за всички изчислени

[28] Долният индекс на всяка една от метриките, оценяващи представянето, дава информация за това за кой клас се отнася съответната статистика – (‘pos’ – позитивен клас; ’neg’- отрицателен).

мерки. Извода на база на резултатите върху валидационната извадка е, че тези два класификационни алгоритъма се справят еднакво добре.

Съществува още една стъпка на този последен етап, за да можем да затвърдим наблюденията си и това е използването на тестовата извадка. Нейното предназначение е именно това - да бъде независима извадка, която не е участвала нито в процеса по обучение на класификаторите, нито в процеса по избор на хиперпараметри. По този начин се застраховаме от това, да изпаднем в ситуация на прекомерно нагаждане към данните и получаваме оценка за това, как ще работи модела върху данни, които никога не е виждал.

Таблица 7 ни показва валидационните резултати върху тестовата извадка.

Таблица 7. Резултати от трети етап на моделиране - най-добре представили се модели от всеки вид алгоритъм- детайлна разбивка на резултатите върху тестовата извадка

Резултати върху тестова извадка								
	Брой променливи	Прецизност$_{pos}$	Прецизност$_{neg}$	Чувствителност$_{pos}$	Чувствителност$_{neg}$	F1-мярка$_{pos}$	F1-мярка$_{neg}$	Точност
NB	10 000	0.7566	0.8845	0.9075	0.7081	0.8252	0.7865	0.8078
Linear SVC	10 000	0.8180	0.8897	0.9009	0.7996	0.8575	0.8423	0.8503
LogReg	10 000	0.8177	0.8898	0.9010	0.7991	0.8573	0.8420	0.8501

От **Таблица 7** става ясно, че Наивният Бейсов модел е понижил своята точност. Интересното е, че за другите два алгоритъма констатираме леко повишение във всички изчислени статистики, като разликите помежду им вече са в четвъртият знак след десетичната запетая. Тенденцията неправилно да се класифицират негативни отзиви като позитивни, за сметка на по-големият процент разпознати положителни отзиви наблюдаваме и в тези резултати.

Изводът, който можем да направим е, че най-добро общо представяне на системата за разпознаване на полярността на настроението в потребителски коментари за мобилни приложения можем да постигнем използвайки, както метода на опорните вектори, така и логистичната регресия.

3.7. Сравнение с подобни изследвания

Следва кратко сравнение с резултатите, получени в утвърдени изследвания, занимаващи се с анализ на полярността на настроението, изразено в потребителски отзиви. В това сравнение включваме всякакви сфери (а не само софтуерната), като сме подбрали конкретно изследвания, прилагащи всички или поне някои от алгоритмите, използвани в настоящото.[29] Направена е съпоставка с резултатите, които постигнахме върху независимата тестова извадка, тъй като тя показва най-точно как ще се държат моделите, когато ги приложим върху нови данни.

Таблица 8. Сравнение с резултати в други изследвания на полярността на настроението, изразено в потребителски отзиви

Изследване	Данни	NB	SVM	LogReg
(Pang, Lee, & Vaithyanathan, 2002, July)	Отзиви за филми	81.50%	82.90%	81.00%
(Dave, Lawrence, & Pennock, 2003, May)	Отзиви за продукти в Amazon и CNET	87.00%	87.20%	-
(Smeureanu & Bucur, 2012)	Отзиви за филми	79.93%	-	-
(Narayanan, Arora, & Bhatia, 2013, October)	Отзиви за филми	88.80%	-	-
(Ortigosa, Martín, & & Carro, 2014)	Данни за статуси във Facebook	83.13%	83.27%	-
(Gezici, Yanıkoğlu, Tapucu, & Saygın, 2012)	Отзиви за хотели в TripAdvisor	-	80.85%	81.45%
Система, предложена в настоящата работа	Отзиви за приложения в Google App Store	**80,78%**	**85.03%**	**85.01%**
Средна стойност на постигнатата точност	**-**	**84.07%**	**83.85%**	**82.49%**

От **Таблица 8** става ясно, че като цяло резултатите, постигнати в настоящото изследване са подобни на тези в други такива изследвания. Точността на Бейсовият

[29] Базовите алгоритми са едни и същи, но в сравнението са включени различни техни разновидности в зависимост от това, какво е било използвано от авторите в цитираните изследвания (например, не само линеен метод на опорните вектори ами и такъв с радиално-базисна функция).

класификатор е с около 3% по-ниска от средната стойност на точността, измерена в избраните изследвания. За сметка на това, представянето, което постигаме с другите два алгоритъма е по-високо с около 1-2.5%, в сравнение с това в останалите.

3.8. Възможности за подобрение на създадената система за извличане на емоцията в текст

Възможностите за подобрение на създадената система в настоящата разработка можем да разграничим на три основни вида:

- ✓ Промени в обработката на естествен език – разширяване на използваният набор от техники за подготвяне на текста в подходящ за количествен анализ вид;
- ✓ Обогатяване на множеството от обясняващи променливи, използвани за създаване на класификационните модели;
- ✓ Изследване на представянето на други класификационни алгоритми от машинното самообучение или използване на различни подходи.

Следва да разгледаме в детайл всяко едно от горепосочените нива на анализа и да поясним, как то би могло да бъде използвано с цел подобрение на системата за разпознаване на полярността на настроението в потребителски отзиви.

Относно обработката на данните, възможно усъвършенстване би било използването на по-сложни техники за справяне с изкуствено удължени думи, както в изследването на Broody et al. (Brody & Diakopoulos, 2011, July). Друга мярка която допускаме, че ще има благоприятен ефект, е преминаването на текста под някаква форма на проверка на правописа (Miner, 2012). По този начин бихме могли да сведем значително грешните изписвания на думите, чието наличие неминуемо води до загуба на информация.

Относно използваните обясняващи променливи – възможностите за подобрение са много. Първата от тях е да се вземе предвид отрицанието в текста. В глава I вече посочихме разработки, в които това е осъществено, а в емпиричното изследване показахме защо това е важно. Na et al. (Na, Sui, Khoo, Chan, & Zhou, 2004) докладват за около 3% увеличение в точността на класификационният модел за разпознаване на полярността на емоцията, изразена в потребителски отзиви за разнообразни продукти,

именно чрез ползване на техники за справяне с отрицанието в текст. Zhu et al. също докладват за такова подобрение (Zhu, Guo, Mohammad, & Kiritchenko, 2014).

Друга възможност, която би могла да бъде тествана, е използването само на субективни фрази като обясняващи променливи и премахването на всички останали думи в текста, за които се смята, че не носят емоционална натовареност. Това бихме могли да постигнем чрез осъществяване на морфологичен разбор на изречението и оставяне само на прилагателни имена, съществителни имена и глаголи, както правят Fang et al. (Fang & Zhan, 2015), за които вече споменахме в глава I. Положително влияние би могло да окаже и премахването на често употребявани думи, свързани с контекста и средата, но не носещи емоция. В глава III се натъкнахме на такива думи. Тяхната дискриминираща способност е малка в контекста на задачата, която разрешаваме и това само би влошило анализа.

Друга идея за усъвършенстване на системата е използването на цели групи (клъстери) от думи като обясняващи променливи в класификацията. Тези клъстери могат да обединяват думи, имащи близко семантично значение. Такъв подход са предприели Agarwal et al. (Agarwal & Mittal, 2014), като според техните резултати, това е довело до подобрение на системата за разпознаване на настроението.

Както споменахме и по-горе, трета алтернатива е изпробването на различни техники за създаването на подобна система. Бихме могли да тестваме представянето на други алгоритми на машинното самообучение, като метода на случайната гора или метода на *k* най-близки съседи (*k-nearest neighbors algorithm*).

Както стана ясно в глава II, в това изследване използваме един генеративен и два дискриминативни подхода за класификация на полярността на настроението. Не можем да не споменем обаче, за наличието и на методи на дълбокото учене (*deep learning*), представляващи клас алгоритми на машинното самообучение, с които може да бъде постигната много висока предвиждаща точност. Изкуствените невронни мрежи (*artificial neural networks*) могат да бъдат използвани за класификация на полярността на настроението, като тестването на подобен подход също представлява възможност за подобрение на представения в настоящата разработка (Ghiassi, Skinner, & Zimbra, 2013).

Заключение

В настоящият труд си поставихме за цел и реализирахме създаването на система за разпознаване на полярността на потребителското настроение, изразено в отзиви за мобилни приложения.

Във връзка с така поставената цел на нашето изследване, ние осъществихме обстоен литературен преглед в областта на анализа на настроението в текст. Позиционирахме този вид анализ в обширното поле на анализа на текст и посочихме някои от основните моменти, които трябва да се вземат предвид в подобен род изследвания. Разгледахме утвърдени разработки в различни домейни, като обърнахме специално внимание на такива в сферата на мобилните приложения, тъй като именно този вид данни са обект на настоящата работа. Направихме обзор на поставените цели в тези изследвания, използваните техники за обработка на естествен език, приложените алгоритми и получените резултати. Изброихме част от многобройните приложения на анализа на настроението в света на бизнеса (политическата и социална сфера също не бива да бъдат пренебрегвани), като подчертахме, че в известна степен именно това обуславя неговата актуалност и в академичният свят.

За да постигнем целта на изследването създадохме методология за превръщане на текста в подходящ за количествен анализ вид и предприехме подход за разрешаване на настоящата задача чрез методи на машинното самообучение. Дефинирахме множество от обясняващи променливи, които да използваме в анализа. По време на този процес не се ограничихме само до думите в текста, а извлякохме и други променливи от него, носещи информация за емоционалната му натовареност. Последва процедура по подбор на обясняващите променливи с цел отсяване на тези, притежаващи по-голяма предвиждаща способност. Изследвахме поведението на три различни класификационни алгоритъма в решението на настоящата задача, като използвахме и техники за тяхното фино настройване с цел подобряване на представянето им. Валидирахме създадените модели върху независима извадка, за да проверим устойчивостта на системата. Интерпретирахме своите резултати, сравнихме ги с тези в утвърдени изследвания и очертахме кои са възможностите за подобрение на системата.

Резултатите, постигнати в настоящото изследване са подобни на тези в други сходни анализи. Точността на създадената система за разпознаване на полярността на настроението е **85%**, като това надвишава средната такава, докладвана от изследователи в сферата, имащи подобен на нашият подход.

Важно е да отбележим, че бяха формирани и две хипотези, които проверихме в рамките на настоящото емпирично изследване. Нашите изводи са, че използването на емотикони и други допълнителни характеристики (извлечени от текста и различни от думите в него) води до, макар и малко, реално съществуващо подобрение на представянето на системата. Също така, емпиричното ни изследване показва, че метода на опорните вектори и логистичната регресия се справят еднакво добре в разрешаването на задачата за разпознаване на полярността на настроението.

В заключение, ще очертаем и приносите на настоящият труд. В това изследване предлагаме методология за осъществяването на обработка и привеждане на текстови данни, част от домейна на мобилните приложения, в подходящ за количествен анализ вид. Методологията е авторско съчетание на различни техники от обработката на естествен език и е пригодена напълно за задачата, която разрешаваме в настоящият труд. Тя взима предвид спецификите на анализираният текст в този домейн, като това я прави приложима и в други изследвания, изискващи обработката на подобни данни. Някои от наблюденията ни върху данните (след прилагане на техниките за обработка) съвпадат с тези в други анализи на съдържанието на текст, част от домейна на потребителските отзиви за мобилни приложения. Това потвърждава устойчивостта на наблюденията ни в тази разработка.

В настоящият труд изследваме и представянето на три алгоритъма на машинното самообучение. Мотивът за това е, че резултатите от подобни изследвания могат да бъдат предвидени в много малка степен, тъй като зависят доста от домейна на текста, но и от други фактори, споменати по-рано. Така, нашите изводи и резултати могат да послужат за сравнителен анализ на представянето на основни алгоритми на машинното самообучение, използвани за разпознаване на емоцията в потребителски отзиви - както в домейна на мобилните приложения, така и като цяло в сферата.

Насока за бъдещо развитие на настоящата разработка е да бъде анализирано поведението на методи от дълбокото учене на данни за разрешаване на задачата за разпознаване на полярността на настроението, изразено в отзиви за мобилни приложения.

Библиография

Agarwal, B., & Mittal, N. (2014). Semantic feature clustering for sentiment analysis of English reviews. *IETE Journal of Research, 60(6)*, 414-422.

Brody, S., & Diakopoulos, N. (2011, July). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. *In Proceedings of the conference on empirical methods in natural language processing* (стр. 562-570). Association for Computational Linguistics.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). *A Practical Guide to Sentiment Analysis.*

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter, 6(1)*, 1-6.

Chen, L. S., Liu, C. H., & Chiu, H. J. (2011). A neural network based approach for sentiment classification in the blogosphere. *Journal of Informetrics, 5(2)*, 313-322.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research, 43(3)*, 345-354.

Ciurumelea, A., Schaufelbühl, A., Panichella, S., & Gall, H. C. (2017, February). Analyzing reviews and code of mobile apps for better release planning. *In Software Analysis, Evolution and Reengineering (SANER), 2017 IEEE 24th International Conference on* (стр. 91-102). IEEE.

Coelho, L. P., & Richert, W. (2015). *Building machine learning systems with Python.* Packt Publishing Ltd.

Cui, H., Mittal, V., & Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. *AAAI Vol. 6*, pp. 1265-1270.

Dave, K., Lawrence, S., & Pennock, D. M. (2003, May). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In Proceedings of the 12th international conference on World Wide Web* (стр. 519-528). ACM.

Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *Journal of retailing, 84(2)*, 233-242.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data, 2(1)*, 5.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33(1)*, 1.

Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., & Sadeh, N. (2013, August). Why people hate your app: Making sense of user feedback in a mobile app store. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (стр. 1276-1284). ACM.

Gezici, G., Yanıkoğlu, B., Tapucu, D., & Saygın, Y. (2012). New features for sentiment analysis: Do sentences matter? *CEUR Workshop Proceedings.*

Ghiassi, M., Skinner, J., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications, 40(16)*, 6266-6282.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. *In International encyclopedia of statistical science*, 977-979.

Gräbner, D., Zanker, M., Fliedl, G., & Fuchs, M. (2012). Classification of customer reviews based on sentiment analysis. *na*, 460-470.

Grano, G., Di Sorbo, A., Mercaldo, F., Visaggio, C. A., Canfora, G., & Panichella, S. (2017, September). Android apps and user feedback: a dataset for software evolution and quality improvement. *Proceedings of the 2nd ACM SIGSOFT International Workshop on App Market Analytics* (стр. 8-11). ACM.

Gu, X., & Kim, S. (2015, November). " What Parts of Your Apps are Loved by Users?"(T). *In Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference on* (стр. 760-770). IEEE.

Guzman, E., & Maalej, W. (2014, August). How do users like this feature? a fine grained sentiment analysis of app reviews. *Requirements Engineering Conference (RE), 2014 IEEE 22nd International* (стр. 153-162). IEEE.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research, 5(Oct)*, 1391-1415.

Hofmann, M., & Chisholm, A. (2016). *Text Mining and Visualization: Case Studies Using Open-source Tools (Vol. 40).* CRC Press.

Hogenboom, A., Bal, D., Frasincar, F., Bal, M., De Jong, F., & Kaymak, U. (2015). Exploiting Emoticons in Polarity Classification of Text. *J. Web Eng., 14(1&2)*, 22-40.

Hoon, L., Vasa, R., Martino, G. Y., Schneider, J. G., & Mouzakis, K. (2013, November). Awesome!: conveying satisfaction on the app store. *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (стр. 229-232). ACM.

Hoon, L., Vasa, R., Schneider, J. G., & Mouzakis, K. (2012, November). A preliminary analysis of vocabulary in mobile app user reviews. *In Proceedings of the 24th Australian Computer-Human Interaction Conference* (стр. 245-248). ACM.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Indriani, F., & Nugrahadi, D. T. (2016, October). Comparison of Naive Bayes smoothing methods for Twitter sentiment analysis. *In Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on* (стр. 287-292). IEEE.

Liang, T. P., Li, X., Yang, C. T., & Wang, M. (2015). What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. *International Journal of Electronic Commerce, 20(2)*, 236-260.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies, 5(1)*, 1-167.

Maalej, W., Kurtanović, Z., Nabil, H., & Stanik, C. (2016). On the automatic classification of app reviews. *Requirements Engineering, 21(3)*, 311-331.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

Manning, C., Raghavan, P., & Schutze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press.

Martin, W., Sarro, F., Jia, Y., Zhang, Y., & Harman, M. (2017). A survey of app store analysis for software engineering. *IEEE transactions on software engineering, 43(9)*, 817-847.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. *In AAAI-98 workshop on learning for text categorization* , (Vol. 752, No. 1, pp. 41-48).

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5(4)*, 1093-1113.

Miner, G. (2012). *Practical text mining and statistical analysis for non-structured text data applications.* Academic Press.

Mohammad, S. M. (2012, June). # Emotional tweets. *In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (стр. 246-255). Association for Computational Linguistics.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of machine learning.* MIT press.

Na, J. C., Sui, H., Khoo, C. S., Chan, S., & Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *I.C. McIlwaine (Ed.), Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference*, (стр. 49-54). Wurzburg Germany: Ergon Verlag.

Narayanan, V., Arora, I., & Bhatia, A. (2013, October). Fast and accurate sentiment classification using an enhanced Naive Bayes model. *International Conference on Intelligent Data Engineering and Automated Learning* (стр. 194-201). Berlin, Heidelberg: Springer.

Ortigosa, A., Martín, J. M., & & Carro, R. M. (2014). Sentiment analysis in Facebook and its application to e-learning. *Computers in Human Behavior, 31*, 527-541.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2(1–2)*, 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (стр. 79-86). Association for Computational Linguistics.

Panichella, S., Di Sorbo, A., Guzman, E., Visaggio, C. A., Canfora, G., & Gall, H. C. (2015, September). How can i improve my app? classifying user reviews for software maintenance and evolution. *Software maintenance and evolution (ICSME), 2015 IEEE international conference* (стр. 281-290). IEEE.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk.

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302)*, 157-175.

Pedregosa, F. V., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12(Oct)*, 2825-2830.

Peng, R. D., & Matsui, E. (2016). *The Art of Data Science: A Guide for Anyone Who Works with Data.* Lulu. com.

Rain, C. (2013). Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning. Swarthmore College.

Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems, 89*, 14-46.

Read, J. (2005, June). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. *In Proceedings of the ACL student research workshop* (стр. 43-48). Association for Computational Linguistics.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18(11)*, 613-620.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning.* Springer Science & Business Media.

Shai, S. S., & Shai, B. D. (2014). *Understanding machine learning: from theory to algorithms.*

Smeureanu, I., & Bucur, C. (2012). Applying supervised opinion mining techniques on online user reviews. *Informatica economica, 16(2), 81*.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

Tripathy, A., Agrawal, A., & Rath, S. K. (2015). Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science, 57*, 821-829.

Turney, P. D. (2002, July). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *In Proceedings of the 40th annual meeting on association for computational linguistics* (стр. 417-424). Association for Computational Linguistics.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research, 37*, 141-188.

Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal, 2(6)*, 282-292.

Vohra, S. M., & Teraiya, J. B. (2013). A comparative study of sentiment analysis techniques. *Journal JIKRCE, 2(2)*, 313-317.

Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social science computer review, 19(3)*, 324-347.

Weiss, S. M., Indurkhya, N., & Zhang, T. (2015). *Fundamentals of predictive text mining.* Springer.

Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (стр. 42-49). ACM.

Zhai, C. (2008). Statistical language models for information retrieval. *Synthesis Lectures on Human Language Technologies, 1(1)*, 1-141.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS), 22(2)*, 179-214.

Zhang, L., Hua, K., Wang, H., Qian, G., & Zhang, L. (2014). Sentiment analysis on reviews of mobile users. *Procedia Computer Science, 34*, 458-465.

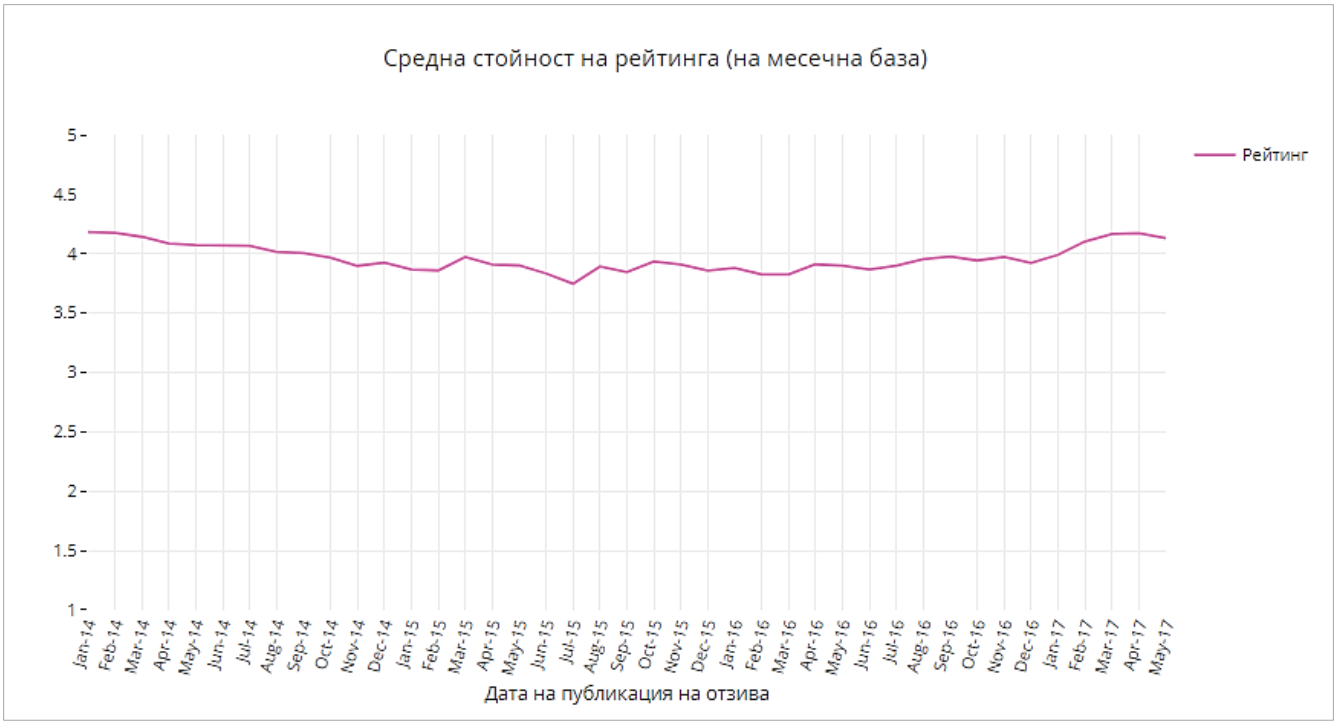
Zhu, J., & Hastie, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics, 5(3)*, 427-443.

Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014). An empirical study on the effect of negation words on sentiment. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* , (стр. 304-313).

Приложения

Таблица 9. Категории на приложенията, попадащи в извадката

App Category	Apps	Total Reviews
Books & Reference	19	15892
Business	1	1172
Comics	4	2287
Communication	33	31219
Education	12	1291
Entertainment	5	2584
Finance	7	621
Games	30	20378
Health and Fitness	3	1149
Libraries and Demo	3	990
Lifestyle	4	246
Maps and Navigation	10	1411
Music and Audio	17	3025
News and Magazines	6	1988
Personalization	18	12037
Photography	7	4275
Productivity	45	8361
Shopping	2	2647
Social	7	6146
Tools	139	151509
Travel and Local	9	984
Video Players and Editors	12	15352
Weather	2	2501
TOTAL	395	288065

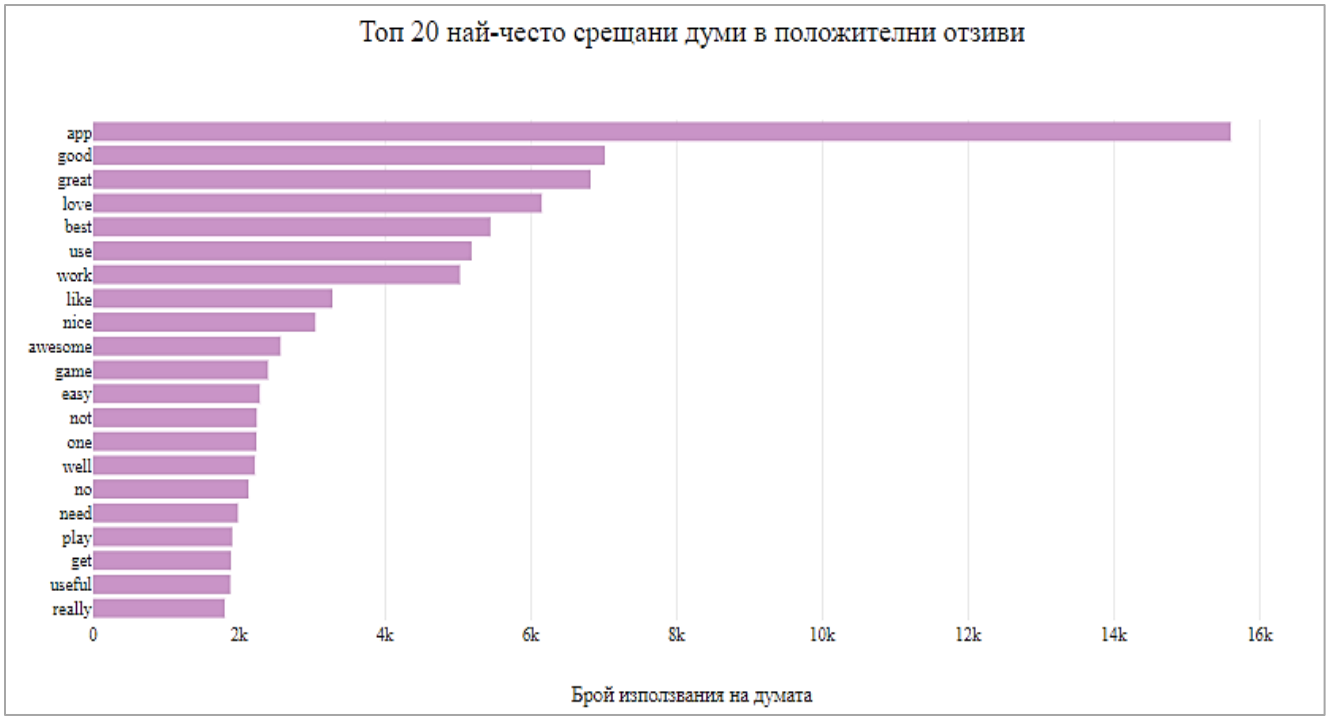


Фигура 10. Средна стойност на рейтинга на приложенията на месечна база (Януари 2014 – Май 2017)

Таблица 10. Описателни статистики на дължината на отзивите (измерена в брой думи) спрямо техният клас

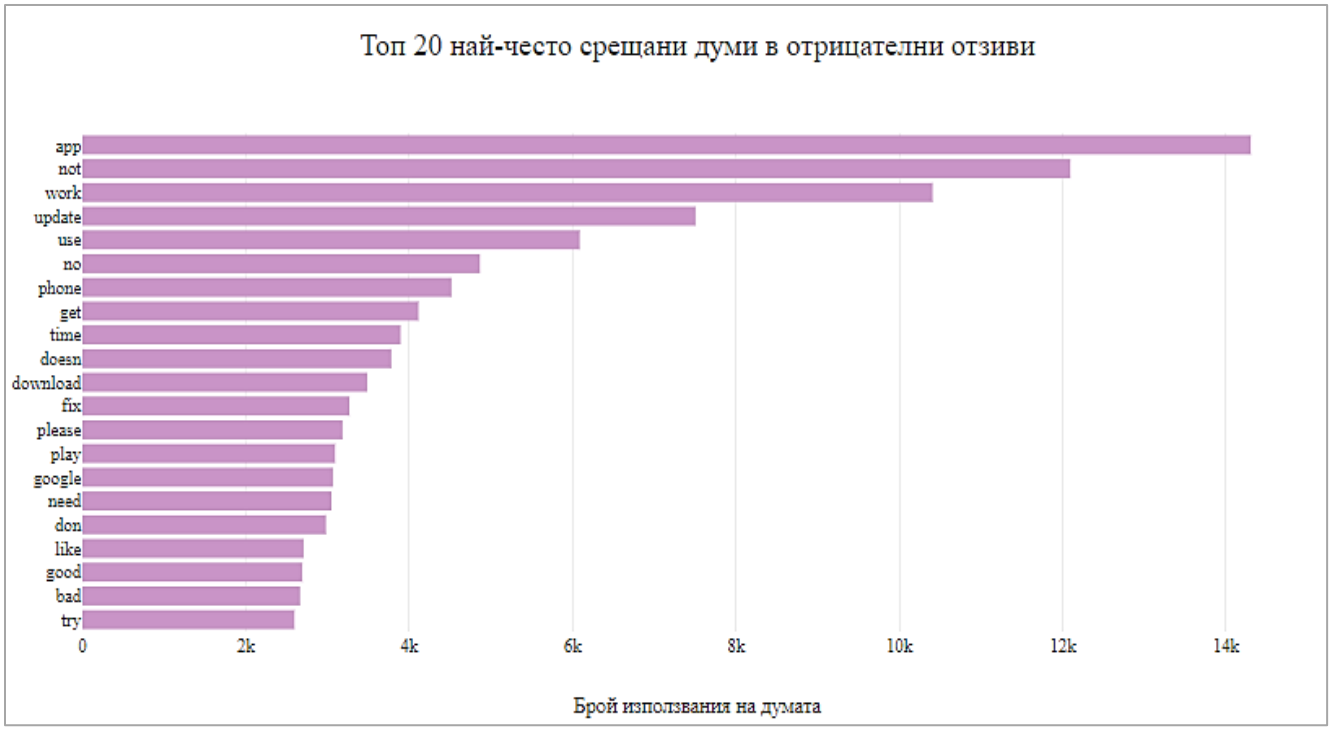
Описателни статистики на дължината на отзивите в извадката, използвана за разработването на системата за разпознаване на настроението		
	Положителни отзиви	Отрицателни отзиви
Брой наблюдения	48 000	48 000
Средна стойност	11.92	17.37
Стандартно отклонение	15.98	19.64
Минимум	1.0	1.0
25-ти персентил	3.0	5.0
50-ти персентил	6.0	11.0
75-ти персентил	14.0	23.0
Максимум	330.0	418.0
Ексцес	25.60	17.45
Асиметрия	3.735	2.996

Фигура 11. Разпределение на на дължината на всички отзиви (без да взимаме предвид класа)

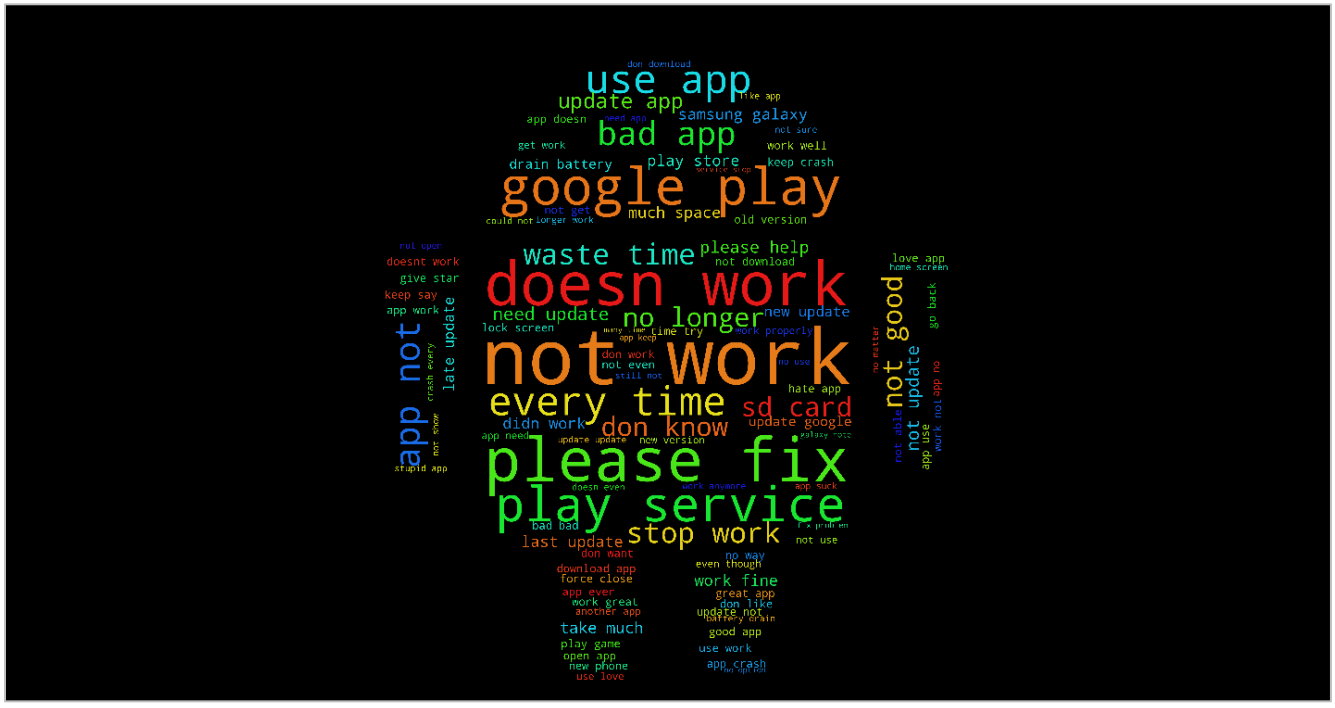


Фигура 12. Топ 20 най-често срещани думи (униграми) в положителни отзиви

Фигура 13. Най-често срещани биграми в положителни отзиви



Фигура 14. Топ 20 най-често срещани думи (униграми) в отрицателни отзиви



Фигура 15. Най-често срещани биграми в отрицателни отзиви

Таблица 11. Резултати от хи-квадрат тест – 100-те най-дискриминативни променливи за всяка версия от експеримента

Версия 1		
Дума	Хи-квадрат статистика	Вероятностна стойност
not	3218.18	0
best	2255.94	0
great	2008.95	0
love	1993.38	0
update	1656.76	0
doesn	1379.75	5.28E-302
awesome	1198.53	1.27E-262
not work	1165.27	2.15E-255
fix	1080.52	5.65E-237
good	1032.90	1.27E-226
bad	949.83	1.45E-208
crash	916.31	2.80E-201
easy	906.04	4.78E-199
nice	880.47	1.73E-193
doesn work	860.25	4.30E-189
work	840.88	7.00E-185
stop	774.52	1.87E-170
hate	750.23	3.58E-165
excellent	689.42	5.98E-152
useless	666.53	5.69E-147
useful	663.07	3.21E-146
great app	658.83	2.69E-145
time	658.03	4.00E-145
even	632.27	1.60E-139
simple	629.06	7.99E-139
suck	617.58	2.51E-136
don	603.81	2.48E-133
phone	585.20	2.78E-129
amaze	570.83	3.71E-126
waste	560.07	8.11E-124
please fix	553.34	2.36E-122
please	539.08	2.99E-119
error	538.14	4.78E-119
easy use	534.34	3.21E-118
thank	524.02	5.64E-116
thanks	521.55	1.95E-115
perfect	520.34	3.57E-115

Версия 2		
Дума	Хи-квадрат статистика	Вероятностна стойност
not	3218.18	0
best	2255.94	0
great	2008.95	0
love	1992.48	0
update	1658.21	0
doesn	1379.75	5.28E-302
awesome	1198.53	1.27E-262
not work	1166.25	1.32E-255
fix	1080.52	5.65E-237
good	1031.53	2.52E-226
bad	950.77	9.00E-209
crash	916.31	2.80E-201
easy	906.04	4.78E-199
nice	880.47	1.73E-193
doesn work	860.25	4.30E-189
hasquestion	842.90	2.54E-185
work	841.96	4.07E-185
stop	774.52	1.87E-170
hate	750.23	3.58E-165
excellent	689.42	5.98E-152
useless	666.53	5.69E-147
useful	663.07	3.21E-146
great app	658.83	2.69E-145
time	658.03	4.00E-145
even	632.27	1.60E-139
simple	630.00	5.00E-139
suck	617.58	2.51E-136
don	603.81	2.48E-133
phone	585.20	2.78E-129
amaze	563.94	1.17E-124
waste	560.07	8.11E-124
please fix	553.34	2.36E-122
please	538.36	4.29E-119
error	538.14	4.78E-119
easy use	534.34	3.21E-118
thank	524.02	5.64E-116
thanks	521.55	1.95E-115

try	498.50	2.01E-110
download	491.02	8.54E-109
no	484.58	2.15E-107
uninstall	473.28	6.18E-105
slow	467.74	9.96E-104
cant	449.67	8.50E-100
take	435.37	1.10E-96
get	417.72	7.67E-93
best app	413.92	5.14E-92
say	373.52	3.20E-83
show	344.26	7.54E-77
space	338.05	1.69E-75
force	337.78	1.94E-75
stupid	332.58	2.64E-74
cool	326.37	5.92E-73
annoy	324.39	1.61E-72
uninstalled	317.03	6.44E-71
delete	316.57	8.10E-71
remove	316.21	9.71E-71
open	310.21	1.97E-69
load	305.87	1.74E-68
start	287.53	1.72E-64
waste time	287.01	2.23E-64
every time	278.91	1.29E-62
install	275.34	7.79E-62
crap	274.42	1.23E-61
bad app	273.58	1.88E-61
job	270.07	1.09E-60
go	264.68	1.64E-59
free	261.73	7.22E-59
didn	258.25	4.12E-58
stop work	254.22	3.12E-57
nothing	251.33	1.33E-56
back	248.38	5.86E-56
anything	248.17	6.52E-56
anymore	241.46	1.89E-54
instal	226.04	4.35E-51
screen	224.97	7.47E-51
memory	224.14	1.13E-50
fantastic	222.87	2.14E-50
good app	220.43	7.28E-50
app not	220.24	8.04E-50
awesome	213.63	2.22E-48

perfect	520.34	3.57E-115
try	498.50	2.01E-110
download	492.40	4.29E-109
no	484.58	2.15E-107
uninstall	473.28	6.18E-105
slow	467.74	9.96E-104
cant	449.67	8.50E-100
take	435.37	1.10E-96
get	418.55	5.06E-93
best app	413.92	5.14E-92
say	373.52	3.20E-83
show	344.26	7.54E-77
space	338.05	1.69E-75
force	337.78	1.94E-75
stupid	332.58	2.64E-74
annoy	330.97	5.91E-74
cool	326.37	5.92E-73
uninstalled	317.03	6.44E-71
delete	316.57	8.10E-71
remove	316.21	9.71E-71
open	309.09	3.45E-69
load	308.56	4.50E-69
start	287.53	1.72E-64
waste time	287.01	2.23E-64
every time	278.91	1.29E-62
install	275.34	7.79E-62
crap	274.42	1.23E-61
bad app	273.58	1.88E-61
job	270.07	1.09E-60
go	264.68	1.64E-59
free	261.73	7.22E-59
didn	258.25	4.12E-58
stop work	254.22	3.12E-57
smileyhappy	253.47	4.56E-57
nothing	251.33	1.33E-56
hasexclamation	248.87	4.57E-56
back	248.38	5.86E-56
anything	248.17	6.52E-56
anymore	241.46	1.89E-54
instal	226.04	4.35E-51
screen	224.97	7.47E-51
memory	224.14	1.13E-50
fantastic	222.87	2.14E-50

app		
well	213.44	2.43E-48
galaxy	206.36	8.56E-47
wtf	206.34	8.62E-47
close	204.62	2.05E-46
no ad	204.36	2.34E-46
nice app	203.79	3.11E-46
uninstalling	202.98	4.68E-46
dont	202.62	5.61E-46
not good	201.94	7.86E-46
problem	201.31	1.08E-45
doesnt	197.65	6.80E-45
poor	194.95	2.64E-44
exactly	192.46	9.25E-44
google	192.01	1.16E-43
terrible	191.89	1.23E-43
unable	191.67	1.38E-43
happen	190.18	2.90E-43
app love	182.20	1.60E-41
longer	176.33	3.06E-40
fast	175.74	4.14E-40

app not	222.15	3.07E-50
good app	220.43	7.28E-50
well	213.89	1.94E-48
awesome app	213.63	2.22E-48
smileysad	211.99	5.06E-48
galaxy	206.36	8.56E-47
wtf	206.34	8.62E-47
close	204.62	2.05E-46
no ad	204.36	2.34E-46
nice app	203.79	3.11E-46
uninstalling	202.98	4.68E-46
dont	202.62	5.61E-46
problem	201.31	1.08E-45
not good	199.59	2.56E-45
doesnt	197.65	6.80E-45
poor	194.95	2.64E-44
happen	194.84	2.79E-44
exactly	192.46	9.25E-44
google	192.01	1.16E-43
terrible	191.89	1.23E-43

Таблица 12. Резултати от първи етап на моделиране – макро F1-мярка за всички обучени модели

Алгоритъм	Стойност на хиперпараметъра	Версия 1			Версия 2		
		Брой обясняващи променливи			Брой обясняващи променливи		
		5 000	7 000	10 000	5 000	7 000	10 000
NB	0.01	0.8065	0.8058	0.8053	0.8092	0.8087	0.8083
NB	0.05	0.8068	0.8064	0.8067	0.8093	0.8094	0.8094
NB	0.10	0.8069	0.8064	0.8067	0.8095	0.8095	0.8097
NB	0.30	0.8072	0.8069	0.8058	0.8095	0.8095	0.8092
NB	0.50	0.8071	0.8060	0.8051	0.8095	0.8095	0.8081
NB	0.70	0.8065	0.8058	0.8047	0.8089	0.8084	0.8080
NB	1.00	0.8060	0.8043	0.8037	0.8081	0.8071	0.8070
Linear SVC	0.001	0.8260	0.8260	0.8258	0.8272	0.8277	0.8275
Linear SVC	0.002	0.8320	0.8325	0.8326	0.8344	0.8351	0.8353
Linear SVC	0.004	0.8363	0.8371	0.8374	0.8391	0.8399	0.8400
Linear SVC	0.01	0.8395	0.8408	0.8405	0.8409	0.8427	0.8429
Linear SVC	0.02	0.8412	0.8430	0.8438	0.8433	0.8447	0.8457
Linear SVC	0.03	0.8429	0.8437	0.8453	0.8440	0.8449	0.8464
Linear SVC	0.06	0.8430	0.8439	0.8455	0.8440	0.8447	0.8458
Linear SVC	0.13	0.8425	0.8427	0.8433	0.8426	0.8432	0.8438
Linear SVC	0.25	0.8399	0.8403	0.8398	0.8409	0.8419	0.8412
Linear SVC	0.50	0.8387	0.8379	0.8358	0.8390	0.8393	0.8380
Linear SVC	1	0.8359	0.8352	0.8334	0.8371	0.8364	0.8345
Linear SVC	2	0.8344	0.8322	0.8295	0.8350	0.8336	0.8305
Linear SVC	4	0.8326	0.8295	0.8260	0.8341	0.8304	0.8266
Linear SVC	8	0.8318	0.8277	0.8234	0.8330	0.8288	0.8226
Linear SVC	16	0.8307	0.8267	0.8215	0.8327	0.8277	0.8195
Linear SVC	32	0.8289	0.8254	0.8189	0.8308	0.8263	0.8175
Linear SVC	64	0.8241	0.8248	0.8160	0.8279	0.8248	0.8152
Linear SVC	128	0.8223	0.8147	0.7907	0.8199	0.8241	0.8107
Linear SVC	256	0.7348	0.8079	0.7919	0.8152	0.8009	0.7996
Linear SVC	512	0.6954	0.7905	0.7571	0.7890	0.7868	0.7985
LogReg	0.001	0.7958	0.7963	0.7964	0.7971	0.7973	0.7973
LogReg	0.01	0.8238	0.8242	0.8242	0.8255	0.8262	0.8263
LogReg	0.1	0.8398	0.8408	0.8408	0.8420	0.8437	0.8437
LogReg	1	0.8426	0.8444	0.8443	0.8438	0.8445	0.8453
LogReg	10	0.8372	0.8363	0.8351	0.8380	0.8382	0.8361
LogReg	100	0.8317	0.8283	0.8253	0.8321	0.8297	0.8251
LogReg	1000	0.8296	0.8254	0.8192	0.8300	0.8257	0.8197