Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Даниела Иванова Петрова

ИЗПОЛЗВАНЕ НА СЕМАНТИЧНИ ТЕХНОЛОГИИ В ЕДИН КЛАС СОФТУЕРНИ ПРИЛОЖЕНИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

по докторска програма: Автоматизирани системи за обработка на информация и управление
професионално направление:5.3.”Комуникационна и компютърна техника”

Научен ръководител: Доц. д-р инж. ВИОЛЕТА БОЖИКОВА

Рецензенти:
1. .....
2. .....

Варна 2023 г.

Дисертационният труд е обсъден на ……2023 г. в катедра „Софтуерни и интернет технологии“ на катедрен съвет и насочен за защита.

Автор: Даниела Иванова Петрова

Заглавие: Използване на семантични технологии в един клас софтуерни приложения

Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Даниела Иванова Петрова

ИЗПОЛЗВАНЕ НА СЕМАНТИЧНИ ТЕХНОЛОГИИ В ЕДИН КЛАС СОФТУЕРНИ ПРИЛОЖЕНИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

Варна 2023 г.

Дисертационният труд съдържа 142 страници, включително 22 фигури, 53 таблици, 31 формули, оформени в 4 глави, приноси на дисертационния труд, списък с публикациите на автора по темата на дисертационния труд и списък на използваната литература от 116 заглавия, от които 10 на кирилица и 106 на латиница.

Защитата на дисертационния труд ще се състои на ..... г. от ..... ч. в ..... на открито заседание на жури сформирано със заповед на Ректора № ...../..... г.

Материалите по защитата (дисертацията, рецензиите и становищата) са на разположение на интересуващите се в Докторантски център, стая 318 НУК.

1. Актуалност на проблема

С развитието на Интернет, WWW и мобилните технологии през последните години ставаме свидетели на нарастващата популярност на уеб приложенията (клас софтуерни приложения). Същевременно количеството на текстова информация, генерирана от тези уеб приложения всеки ден, се покачва драстично. Това огромно количество от предимно неструктуриран текст не може просто да бъде обработено и възприето от компютрите. За тази цел са необходими ефикасни (точни) и ефективни (бързи)[4] техники и алгоритми. Нарастващият обем на информация, както и непрекъснато увеличаващият се брой потребители на тази информация, определят все повече нуждата от категоризиране, описание, подреждане и класифициране на информацията в общо хранилище, превръщайки неструктурирания текст в структуриран.

През последните години семантичният уеб се появи и наложи като основа, която чрез набор от технологии и стандарти, задава общ формат на данните и протоколите за тяхната обмяна в уеб пространството. Част от семантичния уеб са и текстово добиване на знание (text-mining – TM) и анализа на настроения (sentiment analysis – SA), които изпълняват задачите за извличане на значима информация от текст, както и определянето на полярността на мнението в даден текст. Тези задачи набират все повече внимание поради огромното количество на текстови данни, които се генерират под различни форми, като социални мрежи, заявления за патенти, застрахователни данни, здравно-осигурителни данни, новини, коментари и мнения за продукти и други.

Според направеното от автора проучване за наличието на разработки и изследвания в сферата на извличането на мнения от текстове на български език може се стигне до заключението, че техният брой е все още незадоволителен. За разлика от английския език, за който има богат набор от инструменти и бази от данни (БД), основен проблем за българския език е почти пълната липса на средства за обработка на текстове на български език, както и готови набори от данни, които могат свободно да се използват в изследванията.

2. Цели и задачи на изследването

Основна научноизследователска цел на дисертацията е създаването на алгоритми и подходи за извличане на мнения (настроения, чувства) и анализ на текстове на български език - нещо, което липсва или е в оскъдни количества в публикуваните до момента и открити от докторанта научни проучвания в България. За реализация на тази цел са формулирани следните основни задачи:

1. Да се изследва текущото развитие на технологиите свързани с анализа на мнения, в частност за българския език.
2. Да се разработят БД с коментари на български език, за целите на дисертацията и бъдещо ползване.
3. Да се разработи собствен списък от „стоп думи“ за целите на предварителната обработка на данните и възможност за бъдещо ползване.
4. Да се предложи процедура по предварителна обработка на данните, в зависимост от спецификите на българския език.
5. Да се извърши анализ на сложността на текстовете в разработените БД.
6. Да се изследват, оценят и сравнят съществуващи методи, алгоритми и съвременни модели за класификация на текст в машинното обучение и да се предложи нов, по-точен алгоритъм.

3. Обект и предмет на изследване

Обект на изследване са методи и алгоритми на машинното обучение за решаване на класификационни задачи, в частност анализ на мнения в текстове на български език.

Предмет на изследване са коментари на потребители на български език.

4. Методи на изследване

За постигане на поставените цели и задачи са приложени следните методи:

Емпирични методи: Анализ на публикации, доклади, книги и резултати от дейности; Теоретични методи: Индукция, Анализ, Синтез, Сравнение, Извличане на общи черти, Обобщение и др.

5. Място на изследване

Изследванията са проведени в лабораториите на катедра СИТ при ТУ – Варна.

6. Научна новост на изследването

Създаден е подобрен вариант на алгоритъма за създаване на база данни и предварителната й обработка в спецификите на българския език.

Предложен е алгоритъм за извличане на мнение от коментари на български език, който дава най-добри резултати, спрямо останалите, извършени експериментални изчисления. Този алгоритъм представлява метод на ансамбъла, който комбинира Метод на опорните вектори - Support Vector Machines (SVM) и рекурентни невронни мрежи (LSTM), като в SVM модела е включен механизмът на внимание, и приложен мета класификатор Случайна Гора (Random Forest-RF), в комбинация с частичен метод на лексикона.

7. Практическа ценност на изследването

Към резултатите с приложна насоченост могат да се посочат разработването на две бази от данни (БД) с мнения и коментари на български език (с приблизително по 100 000 коментара всяка). Авторът претендира да е създал двете най-големи БД с коментари на български език до момента, които могат да се използват и за бъдещи научни изследвания. Неговите изследвания са единствени по рода си, базирани на такива обширни набори от коментари на български език. Разработен е и списък от стоп думи, който може да се използва за предварителната обработка на текстове в бъдещи анализи и изследвания на текстове на български език.

8. Апробация на изследването

Основните резултати от изследванията са докладвани и публикувани в следните международни научни форуми и издания:

Конференции:

- International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) - 11-18 Ноември 2022, Малдиви
- International Conference on Biomedical Innovations and Applications, 2-4 Юни 2022, Варна, България
- Integrated Emerging Methods of Artificial Intelligence & Cloud Computing (IEMAICLOUD) 26-28 Април 2021, Лондон
- Interdisciplinary Conference on Mechanics, Computers and Electrics (ICMECE) - 27-28 Ноември 2021, Анкара, Турция

Списания:

- Списание КОМПЮТЪРНИ НАУКИ И ТЕХНОЛОГИИ - Година XVIII, Брой 1/2020, стр 126, ISSN 1312-3335
- Годишник на ТУ-Варна, България, 6(2), 57-62, ISSN :2603-316X

Докторантът и неговия научен ръководител имат и едно цитиране на тяхна обща публикация:
Petrova, D., Bozhikova V. (2022) Random forest and recurrent neural network for sentiment analysis on texts in Bulgarian language, International Conference on Biomedical Innovations and Applications, Varna, Bulgaria, 2-4 June 2022, **DOI:** 10,1109/BIA52594.2022.9831326, **Electronic ISBN:**978-1-6654-4581-8, 66-69стр
От
Ananthajothi K., Karthikayani K., Prabha R., Explicit and implicit oriented Aspect-Based Sentiment Analysis with optimal feature selection and deep learning for demonetization in India (2022) Data and Knowledge Engineering, 142, art. no. 102092, **DOI:**10.1016/j.datak.2022.102092

9. Публикации по дисертационния труд

Основните етапи от разработването на дисертационния труд са отразени в шест заглавия, три от които в съавторство, четири - индексирани по SCOPUS, списък на които е приложен в края на автореферата.

Глава Първа: Обзор

В Глава 1 е направен преглед на тенденциите в развитието на уеб технологиите; разгледани са различните методи и подходи, използвани в TM и SA, и тяхното прилагане до момента; представен е и обзор на извършените изследвания върху текстове на български език в научните среди в България.

Изводи към Глава Първа

От направеното изследване върху съществуващите разработки и различните методи, които са приложени, може да се заключи:

- Налице е голямо разнообразие на алгоритми и методи, които могат да се използват за извличането на мнения и откриването на чувства в даден текст. Има множество проучвания, който търсят или комбинират различни методи за анализ на мнения на английски език. Разработен е богат набор от техники, алгоритми, инструменти и БД, които да се използват в предварителната обработка и в последващите изследвания, но отново за английски език.
- Българският език, обаче, се явява „слабо развит“ език. Липсват или са малко различните инструменти за предварителна обработка на данните. Самите данни, по-конкретно за задачата за извличане и анализ на мнения, са оскъдни или почти липсват. Малко са и извършените изследвания върху текстове на български език.
- Всичко това предполага нуждата от провеждането на подробно изследване и търсенето на най-подходящия алгоритъм за извличането на мнения в спецификите и особеностите на българския език, както и създаването БД с коментари на български език.

Глава Втора: Алгоритми, методи и метрики

В Глава 2 е изложена формалната постановка на задачата за анализ на мнения чрез класификация; извършен е подбор на инструменталните средства; описан е алгоритъмът на изследването и са показани методите и метриките за оценка на класификационните модели.

В таблица 1 са изведени предимствата, недостатъците и приложенията на избраните за целите на дисертационния труд основни методи в машинното обучение. Всеки от тези методи има свои силни и слаби страни, което ги

прави подходящи за различни видове задачи по машинно обучение. Изборът на метод зависи от конкретната задача, характеристиките на данните и наличните изчислителни ресурси.[95, 105]

Описание на метода	Предимства	Недостатъци	Приложение
Naïve Bayes			
Тип: Вероятностен Класификатор **Предположение**: Предполага, че признаците за даден клас са условно независими.	✓ Прост и изчислително ефективен ✓ Добре се справя с високо размерни данни ✓ Често дава добри резултати за класификация на текст и филтриране на спам	- Предполага независимост между признаците (което не винаги е така) - Може да не се справи добре, когато предположението за независимост е нарушено.	• Класификация на текст (например, детекция на спам, анализ на настроения) • Категоризация на електронна поща • Категоризация на документи
Логистична регресия (Logistic regression)			
Тип: Супервизиран Класификатор **Модел**: Сигмоидната функция изобразява линейна комбинация на признаците към вероятността.	✓ Прост и интерпретируем. ✓ Ефективен за бинарна. класификация ✓ Предоставя вероятности като изход.	- Предполага линейна връзка между признаците и целевата променлива. - Може да не се справи добре със сложни, нелинейни връзки.	• Медицинска диагностика • Прогнозиране на напускането на клиенти. • Оценка на кредитоспособност.
Метод на опорните вектори (Support Vector Machines – SVM)			
Тип: Супервизиран Класификатор **Цел**: Намира хиперплоскост, която максимизира разстоянието между класовете	✓ Ефективен във високо размерни пространства ✓ Може да се справи както с линейни, така и с нелинейни класификационни задачи, използвайки	- Чувствителен към избора на функция на ядрото и хиперпараметри - Може да бъде изчислител-	• Класификация на изображения. • Класификация на текст. • Разпознаване на ръкописни

	различни ядра. ✓ Добър при обработка на малки и средно големи набори от данни.	но интензивен при големи данни	символи.
Random Forest			
Тип: Ансамбълен Класификатор (комбинира множество решаващи дървета) Комбинира Bootstrap Aggregation и случайно подреждане на признаци.[99]	✓ Висока точност и устойчивост срещу прекомерно обучение. ✓ Справя се както с класификационни, така и с регресионни задачи. ✓ Може да се справи с високоразмерни данни и липсващи стойности.	- Обучението може да бъде бавно. - По-малко интерпрети-руем спрямо едно решаващо дърво.	• Класификация на изображения. • Прогнозиране на поддръжка. • Финанси- за детекция на измами
Recursive Neural Network			
Тип: Невронна мрежа с обратни връзки. Подходящ за последователни данни, където редът има значение.	✓ Може да моделира ефективно последователни и времеви редове данни. ✓ Справя се с входни последователности с променлива дължина. ✓ Добър за задачи като разпознаване на реч, моделиране на езика и анализ на настроения.	- Страда от проблема с изчезващ градиент. - Склонен към проблеми с дългосрочни зависимости. - Изисква изчислителни ресурси.	• Обработка на естествен език (NLP). • Разпознаване на реч. • Прогнозиране на времеви редове.
Stacking			
Тип: Ансамбълен метод Комбинира множество базови модели чрез обучение на мета-модел.	✓ Може да улавя разнообразни моделни шаблони в данните, като комбинира различни модели. ✓ Обикновено	- Изисква допълнител-ни изчислител-ни ресурси и обучение на модели.	• Всякакви класифика-ционни или регресионни задачи, където е важна производител-

	представя по-добри резултати от индивидуалните модели. ✓ Предоставя гъвкавост при избора на модели.	-	Сложност при избора на правилни базови модели и мета-модела.	ността.

Изводи към Глава Втора

- Според направеното проучване не съществуват големи БД с коментари на български език, които биха могли да се използват за текущото изследване. Това доведе до решението да се създаде собствена БД с коментари на български език към настоящата разработка, което в последствие прерасна в създаването на две БД с по около 100 000 записа.
- Избрани са два различни вида векторизатора Count Vectorizer и Tf-idf векторизатор, за да се определи кой работи най-добре за дадена задача.
- За извършване на анализ на сложността на текстовете е избрана формулата на индекса Gunning Fog, адаптирана за български език.
- За целите на дисертационния труд са избрани следните методи за класификация и анализ на настроенията: Наивен Бейсов класификатор; Логистична регресия; Метод на опорните вектори; Метод случайна гора; Рекурсивни невронни мрежи; Метод на Ансамбъла.
- За метод за валидиране е избран метода Train-Test-Split, като експериментални изследвания са извършени върху различни комбинации от съотношения в данните за обучение и тестване (70%-30%; 80%-20%), за да се открие кое е по-подходящото.
- За основна метрика за оценка на класификационните модели е избран показателя Точност.

Глава Трета: Създаване на БД и предварителна обработка на данните

Глава 3 представя етапите по създаването от автора на двете БД с коментари на български език, както и списъка от „стоп думи“. Предложен е подобрен вариант на алгоритъма за предварителна обработка на данните, в спецификата на българския език. Извършен е анализ на текстуалната сложност на двете БД.

3.1. Избор на подходящ набор от текстови документи

Както вече бе казано, според направеното от докторанта проучване за българския език, почти няма готови БД с коментари в каквато и да е било сфера, които биха могли да се използват за експерименталните изследвания на автора. Единствените открити от автора са 10 000 коментара на филми, разработени от Борислав Капукаранов и Преслав Наков [43]. Това наложи създаването на собствена база данни с мнения. За начало на изчисленията на текущата разработка бяха избрани данни за мнения на клиенти на хотели, тъй като това е една от най-обширните теми, по която хората пишат коментари. За целта, като източник на информация, бе използван сайтът www.booking.com, тъй като в него информацията е най-добре подредена и е в голямо количество. За да не се налага ръчно сваляне на данните, бе използвана програмата Octoparse 8.0, която позволява лесно обхождане на Интернет страниците и сваляне на необходимата информация. Коментарите в booking.com са подредени в две графи, положителни и отрицателни, за всеки един хотел, като понякога потребителите са оставили само положителен или само отрицателен коментар. Това е и причината положителните коментари да са в пъти повече от отрицателните. Бяха свалени положителни и отрицателни коментари за хотели и къщи в различни курорти и градове в България: Варна, Бургас, София, Пловдив, Велико Търново, Русе, Боровец, Банско, Пампорово, Златни пясъци, Албена, Слънчев бряг, Несебър, Поморие, Обзор, Трявна, Свети Влас и Приморско. Наложи се допълнителна ръчна обработка на данните, тъй като много от отбелязаните от потребителите като отрицателни коментари съдържаха изречения от рода: „*Няма нищо*“, „*Всичко ни хареса*“, „ *Бяхме много доволни*“, „*Няма нещо, от което да не сме доволни*“ и др. Такива коментари, които са напълно положителни, но реално маркирани като отрицателни, биха объркали системата и биха дали грешни резултати. Поради тази причина, бе прегледана цялата база данни за подобни коментари и те бяха изтрити. Допълнително, поради грешка в програмата за сваляне на данните, когато потребител е оставил само отрицателен коментар за хотела, той се сваля като положителен. Отново бе прегледана цялата база данни и такива коментари бяха отстранени от графа положителни и бяха преместени в графа отрицателни коментари. След така проведеното машинно сваляне и ръчно доизчистване на коментарите от посетители на хотели, беше съставна база данни от 31 720 положителни коментари и 17 230 отрицателни коментара на български език, отделени в два отделни файла, като всеки

коментар е на отделен ред и може да съдържа от една дума до няколко изречения. За да се увеличи обемът на базата данни, допълнително беше открит нов източник на информация – Интернет страницата за ваучери – grabo.bg, в която могат да се открият коментари за 2400 хотела и къщи за гости в България. Свалени бяха всички налични мнения в тази категория. Начинът, по който са структурирани коментарите в booking.com, се различава от този на grabo.bg. В първия, оставените мнения са разделени в две отделни полета – положително и отрицателно на всеки един потребител. Докато във втория сайт коментарът е един и неговото настроение се означава със звездички – от 1 (за напълно негативните) до 5 (за най-положителните). Поради тези различия данните трябваше да се унифицират, като коментарите с 4 и 5 звездички бяха определени като положителни, а тези с 1, 2 или 3 звездички – като отрицателни. Отново прави впечатление разликата между броя на положителните и броя на отрицателните коментари. Една от причините за това е, че в много от случаите в booking.com бе попълнена само графата за положителни отзиви, а тази на отрицателните или е оставена празна, или са написани коментари от рода „*нямам забележки*“, „*всичко беше наред*“ и т.н. Като краен резултат се получи база от данни с 100 082 коментара, 72 078 – положителни и 28 004 – отрицателни.

На по-късен етап от работата по текущата дисертация, отчасти поради не достатъчно високите резултати на предсказване, използвайки първоначалната база данни, бе решено да се състави втора база на български език, с данни от друга сфера. За източник на информацията отново бе избран Интернет портала за ваучери grabo.bg. Този път бяха избрани коментари от бизнес сфери, различни от хотелиерството – различни културни и развлекателни събития, ресторанти, козметични и спа процедури, медицински услуги, магазини за хранителни и нехранителни стоки и други. Отново, след първоначалното им сваляне и събиране, всички коментари с оценка 4 и 5 звездички, бяха маркирани като положителни, а останалите - като негативни. Това бе извършено с цел да се уеднакви структурата на двете БД. По този начин се получи предварителна база от 105 052 коментара, 90 384, от които положителни и 15 062 отрицателни.

Като краен резултат авторът е разработил две големи самостоятелни БД с коментари на български език, които могат да се ползват свободно за бъдещи изследвания и анализи, със съответно 100 082 и 105 052 коментара.[108] В Таблица 1 може да се види

разпределението на положителните спрямо отрицателните коментари в двете БД.

Оттук нататък първоначалната база данни с коментари за хотели за по-кратко ще бъде наричана База 1, а базата с мнения за други обекти и услуги - База 2.

База 1 – хотели	База 2 – други услуги
	
100 082	**105 052**

Таблица 1 - Разпределение на положителни и отрицателни коментари в двете БД

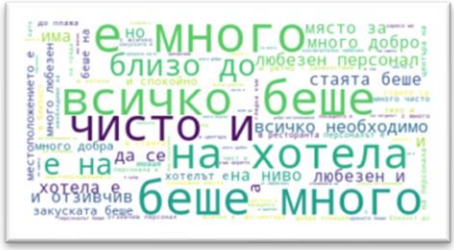
3.2. Предварителна обработка на текста

Предварителната обработка на текстове стандартно включва следните няколко дейности:

1. Преобразуване на текста към малки букви;
2. Преобразуване на изреченията в поредица от думи;
3. Премахване на празните места и пунктуацията;
4. Изчистване от думи и букви на друг език, освен кирилица. Бяха премахнати и всички коментари, които не са на български език. Тъй като много хора пишат на български език, но с латински букви, за да не се загубят тези отзиви, те бяха преобразувани на кирилица.;
5. Премахване на e-mail адреси или уеб сайтове;
6. Премахване на стоп думи - Това са група от думи и предлози, които не носят значение при изследването на текстовете, но в действителност са най-срещаните думи в текстовете, като: в, на, под, всичко, няколко и т.н. Те не се отразяват на откриването на настроение в текста и трябва да се премахнат.
7. Stemming или lemming

Последните две стъпки в предварителната обработка на данните са препоръчителни, независимо от езика, на който са коментарите, но прилагането им за български текстове отново е свързано с трудности. Това са премахването на „стоп думите“ (изчистването на текстовете от думи, които не носят смисъл и значение за предстоящия анализ, като предлози, междуметия,

местоимения и паразитни думи) и т.н „stemming” или “lemming” (довеждане на думите до техните основни форми или корени, за да се унифицират). И при двете почти няма готови приложения или бази, които да могат да се използват (за разлика от английския език, например, за които има множество списъци със „стоп думи“ и програми за превръщането на думите в техните корени). Това наложи създаването на собствен списък от „стоп думи“, които да бъдат премахнати от коментарите. Преди тяхното премахване, 30те най-често срещани думи в База 1 имат следния вид Таблица 2.



Фиг. 1 - Най-често срещани думи в База 1 преди премахване на "стоп думи"

дума	брой срещания	дума	брой срещания
\n	31727	**има**	2733
и	28309	**не**	2722
на	15239	**са**	2700
е	13887	**-**	2269
много	10712	**хотела**	2219
беше	9103	**изключително**	2182
за	8883	**място**	1957
в	8221	**любезен**	1949
с	7369	**стаята**	1862
от	5441	**чисто**	1662
се	4847	**към**	1611
всичко	4498	**ми**	1562
да	4150	**че**	1536
до	3829	**ще**	1506
ни	2826	**си**	1495

Таблица 2 - 30те най-често срещани думи в База 1 преди премахване на „стоп думи“

Вижда се, че сред най-често срещаните думи има твърде много предлози, местоимения и други части на речта, които не са необходими за настоящото изследване. Поради тази причина се налага изчистването на тези стоп-думи. В момента от разработването на настоящата дисертация, в който

се наложи необходимостта от използване на списък със стоп-думи, такъв съществуващ списък не беше известен на автора. Това и наложи създаването на такъв самостоятелно. За целта бяха използвани списък от думи на български и английски език, които се игнорират от търсещите машини, [7] и ръчно добавени думи, предлози, местоимения, препинателни знаци, различни времена на спомагателния глагол съм и други. **В резултат авторът създаде собствен списък от „стоп думи“, който може да се използва в обработката на данни на български език в бъдещи анализи и изследвания и може да се види в Приложение 1.**[108]

След премахването на думите от този списък се получи следният резултат от най-често срещани думи в списъка с положителните коментари - Таблица 3.

дума	брой срещания	дума	брой срещания
място	3781	**закуската**	1681
чисто	3541	**центъра**	1547
хотела	3290	**отзивчив**	1534
персонал	3272	**храната**	1509
любезен	2790	**персоналът**	1492
стаята	2442	**близо**	1477
изключително	2250	**закуска**	1467
добра	1956	**препоръчвам**	1443
персонала	1922	**уютно**	1378
локация	1918	**ниво**	1358
местоположение	1876	**стаите**	1348
хотел	1831	**вкусна**	1329
местоположението	1792	**спокойно**	1283
добро	1730	**хареса**	1265
удобно	1722	**отлично**	1259

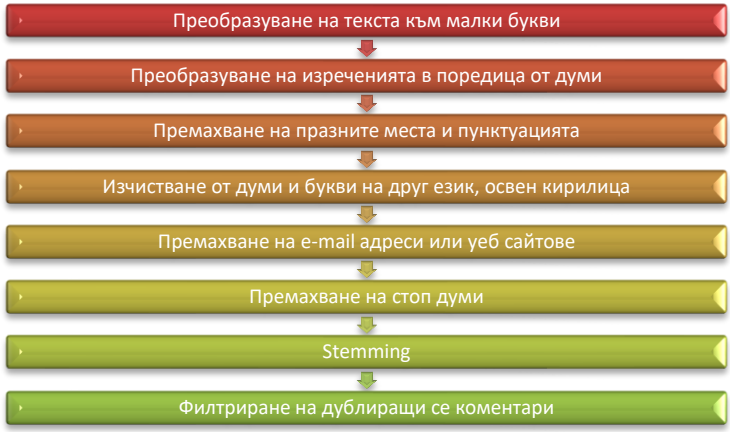
Таблица 3 - 30те най-често срещани думи в База 1 след премахване на „стоп думи“

На база направените корекции върху базите от данни, авторът предлага следната последователност в почистването и предварителната обработка на данни на български език за използването им за анализ на мнения и чувства.

След прилагане на гореспоменатите стъпки на предварителна обработка, размерът на двете БД се намали с по около 10 000 записа (Таблица 4):

Коментари	База 1	База 2
Положителни	63 714	84 489
Отрицателни	25 624	14 357
Общо	**89 341**	**98 846**

Таблица 4 - БД след предварителната обработка



Фиг. 2 - Последователност в почистване и предварителната обработка на данни на български език

Направена е последваща унификация на двете БД, за да бъдат еднакви началните условия за прилагане на различните методи. Това се налага за да могат да бъдат коректно сравними резултатите, които дават изследваните методи върху двете БД. За целта са уеднаквени броят на положителните коментари в двете бази и съответно този на отрицателните. Новополучените БД са с по-малко брой записи, но наличието на еднакъв брой коментари в двете бази позволява правенето на по-точни изводи при анализа на текстуалната сложност, както и при анализа на мнения (Таблица 5).

Коментари	База 1	База 2
Положителни	63 714	63 714
Отрицателни	14 357	14 357
Общо	**78 071**	**78 071**

Таблица 5 - БД след предварителна обработка и унификация

Преди да се прилагат различните методи за класификация данните трябва да се разделят на данни за обучение на моделите и данни за тестване. В разглежданата литература, най-често срещаните съотношения в разделянето на данните са 80%-20% и 70%-30%. За да се прецени кое е най-подходящото съотношение, всички изчисления, свързани с текущата дисертация са извършвани със следните разпределения на данните (Таблица 6):

Обучителни данни	Тестови данни
70% (54 649)	**30%** (23 421)
80% (62 456)	**20%** (15 614)

Таблица 6 - Разпределение на данните на обучителни и тестови

3.1. Анализ на текстуалната сложност

С цел сравняването на резултатите от двете БД беше извършен анализ на текстуалната сложност. Беше приложена адаптираната за български език формула на индекса Gunning Fog и към двете БД. Първо бяха преброени думите, изреченията и сложните думи във всяко потребителски мнение. След това се изчислява F (мярка за четливост на текст) за всеки коментар и накрая се изчислява среден индекс за всяка база данни. Резултатите са показани в следващата таблица (Таблица 7):

Gunning Fog	База 1	База 2
F	11.32	12.40

Таблица 7 - Резултати от анализ на текстуалната сложност

Разликата в резултатите не е голяма, но според дефиницията на Boehm те са поставени в два различни сегмента: База 1 с F под 12 – за бизнес документи и База 2 с F над 12 – за софтуерна документация. Изглежда, че в База 2 има по-сложни текстове, но в същото време дава по-добра крайна точност на анализа на настроенията, което е донякъде обърквашо. Това провокира автора да проучи допълнително съдържанието и сложността на всяка база данни.

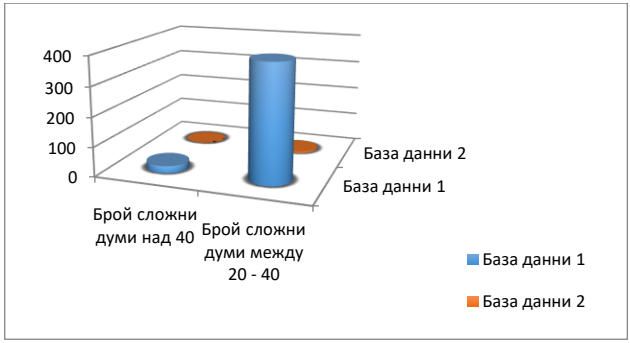
При изчисляването и преброяването на сложните думи и изречения във всеки коментар във всяка база данни се откроиха интересни резултати. Те са обобщени в Таблица 8 и фигура 15:

	База 1	База 2
Брой изречения над 20	22	0
Брой изречения между 10 - 20	732	51
Брой 1 изречение	43 789	43 986
Брой сложни думи над 40	24	0
Брой сложни думи между 20 - 40	397	4
0 на брой сложни думи	14 219	11 720

Таблица 8 - Брой изречения и брой сложни думи в изречения

Очевидно е, че в База 1 има коментари с повече изречения в тях в сравнение с База 2: База 1 съдържа **754** потребителски мнения с повече от 10 изречения във всеки коментар, докато База 2 - само **51**. По отношение на

съдържанието на сложни думи База 1 отново води с **421** коментара с повече от 20 сложни думи в тях, докато в База 2 има само **4** такива. Пакетите от думи, използвани в двете БД, също се различават - в База 1 има значително по-сложни думи в сравнение с тези в База 2. Общият брой сложни думи в База 1 е 793 174, докато в База 2 е 606 947.



Фиг. 3 - Разпределение на брой сложни думи в двете БД

Допълнително, се забелязва значителна разлика и в броя на всички думи в двете БД, отново в превес за База 1. Всички думи в База 1 са 831 469, а в База 2 - 638 104. Така корпусът от думи на База 1 е по-голям с приблизително 200 00 думи. (Таблица 9)

База 1 – брой сложни думи
793 174 от 831 469 думи общо

База 2 – брой сложни думи
606 947 от 638 104 **думи общо**

Таблица 9 – Брой сложни думи в двете БД

Поради болшинството други коментари с 1 изречение и без сложни думи, както и вследствие на осредняването на получените резултати, разликата между F на двете БД не е значителна, но след по-задълбочен поглед в пакетите от думи на двата набора от данни, може да се види реалната разлика между тях.

Изводи към Глава Трета

- Поради липсата на подходящ БД на български език за реализация на целта на дисертационния труд, авторът е разработил две собствени БД с коментари на български език, които могат да се ползват свободно за бъдещи изследвания и анализи, със съответно 89 341 и 98 846 коментара.
- Разработен е собствен списък от „стоп думи“ за целите на предварителната обработка на данните и с възможност да се използва в бъдещи изследвания.
- Предложен е алгоритъм за почистване и предварителна обработка на данни на български език за използването им в разработения от автора подход за анализ на мнения и чувства, представен в Глава 4.
- Извършен е анализ на текстуалната сложност на двете БД, който да аргументира откритите разлики в резултатите, получени от двете БД.

Глава Четвърта: Експериментални изследвания

Глава 4 съдържа експерименталните изследвания, извършени от автора. Описани са всички стъпки, през които е преминал докторантът и резултатите, които е получил.

4.1. Прилагане на различни методи върху БД на български език

Naïve Bayes класификатор

Преди да се приложи наивният класификатор на Бейс върху обработените данни, те се размесват произволно и се разделят на обучителни и тестови данни. Както бе вече бе уточнено, за да се провери кое разпределение на данните дава по-точни резултати, всички изчисления са правени в два варианта: 70%-30% и 80%-20% съответно за обучителни и тестови данни. Допълнително, изчисленията са извършени както върху данни, които не са преминали stemming, така и върху данни преминали през stemming в предварителната им обработка.

В Таблица 10 са изведени резултатите от изчисленията, като са дадени осреднени стойности за всяко едно разпределение.

Разпределение на данните	База 1 Accuracy (точност)	База 2 Accuracy (точност)
70%-30%	**0,841**	**0,864**
80%-20%	0,827	0,861

Таблица 10 - Резултати с Naive Bayes класификатор без stemming

Разпределение на данните	База 1 Accuracy (точност)	База 2 Accuracy (точност)
70%-30%	**0,836**	**0,858**
80%-20%	0,822	0,855

Таблица 11 - Резултати с Naive Bayes класификатор след stemming

От направените изчисления може да се заключи, че по-добри резултати дава разпределение на данните 70%-30%, както и обработването на данни без stemming. Освен това се вижда, че База 2 има с около 2% по-високи резултати спрямо База 1 и в двата варианта на разпределение на данните.

Логистична регресия

Изчисленията са извършени по няколко алгоритъма:

1) Count Vectorizer → логистична регресия
2) Премахване на наставки → Count Vectorizer → логистична регресия
3) Tf-idf векторизатор → логистична регресия
4) премахване на наставки → Tf-idf векторизатор → логистична регресия

5) Премахване на наставки → Count Vectorizer → прилагане на механизъм на внимание → логистична регресия
6) Премахване на наставки → Tf-idf векторизатор → прилагане на механизъм на внимание → логистична регресия

За всички алгоритми са изпробвани различни комбинации от основните параметри на методите. За да се провери дали се откриват връзки между отделните думи, при векторизирането на данните бе използвана възможността за образуване на двойки думи (би-грами), след което отново бе приложена логистична регресия.

Данни	Count Vectorizer				Tf-idf			
	Stemmed		Not stemmed		Stemmed		Not stemmed	
	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams
База 1	**0,889**	0,887	0,889	**0,895**	0,898	**0,901**	0,897	**0,899**
База 2	0,928	**0,933**	0,925	**0,933**	0,936	**0,940**	0,934	**0,940**

Таблица 12 - Обобщение на резултатите след логистична регресия

и разпределение на данните 70%-30%

За да се направи опит да се повишат резултатите от модела, е приложен и механизъм за внимание. Механизмът за внимание е техника, която се използва широко в машинното обучение и задачите по обработка на естествения език, за да се зададат различни нива на важност или внимание към различни части от входните данни. В контекста на класификация на текст, механизмът за внимание може да помогне на модела да се фокусира върху конкретни думи или характеристики, които са по-релевантни за точни прогнози.

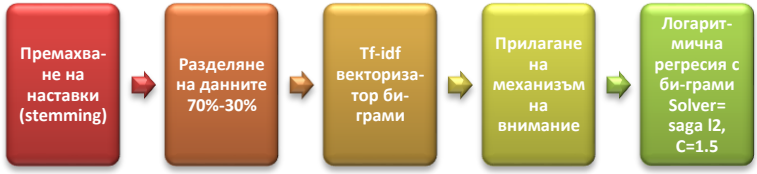
В Таблица 13 са обобщени резултатите, които се получават след прилагането на LogisticRegression('l2',C=1.5,solver='saga') в комбинация с механизъм за внимание.

Данни	Count Vectorizer				Tf-idf			
	Stemmed		Not stemmed		Stemmed		Not stemmed	
	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams
База 1	0,893	**0,897**	0,893	0,896	0,899	**0,901**	0,897	**0,898**
База 2	0,929	**0,939**	0,930	**0,936**	0,936	**0,941**	0,933	**0,937**

Таблица 13 - Обобщени резултати след приложен механизъм на вниманието, разпределение на данните 70%-30%

Въпреки че разликата е минимална, резултатите са повишили точността си след прилагането на механизъм на вниманието. Останалите тенденции в точността са се запазили.

От направените изследвания може да се заключи, че най-добри резултати за извеждането на мнения чрез логистична регресия дават следните алгоритмични стъпки:



Фиг. 4 - Алгоритмични стъпки за извеждането на мнения чрез логистична регресия

Метод на опорните вектори

За разработения модел на метода на опорните вектори са изследвани две различни функции Linear SVC(Support Vector Classificator) и C-Support Vector Classification, както и няколко параметъра, които функциите приемат. Първият **параметър е C** (regularization parameter) - параметър за регулиране. Силата на регуляризацията е обратно пропорционална на C, като за C са взети следните стойности [0,01, 0,05, 0,25, 0,5, 1, 1.5]. Другият **параметър е kernel**, който указва типа на ядрото, което да се използва в алгоритъма: [squared hinge и hinge] за LinearSVC и съответно [linear, poly, sigmoid, rbf] за SCV.

Използвано е същото разделение на данните, както при логистичната регресия и отново изчисленията са извършени чрез няколко алгоритъма, с и без прилагането на би-грами:

1) Count Vectorizer → метода на опорните вектори
2) Премахване на наставки → Count Vectorizer → метода на опорните вектори
3) Tf-idf векторизатор → метода на опорните вектори
4) Премахване на наставки → Tf-idf векторизатор → метода на опорните вектори
5) Премахване на наставки → Count Vectorizer → прилагане на механизъм на внимание → метода на опорните вектори
6) Премахване на наставки → Tf-idf векторизатор → прилагане на механизъм на внимание → метода на опорните вектори

Резултатите от обучаването на системите чрез метода на опорните вектори са обобщени в следващата Таблица 14, като са изведени най-

високите резултати спрямо различните настройки на параметрите, изследвани и анализирани по-горе. Тъй като премахването на наставките не дава големи разлики в резултатите, то е приложено само за метода и параметрите, които дават най-висок резултат и това са linear SVC, kernel=hinge при разделяне на данните 70%-30%. Изследван е и другият векторизатор Count Vectorizer, но тъй като очакваните стойности са по-ниски, с него отново са направени изчисления само за параметрите, които дават най-добри резултати. Данните са обобщени в таблица 30.

Данни	**Count Vectorizer**				**Tf-idf**			
	Stemmed		Not stemmed		Stemmed		Not stemmed	
	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams
База 1	0,**886**	0,885	0,882	0,881	0,896	**0,898**	0,892	0,894
База 2	0,923	**0,930**	0,919	0,926	0,929	**0,937**	0,925	0,933

Таблица 14 – Обобщени резултати след SVM, разпределение на данните 70%-30%

Отново резултатите са със сходни стойности, в зависимост от използвания векторизатор и би-грами, но разликата между резултатите на двете БД е значителна с предимство за База 2. Най-високи стойности се постигат чрез използването на stemming в предварителната обработка и прилагане на Tf-idf векторизатор с би-грами, съответно 89,9% за База 1 и 93,7% за База 2.

За да се повишат резултатите от модела отново е приложен и механизъм за внимание. Механизмът за внимание помага на модела да се фокусира върху конкретни думи или характеристики, които са по-релевантни за точни прогнози. За спецификите на данните и конкретната задача за бинарна класификация,бе разработен персонализиран механизъм на внимание, специално за метода на опорните вектори. Този механизъм присвоява тегла на вниманието към характеристики (features) (или атрибути) на входните данни въз основа на тяхната важност за SVM класификатора.

В Таблица 15 са обобщени резултатите, които се получават след прилагането на **linear SVC, kernel=hinge** в комбинация с механизъм за внимание.

Данни	**Count Vectorizer**				**Tf-idf**			
	Stemmed		Not stemmed		Stemmed		Not stemmed	
	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams	Uni-grams	Bi-grams
База 1	0,**890**	0,887	0,886	0,887	0,898	**0,901**	0,898	0,899
База 2	0,928	**0,931**	0,924	**0,**932	0,936	**0,940**	0,934	0,939

Таблица 15 - Обобщени резултати след attentionSVM, разпределение на данните 70%-30%

Ако се сравнят Таблица 14 и Таблица 15, може да се види, че след прилагане на механизмите за внимание, резултатите се повишават с около 1%, като най-добрите резултати за База 1 са вече над 90%, а за База 2 – 94%. Останалите тенденции за разлика между двете БД, както и различията между приложени stemming и би-грами, се запазват.

Най-добри резултати чрез метод на опорните вектори дават следните алгоритмични стъпки, които авторът предлага за използване при изследване на коментари на български език:

Фиг. 5 - Алгоритмични стъпки за извеждането на мнения чрез SVM

Random forest algorithm

За прилагането на алгоритъма Random forest отново са използвани двата вида разпределние на данните 70%-30% и 80%-20%. Бяха извършени множество изчисления за оптимизиране на алгоритъма чрез различни комбинации от параметри, в резултат на които комбинацията от следните параметри дава най-високи резултати: **n_estimators=100, max_depth=80, min_samples_leaf=5, min_samples_split=12**.

Данни 70%-30%	**Random forest**	
	Stemmed	Not stemmed
База 1	**0,833**	0,825
База 2	**0,841**	0,829

Таблица 16 - Резултати с алгоритъм Random forest при данни 70%-30%

Данни 80%-20%	**Random forest**	
	Stemmed	Not stemmed
База 1	**0,827**	0,818
База 2	**0,836**	0,827

Таблица 17 - Резултати с алгоритъм Random forest при данни 80%-20%

От Таблица 16 и Таблица 17,в които са поместени резултатите от изчисленията за двете бази и двете разпределения, се вижда, че по-високи резултати дава разпределение 70%-30%, както и данните преминали stemming.

Рекурентни невронни мрежи

За обучаването на невронната мрежа бяха използвани три слоя:

- Слой за вграждане (Embedding Layer)
- Двупосочен LSTM слой (Bidirectional LSTM)
- Плътен слой (Dense Layer)

Данни 70%-30%	**RNN**	
	Stemmed	Not stemmed
База 1	**0,8881**	0,8878
База 2	**0,9303**	0,9317

Таблица 18 - Резултати с невронни мрежи, разпределение на данните 70%-30%

Данни 80%-20%	**RNN**	
	Stemmed	Not stemmed
База 1	**0,912**	0,909
База 2	**0,944**	0,943

Таблица 19 - Резултати с невронни мрежи, разпределение на данните 80%-20%

Резултатите и за двете БД са най-високите спрямо всички предходни модели – 91,2% акуратност за База 1 и 94,4% за База 2. Тенденцията да има разлика между двете бази се запазва, макар и тук да е по-малка – едва 3%. С малко по-високи стойности са резултатите при разпределение на данните на 80% за обучение и 20% за тестване. Както и предварително обработените данни чрез stemming, макар и с малко, дават по-високи резултати.

4.2. Метод на ансамбъла att_SVM+biLSTM+lex_RF за коментари на български език

Като следваща стъпка е предложен нов подход, обединяващ два метода за машинно обучение, който е изпълнен чрез метод на ансамбъла, наречен Stacking. Той включва комбинирането SVM, към който е приложен механизъм на внимание, с biLSTM, и тяхното обединяване чрез stacking с RF мета-класификатор, в който частично в включен и метод на лексикона **(att_SVM+biLSTM+lex_RF**). Stacking представлява обучението на няколко модела за машинно обучение и комбинирането на техните предсказвания с цел подобряване цялостната точност. За целта е направено сравнения с три комбинации от методи, които са прилагани за текстове на английски език. [13, 54]

- SVM и RF;
- SVM и biLSTM;
- SVM и biLSTM с механизъм на внимание;

Избрани са тези три комбинации, тъй като това са моделите, с които се постига най-висока точност в изследванията до момента. Използвани са параметрите за SVM, RF и RNN, които дават най-добри резултати в предходните изчисления.

SVM и RF:

В първата комбинация, се извършва TF-IDF векторизация на обучаващите данни. Обучават се отделните модели SVM (с LinearSVC функция и параметри kernel= hinge и C=1) и RF случаен горски класификатор и чрез използване на крос-валидация се комбинират техните прогнози. Прилага се мета-класификатор от тип логистична регресия върху комбинираните прогнози.

Фиг. 6 – Метод на ансамбъла SVM+ RF

SVM + Random forest				
Данни 70%-30%	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
База 1	**0,902**	**0,903**	0,892	0,899
База 2	0,942	**0,943**	0,931	0,939

Таблица 20 - Резултати след прилагане на Комбиниран модел SVM и Random forest, при разпределение на данните 70%-30%

SVM + Random forest				
Данни 80%-20%	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
База 1	**0,903**	**0,904**	0,892	**0,903**
База 2	0,942	**0,943**	0,933	0,939

Таблица 21 - Резултати след прилагане на Комбиниран модел SVM и Random forest, при разпределение на данните 80%-20%

От Таблица 20 и Таблица 21 се вижда, че разликата между двата вида разпределяне на данните за обучение и тестване, е пренебрежимо малка. Забелязва се и малка разлика между данните, преминали stemming и тези, които не са, в полза на необработването на данни. Най-високи резултати,

обаче, генерира комбинацията от би-грами и неприлагане на stemming, съответно 90,3% за База 1 и 94,3% за База 2.

SVM и biLSTM:

При втората комбинация от модели са използвани SVM с LinearSVC функция и параметри kernel= hinge и C=1 и двуслоен LSTM модел със следните слоеве:

- Слой за вграждане (Embedding Layer): Този слой преобразува входните цели числа в плътни вектори с фиксиран размер (в този случай 32) и представя всяка дума във входната последователност.
- Bidirectional LSTM слой (първи слой): Този слой обработва входната последователност, както в направление напред, така и назад, и връща последователности за всяка времева стъпка.
- Bidirectional LSTM слой (втори слой): Подобно на първия слой, този слой също обработва последователностите в две посоки, но не връща последователности за всяка времева стъпка. Вместо това той предоставя обобщено представяне на цялата последователност.
- Dense слой: Този слой има 64 units (юнита) и прилага функцията на активация ReLU. Той въвежда нелинеарност в модела и помага за научаването на сложни модели в данните.
- Dense слой (Изходен слой): Този слой има единичен юнит със сигмоидна функция на активация, която извежда вероятност между 0 и 1, представяща предсказанието за настроението на входната последователност.

Като цяло, този модел използва два LSTM слоя с бидирекционално обработване, за да се улавя информацията от последователността в двете посоки. Dense слоевете помагат за допълнителна обработка и правене на предсказания на база на научените представяния.

След като и двата модела са обучени, техните прогнози се комбинират с помощта на метакласификатор, използващ логистична регресия с параметри l2, solver='saga', C=1.

Фиг. 7 - Метод на ансамбъла SVM+ biLSTM

Отново са направени изчисления, както за разпределяне на данните в съотношение 80%-20% и 70%-30%, така и за би-грами и униграми.

SVM + biLSTM model				
Данни 70%-30%	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
База 1	0,897	**0,898**	0,896	**0,898**
База 2	0,939	**0,948**	0,936	**0,942**

Таблица 22 - Резултати след прилагане на комбиниран модел SVM + biLSTM с разпределение на данните 70%-30%

SVM + biLSTM model				
Данни 80%-20%	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
База 1	**0,900**	0,897	0,899	**0,902**
База 2	0,940	**0,949**	0,935	**0,943**

Таблица 23 – Резултати от прилагане на комбиниран модел SVM + biLSTM с разпределение на данните 80%-20%

От Таблица 22 и Таблица 23 може да се направи извод, че макар и с малка разлика в резултатите, разпределението на данните 80%-20%, съответно за обучение и тестване, дава по-високи крайни резултати. Освен това, прави впечатление, че за разлика от другите методи, комбинацията между SVM и biLSTM, където данните не са претърпели stemming, дава по-високи резултати от тези, където предварително са премахнати наставките на думите. Тази тенденция се запазва и за двете БД, независимо дали се работи само с единични думи (uni-grams) или с комбинация от двойки думи (bi-grams).

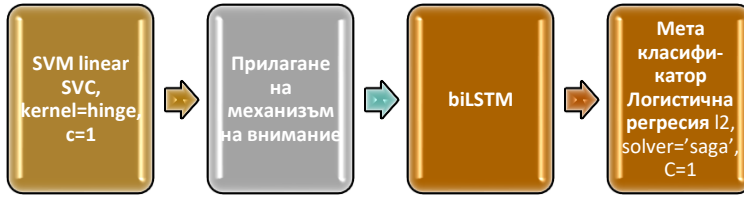
SVM и biLSTM с механизъм на внимание

В предложения модел се комбинират SVM и biLSTM, като в рекурсивната невронна мрежа се използва персонализиран механизъм за внимание, който представлява форма на базирано на съдържание внимание. Конкретно, подходът реализира механизъм за внимание със следните характеристики:

- Механизмът за внимание се реализира като персонализиран слой в Keras, наречен `Attention`.
- Механизмът използва невронна мрежа с пряк пропуск, за да изчисли оценки за внимание между векторите за заявка (скрито състояние) и ключовите вектори (кодирана входна последователност).

- Оценките за внимание се трансформират с помощта на функция софтмакс, за да се получат теглата за внимание.
- Контекстният вектор се изчислява като умножение на сумата на векторите за стойност (кодирана входна последователност) с теглата за внимание.
- Контекстният вектор се използва в предсказанието на модела.

Допълнително са изследвани различните комбинации от активанционни функции в двата слоя на невронната мрежа, за да се открие комбинацията с най-високи резултати. Подобен анализ на настройката на хиперпараметрите е направил Amankumar в своята разработка върху коментари на английски език.[12]



Фиг. 8 - Метод на ансамбъла SVM+ att_biLSTM

SVM + att_biLSTM model 70%-30% База 1				
Activation Function	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
activation='relu' activation='sigmoid'	**0,9063**	0,9028	0,9005	0,8978
activation='softmax' activation='sigmoid'	**0,9053**	0,9013	0,8998	0,8978
activation='softmax' activation='softmax'	**0,9052**	0,9039	0,8998	0,9000

Таблица 24 – SVM+ att_biLSTM модел 70%-30% БД1

SVM + att_biLSTM model 80%-20% База 1				
Activation Function	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
activation='relu' activation='sigmoid'	**0,9075**	0,9051	0,9062	0,9017
activation='softmax' activation='sigmoid'	**0,9073**	0,9085	0,9026	0,9041
activation='softmax' activation='softmax'	**0,9055**	0,9069	0,9017	0,9014

Таблица 25– SVM + att_biLSTM модел 80%-20% БД1

SVM + att_biLSTM model 70%-30% База 2				
Activation Function	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
activation='relu' activation='sigmoid'	**0,9416**	0,9385	0,9378	0,9369
activation='softmax' activation='sigmoid'	**0,9424**	0,9416	0,9375	0,9373
activation='softmax' activation='softmax'	**0,9432**	0,9431	0,9379	0,9369

Таблица 26 – SVM + att_biLSTM модел 70%-30% БД2

SVM + att_biLSTM model 80%-20% База 2				
Activation Function	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
activation='relu' activation='sigmoid'	**0,9423**	0,9377	0,9412	0,9371
activation='softmax' activation='sigmoid'	**0,9421**	0,9411	0,9414	0,9351
activation='softmax' activation='softmax'	**0,9433**	0,9426	0,9375	0,9380

Таблица 27 – SVM + att_biLSTM модел 80%-20% БД2

От резултатите, поместени в таблици 45-48, може да се види, че съществена разлика в точността за двата вида разпределение на данните няма. Комбинацията от би-грами и прилагане на stemming отново дават най-високи резултати. Наблюдава се разлика за двете бази в комбинациите от активационни функции, който дават по-висока точност. Докато за База 1 комбинацията от *relu* и *sigmoid* генерират най-високи резултати – **90,75%,** то за База 2 тази комбинация е от *softmax* и за двата слоя – **94,33**%.

Анализ на настроения и характеристики за настроения чрез SVM с механизъм на внимание + biLSTM и мета класификатор RF (att_SVM+biLSTM+lex_RF).

В разработения алгоритъм отново се комбинират SVM и BILSTM, но този път механизмът на внимание е включен в SVM модела. Тъй като обикновено механизмите на внимание са част от невронни мрежи, за да се имплементира сходен механизъм като част от SVM, е дефиниран специален клас, който комбинира метода на опорните вектори с механизъм за внимание. Методът fit на този клас обучава модел на линеен SVM, изчислява теглата на

внимание въз основа на коефициентите на SVM и съхранява обучения модел. След което методът predict прави предсказания с изучения SVM модел.

След задълбочен анализ на резултатите от предишните изчисления бе открито, че логистичната регресия, използвана като мета класификатор, дава по-ниски крайни резултати спрямо използването на двата модела преди нея самостоятелно. Например, за База 2 SVM дава 0,9399, BILSTM - 0,9419, а след тяхното обединяване чрез логистична регресия, крайната точност е по-ниска . 0,9389. Това доведе до търсенето на други варианти за мета класификатори. Бяха изследвани Gradient Boosting Classifier и Random Forest Classifier. Оказва се, че и двата се представят по-добре от логистичната регресия, но най-добри резултати дава Random Forest Classifier.

Допълнително, преди прилагането на мета-класификатора, е интегриран частично лексиконен подход, като изчислява оценки на лексикона за всички коментари и ги комбинира с прогнозите на моделите SVM и biLSTM. Комбинираните данни се използват за обучение на мета-класификатор от тип Random Forest, за да се направят окончателни прогнози. Този подход използва както прогнозите на моделите, така и оценките на лексикона, за да подобри точността на класификацията. Интегрираният лексиконен подход е частично приложен само за негативни думи, тъй като в процеса на изследване бе открит само лексикон с думи, носещи отрицателен смисъл [91]. Използването на този лексикон донякъде компенсира значително по-малкия брой отрицателни коментари в двете БД спрямо положителните.

Резултатите от предложения алгоритъм **att_SVM+biLSTM+lex_RF** са описани в Таблица 28, Таблица 29, Таблица 30, Таблица 31. За да се види положителният ефект от комбинирането на двата метода, в таблицата са включени и резултатите, които двата метода дават поотделно, преди тяхното обединяване чрез мета класификатора Random Forest.

att_SVM+biLSTM+lex_RF 70%-30% База 1				
Models	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
SVM	0.9006	0.8990	0.8981	0.8976
LSTM	0.9008	0.8963	0.9001	0.8977
Random Forest	**0.9125**	0.9051	0.9081	0.9079

Таблица 28 - att_SVM+biLSTM+lex_RF модел 70%-30% БД1

att_SVM+biLSTM+lex_RF 80%-20% База 1				
Models	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
SVM	0.9009	0.8988	0.8993	0.8985
LSTM	0.9008	0.8962	0.9001	0.8974
Random Forest	**0.9105**	0.9055	0.9044	0.9051

Таблица 29 - att_SVM+biLSTM+lex_RF модел 80%-20% БД1

att_SVM+biLSTM+lex_RF 70%-30% База 2				
Models	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
SVM	0.9403	0.9399	0.9365	0.9339
LSTM	0.9421	0.9396	0.9359	0.9411
Random Forest	**0.9450**	0.9433	0.9414	0.9415

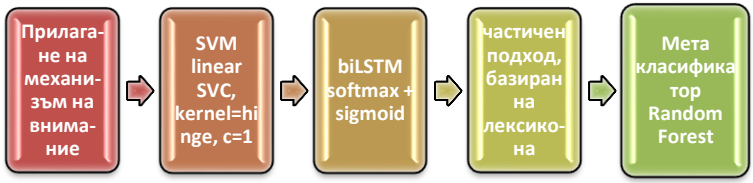
Таблица 30 - att_SVM+biLSTM+lex_RF модел 70%-30% БД2

att_SVM+biLSTM+lex_RF 80%-20% База 2				
Models	Bi-grams		Uni-grams	
	Stemmed	Not stemmed	Stemmed	Not stemmed
SVM	0.9397	0.9409	0.9362	0.9346
LSTM	0.9396	0.9368	0.9381	0.9357
Random Forest	**0.9475**	0.9435	0.9452	0.9407

Таблица 31 - att_SVM+biLSTM+lex_RF модел 80%-20% БД2

В таблици 49-52 се откроява подобрение в крайната точност на предложения модел. За първи път за База 1 се наблюдава резултат над 91% точност. Отново прилагането stemming и използването на би-грами дава по-висока точност. Но на база на резултатите не може да се направи твърдо заключение кое разпределение на данните е по-подходящо - докато База 1 дава по-високи резултати при разпределение 70%-30%, то База 2 постига малко по-високи резултати при разпределение 80%-20%. Така за База 1 най-високата точност **91,25%** е постигната при разпределение 70%-30%, stemming и би-грами, при База 2 при разпределение 80%-20%, stemming и би-грами най-високите резултати са **94,75%.**

Най-добри резултати чрез метод на ансамбъла дават следните алгоритмични стъпки, които авторът предлага за използване при изследване на коментари на български език att_SVM+biLSTM+lex_RF:



Фиг. 9- Предложени алгоритмични стъпки за анализ на настроенията, чрез метод на ансамбъла att_SVM+biLSTM+lex_RF

Изводи към Глава Четвърта

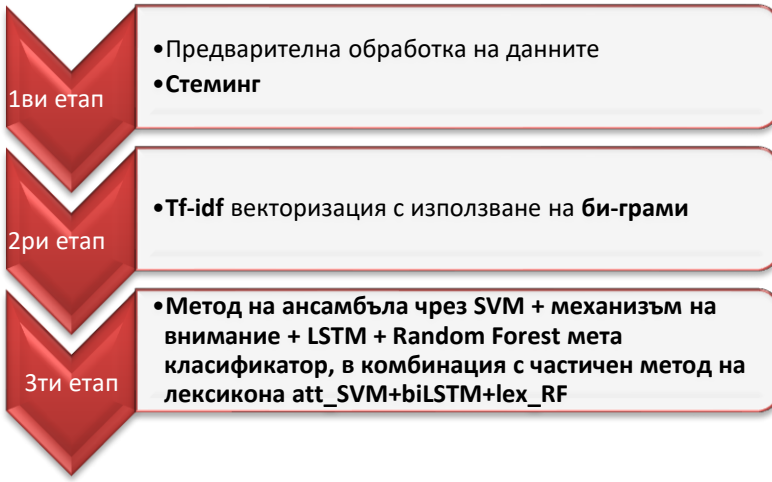
- База 2 дава с около 4% по-високи резултати спрямо База 1, което произтича от по –големия брой дълги изречения и сложните думи в База 1;
- Прилагането на Tf-IDF векторизатор връща по-високи стойности спрямо Count Vectorizer, което обуславя неговото прилагане в разработения алгоритъм;
- Използването на би грами спрямо униграми в Tf-IDF векторизатора води до по-високи резултати, поради което те са включен в разработения алгоритъм;
- В болшинството от случаите прилагането на stemming повишава точността на резултатите, което обуславя неговото прилагане в разработения алгоритъм;
- Най-висока точност и при двете БД генерира метода biLSTM, което определя използването му в предложения алгоритъм.

Резултатите от всички изчисления, постигнати посредством използваните модели, са обобщени в **Error! Reference source not found.**. Най-точен за класифициране на текстове спрямо тяхното настроение се оказва предложеният от автора алгоритъм **att_SVM+biLSTM+lex_RF**, който дава точност съответно **91.3%** за База 1 и **94.8%** за База 2.

Модел на Класификация Tf –IDF Bi-grams	База 1		База 2	
	Без stemming	Stemming	Без stemming	Stemming
Naïve Bayes	0,841	0,836	0,866	0,858
Логистична Регресия	0.899	0.901	0.940	0.940
Логистична Регресия +att	0,898	0,901	0,937	0,941
Support Vector Machine	0.894	0,898	0,933	0,937
SVM+ att	0,899	0,901	0,939	0,940
Random forest	0,859	0,871	0,902	0,914
LSTM	0,904	**0,907**	0,941	**0,946**
SVM + RF	0,904	0,903	0,943	0,942
SVM+LSTM 80-20	0,897	0,900	**0,949**	0,940
SVM+LSTM+attention	0,905	**0,907**	0,942	**0,943**
att_SVM+biLSTM+lex_RF	0.905	**0.913**	0.944	**0,948**

Таблица 32 - Обобщение на резултатите от изчисленията

От тези резултати може да се изведе обобщен подход за обработка и изследване на данните, който дава най-голяма точност, на база извършените от автора изчисления, при предсказване на мнението на потребителите в текстове на български език:



фигура 1 - Алгоритъм за обработка и изследване на коментари на български език

Приноси по дисертационния труд

1. Разработени са две БД с мнения и коментари на български език (с приблизително по 100 000 коментара всяка). Авторът претендира да е създал двете най-големи БД с коментари на български език до момента, които могат да се използват и за бъдещи научни изследвания. Неговите изследвания са единствени по рода си, базирани на такива обширни набори от коментари на български език.
2. Разработен е списък от стоп думи, който може да се използва за предварителната обработка на текстове в бъдещи анализи и изследвания на текстове на български език.
3. Предложен е подобрен вариант на алгоритъма за създаване на база данни и предварителната й обработка в спецификите на българския език.
4. Осъществен е сравнителен анализ на точността на различни методи и алгоритми за анализ на мнения чрез класификация на текстове на български език.
5. Предложен е алгоритъм за извличане на мнение от коментари на български език, който дава най-добри резултати, спрямо останалите, извършени експериментални изчисления. Този алгоритъм представлява метод на ансамбъла, който комбинира SVM и LSTM, като в SVM модела е включен механизмът на внимание, и приложен мета класификатор Random Forest, в комбинация с частичен метод на лексикона.

Списък на публикациите по дисертацията

Списъкът на авторските публикации се състои от шест заглавия, три от които в съавторство, четири - индексирани по SCOPUS:

1. Д.Петрова, Overview of the methods used for opinion mining and sentiment analysis and their application in bulgarian language so far, Обзор на методите, използвани за анализ на мнения и извличане на чувства и тяхното приложение за български език до момента, Списание КОМПЮТЪРНИ НАУКИ И ТЕХНОЛОГИИ - Година XVIII, Брой 1/2020 – стр 126, ISSN 1312-3335

2. D.Petrova, Comparative assay on sentiment analysis on two databases in Bulgarian language, ICMECE 27-28.11.2021 Ankara, Turkey, ISBN:978-625-409-707-2, pp. 43-47.

3. D.Petrova, Automatic Sentiment Analysis on Hotel Reviews in Bulgarian – Basic Approaches and Results, IEMAICLOUD 26-28.04.2021 London,UK, ISBN:978-3-030-92904-6 pp.48-56. DOI: 10,1007/978-3-030-92905-3_5

4. Petrova, D., & Bojikova, V. (2022, December 31). Development of two data bases with comments in Bulgarian language and application of supervised learning approaches on them for comparative sentiment analysis. A brief overview. ANNUAL JOURNAL OF TECHNICAL UNIVERSITY OF VARNA, BULGARIA, 6(2), 57-62. https://doi.org/10,29114/ajtuv.vol6.iss2.261, ISSN :2603-316X

5. Petrova, D., Bozhikova V. (2022) Random forest and recurrent neural network for sentiment analysis on texts in Bulgarian language, International Conference on Biomedical Innovations and Applications, Varna, Bulgaria, 2-4 June 2022, **DOI:** 10,1109/BIA52594.2022.9831326, **Electronic ISBN:**978-1-6654-4581-8, 66-69стр

6. D.Petrova, V.Bozhikova (2022) Sentiment and complexity analysis on two databases in Bulgarian language – final estimation, ICECCME, 16-18 November, Maldives, DOI: 10,1109/ICECCME55909.2022, **Electronic ISBN:**978-1-6654-7095-7

Научно-изследователска работа по други договорни теми и задачи:

- **ПД6/2021г.:** „Автоматизиран анализ на мнения в текстове на български език", проект в помощ на докторанти с ръководител доц.д-р В.Божикова

Научноизследователски проекти:

- **НП12:** „Класификация на текст на български език чрез методи на машинното обучение ", 2022г. с ръководител доц.д-р Н. Калчева

Благодарности

Благодаря на научния си ръководител доц.др. Виолета Божикова, че ме прие за свой докторант, за търпеливото, задълбочено и многократно редактиране, за ценните съвети и за цялото отделено време и съвместна работа по настоящата дисертация, съвместните публикации и участия в научни проекти.

Бих искала също да изкажа благодарност на колегите си от катедра Софтуерни и Интернет Технологии, които през изминалите четири години ми помогнаха със съвети и напътствия.

Работата по настоящата дисертация се осъществи и благодарение на научно - изследователския проект ПД6/2021 в помощ на докторанти към Технически Университет – Варна.

Благодаря на семейството си за търпението и насърчаването в продължителната работа по тази дисертация.

Annotation

The main research goal of the dissertation is the creation of algorithms and approaches for the text mining and sentiment analysis of texts in Bulgarian - something that is missing or is in scarce quantities in the scientific studies published to date and discovered by the doctoral student in Bulgaria.

Two databases with opinions and comments in Bulgarian (with approximately 100,000 comments each) have been created to be used for the experiments in this dissertation. The author claims to have created the two largest databases with comments in Bulgarian to date, which can also be used for future scientific research. A list of stop words has also been created, which can be used for the preliminary processing of texts in future analyzes and researches of texts in the Bulgarian language.

An improved version of the algorithm for creating a database and its pre-processing in the specifics of the Bulgarian language has been developed.

An algorithm, called **att_SVM+biLSTM+lex_RF,** for sentiment analysis of comments in Bulgarian is proposed, which gives the best results compared to other experimental calculations. This algorithm is an ensemble method that combines Support Vector Machines (SVM) and Recurrent Neural Networks (LSTM), with the attention mechanism included in the SVM model, and then Random Forest as meta classifier, in combination with a partial lexicon method.