



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA
U NOVOM SADU




Срђан Стјепановић

**Класификација сајбер
насиља у тексту
употребом неуронских
мрежа и статистичких
модела машинског учења**

Дипломски рад
- Основне академске студије -

Нови Сад, 2024.

	УНИВЕРЗИТЕТ У НОВОМ САДУ ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА 21000 НОВИ САД, Трг Доситеја Обрадовића 6	Датум:
	ЗАДАТАК ЗА ИЗРАДУ ДИПЛОМСКОГ (BACHELOR) РАДА	Лист:
		1/1

(Податке уноси предметни наставник - ментор)

Врста студија:	Основне академске студије
Студијски програм:	Софтверско инжењерство и информационе технологије
Руководилац студијског програма:	проф. др Мирослав Зарић

Студент:	Срђан Стјепановић	Број индекса:	SV16/2020
Област:	Електротехничко и рачунарско инжењерство		
Ментор:	Др Јелена Сливка, ванредни професор		

НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА:

- проблем – тема рада;
- начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна;
- литература

НАСЛОВ ДИПЛОМСКОГ (BACHELOR) РАДА:

Класификација сајбер насиља у тексту употребом неуронских мрежа и статистичких модела машинског учења

ТЕКСТ ЗАДАТКА:

Изградити модел за класификовање текста на тип сајбер насиља техникама машинског учења: 1. Анализирати стање у области. 2. Изградити спецификацију захтева софтверског решења. 3. Изградити спецификацију дизајна софтверског решења. 4. Имплементирати софтверско решење према израђеној спецификацији. 5. Тестирати имплементирано софтверско решење. 6. Документовати (1), (2), (3), (4) и (5).

Руководилац студијског програма:	Ментор рада:

Примерак за: ☐ - Студента; ☐ - Ментора

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, РБР:	
Идентификациони број, ИБР:	
Тип документације, ТД:	монографска публикација
Тип записа, ТЗ:	текстуални штампани документ
Врста рада, ВР:	дипломски рад
Аутор, АУ:	Срђан Стјепановић
Ментор, МН:	др Јелена Сливка, ванредни професор
Наслов рада, НР:	Класификација сајбер насиља у тексту употребом неуронских мрежа и статистичких модела машинског учења
Језик публикације, ЈП:	српски
Језик извода, ЈИ:	српски / енглески
Земља публикавања, ЗП:	Србија
Уже географско подручје, УГП:	Војводина
Година, ГО:	2024
Издавач, ИЗ:	ауторски репринт
Место и адреса, МА:	Нови Сад, Факултет техничких наука, Трг Доситеја Обрадовића 6
Физички опис рада, ФО:	9 поглавља / 49 странице / 0 цитата / 8 табела / 4 слике / 2 графикона / 0 прилога
Научна област, НО:	Софтверско инжењерство и информационе технологије
Научна дисциплина, НД:	Софтверско инжењерство
Предметна одредница / кључне речи, ПО:	Класификација сајбер насиља, класификација текста, машинско учење
УДК	
Чува се, ЧУ:	Библиотека Факултета техничких наука, Трг Доситеја Обрадовића 6, Нови Сад
Важна напомена, ВН:	
Извод, ИЗ:	Рад представља решење класификације насилног текста на типове сајбер насиља помоћу класификационих модела машинског учења (метод потпорних вектора, метод насумичне шуме, <i>XGBoost</i> и <i>bagging</i> модела машинског учења), као и помоћу до-тренирања <i>BERT</i> трансформатора.
Датум прихватања теме, ДП:	
Датум одбране, ДО:	
Чланови комисије, КО:	
председник	др Милан Сегединац, ванредни професор
члан	др Жељко Вуковић, доцент
ментор	др Јелена Сливка, ванредни професор
Потпис ментора	

KEY WORDS DOCUMENTATION

Accession number, ANO :	
Identification number, INO :	
Document type, DT :	monographic publication
Type of record, TR :	textual material
Contents code, CC :	bachelor thesis
Author, AU :	Srdan Stjepanović
Mentor, MN :	Jelena Slivka, associate professor, PhD
Title, TI :	Cyberbullying text classification using neural networks and statistical machine learning models
Language of text, LT :	Serbian
Language of abstract, LA :	Serbian / English
Country of publication, CP :	Serbia
Locality of publication, LP :	Vojvodina
Publication year, PY :	2024
Publisher, PB :	author's reprint
Publication place, PP :	Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6
Physical description, PD :	9 chapters / 49 pages / 0 quotes / 8 tables / 4 pictures / 2 graph / 0 attachments
Scientific field, SF :	Software Engineering and Information Technologies
Scientific discipline, SD :	Software Engineering
Subject / Keywords, S/KW :	Cyberbullying classification, NLP, text classification
UDC	
Holding data, HD :	Library of the Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad
Note, N :	
Abstract, AB :	This paper presents a solution for cyberbullying text classification using statistical machine learning models (Support Vector Machines, Random Forest, XGBoost, and bagging), as well as fine-tuning the BERT transformer model.
Accepted by sci. Board on, ASB :	
Defended on, DE :	
Defense board, DB :	
president	Milan Segedinac, associate professor, PhD
member	Željko Vuković, assistant professor, PhD
mentor	Jelena Slivka, associate professor, PhD
Mentor's signature	

САДРЖАЈ

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА	4
KEY WORDS DOCUMENTATION	5
1. УВОД	9
2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ	11
3. ТЕОРИЈСКИ ПОЈМОВИ И ДЕФИНИЦИЈЕ	17
3.1 TF-IDF	18
3.2 GloVe	19
3.3 Метода потпорних вектора (SVM)	19
3.4 Стабло Одлучивања	21
3.5 <i>Bagging</i> ансамбл модел	22
3.6 Метод случајне шуме	22
3.7 XGBoost	23
3.8 Класична вештачка неуронска мрежа	23
3.9 BERT трансформер модел	24
3.10 <i>Transfer learning</i>	25
4. МЕТОДОЛОГИЈА	27
4.1 Модул за претпроцесирање скупа података	28
4.2 Модул за генерисање скупа атрибута	29
4.3 Модул за класификацију употребом статистичких модела	29
4.4 Модул за класификацију употребом BERT трансформер модела	30
5. ЕКСПЕРИМЕНТИ	33
5.1 Скуп података	33
5.2 Одабир хипер-параметара статистичких модела машинског учења	35
5.3 Одабир хипер-параметара BERT трансформер модела	37
5.4 Евалуација	37
6. РЕЗУЛТАТИ И ДИСКУСИЈА	39
6.1 Резултати статистичких модела	39
5.2 Резултати BERT трансформер модела	41
8. ЗАКЉУЧАК	43
9. ЛИТЕРАТУРА	45
10. БИОГРАФИЈА	49

1. УВОД

Услед нагле дигитализације друштва наша способност да процесирамо и разумемо велике количине података постаје неизмерно важна. Већина података које користимо је неструктурирана. Процењује се да је 90% података на интернету неструктурирано, а већину тих података чини текст [1]. Напретком обраде природног језика (енгл. NLP – *Natural Language Processing*), процесирање текстуалног садржаја је веома поједностављено. Један од главних дијелова обраде природног језика је класификација текста, која има за циљ сврставање текста у једну или више предефинисаних категорија.

Популаризацијом социјалних мрежа повећала се комуникација између људи путем интернета, што је довело до изразитог повећања сајбер насиља на друштвеним мрежама и другим средствима комуникације [2]. Овај раст дискриминације у текстуалним подацима довео је до потребе да се ограничи овај вид насиља. Филтрирање текстуалног садржаја бисмо могли аутоматизовати употребом модела за класификацију текста. Модел би био уграђени у апликацију где би се аутоматски брисао непожељан садржај, а самим тим би се сузбијало ширење сајбер насиља.

У овом раду, вршена је класификација твитова (енгл. *Tweets*), преузетих са друштвене мреже Твитер (енгл. *Twitter*), на типове сајбер насиља које испољавају својим садржајем [3]. Издвојено је шест најчешћих класа сајбер насиља које су познате моделу, а те класе су *Age*, *Ethnicity*, *Gender*, *Religion*, *Other Types* и *Not Cyberbullying*. Систем као улазни податак прима текстуални садржај твита, а као излаз враћа препознати тип насиља израженог у улазним подацима.

Приликом рјешавања овог проблема, имплементирано је више линеарних класификатора заједно са различитим методама векторизације. Такође, вршено је и до-тренирање (енгл. *fine-tuning*) трансформера који представља тренутни *state-of-the-art* приступ за овај проблемски домен. Од технологија коришћене су *scikit-learn*¹ и *PyTorch*² библиотеке за имплементацију класификационих модела.

Евалуација модела вршена је евалуацијом квалитета рада класификационог модела. Квалитет векторизације није евалуиран посебно. Разлог за то је тај што векторизација улазних података представља екстракцију обележја, сходно томе се векторизација

¹ <https://scikit-learn.org/stable/modules/classes.html>

² <https://pytorch.org/>

евалуира у склопу читавог модела [4]. Како је улазни скуп података балансиран, као мјере перформансе користе се тачност, прецизност, одзив и макро Ф-мера.

У евалуацији спроведеној у раду се показало да до-тренирани BERT трансформер даје најбоље резултате приликом класификације (86% макро Ф-мера). Такође, сви линеарни класификатори дају добре резултате (1% маљу макро Ф-меру од BERT трансформер модела) и показали су се као добро решење поготово у ситуацијама када су нам рачунарски ресурси ограничени. Квалитет рада линеарних класификатора доста зависи од претпроцесирања података, као и методе за векторизацију обележја. Резултати у литератури³ су такође показали да је BERT најбољи модел. Перформансе BERT модела у литератури су мало ниже (85% макро Ф-мера), али важно је нагласити да је до-трениран само последњи слој модела, док је у нашем раду коришћена *BitFit* [12] метода до-тренирања.

У поглављу 2 биће представљен преглед стања у области заједно са најбитнијим радовима у овом проблемском домену, као и детаљним описом радова на који се овај рад ослања. У поглављу 3 биће описани теоријски појмови и дефиниције потребне за разумевање овог рада. Поглавље 4 посвећено је опису методологије као и опису тока експеримента. Даље, у поглављу 5, приказане су поставке експеримената који су извршени док се у поглављу 6 представљају и дискутују резултати ових експеримената. На самом крају, у поглављу 7, даје се закључак о овом раду.

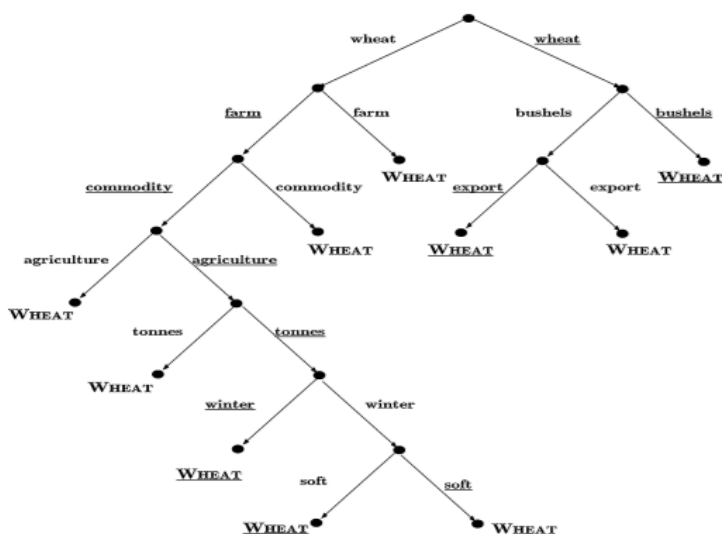
³ <https://www.kaggle.com/code/prasadjawale/bert-cyberbullying-classifer>

2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ

У овом поглављу је наведен историјат развоја приступа за класификацију текста као и тренутни *state-of-the-art* приступ код рјешавања овог доменског проблема. Приликом избора релевантних радова вођено је рачуна о важности рада у овој области, као и да су представљене методологије имале довољно добре резултате приликом класификације.

Прве технике за класификовање текста заснивале су се на ручно креираним правилима [5]. У једној од најпознатијих публикација на ову тему, *Hayes* и сарадници [6] су развијали систем за класификацију новинских чланака према њиховом садржају креиран за *Reuters* новинску агенцију. Систем је комбиновао правила која су вршила и синтатичку и семантичку анализу. На слици 2.1 приказан је примјер групе правила коришћених у овом раду. Аутори су постигли боље резултате чак и од класификатора развијених помоћу *state-of-the-art* методологија машинског учења тог времена. Упркос томе, резултати постигнути у овом раду се не могу узети као генералне перформансе оваквих система, јер ни један други класификатор није био тестиран на овом скупу податка [5]. Главна мана овог приступа је *knowledge acquisition bottleneck*. Правила се ручно креирају уз помоћ доменског експерта, што доводи до тога да су правила превише прилагођена корпусима на основу којих су настала и веома тешко се адаптирају на нове корпусе. Из тих разлога перформансе оваквих система варирају у зависности од домена. Крајем 20. вијека ова методологија је замењена методама машинског учења, које су донијеле бољу генерализацију и скалабилност.

У раду [9] аутори предлажу употребу метода наивног Бајеса (енгл. *Naïve Bayes*, NB) приликом класификације текста. Због своје једноставности и ефикасности NB је био један од најпопуларнијих приступа тог времена. *McCallum* и *Nigam* су вршили поређење између MNB (енгл. *Multinomial Naïve Bayes*) и BNB (енгл. *Bernoulli Naïve Bayes*) на *Newsgroups* скупу података сачињеном од 20 хиљада докумената балансираних између 20 класа. MNB је постигао просечну тачност од 89.2%, док је BNB постигао 84.5%. На основу постигнутих резултата закључено је да је број појављивања ријечи у документу бољи приступ овом проблему. Упркос добрим резултатима, NB методи се заснивају на претпоставци о независним обележјима. Ова претпоставка у пракси често није тачна, што често може лимитирати перформансе NB класификатора.



Слика 2.1. Примјер стабла одлучивања, гране представљају термине у правилима, а листови категорије

Joachims [7] је био један од првих који је применио метод потпорних вектора (енгл. *Support Vector Machines*, SVM) за класификацију текста. Приликом екстракције обележја користио је *Bag-of-Words* [9] као и TF-IDF (енгл. *Term Frequency-Inverse Document Frequency*) [8]. У датом раду класификација је вршена на *Reuters* скупу података сачињеном од новинских чланака који се категоришу на теме према њиховом садржају. Приликом евалуације овог модела коришћена је *PRBEP* метрика (енгл. *Precision/Recall Breakeven Point*). Модел је постигао *PRBEP* од 86.4% и тако остварио значајно боље резултате од других традиционалних метода машинског учења као што су k-NN (енгл. *K-Nearest Neighbor*) и NB (енгл. *Naïve Bayes*). Овим је SVM модел показао своју робусност и способност да ефикасно решава проблеме високе димензионалности. Главна мана овог приступа је временска захтевност као и велика потрошња меморије приликом чувања потпорних вектора. Такође, ефикасност класификатора је доста зависила од екстракције обележја. Већина метода за векторизацију се заснивала на фреквенцији ријечи у документу, чиме није праћена семантичка сличност ријечи.

Mikolov и сарадници су 2013. представили *Word2Vec* алгоритам, који представља временски ефикасан метод за креирање векторске

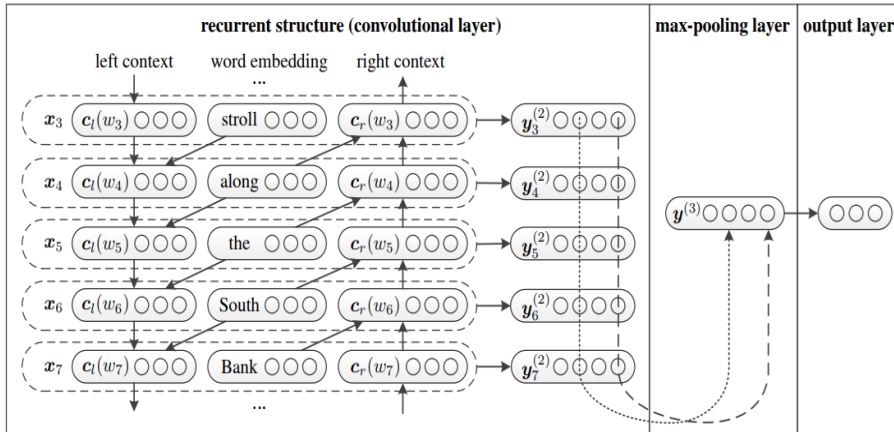
репрезентације речи [10]. *Word2Vec* користи неуронску мрежу за учење векторских репрезентација ријечи, гдје су семантички сличне ријечи представљене сличним векторима. Недуго затим, *Pennington* и сарадници представљају *GloVe* као нови *Word Embedding* који поред локалног контекста узима у обзир и контекст на нивоу документа, чиме боље репрезентује односе између ријечи [11]. Развојем *Word Embedding* метода ријешена су ограничења традиционалних метода за векторизацију која се огледају у немогућности креирања семантичке везе између ријечи. Иако овакав начин векторизације занемарује контекст речи у реченици, довео је до употребе неуронских мрежа у многим NLP (енгл. *Natural Language Processing*) задацима.

Наредних година је расла употреба неуронских мрежа за решавање проблема класификације текста. *Yoon Kim* је био један од првих који је приступио проблему употребом конволутивне неуронске мреже (енгл. *Convolutional Neural Network*, CNN) [13]. Предложени систем је остварио *state-of-the-art* резултате за већину скупова податка коришћених у експерименту. CNN модели су се показали као одлични у учењу локалних обележја односно фраза и зависности између речи, али главни недостатак је био немогућност разумевања контекста који се протеже кроз цијели текст.

Siwei Lai и сарадници су 2015. употребили рекурентну конволутивну неуронску мрежу (енгл. *Recurrent Convolutional Neural Networks*, RCNN) [14]. Употребом RCNN архитектуре комбинује се способност CNN мреже да учи локална обележја и способност рекурентних неуронских мрежа (енгл. *Recurrent Neural Network*, RNN) да разумеју контекст целог текста. На слици 2.2 приказана је предложена архитектура модела. Евалуација је вршена на *20Newsgroups* скупу података уз помоћ макро Ф-мере. Постигавши резултат од 96.49%, модел је надмашио CNN мрежу за 2% на истом скупу податка.

Представљањем трансформер модела дошло је до новог искорака у свим NLP проблемима па тако и у класификацији текста. Механизам вишеструке пажње (енгл. *Attention Mechanism*), представљен 2017. од стране *Ashish Vaswani* и сарадника [15], био је кључан у развоју LLM (енгл. *Large Language Model*) модела. Затим је 2019. представљен BERT модел од стране *Devlin* и сарадника [16]. Претходни трансформер модели су процесирали текст искључиво у једну страну, док је BERT модел бидирекциони трансформер који разуме контекст у оба правца текста. Овај модел је поставио *state-of-the-art* резултате у

многим *benchmark* тестovima. Бидирекциони приступ у LLM моделима покренуо је даљи развој и низ нових решења у овој области.



Слика 2.2. Структура RCNN мреже. Представљен је парцијални примјер реченице „A sunset stroll along the South Bank affords an array of stunning vantage points”. Индекс представља позицију одговарајуће ријечи у оригиналној реченици.

У решењу [17] употребљени су статистички модели као и BERT трансформер модел за класификацију сајбер насиља у тексту, а затим су се поредили резултати различитих приступа. Као представници статистичких модела изабрани су SVM и метод насумичне шуме (енгл. *Random Forest*, RF) заједно са TF-IDF начином векторизације. Евалуација је вршена уз помоћ два скупа података. Први скуп података је *Formspring*⁴ бинарни скуп података, лоше балансиран између насилног и ненасилног текста. Други скуп података је креиран комбинацијом *Formspring* и *Twitter*⁵ скупова података и балансиран је између двије класе. На балансираном скупу података BERT модел је постигао макро Ф-меру од 86%, док су SVM и RF модели постигли 85% и 83% респективно. На небалансираном скупу података BERT је остварио знатно боље резултате постигавши микро Ф-меру од 81%, док је SVM остварио 66%, а RF 46%. Резултати показују да су трансформер модели супериорни у односу на

⁴ <https://github.com/sweta20/Detecting-Cyberbullying-Across-SMPs>

⁵ <https://drive.google.com/drive/folders/1oB2fan6GVGG83Eog66Ad4wK2ZoOjwu3F>

традиционалне методе класификације због своје способности да разумеју ужи и шири контекст текста. Традиционални методи су се показали корисни у ситуацијама када је скуп података балансиран, али њихове перформансе знатно зависе од природе проблема.

3. ТЕОРИЈСКИ ПОЈМОВИ И ДЕФИНИЦИЈЕ

Методе машинског учења коришћене приликом класификације сајбер насиља можемо подијелити на метод заснован на неуронским мрежама и методе статистичких модела. У овом поглављу биће представљене теоријске основе неопходне за разумевање ових поступака.

Приликом употребе статистичких модела тестирана су два начина векторизације:

- TF-IDF (енгл. *Term Frequency-Inverse Document Frequency*)
- GloVe (енгл. *Global Vectors for Vector Representation of Words*)

Поред различитих начина векторизације, као модел за класификацију текста тестирана су четири приступа:

- Метод потпорних вектора (енгл. *Support Vector Machines, SVM*)
- Метод случајних шума (*Random Forest*)
- XGBoost (енгл. *Extreme Gradient Boosting*)
- *Bagging*

Уз наведене статистичке методе класификације, тестирано је и до-тренирање (енгл. *fine-tuning*) BERT трансформер модела.

У складу са тиме, у овом поглављу ће прво бити објашњени TF-IDF (поглавље 3.1) и GloVe (поглавље 3.2) методи за векторизацију. Након тога, у поглављу 3.3 биће објашњен SVM модел. Поглавља 3.4 и 3.5 ће пружити објашњење о стаблу одлучивања (енгл. *Decision Tree*) и *Bagging* ансамбл моделу. Разлог увођења поглавља посвећеном *Decision Tree* моделу је тај што овај модел представља срж *Random Forest* и *XGBoost* модела и увид у њихово функционисање је умногоме олакшан схватањем рада позадинског модела. У поглављу 3.6 и 3.7 ће бити објашњени *Random Forest* и *XGBoost* модели. Затим ће бити објашњене класичне вештачке неуронске мреже (поглавље 3.8) које представљају основу за разумевање BERT трансформер модела, представљеном у поглављу 3.9. На крају, у поглављу 3.10,

биће објашњен *transfer learning* приступ за обучавање неуронских мрежа.

3.1 TF-IDF

TF-IDF (енгл. *Term Frequency-Inverse Document Frequency*) је статистичка мера која се користи за процену важности речи у документу у односу на корпус [18]. Способност овог метода да балансира између фреквенције термина и фреквенције документа чини га основом у анализи текста. TF-IDF је сачињен из два дела:

- TF (енгл. *Term Frequency*)
- IDF (енгл. *Inverse Document Frequency*)

TF мери колико често се одређени термин појављује у документу. Претпоставка је да су термини који се чешће појављују унутар документа важнији за значење самог документа. Постоји више мера којим може да се изрази TF:

- Број понављања речи у документу
- Број понављања речи у документу подијељен дужином документа
- Логаритамски скалиран број понављања речи у документу

Са друге стране, IDF мери колико је важан одређени термин широм целог корпуса. Овај дио TF-IDF метода смањује тежину термина који се појављују у многим документима, јер су мање информативни о било ком јединственом документу. Једначина (1) представља однос укупног броја (N) докумената у корпусу (D) и броја докумената (d) у којима се појављује термин (t):

$$IDF(t, D) = \log\left(\frac{N}{count(d \in D: t \in d)}\right) \quad (1)$$

TF-IDF мера се добија множењем TF и IDF вредности. Ова комбинована мера одражава важност термина унутар одређеног документа и широм корпуса. Главна мана овог метода је статистички приступ који посматра термине независно, чиме се занемарује семантичка структура документа.

3.2 GloVe

GloVe (енгл. *Global Vectors for Vector Representation of Words*) је алгоритам ненадзираног учења развијен од стране истраживача са *Stanford* универзитета за добијање векторске репрезентације речи [11]. Свака реч је представљена вектором, где растојање и правац између вектора енкодирају семантичке односе речи. *Glove* је трениран на великом корпусу текста како би научио векторске репрезентације речи које добро предвиђају вјероватноћу заједничког појављивања речи. За разлику од модела као што је *Word2Vec*, који користи локални контекст за тренинг, *Glove* користи глобалне статистике заједничког појављивања речи. То значи да *Glove* може да хвата шире семантичке односе узимајући у обзир цео корпус.

Велики корпус текста користи се за изградњу матрице X , где $X_{i,j}$ представља број пута када се реч j појављује у контексту речи i . *Glove* користи *Weighted Least Squares* функцију (2) која минимизује разлику између скаларног производа два вектора која представљају речи и логаритма броја њиховог заједничког појављивања:

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}) \quad (2)$$

гдје w_i и b_i представљају вектор и *bias* речи i , док \tilde{w}_j и \tilde{b}_j представљају представљају вектор и *bias* речи j . Функција f (3) представља отежињену функцију, која уз помоћ хипер-параметара x_{max} и α додељује мање тежине ретким заједничким појављивањима:

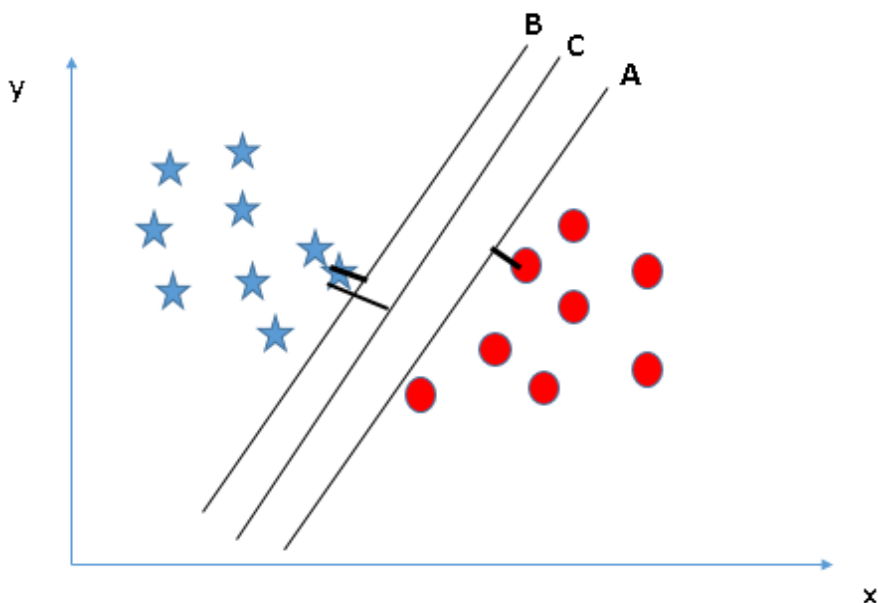
$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^\alpha, & x < x_{max} \\ 1, & x \geq x_{max} \end{cases} \quad (3)$$

Из овога можемо закључити да је *GloVe* веома моћан *Word Embedding* који пружа унапређења у односу на *Word2Vec*. Због своје могућности да разуме шири контекст често је бољи избор уколико немамо ресурсе да до-тренирамо модел на специфичан скуп података.

3.3 Метода потпорних вектора (SVM)

Метода потпорних вектора (енгл. *Support Vector Machine*, SVM) је линеарни модел за рјешавање проблема класификације и регресије.

Може да решава линеарне и нелинеарне проблеме и добро ради на широком спектру проблема (нарочито класификационих), што га чини једним од најмоћнијих не-неуронских модела [19]. Идеја SVM модела је једноставна: алгоритам креира хипер-раван која раздваја податке у класе, такву да маргина одлуке буде максимална. Односно, циљ је да подаци сваке класе буду удаљени што више могуће од границе одлуке.



Слика 3.3.1 Формирање границе одлуке SVM класификатора

Овде се може јавити проблем када подаци нису линеарно сепарабилни у њиховом првобитном облику. Постоје два рјешења за то. Прво је једноставно релаксирање границе одлуке, гдје дозвољавамо малу грешку у нади да ће алгоритам искомвергирати са адекватном границом одлуке. Друго је да се подаци трансформишу у други векторски простор у којем је могуће одредити адекватну хипер-раван која ће раздвојити податке, а затим применом математичких трансформација вратити и податке и хипер-раван у оригинални векторски простор. Из самог поступка се може закључити да овај процес може бити изузетно рачунарски захтеван код података који имају велику димензуоналност. Међутим, код SVM модела је могуће применити такозвани „кERNEL трик“ помоћу којег можемо одредити хипер-раван у другом векторском простору без да морамо све податке експлицитно трансформисати у исти.

Битно је нагласити да је SVM иницијално концепиран као бинарни класификатор, али се проблем класификације у више класа може адресирати преко OVO (енгл. *one-vs-one*) или OVR (енгл. *one-vs-rest*) приступа.

3.4 Стабло Одлучивања

Стабло одлучивања (енгл. *Decision Tree*) је алгоритам надгледаног машинског учења који се користи за класификацију или регресију на основу тога како је одговорено на научен скуп питања [20]. У овом алгоритму се не уводе никакве претпоставке о циљној функцији, што чини овај алгоритам флексибилним и робусним. Стабло одлучивања имитирају људско размишљање, тако да је научницима генерално лако да схвате и протумаче резултате једноставном визуелизацијом чворова и правила у њима. Овај алгоритам не мора увијек да даје егзактну вриједност класификације већ, умјесто тога, може представити опције тако да корисник може сам да донесе смислену одлуку. Градивни блокови овог алгоритма су:

- Коренски чвор (енгл. *root node*) – представља почетну тачку у одлучивању
- Операција раздвајања (енгл. *splitting*) – представља операцију разбијања једног чвора на више под-чворова
- Чвор одлуке (енгл. *decision node*) – чвор који може преусмерити ток извршавања
- Лист стабла (енгл. *leaf node*) – представља могући резултат
- *Pruning* - операција уклањања чворова из стабла
- Грана одлуке (енгл. *decision branch*) – под-стабло стабла одлучивања

Основа стабла одлучивања је коренски чвор. Из коренског чвора тече низ чворова одлуке који описују одлуке које треба донети. Потомци чворова одлуке могу бити други чворови одлуке или листови стабла. Сваки чвор одлуке представља питање или могуће одговоре. Свако под-стабло стабла одлучивања називамо „граном“. Изградња стабла одлучивања се спроводи приликом тренирања модела, у којој се уче атрибути и услови који ће произвести стабло. Затим се стабло орезује (*pruning*) да би се уклониле небитне гране које би могле да

негативно утичу на перформансе. *Pruning* укључује уочавање *outlier*-а, тачака података далеко изван норме, који би могли да доведу до одбацавања прорачуна дајући превелику тежину ретким појавама у подацима. У зависности који проблем рјешавамо, стабла одлучивања се дијеле на категоричке и континуалне, а у овом раду бавићемо се само категорицим стаблима.

3.5 *Bagging* ансамбл модел

Ансамбл модели су технике машинског учења које комбинују више једноставних модела како би се креирао јединствен моћан модел. Ова техника машинског учења је креира да побољша стабилност и тачност статистичких модела, са циљем да смањи варијансу и помогне у избегавању преприлагођавања скупу података. Основни принцип иза овог модела је једноставан: велики број некорелираних слабих предиктора ће имати боље перформансе од појединачних модела. Разлог за овај ефекат је што модели међусобно „штите“ један другог од својих појединачних грешака (све док не греше стално у истом правцу). *Bagging* може користити различите слабе предикторе као свој основни.

Bagging тип ансамбла приликом тренирања, сваком засебном стаблу изнова семплује тренинг скуп који је исте дужине као и оригинални скуп, с тим што је креиран узимањем насумичних инстанци из оригиналног скупа на начин да једна инстанца може бити селектована више пута (комбинације са понављањем). Додатно, за сваки скуп податка се узима насумичан подскуп обележја, што доводи до веће варијансе између појединачних модела и осигурава међусобну некорелираност.

3.6 Метод случајне шуме

Метод случајне шуме (енгл. *Random Forest*), као што име модела имплицира, састоји се од великог броја појединачних стабала одлучивања која функционишу као *bagging* ансамбл [21]. Овај модел може бити употребљен за решавање регресије и класификације. Приликом класификације свако појединачно стабло даје своју предикцију класе, а класа са највише гласова се усваја као коначна предикција нашег модела. Главно унапређење овог модела у односу на друге *bagging* моделе је та што поред семпловања различитих скупова за сваки слаби модел, модел такође врши и *feature bagging*. То значи да се за сваки скуп податка узима насумичан подскуп обележја, што

доводи до веће варијансе између појединачних модела и осигурава међусобну некорелираност.

3.7 XGBoost

XGBoost (енгл. *Extreme Gradient Boosting*) се састоји од великог броја појединачних стабала одлучивања која функционишу као *boosting* ансамбл [24]. *Boosting* ансамбл модели се креирају итеративним обучавањем слабих модела. Сваки слаби класификатор покушава да исправи грешке које је правио претходни класификатор. Приликом обучавања сваког појединачног модела семплује се тренинг скуп, али за разлику од *bagging* приступа, примерима на којима су претходни модели грешили се дају веће тежине тако да ти примери имају већу вероватноћу да буду изабрани.

За разлику од *AdaBoost* метода, у *gradient boosting* методу, коришћеном у *XGBoost*-у, слаби модели немају експлицитно додељене тежине које се узимају у обзир приликом класификације. У овом случају се сваки слаби модел вреднује на основу способности да исправи грешке својих претходника. Због своје особине да се стално прилагођавају погрешно класификованим примерима, *Boosting* ансамбл модели захтевају висок квалитет података. Такође, *Boosting* модели нису отпорни на преприлагођавање, али су знатно робустнији у односу на основне моделе.

XGBoost је конкретна имплементација *gradient boosting* метода са многим унапређењима и оптимизацијама које га чине једним од најбољих не-неуронских приступа. Овај ансамбл модел користи *Lasso* [22] и *Ridge* [23] регуларизацију за спречавање преприлагођавања. Такође, *XGBoost* је имплементиран да одлично решава проблем недостајућих вредности и за разлику од традиционалних имплементација, оптимизован је за коришћење великих скупова података.

3.8 Класична вештачка неуронска мрежа

Вјештачка неуронска мрежа је модел инспирисан неуронском мрежом људског мозга. „Неурон“ у неуронској мрежи је математичка функција која прикупља и класификује информације према специфичној архитектури. Неуронска мрежа садржи слојеве међусобно повезаних чворова гдје је сваки чвор у литератури познат и као перцептрон. Битно је напоменути да постоје три различита типа слојева:

- улазни слој - садржи улазне информације потребне за одлучивање, односно, класификацију
- скривени слој - на основу кога мрежа врши пропратне калкулације и ствара везе са потенцијалним излазима
- излазни слој - садржи резултат калкулације

Сваки неурон првог (улазног) слоја, комбинован са низом одређених коефицијента (тежинама) представља директан улаз у други (скривени) слој. Вредности чворова у овом слоју рачунају се као сума производа претходно поменутих тежина и вредности одговарајућих улазних чворова. Чвор доводи сигнал произведен овом математичком операцијом у функцију активације која је по правилу нелинеарна. Аналогно, претходно добијене вредности у комбинацији са тежинама представљају улаз у следећи скривени слој (може их бити произвољан број са произвољно неурона у сваком од њих) или излазни слој, а вредности индивидуалних чворова рачунају се на исти начин као и у претходном слоју. Из овога се може закључити да неуронска мрежа има велику сличност са статистичким методама као што су уклапање криве и регресиона анализа. Међутим, велики број повезаних чворова праћених нелинеарним трансформацијама на крају даје простора за далеко већу флексибилност него што је то случај код традиционалних модела, на уштрб већег броја параметара и потребе за доста већим скупом податка како бисмо избегли преприлагођавање.

3.9 BERT трансформер модел

BERT је трансформер модел ненадгледано трениран на великом корпусу текста који, након до-тренирања, пружа могућност креирања модела који остварује *state-of-the-art* резултате у многим NLP (енгл. *Natural Language Processing*) задацима [16].

Иако су трансформер модели тип неуронских мрежа, кључна предност трансформер модела је увођење механизма вишеструке пажње (енгл. *Attention Mechanism*) [15]. Овај механизам покушава да симулира људску пажњу усмерену на појединачне ствари, тако што рачуна тежине за сваку реч у реченици. Рачунањем тежина помаже моделу да схвати контекст сваке речи базирано на свим другим речима у реченици. Такође, поред механизма вишеструке пажње, трансформер модели користе позиционално енкодирање, како би модел био свестан позиције сваке речи у реченици. Поменути механизми пружају трансформер моделима могућност да много боље разумеју зависности

између веома удаљених речи у односу са неке друге архитектуре неуронских мрежа.

Главно унапређење BERT модела у односу на претходне трансформер моделе је могућност бидирекционог разумевања текста. Тачније, BERT трансформер модел узима у обзир и претходне и наредне токене приликом предикције речи, док старији трансформер модели посматрају искључиво претходне или наредне токене. Ово омогућава дубље разумевање контекста цијелог текста.

BERT трансформер модел је сачињен искључиво од трансформер енкодера. Модел постоји у двије варијанте, у зависности од броја слојева:

- BERT-Base (12 слојева, 110 000 000 параметара)
- BERT-Large (24 слоја, 340 000 000 параметара)

Такође, оба модела могу да буду у *cased* и *uncased* варијанти у зависности да ли модел подржава велика слова.

Приликом обучавања модела, врши се тренирање на два различита проблема:

- *Masked Language Modeling* (MLM)
- Предвиђање следеће реченице (енгл. *Next Sentence Prediction*, NSP)

Приликом тренирања на MLM проблему, 15% речи у свакој реченици је насумично маскирано и модел предвиђа маскиране термине. Ово се значајно разликује од претходних LLM (енгл. *Large Language Model*) модела, који су предвиђали реч на основу претходних речи. Односно, ово је кључна разлика која је омогућила бидирекционалност BERT модела.

3.10 Transfer learning

Архитектуре са великим бројем слојева могу потрошити доста времена и рачунарских ресурса при тренирању модела. Такође, за многе примене не постоје велики, специјализовани, анотирани скупови података на којима би се могли тренирати дубоки модели.

Код *transfer learning* приступа, идеја је искористити претходно истрениран модел, који ради за сличан проблемски домен и прилагодити га за сопствене потребе [17]. Прилагођавање се врши тако што се отклони оригинални излазни слој и замени се са неким од

излазних слојева који одговарају нашем приступу. Затим је потребно истренирати нови излазни слој над нашим скупом података.

У зависности од величине скупа података, могуће је применити *fine-tuning* метод који представља специфичан метод *transfer learning* приступа. Приликом *fine-tuning* метода, поред излазног слоја, могуће је тренирати све или одређене слојеве из постојећег модела како би се исти што боље прилагодио жељеном домену.

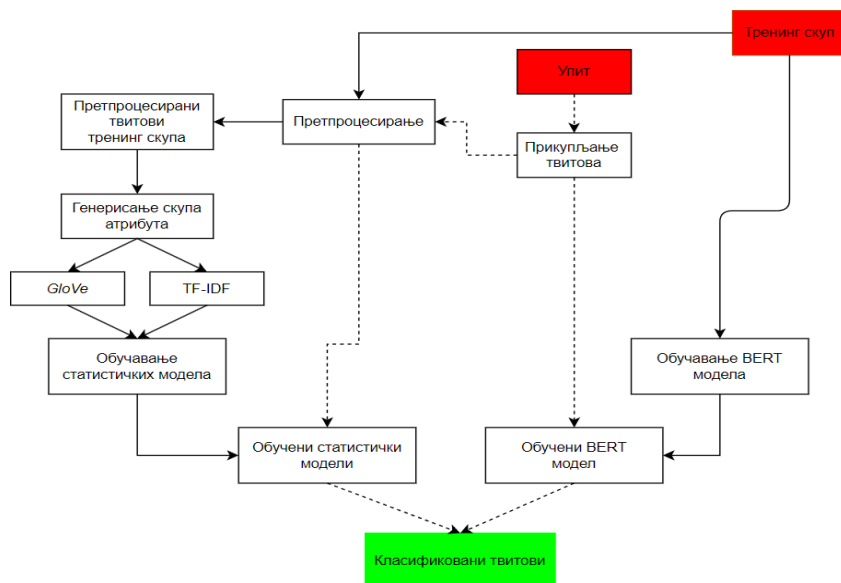
4. МЕТОДОЛОГИЈА

У овом поглављу је представљена имплементација система за класификацију твитова (енгл. *Tweets*) на типове сајбер насиља које изражавају. Улаз у систем представља твит, односно текст објављен на друштвеној мрежи Твитер (енгл. *Twitter*), док излаз из система представља један од типова сајбер насиља испољеног у тексту. Излаз који систем враћа може да буде једна од шест класа сајбер насиља и то *Age, Ethnicity, Gender, Religion, Other Types* и *Not Cyberbullying*.

На слици 4.1 можемо видети ток податка реализованог система. Важно је приметити да имамо две главне гране у нашем систему у зависности од тога да ли користимо статистичке моделе или BERT трансформер модел. Приликом класификације уз помоћ статистичких модела, прво се врши претпроцесирање текста, затим се врши генерисање скупа атрибута и подаци се тек онда прослеђују моделима за класификацију. Са друге стране, приликом класификације BERT трансформер моделом, изостављамо кораке претпроцесирања и екстракције обележја и директно улазни текст прослеђујемо BERT трансформер моделу који врши класификацију. Систем можемо поделити на четири главна модула:

- Модул за претпроцесирање скупа података
- Модул за генерисање скупа атрибута
- Модул за класификацију употребом статистичких модела
- Модул за класификацију употребом BERT трансформер модела

Сви модели су обучени и евалуирани над истим скуповима података. Улазни скуп података је подељен на тренинг, валидациони и тест скуп у односу 80/20/10.



Слика 4.1 Шематски приказ класификације сајбер насиља у твитовима. Пуним линијама је представљен процес тренинга, а испрекиданим линијама процес класификације

4.1 Модул за претпроцесирање скупа података

Улаз у овај модул представљају твитови у изворном облику. У оквиру модула се итерира кроз све примере улазног скупа и примењују се следеће операције претпроцесирања:

- Избацивање свих карактера који не припадају енглеском алфabetу, а да нису размак
- Претварање текста у мала слова
- Избацивање стоп-речи које се налазе у тексту
- Лематизација речи улазног текста

Приликом имплементације овог дијела искоришћено је више библиотека. За избацивање знакова који нису дио алфавета користила се *re*⁶ (енгл. *Regular expression operations*) библиотека. За избацивање

⁶ <https://docs.python.org/3/library/re.html>

стоп-речи, као и за лематизацију употребљена је *nltk*⁷ (енгл. *Natural Language Toolkit*) која представља једну од водећих библиотека за обраду природног језика.

4.2 Модул за генерисање скупа атрибута

Улаз у овај модул представљају претпроцесирани твитови који су излаз из модула за претпроцесирање скупа података. Генерисање скупа атрибута се врши уз помоћ два приступа:

- TF-IDF (енгл. *Term Frequency-Inverse Document Frequency*)
- GloVe (енгл. *Global Vectors for Vector Representation of Words*)

Важно је нагласити да се модул за генерисање атрибута користи искључиво у комбинацији са статистичким моделима за класификацију, док се текст у изворном облику прослеђује BERT трансформер моделу. У нашем експерименту су комбинована оба метода за векторизацију заједно са свим статистичким моделима.

TF-IDF векторизатор је имплементиран употребом *TfidfVectorizer* из *scikit-learn* библиотеке. Одређен је максималан број обележја које векторизатор издваја постављањем „max_features“ параметра на 9300.

Претварање улазних твитова у векторску репрезентацију извршено је помоћу *Glove Embedding* методе. Имплементација је извршена помоћу *numpy*⁸ и *tensorflow*⁹ библиотеке. Употребљени *Embedding* је трениран на корпусу сачињеном од 6 000 000 000 токена, док су токени представљани векторима дужине 300.

4.3 Модул за класификацију употребом статистичких модела

Улаз у овај модул представља резултат једног од метода за генерисање скупа атрибута, а као излаз модул враћа предикцију класе која означава један тип сајбер насиља израженог у улазном тексту. Овај модул је имплементиран на четири начина, тј., као модел за класификацију текста тестирана су четири приступа:

⁷ <https://www.nltk.org/>

⁸ <https://numpy.org/doc/>

⁹ https://www.tensorflow.org/api_docs

- Метод потпорних вектора (енгл. *Support Vector Machines*, SVM)
- Метод случајних шума (*Random Forest*)
- XGBoost (енгл. *Extreme Gradient Boosting*)
- *Bagging*

Сви модели су обучени и евалуирани над истим скупом података. Важно је напоменути и да је сваки приступ за класификацију комбинован и са TF-IDF и са *GloVe* методом за екстракцију обележја. Са оптимизацију параметара свих модела коришћен је *GridSearchCV* из *scikit-learn* библиотеке.

SVM модел је имплементиран коришћењем SVC (*C-Support Vector Classification*) из *scikit-learn* библиотеке. Коришћен је стандардни RBF кернел са свим подразумеваним параметрима.

Random Forest класификатор имплементиран је употребом *RandomForestClassifier* класе из *scikit-learn* библиотеке. Коришћено је 100 естиматора. Извршена је паралелизација тренинга и извршавања модела над свим расположивим језгрима процесора постављањем „*n_jobs*“ параметра на -1.

XGBoost је имплементиран употребом *XGBClassifier* класе из *xgboost* библиотеке. Коришћено је 100 стабала у ансамбл моделу. Изабрано је да ансамбл модел користи 70% тренинг скупа за обучавање сваког стабла постављањем параметра „*subsample*“ на 0.7. Такође, ограничена је употреба броја обележја које се користе за обуку стабала постављањем „*colsample_bytree*“ параметра на 0.6. Овим параметрима смо спречавали преприлагођавање ансамбл модела.

Bagging ансамбл модел је имплементиран употребом *BaggingClassifier* из *scikit-learn* библиотеке. Као основни модел овог ансамбла коришћен је SVM модел са свим подразумеваним параметрима. Коришћено је 10 естиматора због ограничених ресурса који су употребљени за тренирање модела.

4.4 Модул за класификацију употребом BERT трансформер модела

Приликом имплементације BERT трансформер модела коришћен је BERT-*base* модел који поседује 12 слојева са 110 000 000 параметара. Тип овог модела је био *uncased*, што значи је модел занемарује велика слова. Поред постојеће архитектуре додан је

потпуно повезани слој са 512 неурона након којег следи стандардан *softmax* слој са 6 могућих категорија. За имплементацију архитектуре коришћена је *PyTorch*¹⁰ библиотека. Коришћен је *AdamW* оптимизациони алгоритам [26] из *transformers*¹¹ библиотеке, док је *learning rate* износио 0.00001. Функција грешке која је искоришћена је *NLLLoss* (енгл. *Negative Log-Likelihood Loss*) функција из *torch* библиотеке. Као активациона функција коришћена је *ReLU* функција.

Приликом анализе скупа података утврђено је да је дужина већине улазних твитова мања од 40 речи, тако да је максималан број токен постављен на 40. Овим смо знатно смањили потребне меморијске ресурсе и омогућили *batch* процесирање.

BERT трансформер модел је до-трениран употребом *BitFit* [12] методе до-тренирања. Овом методом смо, поред слојева које смо додали на постојећи архитектуру, до-тренирали други слој трансформер модела заједно са свим *bias* параметрима модела. Овим смо вршили до-тренирање само 0.1% свих параметара у трансформер моделу и знатно смањили потребне ресурсе тренирања, али упркос томе задржали добре перформансе модела.

¹⁰ <https://pytorch.org/docs/stable/library.html>

¹¹ <https://pypi.org/project/transformers/>

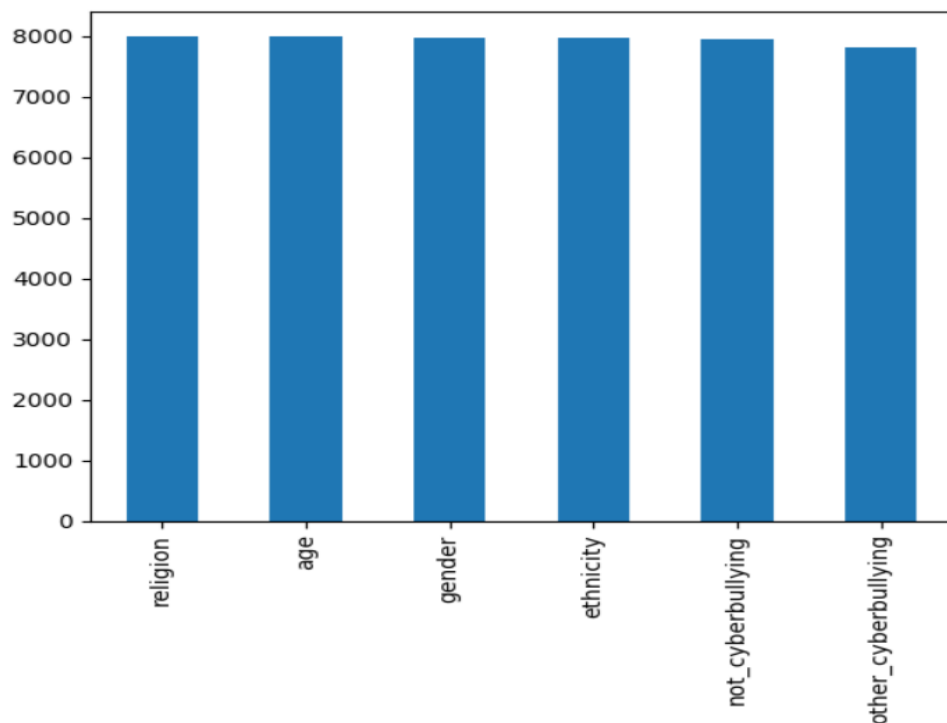
5. ЕКСПЕРИМЕНТИ

Ово поглавље се бави описом (главних) експеримената који су извршени у току развијања система. У потпоглављу 5.1 биће описан скуп података који је коришћен за тренинг, валидацију и евалуацију класификационих модела који чине трећи и четврти модул система. У 5.2, и 5.3 биће описан процес одабира хипер-параметара статистичких модела и BERT трансформер модела. Циљ ових експеримената је да се покажу реалне перформансе система за квалитет класификације сајбер насиља. Експерименти у којима је експериментисано са различитим вредностима хипер-параметара модела нису укључена у овом поглављу, како би фокус био на коначној евалуацији система на тест скупу података. На крају, у потпоглављу 5.7 описан је процес евалуације појединачних класификационих приступа. Важно је напоменути да је тренинг свих модела извршен на *Google Colab* платформи употребом GPU инстанце. Коришћена верзија *Python* програмског језика је 3.10.12.

5.1 Скуп података

За потребе тренирања, евалуације и тестирања класификационих метода коришћен је јавно доступан *Cyberbullying Classification* скуп података [3]. Скуп података је сачињен од твитова (енгл. *Tweets*) у којима се изражава насилан садржај. Сваки твит је лабелиран типом сајбер насиља на који се односи. Скуп података је сачињен од шест класа, и то *Age*, *Ethnicity*, *Gender*, *Religion*, *Other Types* и *Not Cyberbullying*.

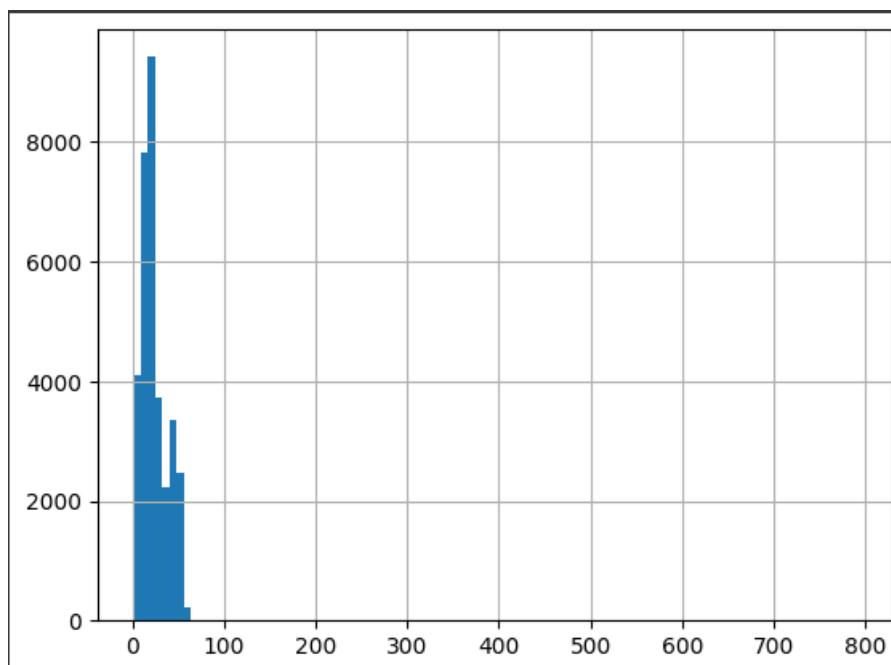
Скуп података садржи 47 652 инстанце. Дистрибуција класног обележја у овом скупу података је балансирана, где најзаступљенија класа има 7998 а најмање заступљена 7798 инстанци. Циљне лабеле су енкодиране за потребе обучавања техником *label encoding*-а. *Frequency plot* класа овог скупа података је приказан на графикону 5.1.1.



Графикон 5.1.1 Дистрибуција класног обележја

Твитови често садрже специјалне карактере, линкове и други садржај који није важан приликом класификације. Из тог разлога је вршено претпроцесирање података описано у потпоглављу 4.1.

Просечан број токена у претпроцесираном твиту је 23. Најкраћа инстанца садржи један токен, док најдужа садржи 79 токена. Графикон 5.1.2 приказује хистограм дужине инстанци у скупу података.



Графикон 5.1.2 Хистограм броја токена у процесираним твитовима

5.2 Одабир хипер-параметара статистичких модела машинског учења

У овом експерименту су за класификацију разматрана четири статистичка модела машинског учења: метод потпорних вектора, метод насумичних шума, XGBoost и *Bagging*. За екстракцију обележја коришћени су TF-IDF и *GloVe* методи за векторизацију. Максималан број обележја за све моделе сем *Bagging* ансамбл модел је постављен да буде 9300. У случају *Bagging* модела је број обележја постављен на 3000 због ограничених рачунарских ресурса. Коришћен је *GridSearchCV* приступ (детаљно описан у поглављу 5.4) за одређивање оптималних хипер-параметара модела на валидационом скупу података. Оптимални хипер-параметри за моделе метод потпорних вектора, метод насумичних шума, XGBoost и *Bagging* се налазе у табелама 5.3.1 до 5.3.4, тим редоследом.

Хипер-параметар	Вредност
<i>C</i>	1.0
<i>gamma</i>	'scale'
<i>shrinking</i>	<i>True</i>
<i>probability</i>	<i>False</i>
<i>tol</i>	1e-3
<i>max_iter</i>	-1
<i>decision_function_shape</i>	'ovr'
<i>kernel</i>	'rbf'

Табела 5.3.1 Оптималне вредности хипер-параметара SVM класификатора одабране на валидационом скупу података

Хипер-параметар	Вредност
<i>criterion</i>	'gini'
<i>min_samples_split</i>	2
<i>min_samples_leaf</i>	1
<i>min_weight_fraction_leaf</i>	0.0
<i>max_features</i>	'sqrt'
<i>max_leaf_nodes</i>	<i>None</i>
<i>bootstrap</i>	<i>False</i>
<i>n_estimators</i>	100
<i>warm_start</i>	<i>True</i>

Табела 5.3.2 Оптималне вредности хипер-параметара *Random Forest* класификатора одабране на валидационом скупу података

Хипер-параметар	Вредност
<i>n_estimators</i>	100
<i>subsample</i>	0.7
<i>colsample</i>	0.6
<i>learning_rate</i>	0.3
<i>max_depth</i>	6

Табела 5.3.3 Оптималне вредности хипер-параметара *XGBoost* класификатора одабране на валидационом скупу података

Хипер-параметар	Вредност
<i>n_estimators</i>	10
<i>estimator</i>	SVM
<i>max_features</i>	1.0
<i>max_samples</i>	1.0

Табела 5.3.4 Оптималне вредности хипер-параметара *Bagging* класификатора одабране на валидационом скупу података

5.3 Одабир хипер-параметара BERT трансформер модела

У овом експерименту тестиран је BERT трансформер модел као класификациони модел. Приликом тренирања модела, величина *mini batch*-а је подешена на 32 и извршено је 10 епоха. Вредности ових параметара добијене су емпиријским путем на валидационом скупу података.

5.4 Евалуација

У овом поглављу је приказана евалуација тренираних класификатора над тест скупом. Као метрике перформанси коришћене су тачност, прецизност, одзив и макро Ф-мера.

Тест скуп је добијен насумичним узорковањем оригиналног скупа података. Подела је извршена на тренинг, валидациони и тест скуп у односу 80/20/10. Након подешавања хипер-параметара коришћењем валидационог скупа, модел је поново трениран на тренинг скупу и извршена је евалуација на тест скупу. Подела је извршена помоћу *train_test_split* функције из *scikit-learn* библиотеке, где је *random state* за одабир псеудо-случајних инстанци био „123“.

Такође, битно је напоменути да је одабир параметара статистичких модела извршен уз помоћ *GridSearchCV* класе која бира најбољу комбинацију параметара уз помоћ унакрсне валидације. У мом случају коришћен је стратификовани *10-fold-cross-validation*.

6. РЕЗУЛТАТИ И ДИСКУСИЈА

У овом поглављу приказани су резултати експеримената описаних у претходном поглављу (5.2 и 5.3).

6.1 Резултати статистичких модела

Сви статистички модели су тестирани са TF-IDF и *GloVe* начином векторизације. Резултати свих тестираних статистичких модела су представљени у табели 6.1.1.

Класа	Тачност	Макро Ф-мера
<i>GloVe</i> + <i>Random Forest</i>	0.76	0.76
<i>GloVe</i> + SVM	0.79	0.79
<i>GloVe</i> + <i>XGBoost</i>	0.78	0.79
<i>GloVe</i> + <i>Bagging</i>	0.81	0.81
TF-IDF + <i>Random Forest</i>	0.83	0.83
TF-IDF + <i>Bagging</i>	0.83	0.83
TF-IDF + SVM	0.84	0.84
TF-IDF + <i>XGBoost</i>	0.84	0.85

Табела 6.1.1 Резултати тестирања статистичких модела

Приликом употребе *GloVe Embedding*-а као метод за векторизацију тачност и макро Ф-мера опадају на свим моделима. Како употреба *GloVe*-а уноси разумевање семантичких веза речи, тешко је утврдити зашто TF-IDF даје боље резултате. *Random Forest* класификатор заједно са *GloVe* методом за векторизацију остварује најлошију макро Ф-меру од 0.76, док *XGBoost* у комбинацији TF-IDF обележјима остварује најбољи резултат од 0.85. Највећа добијена макро Ф-мера представља унапређење од 4% у односу на најбољи резултат статистичких модела из постојећих решења¹².

Резултати остварени од стране *Random Forest* модела су лошији од оних које је остварио SVM модел. Такви резултати нису били очекивани с обзиром на то да за ансамбл моделе важи генерално прихваћена претпоставка да би требали радити боље од осталих традиционалних модела. Иако је *Bagging* ансамбл модел искоришћен са SVM моделом као основним у циљу побољшања резултата SVM

¹² <https://www.kaggle.com/code/thisishusseinali/cyberbullying-text-classification>

модела, неочекивано овај експеримент постиже лошије резултате. Ипак, важно је нагласити да је коришћен мали број естиматора и да је број обележја које векторизатор издваја три пута мањи него код других статистичких модела. Лоши резултати су највише условљени ограничењем рачунарских ресурса који су били на располагању.

Класа	Прецизност	Одзив	Ф-мера
<i>Not Cyberbullying</i>	0.65	0.50	0.56
<i>Other Types</i>	0.60	0.81	0.69
<i>Gender</i>	0.91	0.83	0.87
<i>Religion</i>	0.97	0.95	0.96
<i>Age</i>	0.99	0.97	0.98
<i>Religion</i>	0.99	0.99	0.99

Табела 6.1.2 Резултати тестирања за сваку класу (TF-IDF у комбинацији са XGBoost моделом)

Табела 6.1.2 представља резултате за сваку појединачну класу класификације употребом XGBoost модела и TF-IDF метода за векторизацију јер та комбинација постиже највећу Ф-меру од свих статистичких метода. Можемо приметити да овај модел најлошије класификују класе *Not Cyberbullying* и *Other Types*, док модел постиже драстично боље резултате приликом класификације других класа. Узрок лоше класификације *Not Cyberbullying* класе представља доста инстанци погрешно лабелираних као *Other Types*, а које треба да припадају *Not Cyberbullying* класи. Постоји још простора за унапређење модела уколико бисмо повећали број естиматора и додатно подесили хипер параметре.

6.2 Резултати BERT трансформер модела

Поређење перформанси најбољег статистичког модела и BERT трансформер модела су представљени у табели 6.2.1.

Класа	Тачност	Макро Ф-мера
TF-IDF + XGBoost	0.84	0.85
BERT	0.86	0.86

Табела 6.2.1 Поређење перформанси најбољег статистичког модела и BERT трансформер модела

BERT трансформер модел постиже тачност и макро Ф-меру од 0.86 и тако остварује најбоље резултате у овом раду. Модел потврђује супериорност овог приступа над статистичким моделима, иако је макро Ф-мера већа за само 1% у односу на *XGBoost*. Можемо закључити да статистички модели у неким случајевима могу пружити одличне резултате, нарочито узимајући у обзир мању рачунарску захтевност у односу на трансформер моделе.

BERT трансформер остварује бољи резултат него у постојећим решењима³ за 1%. Такође, овај модел постиже боље резултате и од напреднијег ROBERTA модела, који у постојећим решењима¹³ постиже макро Ф-меру од 0.85.

Класа	Прецизност	Одзив	Ф-мера
<i>Not Cyberbullying</i>	0.67	0.61	0.64
<i>Other Types</i>	0.70	0.74	0.72
<i>Gender</i>	0.89	0.89	0.89
<i>Religion</i>	0.92	0.96	0.94
<i>Age</i>	0.98	0.97	0.97
<i>Religion</i>	0.98	0.98	0.98

Табела 6.2.2 Резултати тестирања на свакој појединачној класи (BERT)

Резултати за сваку појединачну класу предикције BERT трансформер модела су представљени у табели 6.2.2. Као и остали модели, BERT трансформер модел најлошије класификује *Not Cyberbullying* и *Other Types* класе. Поред грешака у скупу података примећени су додатни потенцијални узроци погрешних класификација. Модел често погрешно класификује улаз као *Gender* уколико улазни текст садржи речи у женском роду. Такође, примећено је да модел има проблем са текстом који садржи кључне речи које представљају насиље („*Bullying*“, „*Racism*“ и др.). Овакве примере модел класификује као неки тип насиља, иако текст представља *Not Cyberbullying* класу. Пад перформанси *Religion* класе је последица кључних речи које се односе на веру, самим тим модел тежи да улаз

¹³ <https://www.kaggle.com/code/khaledabdelgaber/my-roberta-for-bullying-classification>

класификује као *Religion* уколико улазни текст садржи неку од тих речи.

Перформансе модела се значајно могу унапредити побољшањем анотација у тренинг скупу. Такође, додатно подешавање хипер-параметара је могуће, као и до-тренирање свих параметара BERT трансформер модела, јер је у овом раду коришћена *BitFit* метода до-тренирања која селективно бира параметре модела чије тежине се тренирају.

7. ЗАКЉУЧАК

У овом раду представљен је систем за класификацију твитова (енгл. *Tweets*) на тип сајбер насиља који испољавају. Такав систем би могао представљати основу за аутоматско филтрирање садржаја на интернету, а самим тим и за сузбијање сајбер насиља. Систем се састоји из два модула:

- Модул за класификацију употребом статистичких модела
- Модул за класификацију употребом BERT трансформер модела

Први модул користи TF-IDF и *GloVe Embedding* као методе за векторизацију текста. У овом случају су за класификацију разматрана четири статистичка модела машинског учења: метод потпорних вектора, метод насумичних шума, XGBoost и *Bagging*.

У другом модулу коришћен је BERT трансформер модел који представља *state-of-the-art* технологију за овај проблемски домен [16]. Сви модели су обучавани и евалуирани над *Cyberbullying Classification* скупом података [3].

BERT модел је постигао најбољу Ф-меру од 0.86, XGBoost је био на другом месту са 0.85, док је SVM остварио Ф-меру од 0.84. Најлошије резултате пружили су модел насумичне шуме и *Bagging* модел са Ф-мером од 0.83. Важно је нагласити да су овде приказани резултати статистичких модела остварени уз помоћ TF-IDF векторизације јер се то показало као боље у односу на *GloVe Embedding* векторизацију.

Треба се осврнути и на то да систем најлошије класификује класе *Not Cyberbullying* и *Other Types*. Главни узрок представља доста погрешно лабелираних *Not Cyberbullying* инстанци класом *Other Types*. Модел често погрешно класификује улаз као *Gender* уколико улазни текст садржи речи у женском роду. Такође, примећено је да модел има проблем са текстом који садржи кључне речи које представљају насиље („*Bullying*“, „*Racism*“ и др.). Овакве примере модел класификује као неки тип насиља, иако текст представља *Not Cyberbullying* класу. Проблем праве и кључне речи које се односе на веру, јер погрешно наводе моделе да улаз класификују као *Religion* уколико улазни текст садржи неку од тих речи.

Перформансе система би се могле унапредити исправљањем грешака у анотацијама скупа података. Такође, могућа су унапређења у избору хипер-параметара. Као следећи корак у развоју овог система

би се могло размотрити тестирање још напреднијих трансформер модела и великих језичких модела.

8. LITERATURA

- [1] Rydning, J. (2023). *Worldwide Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast, 2023–2027* (Document Number: US50397723). International Data Corporation (IDC).
- [2] Peterson, J. K., & Densley, J. (2016). Is social media a gang? Toward a selection facilitation or enhancement explanation of cyber violence. *Aggression and Violent Behavior*.
- [3] Cyberbullying Classification скуп података (датум приступа 29.05.2024.)
<https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification/data>
- [4] Wang, Y., Dong, L., Jiang, X., Ma, X., Li, Y., & Zhang, H. (2021). KG2Vec: A node2vec-based vectorization model for knowledge graph. *PLOS ONE*, 16(3), e0248552.
- [5] Sebastiani, F. (2002). Automatic text categorization: A review. *ACM Computing Surveys*, 34(1), 1-47.
- [6] Hayes, P. J., Andersen, P. M., Nirenburg, I., & Schmandt, L. M. (1990). Intelligent text retrieval: Using an inference network. *Information Processing & Management*, 26(3), 343-368.
- [7] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Proceedings of the 10th European Conference on Machine Learning (ECML-98)* (pp. 137-142). Berlin, Heidelberg: Springer.
- [8] Salton, G., Fox, E. A., & Wu, H. (1983). *The Smart Retrieval System—Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.
- [9] McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization* (pp. 41-48).
- [10] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

- [11] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- [12] Zaken, E. B., Ravfogel, S., & Goldberg, Y. (2021). BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.
- [13] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- [14] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 2267-2273).
- [15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)* (pp. 5998-6008).
- [16] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).
- [17] Ogunleye, B., & Dharmaraj, B. (2023). The use of a large language model for cyberbullying detection. *Analytics*, 2(3), 694-707.
- [18] Aizawa, A. (2002). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1), 45-65.
- [19] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13(4), 18-28.
- [20] Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, 15(3), 142-147.
- [21] Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197-227.

- [22] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288.
- [23] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- [24] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.
- [25] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.
- [26] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

9. БИОГРАФИЈА

Срђан Стјепановић је рођен 09.08.2001. у Дервенти, где је стекао основно и средње образовање. Школске 2020/21 године се уписује на Факултет Техничких Наука на студијски програм Софтверско Инжењерство и Информационе Технологије. Положио је све испите предвиђене планом и програмом и стекао услов за одбрану завршног рада.