Augmentacija podataka za klasifikaciju kratkih tekstova

I. Hrga
Sveučilište u Rijeci/Fakultet informatike i digitalnih tehnologija, Rijeka, Hrvatska
ingrid.hrga@gmail.com

***Sažetak* - Augmentacija podataka postala je neizostavni korak u procesu učenja većine sustava temeljenih na dubokim neuronskim mrežama. Velike količine podataka potrebne da bi sustav uspješno naučio rješavati zadatke često je teško pribaviti u dovoljnoj količini i kvaliteti. Zbog toga se pribjegava različitim postupcima kojima se primjenom raznovrsnih transformacija može višestruko povećati osnovni skup podataka. U području računalnog vida već su uvriježene brojne tehnike koje relativno lako transformiraju sliku bez da se promijeni što ta slika predstavlja. Međutim, u području obrade prirodnog jezika takve su transformacije znatno manje zastupljene. Već i promjena samo jednog slova može znatno promijeniti smisao rečenice, stoga je potrebno puno više opreza prilikom odabira pogodnih transformacija. U ovom radu uspoređuju se postupci za augmentaciju tekstualnih podataka. Uz prikaz njihovih prednosti i nedostataka, analizira se kako se promjena pojedine tehnike odražava na rezultate klasifikacije kratkih tekstova.**

Ključne riječi - augmentacija podataka; klasifikacija teksta; konvolucijske neuronske mreže; WordNet sinonimi; kontekstualni vektori riječi

I. UVOD

Otkako su duboke neuronske mreže pokazale sposobnost rješavanja najraznovrsnijih zadataka iz područja računalnog vida ili obrade prirodnog jezika [1, 2], sve se više autora bavi problemom učenja iz ograničenih skupova podataka [3]. Nedovoljna količina podataka za učenje u odgovarajućoj kvaliteti dovodi do problema prenaučenosti (eng. *overfitting*), što se negativno odražava na sposobnost generalizacije modela.

Postoje različiti načini za ublažavanje prenaučenosti [4], stoga se neki autori usredotočuju na razvoj naprednijih tehnika za prijenos znanja [5], drugi se usmjeravaju na unaprjeđenje metoda za regularizaciju modela [6]. Međutim, veće količine podataka za učenje u pravilu predstavljaju najpoželjnije rješenje problema prenaučenosti, u čemu važnu ulogu ima augmentacija podataka [7].

Augmentacija podataka obuhvaća tehnike za povećanje veličine i raznovrsnosti osnovnog skupa bez potrebe za prikupljanjem novih podataka [7]. Takve tehnike primjenom različitih transformacija stvaraju varijacije postojećih podataka. Osim toga, u svrhu proširenja osnovnog skupa mogu se koristiti i sintetički primjeri dobiveni generativnim modelima ili primjenom ostalih naprednih tehnika koje stvaraju sasvim nove podatke [8]. Međutim, bez obzira na to koji se način za proširenje osnovnog skupa koristi, važno je sačuvati vezu između podataka i dodijeljenih im oznaka, poput klasa. U protivnome, u podatke se uvodi nepotreban šum, što u konačnici, unatoč većem skupu za učenje, može rezultirati lošijim performansama modela.

Pored navedenoga, augmentacija podataka može pomoći i u situacijama kada postoji znatna neravnoteža među pojedinim klasama. Manjinska klasa tad se augmentira transformiranim ili sintetičkim primjerima, čime se pridonosi uravnoteženju skupa podataka [9]. Augmentacija smanjuje potrebu za prikupljanjem i označavanjem podataka, što je često dugotrajan i zamoran proces, a ponekad nije niti moguć, npr. ako zahtijeva specifično ekspertno znanje ili se provodi nad osjetljivim podacima. Generiranjem sintetičkih primjera iz distribucije jednake ili slične onoj osnovnog skupa, moguće je nadopuniti ili zamijeniti originalne podatke.

Neupitna je korisnost augmentacije te je ona postala neizostavni korak procesa učenja u području računalnog vida. Pomoću jednostavnih transformacija slika poput rotacije, translacije, promjena boje [10] te kompleksnijih poput prekrivanja dijelova slike [11] ili interpolacijom više različitih slika [12], postižu se bolje performanse modela uz veću robusnost na varijacije u podacima [7].

Međutim, u radu s tekstualnim podacima takvi su postupci znatno manje zastupljeni. Za razliku od slika, gdje većina promjena vrijednosti pojedinih piksela ili objekata neće utjecati na to što ta slika predstavlja (npr. mačka sa zelenim krznom i dalje će biti prepoznata kao mačka), u radu s tekstom već i promjena jednog znaka može promijeniti smisao čitave rečenice. Diskretna priroda jezika onemogućava stvaranje postepenih varijacija. Osim toga, poredak riječi u rečenici utječe na smisao rečenice, a smisao također ovisi i o kontekstu. Promjenom konteksta može se znatno promijeniti značenje riječi. Kompleksna struktura teksta i veza među riječima čine augmentaciju znatno izazovnijom te je potrebno puno više opreza prilikom odabira prikladnih transformacija.

U ovom radu uspoređuju se odabrane tehnike za augmentaciju tekstualnih podataka. Dan je pregled postupaka temeljenih na promjenama pojedinih znakova, zamjenama riječi baziranih na WordNet [13] sinonimima te tehnikama koje manipuliraju reprezentacijom teksta u vektorskom prostoru. Uz prikaz njihovih prednosti i nedostataka, empirijski je provjereno kako se primjena

pojedine tehnike odražava na rezultate klasifikacije kratkih tekstova.

Rad je strukturiran na sljedeći način: nakon uvodnog dijela, dan je kratki pregled najčešće korištenih tehnika za augmentaciju tekstualnih podataka prikladnih za problem klasifikacije, uz dodatni prikaz mreža koje se u tu svrhu koriste. U trećem poglavlju opisani su podaci, odabrane transformacije i ostale postavke eksperimenta. U četvrtom poglavlju prikazani su rezultati klasifikacije modela treniranih na originalnim podacima, onim augmentiranim te njihovim kombinacijama. Na kraju je dan zaključak, uz nekoliko smjernica za buduća istraživanja.

II. AUGMENTACIJA PODATAKA I KLASIFIKACIJA

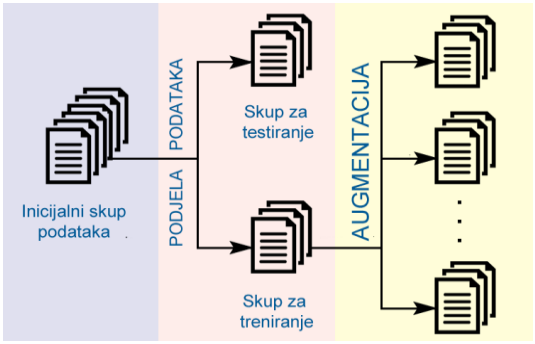
A. *Augmentacija podataka*

Većina transformacija koje se u svrhu augmentacije koriste u području obrade prirodnog jezika inspirirana je augmentacijama uvriježenima u području računalnog vida. Na primjer, tehnikama koje mijenjaju pojedine znakove odgovaraju promjene vrijednosti pojedinih piksela, dok se promjene na razini riječi mogu poistovjetiti s promjenama pojedinih objekata na slici. Međutim, za razliku od računalnog vida, gdje je augmentacija podataka tijesno povezana s procesom učenja, tekstualne se transformacije uglavnom pripremaju unaprijed, prije samoga učenja (Slika 1).

Tehnike augmentacije namijenjene radu s tekstualnim podacima mogu se kategorizirati na više načina [3, 9, 14]: 1) s obzirom na granularnost jedinica transformacije mogu se podijeliti na one koje se primjenjuju na pojedinim znakovima, riječima, frazama ili rečenicama, 2) na temelju reprezentacije teksta dijele se na one koje izravno mijenjaju sirove podatke te one koje manipuliraju njihovom vektorskom reprezentacijom, 3) augmentacije mogu biti bazirane na pravilima, na unaprijed treniranom modelu ili dobivene učenjem novog modela, 4) mogu biti ovisne o jeziku ili neovisne o jeziku, 5) prilagođene određenom zadatku ili univerzalne, 6) determinističke ili stohastičke.

Najjednostavnije su tehnike koje mijenjaju pojedine znakove. Takve promjene mogu biti neovisne o jeziku, poput brisanja, zamjene ili ubacivanja slučajno odabranog znaka, ili mogu ovisiti o jeziku teksta, poput simulacije pogrešaka prilikom tipkanja na temelju rječnika pogrešaka. Ovakve promjene, zbog uvođenja šuma stvaranjem nepostojećih riječi, rezultiraju velikim vokabularom, a iako su jednostavne za implementaciju, u radovima nisu značajnije zastupljene.

S druge strane, autori glavnine radova predlažu transformacije na razini pojedinih riječi ili pojavnica (eng. *token*) među kojima su najzastupljenije zamjene slučajno odabranih riječi njihovim sinonimima tj. riječima jednakog ili bliskog značenja. Moguće su i zamjene antonimima tj. riječima suprotnog značenja, međutim, takve promjene mijenjaju smisao rečenice, stoga nisu prikladne za problem klasifikacije.



Slika 1. Shema općenitog postupka augmentacije podataka. Nakon podjele podataka, transformiraju se samo podaci za treniranje dok se podaci za testiranje koriste u izvornom obliku.

U literaturi su zastupljeni različiti pristupi pri odabiru sinonima. U [15] autori koriste WordNet, leksičku bazu međusobno povezanih sinonima za engleski jezik. Takve su zamjene brze i jednostavne, međutim, baze sinonima ne postoje za sve jezike ili ne pokrivaju dovoljan broj riječi, o čemu ovisi i kvaliteta augmentacije. Npr. WordNet obuhvaća 117000 skupova sinonima (eng. *synsets*), dok hrvatska inačica, Crown [16] u verziji 2.0 obuhvaća samo 23122 sinonimska skupa. Zato drugi autori [17] predlažu odabir sinonima na temelju kosinusne sličnosti vektorskih reprezentacija riječi. U tu se svrhu najčešće koriste word2vec [18], GloVe [19] ili fasttext [20] vektori. Nedostatak ovih pristupa je to što ne uzimaju u obzir kontekst te će riječ u slučaju polisemije, tj. u slučaju kada ona ima više značenja, biti jednako reprezentirana (vektorom istih vrijednosti) za svako od značenja. Npr. riječi glava i lubanja dovoljno su bliske, ali zamjena "glavica luka" → "lubanja luka" ne bi bila prihvatljiva. Zbog toga se prednost često daje kontekstualnim vektorima koje stvaraju jezični modeli unaprijed trenirani na iznimno velikim korpusima, s ciljem predviđanja prikrivene riječi u rečenici. Njihova je prednost što pri tome kao kontekst uzimaju čitavu rečenicu. To znači da će pojedina riječ imati poseban vektor za svako značenje, a za zamjenu bit će odabrana ona riječ koja odgovara smislu čitave rečenice. Primjeri takvih jezičnih modela su BERT [2] ili DistillBERT [21] kao njegova umanjena verzija.

Jedna od ranijih primjena augmentacije može se pronaći u [22]. Autori u [23] koriste augmentaciju sinonimima za treniranje konvolucijskih neuronskih mreža, dok su autori u [15], pod nazivom *Easy Data Augmentation (EDA)*, obuhvatili kombinaciju slučajnih zamjena sinonimima s ubacivanjem sinonima, te brisanjem slučajno odabranih riječi ili zamjenom njihovog poretka. Ovakve jednostavne transformacije pokazale su se učinkovitima u treniranju jednostavnijih modela na malim skupovima podataka.

Druga skupina popularnih tehnika stvara varijacije rečenica parafraziranjem, pri čemu se u tu svrhu najčešće koristi prevođenje [24]. Augmentirana rečenica dobije se prevođenjem izvorne rečenice na pivotalni jezik, a zatim se takav prijevod ponovno prevodi na izvorni jezik. Moguće je koristiti jedan pivotalni jezik ili više različitih jezika, čime se augmentirana rečenica dodatno udaljava od izvorne.

Potpuniji pregled dodatnih tehnika može se pronaći u [3, 9, 14].

B. *Klasifikacija teksta*

Klasifikacija teksta je postupak kojim se nekom tekstu dodjeljuje oznaka iz konačnog skupa unaprijed poznatih oznaka. Koristi se u zadacima poput analize sentimenta, gdje se na temelju izraženog stava tekst označava polaritetom (npr. pozitivan ili negativan), zatim u zadacima filtriranja informacija na relevantne i irelevantne, označavanju imenovanih entiteta (eng. *Named Entity Recognition - NER*) poput osoba, lokacija ili organizacija, te označavanju vrsta riječi u rečenici (eng. *Part-of-Speech (POS) tagging*), uz brojne druge zadatke.

Veliki jezični modeli bazirani na arhitekturi transformera [25] pokazuju zadivljujuće rezultate u raznim zadacima obrade jezika [2, 26], međutim, njihova primjena nije uvijek preporučljiva niti moguća zbog velikih zahtjeva za resursima. Stoga se kao manje zahtjevna alternativa i dalje koriste inačice konvolucijskih neuronskih mreža (CNN), poput onih usmjerenih na riječi [27] ili usmjerenih na pojedine znakove [23]. Drugu grupu popularnih modela čine oni temeljeni na povratnim mrežama, poput LSTM [28] ili BiLSTM [29] koja procesiranje sekvencijalnih podataka vrši u oba smjera, unaprijed i unazad. Osim toga, korpusi na kojima su trenirani veliki jezični modeli već sadrže brojne varijacije riječi i njihovih kombinacija, zbog čega se pokazalo da augmentacija podataka u tom slučaju nije potrebna, a može dovesti i do lošijih rezultata [30]. Raniji radovi [29] već su ukazali na prednosti jednostavnijih modela u situaciji kad se raspolaže s malom količinom podataka.

Ovaj je rad najsličniji [31] u kojem su autori usporedili različite tehnike augmentacije, ali na problemu klasifikacije slika. Osim toga, ovdje nije korišten prijenos znanja već su mreže trenirane od početka. Ovaj rad ima određenih sličnosti i s [3], u kojem su autori usporedili tehnike za augmentaciju teksta, ali koristeći BERT-base model, te [15] u kojem su autori koristili samo WordNet za odabir sinonima. Za razliku od toga, ovdje je uspoređeno više različitih pristupa u odabiru sinonima kako bi se utvrdilo odražava li se dodatna kompleksnost pozitivno i na rezultate klasifikacije.

III. METODE I EKSPERIMENT

A. *Podaci*

Kao podloga za transformacije upotrijebljen je „Subjectivity" [32], standardni skup za provjeru performansi klasifikatora. Skup se sastoji od 10000 recenzija i sažetaka filmova na engleskom jeziku. Podaci su prikupljeni iz dva izvora: Rotten Tomatoes[1] i Internet Movie Database[2]. Prvi izvor smatra se subjektivnim dok se drugi smatra objektivnim, a zadatak je klasificirati rečenice na objektivne i subjektivne. Skup je uravnotežen te ima po 5000 primjera svake od navedenih dviju klasa. S obzirom da nije unaprijed podijeljen na podskupove za treniranje, validaciju i testiranje modela, u podjeli podataka slijedio se primjer iz [15] te je slučajnim odabirom izdvojeno 1000 primjera za testiranje, a ostatak je korišten za treniranje.

[1] http://www.rottentomatoes.com

[2] http://www.imdb.com

Iz podskupa za treniranje slučajnim su odabirom dodatno izdvojena dva manja, veličine 100 i 1000 primjera, kako bi se simulirala situacija s ograničenim podacima. Pri tome je svaki podskup pravi podskup tj. svi primjeri iz manjega nalaze se u većem, a uz to je sačuvan i postojeći omjer među klasama. Tijekom procesa učenja izdvajalo se 20% primjera za validaciju. Statistike skupova za treniranje dane su u Tablici I.

B. *Tehnike augmentacije*

U eksperimentu je korišteno sedam transformacija generiranih pomoću knjižnice nlpaug [33]. To su:

Na razini pojedinog znaka:

1. Keyboard" (u daljnjem tekstu KB) - simulacija pogrešaka prilikom tipkanja zamjenom slučajno odabranih znakova (1-3) drugim bliskim znakovima.

Na razini riječi:

2. „Synonym" (WN) - zamjena slučajno odabranih riječi (1-3) njihovim sinonimima. Za odabir sinonima korištena je leksička baza WordNet.
3. „Word embedding" (WE) - zamjena slučajno odabranih riječi (1-3) sinonimima. Sinonimi su birani na temelju kosinusne sličnosti riječi reprezentiranih pomoću GloVe vektora riječi.
4. „Contextual word embedding" (CWE) - zamjena slučajno odabranih riječi (1-3) sinonimima koji odgovaraju kontekstu rečenice, za što je korišten model DistillBERT.
5. „Delete word" (DW) - brisanje slučajno odabranih riječi (1-2).
6. „Swap word" (SW) - zamjena poretka slučajno odabranih riječi (1-3).

Na razini rečenice:

7. „Backtranslation" (BT) – parafraziranje rečenice prevođenjem s engleskog jezika na njemački te ponovno na engleski. Za ovu kombinaciju jezika pokazano je da daje dobre rezultate [34], a za prijevod su korišteni modeli Facebook FAIR's WMT19 (wmt19-en-de i wmt19-en-de) [35].

Primjeri pojedinih augmentacija dani su u Tablici II.

C. *Model i eksperimentalne postavke*

S obzirom na to da cilj rada nije postizanje najboljih performansi na zadatku klasifikacije, već izmjeriti relativni dobitak u točnosti primjenom različitih tehnika augmentacije, za klasifikaciju je upotrijebljena jednostavna konvolucijska mreža prilagođenu radu s

TABLICA I. STATISTIKE SKUPOVA ZA TRENIRANJE

Veličina skupa za treniranje	Prosječna duljina rečenice	Veličina vokabulara
100	23,98	1833
1000	24,11	5956
9000	24,62	21853

TABLICA II. PRIMJERI ODABRANIH TRANSFORMACIJA

Vrsta transformacije	Rezultat
Originalna rečenica	the movie begins in the past where a young boy named sam attempts to save celebi from a hunter .
KB	the mo**D**ie be**r**ins in the past where a y**oJ**ng boy named sam at**f**empts to save **v**elebi from a hunter.
WN	the movie begins in the past where a young **male child** named sam attempts to **relieve** celebi from a **huntsman**.
WE	the **films continues** in the past **outside** a young boy named sam attempts to save celebi from a hunter.
CWE	**same** movie **ends** in **centuries** past where **handsome** young boy named sam attempts to **differentiate himself from monster hunter.**
DW	the _ begins in the past where a young boy named sam attempts to save celebi from _ hunter.
SW	the movie begins **the in where past** a young boy named sam attempts **save to** celebi from a hunter.
BT	**T**he film begins in the past **when** a **little** boy named **S**am **tries** to save **celebs** from a hunter.

KB – "Keyboard", WN - "Synonym", WE - "Word embedding", CWE - "Contextual word embedding", DW - "Delete word", SW - "Swap word", BT - "Backtranslation"

tekstualnim podacima. Iako određene varijante povratnih mreža, kao što je to BiLSTM, daju ponešto bolje rezultate u odnosu na CNN [29], učenje takvih mreža traje duže te je iz tog razloga odabrana CNN mreža.

Mreža se sastoji od dva konvolucijska sloja koji su praćeni slojevima za sažimanje po maksimumu (eng. *max pooling*), odnosno globalnim sažimanjem po maksimumu (eng. *global max pooling*), u skladu s preporukom u [36]. Veličina filtera je 5 dok je broj mapa značajki 128. Za reprezentaciju teksta korišteni su GloVe vektori riječi (glove6B) dimenzionalnosti d=300. Prednost GloVe vektora u odnosu na word2vec je to što uzimaju u obzir supojavljivanje riječi (eng. *co-occurrence*), a u odnosu na fasttext vektore manje su memorijski zahtjevni. Na kraju mreže nalazi se linearni sloj uz sigmoidnu aktivacijsku funkciju za binarnu klasifikaciju. U konvolucijskim slojevima aktivacijska funkcija je ReLU, a za regularizaciju je upotrijebljen dropout [6] uz p=0.5. Za optimizacijski algoritam odabran je Adam [37]. Navedene vrijednosti hiperparametara predstavljaju uobičajene postavke korištene u literaturi [15].

Mreža je trenirana 20 epoha uz ranije zaustavljanje ako kroz tri epohe nije došlo do smanjenja validacijske pogreške. Svaki je eksperiment ponovljen tri puta te je izračunata prosječna točnost na skupu za testiranje.

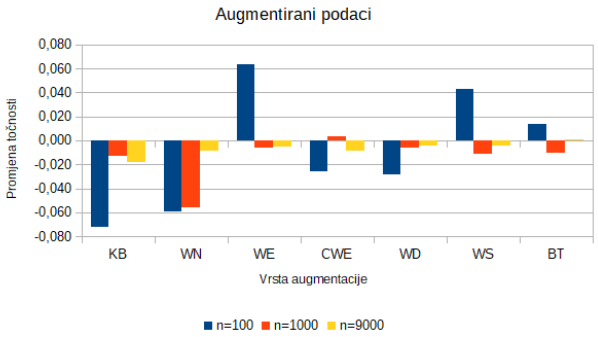
Mreža je trenirana na podskupovima veličine 100, 1000 i 9000 primjera, na originalnim podacima, podacima dobivenim pojedinim transformacijama te kombinacijom originalnih i augmentiranih podataka u dva omjera: u prvom je slučaju za svaki originalni primjer kreirana n=1 transformacija, dok je u drugom slučaju n=3.

IV. REZULTATI

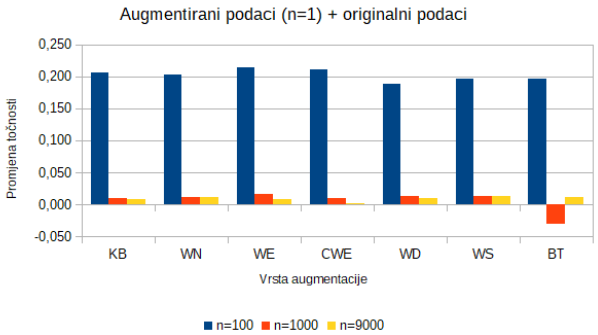
Točnost modela treniranog samo na originalnim podacima veličine 100, 1000 i 9000 primjera iznosi redom: 58,0% (±4,5), 86,5% (±0,7), te 91,4% (±0,3). Najbolji rezultat iznosi 93,6% (±2,3) te je ostvaren modelom treniranim na najvećem skupu koji se sastojao od kombinacije originalnih i KB augmentiranih podataka u omjeru 1:3. Rezultatom slijede kombinacije originalnih podataka s WS (92,6%, ±0,8), i WN augmentacijama u omjeru 1:1 (92,5, ±0,8).

Rezultati klasifikacije izraženi kao razlika točnosti u odnosu na modele trenirane samo na osnovnom skupu prikazani su na slikama 2-4. Može se uočiti da su razlike izraženije na malom skupu (n=100) dok se s povećanjem veličine skupa za treniranje one gube.

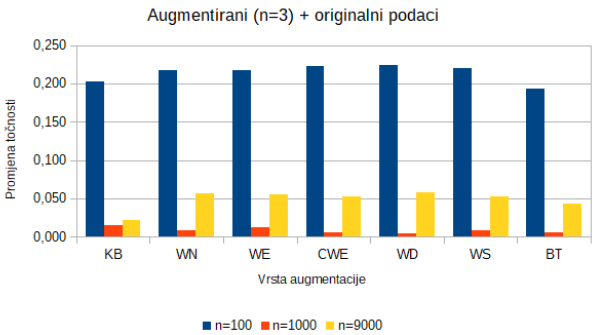
Ako se promatraju rezultati treniranja samo na augmentiranim podacima (Slika 2), može se primijetiti da su oni gotovo u pravilu lošiji od rezultat modela treniranog samo na osnovnom skupu. Iako se ne radi o značajnijim odstupanjima, ipak su te razlike ponešto veće na najmanjem skupu. Međutim, ako se augmentiranim podacima dodaju i originalni (Slika 3), tada dolazi do jasnog poboljšanja točnosti, posebno na najmanjem skupu.



Slika 2. Razlika točnosti modela treniranih na augmentiranim podacima u odnosu na točnost klasifikacije modela treniranih samo na osnovnom skupu veličine 100, 1000 i 9000 primjera.



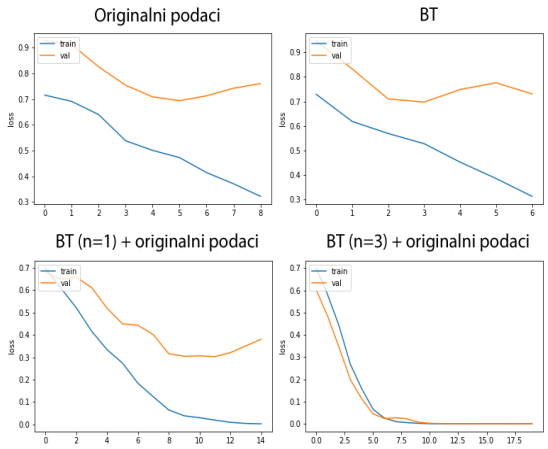
Slika 3. Razlika točnosti modela treniranih na skupovima originalnih podataka kojima su dodani augmentiranim podaci u jednakom omjeru, u odnosu na točnost klasifikacije modela treniranih samo na osnovnom skupu veličine 100, 1000 i 9000 primjera.



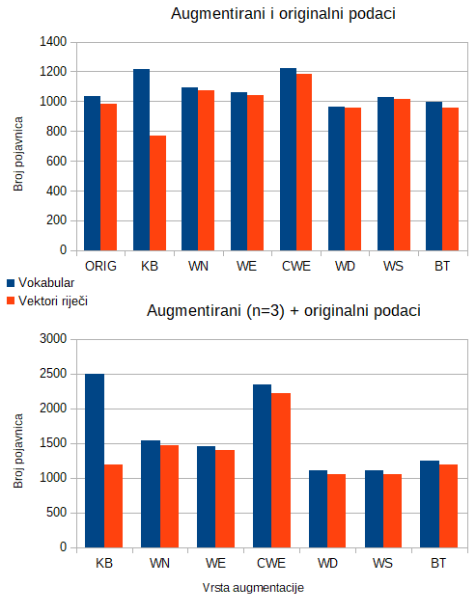
Slika 4. Razlika točnosti modela treniranih na skupovima originalnih podataka kojima su dodani augmentiranim podaci u omjeru 1:3, u odnosu na točnost klasifikacije modela treniranih samo na osnovnom skupu veličine 100, 1000 i 9000 primjera.

Ipak, daljnjim dodavanjem augmentiranih podataka, premda generiranih na način da se postigne što veća raznolikost, nije značajnije pridonijelo većoj točnosti klasifikacije (Slika 4). Iako je s porastom omjera augmentacije s 1:1 na 3:1 došlo do kretanja u pozitivnom smjeru, značajniji je učinak postignut samo na najvećem skupu. Može se primijetiti da su dodatni podaci temeljeni na kontekstualnim vektorima pokazali ponešto povoljniji učinak od ostalih tehnika augmentacije, ali samo u kombinaciji s originalnim podacima. Međutim, generiranje takvih podataka trajalo je nekoliko desetaka puta duže u odnosu na pronalaženje WordNet sinonima. Zanimljivo je da su rezultati na skupu srednje veličine najmanje reagirali na tehnike augmentacije, neovisno o veličini skupa za učenje.

S druge strane, ako se uzmu u obzir krivulje validacijskog gubitka, može se uočiti jasno poboljšanje, odnosno smanjenje prenaučenosti s proširenjem skupa augmentiranim podacima. Slika 5 prikazuje krivulje za primjer augmentacije prevođenjem. Može se uočiti da je u slučaju treniranja samo na originalnim podacima do prenaučenosti dolazilo već tijekom prvih nekoliko epoha.



Slika 5. Krivulje validacijskog gubitka za veličinu osnovnog skupa od 100 primjera i modele trenirane a) samo na originalnim podacima (gore lijevo), b) samo na augmentiranim podacima gore desno), c) na kombinaciji originalnih i augmentiranih podataka (n=1) (dolje lijevo), d) na kombinaciji originalnih i augmentiranih podataka (n=3) (dolje desno). BT - „Backtranslation“.



Slika 6. Veličina vokabulara i pokrivenost vokabulara vektorima riječi skupa za učenje od 100 primjera. U gornjem prikazu uspoređuju se vokabulari originalnih podataka i njihovim transformacija, a u donjem dijelu kombinacije originalnih i augmentiranih podataka (omjer 1:3).

Ako se promatraju pojedine tehnike augmentacije, može se vidjeti da su one prilično ujednačene. Iznenađujuće, parafraziranje prevođenjem s engleskog na njemački i obratno, pokazalo je lošije rezultate od očekivanog. A kad se uzme u obzir da je generiranje takvih podataka trajalo 100-200 puta duže nego npr. KB augmentiranih, postavlja se upitnost primjene takve tehnike, barem u slučaju općenitih tekstova. S druge strane, s obzirom na to da kvaliteta augmentacije ovisi o odabranom modelu, ali i o odabranoj kombinaciji jezika, trebalo bi ispitati dodatne modele te kombinacije leksički udaljenijih jezika, npr. engleskog i japanskog, što bi se moglo pozitivno odraziti na raznovrsnost rečenica.

Slika 6 prikazuje odnos veličine vokabulara i pokrivenosti riječi iz vokabulara GloVe vektorima. S obzirom na to da su korišteni vektori koji prepoznaju samo 400000 riječi, to se odrazilo i na uspješnost tehnika augmentacije. Može se vidjeti da je najveći vokabular dobiven s augmentacijom KB, što je očekivano jer se kod takve augmentacije zamjene vrše znakovima a ne čitavim riječima. Tako se stvara veliki broj nepostojećih riječi koje neće biti pronađene među GloVe vektorima. S druge strane, CWE augmentacija je također stvarala vokabular veći od prosjeka, ali to se proširenje dobilo stvarnim riječima. Moguće da je ta dodatna raznolikost doprinijela nešto većoj točnosti klasifikacije. Iako, promatrajući WD augmentaciju, može se uočiti ispodprosječni vokabular, ali rezultat koji parira najboljem u slučaju kombinacije s originalnim podacima u omjeru 3:1.

V. ZAKLJUČAK

U radu su uspoređene odabrane tehnike augmentacije na primjeru klasifikacije teksta pri čemu su korišteni skupovi za učenje različitih veličina. Rezultati su pokazali da je augmentacija neupitno korisna na skupovima jako ograničene veličine, ali da se ta korist gubi s porastom broja dostupnih primjera za učenje. Također, postoji

granica iznad koje više nije preporučljivo dodatno povećavanje osnovnog skupa jer dobitak u performansama uglavnom ne kompenzira dodatno vrijeme potrebno za pripremu i učenje. Tehnike augmentacije, iako međusobno različite, pokazale su prilično ujednačene rezultate. Stoga se za male skupove općenitog sadržaja, osim jednostavnijih modela, može preporučiti i jednostavnije tehnike augmentacije. U ovom slučaju kompleksnije tehnike nisu opravdale dodatne zahtjeve za resursima. Međutim, određeno ograničenje istraživanja je oslanjanje na vektore riječi relativno malog kapaciteta. Stoga bi se za potpuniju sliku istraživanje trebalo proširiti i na ostale varijante reprezentacije teksta. Uz to, iako su u radu korišteni podaci na engleskom jeziku, odabrane tehnike augmentacije mogle bi se primijeniti i u slučaju drugih jezika, poput hrvatskog, uz pretpostavku da postoje odgovarajući resursi, kao što su to baze sinonima ili modeli za strojno prevođenje, a o kvaliteti navedenih resursa ovisit će i kvaliteta augmentacije. Osim toga, buduća istraživanja mogla bi se usmjeriti i na multimodalnu augmentaciju, što predstavlja još nedovoljno istraženo područje.

REFERENCE

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang, "An empirical survey of data augmentation for limited data learning in NLP," *arXiv preprint arXiv:2106.07499*, 2021.

[4] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," Artificial Intelligence Review, pp. 1–48, 2021.

[5] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[6] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[8] S. I. Nikolenko, "Synthetic data for deep learning," arXiv preprint arXiv:1909.11512, 2019.

[9] S. Y. Feng *et al.*, "A survey of data augmentation approaches for NLP," *arXiv preprint arXiv:2105.03075*, 2021.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[11] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[13] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[14] B. Li, Y. Hou, and W. Che, "Data augmentation approaches in natural language processing: A survey," *AI Open*, vol. 3, pp. 71–90, 2022.

[15] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[16] I. Raffaelli, M. Tadic, B. Bekavac, and Ž. Agic, "Building croatian wordnet," in *Proceedings of GWC*, 2008, pp. 349–360.

[17] W. Y. Wang and D. Yang, "That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviours using# petpeeve tweets," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2257–2563.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[19] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[20] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[22] S. Kolokiyets Oleksandr andBethard and M.-F. Moens, "Model-portabilty experiments for textual temporal analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 2011, vol. 2, pp. 271–276.

[23] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[24] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[25] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention," *arXiv preprint arXiv:2006.03654*, 2020.

[27] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, pp. 1746–1751.

[28] J. Schmidhuber, S. Hochreiter, and others, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] A. Adhikari, A. Ram, R. Tang, and J. Lin, "Rethinking complex neural network architectures for document classification," 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), 2019, pp. 4046–4051.

[30] A. Ezen-Can, "A Comparison of LSTM and BERT for Small Corpus," arXiv preprint arXiv:2009.05451, 2020.

[31] I. Hrga and M. Ivasic-Kos, "Effect of Data Augmentation Methods on Face Image Classification Results," in *Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods - Volume 1*, 2022, pp. 660–667.

[32] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in *Proceedings of the ACL*, 2004.

[33] E. Ma, "NLP Augmentation." 2019. [Online]. Available: https://github.com/makcedward/nlpaug

[34] M. Fadaee, A. Bisazza, and C. Monz, "Data augmentation for low-resource neural machine translation," arXiv preprint arXiv:1705.00440, 2017.

[35] N. Ng, K. Yee, A. Baevski, M. Ott, M. Auli, and S. Edunov, "Facebook FAIR's WMT19 news translation task submission," arXiv preprint arXiv:1907.06616, 2019.

[36] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," *arXiv preprint arXiv:1510.03820*, 2015.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.