Univerzitet u Nišu

ELEKTRONSKI FAKULTET

Katedra za računarstvo

Data augmentacija nad tekstom

SEMINARSKI RAD

Studijski program: Veštačka inteligencija i mašinsko učenje

Student:

Milica Jovanović 1636

Niš, Februar 2024.

Uvod

U obradi prirodnog jezika (NLP), koncept augmentacije podataka se pojavio kao tehnika usmerena na poboljšanje performansi i robusnosti modela mašinskog učenja. Augmentacija podataka odnosi se na proces veštačkog proširivanja veličine i raznolikosti trening skupa podataka primenom različitih transformacija na postojeće podatke. Ova metoda je naročito korisna kada je dostupna količina podataka ograničena ili kada se modeli suočavaju sa izazovima poput overfittinga i raznolikosti ljudskog jezika.

Kako raste potražnja za sofisticiranim NLP aplikacijama, sve je očiglednija potreba za robusnim trening skupovima podataka. U mnogim realnim scenarijima, sticanje velikih količina označenih podataka nije samo dugotrajno, već i skupo. Augmentacija podataka pruža održivo rešenje za ovaj problem omogućavajući istraživačima da maksimalno iskoriste svoje postojeće skupove podataka, čime se ubrzava ciklus razvoja NLP aplikacija.

U ovom radu se istražuje uloga augmentacije u povećanju kvaliteta i otpornosti NLP modela. Analizira se proces pripreme teksta, uključujući uklanjanje šuma, normalizaciju i lematizaciju, kako bi se podaci prilagodili za različite tehnike augmentacije. Detaljno su obrađeni tipovi tehnika augmentacije, uključujući metode u prostoru karakteristika, prostoru podataka i napredne pristupe zasnovane na generativnim modelima poput GAN-ova i varijacionih autoenkodera.

1. Uloga augmentacije podataka

U obradi prirodnog jezika, metode augmentacije podataka igraju ključnu ulogu u poboljšanju performansi modela mašinskog učenja. Ove tehnike omogućavaju generisanje dodatnih podataka, čime se povećava raznovrsnost jezičkih uzoraka i smanjuje overfittting modela. Na primer, korišćenje sinonima, parafraziranje ili menjanje strukture rečenica su popularne strategije koje se koriste za obogaćivanje postojećih tekstualnih podataka. Ove metode ne samo da pomažu u stvaranju većeg obima podataka, već i omogućavaju modelima da bolje generalizuju, čime se povećava njihova efikasnost u stvarnim aplikacijama.

Data augmentacija ima ključnu ulogu u rešavanju zadataka povezanih sa treniranjem modela mašinskog učenja, posebno u NLP-u, gde dostupnost visokokvalitetnih označenih podataka može biti veoma izazovno. U mnogim slučajevima, prikupljanje velikih skupova podataka nije samo dugotrajno, već i skupo, posebno u specijalizovanim domenima kao što su medicinski ili pravni tekstovi. Primenom tehnika augmentacije podataka, može se efikasno povećati obim trening podataka bez potrebe za dodatnim resursima. Ovo je posebno važno kada model može naići na retke ili nedovoljno zastupljene klase tokom zaključivanja. Augmentacija podataka pomaže u ublažavanju overfitting-a, čestog problema u mašinskom učenju gde modeli dobro rade na trening podacima, ali ne uspevaju da se generalizuju na nepoznate podatke. Uvođenjem varijabilnosti u trening skup podataka, modeli mogu da nauče da prepoznaju obrasce i karakteristike koje su reprezentativnije za distribucije realnih podataka. To dovodi do poboljšanih performansi modela, kao i povećane robusnosti protiv buke i adverzarnih primera.

1.1. Overfitting

Overifitting je čest problem u mašinskom učenju gde model uči da izvanredno dobro radi na trening podacima, ali ne uspeva da se generalizuje na nove, nepoznate podatke. Ovo je posebno često u NLP-u, gde kompleksnost i bogatstvo jezika mogu dovesti do toga da modeli pamte specifične primere, umesto da uče osnovne obrasce. Augmentacija podataka rešava ovaj problem uvođenjem varijabilnosti u trening skup. Generisanjem novih primera kroz tehnike augmentacije podataka, modeli su izloženi širem spektru lingvističkih struktura i značenja. Ova izloženost pomaže u jačanju sposobnosti modela da se generalizuje, što na kraju dovodi do poboljšanih performansi na realnim zadacima.

1.2. Raznolikost ljudskog jezika

Ljudski jezik je po svojoj prirodi kompleksan i promenljiv, pod uticajem faktora kao što su kultura, kontekst i individualni izraz. Ova varijabilnost se može manifestovati na brojne načine, uključujući upotrebu idiomatskih izraza, sleng i različite sintaktičke strukture. Na primer, fraza *"pasti s nogu"* je idiom koji znači *biti iscrpljen*, a njegovo značenje možda nije odmah jasno modelu treniranom na literalnijem jeziku. Pored toga, isti sentiment se može izraziti na mnogo načina, kao što su *"jedva stojim na nogama"* ili *"toliko sam umoran da ne mogu više"*. Tehnike augmentacije podataka mogu pomoći da se uoči ova raznolikost stvaranjem višestrukih reprezentacija istog osnovnog značenja. To ne samo da obogaćuje trening skup podataka, već oprema modele sposobnošću da razumeju i interpretiraju jezik fleksibilnije.

1.3. Poboljšanje robustnosti modela

Još jedna značajna prednost augmentacije podataka je poboljšanje robusnosti modela. U realnim primenama, modeli često nailaze na bučne ili neorganizovane podatke koji se mogu značajno razlikovati od čistih, korigovanih skupova podataka korišćenih za treniranje. Uvođenjem uvećanih podataka koji simuliraju različite oblike buke - kao što su tipografske greške, gramatičke greške ili neformalni jezik - modeli mogu da nauče da se efikasnije nose sa takvim varijacijama. Ova robusnost je ključna za primene kao što su četbotovi ili virtuelni asistenti, gde korisnici možda neće uvek komunicirati na standardnom jeziku. Treniranjem na raznovrsnijem skupu podataka, modeli postaju bolje opremljeni da razumeju i reaguju na širi spektar ulaznih podataka, što na kraju dovodi do zadovoljnijeg korisničkog iskustva.

Osim toga, primena naprednih tehnika kao što su generativni modeli ili transformatori može dodatno unaprediti proširenje podataka. Generativni modeli, kao što su Variational Autoencoders (VAE) i Generative Adversarial Networks (GAN), omogućavaju kreiranje novih podataka koji su statistički slični originalnim, ali istovremeno unose varijacije koje mogu pozitivno uticati na učenje modela. Korišćenjem ovih pristupa, mogu se stvoriti opširniji skupovi podataka koji obuhvataju različite jezičke i kulturne aspekte, što je posebno važno za primene u višejezičnim ili kulturno specifičnim okruženjima.

2. Priprema teksta za augmentaciju

Tehnike za augmentaciju teksta predstavljaju metode kojima se obrađuju podaci koji se u originalnom obliku nalaze u tekstualnom formatu. S obzirom na to da nad podacima u tekstualnom formatu nije moguće direktno primeniti matematičke metode mašinskog učenja ili trenirati model, neophodno je transformisati te podatke u pogodan format za obradu.[2] Taj format je vektorski ili tenzorski prostor, u kojem se tekstualni kontekst skupa podataka reprezentuje.

Da bi se došlo do vektorskog oblika podataka prvo je potregno proći kroz sledeće metode preprocesiranja podataka[2]:

- Uklanjanje šuma
- Normalizacija
- Tokenizacija
- Stemovanje
- Lematizacija

Na kraju obrađeni tekstualni podaci postaju vektori, koji se u vektorskom ili tenzorskom prostoru predstavljaju kao nizovi brojeva određene dužine. Vrednosti koje ti vektori nose treba na kvalitetan način da opišu tekstualni podatak, uz očuvanje svojstava poput redosleda reči u rečenici, kontekstualnih zavisnosti između reči, dužine tekstova, kao i kontekstualnih sličnosti između njih.

2.1. Uklanjanje šuma

Uklanjanje šuma podrazumeva uklanjanje znakova, cifara i delova teksta koji mogu ometati analizu. Uklanjanje šuma je jedan od najvažnijih koraka u pretprocesiranju teksta. Postoji više načina za uklanjanje šuma. Ovo uključuje uklanjanje interpunkcije, specijalnih karaktera, brojeva, HTML formatiranja, uklanjanje ključnih reči specifičnih za domen (npr. „RT“ za retvit), uklanjanje izvornog koda, uklanjanje zaglavlja i drugo. Veoma je zavisno od domena. Na primer, u Tvitovima, šum mogu biti svi specijalni karakteri osim heštagova, jer oni označavaju pojmove koji mogu karakterisati Tvit [2]. Problem sa šumom je što može dovesti do nekonzistentnih rezultata.

	raw_word	stemmed_word
0	..trouble..	..trouble..
1	trouble<	trouble<
2	trouble!	trouble!
3	<a>trouble	<a>trouble
4	1.trouble	1.troubl

Slika 1. Primer izlaza sa šumom

Može se primetiti da sve izvorne reči iznad sadrže neki šum oko sebe. Kada se primeni stemming na ove reči, rezultat stemming-a ne izgleda lepo. Nijedna reč nema ispravan koren.[2] Međutim, sa malo čišćenja, rezultati izgledaju mnogo bolje što se može uočiti na slici 2.

	raw_word	cleaned_word	stemmed_word
0	..trouble..	trouble	troubl
1	trouble<	trouble	troubl
2	trouble!	trouble	troubl
3	<a>trouble	trouble	troubl
4	1.trouble	trouble	troubl

Slika 2. Primer izlaza nakon redukcije šuma

2.1.2. Uklanjanje zaustavnih reči

Zaustavne reči (stop words) su skup uobičajeno korišćenih reči u nekom jeziku. Primeri zaustavnih reči na srpskom jeziku su „šta“, a“, „je“, i“ i slično. Intuicija iza korišćenja zaustavnih reči je ta da, uklanjanjem reči sa niskom informativnošću iz teksta, možemo se fokusirati na one važne.

Na primer, u kontekstu pretraživačkog sistema, ako je upit „šta je pretprocesiranje teksta?“, sistem za pretragu treba da pronađe dokumente koji govore o pretprocesiranju teksta, a ne o pojmovima „šta je“. Ovo se postiže tako što se reči sa

liste zaustavnih reči sprečavaju da budu analizirane. Zaustavne reči se često koriste u pretraživačkim sistemima, aplikacijama za klasifikaciju teksta, modeliranju tema, ekstrakciji tema i drugim zadacima.

Uklanjanje zaustavnih reči, iako je efikasno u sistemima za pretragu i ekstrakciju tema, pokazalo se kao manje kritično u klasifikacionim sistemima. Ipak, pomaže u smanjenju broja karakteristika koje treba razmotriti, što olakšava održavanje modela u razumnim veličinama. [2]

Primer uklanjanja zaustavnih reči, sve zaustavne reči se zamenjuju lažnim karakterom, npr. sa „W“:

```
original sentence =  this is a text full of content and we need to clean it up
sentence with stop words removed=  W W W text full W content W W W W clean W W
```

Slika 3. Primer sa uklonjenim stop rečima

Stop liste mogu poticati iz unapred definisanih setova, ili se može kreirati prilagođena lista za izabrani domen. Neke biblioteke (npr. sklearn) omogućavaju uklanjanje reči koje se pojavljuju u određenom procentu dokumenata, što takođe može dati efekat uklanjanja zaustavnih reči. [2]

2.2. Normalizacija

Veoma zanemaren korak u pretprocesiranju teksta je normalizacija teksta. Normalizacija teksta je proces pretvaranja teksta u kanonski (standardni) oblik. Na primer, reči kao što su „gooood“ i „gud“ mogu biti transformisane u „good“, njihov kanonski oblik. Drugi primer je mapiranje skoro identičnih reči kao što su „stopwords“, „stop-words“ i „stop words“ u „stopwords“.
Normalizacija teksta je važna za tekstove koji sadrže šum, kao što su komentari na društvenim mrežama, tekstualne poruke i komentari na blogovima gde su prisutne skraćenice, pravopisne greške i reči van rečnika.[2]

Raw	Normalized
2moro 2mrrw 2morrow 2mrw tomrw	tomorrow
b4	before
otw	on the way
:) :-) ;-)	smile

Slika 4. Reči pre i posle normalizacije

Ne postoji standardni način za normalizaciju teksta. Obično zavisi od zadatka. Neki uobičajeni pristupi normalizaciji teksta uključuju mapiranja rečnika (najlakše), statističke mašinske prevode (SMT) i metode zasnovane na ispravljanju pravopisa.

2.2.1. Prebacivanje u mala slova

Konverzija svih podataka u mala slova jedna je od najjednostavnijih i najefikasnijih formi pretprocesiranja teksta. Primenjiva je na većinu problema u obradi teksta, a može pomoći posebno kada skup podataka nije prevelik i kada je potrebna doslednost u očekivanom izlazu.

Evo primera kako konverzija u mala slova rešava problem retkosti, gde se iste reči sa različitim velikim slovima mapiraju u isti oblik sa malim slovima:[2]

Raw	Lowercased
Canada CanadA CANADA	canada
TOMCAT Tomcat toMcat	tomcat

Slika 5. Primer prebacvanja velikih u mala slova

Najjednostavniji način za to u Python-u jeste korišćenje ugrađene funkcije lower(). Ova metoda konvertuje sva velika slova u stringu u mala slova i vraća rezultat.[3]

2.2.2. Standardizacija teksta

Standardizacija teksta je proces pretvaranja neformalnog teksta u formalan i lako obradiv format. Cilj je da se različiti oblici reči, skraćenice, ili neformalni izrazi prepoznaju i pretvore u standardizovane oblike kako bi se omogućila preciznija analiza teksta u narednim fazama obrade. To je posebno važno u analizama gde tekstualni podaci dolaze iz neformalnih izvora kao što su društvene mreže, blogovi, ili recenzije korisnika.

Tekstualni podaci često sadrže skraćenice i tipografske greške, koje je potrebno standardizovati kako bi modeli mogli da razumeju značenje. Na primer, „btw” treba prepoznati kao „by the way.” Standardizacija pomaže u stvaranju konzistentnih podataka, smanjujući varijacije koje mogu nastati zbog različitih stilova pisanja ili izbora reči. Kada se tekst standardizuje, postaje lakše modelima mašinskog učenja

da obrade i analiziraju podatke, čime se poboljšava tačnost zadataka kao što su klasifikacija, prepoznavanje entiteta, ili sentiment analiza.

2.2.3. Ispravljanje pravopisa

Ispravljanje pravopisa je proces identifikacije i ispravljanja tipografskih i pravopisnih grešaka u tekstualnim podacima. Ovo je posebno važno kod obrade podataka koji dolaze iz neformalnih izvora, kao što su blogovi, komentari na društvenim mrežama, ili recenzije korisnika, gde su greške česte.

Greške u kucanju ili pravopisne greške mogu stvoriti više verzija iste reči. Na primer, „proccessing“ i „processing“ bi se tretirale kao različite reči, iako imaju isto značenje. Ispravljanje pravopisa smanjuje ovakve varijacije, što olakšava obradu podataka.[2] Kada se pravopis ispravi, modeli mašinskog učenja mogu preciznije analizirati podatke jer nema grešaka koje bi ih zbunile. Ovo može direktno poboljšati performanse modela za klasifikaciju, ekstrakciju entiteta, ili analizu sentimenta.

Pre nego što se primeni ispravljanje pravopisa, važno je prvo rešiti skraćenice i kolokvijalne izraze, jer u suprotnom, sistem za ispravku može napraviti greške. Na primer, ako se „ur“ ne zameni s „your“ pre ovog koraka, sistem bi mogao da ispravi „ur“ u „or,“ što bi unelo dodatne greške u tekst. Za ispravku pravopisa mogu se koristiti jednostavni alate kao što je TextBlob, koji automatski prepoznaje i ispravlja greške u tekstu.

2.3. Tokenizacija

Tokenizacija je možda najvažniji korak u pretprocesiranju teksta, zajedno sa kodiranjem teksta u numeričku reprezentaciju, pre korišćenja NLP-a i jezičkih modela. To podrazumeva razbijanje svakog unosa teksta na vektor delova ili tokena. U najjednostavnijem scenariju, tokeni su najčešće povezani sa rečima, ali u nekim slučajevima, kao kod složenih reči, jedna reč može rezultirati sa više tokena. Određeni znaci interpunkcije (ako nisu prethodno uklonjeni kao šum) takođe ponekad mogu biti identifikovani kao zasebni tokeni.[4]

Rečenica: "Mašinsko učenje je fascinantno!", nakon procesa tokenizacije, bi se mogla podeliti na sledeće tokene:

1. Tokenizacija po rečima:
    - ["Mašinsko", "učenje", "je", "fascinantno", "!"]

U ovom slučaju, svaka reč postaje poseban token, uključujući i znak interpunkcije "!".

2. Sub-rečna tokenizacija (kada se koriste modeli koji dele reči na manje delove):

- ["Mašins", "ko", "učenje", "je", "fasci", "nant", "no", "!"]
- Ovde je reč "fascinantno" podeljena na nekoliko delova jer model prepoznaje složene reči kao više tokena, što pomaže u generalizaciji reči koje nisu često viđene u skupu podataka.

Tokenizacija je ključna zato što omogućava modelima da tekstualne podatke pretvore u formu koja je pogodna za dalju obradu, poput numeričkog kodiranja za algoritme učenja.

2.4. Stemming

Stemming je proces smanjenja infleksija u rečima (npr. troubled, troubles) na njihov osnovni oblik (npr. trouble). "Koreni" u ovom slučaju možda nisu stvarni koreni reči, već samo kanonski oblik originalne reči. Stemming koristi grubi heuristički proces koji skraćuje krajeve reči u nadi da će ispravno transformisati reči u njihov osnovni oblik. Tako bi reči trouble, troubled i troubles mogle biti konvertovane u troubl umesto u trouble jer su krajevi jednostavno odsečeni.

Postoje različiti algoritmi za stemming. Najpoznatiji algoritam, koji se pokazao efikasnim za engleski jezik, je Porterov algoritam. [4]

Stemming je koristan za rešavanje problema retkosti podataka i za standardizaciju vokabulara. Bio je uspešan u aplikacijama za pretragu. Ideja je da, ako se pretražuje „deep learning classes“, treba da se pronađu dokumenti koji spominju „deep learning class“ kao i „deep learn classes“, iako ovo poslednje zvuči nepravilno. Potrebno je da se pokriju sve varijacije reči kako bi se dobili najrelevantniji rezultati.

2.5. Lemmatizing

Lemmatizacija na prvi pogled izgleda vrlo slično stemmovanju, gde je cilj ukloniti infleksije i mapirati reč na njen osnovni oblik. Razlika je u tome što lemmatizacija pokušava to da uradi na ispravan način. Ne seče jednostavno krajeve, već zapravo transformiše reči u njihov pravi koren. Na primer, reč „better“ bi se mapirala u „good“. Može koristiti rečnik kao što je WordNet za mapiranje ili neku specijalnu metodu zasnovanu na pravilima.

Zavisno od algoritma, lemmatizing može biti znatno sporija u poređenju sa osnovnim stemmovanjem, a možda će biti potrebno znati deo govora reči kako bi se dobio ispravan lemat. Lemmatizacija nema značajan uticaj na tačnost klasifikacije teksta sa neuronskim arhitekturama.

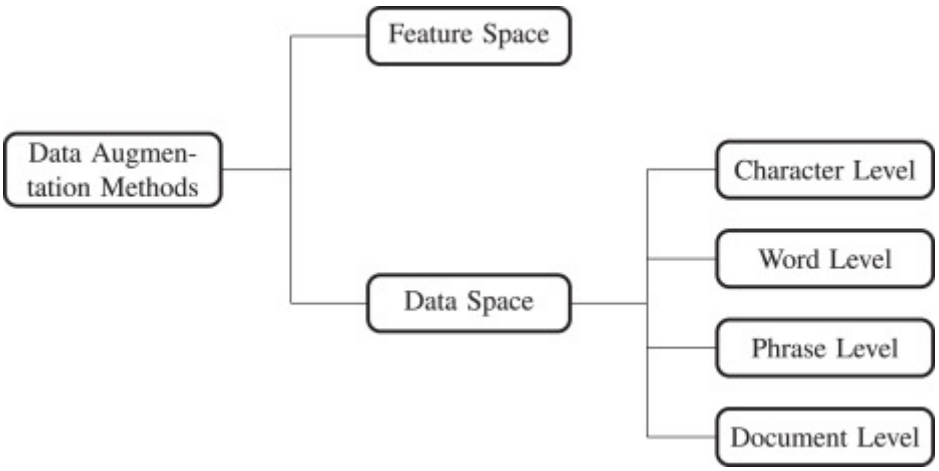
Na primer kod reči automobili i kuće, lemmatizacija prepoznaje da je "automobili" množina od "automobil" i vraća osnovni oblik reči, takođe prepoznaje da je "kuće" oblik reči "kuća" i vraća osnovni oblik reči.

Stemming jednostavno uklanja sufiks "i" kako bi dobio osnovni oblik reči, tako automobili postaje automobil (što je u ovom slučaju osnovan oblik reči), a reč kuće postaje kuć.

3. Tipovi tehnika augmentacije podataka

Spektar tehnika za proširenje podataka za tekst je i raznovrstan i složen, obuhvatajući niz metoda koje se mogu grubo kategorizovati u nekoliko različitih tipova. Svaka kategorija ima jedinstvenu svrhu i karakteriše se specifičnim mehanizmima transformacije.

Na sledećoj slici je prikazana podela augmentacije tekstualnih podataka[5]. Ona prvo deli metode augmentacije podataka u NLP-u prema prostoru na kojem se odvija proces augmentacije. Metode augmentacije u prostoru karakteristika (Feature Space), koje se oslanjaju na reprezentaciju podataka, kao što su word embeddings ili aktivacionii vektori neuronskih mreža. Metode augmentacije u prostoru podataka (Data Space) deluju direktno na tekstualne podatke u njihovom sirovom obliku, primenjujući transformacije na karaktere, reči, fraze ili same dokumente.[6]



Slika 6. Graf tipova data augmentacije

3.1. Metode augmentacije u prostoru karakteristika

Feature space data augmentation tehnike se najčešće koriste kada su podaci već predstavljeni u nekom kompresovanom obliku. Tehnike augmentacije podataka u prostoru karakteristika (feature space data augmentation) odnose se na metode kojima se veštački uvećava skup podataka direktno u prostoru karakteristika, a ne u sirovom obliku podataka. U kontekstu obrade prirodnog jezika (NLP), to znači da se

augmentacija ne primenjuje na sam tekst, već na njegove reprezentacije, poput vektorskih ili numeričkih oblika (embeddinga).

3.1.1. Dodavanje šuma

Ova tehnika podrazumeva dodavanje nasumičnih varijacija u vektorske reprezentacije podataka. Na primer, mala količina nasumičnog šuma može biti dodata u vektore karakteristika kako bi se simulirale varijacije u podacima. To pomaže modelu da postane otporniji na male varijacije u ulaznim podacima i generalizuje bolje na nepoznate primere.

3.1.2. Interpolacija

Za tekstualne podatke, metode interpolacije su većinom ograničene na prostor karakteristika, pošto ne postoji intuitivan način za kombinovanje dve različite tekstualne instance. Ipak, primena u prostoru karakteristika izgleda razumno, budući da interpolacija skrivenih stanja dve rečenice stvara novu koja sadrži značenje obe originalne rečenice. Pored ovoga, iz perspektive zasnovane na učenju, metode interpolacije imaju visoku vrednost za modele mašinskog učenja. Moguća objašnjenja za uspeh metoda interpolacije mogu proizaći iz balansiranja klasa, zaglađivanja odluke (regularizacija) i poboljšanja reprezentacija.

3.1.2.1. Interpolacija između primera (Mixup)

Ova metoda kombinuje vektore karakteristika dva različita primera tako što ih meša (linearna interpolacija) i koristi kao novi uzorak. Na taj način model dobija uvećani skup podataka koji sadrži kombinovane karakteristike postojećih podataka.

3.1.2.2. SMOTE Interpolacija

SMOTE je metoda koja generiše nove uzorke za manjinske klase u skupu podataka. Umesto da jednostavno duplicira postojeće uzorke manjinske klase, SMOTE koristi interpolaciju da kreira nove uzorke između postojećih. To se postiže tako što se za svaki uzorak manjinske klase identifikuju njegovi najbliži susedi (npr. korišćenjem udaljenosti kao što je Euklidova udaljenost) i generišu nove tačke na osnovu linearnih kombinacija postojećih tačaka i njihovih suseda.

3.1.3. Permutacija i pertubacija u prostoru karakteristika

Ova tehnika uključuje reorganizaciju ili blagu promenu redosleda karakteristika unutar vektora, bez značajnog narušavanja originalnog značenja ili strukture. Cilj je generisati varijacije koje pomažu modelu da postane robusniji.

3.1.4. Smanjenje dimenzionalnosti i projekcija

U nekim slučajevima, smanjenje dimenzionalnosti (npr. pomoću PCA ili t-SNE metoda) može se koristiti za kreiranje sažetih verzija vektorskih reprezentacija, koje se potom mogu ponovo projikovati u prostor veće dimenzionalnosti, stvarajući nove varijacije podataka.

3.1.5. Embedding Transformations

Ova tehnika koristi preoblikovanja vektorskih reprezentacija reči ili rečenica, kao što su rotacije, translacije ili skaliranje u prostoru vektora. Time se generišu novi primeri koji zadržavaju semantičku sličnost sa originalnim podacima, ali se razlikuju u numeričkom obliku.

3.2. Metode augmentacije u prostoru podataka (Data Space)

Tehnike augmentacije podataka se primenjuju na tri nivoa podataka:

- Nivo karaktera
- Nivo reči
- Nivo rečenica
- Nivo dokumenta

3.2.1. Nivo karaktera

Najniži nivo augmentacije podataka je augmentacija na nivou karaktera. Augmentacija na nivou karaktera je jedna od osnovnih tehnika uvećavanja podataka u obradi prirodnog jezika, posebno korisna za jezike sa velikim brojem karaktera ili za zadatke gde je važna ortografija. Ova vrsta augmentacije se fokusira na manipulaciju pojedinačnim karakterima unutar teksta, kako bi se generisale nove verzije podataka koje zadržavaju osnovno značenje, ali su dovoljno različite da doprinesu generalizaciji modela. Postoje dva osnovna načina za augmenaciju podataka na nivou karaktera to su dodavanje šuma i manipulacija karakterima.

Najčešće metode augmentacije na nivou karaktera uključuju:

- Zamena karaktera sinonimima.
    - Zamena slova sličnog zvuka ili značenja (npr. zamena 'a' sa 'e' u nekim rečima). Ovo je posebno korisno za jezike sa velikim brojem dijalekatskih varijacija.
    - Primer: Ako se u reč "kafa" zameni "a" sa "e", dobija se "kefa".
- Brisanje karaktera.
    - Slučajno brisanje pojedinačnih karaktera unutar reči kako bi se simulirale greške pri kucanju ili optičko prepoznavanje karaktera.
    - Ako se u reč "telefon" izbriše "o", dobija se "telefn".

- Ubacivanje karaktera.
    - Dodavanje novih karaktera na slučajna mesta unutar reči.
    - Ako se u reč "auto" ubaci "p" između "a" i "u", dobija se "aputo".
- Zamena karaktera slučajnim karakterima.
    - Zamena pojedinačnih karaktera nasumičnim karakterima ili brojevima.
    - Ako se u reč "pas" zameni "p" sa "q", dobija se "qas".
- Zamena karaktera sličnim karakterima.
    - Zamena karaktera sličnim karakterima iz drugih jezika (npr. zamena 'a' sa 'ä' u nemačkom).

Augmentacija na nivou karaktera može pomoći modelima da postanu robusniji na različite stilove pisanja i varijacije u kvalitetu slike. Generisanje novih parova rečenica sa različitim varijacijama u pravopisu može pomoći modelima mašinskog učenja da bolje generalizuju i da se nose sa šumom u podacima. Takođe, augmentacija može pomoći modelima da prepoznaju različite načine izražavanja istog sentimenta, čak i kada postoje greške u pravopisu.

3.3. Augmentacija na nivou reči

Augmentacija na nivou reči podrazumeva tehnike koje se koriste za generisanje novih reči ili varijacija postojećih reči kako bi se obogatio skup podataka za obuku modela mašinskog učenja. Ove tehnike mogu pomoći u povećanju raznolikosti podataka, smanjenju overfitting-a i poboljšanju generalizacije modela.

Easy Data Augmentation (EDA) je jedan primer jednostavnih tehnika manipulacije, koje kombinuju nekoliko manipulacija u jedinstvenu metodu [5]. Ova metoda se sastoji od primene skupa jednostavnih operacija na originalni tekst kako bi se generisali novi sintetički tekstovi.Za svaku rečenicu u skupu za testiranje, nasumično se bira i izvodi jedna od sledećih operacija:

3.3.1. Zamena sinonima

Jedna od najjednostavnijih, ali efikasnih tehnika uključuje zamenu reči u datom tekstu njihovim sinonimima. Korišćenjem sinonima ili promene reči alatima kao što su Word2Vec ili GloVe, može se generisati više varijacija rečenice, zadržavajući njeno originalno značenje. Ova metoda osim sto zadržava originalno značenje, uvodi i leksičku raznolikost. Na primer, u rečenici "Pas trči brzo," reč "brzo" se može zameniti sa "hitro," čime se dobija: "Pas trči hitro.", čime se proširuje dataset bez promene osnovnog semantičkog sadržaja. Ova tehnika je posebno korisna u situacijama kada model treba da nauči prepoznavanje različitih izraza iste ideje, što je ključno za zadatke poput analize sentimenta i klasifikacije teksta.

3.3.2. Slučajno ubacivanje

Nasumično umetanje podrazumeva dodavanje novih reči u postojeće rečenice, što može povećati složenost i bogatstvo teksta. Ova tehnika zahteva pažljivo razmatranje kako bi se osiguralo da su umetnute reči kontekstualno odgovarajuće i da ne narušavaju koherentnost originalne rečenice. Na primer, umetanje reči "brzo" u rečenicu "Pas je trčao" daje "Pas je trčao brzo," čime se proširuje dataset uz održavanje gramatičkog integriteta. Dodavanjem prideva ili priloga, ova tehnika može pomoći modelima da razumeju nijanse jezika, što je često prisutno u prirodnom jeziku.

3.3.3. Slučajno brisanje

Ova tehnika podrazumeva slučajno uklanjanje reči iz rečenice, što može simulirati greške pri kucanju ili skraćivanje. Na primer, u rečenici "Devojka igra fudbal," može se izbrisati reč "fudbal," čime se dobija: "Devojka igra.". Ova metoda podstiče modele da se fokusiraju na najvažnije delove rečenice i može biti posebno korisna u zadacima gde je sažetost od suštinskog značaja, poput sumiranja teksta.

3.3.4. Slučajna zamena reči

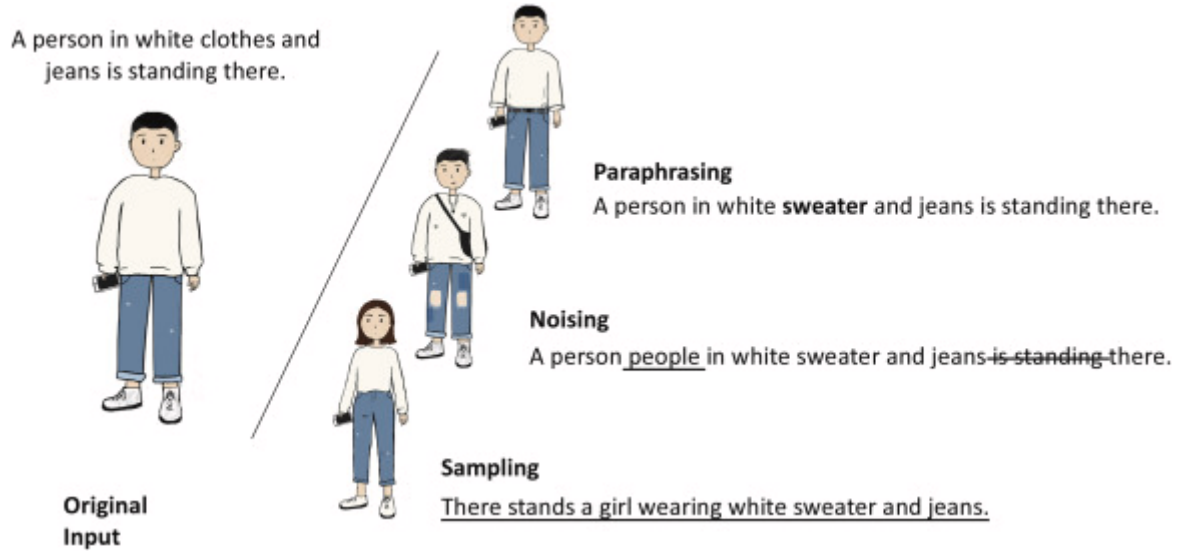
Slučajna zamena reči je metoda augmentacije teksta u kojoj se nasumično biraju dve reči u rečenici i zamene svojim mestima. Cilj ove metode je da generiše nove varijante rečenica zadržavajući njihovo značenje, dok uvodi male promene u strukturi kako bi se poboljšala raznolikost u obuci modela. Na primer, rečenica: "Pas trči po parku." može se promeniti u "Park trči po pasu." Iako ovde značenje može postati nejasno, uvek se bira ograničeni broj zamena koje i dalje zadržavaju razumljivost u mnogim slučajevima.

3.4. Augmentacija na nivou rečenica

Augmentacija rečenica se fokusira na generisanje novih rečenica iz postojećih tekstualnih podataka. Ona se razlikuje od augmentacije karaktera i reči po tome što se fokusira na transformaciju celih rečenica umesto pojedinačnih karaktera ili reči. Ova tehnika ima za cilj generisanje novih rečenica koje zadržavaju suštinu originalnih, ali sa promenjenom strukturom. Na osnovu validnosti, od augmentiranih podataka se očekuje da budu raznovrsni kako bi se poboljšala generalizacija modela u zadacima koji slede. Ovo se odnosi na raznolikost augmentiranih podataka.

Metode augmentacije podataka se mogu grubo podeliti u tri kategorije prema raznolikosti njihovih augmentiranih podataka: parafraziranje, dodavanje šuma i uzorkovanje.[7]

- **Metode zasnovane na parafraziranju** generišu augmentirane podatke koji imaju ograničenu semantičku razliku u odnosu na originalne podatke, na osnovu pravilnih i ograničenih promena u rečenicama. Augmentirani podaci prenose veoma slične informacije kao originalni oblik.
- **Metode zasnovane na dodavanju šuma** dodaju diskretni ili kontinualni šum pod uslovom da se obezbedi validnost. Poenta ovih metoda je da poboljšaju robusnost modela.
- **Metode zasnovane na uzorkovanju** savladaju raspodelu podataka i uzimaju uzorke novih podataka unutar njih. Ove metode generišu raznovrsnije podatke i zadovoljavaju više potreba zadataka koji slede, na osnovu veštačkih heuristika i treniranih modela.



Slika 7. Primer augmentacije na novou rečenica

3.4.1. Metode zasnovane na parafraziranju

Jedan od načina augmentacije teksta na nivou rečenica je parafraziranje. Parafraziranje je način preformulisanja teksta kako bi se očuvalo izvorno semantičko značenje. Konstruisanje i identifikacija parafraza su oblasti od velikog značaja u istraživanjima obrade prirodnog jezika (NLP), kao što su sažimanje ili odgovaranje na pitanja. Zbog svojih karakteristika preformulisanja rečenica, parafrazatori mogu da se ponašaju poput dobrih algoritama za augmentaciju podataka (DA) [7], jer uvode leksičku raznovrsnost, zadržavajući vernost originalnom značenju.

Jedan od načina parafraziranja je zamena p reči korišćenjem WordNet-a. Za svaku rečenicu se biraju sve reči koje je moguće zameniti i nasumično se bira p reči koje se menjaju. Široko korišćena metoda augmentacije teksta nazvana EDA (Easy Data

Augmentation Techniques) takođe zamenjuje originalne reči njihovim sinonimima koristeći WordNet.

Drugi način prafraziranje je korišćenje unapred treniranih modela dubokog učenja. Ova metoda prevazilazi ograničenja u pogledu raspona zamene i vrsta reči u metodi zasnovanoj na tezaurusu. Koristi unapred trenirane embeding modele, kao što su Glove, Word2Vec, FastText, itd., i zamenjuje originalnu reč u rečenici njenim najbližim susedom u embedding prostoru.

Slika 8. Metod parafraziranja

3.4.2. Metode zasnovane na dodavanju šuma

Metode zasnovane na dodavanju šuma kod augmentacije rečenica podrazumevaju ubacivanje slučajnih grešaka, suvišnih reči ili manjih promena unutar rečenice kako bi se povećala raznovrsnost podataka, uz očuvanje ključnog značenja. Cilj ovih metoda je da model postane otporniji na greške i varijacije koje se mogu pojaviti u stvarnim podacima, poboljšavajući njegovu robusnost.

Najčešće korišćene metode zasnovanih na dodavanju šuma su:

- Nasumična zamena reči
    - Dodaju se ili menjaju reči u rečenici, obično reči koje nisu ključne za značenje rečenice.
    - Na primer rečenica "On ide u školu." može postati "On zapravo ide u školu."
- Nasumična permutacija (random swap)
    - Menjaju se mesta reči u rečenici kako bi se unela mala greška ili promena u redosledu.
    - Na primer, rečenica: "Pas trči po parku." može se promeniti u "Park trči po psu."
- Umetanje nasumičnih reči
    - Ubacuju se reči koje nisu neophodne, ali zadržavaju gramatičku ispravnost rečenice.
    - Na primer, rečenica:"Ona voli da peva." može se promeniti u "Ona stvarno voli da peva."
- Brisanje reči
    - Izostavljanje slučajnih reči iz rečenice da bi se stvorio efekat nepotpunosti, ali dovoljno informacija ostaje da se razume značenje.

    - Na primer, rečenica: "Dečak igra fudbal svakog dana." može se promeniti u "Dečak igra fudbal."
- Umetanje tipografskih grešaka
    - Namerno ubacivanje manjih pravopisnih ili tipografskih grešaka u rečenicu, kako bi se model učinio otpornim na greške pri unosu.
    - Na primer, rečenica: "On je danas pročitao knjigu." može se promeniti u "On je dnas pročitao knjigu."
- Parcijalno zamagljivanje reči
    - Zamenjuje se deo reči slučajnim simbolima ili greškama, zadržavajući većinu značenja, ali dodajući šum.
    - Na primer, rečenica: "On voli čokoladu." može se promeniti u "On v0li čokoladu."

3.4.3. Metode zasnovane na uzorkovanju

Metode zasnovane na uzorkovanju kod augmentacije rečenica generišu nove rečenice tako što uzimaju uzorke iz distribuiranog skupa podataka ili modela. Ove metode se oslanjaju na uzorkovanje iz statističkih modela ili distribucija podataka kako bi kreirale nove, slične rečenice koje zadovoljavaju određene jezičke i semantičke karakteristike. Cilj ovih metoda je stvaranje raznovrsnih podataka koji pokrivaju širok spektar varijacija u jeziku.

Evo nekoliko uobičajenih metoda zasnovanih na uzorkovanju:

- Uzorkovanje iz jezičkog modela (Language Model Sampling)
    - Ova tehnika koristi unapred trenirane jezičke modele, kao što su GPT-2, XLNet, ili BERT, kako bi generisala nove rečenice na osnovu uzorkovanja iz raspodela koje model nauči. Model generiše rečenice koje su statistički verovatne na osnovu skupa podataka na kojem je treniran.
    - Primer: "Pas trči po parku." "Pas šeta kroz zelenilo."
- Uzorkovanje iz distribucije sinonima
    - Ova tehnika koristi unapred definisane distribucije sinonima iz resursa kao što je WordNet ili ugradnje reči (npr. Word2Vec, GloVe). Rečenice se generišu tako što se reči iz rečenice nasumično zamenjuju sinonimima iz njihovih raspodela.
    - Primer: "Mačka spava na kauču." "Mačka drema na sofi."
- Embedding Sampling
    - Umesto da se koriste direktni sinonimi, uzorkovanje iz prostora semantičkih ugradnji koristi metode kao što su Word2Vec, GloVe, ili FastText. Reč se menja najbližim susedom u prostoru ugradnji, čime se generiše nova rečenica slična originalu, ali s različitim izborom reči.
    - Primer: "Učenik brzo rešava zadatke." "Đak brzo rešava probleme."

- Sequence Sampling
    - Ova tehnika koristi stohastičke modele za generisanje novih sekvenci reči. Sekvencijalni modeli poput LSTM-a ili transformera mogu generisati nove rečenice birajući sledeću reč na osnovu raspodele prethodnih reči.
    - Primer: "On je kupio novi bicikl." "On je nabavio sjajan bicikl."
- Statistical Sampling
    - Ova metoda uključuje uzorkovanje iz statističkih modela kao što su n-gram modeli ili Markovljevi lanci. Na osnovu učestalosti pojavljivanja reči u korpusu, modeli generišu nove rečenice koje liče na one u korpusu.
    - Primer: "Auto se brzo kreće niz ulicu." "Vozilo se brzo kotrlja niz put."

Razlika između metoda augmentacije na nivou reči i augmentacije na nivou rečenica je ta što se na nivou reči posmatra samo ta reč koja se menja, dok se na nivou rečenice posmatra cela rečenica i njen kontekst.

3.5. Augmentacija na nivou dokumenta

Augmentacija na nivou dokumenta obuhvata tehnike koje transformišu čitav dokument ili velike segmente teksta, a ne samo pojedinačne reči ili rečenice. Ove metode ciljaju na promenu strukture, toka i rasporeda informacija u dokumentu, ali zadržavaju osnovno značenje i ključne informacije. Takve metode su korisne za poboljšanje modela koji rade na nivou paragrafa, članaka, eseja ili drugih dugih formi teksta.

Uobičajene metode za augmentaciju na nivou dokumenta:

- Parafraziranje na nivou dokumenta
    - Ova metoda koristi napredne modele za obradu prirodnog jezika (kao što su GPT-2, T5, ili BART) da preformulišu čitav dokument, zadržavajući njegovo osnovno značenje, ali menjajući strukturu, formulacije i ponekad redosled rečenica ili pasusa.
- Promena redosleda paragrafa
    - Ova tehnika menja redosled pasusa ili segmenata unutar dokumenta, zadržavajući logičan tok. Korisna je za obuku modela koji treba da se nose sa različitim stilovima pisanja i strukturama teksta.
- Dodavanje ili uklanjanje pasusa
    - Nasumično se dodaju ili uklanjaju delovi teksta koji nisu ključni za osnovno značenje, kao što su uvodne ili zaključne napomene. Ova metoda pomaže u variranju dužine i stila teksta.
- Sinteza dokumenata
    - Ova tehnika kombinuje dva ili više različitih dokumenata ili segmenata iz različitih izvora u jedan novi dokument. Čime se povećava raznolikost

i složenost skupa podataka, omogućavajući modelima da se bolje prilagode kombinovanim informacijama.

- Izmena stila pisanja
    - Ova tehnika menja stil pisanja celog dokumenta, prilagođavajući ga različitim tonovima, formalnostima ili stilovima (npr. naučni, novinarski, marketinški stil).
    - Koristi se za obuku modela da prepoznaju različite stilove pisanja.
- Preuređivanje sadržaja
    - Preuređivanje delova dokumenta kako bi se promenio tok informacija, npr. promene u redosledu izlaganja argumenata, zaključaka ili primera.

3.5.1. Back translation

Povratno prevođenje je sofisticiranija tehnika koja podrazumeva prevođenje rečenice na drugi jezik, a zatim vraćanje na originalni jezik. Ova metoda ne samo da uvodi varijabilnost, već često rezultira suptilnim promenama koje mogu obogatiti skup podataka. Na primer, fraza "Uživam u čitanju knjiga" može se prevesti na francuski, a zatim vratiti na srpski, što daje varijacije poput "Uživam u čitanju literature." Ova metoda ne samo da generiše nove primere za obuku, već uvodi i jezičke varijacije koje mogu poboljšati otpornost modela na različite načine izražavanja. Ova tehnika je posebno korisna u višejezičnim kontekstima, gde izloženost različitim jezičkim strukturama može poboljšati prilagodljivost i performanse modela na različitim jezicima.

3.6. Generative Models

Generativni modeli, posebno oni zasnovani na arhitekturama dubokog učenja, postali su značajni u oblasti proširenja podataka. Modeli kao što su Generative adversarial networks (GANs) i Varijacion autoenkoderi (VAEs) mogu se koristiti za generisanje novih tekstualnih podataka. Ovi modeli koriste složene algoritme da uče iz postojećih podataka i stvaraju nove izlaze koji mogu značajno unaprediti skupove podataka za obuku.

3.6.1. Generative Adversarial Networks (GANs)

GANs se sastoje od dve neuronske mreže, generatora i diskriminatora, koje se treniraju istovremeno. Generator kreira nove tekstualne uzorke, dok diskriminator procenjuje njihovu autentičnost. Kroz ovaj suparnički proces, GANs mogu proizvoditi visokokvalitetni sintetički tekst koji može proširiti postojeće skupove podataka. Generator uči da stvara tekst koji imitira podatke iz obuke, dok diskriminator uči da razlikuje stvaran od generisanog teksta. Ova dinamika dovodi do kontinuiranog

poboljšanja kvaliteta generisanih uzoraka, što čini GANs moćnim alatom za proširenje tekstualnih podataka.

3.6.2. Variational Autoencoders (VAEs)

Varijacioni autoenkoderi (VAEs) su još jedna klasa generativnih modela koji uče da kodiraju ulazne podatke u latentni prostor, a zatim ih dekodiraju nazad u originalni prostor. Uzimanjem uzoraka iz ovog latentnog prostora, VAEs mogu generisati nove tekstualne uzorke koji zadržavaju statističke osobine podataka iz obuke. VAEs su posebno korisni za generisanje raznovrsnih izlaza, jer mogu istraživati varijacije u latentnom prostoru kako bi proizveli različite, ali koherentne rečenice. Ova sposobnost omogućava kreiranje bogatih skupova podataka koji mogu poboljšati performanse različitih zadataka obrade prirodnog jezika.

3.7. Adversarijalne tehnike

Adversarijalne tehnike uključuju kreiranje primera koji su specifično dizajnirani da izazovu otpornost modela mašinskog učenja. Ove tehnike mogu biti korisne u augmentaciji skupova podataka tako što uvode adversarijalne primere koji simuliraju potencijalne izazove iz stvarnog sveta. Izlaganjem modela ovim izazovnim scenarijima, može se poboljšati njihova sposobnost generalizacije i postizanja dobrih rezultata u različitim uslovima.

Napredne tehnike uključuju korišćenje generativnih modela, kao što su GPT-3 ili BERT, za kreiranje potpuno novih rečenica na osnovu zadanog upita. Ovi modeli mogu generisati kontekstualno relevantan tekst koji se može koristiti za augmentaciju postojećih skupova podataka. Ovaj pristup ne samo da povećava količinu podataka, već i poboljšava njihovu kvalitetu iskorišćavanjem mogućnosti najsavremenijih jezičkih modela.

3.7.1. Generacija adversarijalnog teksta

Ovaj pristup uključuje izmenu postojećih uzoraka teksta na način koji je neprimetan ljudima, ali može zbuniti modele mašinskog učenja. Na primer, blage izmene u strukturi rečenice ili izboru reči mogu stvoriti adversarijalne primere koji testiraju granice tačnosti modela. Tehnike poput perturbacija na nivou karaktera, gde se pojedinačni karakteri menjaju ili zamenjuju, mogu stvoriti suptilne promene koje mogu dovesti do pogrešne klasifikacije od strane modela. Ova metoda ne samo da pomaže u identifikaciji ranjivosti modela, već i doprinosi obuci modela da budu otporniji na takve adversarijalne napade.

3.7.2. Generacija tekstualnih implikacija i kontradikcija

Druga adversarijalna tehnika uključuje generisanje primera koji ili impliciraju ili kontradikciju u odnosu na dato stanje. Kreiranjem parova rečenica gde jedna logički sledi iz druge ili je direktno oprečna, modeli se mogu obučiti da bolje razumeju odnose između različitih delova teksta. Ovo može biti posebno korisno u zadacima kao što je natural language inference, gde je razumevanje nijansi implikacije i kontradikcije ključno.

3.7.3. Contextual Word Embeddings

Sa pojavom naprednih jezičkih modela kao što su BERT i GPT, kontekstualni vektori reči postali su moćan alat za augmentaciju podataka. Korišćenjem ovih modela, reči se mogu zameniti kontekstualno relevantnim alternativama koje se generišu na osnovu okolnog teksta. Ovaj pristup osigurava da se semantička koherentnost rečenice očuva dok se uvodi varijabilnost koja može poboljšati skup podataka za obuku. Korišćenje kontekstualnih vektora omogućava nijansiranije razumevanje jezika, omogućavajući modelima da shvate suptilne razlike u značenju na osnovu konteksta. Korišćenje kontekstualnih vektora, kao što su ELMo ili BERT, omogućava generisanje augmentiranog teksta koji odražava nijanse korišćenja reči u različitim kontekstima. Zamenom reči sa njihovim kontekstualno prikladnim ekvivalentima, ova metoda može stvoriti varijacije koje su bliže načinu na koji se jezik koristi u praksi.

4. Primena

U ovom poglavlju će se spomenuti neke od najpopularnijih primena data augmentacije.

4.1. Klasifikacija teksta

Augmentacija podataka igra ključnu ulogu u zadacima klasifikacije teksta, gde se modeli obučavaju da klasifikuju dokumente na osnovu njihovog sadržaja. Korišćenjem tehnika augmentacije, praktičari mogu stvoriti uravnoteženije skupove podataka koji smanjuju efekte neuravnoteženosti klasa, što na kraju dovodi do pouzdanijih rezultata klasifikacije. Ovo je posebno korisno u scenarijima gde su određene klase možda nedovoljno zastupljene, omogućavajući modelima da uče iz sveobuhvatnijeg skupa primera.

Za zadatke klasifikacije teksta, kao što su detekcija spama ili kategorizacija tema, augmentacija podataka može poboljšati raznolikost primera za obuku. Ovo je posebno korisno u slučajevima kada su određene klase nedovoljno zastupljene, osiguravajući da je model izložen široj paleti primera.

4.2. Analiza sentimenta

U oblasti analize sentimenta, augmentacija podataka može značajno poboljšati otpornost modela koji su zaduženi za klasifikaciju sentimenta izraženog u tekstualnim podacima. Augmentacijom skupova podataka za obuku sa raznolikim izrazima sentimenta, modeli se mogu obučiti da prepoznaju širi spektar jezičkih nijansi, čime se poboljšava njihova prediktivna tačnost. Ovo je posebno važno u aplikacijama kao što je praćenje društvenih mreža, gde se izraženi sentiment može značajno razlikovati u zavisnosti od konteksta i formulacije.

U analizi sentimenta, augmentacija podataka može pomoći u poboljšanju sposobnosti modela da generalizuje kroz različite izraze sentimenta. Augmentacijom skupova podataka sa raznolikom formulacijom i sinonimima, modeli mogu naučiti da preciznije identifikuju sentiment, čak i kada je jezik korišćen u primerima za obuku različit.

4.3. Mašinski prevod

U oblasti mašinskog prevođenja, augmentacija podataka može poboljšati kvalitet prevoda izlaganjem modela širem spektru jezičkih struktura i izraza. Tehnike kao što je back translation mogu biti posebno korisne u ovom kontekstu, jer uvode varijabilnost dok čuvaju semantičku suštinu originalnog teksta. Ovo ne samo da poboljšava fluentnost prevoda, već takođe pomaže modelima da bolje razumeju idiomatske izraze i kulturne nijanse.

U mašinskom prevođenju, augmentacija podataka može se koristiti za poboljšanje otpornosti prevodnih modela. Generisanjem parafraziranih rečenica ili korišćenjem back translation, modeli mogu naučiti da se nose sa širim spektrom struktura rečenica i jezičkih nijansi, što dovodi do tačnijih prevoda.

4.4. Prepoznavanje imenovanih entiteta

Augmentacija podataka se takođe može primeniti na zadatke prepoznavanja imenovanih entiteta (NER), gde je cilj identifikovati i klasifikovati entitete unutar teksta. Augmentacijom podataka za obuku sa varijacijama rečenica koje sadrže imenovane entitete, modeli mogu poboljšati svoju sposobnost prepoznavanja entiteta u raznolikim kontekstima.

4.5. Pretraživanje informacija

Još jedna značajna primena augmentacije podataka je u sistemima za pretragu informacija, gde je cilj poboljšati relevantnost rezultata pretrage. Augmentacijom skupova podataka upita sa varijacijama pretraživačkih termina i fraza, modeli se mogu obučiti da bolje razumeju nameru korisnika i preuzmu

relevantnije dokumente. Ovo može dovesti do poboljšanog zadovoljstva korisnika i boljih performansi pretraživača i sistema preporuka.

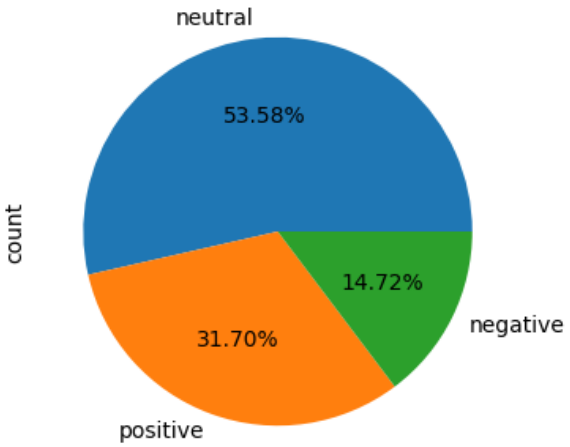
4.6. Chatbotovi i Conversational Agents

U razvoju chatbotova i konverzacionih agenata, augmentacija podataka može pomoći u kreiranju zanimljivijih i raznovrsnijih dijaloga. Generisanjem različitih odgovora i pitanja, programeri mogu obučiti modele koji su bolje opremljeni da se nose sa širokim spektrom interakcija sa korisnicima.

5. Praktičan rad

Praktični deo ovog rada bavi se primenom tehnika data augmentacije na tekstualnim podacima, sa posebnim fokusom na analizu sentimenta. Kao ulazni skup podataka korišćen je dataset koji sadrži rečenice sa pridruženim oznakama sentimenta (*positive*, *negative*, *neutral*). Cilj rada je da se kroz augmentaciju proširi dataset, unapredi raznovrsnost podataka i omogući bolja generalizacija modela za analizu sentimenta.

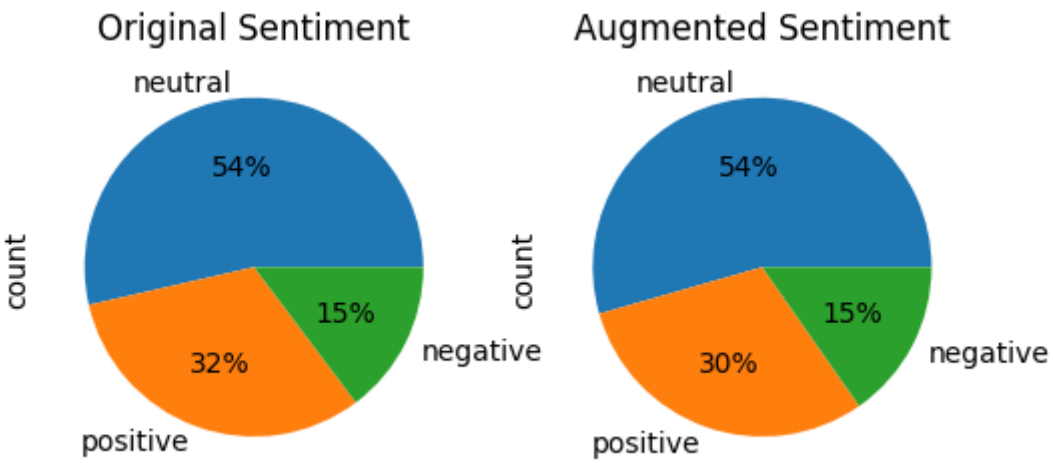
Dataset se sastoji od dve kolone: tekstualnih podataka koji predstavljaju izjave korisnika i kategorija sentimenta (*positive*, *negative*, *neutral*). Prvi korak u obradi bio je prikaz osnovnih statistika dataset-a, uključujući ukupan broj primera, distribuciju po kategorijama sentimenta i primere rečenica za svaku kategoriju. Rezultati ove analize prikazani su na sledećoj slici.



Slika 10. Raspodela sentimeta u datasetu

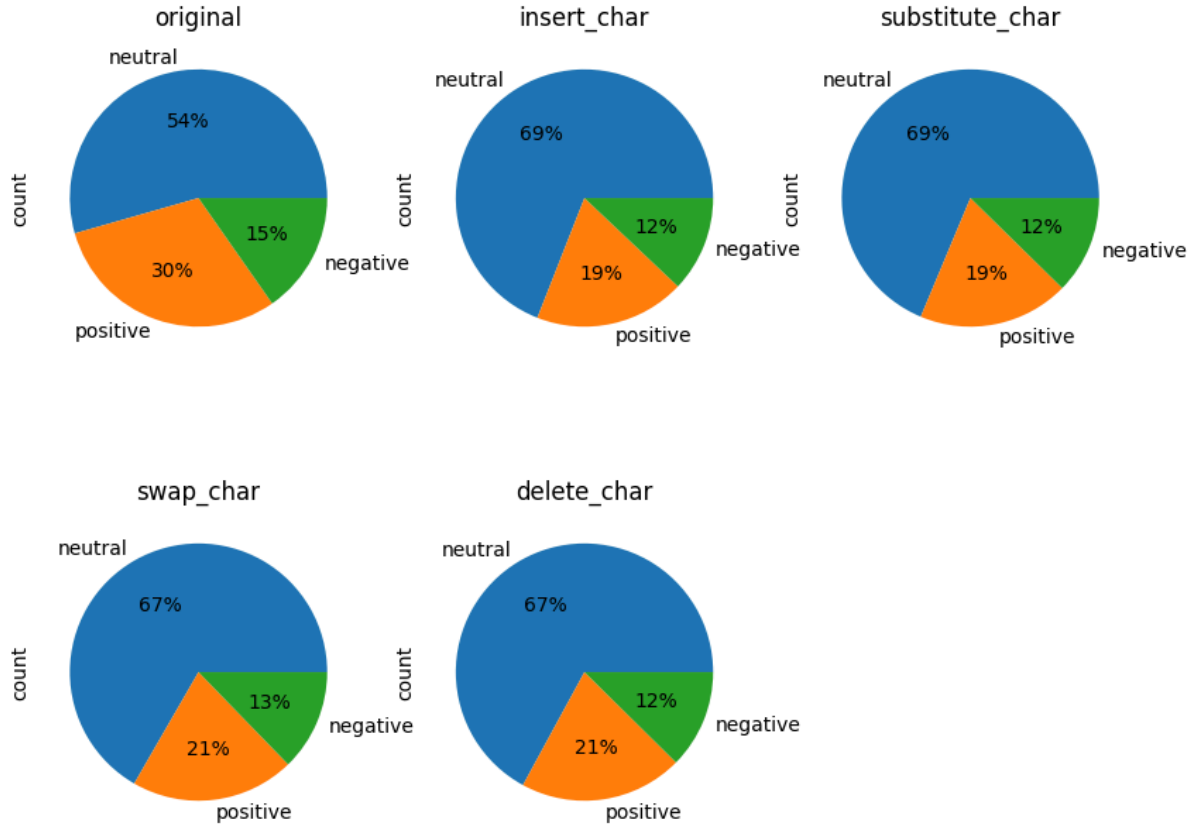
Sledeći korak u obradi tekstualnih podataka je uređivanje teksta kako bi modeli mogli da rade sa datim podacima. Tekst se prvo tokenizuje, odnosno deli na pojedinačne reči i simbole, pri čemu se svi karakteri prebacuju u mala slova kako bi se obezbedila uniformnost. Zatim se uklanjaju česte reči koje nemaju značajnu informativnu vrednost, poznate kao stop reči, pomoću unapred definisanog skupa takvih reči za engleski jezik. Nakon toga, primenjuje se lematizacija, čime se smanjuje redundantnost i obezbeđuje semantička konzistentnost. Na kraju, obrađene reči se spajaju nazad u jedinstven tekstualni niz, čime se kreira preprocesiran tekst spreman za analizu ili primenu u modelima mašinskog učenja.

Prvo je primenjena analiza nad čistim podacima, na sledećoj slici je prikazan rezultat analize.

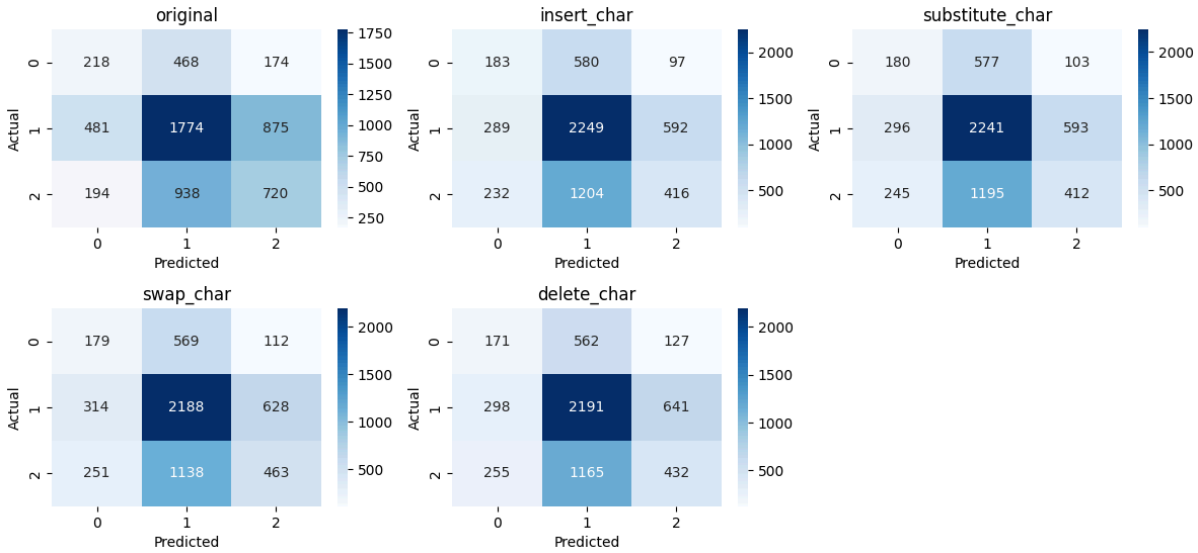


Slika 11. Analiza nad originalnim datasetom

Augmentacija karaktera - EDA

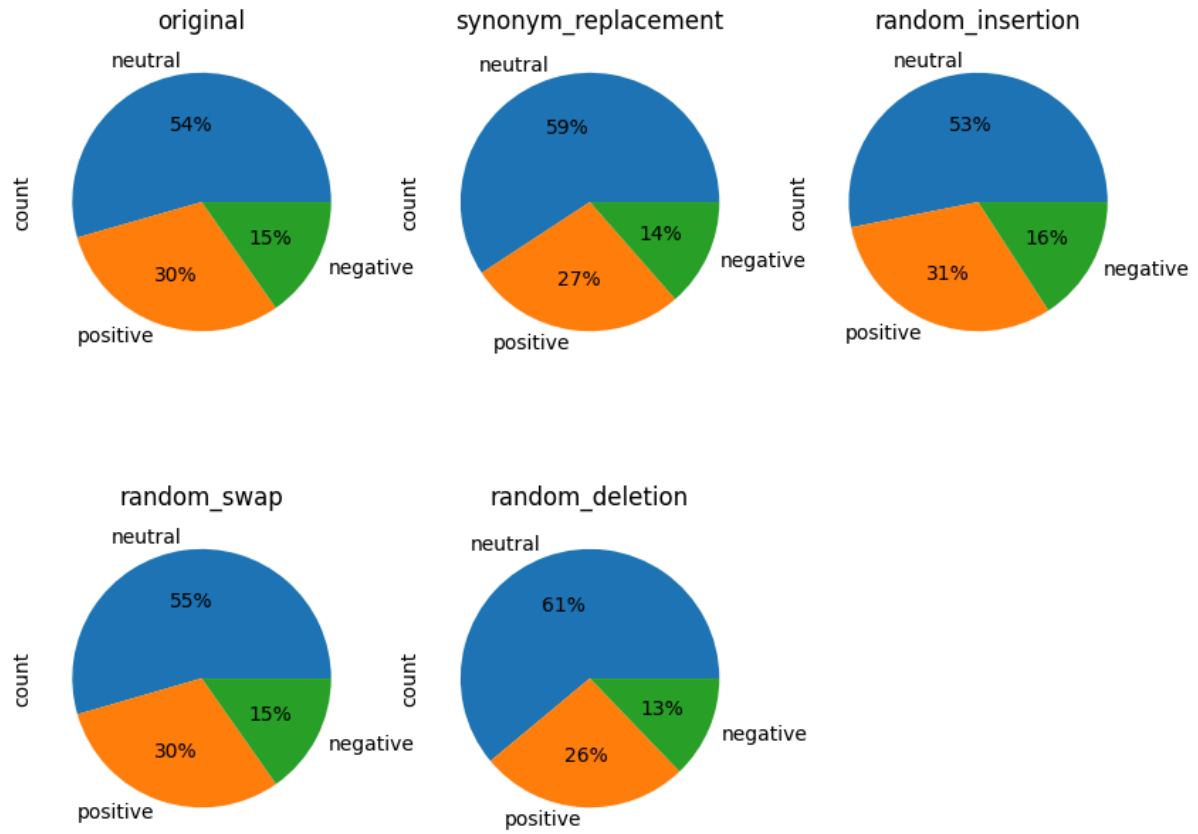


Slika 12. Analiza nad sa EDA augmentacijom karaktera

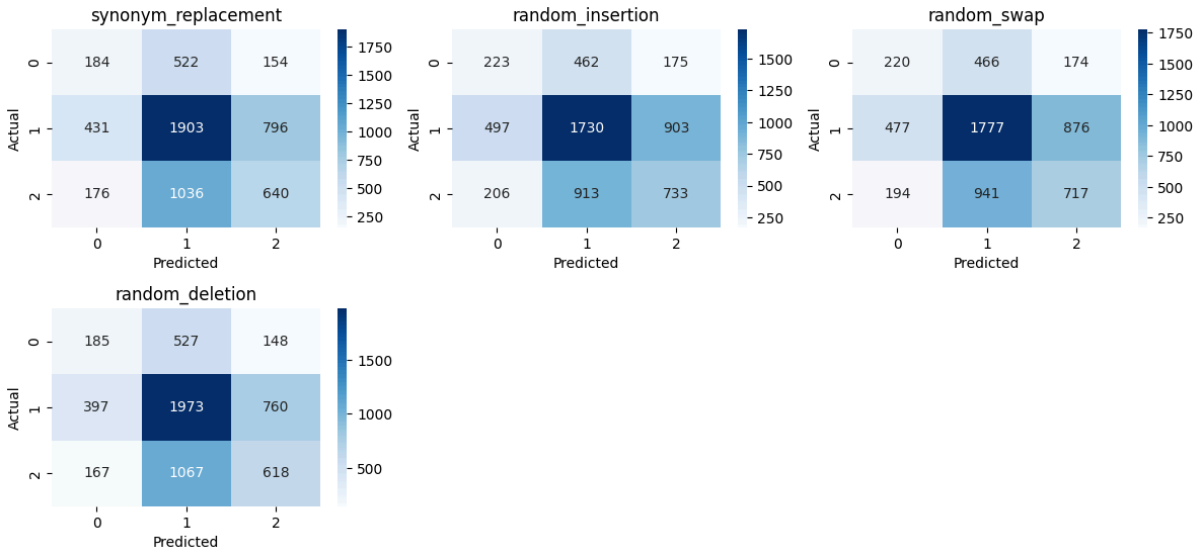


Slika 13. Matrica konfuzije za analizu sa EDA augmentacijom karaktera

Augmentacija reči - EDA

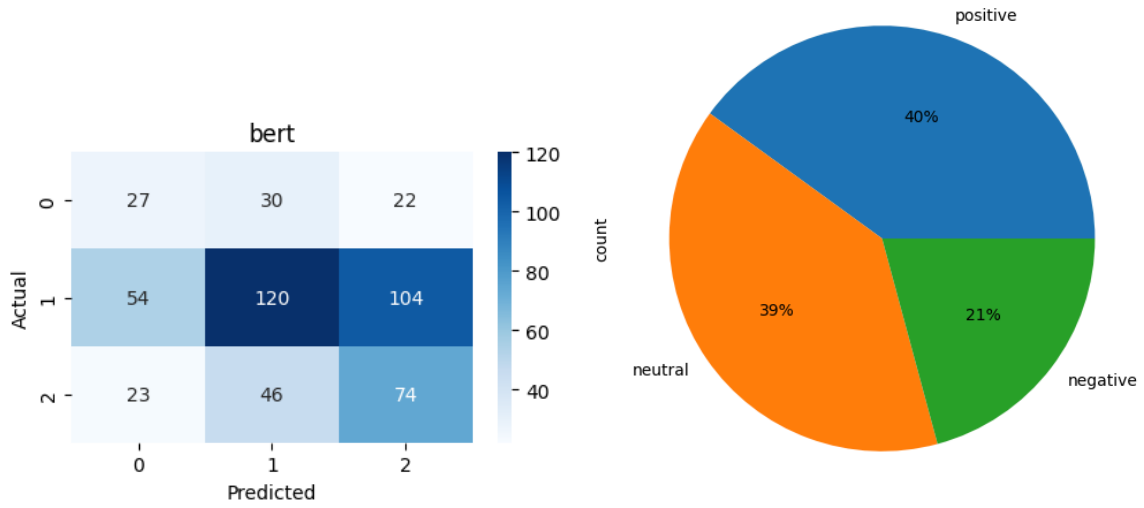


Slika 14. Raspodela nakon analize sa EDA augmentacijom reči



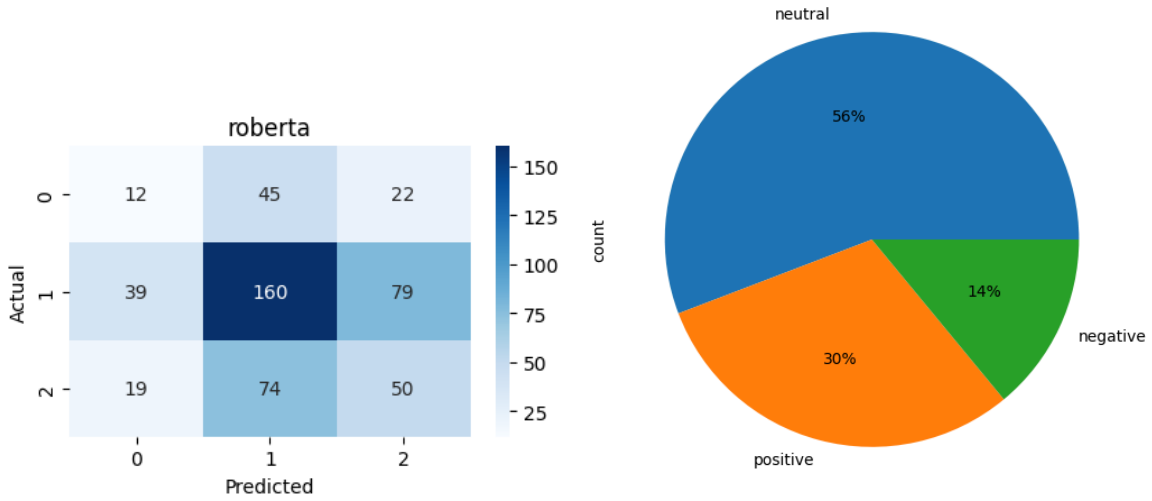
Slika 15. Matrca konfuzije nakon analize sa EDA augmentacijom reči

Contextual word embeddng za reči - BERT



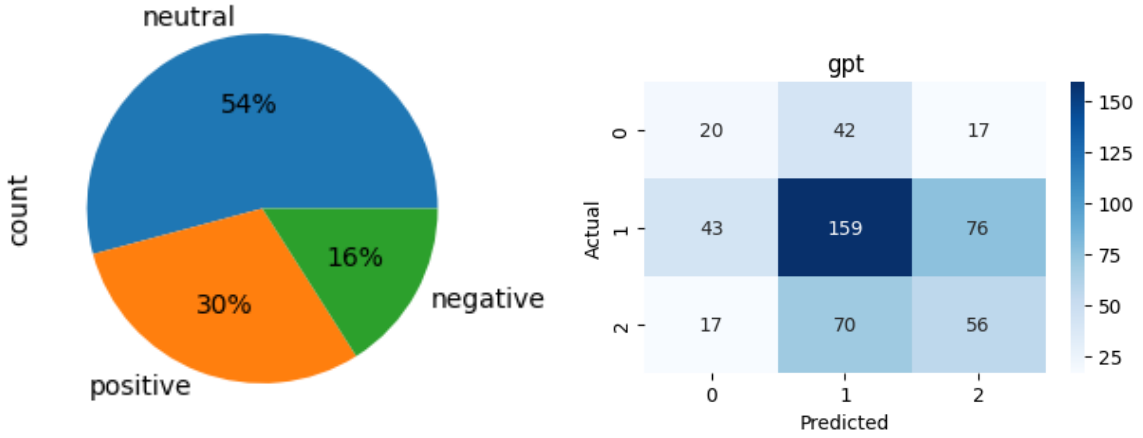
Slika 16. Matrica kofuzije i raspodela nakon analize sa BERT modelom za augmentaciju reči

Contextual word embedding za reči - RoBERTa



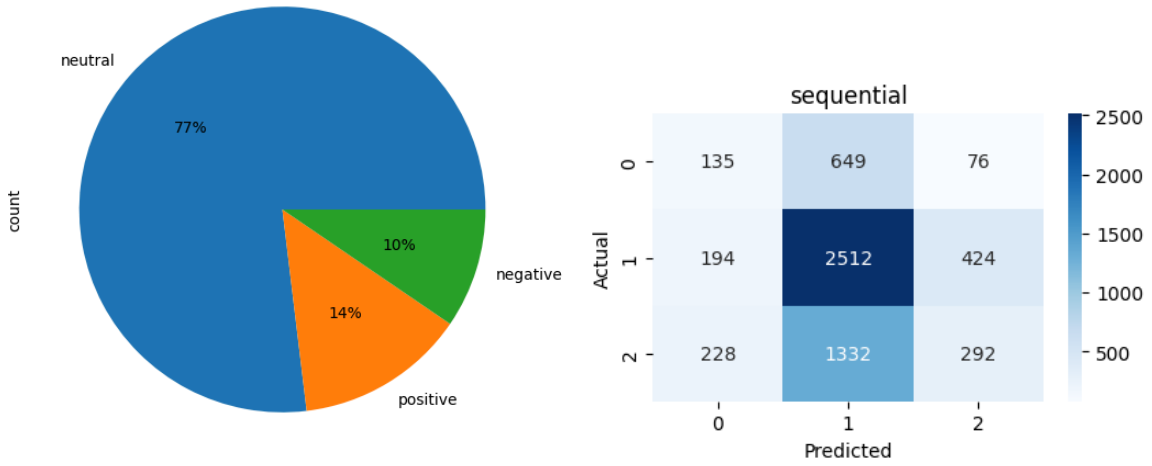
Slika 17. Matrica kofuzije i raspodela nakon analize sa RoBERTa modelom za augmentaciju reči

Contextual word embedding za rečenice



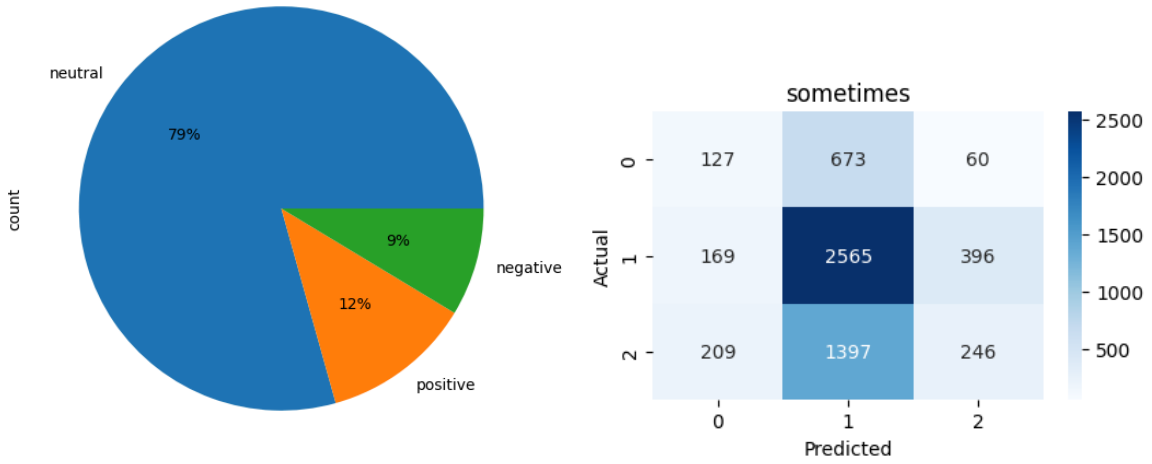
Slika 18. Matrica kofuzije i raspodela nakon analize sa Contextual word embedding za augmentaciju rečenica

Sequential pipeline



Slika 19. Matrica kofuzije i raspodela nakon analize sa Sequiential pipeline

Sometimes pipeline



Slika 20. Matrica kofuzije i raspodela nakon analize sa Sometimes pipeline

Konačan rezultat

	Augmentation Type	Accuracy
13	SubstituteChar RandomWord InsertChar	0.519514
14	DeleteChar InsertChar RandomWord	0.508901
15	DeleteChar SubstituteChar SubstituteWord	0.507874
12	WordSynonym WordRandom CharSubstitute	0.499144
2	substitute_char	0.486648
3	swap_char	0.484423
1	insert_char	0.484081
4	delete_char	0.479288
8	random_deletion	0.472270
11	gpt	0.470000
0	original	0.464225
7	random_swap	0.463198
5	synonym_replacement	0.462170
6	random_insertion	0.458233
16	DeleteChar SubstituteChar Bert	0.457449
10	roberta	0.444000
9	bert	0.442000

Slika 21. Konačan rezultat

Uočava se da su tehnike koje kombinuju različite metode augmentacije, poput "sequential" i "sometimes," dale najbolje rezultate, sa tačnošću od 50.31% i 50.29%, respektivno. Ovi rezultati sugerišu da kombinacija augmentacionih tehnika može obezbediti veću raznovrsnost podataka i doprinose boljoj generalizaciji modela.

Jednostavne transformacije na nivou karaktera, kao što su umetanje karaktera ("insert_char"), zamena karaktera ("substitute_char") i brisanje karaktera ("delete_char"), takođe su pokazale relativno visok učinak, sa tačnostima u opsegu od 47% do 49%. Ove tehnike čine minimalne izmene u rečenicama, zadržavajući semantički smisao, što verovatno omogućava modelu da uči na raznolikim primerima bez gubitka originalnog značenja.

S druge strane, tehnike koje koriste naprednije modele za generisanje teksta, kao što su BERT, RoBERTa i GPT, pokazale su nižu tačnost, posebno BERT i RoBERTa, sa tačnostima od 44.2%, dok je GPT imao nešto bolji rezultat od 47%. Ovi rezultati mogu ukazivati na to da složeniji modeli, iako generišu rečenice bogate semantičkim

varijacijama, ponekad odstupaju od osnovne strukture podataka, što može dovesti do smanjene performanse modela za analizu sentimenta.

Na osnovu ovih rezultata može se zaključiti da kombinacija jednostavnih augmentacija ili njihova primena u sekvenci daje najznačajniji doprinos poboljšanju modela za analizu sentimenta.

Zaključak

Data augmentacija nad tekstom predstavlja jedan od ključnih pristupa za prevazilaženje ograničenja u obuci NLP modela, posebno u situacijama sa nedovoljno podataka. Analizom različitih metoda augmentacije pokazano je da ova tehnika ne samo da povećava količinu podataka već i poboljšava robustnost modela, omogućavajući im da bolje generalizuju i rešavaju kompleksne zadatke.

Metode augmentacije u feature space-u, poput dodavanja šuma i interpolacije, omogućavaju generisanje novih podataka uz zadržavanje ključnih semantičkih informacija. S druge strane, augmentacija u prostoru podataka, na nivou reči, rečenica i dokumenata, omogućava modelima da bolje razumeju jezičke nijanse i varijacije. Napredni pristupi, poput generativnih modela i adversarijalnih tehnika, dodatno proširuju mogućnosti augmentacije, pružajući inovativna rešenja za izazove u obradi teksta.

Kroz praktične primenu u analizi sentimenta, prikazano je kako različiti metodi augmentacije podataka utiču na analizu sentimenta. Izvršena je analiza dobijenih rezultata i prikazano koji metodi i koliko poboljšavaju ili pogoršavaju analizu.

Literatura

[1] M. Bayer, M. A. Kaufhold, and C. Reuter, "A Survey on Data Augmentation for Text Classification," 2021. [PDF]

[2] https://www.kdnuggets.com/2019/04/text-preprocessing-nlp-machine-learning.html

[3] A. Kulkarni and A. Shivananda, *Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python*, PUB, 2019.

[4] "Cleaning and Preprocessing Text Data in Pandas for NLP Tasks," KDnuggets, 2020. [Online]. Available: https://www.kdnuggets.com/cleaning-and-preprocessing-text-data-in-pandas-for-nlp-tasks.

[5] A. B. Author et al., "Title of the article," *Journal Name*, vol. 10, no. 2, pp. 45-67, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622008523#b32.

[6] "Text Data Augmentation in Natural Language Processing with TextAttack," Analytics Vidhya, Feb. 2022. [Online]. Available: https://www.analyticsvidhya.com/blog/2022/02/text-data-augmentation-in-natural-language-processing-with-texattack/.

[7] "Title of the article," *Journal Name*, vol. 5, no. 1, pp. 100-110, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666651022000080#sec2.

[8] Contextual augmentation: Data augmentation by words with paradigmatic relations. arXiv preprint arXiv:1805. 06201. Wei, J.W. , & Zou, K.

[9] Jason Weston, "Understanding Data Augmentation For Text-5 Techniques With Examples," Jason Weston Blog, URL: https://jasonweston. com/understanding-data-augmentation-for-text-5-techniques-with-examples/2.

[10] Ahmadi, Sina, Daneshfar, Fatemeh, Hameed, Razhan, "Transfer Learning for Low-Resource Sentiment Analysis", 2023, http://arxiv.org/abs/2304.04703 (accessed: 17 Oct, 2024).

[11] (2019). EDA: Easy data augmentation techniques for boo.