**Sveučilište u Rijeci  
Fakultet Informatike i Digitalnih Tehnologija**

**A picture containing text, antenna

Description automatically generated**

**Projektni zadatak iz kolegija**

**Metode obrade prirodnog jezika**

**Text Summarizer**

**Autori:** Ivan Vrsalović,  
**Mentor:** prof. dr. sc. Sanda Martinčić-Ipšić

Rijeka, 9.01.2024

Sadržaj

[Sadržaj 2](#_Toc155864534)

[Uvod 3](#_Toc155864535)

[Opis Skupa Podataka 4](#_Toc155864536)

[Predprocesiranje podataka 5](#_Toc155864537)

[Embedding 7](#_Toc155864538)

[Arhitektura modela 9](#_Toc155864539)

[Korištenje modela 10](#_Toc155864540)

[Procesi pred generiranje sažetka 11](#_Toc155864541)

[Generiranje sažetka 13](#_Toc155864542)

[Evaluacija: Rouge score 14](#_Toc155864543)

[T5 model 15](#_Toc155864544)

[Opis modela 15](#_Toc155864545)

[Učitavanje T5 modela 15](#_Toc155864546)

[Treniranje T5 modela 15](#_Toc155864547)

[Generiranje sažetka pomoću T5 modela 16](#_Toc155864548)

[Usporedba modela 17](#_Toc155864549)

[Zaključak 18](#_Toc155864550)

# Uvod

U ovom projektu fokusiramo se na implementaciju Text Summarizer-a (sažimatelja teksta), alata koji ima sposobnost sažimanja velikih količina teksta na suštinske informacije. Koristimo Kaggle-ov CNN Daily News Dataset, bogat izvor vijesti prikupljen s CNN-a, kako bismo izgradili model za automatsko stvaranje sažetaka tekstova.

U sklopu našeg projekta Text Summarizer-a, implementirali smo složenu arhitekturu s enkoder-dekoder modelom, prilagođavajući ga specifičnim zahtjevima sažimanja teksta. Uz vlastiti model, korišten je i T5 (Text-to-Text Transfer Transformer) pretrenirani model. Kako bi usporedili rezultate i kvalitetu.

# Opis Skupa Podataka

Kaggle-ov CNN DailyMail News Dataset, predstavlja bogat izvor vijesti prikupljen s CNN-a. Ovaj skup podataka sadrži širok spektar vijesti iz različitih područja, uključujući politiku, sport, zabavu, ekonomiju i druge relevantne teme.

Skup podataka sastoji se od .csv datoteka „train.csv“ i „test.csv“ koji ćemo iskoristi za treniranje i testiranje. Svaka datoteka je u obliku 3. stupca „ID“, „Article“ i „Highlights“. Svaki članak je zapravo jedan red u datoteci u kojem je prvi stupac jedinstvena oznaka tj. ID, Drugi stupac sam tekst cijelog članka, te 3. stupac koji je već predodređeni sažetak članka iz drugog stupca.

Primjer:

|  |  |  |
| --- | --- | --- |
| ***ID*** | ***Article*** | ***Highlights*** |
| 92c514c913c0 | Ever noticed how plane seats appear to be getting smaller and smaller? With increasing numbers of people taking to the skies… | Experts question if packed out planes are putting passengers at risk .  U.S consumer advisory group says minimum space must be stipulated . |
| a248f9cd536d727f27a45a | "A drunk teenage boy had to be rescued by security after jumping into a lions' enclosure at a zoo in western India…. | Drunk teenage boy climbed into lion enclosure at zoo in west India .  Rahul Kumar, 17, ran towards animals shouting 'Today I kill a lion!'  Fortunately he fell into a moat before reaching lions and was rescued ." |

Skup podataka sadrži značajan broj članaka, točnije train.csv sadrži 287113 članka, dok test.csv sadrći 11490 članaka, što osigurava dovoljnu količinu podataka za treniranje i evaluaciju modela.

# Predprocesiranje podataka

Učitavamo podatke iz CSV datoteka za trening i testiranje te reduciramo veličinu skupa podataka na četvrtinu originalne veličine zbog ograničenja hardverskih resursa. Koristimo biblioteku *pandas* za čitanje CSV datoteka i manipulaciju podacima.

Zatim definiramo rječnik kontrakcija koji će nam omogućiti da engleske skraćenice poput „can't“ raspišemo u „can not“ i slično. Funkciju clean\_text za čišćenje teksta, uključujući zamjenu kontrakcija, uklanjanje URL-ova, HTML tagova, specijalnih znakova i stop riječi. Rječnik stop riječi se preuzet iz biblioteke nltk, te uključuje sve riječi bez većeg značaja (The, he, she, a, an, … )



Kod 1 Čišćenje teksta



Kod 2 WordCloud najčešćih riječi u sažetcima

Slika 1 Najčešće riječi u skupu

Tokenizacija i Prikaz Vokabulara

Koristimo *train\_test\_split* iz scikit-learn biblioteke kako bismo podijelili podatke na trening i test skupove. Nakon podjele tokeniziramo podatke pomoću „*Tokenizer-a“* iz iz Keras biblioteke kako bismo izvršili tokenizaciju teksta i sažetka. Tokenizacija je postupak pretvaranja sekvenci riječi u sekvence brojeva (indeksa). Tokenizer automatski generira rječnik svih jedinstvenih riječi (tokena) iz teksta i dodjeljuje svakoj riječi odgovarajući indeks. Kako bi smo bolje razumijeli distribuciju riječi u našem korpusu, brojimo rijetke riječi (ispod 4 ponavljanja), te brojimo ukupan broj pojavljivanja svake riječi.



Kod 3 Opis train skupa

Isti proces ponavljamo za train i test skup riječi



Kod 4 Opis test skupa

Kako bi osigurali da su ulazni podaci pogodi za model, moramo ih tokenizrati i pretvoriti u sekvencu, a svaku sekvencu moramo osigurati da su iste duljine pa ujedno primjenjujemo *pad\_sequences funkciju* koja dodaje nule na kraj sekvenciji kako bi sve bile jednake duljine koju smo unaprijed zadali. Isti postupak moramo primijeniti na x i y podatke.



Kod 5: Tokenizacije i pretvorba sekvenci

# Embedding

Embedding je tehnika u strojnom učenju koja se koristi za pretvaranje riječi ili entiteta u vektorski prostor, gdje slične riječi imaju slične vektore. Ovaj postupak omogućuje modelima da nauče reprezentacije riječi temeljene na njihovom kontekstu u skupu podataka. Embeddingi su ključni za prijenos značenja riječi u oblik pogodan za analizu i obradu od strane algoritama strojnog učenja.

U ovom projektu korišten je glove.6B.100d.txt datoteka koja sadrži pre-trenirane vektorske reprezentacije riječi (embeddinge) dobivene pomoću GloVe (Global Vectors for Word Representation) modela na velikom korpusu teksta. Konkretno, "100d" u nazivu označava da svaki vektor ima 100 dimenzija. Ovaj model omogućuje da se riječi predstave kao gusto raspoređeni vektori u višedimenzionalnom prostoru.



Kod 6 Učitavanje embedding datoteke

Sada možemo stvoriti matricu t\_embed, koja će se koristiti kao embedding matrica za tekstualne podatke. Prvo se stvara prazna matrica dimenzija (t\_max\_features, embed\_dim), gdje t\_max\_features predstavlja maksimalan broj riječi koje će se uzeti u obzir, a embed\_dim dimenzionalnost embeddinga.

t\_tokenizer.word\_index, koji mapira riječi na njihove indekse, te se za svaku riječ provjerava postoji li njezin vektor u prethodno stvorenim embeding\_index. Ako riječ ima vektor i njezin indeks je manji od t\_max\_features, tada se vektor dodjeljuje odgovarajućem retku u matrici t\_embed. Ovaj postupak rezultira embedding matricom koja će se koristiti kao ulaz za embedding sloj modela tokom treniranja.



Kod 7 Stvaranje embedded matrice

# Arhitektura modela

Model koristi arhitekturu s dvosmjernim LSTM enkoderom i dekoderom za izradu sustava sažimanja teksta. Enkoder prima ulazne sekvence teksta i koristi embedding kako bi predstavio riječi u skrivenom (skup skrivenih značajki koje mreža nauči iz ulaznih podataka) prostoru. Dvosmjerni LSTM enkoder generira skrivene stanice (enc\_h i enc\_c), koje se zatim prosljeđuju dekoderu.

Dekoder prima oznake sekvenci koje predstavljaju sažetak, također koristeći embedding matricu. LSTM dekoder generira sekvence izlaza i koristi TimeDistributed sloj s mekom aktivacijom kako bi predvidio vjerojatnosti riječi u sažetku. Model je treniran tako da minimizira razliku između stvarnih i predviđenih vjerojatnosti riječi.

Ova arhitektura omogućuje modelu naučiti sažimati tekst izračunavanjem značajki u latentnom prostoru pomoću LSTM enkodera, a zatim dekodiranjem tih značajki kako bi generirao sažetak.



Kod 8 Definiranje modela

Slika na kojoj se prikazuje tekst, snimka zaslona, račun, Font

Opis je automatski generiran

Slika 1 Arhitektura modela

Sada imamo pripremljene podatke i spremni smo za treniranje modela.



Kod 9 Treniranje modela

# Korištenje modela

Kako bi primijenili naš trenirani model na stvaranju sažetaka iz tekstova, koristimo Model klasu u Kerasu kako bismo definirali dvije odvojene instance za enkoder i dekoder.

Prva naredba stvara enkoder model (enc\_model) koji ima istu arhitekturu kao enkoder unutar treniranog modela, s tim da uzima samo ulazni tekst i generira izlazne stanice enkodera (enc\_h i enc\_c). Ovo će biti korišteno kao početno stanje dekodera tijekom generiranja sažetka.

Sljedeće definiramo ulazna stanja dekodera (dec\_init\_state\_h i dec\_init\_state\_c). Ova stanja će biti inicijalna stanja dekodera kada počnemo generirati sažetak. Zatim slijedi dio dekodera. Prva naredba dekodira ugrađeni tekst koristeći početna stanja definirana prije. Nakon toga, izlazni tenzor dekodera (dec\_out) se proslijeđuje kroz gusti sloj (dec\_dense) kako bi se dobio konačni sažetak.



Kod 10 Definicija enkodera i

I zatim spremimo dobiveni enkoder, dekoder i tokenizer.



Kod 11 Spremanje modela i tokenizera

## Procesi pred generiranje sažetka

Kako bi generirali sažetak od unesenog teksta, prvo vršimo predprocesiranje teksta, koristeći iste korake kao i prilikom treniranja modela na skupu podataka. Koristimo prethodno trenirani tokenizer (s\_tokenizer) i modele za enkoder (enc\_model) i dekoder (dec\_model), koji su prethodno spremljeni. Zatim, unosni tekst pretvaramo u sekvencu tokena koristeći tokenizer, nakon čega ga nadopunjujemo kako bi imao odgovarajuću duljinu.

Nakon pripreme ulazne sekvence, koristimo enkoder model da bismo dobili početno stanje dekodera. Zatim, inicijaliziramo praznu sekvencu (next\_token) s početnim tokenom za dekoder. U petlji, koristimo dekoder model da generiramo nove tokene dok ne dođemo do kraja sažetka ili dosegnemo maksimalni broj koraka. Svaki novi token dodajemo u izlaznu sekvencu, a dobiveni token koristimo kao ulaz za sljedeći korak u petlji. Postupak se ponavlja dok ne generiramo cijeli sažetak ili dostignemo maksimalni broj koraka u petlji. Na kraju, vraćamo generirani sažetak.



Kod 12 Generiranje sažetka

S time imamo sve potrebne elemente za generiranje sažetka.

## Generiranje sažetka

Tekst članka:  
Reports of Cristiano Ronaldo's decline have been greatly exaggerated: The Real Madrid forward scored five goals against Granada, taking his tally in La Liga to 36 for the season - more than half of Europe's top teams. Granada were torn to shreds by Ronaldo and his Real team-mates as the European champions won 9-1 at the Bernabeu, and their Portuguese talisman managed the first five-goal haul of his career. His record of 36 league goals for the season overtook Lionel Messi's 32 strikes, and put him in the incredible position of having scored more goals than 53 of the 98 teams in Europe's top five leagues this season. Cristiano Ronaldo scored an incredible five goals against Granada to lift his tally for the season to 36 . Granada were torn to shreds by Ronaldo and his Real team-mates as the European champions won 9-1 . The Portuguese star now has more goals than 53 teams in Europe's top leagues, including nine in England . There are nine teams in the Barclays Premier League who have managed fewer goals in the league than Cristiano Ronaldo has this season. They are: . Stoke City, Newcastle United, West Bromwich Albion, Hull City, Aston Villa, Sunderland, Queens Park Rangers, Burnley and Leicester City. Ronaldo took some flak last month after a series of poor performances for Real, but his goalscoring has never waivered and even through a difficult period he managed six goals in 10 games, the sort of record most strikers can only dream of. And his displays across the season have brought more goals than the likes of Newcastle United, Malaga, Borussia Dortmund, Torino and Lille. The first three of his five goals against Granada marked the quickest hat-trick of his career, coming in just eight first half minutes. Another followed 16 minutes later, and his last came 35 minutes after that. Ronaldo's displays across the season have brought more goals than the likes of Borussia Dortmund and Lille . It all adds up to a brilliant Easter Sunday for the world's best player, and he spoke of his delight afterwards . Lionel Messi and Barcelona came out on top in El Clasico but he has fewer league goals than Ronaldo . Borussia Dortmund, who can boast Marco Reus among their ranks, have fewer goals than the Real talisman . While Ronaldo overtook Messi on the goalscoring charts, the hat-trick - which could have been doubled - levelled the Argentine's record of 24 La Liga trebles. He almost matched Real legend Alfredo Di Stefano's 28 hat-tricks for Los Blancos in what was their biggest league win in 48-and-a-half years. It all adds up to a brilliant Easter Sunday for the world's best player, and he posted on Twitter after the game, saying: 'Happy to have scored 5 goals with this excellent team work. Thank you for all your support.' If the critics still see Ronaldo as greedy, it certainly paid off at the Santiago Bernabeu on Sunday.

Sažetak:

cristiano ronaldo has been a 1 1 1 in the premier league cup cup in the premier league league will be a season of the premier league cup will be a season of the premier league cup in the premier league will be a season of the premier league

Članak:   
Liverpool target Neto is also wanted by PSG and clubs in Spain as Brendan Rodgers faces stiff competition to land the Fiorentina goalkeeper, according to the Brazilian's agent Stefano Castagna. The Reds were linked with a move for the 25-year-old, whose contract expires in June, earlier in the season when Simon Mignolet was dropped from the side. A January move for Neto never materialised but the former Atletico Paranaense keeper looks certain to leave the Florence-based club in the summer. Neto rushes from his goal as Juan Iturbe bears down on him during Fiorentina's clash with Roma in March . Neto is wanted by a number of top European clubs including Liverpool and PSG, according to his agent . It had been reported that Neto had a verbal agreement to join Serie A champions Juventus at the end of the season but his agent has revealed no decision about his future has been made yet. And Castagna claims Neto will have his pick of top European clubs when the transfer window re-opens in the summer, including Brendan Rodgers' side. 'There are many European clubs interested in Neto, such as for example Liverpool and Paris Saint-Germain,' Stefano Castagna is quoted as saying by Gazzetta TV. Firoentina goalkeeper Neto saves at the feet of Tottenham midfielder Nacer Chadli in the Europa League . 'In Spain too there are clubs at the very top level who are tracking him. Real Madrid? We'll see. 'We have not made a definitive decision, but in any case he will not accept another loan move elsewhere.' Neto, who represented Brazil at the London 2012 Olympics but has not featured for the senior side, was warned against joining a club as a No 2 by national coach Dunga. Neto joined Fiorentina from Atletico Paranaense in 2011 and established himself as No1 in the last two seasons.

Stvarni sažetak:

Fiorentina goalkeeper Neto has been linked with Liverpool and Arsenal .

Neto joined Firoentina from Brazilian outfit Atletico Paranaense in 2011 .

He is also wanted by PSG and Spanish clubs, according to his agent .

CLICK HERE for the latest Liverpool news .

Generirani:  
manchester united have been linked with a move to the club s first season in the premier league club has been linked with a move to the club s side

## Evaluacija: Rouge score

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrika** | **Recall** | **Precision** | **F1 score** |
| **Rouge-1** | 0.075 | 0.286 | 0.115 |
| **Rouge-2** | 0.011 | 0.037 | 0.016 |
| **Rouge-l** | 0.071 | 0.273 | 0.109 |

## 

# T5 model

## Opis modela

T5, što znači "Text-to-Text Transfer Transformer", je model transformer arhitekture razvijen od strane Googlea. Ova arhitektura je posebna po tome što se može koristiti za različite jezične zadatke formulirane kao zadaci pretvorbe teksta u tekst. Umjesto da bude dizajniran za specifičan zadani zadatak, T5 se trenira za rješavanje raznolikih jezičnih problema, uključujući strojno prevođenje, generiranje teksta, odgovaranje na pitanja, i mnoge druge.

## Učitavanje T5 modela

Svi koraci prije stvaranja modela su isti, bez obzira radimo li s vlastitim modelom ili koristimo već gotov model poput T5. Razlika leži u tome što, kada koristimo T5, jednostavno ga učitavamo pomoću odgovarajuće naredbe, čime dobivamo već spreman model koji se može koristiti za različite jezične zadatke. Ova praktičnost omogućuje bržu i jednostavniju obradu prirodnog jezika.

Kako bi učitali model moramo koristiti naredbu iz biblioteke tensorflow kako za učitavanje modela. U ovom projektu korišten je model T5-small, koji je manja verzija T5 modela, tj. manje hardverski zahtjevan.



Kod 13 Učitavanje T5 modela

## Treniranje T5 modela

Zatim moramo obraditi naš skup podataka, kako bi ga mogli dati modelu na treniranje u obliku koji je čitljiv.



Kod 14 Priprema podataka za modeliranje sekvenci

Sada smo spremni za treniranje modela



Kod 15 Treniranje T5 modela

## Generiranje sažetka pomoću T5 modela

Generiranje sažetka radimo pomoću transformers biblioteke za izradu pipeline-a za sumarizaciju teksta pomoću T5 modela. Ulazni tekst se dohvaća iz skupa podataka, a zatim se koristi cjevovod za generiranje sažetka.



Kod 16 Generiranje sažetka pomoću T5 modela

Članak:

BREMEN, Germany -- Carlos Alberto, who scored in FC Porto\'s Champions League final victory against Monaco in 2004, has joined Bundesliga club Werder Bremen for a club record fee of 7.8 million euros ($10.7 million). Carlos Alberto enjoyed success at FC Porto under Jose Mourinho. "I\'m here to win titles with Werder," the 22-year-old said after his first training session with his new club. "I like Bremen and would only have wanted to come here." Carlos Alberto started his career with Fluminense, and helped them to lift the Campeonato Carioca in 2002. In January 2004 he moved on to FC Porto, who were coached by José Mourinho, and the club won the Portuguese title as well as the Champions League. Early in 2005, he moved to Corinthians, where he impressed as they won the Brasileirão,but in 2006 Corinthians had a poor season and Carlos Alberto found himself at odds with manager, Emerson Leão. Their poor relationship came to a climax at a Copa Sul-Americana game against Club Atlético Lanús, and Carlos Alberto declared that he would not play for Corinthians again while Leão remained as manager. Since January this year he has been on loan with his first club Fluminense. Bundesliga champions VfB Stuttgart said on Sunday that they would sign a loan agreement with Real Zaragoza on Monday for Ewerthon, the third top Brazilian player to join the German league in three days. A VfB spokesman said Ewerthon, who played in the Bundesliga for Borussia Dortmund from 2001 to 2005, was expected to join the club for their pre-season training in Austria on Monday. On Friday, Ailton returned to Germany where he was the league\'s top scorer in 2004, signing a one-year deal with Duisburg on a transfer from Red Star Belgrade. E-mail to a friend .

Sažetak:

Carlos Alberto joins Werder Bremen for a club record fee $10 million . 22-year-old scored in FC Porto's champions league final win against Monaco 2004 . he started career with fluminense and helped lift campeonato .

# Usporedba modela

Skraćeni index članak:

Pobjednik Života na vagi je Mislav Šepić

Index Magazin

22:26, 04. siječnja 2024.

Pobjednik Života na vagi je Mislav Šepić

Foto: Press

U STUDIJU zagrebačkog Jadran filma, pod voditeljskom palicom Marijane Batinić, natjecatelji sedme sezone Života na vagi pokazali su u punom sjaju da su se isplatili njihovi trud i zalaganje u proteklih pet mjeseci. Prvo su svoje rezultate pokazali nefinalisti koji su se borili za nagradu od 7500 eura, a potom su na scenu stupiti finalisti u borbi za 20.000 eura - Nina, Silvija, Mislav Šepić, Goran i tajni finalist Mislav Vlašić.

....

Prvi je na vagu stao Mislav Vlašić. Kad je došao u Život na vagi, Vlašić je imao 213.5 kilograma, izgubio je 88.9 kilograma, što je rekord emisije! „Nikad nismo imali natjecatelja s više od 200 kilograma koji je napravio ovakav rezultat!“ rekao je ponosno Edo za Mislava Vlašića koji je izgubio 41.6% tjelesne mase.

Na petome mjestu završio je Mislav Vlašić, četvrta je njegova djevojka Silvija, fantastičan rezultat ostvarila je Nina, drugi je bio Goran, a pobjednik sedme sezone Života na vagi je Mislav Šepić.

**Sažetak T5 modela:**

pobjednik ivota na vagi je Mislav epi +10 Foto: Press U STUDIJU zagrebakog Jadran filma . pokazali su u punom sjaju da su se isplatili zalaganje u proteklih pet mjeseci .

**Sažetak T5 modela sa prijevodom članka na engleski**

mislav epi weighed 173.9 kilograms five months . he lost 46.9% of her body mass . "it was really worth it" says one contestant.

**Sažetak modela**

the world cup is the world cup of the world cup of the world cup is the world cup of the world cup is the world cup of the world cup

**Sažetak modela sa prijevodom članka na engleski**

the of the is a new york city council department the new york city council is a new york city council is a new york city of the new york city council is a new york city of the new york city

# Zaključak

Izgradnja vlastitog „text summarizer-a“ predstavlja zahtjevan zadatak koji zahtijeva duboko razumijevanje jezičnih struktura, tehnologije procesiranja prirodnog jezika (NLP) i kompleksnosti zadatka sažimanja teksta. U ovom projektu, uspoređivao sam performanse vlastitog modela s T5-small (Text-To-Text Transfer Transformer) modelom, koji je jedan od naprednih modela u području NLP-a.

Prvi izazov je u prikupljanju i obradi obimnog i raznovrsnog skupa podataka za obuku modela. Postizanje visokih performansi zahtijeva rad s velikim skupom podataka, često u rasponu tolikom da vlastito računalo ne može podnijeti obradu takvih podataka. Ovaj proces zahtijeva velike računalne i prostore za pohranu, što može biti izazovno i financijski zahtjevno.

Drugi izazov leži u podešavanju modela kako bi se postigla optimalna performansa. Ovaj proces zahtijeva duboko razumijevanje arhitekture modela, kao i eksperimentalni pristup kako bi se pronašle najbolje vrijednosti parametara.

U usporedbi s vlastitim modelima, veći modeli koje su već napravljeni, poput T5 često pružaju bolje rezultate. Ovi napredni modeli sposobni su za shvaćanje konteksta, semantičkih veza i složenosti jezika na način koji manji modeli će teško postići.

U zaključku, izgradnja vlastitog „text summarizer-a“ je kompleksan proces koji zahtijeva napredno razumijevanje NLP-a i pristup resursima. Upotreba velikih modela poput T5 može značajno poboljšati performanse summarizacije teksta, ali istovremeno ubrzati razvoj i kvalitetu. Za mnoge projekte, korištenje već gotovih velikih modela može biti bolja opcija zbog njihove veće efikasnosti i dostupnosti.