Obrada prirodnog jezika - NLP

Emilija Krstonošić

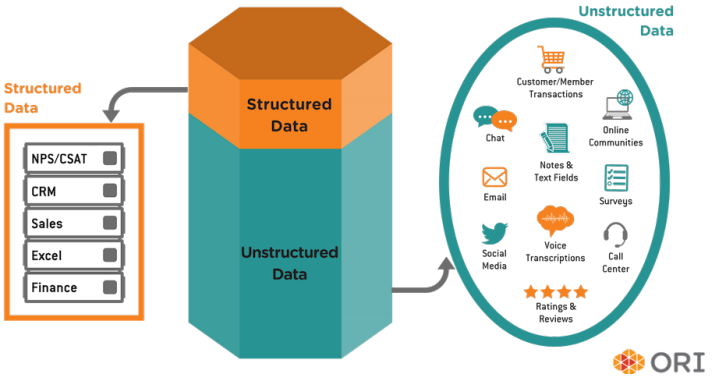
April 2020

1 Podaci na prirodnom jeziku

Jezik je veoma važan deo naše prirode i jedan je od temeljnih obeležja ljudskog bića. Konstantno smo okruženi jezikom; bilo u verbalnom ili pisanom obliku, sve što svakodnevno ispoljavamo pomoću njega nosi ogromne količine različitih podataka. To je način na koji mi interagujemo sa okolinom, iskazujemo mišljenje, stavove, ideje, pa čak i emocije. Prirodan jezik nam služi kao odličan alat pomoću kojeg interpretiramo i prenosimo informacije između sebe ili kroz neki vid medija. On nije savršeno definisan, ima puno nedostataka, diverznosti i izuzetaka, kao direktna posledica toga što je nastajao spontano i evoluirao vremenom pod različitim uticajima. Na svetu ih imamo veliki broj gotovo kompletno različitih, a ukupno preko 6500. Dok neki dijalekti imaju sintaksu bližu nekoj normi, ostali su podložniji promenama forme, stilskim figurama i rečeničnim spletkama, čineći sam koncept sagledanja generalnih karakteristika prirodnog jezika veoma složenim i neegzaktnim.

Nasuprot njima postoje i veštački konstruisani jezici, kao što su mašinski, programski, znakovni i jezici posebne namene, kod kojih su pravila jasno definisana, a struktura uvek jednoznačno razumljiva. Kompjuteri mogu precizno analizirati podatke tog tipa jer se dobro struktuirani podaci uvek mogu validno predstaviti u numeričkom obliku koji je računaru razumljiv. To nije slučaj i sa prirodnim jezikom; najveći deo dostupnih izvora na prirodnom jeziku (i podataka generalno) nalazi se u nestruktuiranoj tekstualnoj formi koja je neuniformna i, samim tim, teška i nepogodna za manipulisanje i dalje procesiranje. Umesto tabularnog prikaza prisutnog u relacionim bazama[1] inicijalni podaci na prirodnom jeziku predstavljeni su u vidu stringova čiji sadržaj sagledamo kroz sledeća tri kriterijuma: **sintakse**, **semantike** i **pragmatike**.

Zahvaljujući rapidnom razvoju *big data* polja i nauke o podacima (eng.- *Data science*), prikupljanje i čišćenje sirovih podataka prenetih na prirodnom jeziku postalo je znatno lakše. Oni su prisutni gotovo svuda u različitoj formi - u pisanoj fizičkoj, vokalnoj u vidu audio / video zapisa i na internetu, gde nam je omogućen pristup ogromnim setovima podataka, ekstraktovanih iz već široko korišćenih online medija, društvenih mreža, portala itd.
Danas, raznim računarskim alatima i algoritmima imamo mogućnost da relativno lako pripremimo tekstualne korpuse za dalju analizu i obradu.



Slika 1. (Ne)struktuirani podaci

[1] Relaciona baza podataka - tip baze kod koje se organizacija podataka zasniva na skupovima odnosa članova između kojih se definišu određene veze.

2 Uvod

Obrada prirodnog jezika (eng.- Natural Language Processing / NLP) je grana istraživanja na raskršću računarskih nauka, veštačke inteligencije (ML, AI) i lingvistike koja se tiče razvoja aplikacija i servisa sposobnih da razumeju ljudski govor. To je disciplina koja povezuje *data science*[2] sa ljudskim jezikom i nalazi primenu u širokom spektru oblasti inženjerstva i industrije. U literaturi se često nailazi i na termine "računarska lingvistika"(eng.- *computational linguistics*) i text mining. Algoritmi obrade prirodnog jezika pomažu računaru da primi, razume, analizira i potencijalno interpretira ljudski govor. Razvojem NLP-a otvara se mogućnost za bolju, dostupniju komunikaciju između čoveka i mašine. Razumevanje kompleksnijih i neretko dvosmislenih jezičkih formi važan je aspekt AI-a koji još uvek nije sasvim usavršen zbog same specifičnosti ljudskog govora. Veliki izazov predstavlja prepoznavanje figurativnog govora, ironije, frazeologizama, višeznačnosti u izražavanju, interesa i emocija koji stoje iza neke izjave ili teksta.

Zadatak NLP-a deli se u tri kategorije - probleme prepoznavanja prirodnog jezika u govoru, probleme poimanja prirodnog jezika u datom tekstu i probleme reprodukovanja istog, u vidu smislenih odgovora. Kod ekstraktovanja jezika iz govora potrebno je uključiti i algoritme za obradu signala, detekciju glasa, analize spektograma itd. što je velik zadatak sam po sebi, stoga ću se za sada držati samo pisane forme. Tekst se analitički može posmatrati na dva nivoa - površno, kao sekvencu zasebnih entiteta, ili kroz kontekst, u nekom širem pogledu. Znatno je teže sagledati tekst u celosti i ulaziti u njegovu pozadinu kako bi se razaznalo mišljenje ili tema koju on prenosi, ali jedino ga je na taj način moguće razumeti u potpunosti. Što se tiče generisanja teksta i odgovora na zahteve date na prirodnom jeziku, mašina mora najpre suštinski da razume prirodu pitanja ili izjave koja joj je upućena, kao i šta se to od nje tražilo, a zatim da to uspešno interpretira i sažima u odgovarajuću reakciju. Da bi se postigao ubedljiv model veštačkog govornog agenta, mašina mora imati razumevanje i za kognitivne aspekte ljudskog mozga - stvari za koje je zainteresovan, poznavanje kulture i aktuelnosti koje spadaju u *world knowledge*, uključujući taksonomiju objekata i apstraktna razmišljanja u našem svetu, što sve predstavlja značajne izazove.

Najčešća primena NLP modela je realizacija *chatbotova*, govornih asistenata, sistema za prepoznavanje govora, mašinskog dinamičkog prevoda, automatskog ispravka teksta, prepoznavanje i izbacivanje nepoželjnih sadržaja itd.. Sve većim razvojem ovog polja postižu se značajni rezultati i u automatizaciji unutar oblasti zdravstva, medija, finansija i ljudskih resursa, između ostalog.

2.1 Istorijat i razvoj

Kompjuterska obrada prirodnog jezika započinje pedesetih godina prošlog veka, kada se i pitanje veštačke inteligencije po prvi put pominje u nauci. Prvi problemi kojima se NLP bavio bili su vezani za mašinski prevod rečenica sa nekog stranog jezika na engleski. Ovaj model, kao i većina drugih računarskih sistema u to vreme, zasnivao se na setu ručno napisanih generalnih pravila. Sistem je funkcionisao sa neznatnim brojem rečeničnih konstrukcija, no čim su naišle nešto kompleksnije, dvosmislene forme, mašina nije mogla da ispuni zahteve i uspešno ih prevede. Vrlo mali pomak je usledio u rešavanju problema prevoda teksta sa ovim pristupom sve do kasnih 1980-ih, pojavom prvih **statističkih modela**. Šezdesetih i sedamdesetih godina pojavljuje se više notabilnih sistema - mahom jednostavni četboti i osnovni govorni asistenti (npr. simulacije psihoterapeuta davanjem nekih od generičnih odgovora na prinete "žalbe"). Počevši od kasnih osamdesetih, ustupljena je revolucija u ovom polju uvođenjem algoritama mašinskog učenja u do tada poznate koncepte NLP-a. Usled velikog koraka ka povećanju računarske moći i razvoju moderne lingvistike (korpus lingvistika, bolje segmentisanje teorije i jasnije definisana pravila) ovaj pristup se odmah pokazao uspešnijim od dotadašnjeg rada sa manuelno unošenim pravilima. Neki od najranije korišćenih algoritama mašinskog učenja, kao što su **stabla odluke** (eng.: *decision trees*), predstavljali su sisteme velikog broja hardkodovanih *IF->THEN* orijentisanih uslova koji su se nadograđivali na već postojeća ručno napisana pravila. Neki finiji, detaljniji procesi i klasifikacije rečeničnih članova omogućene su upotrebom skrivenih Markovljevih modela[3](**HMM**).

Sve više, istraživanje se orjentisalo ka statističkim modelima odlučivanja zasnovanog na empirijskoj verovatnoći, povezivanjem znanja o već poznatim konceptima sa ulaznim podacima.

[2]Presek komunikacija, programiranja, matematike i statistike
[3]Statistički Markovljev model čiji je cilj da "uči dok posmatra"

Mnogi sistemi za prepoznavanje govora i dalje se oslanjaju na primere takvih modela jer su generalno pogodniji za rad sa datim nepoznatim ulazima koji, na primer, sadrže neku slučajnu grešku. Skoriji radovi fokusiraju se više na nenadgledane i polunadgledane algoritme učenja. Takvi algoritmi mogu da koriste sirove podatake za trening koji nisu nužno pribeleženi na željene izlaze, već uče i donose odluke iz iskustva (prethodnih iteracija). To je generalno teže postići u odnosu na nadgledane mreže, ali zato imamo znatno veći dostupan trenažni skup podataka (bilo koji sadržaj na internetu, između ostalog) i mnogo manje manuelnih postavki koje je potrebno prethodno napomenuti. U poslednjih deset godina, duboko i reprezentaciono učenje (eng.: *feature learnining*), koje analizira podatke transformišujući ih izvlačenjem njegovih obeležja, olakšalo je drastično izvršavanje zadataka nalik klasifikaciji i predikciji fičera u tekstu. Ovaj pristup postao je najšire korišćen zbog toga što ima najveće mogućnosti i obezbeđuje najpreciznije $state - of - the - art$ rezultate. Uz pomoć prediktivnog modelovanja, 2018. godine izbačen je i prvi rad (roman) "1 the Road", izgenerisan isključivo sa strane AI-a uz pomoć senzorike.

Danas, umesto proste interpretacije govora i teksta samo iz ključnih reči, moguće je sve bolje izvlačiti njegovo pravo značenje na kognitivan način, pa čak i detektovati neke od komplikovanijih formi govora poput ironije, slenga, žargona, itd..

2.1.1 Turingov test

Takozvani "Tjuringov test" način je testiranja inteligencije veštački konstruisanih sistema. Predlog je ogleda za procenu mašinske sposobnosti da demonstrira inteligentno ponašanje ekvivalentno, ili prosto nerazlučivo od ljudskog. Princip je opisao Alan Tjuring 1950. godine u svom radu na temu veštačke inteligencije, "*Računske mašine i inteligencija*", gde se ujedno po prvi put pojavljuje i pojam obrade prirodnog jezika kao zaseban predmet istraživanja. Naučnik je predložio sistem gde čovek analizira razgovor na prirodnom jeziku sa mašinom dizajniranom da generiše odgovore slične ljudskim (*chatbot*). Evaluator koji učestvuje u testu stupa u konverzaciju putem tekstualnih kanala sa mašinom na jednoj, i čovekom na drugoj strani. Ukoliko sudija na kraju eksperimenta ne može sa sigurnošću da odredi koji sagovornik je čovek a koji mašina, smatra se da je test uspešan, a mašina se procenjuje kompetentnom da "misli". Rezultat ne zavisi od sposobnosti mašine da daje tačne odgovore na sva postavljena pitanja, već mogućnosti veštačke inteligencije da se kroz lingvističke aspekte uverljivo predstavi kao čovek. Da bi se to izvršilo, potrebno je da mašina najpre kompletno razume govorni jezik ljudi, a zatim da ga reprodukuje i segmentuje u odgovarajuće smislene izlaze. To i dalje nije postignuto i nijedna mašina još nije uspešno prošla test, no razvojem alata i velikim pomacima u oblastima NLP-a postaje moguće usavršavanje ovih procesa i dolasku do što elegantnijih rešenja.

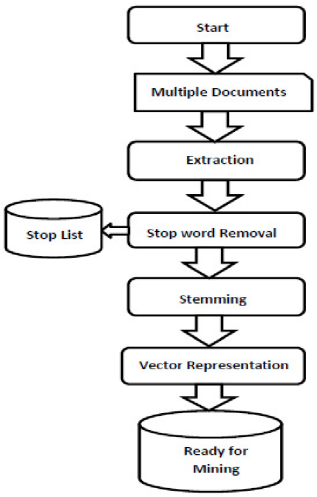
3 Predprocesiranje teksta

Najveći deo dostupnih tekstualnih podataka / fajlova nalazi se u obliku rečenica (dokumenata), članaka, tvitova, objava, komentara... Kako bi se ti podaci inicijalno transformisali i doveli do elementarnije forme, koriste se razni NLP alati za predprocesiranje teksta. Dobro očišćen tekst pospešuje performanse modela i tačnost izlaznih podataka. Neke od važnih operacija koje se koriste za početno normalizovanje teksta su [*]:

- Isključivanje svih brojeva, specijalnih karaktera, linkova, hash ili html tagova itd. i zamena sa **place holderima**,
- **Tokenizacija** - input se razbija na manje jedinice, izoluju se pojedinačni delovi rečenice, reči i znaci interpunkcije[4] kao zasebni objekti.
- **Stemovanje** - izdvojene reči vraćaju se u njihov osnovni oblik (nominativ jednine/infinitiv), uklanjaju se prefiksi/sufiksi... Svaka reč se posmatra individualno, bez šire povezanosti.
- **Lematizacija** - prebacuje reči u njihov rečnički oblik, grupišući korene sličnih.
- **Stop reči** - izbacuju se sve reči malog semantičkog značaja (poruka ostaje ista i bez njih). To su uglavnom zamenice, predlozi, prilozi, rečce...

[4] Uglavnom se isključuju, mada kod naprednijih analiza delova teksta mogu biti i korisni

Osim osnovnih operacija, izvršavaju se još neki procesi koji se tiču obrade sintaksičkih, semantičkih i pragmatičkih obeležja teksta. Na taj način još bolje sagledavamo njegovu dubinu u celosti.



Slika 2. Koraci predproces.

Pri analizi **sintakse** sagledaju se konstrukcije reči, gramatička ispravnost, način na koji su postavljene u rečenici i njihova pojedinačna funkcija. U to spadaju razni algoritmi parsiranja, obeležavanje rečeničnih članova (tzv. *Parts-of-speech / POS tagging*), građenje sintaksnih drveća zavisnosti, itd. POS tagging nam je posebno koristan jer određuje kojoj vrsti reči token pripada, omogućavajući nam da ih dalje grupišemo po određenom osnovu (proces koji nazivamo *chunking* ili *plitko parsiranje*). To uglavnom rešava čestu zabunu kod pojave homonima, sinonima itd.
Semantička analiza obuhvata dosta kompleksnije pristupe jer, osim značenja pojedinačnih delova rečenice, vodi računa i o tome kako se one međusobno povezuju unutar te sekvence. Način na koji definišemo *razumevanje* informacija je jedan od glavnih problema računarske lingvistike sam po sebi. Pri kvalitetnoj analizi semantike teksta susrećemo se sa raznim zadacima kao što su obeležavanje uloge rečeničnih članova (eng.: *Semantic Role Labelling*), ekstrakcija relacija između tokena, prepoznavanje opštepoznatih naziva kompanija, ljudi, proizvoda, lokacija itd. (eng.: *Named Entity Extraction / NER*) i *Word Sense Disambiguation* iliti razjašnjenje smisla reči posmatrajući dati kontekst.
Pragmatika analizira celi tekst kao jedno, vodeći računa o implikaturi unutar njega. Ona gleda na tekst globalno, kao sekvencu rečenica. Pragmatičkim pristupom možemo da razaznamo načine kazivanja česte u komunikaciji, kao što su narativi, reference, menjanje ili nadovezivanje na neku temu... neki od programa koji se tiču raslojavanja pragmatike u tekstu su segmentacija tematike, građenje leksičkih lanaca, sumiranje teksta, prepoznavanje koreferencije ili anafore itd...

* Neki od primera priloženi su u jupyter svesci *

3.1 Prikupljanje podataka

Suštinski, za obradu teksta može se koristiti bilo kakav izvor, od običnih tekstualnih fajlova do čitavih web sajtova. Zbog toliko dostupnih medija na internetu, primeri realne upotrebe prirodnog jezika se možda i najbolje mogu ukazati na raznim sajtovima, forumima i društvenim mrežama. Najlakši način da prikupimo data setove željenih sadržaja sa interneta je *web scraping*.
Web scraping (*harvesting*) je tehnika za ekstraktovanje velikih količina podatka sa sajtova na mreži pomoću koje ih možemo sačuvati u lokalnom fajlu na našem računaru ili u bazi podataka, uglavnom u formatu tabele. To nam značajno olakšava pripremu korpusa koji je odgovarajuć za naš projekat. Ovo je obično automatizovan proces koji izvršava *bot* (softverski program koji automatski izvršava određene zadatke) napisan u Pythonu (najčešće) ili Node-u, Ruby-u, C/C++-u...

4 Metod - automatski sistemi

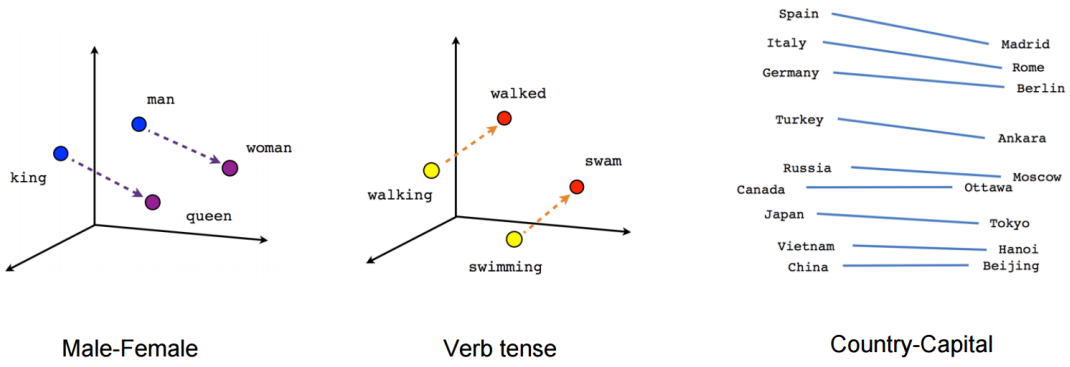
U današnje vreme gotovo svi sistemi u upotrebi teže nekom vidu automatizacije. Mašine su zadužene da obavljaju gomilu naših rutinskih poslova, a generalni koncepti uglavnom se mogu unaprediti tako da vremenom samostalno nalaze način da dođu do optimalnih rešenja određenih problema. Konstruisanje tog vida sposobnosti mašine da uči, analizira i donosi odluke, nazivamo **veštačka inteligencija** (*Artificial Intelligence*). Osim što nam ona u većem broju slučajeva obezbeđuje da neke zadatke odrađujemo brže i sa većom preciznošću, AI je sposoban da vodi računa o mnogo širem spektru uslovnih elemenata i parametara koji učestvuju u sistemu od čoveka, stoga ima veće pokriće nad konkretnim zadacima.
Prvi konceptualni pristupi ka obradi prirodnog jezika bili su orjentisani na rešavanju mnoštva pojedinačnih slučajeva. To su bile metode zasnovane na skupovima ručno sastavljenih pravila i baza znanja velike složenosti (eng.: ***rule-based NLP***).

Generalno govoreći, rule-based metode bile su poprilično tačne, mada veoma krhke. Na primerima koji nisu bili pokriveni postojećim, unetim pravilima, model je neizbežno imao veoma loše performanse. Godinama se nije mogao napraviti značajan pomak u ovim oblastima jer je izrada sveobuhvatnih baza znanja za određenu problematiku izuzetno naporan i dug posao i često zahteva dodatno ekspertno znanje. Osim toga, jezik je fonomen koji se zbog svoje prirode konstantno menja, stoga je neophodno iznova ažurirati realizovan sistem, što je, naravno, veoma nepraktično.

Najznačajniji progres u izradi NLP algoritama ukazao se tzv. statističkom revolucijom, kada se pojavio novi pristup u rešavanju repetetivnih zadataka, mašinsko učenje.
Mašinsko učenje (eng.: *machine learning*) je oblast veštačke inteligencije koja se bavi izgradnjom računarskih sistema koji uče iz iskustva. Najviše korišćeni ML modeli za NLP spadaju u kategoriju nadgledanih (*supervised*) metoda. To znači da su podaci korišćeni da se pohrani model prethodno anotirani i služe kao primeri za dalje uočavanje takvih obrazaca. Označavanje podataka (koji su najčešće inicijalno predprocesirani [*]) takođe podrazumeva i njihovo prilagođavanje računarskom sistemu. Prebacivanje teksta u numerički oblik može se odraditi na nekoliko konceptualno različitih načina. Jedan od češćih jednostavnijih modela da se to postigne jeste tzv. *bag-of-words* model.
Bag-of-words sistem za vizuelizaciju teksta pokazao je uglavnom uspešne performanse kod ekstrakcije obeležja za dalju klasifikaciju i prostu predikciju. Ne vodeći računa o gramatičkim karakteristikama, svakoj reči se dodeljuje broj koji predstavlja koliko puta se ona pojavila unutar nekog teksta. Na taj način možemo da adaptiramo korpus za tabelarnu strukturu, ali usput uskraćujemo i dosta preciznosti, iz razloga što se u "vreći reči" gube informacije o položaju pojedinih elemenata u odnosu na druge. Ovaj problem delimično se rešava *n-gramima* (najčešće bigrami) koji parsiraju tekst u logičke jedinice ili grupišu parove reči koji se često ukazuju susedno jedno od drugih. Tako je konstruisan tzv. *kontinualni* bag-of-words (**CBOW**) model sposoban za predikciju targetnih reči na osnovu njihove okoline tj. bliskog konteksta. No, ovakav pristup nije bio najpogodniji onda kada nam je cilj nešto detaljnija analiza teksta, naročito sa semantičkog aspekta.
Mnogo zahvalniji koncept od navedenog jeste upotreba **vektorske reprezentacije reči**. Vektori reči od velike su pomoći tehnologiji na putu ka razumevanju odnosa između rečeničnih članova i uočavanju njihovih semantičkih sličnosti. Računar uz pomoć AI-a uči da mapira njihove vrednosti u multidimenzionalni prostor, obrazujući sistem reprezentacije koji nazivamo ***word-embbedings***.

4.1 Word Embedding sistemi

Jedan od ključnih proboja za performanse nenadgledanih mreža i metoda dubokog učenja pri obradi prirodnog jezika bili su *word embedding* sistemi. Word embeddings su načini za reprezentaciju reči u numeričkom obliku koji gradi globalnu apstrakciju ljudskog govora za mašinu. Oni predstavljaju nizove realnih brojeva, gde se svaka vrednost unutar niza odnosi na dimenziju vektorskog prostora gde je reč mapirana. Posmatrajući reči na taj način, gradimo sistem koji konstruiše odnose izmedju reči iz istog konteksta, dodeljivajući im međusobno približne skalarne vrednosti vektora. Unutar "dimenzija" vektorskog prostora (zovemo ih i «fičeri»), razmaci između ucrtanih vrednosti (distance između tačaka) od velikog su značaja. Reči sličnih semantičkih karakteristika i značenja smeštaju se na bliske prostore (jer su im slični vektori), tako gradeći smislene relacije između sebe.



Slika 3. Primeri povezivanja vektora reči

Vektorski prostor potrebno je prvo istrenirati na početnom vokabularu (ili leksikonima) poznatih reči kako bi se inicijalizovao. Za to se uglavnom koriste plitke neuronske mreže, a dobra stvar jeste da se to može postići i kao usputni proizvod rešavanja problema klasifikacije na neanotiranom korpusu teksta (nije potreban dodatni napor za manuelnu anotaciju). Neki od dobrih primera već utreniranih implementacija za predstavljanje reči su $Word2Vec$ i $GloVe$ (global vectors).
Word2Vec jedan je od najpopularnijih tehnika učenja word embeddinga poslednjih godina, kreiran 2013. godine u *Google*-u. To je dvoslojna neuronska mreža koja procesuje ulazni tekstualni korpus obrazujući ga u set fičer vektora. Suštinski predstavlja skup **prediktivnih modela** koji uspevaju da povežu reči unutar datog konteksta i od njih naprave odgovarajuću numeričku reprezentaciju. Najzastupljeniji među njima su već pomenuti CBOW i tzv. *skip-grami*). Iako je W2V plitka mreža, forma teksta koju dobijemo na njenom izlazu dobro je razumljiva dubokim neuronskim mrežama što se odlično pokazalo u praksi.
Ove tehnike generalno su veoma zgodne i esencijalne za rešavanje većine trenutnih NLP problema, jer nam daju znatno više informacija o rečima, njihovom značenju i implikaturi od drugih, tradicionalnih reprezentacija kao što su *BoW*, *One-Hot-Representation*, *Term Frequency–Inverse Document Frequency* (TF-IDF) i slično..

4.2 Neuronske mreže

Nakon što podatke koje želimo da koristimo prevedemo u oblike s kojima računar može fino da radi, slede algoritmi za klasifikaciju i regresiju teksta. Neki od široko korišćenih NLP nadgledanih automatskih algoritama za to jesu metode potpornih vektora (*Support Vector Machines*), zbir nasumičnih drveća odluke (*random forrest metod*), princip maksimalne entropije[5], Bajesovi klasifikatori (npr. Naive Bayes), Bajesovske mreže i logistička regresija. Navedeni algoritmi daju sasvim razumne rezultate na velikom setu obeleženih podataka, ali mahom pokazuju manju efikasnost i gube preciznost kod specifičnijih zadataka baziranim na sirovom tekstu, izazivajući česte greške pri pojavi spletki i kompleksnijih formi. Rešenja koja su se vremenom pokazala superiornije od ovih tradicionalnih metoda mašinskog učenja tiču se neuronskih mreža pohranjenih vektorima reči (word embeddings-ima).

Neuronska mreža je oblik veštačke inteligencije, implementacija sistema koji se sastoji od izvesnog broja međusobno povezanih čvorova, procesorskih jedinica koje nazivamo veštačkim neuronima. Glavna prednost takvog sistema jeste paralelna obrada podataka, pristup znatno efikasniji od regularnih algoritama. Masovna paralelnost znači da istovremeno rade više procesorskih jedinica, a rezultati njihove obrade pri svakom ponavljanju prelaze na sledeće neurone. Svaki od neurona ima lokalnu memoriju u kojoj pamti podatke koje obradi i tako obezbeđuje da nove iteracije obrade deluju na osnovu prethodnog "iskustva". Tako se postiže samoodrživ sistem sposoban da prepozna i relativizuje opšte slučajeve i paterne ponašanja. Koncept po kome određena neuronska mreža uči i radi definišu obrasci po kojima je ona trenirana. Proces kojim se mreža trenira zove se *algoritam obučavanja*. Kroz ovu proceduru se na sistematičan način menjaju, tj. ažuriraju tzv. sinaptičke težine (težinski koeficijenti veza) u cilju dostizanja željenih performansi i skladnih izlaza. Generalizacija po tim obrascima znači sposobnost produkovanja zadovoljavajućeg izlaza neuronske mreže i za one ulaze koji nisu bili prisutni u toku obučavanja, što je direktno i opšti cilj bilo kog neuralnog algoritma. Arhitektura neuronske mreže predstavlja specifičnu povezanost njenih neurona u jednu celinu. Njena struktura segmentuje se u slojeve čiji broj može da varira. Prvi sloj je ulazni, poslednji izlazni, a slojevi između njih nazivaju se "skriveni" slojevi. Prvi ili ulazni sloj predstavlja tok kojim se izabrani podaci distribuiraju daljim slojevima, sledeći (skriveni) slojevi prosleđuju validno obrađene podatke do izlaznog sloja gde zapravo dobijamo konačan rezultat. Svi slojevi međusobno su potpuno povezani, a što je više skrivenih slojeva to je i sistem složeniji. Slojevi vrše komunikaciju tako što se izlaz svakog neurona povezuje sa ulazima svih neurona narednog sloja. Svaki čvor ima nekoliko ulaza i jedan izlaz.

Postoji mnogo vrsta neuronskih mreža podeljenih po različitostima u načinu rada. Različiti tipovi mreža ispunjavaju drugačije performanse pri rešavanju različitih zadataka; neke su adekvatnije i bolje konfigurisane za primeniti na određene probleme od drugih. Složenija neuronska mreža ne mora nužno da implicira da daje bolje ili preciznije rezultate od neke jednostavnije, već zavisi od toga kako je problem kome pristupamo definisan. Za obradu prirodnog jezika koriste se pretežno algoritmi ***dubokog učenja***, koji su u stanju da prisvoje diverzne nivoe apstrakcije.

[5]Stanje potpune neorganizovanosti (ja)

Neke od trenutno najboljih modela dubokih neuronskih mreža prikladnih za NLP su rekurentne, konvolutivne, rekurzivne i LSTM mreže.

4.2.1 Rekurentne neuronske mreže

RNN), dugo su bile najbolje rešenje za probleme klasifikacije, sumiranja, parafraziranja i generisanja (mašinski predvod npr.). Pristup sa ovim

efetkivne pri procesiranju sekvencijalnih informacija (seq2seq)

rekurzivno obrađuje svaku instancu ulazne sekvence uslovljeno prethodno obrađenim rezultatom

Sekvence su obično predstavljene fiksiranim nizom tokena koji pohranjuju rekurentnu jedinicu jedan po jedan (sekvencijalno)

Glavna prednost RNN-a je njen kapacitet da memoriše rezultate prethodnih iteracija obrade i onda iskoristi stečene informacije pri trentnom proračunu. To čini rekurentne modele sasvim prikladnim za za modelovanje okvirnih zavisnosti iz ulaza proizvoljne dužine kreirajući podesne kompozicije za njih.

Koriste se u proučavanju raznih zadataka obrade govornog jezika nalik mašinskom prevođenju, modelovanju jezika (izdvajanje notabilnih fičera), klasifikacije hijerarhije reči i rečenica i sl..

4.2.2 Konvolucione

4.2.3 LSTM

attention

4.2.4 Trasformeri

BERT, state of the art

5 Alati

NLP je poprilično aktuelna i popularna oblast računarstva, tako da postoji širok spektar dostupnih (open-source) alata za razvoj. Suštinski, mnogi programski jezici podržavaju pisanje koda za čišćenje tekstualnih podataka i obradu prirodnog jezika, kao na primer R, Python, Node, Java, C++, Scala itd., mada nisu svi jednako ažurni i zgodni za to. Kao i za većinu drugih aplikacija neuronskih mreža, mašinskog učenja i veštačke inteligencije, programski jezik Python se pokazao kao najjednostavniji i pogodniji za dalju implementaciju zbog širokog izbora korisnih alata i dostupnih biblioteka pisanih na istom jeziku. Neki od najpopularnijih Python biblioteka koje se trenutno koriste za razvoj NLP servisa su: (primeri pojedinih implementacija pokazani su u priloženoj *juPyter* svesci):

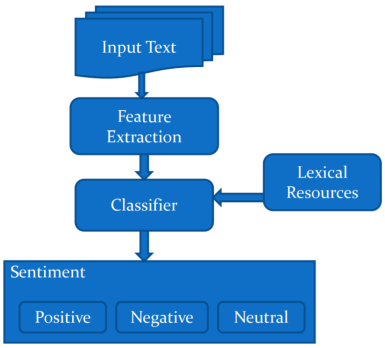
- **NLTK** - Natural Language Tool Kit; biblioteka nesporno najekstenzivnije opremljena svim osnovnim alatima potrebnim za brže i lakše izvršavanje nekih od najčešće korišćenih komponenta NLP-a. Gotovo svi procesi koji bi bili potrebni da se odrade kao što je klasifikacija, tagovanje, parsiranje, semantička rezonacija i predprocesorske stavke, mogu se izvršiti pomoću izvedenih funkcija unutar biblioteke. Svi podaci se, doduše, distribuiraju u formi stringa, što je zgodno za jednostavije konstrukcije, ali znatno otežava rad pri nekim naprednim funkcionalnostima koji zahtevaju parcijalni pristup. Iz tog razloga se algoritmi iz NLTK-a često kombinuju sa bibliotekama kao što je **pandas** radi lakše manipulacije nad podacima. [8]
- **SpaCy** - posmatra podatke kao objekte što ga čini bržim i praktičnijim u nekim slučajevima od NLTK-a. Ima širok raspon složenih naprednih alata za duboko učenje unutar svoje biblioteke i sadrži pre-trenirane vektore reči što olakšava implementaciju word embedding algoritama. [9]
- **PyTorch-NLP** - Koristi se najčešće za Deep learning pristup pri segmentaciji teksta (POS tagging i modelovanje jezika. [10]
- **Scikit-learn** - Biblioteka za mašinsko učenje, sadrži fine alate za klasifikaciju, regresiju, rad sa vektorima...
- **Gensim** - pogodan za topic modeling.
- **TextBlob** . . .

6 Primena

Obrada prirodnog jezika pripomaže u optimizovanju informacionih sistema, kako u svakodnevnici regularnih korisnika, tako i u raznim segmentima biznisa i manje više bilo kojoj organizaciji gde je komunikacija prisutna kao važan element. Neke od već opšteprisutnih primena NLP-a, dostupnih gotovo svim korisnicima informacionih tehnologija, su aplikacije poput **Google Translate**-a koji ima mogućnosti mašinskog prevoda teksta na preko 100 jezika, auto korekcija pogrešno napisanih reči iliti **Auto correct** prisutan na svim smart telefonima, **grammar check**-eri, automatska detekcija i navigacija pomoću govora, servisi za učenje novih jezika, plasiranje targetnih reklama, filterisanje teksta u odvojene kategorije, izvačenje ključnih reči pretrage itd. Osim toga moguće je realizovati korisne skripte za automatsko rezimiranje dokumenata, pojednostavljivanje tekstova, menadžment dijaloga, prikupljanje ciljanih informacija... Na najaktivnijim platformama kao što su google, ios, osx, microsoft, windows implementirani su interaktivni interfejsi, tzv. **virtuelni asistenti** koji nam služe za automatizaciju i brži pristup nekim sadržajima. Chatbotovi raznih namena sve više su u upotrebi kao ekstenzija nekih servisa, a takođe od velike su koristi i u nekim pogledima zdravstva. Nekoliko dobrih primera takvih klijenata; **Endurance** - pomaže pacijentima sa oblicima dementnosti da se prisete onoga što im je važno, servisi za pomoć pri dijagnozi na osnovu, informacija o korisnikovom stanju (simptomima), **Casper** - pomaže ljudima koji pate od insomnije da se smire kroz razgovor itd.. Aplikacije NLP-a značajno su uticale i na razvoj marketinške inteligencije. Algoritmi za ustanovljivanje sentimenata unutar recenzija, elektronske pošte, vesti itd. donose podosta relevantnih informacija u vezi sa zadovoljstvom klijenata, targetnom grupom kupaca, stanju prodaje i sl. čijom se analizom može doći do raspoznavanja odlučujućih faktora za napredak biznisa.

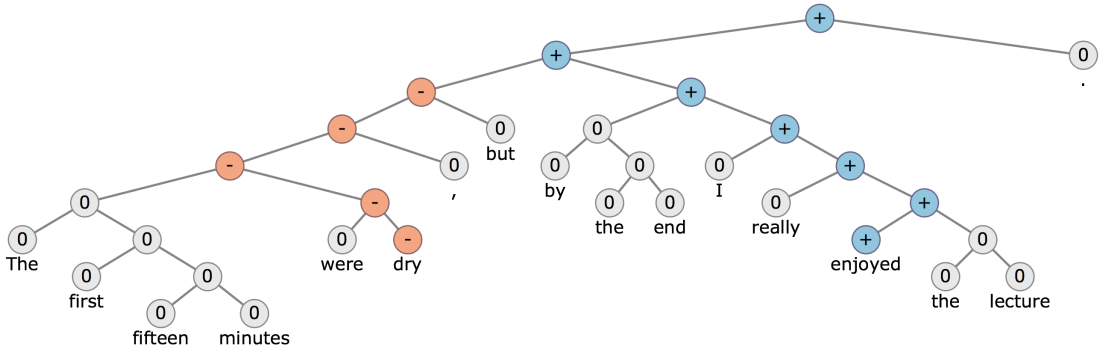
6.1 Analiza Sentimenta

Sentimentalna analiza je proces određivanja iskazne emocije koja stoji iza nekog teksta. Bavi se labeliranjem i *mapiranjem* podataka, odnosno povezivanjem delova teksta sa sentiment klasama kojoj pripadaju. Široko je korišćena kako bi se bolje procenili stavovi, mišljenja i osećanja iza neke izjave, najčešće recenzije proizvoda, servisa, brenda ili objave na društvenoj mreži. Ima više različitih pristupa i tipova sentimentalne analize. Postoje razni alati koji se mogu fokusirati na detekciju *finijih* osećanja (ljutnja, sreća, tuga...), isključivo na polarnost izjave ili na identifikaciju namere (zainteresovanost, indiferentnost, nezainteresovanost...) itd.



Slika 4. A.S.

Ovi algoritmi imaju veliku praktičnu primenu u poboljšanju korisničkih usluga, evaluaciji stava neke ciljne grupe ili karakteristike nekog teksta (npr. novnarskog članka). Kompanije koriste sentimentalnu analizu za automatsko sagledavanje komentara, rezultata anketa, preglede proizvoda itd. ne bi li dobili značajan uvid u kvalitet njihovih proizvoda ili servisa. Na osnovu statistike mišljenja kupaca, brend može korigovati slabije aspekte proizvoda i tako uticati na zadovoljstvo kupaca (samim tim i prodaju).



Slika 5. Parsiranje sentimenta unutar rečenice

6.2 Topic modeling

Topic modeling je tehnika nenadgledanog mašinskog učenja sposobna da preskenira set od više dokumenata pronalažeći paterne reči i fraza unutarnjih i automatski ih klasterovati u zasebne grupe sličnih izraza koji najbolje mogu okarakterisati te tekstove

7 Rezultati obrade

Polje veštačke inteligencije i ML-a konstantno se razvija obzirom da je to jedna mlada i ne sasvim istražena grana nauke. Samim tim, svakim danom se pojavljuje i neko poboljšanje procesiranja prirodnog jezika [11]. Postoji ogroman broj trenutno aktuelnih modela koji funkcionišu dobro za određene zadatke, dok imaju svoje nedostatke sa drugim. Ne postoji kompletan algoritam za potpunu obradu prirodnog jezika, ali stremi se ka tome da se uspešno otklone sve prepreke koje smanjuju preciznost postojanih i da se postigne sistem apsolutno sposoban da učestvuje u složenoj komunikaciji sa ljudskim bićima. Greške do kojih dolazi možemo "predvideti"evaluacijom više gotovih algoritama.

7.1 Boosting algoritmi

Busting algoritmi, za razliku od tradicionalnog mašinskog učenja koji se fokusira na tačnost rezultata samo jednog modela, povezuju više slabije istreniranih tako da svaki model rešava slabosti drugog. Busting algoritmi su generični algoritmi, a ne specifični modeli. Često se primenjuju kod različitih NLP problema koji se sastoje iz više povezanih koncepta, a najpovoljnijim se, za sada, pokazao tzv. **xg Boost**. Extreme Gradient Boosting je algoritam zadužen da kreira nove modele koji predviđaju preostale greške ili greške ranijih modela, te njihovom kombinacijom pravi finalne predikcije. Veoma je koristan kod rada sa posebnim fičerima unutar teksta, a najbolje manipuliše podacima struktuiranim u tzv. dataframe-ove.

8 Literatura

[1] - Ranko Bugarski - *Uvod u Opštu lingvistiku* (1989)

[2] - Alan Turing - *Computing machinery and intelligence* (1950)

[3] - Zsolt Nagy - *Osnove veštačke inteligencije i mašinskog učenja* (2019)

[4] - Joakim Nivre - *On Statistical Methods in Natural Language Processing*, School of Mathematics and Systems Engineering, Växjö University (2016)

dodati izvore

[] - Ramandeep Kaur, Sandeep Kautish - *Sentimental Analysis- from theory to practice* Lap Lambert Academic (2017)

[] - Digvijay Singh, Sourabh Prajapati - *Sarcasm Detection: Step towards Sentiment Analysis* Forsk Technologies (2019)

[8] - https://www.nltk.org/ (9. april 2020)

[9] - https://spacy.io/ (9. april 2020)

[10] - https://pytorch.org/ (10. april 2020)

[11] - http://nlpprogress.com/, repozitorijum sa listovanim aktuelnim algoritmima

Sadržaj