Svojstva velikih količina podataka

Mikulec, Kristina

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Science / Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:217:573591

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Kristina Mikulec

SVOJSTVA VELIKIH KOLIČINA PODATAKA

Diplomski rad

Voditelj rada:
dr. sc. Ognjen Orel

Zagreb, rujan, 2023.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Sadržaj

Uvod

Razvojem tehnologije postalo je sve lakše prikupljati podatke iz raznih izvora, od uređaja koje svakodnevno koristimo, društvenih mreža i web aplikacija do napretka s medicinskim slikama i senzorima. Tako stvorene velike količine podataka (eng. *big data*) postale su novi pokretač razvoja u obradi i skladištenju podataka. Zbog same veličine i brzine kojom takvi skupovi podataka nastaju, a i zbog različitih oblika u kojima se mogu naći, njima se ne može upravljati niti ih analizirati tradicionalnim tehnologijama. Zato je bilo potrebno razviti nove tehnologije i metode za efikasnu obradu.

Veliki podaci mogu se karakterizirati svojim svojstvima, često grupiranima u skupine od više „V" osobina. „6 V" odnosi se na tri osnovna svojstva, količinu (eng. *volume*), brzinu (eng. *velocity*) i raznolikost podataka (eng. *variety*), te tri stečena svojstva, istinitost (eng. *veracity*), vrijednost (eng. *value*) i promjenjivost podataka (eng. *variability*). U ovom radu dan je detaljan osvrt na stečena svojstva, s primjerima metoda i tehnologija za njihovo upravljanje i analizu te primjerima situacija u kojima su ta svojstva prisutna. U praktičnom dijelu provedena je analiza tih svojstava na podacima s društvenih mreža.

Poglavlje 1

Svojstva velikih podataka

Složenost upravljana velikim količinama podataka i njihovom obradom proizlazi iz svojstava koja ih karakteriziraju. Tri osnovna svojstva su ogromne količine podataka koji se moraju obraditi, velika brzina kojom su ti podaci stvoreni i razni oblici u kojima mogu doći. S vremenom su prepoznata i druga svojstva koja proizlaze iz osnovnih, a 3 koja popunjavaju „6 V" osobina su: istinitost, odnosno kvaliteta podataka, promjenjivost podataka i njihovog izvora te vrijednost koja se može pronaći analizom podataka.

Količina podataka

U današnjem svijetu brzi razvoj tehnologija i sve veća povezanost omogućili su ogromno povećanje količine dostupnih podataka. Procjenjuje se da je 2022. godine stvoreno i prikupljeno 97 zetabajta podataka, skoro 50 puta više nego 2010. godine. Za dvije godine taj broj trebao bi prestići 180 zetabajta. [14] Takva ogromna količina podataka zahtijeva posebne metode pohrane i obrade koje se oslanjaju na odgovarajuću tehnologiju. Također, nužni su algoritmi koji mogu izvući korisne uvide iz toliko podataka.

Brzina podataka

Brzina u kontekstu velikih podataka odnosi se na veliku brzinu stvaranja podataka, brzinu unosa podataka i analizu velike količine raznolikih podataka u stvarnom vremenu. Izvori podataka poput senzora, društvenih mreža i sustava trgovanja stalno stvaraju nove podatke koje je potrebno obraditi. Tradicionalni sustavi upravljanja podacima često se ne mogu nositi s brzinom dospijeća podataka. Specijalizirane tehnologije i platforme razvile su se kako bi osiguravale da se podaci prikupe, obrade i pohrane bez opterećenja infrastrukture. Za određena područja vrlo je bitna i analiza podataka u stvarnom vremenu kako bi se donijele pravovremene odluke.

Raznolikost podataka

Analiza velikih podataka obuhvaća podatke iz raznih izvora, ljudskih ili mehaničkih, koji mogu pridonijeti cilju istraživanja. Razni oblici u kojima dolaze mogu se podijeliti na strukturirane, polustrukturirane i nestrukturirane. Strukturiranim podacima pripadaju podaci s definiranom strukturom, formatom i tipom, primjerice podaci u relacijskim bazama podataka ili u CSV datotekama. Polustrukturirani podaci uključuju logove te XML i JSON datoteke. Nestrukturirani podaci mogu uključivati ljudski generirane tekstualne podatke i multimediju.

Istinitost podataka

Istinitost podataka odnosi se na kvalitetu i povjerenje u podatke. Podaci mogu sadržavati krive vrijednosti te vrijednosti mogu nedostajati ili ne pružati precizna mjerenja. Ljudska greška i obmana također mogu narušiti vjerodostojnost podataka. Očuvanje i poboljšanje kvalitete podataka ključno je za donošenje ispravnih odluka i izvlačenje pouzdanih zaključaka. Kontrola pristupa i provjera izvora također pomažu održati istinitost podatka.

Promjenjivost podataka

Podaci se mogu promijeniti kroz vrijeme zbog različitih uzroka, što predstavlja izazove prilikom analize. Sami podaci mogu biti ispravljani ili zamijenjeni novima te poprimiti drugo značenje kroz vrijeme i kontekst. Modeli koji se koriste za analizu moraju se uskladiti s novim značenjem podataka i unaprjeđivati kako bi ostali relevantni. Nadalje, skupovi velikih podataka često dolaze iz različitih izvora te ih je potrebno pravilno integrirati. Sustavi koji ih povezuju moraju biti otporni na promjene radi pouzdane obrade podataka.

Vrijednost podataka

Podaci imaju intrinzičnu vrijednost, no ona izlazi na vidjelo tek nakon kvalitetne obrade. Razvoj prikladnih tehnologija omogućio je jeftiniju pohranu i obradu podataka, povećavajući mogućnosti inovacija i dubljeg uvida u različite dijelove poslovanja te prodornijeg istraživanja. Sve više organizacija koristi se velikim podacima radi efikasnijeg poslovanja i razvoja novih proizvoda. Znanstveno istraživanje, pogotovo u zdravstvu, vidjelo je velike napretke upotrebom velikih količina podataka iz raznih izvora.

1.1 Primjeri velikih podataka

Internet stvari (eng. *Internet of Things* - IoT) objedinjuje izvore senzorskih podataka, od uređaja za praćenje tjelesne kondicije i pametnih kućanskih aparata do industrijskih senzora i sustava za praćenje promjena u vremenskim prilikama. Očekuje se da će vrijednost podataka senzora dostići između 5.5 bilijuna i 12.6 bilijuna dolara do 2030. godine. [21] Ta vrijednost proizlazi iz različitih okruženja u kojima je IoT implementiran, pri čemu okruženje tvornica čini oko 26% tog potencijala, a okruženje zdravstva 10-14%. Veći dio vrijednosti IoT-a očekuje se u primjeni među organizacijama (B2B), čineći otprilike 65% projicirane vrijednosti do 2030. godine. Međutim, raste i broj potrošački orijentiranih IoT (B2C) primjena, posebno u području automatizacije kućanstva.

Digitalizacija financijske industrije uvelike je promijenila globalna tržišta uz pristup sve većoj količini podataka i naprednijim tehnikama obrade. Dostupni podaci mogu biti strukturirani ili nestrukturirani te nude uvide u ponašanje korisnika i pomažu u formuliranju strategija. Neke od mogućih koristi obrade tih podataka su sprječavanje prijevara, ciljano usmjeravanje korisnika, optimizacija performansi i bolja procjena izloženosti riziku. Također, zbog same količine podataka koji se svakodnevno stvaraju na globalnim tržištima i oko njih, sve su popularnije tehnologije temeljene na oblaku za obradu i skladištenje takvih podataka.

Zdravstveni podaci dolaze iz raznolikih izvora koji svaki na svoj način doprinose ukupnoj slici o zdravlju osobe ili stanju zdravstva. Putem sekvenciranja genoma moguće je dobiti sliku genetskog sastava osobe, što služi identifikaciji genetskih varijacija povezanih s osjetljivošću na bolesti i odgovorom na liječenje. Elektronički zdravstveni zapisi bilježe medicinske povijesti pacijenata, a medicinske slike, poput rendgenskih i CT snimaka, pomažu u dijagnosticiranju bolesti pružajući uvid u unutrašnjost tijela bez korištenja invazivnih tehnika. Pametni uređaji i podaci o zdravlju generirani od strane pacijenata pružaju informacije o stanju pacijenata u stvarnom vremenu, kao i njihovoj tjelesnoj aktivnosti, okolišu i načinu života, što omogućava bolju personaliziranu njegu. Integracija takvih podataka zahtjevna je i uvelike se može okoristiti metodama za obradu velikih količina raznovrsnih podataka.

Važan izvor velikih podataka čine društvene mreže, web aplikacije i kritika korisnika. Društvene mreže pružaju obilje sadržaja koje stvaraju korisnici, kroz objave, komentare i druge interakcije. Web aplikacije, od trgovinskih platformi do web stranica za objavu vijesti, generiraju ogromne količine podataka putem korisničkih interakcija, klikova i navika pregledavanja. Kritike pružaju neposredno mišljenje korisnika o proizvodima i uslugama. Zajednička analiza ovih izvora pruža detaljnije uvide u navike korisnika i nastajanje trendova, kao i uvide u ljudsko ponašanje i sentiment tijekom značajnih događaja.

Poglavlje 2

Istinitost podataka

Istinitost podataka odnosi se na stupanj kvalitete i povjerenja u podatke. Podaci moraju biti pouzdani i što preciznije opisivati stvarne događaje i pojave. Trebali bi biti potpuni i bez značajnih neslaganja. Porijeklo podatka može uvelike ukazati na njihovu kvalitetu. Provjera izvora, usporedba s drugim izvorima i jasni zapisi životnog vijeka podatka mogu ukazati na to kojim podacima možemo vjerovati. Potrebno je znati kada i na koji način su prikupljeni podaci. Također je važno povjerenje u organizaciju, osobu ili sustav koji prikuplja podatke i zbog toga je potrebno prethodno znati reputaciju izvora koji se koristi te provoditi redovite provjere.

Osim samog izvora, potrebno je provjeriti i proces prikupljanja, pohrane i obrade podataka. Praćenje promjena uz odgovarajuću dokumentaciju te pravilno upravljanje podacima uz kontrolu pristupa nužni su koraci u očuvanju integriteta podataka. Pravovremena dostupnost podataka ovlaštenim korisnicima također je bitna. Sigurnosne kopije i distribuiranost podataka na više servera ili lokacija osiguravaju da će se podaci očuvati i u slučaju kvara sustava.

2.1 Kvaliteta podataka

Procjena kvalitete podataka ključna je faza prilikom bilo kakve analize podataka. Sirovi podaci nikada se ne mogu koristiti za provođenje analize jer su podložni ljudskim pogreškama i netočnoj automatskoj generaciji, poput podataka senzora. Loši podaci dovode do netočne analize i potencijalno pogrešnih poslovnih odluka ili propuštenih prilika. Istraživanje tvrtke Gartner otkrilo je da organizacije smatraju da loša kvaliteta podataka uzrokuje prosječne gubitke od 15 milijuna dolara godišnje. [29] Porastom količine podataka i broja različitih izvora, procjena kvalitete postaje sve teža. Za bolju procjenu potrebno je

promatrati kvalitetu kroz više dimenzija, poput točnosti, potpunosti, konzistentnosti, pravovremenosti, valjanosti i jedinstvenosti. [7] [35]

Točnost podataka odnosi se na ispravnost i preciznost podataka, odnosno koliko dobro podaci održavaju činjenice i događaje u stvarnom svijetu. Problemi se mogu pojaviti pri samom prikupljanju podataka jer zabilježene vrijednosti mogu biti približne, a ne stvarne. Na primjer, korištenje slabijih senzora koji samo povremeno šalju mjerenja može dovesti do toga da se propuste neke intenzivne, ali kratke promjene u okolini. Također, uzimajući u obzir količinu senzora koji se svakodnevno koriste, raste i vjerojatnost da se neki pokvare ili povremeno ne šalju podatke.

Nepotpuni podaci mogu ozbiljno iskriviti rezultate i dovesti do pogrešnih interpretacija. To postaje posebno izraženo kod prikupljanja podataka iz različitih izvora, gdje neki izvori mogu sadržavati više informacija od drugih. Kod velike količine senzora, raste i vjerojatnost da se neki pokvare ili povremeno ne šalju podatke. U takvim slučajevima, važno je identificirati nedostajuće podatke i odlučiti o najboljem načinu njihovog rješavanja, putem imputacije ili drugih metoda.

Konzistentnost se odnosi na sposobnost usklađivanja podataka bez proturječnosti po svim mjerama od interesa. Nesuglasni podaci mogu se pojaviti zbog više čimbenika, kao što su različiti standardi unosa podataka, različite mjerne jedinice ili čak proturječne informacije.

Pravovremenost podataka znači da su podaci dostupni i ažurirani u odgovarajućem vremenskom okviru. Stvarna svrha analize može zahtijevati trenutačne informacije kako bi se donijele brze odluke, primjerice u financijskom sektoru ili praćenju stvarnih vremenskih događaja. Prikupljanje i ažuriranje podataka u stvarnom vremenu može biti složen izazov, ali je ključno kako bi se osigurala relevantnost i pouzdanost analize.

Valjanost podataka mjeri koliko podaci odgovaraju potrebnom formatu po navedenim pravilima. Formatiranje često uključuje metapodatke koji se sastoje od valjanih tipova podataka, raspona, uzoraka i sl.

Jedinstvenost provjerava broj dupliciranih podataka u skupu, primjerice identifikatora korisnika.

2.2 Ljudski generirani podaci i dezinformacije

Zahvaljujući brzom razvoju interneta, sve je više informacija dostupno sve većem broju ljudi koji uvelike koriste internetske članke i društvene mreže kao ključne izvore informacija o svijetu i svakodnevnih savjeta. Uz to, koncepti kao što su marketing, kritike proizvoda te personalizacija korisničkog iskustva sve više se oslanjaju na analizu tekstualnih podataka kako bi bolje razumjeli potrebe i navike korisnika. Razvoj tehnologija i modela za obradu prirodnog jezika (NLP) uvelike je olakšao analizu ovakvih podataka, iako još postoje mnogi izazovi.

Izazovi obrade prirodnog jezika

Ljudski generirani tekstualni podaci puni su nesigurnosti prilikom obrade unatoč nedavnim napredcima. Prije svega, razumijevanje konteksta i fraza, gdje iste riječi poprimaju različita značenja unutar rečenica, ponekad je teško ljudima, a pogotovo modelima koji se oslanjaju samo na podatke koji su im dostupni. Sličan izazov predstavljaju sinonimi, različite riječi sa sličnim, ali ne nužno identičnim značenjem. To su i homonimi, riječi koje se izgovaraju isto, ali imaju različite definicije, što otežava pretvorbu govorenog jezika u tekst.

Rečenice koje koriste sarkazam i ironiju te imaju višestruka značenja dodatno otežavaju razumijevanje. Jezik koji sadrži neformalne izraze, idiome i kulturološki specifične fraze, kao i izraze specifične za domene ili geografska područja otežava dizajniranje modela za široku primjenu. Za poboljšanje takvih modela potrebno je stalno unaprjeđivanje i dodavanje novih primjera za treniranje modela.

Uz navedene izazove koji proizlaze iz samog jezika, probleme može predstavljati ljudska greška ili namjera. Unatoč napretku u aplikacijama za automatsko ispravljanje i gramatiku, pogrešno napisane ili pogrešno upotrijebljene riječi i dalje predstavljaju problem. Također, ponekad je potrebno analizirati namjeru autora teksta, što još uvijek predstavlja veliki izazov.

Dezinformacije

Porastom izvora informacija na internetu porasla je i količina netočnih i nepotpunih informacija te se one danas šire brže nego ikad. Posljedice toga pogotovo su došle do izražaja u nedavnoj pandemiji COVID-19, tijekom koje se pojavio i izraz „infodemija". Široka upotreba interneta (pojačana time što su mnogi ostali u svojim domovima tijekom pandemije), društvene mreže i potreba za informacijama u stvarnom vremenu stvorile su okruženje pogodno za brzo širenje i istinitih i izmišljenih informacija. Brzina nastanka i širenja informacija te sama količina otežale su provjeru i suzbijanje lažnih informacija.

Provjera izvora informacije i potvrda iz više izvora pridonose vjerodostojnosti vijesti. Primjerice, PolitiFact, stranica za provjeru vijesti, koristi 3 novinara za provjeru svake vijesti. Ipak, količina vijesti koju mogu ručno provjeriti je vrlo mala. Brzina širenja vijesti ili glasine i njen doseg obično se povećavaju što je vijest uzbudljivija. Čak i u suštini istinite vijesti mogu biti izmijenjene da bi se o njima više raspravljalo. Također, potreban je samo jedan prenositelj s velikom publikom na nekoj društvenoj mreži i neka informacija se može proširiti nadaleko te ukorijeniti u svijest publike, bez obzira na točnost.

Kao odgovor na sve težu pravovremenu provjeru informacija, automatsko otkrivanje lažnih vijesti pojavilo se kao praktična primjena NLP-a. Svrha mu je olakšati teret ljudskog uključivanja u identifikaciju i suzbijanje širenja lažnih informacija. Problemi koje je potrebno riješiti mogu se podijeliti na provjeru činjenica i izjava te određivanje glasina, vijesti koje ne možemo dokazati u trenutku analize. Također, stav i namjera, kao i izraženi sentiment autora ili odgovora na neku izjavu mogu pridonijeti provjeri, no to predstavlja složen problem obrade prirodnog jezika.

2.3 Poslovne metode za poboljšanje i očuvanje kvalitete podataka

Analiza porijekla podataka

Analiza porijekla podataka (eng. *data lineage*) odnosi se na proces praćenja životnog tijeka podatka, pružajući jasno razumijevanje odakle su podaci potekli, kako su se promijenili i koja je njihova konačna odredišna točka. [7] Alati za takvu analizu pružaju zapis o podacima tijekom cijelog njihovog životnog ciklusa, uključujući informacije o izvoru i sve transformacije podataka koje su primijenjene tijekom ETL (eng. *Extract, Transform, Load*) procesa.

Ova vrsta dokumentacije omogućuje korisnicima provjeru točnosti i konzistentnosti. Često se koristi kako bi se utvrdila povijest procesiranja te kako bi se pronašle pogreške i njihovo porijeklo. Važan dio ovog procesa su metapodaci koji pružaju informacije poput tipa, formata, strukture, autora, datuma nastanka i datuma promjena podataka te veličine datoteke u kojoj su podaci spremljeni.

Analizira se i odnos među poslovnim i IT aplikacijama u raznim poslovnim domenama. Neki od detalja koji su uključeni mogu biti: gdje se podaci nalaze i kako su pohranjeni, kako se podaci mogu koristiti i tko je odgovoran za ažuriranje, korištenje i izmjenu podataka, uključujući osjetljive osobne podatke, praćenje izmjena podataka te integraciju podataka iz različitih dijelova organizacije.

Profiliranje podataka

Profiliranje podataka (eng. *data profiling*), ili arheologija podataka, je proces pregledavanja i čišćenja podataka kako bi se bolje razumjela njihova struktura i održavali standardi kvalitete podataka unutar organizacije. [7] Podaci se procjenjuju na temelju raznih dimenzija kvalitete koristeći niz poslovnih pravila i analitičkih algoritama. Tijekom profiliranja nastoji se odrediti strukturu i analizirati sadržaj podataka te otkriti veze između različitih skupova podataka.

Glavne zadaće tijekom profiliranja mogu se detaljnije podijeliti na **profiliranje metapodataka** (za otkrivanje strukture i odnosa među skupovima podataka), **profiliranje prezentacije** (za pronalaženje uzoraka), **profiliranje sadržaja** (za procjenjivanje kvalitete), **profiliranje skupova** (analiza distribucije podataka i drugih statistika) te **profiliranje logičkih pravila** (analiza pridržavanja poslovnim pravilima). [22]

Komercijalni alati poput IBM InfoSphere Information Analyzer, Informatica Data Profiling i Oracle Enterprise Data Quality pružaju razne metode za profiliranje podataka. Među njima su: analiza stupaca, identifikacija ključeva, otkrivanje redundantnih podataka, analiza pravila podataka, analiza agregacija, procjena funkcionalne ovisnosti, provjera adresa, standardizacija podataka, analiza metapodataka i druge. Osim komercijalnih alata, javno su dostupne i besplatne opcije poput Talend Open Studio for Data Quality, koji podržava razne statističke metode i analize strukture i sheme podataka.

Upravljanje podacima

Upravljanje podacima (eng. *data governance*) promiče dostupnost, kvalitetu i sigurnost podataka organizacije putem različitih politika i standarda. Ovi procesi određuju vlasnike podataka, sigurnosne mjere za podatke i namjene podataka. Sveukupno, cilj upravljanja podacima je održavanje visokokvalitetnih podataka koji su sigurni i lako dostupni za dublje poslovne uvide. [7] Dodatni izazovi upravljanja velikih količina podataka uključuju samu količinu i raznolikost oblika podataka, različite vrste izvora podataka i procesiranje u stvarnom vremenu.

Različite uloge dodijeljene su svima koji dolaze do podataka u nekom dijelu njihovog životnog ciklusa. Na čelu postoji odbor odgovoran za razvoj i odobravanje strategije upravljanja i standarda. Za razvoj sustava za upravljanje velikim količinama podataka potrebni su arhitekti sustava i inženjeri koji se brinu o pohrani, čišćenju i organizaciji podataka. Analitičari su odgovorni za analizu podataka i pronalaženje uvida korisnih za poslovanje.

Korisnici podataka unose podatke, pristupaju različitim skupovima podataka i generi-

raju izvješća. Među njima, vlasnici podataka su odgovorni za kvalitetu i sigurnost podataka unutar svojih timova. Odgovornost implementacije i nadzora provođenja strategije dijele glavni direktori za podatke, koji sudjeluju i u razvoju strategije upravljanja, te upravitelji, koji sudjeluju u obuci drugog osoblja.

Okvir za upravljanje velikim količinama podataka trebao bi obuhvatiti poslovne ciljeve, a važnost upravljanja potrebno je komunicirati svima u doticaju s podacima, te im dodijeliti već spomenute uloge. Jasna komunikacija, obuka i suradnja nužni su elementi za održavanje efikasnog okvira. Potreban je i skup pravila i standarda koji će osiguravati korektnu i legalnu upotrebu podataka. Uz to, uspostavljanje različitih razina pristupa podacima povećava sigurnost i olakšava kontrolu pristupa i izmjena podataka.

Arhitektura sustava koji se mora nositi s velikim količinama podataka iz raznih izvora može se sastojati od sljedećih slojeva:

- **Sloj podataka** pohranjuje strukturirane i nestrukturirane podatke, na fizičkim lokacijama u bazama ili distribuiranim sustavima ili u oblaku pomoću usluga poput Amazon Web Services. Uključeni su i podaci u stvarnom vremenu koji dolaze iz izvora poput uređaja Interneta stvari i web aplikacija.
- **Sloj integracije i unosa podataka** objedinjuje procese poput ETL-a za pravilnu transformaciju i integraciju podataka.
- **Sloj obrade** sadrži poslovno skladište podataka, u kojem se podaci pretvaraju u oblike prikladne za SQL upite i OLAP poslužitelje, a zatim analiziraju alatima za poslovnu inteligenciju. Za analizu i spremanje podataka u nestandardnim oblicima koriste se tehnologije poput Apache Hadoop i Apache Spark.
- **Sloj analitike i poslovne inteligencije** koristi alate za vizualizaciju podataka i alate za poslovnu inteligenciju temeljene u oblaku za istraživanje, vizualizaciju i interakciju s podacima. Podržava se i analiza u stvarnom vremenu za stalne tokove podataka.

[34]

2.4 Automatizirano čišćenje i provjera podataka

Čišćenje velikih količina podataka

Sa svakodnevnim porastom količine podataka, učinkovitost tradicionalnih pristupa čišćenju podataka postaje ograničena. Postojeći alati bore se sa skalabilnošću, kao i s distribuiranom prirodom mnogih sustava za upravljanje velikim količinama podataka. Veliki problem predstavlja i raznolikost podataka te različite vrste pogrešaka i nedosljednosti u podacima. Često se zahtijeva doprinos stručnjaka za domenu radi razumijevanja i provjere obrađenih podataka, što povećava troškove.

Efikasno čišćenje velikih količina podataka može se postići skalabilnim i automatiziranim metodama. Predloženi su sljedeći sustavi:

- **Cleanix** je paralelni sustav za čišćenje podataka koji se sastoji od 4 glavne faze, počevši sa pronalaženjem i ispravkom nepravilnih vrijednosti i završavajući s rješavanjem suprotstavljenih i dupliciranih podataka. Korisničko sučelje omogućava unos informacija o izvoru podataka, parametrima i odabiru pravila za prilagođeno čišćenje podataka.
- **SCARE (eng. *SCalable Automatic REpairing*)** je skalabilni okvir koji koristi mehanizam za horizontalnu segmentaciju podataka kako bi osigurao skalabilnost i omogućio paralelnu obradu blokova podataka. Strojno učenje koristi se za predviđanje više atributa podataka istovremeno, pri čemu zamjenske vrijednosti prljavih podataka proizlaze iz analize čistih podataka.
- **BigDansing** je sustav koji korisnicima omogućava da specificiraju tok podataka za otkrivanje pogrešaka odabirom pravila koja se dalje prevode u logički plan čišćenja. Podržan je velik broj različitih pravila, a skalabilnost se postiže korištenjem postojećih okvira za paralelnu obradu podataka.
- **KATARA** odstupa od ostalih sustava jer uključuje *crowdsourcing* kao ključni dio čišćenja podataka, uz pouzdane baze znanja. Razlog ovom pristupu je nepovjerenje u automatizirane sustave i strojno učenje koje je podložno greškama u samom modelu koji se koristi. Podaci se provjeravaju koristeći baze znanja i za neispravne podatke se predlaže k najboljih rješenja sastavljenih prema pouzdanim bazama. Slijedi ljudska provjera i odabir najboljeg popravka, kao i prijedlog mogućih popravaka.

[36]

Poboljšanje kvalitete podataka prikupljenih senzorima

Uređaju povezani u Internet stvari generiraju značajne količine podataka iz senzora, no ti su podaci često nepotpuni i nekonzistentni. Primjerice, podaci iz radiofrekvencijske identifikacije često su nepouzdani. Da bi se to ispravilo, posrednički sustavi koriste filtar za ubacivanje nedostajućih vrijednosti, no učinkovitost filtra ovisi o veličini vremenskog prozora koji se koristi. Za određivanje prave veličine za svaku aplikaciju mogu se koristiti statističke metode.

Za rješavanje nesigurnosti u tokovima velike količine podataka predložen je pristup temeljen na načelima kvalitete podataka. Ovaj pristup kvantificira i komunicira kvalitetu podataka putem metapodataka, procjenjujući svojstva poput statusa, razine baterije i koordinata senzora. Procjene kvalitete podataka dijele se s korisnicima putem vizualizacije, zvučnih signala ili izvješća. Ubrzanje obrade postignuto je paralelnim algoritmima koji koriste MapReduce zajedno s tradicionalnim tehnikama ili teoriju grubih skupova za aproksimaciju nedostajućih vrijednosti.

Druge predložene metode uključuju pretvorbu podataka u određene oblike (poput uređenih parova) koristeći probabilističko modeliranje. Funkcija gustoće se zatim koristi za procjenjivanje nesigurnosti svakog objekta, a dodatne metode pomažu ubrzati obradu. Predložen je i model koji uključuje učenje iz dostupnih podataka senzora i predviđanje nedostajućih podataka na temelju ovisnosti između podataka svakog vremenskog razdoblja.

Poboljšanje istinitosti podataka dodatno se može postići pristupom temeljenom na povjerenju u izvor podataka. Određeni senzori su proglašeni pouzdanima na temelju odabranih faktora. Drugi pristup rješava problem napada na sustav koristeći strojno učenje za predviđanje pouzdanosti izvora, pritom analizirajući povijesne podatke iz tog izvora. [18]

Metode za provjeru činjenica i pouzdanosti informacija s društvenih mreža i web aplikacija

Jedan od najvećih izvora za velike količine podataka su web aplikacije. Tako prikupljeni podaci mogu biti strukturirani i nestrukturirani, s pomakom prema upotrebi strukturiranih povezanih podataka za poboljšanu provjeru. Metode poput Bloom filtra koriste se za procjenu vjerodostojnosti povezanih podataka. Vjerodostojnost podataka iz više izvora često se ocjenjuje prema nekoj metodi provjere istinitosti, no tu postoji problem nedostupnosti neke mjere za sve podatke ili pristranosti zbog ljudske procjene. Kombiniranje više metoda s uniformnim ili prilagođenim težinama pokazalo je veću točnost.

Kriptografske tehnike, poput računanja na maskiranim podacima (CMD) i blockchaina, predložene su kako bi se osigurala istinitost podataka istovremeno čuvajući privatnost. Blockchain uvodi tokene i identifikaciju izvora kao poboljšanja kvalitete podataka, dok decentralizirani identifikatori (DID-ovi) olakšavaju autentifikaciju izvora. Modificirana analiza utjecaja i posljedica pogrešaka (FMEA) također je predložena za provjeru podataka s weba, analizirajući rizik neuspjeha na temelju čimbenika poput integracije, nejasnoće, dostupnosti ili konzistentnosti podataka. [18]

Posebnu pažnju pri provjeri dobivaju podaci s društvenih mreža. Uz ocjenjivanje točnosti, koristi se i analiza ponašanja korisnika te analiza širenja informacija. Obrada sadržaja informacija koristi tehnike obrade prirodnog jezika, uz odgovarajuće metode za pretprocesiranje, poput lematizacije i tokenizacije. Za izdvajanje značajki rečenica, pa i cijelog teksta, često se koristi metoda TF-IDF (eng. *Term Frequency - Inverse Document Frequency*).

Za automatiziranu klasifikaciju informacija kao točnih ili netočnih te kao glasina ili informacija koje se mogu provjeriti često se koriste modeli strojnog učenja. Ti uključuju stroj s potpornim vektorima, naivni Bayesov klasifikator, logističku regresiju i stabla odlučivanja. Složeniji modeli temelje se na neuronskim mrežama, poput povratnih i konvolucijskih neuronskih mreža.

Unatoč razvoju kompleksnih sustava klasifikacije informacija, još uvijek su prisutni problemi na koje je potrebno obratiti pažnju pri provjeri informacija. Vijesti često sadrže kombinaciju istinitih i lažnih izjava, zbog čega postojeće stranice za provjeru informacija ocjenjuju članke kao „većinom lažne" ili „poluistinite". Kompleksniji sustavi koji provjeravaju cijeli članak i imaju precizniju klasifikaciju, za razliku od binarne koja se koristi, mogu dati detaljniju ocjenu vijesti. [33]

Poglavlje 3

Vrijednost podataka

Sve više područja koristi se prednostima koje donosi analiza velikih podataka. Od proizvodnje i financijskog sektora do zdravstva i marketinga, veliki podaci pružaju prednost pri donošenju informiranih odluka i ostvarivanju prednosti. U zdravstvu, analiza velike količine podataka omogućuje personaliziranu njegu i istraživanje novih tretmana. U ekonomiji se koristi za analizu tržišta, predviđanje trendova i optimizaciju poslovnih procesa. U financijskom sektoru, analiza velikih podataka doprinosi boljem upravljanju rizikom i donošenju boljih investicijskih odluka.

3.1 Razne primjene analize velikih podataka

Prediktivno održavanje u proizvodnji

Unutar svakog procesa u proizvodnji dobara postoji mnogo dijelova koje je potrebno nadzirati da bi se održala kvaliteta proizvoda te da ne bi došlo do zastoja proizvodnje. Pojava tehnologije senzora omogućila je da strojevi uz obavljanje svojih zadataka komuniciraju svoje stanje i performanse u stvarnom vremenu. Na tome se danas temelji prediktivno održavanje koje primjenom naprednih analitičkih alata i procesa poput strojnog učenja na podacima može identificirati, otkriti i rješavati probleme kako se događaju, ali i predviđati potencijalno buduće stanje opreme.

Za razliku od preventivnog održavanja koje se oslanja na unaprijed definirane rasporede, prediktivno održavanje pruža kontinuirane uvide u stvarno stanje opreme na temelju povijesnih i podataka o kvarovima. Također, koriste se različite metode poput analize zvuka (ultrazvučna akustika), analize temperature (termalna), procjene podmazivanja (ulje, tekućine) i analize vibracija koje mogu otkriti anomalije i pružiti rana upozorenja na potencijalne probleme. Primjerice, porast temperature komponente može ukazivati na blokade

protoka zraka ili trošenje, neobične vibracije mogu sugerirati nepravilan položaj pokretnih dijelova, a promjene u zvuku mogu ukazivati na nedostatke koje ljudsko uho ne može registrirati.

Prednosti koje proizlaze iz strategije prediktivnog održavanja usmjerene su na predviđanje kvarova opreme, smanjenje troškova održavanja i operativnih troškova optimizacijom vremena i resursa te poboljšanje performansi i pouzdanosti opreme. Prema izvješću Deloittea iz 2022. godine, prediktivno održavanje može rezultirati smanjenjem zastoja postrojenja za 5-15% i povećanjem produktivnosti rada za 5-20%. [7] Prediktivno održavanje također ima pozitivan utjecaj na održivost operacija minimiziranjem potrošnje energije i otpada.

Otkrivanje i prevencija financijskih prijevara

Na tržištima se svakodnevno odvijaju brojne transakcije koje je nemoguće pratiti u stvarnom vremenu korištenjem tradicionalnih metoda. Jedna od posljedica sve većeg broja transakcija je teže otkrivanje nedopuštenih transakcija. Da bi se to spriječilo, potrebni su algoritmi i tehnologije koje mogu brzo identificirati transakcije koje odstupaju od norme unutar velike količine trenutnih i povijesnih podataka. Na primjer, nagli veliki prijenos sredstava s računa koji je povijesno pokazivao konzervativne potrošačke navike mogao bi podići uzbunu.

Metode koje se primjenjuju za otkrivanje nepravilnosti su razne i razlikuju se po područjima financijske prijevare. Statističke metode i strojno učenje primjenjuju se za otkrivanje prijevara s kreditnim karticama i osiguranjem. U posljednjih nekoliko godina, modeli dubokog učenja dobivaju sve veću pozornost istraživanja za ovaj zadatak, a čak se koristi i obrada prirodnog jezika. Brzina i preciznost ovih metoda u otkrivanju sumnjivih aktivnosti osnažuje financijske institucije da brzo reagiraju i spriječe pokušaje prijevare prije nego što eskaliraju. [25]

Algoritamsko trgovanje i upravljanje rizikom

Algoritamsko trgovanje koristi računalni program koji slijedi definirani skup uputa (algoritam) za trgovanje. Algoritamsko trgovanje može u teoriji generirati profit brzinom i frekvencijom koja je nemoguća za ljudskog trgovca. Definirani skupovi uputa temelje se na vremenu, cijeni, količini ili bilo kojem matematičkom modelu. Osim prilika za profit trgovca, algoritamsko trgovanje čini tržišta likvidnijima i trgovanje sustavnijim eliminirajući utjecaj ljudskih emocija na trgovačke aktivnosti. [40]

Analiza velike količine informacija unapređuje ovaj proces pružajući obilje tržišnih podataka, povijesnih cijena, reakcija na vijesti i makroekonomskih pokazatelja obrađenih

u stvarnom vremenu za brže i bolje reakcije na promjenjivost tržišta. Korištene strategije temelje se na tehničkoj analizi, statističkim metodama, modelima strojnog učenja, rudarenju teksta i sličnim metodama. Posljedice uključuju ublažavanje rizika zbog preciznijih prognoza ponašanja tržišta i prepoznavanje utjecaja na trendove i ponašanje cijena. Zadnja bitna značajka algoritamskog trgovanja je mogućnost testiranja sustava na povijesnim podacima kako bi se otkrile nepravilnosti prije nego što se sustav pokrene na tržištu. [3]

Sekvenciranje genoma i personalizacija njege

Razvoj tehnologija sljedeće generacije sekvenciranja (eng. *Next Generation Sequencing* - NGS) uveo je napredak u području sekvenciranja genoma, omogućujući različite tehnike poput sekvenciranja cijelog genoma (eng. *Whole Genome Sequencing*), sekvenciranja cijelog egzoma (eng. *Whole Exome Sequencing*), RNK sekvenciranja i drugih. [42] NGS se sastoji od dva osnovna pristupa: sekvenciranje kratkih očitanja, poznato po svojoj ekonomičnosti i preciznosti pogodnoj za studije populacije i otkrivanje kliničkih varijanti, te sekvenciranje dugih očitanja, korisno za izgradnju genoma bez prethodnog znanja točnog redoslijeda fragmenata i otkrivanje izoforma gena.

Analiza podataka NGS-a uključuje masovno paralelno sekvenciranje kratkih očitanja i poravnavanje prema referentnom genomu ili, ako takav ne postoji, generiranje sljedova genoma iz fragmenata. Ovaj korak je ključan za identifikaciju polimorfizama jednog nukleotida (eng. *Single Nucleotide Polymorphisms*), strukturalnim promjenama DNK i drugim genomskim promjenama. Za analizu podataka koriste se razne metode i alati poput Genome Analysis Toolkit.

Sekvenciranje genoma može uvelike pridonijeti personalizaciji njege u različitim medicinskim područjima. Otkrivanje varijacija u genima koji utječu na prihvaćanje lijekova može usmjeravati prilagođene intervencije, sprječavajući nepoželjne reakcije na lijekove. Genetske informacije također pomažu pri odabiru ciljanih tretmana raka analizom DNK tumora. Moguće je i neinvazivno prenatalno testiranje za kromosomske abnormalnosti, a razvija se i tehnologija za procjenu rizika i predviđanje složenijih medicinskih stanja.

Rastuća složenost i količina podataka koje je potrebno analizirati, zajedno s rastom upotrebe sekvenciranja dugih očitanja, predstavljaju nove izazove koji se mogu riješiti tehnologijama za obradu velikih podataka. Za učinkovitu obradu velike količine sirovih podataka često se koriste platforme u oblaku, što olakšava i suradnju i dijeljenje podataka. Razvoju tih tehnologija u medicinske svrhe pridonose i inicijative personalizirane zdravstvene njege. Primjerice, AstraZeneca ima cilj sekvencirati i analizirati dva milijuna genoma, uključujući uzorke kliničkih ispitivanja, do 2026. godine.

Praćenje širenja zaraznih bolesti

Upotreba umjetne inteligencije uz obradu velikih količina podataka u analizi epidemija pokazala je veliki potencijal za kontrolu i prevenciju epidemija, posebno istaknut tijekom globalne pandemije COVID-19. Širenje pandemije 2020. godine otkrilo je ranjivosti unutar sustava javnog zdravstva diljem svijeta, naglašavajući potrebu za poboljšanjem odgovora na hitne slučajeve. Iskorištavanjem ogromnih količina informacija iz nacionalnih elektroničkih baza slučajeva pacijenata, odjeli javnog zdravstva mogu brzo pratiti zarazne bolesti i poboljšati nadzor te dovesti do suzbijanja područja širenja bolesti.

Umjetna inteligencija koristi se za predviđanje i otkrivanje visokorizičnih područja, razdoblja izbijanja i drugih obrazaca vezanih uz epidemije analizirajući velike skupove podataka iz fizičkog i digitalnog svijeta. Time se mogu ubrzati epidemiološka istraživanja, ključan zadatak u razumijevanju epidemija. Koristeći specijalizirane tehnologije može se pojednostaviti prikupljanje, prijenos i organizacija podataka. Napredne tehnike poput obrade prirodnog jezika poboljšavaju izvlačenje novih informacija iz izvješća o slučajevima i sličnih izvora. Osim toga, prediktivna analiza pomoću umjetne inteligencije doprinosi ciljanim preventivnim mjerama temeljenima na sveobuhvatnim podacima o pacijentima.

Tijekom borbe s epidemijom, posebno je važna učinkovita raspodjela zdravstvenih resursa i intervencija. Simulacije se mogu koristiti za procjenu potrebe za resursima i predviđanje trendova bolesti. Iako postoje izazovi u preciznosti modeliranja, kombiniranje medicinskih i društvenih resursa s modelima umjetne inteligencije može donijeti bolja rješenja. Primjena umjetne inteligencije posebno je korisna u farmaceutskom sektoru gdje ubrzava istraživanje i razvoj lijekova. Pomaže u racionalnom dizajnu lijekova, filtriranju spojeva te predviđanju i optimizaciji dizajna strukture lijeka. [26]

Maloprodaja i iskustvo korisnika

Analiza velikih količina podataka postala je nužna za uspjeh uz sve veću konkurenciju u maloprodajnoj industriji. Omogućava poboljšanje iskustva kupaca, optimizaciju marketinških strategija i efikasnije poslovne operacije. Prikupljanje i upravljanje podacima sve više se provodi tehnologijama na temelju oblaka i omogućava trgovcima da analiziraju razne izvore informacija o kupcima.

Omogućeno je stvaranje detaljnih profila kupaca prikupljanjem informacija o navikama, spolu, lokaciji, prisutnosti na društvenim mrežama i više. Na temelju tih informacija stvaraju se sofisticirane marketinške strategije, primjerice identifikacija mikro-influencera za jeftinije promicanje proizvoda. [4] Analiza snimljenih poziva, videozapisa iz trgovina, komentara na društvenim mrežama i ocjena kupaca može otkriti ključne probleme ko-

risničke usluge. Trgovci također mogu strateški pozicionirati proizvode unutar trgovina na temelju analize ponašanja kupaca.

Praćenjem trendova i razumijevanjem demografije i navika kupaca, tvrtke mogu odrediti optimalne cijene proizvoda ili proizvode koji im nedostaju u ponudi. Opskrba i distribucija proizvoda također su područja koja su unaprijeđena, smanjujući rizik nestanka zaliha i negativnog iskustva kupaca. Tehnologije i metode za rad s velikim podacima mogu se i ciljano koristiti tijekom vrlo aktivnih perioda, poput Crnog petka, za brzu obradu povećanog broja kupaca i bolju zaštitu od prijevare.

Upravljanje katastrofama

U području upravljanja i obnove nakon katastrofa, analiza podataka s društvenih mreža i sličnih platformi omogućuje učinkovitu koordinaciju pomoći, kao i procjenu pogođene infrastrukture i socijalno-ekonomske obnove. Olakšavaju se donacije i financijska podrška nakon katastrofe povezujući donatore i primatelje za brzu i personaliziranu dostavu pomoći te omogućavaju prepoznavanje potrebe i učinkovitu koordinaciju donacija. Analizom sentimenta izraženog tijekom i nakon katastrofe efikasnije se mogu identificirati potrebita područja i neadekvatna pomoć.

Podaci s društvenih mreža mogu se iskoristiti za procjenu oštećenja infrastrukture i poremećaja uzrokovanih katastrofama. Obrada prirodnog jezika i strojno učenje korišteni su za analizu tekstualnih i slikovnih podataka s platformi poput Flickra i Twittera, olakšavajući raspoznavanje opsega oštećenja ključne infrastrukture, raspodjelu resursa i napore rekonstrukcije.

Sentiment na društvenim mrežama u vezi s raznim tržištima na pogođenom području može se smatrati pokazateljem socio-ekonomske dinamike oporavka. Turizam, među ostalim djelatnostima, istražen je u kontekstu uloge društvenih mreža u ekonomskom oporavku. Istraživači su koristili podatke označene geolokacijom s Twittera i Flickra kako bi procijenili oporavak turističkih odredišta nakon katastrofe. [31] Razvijeni su i okviri za procjenu statusa oporavka malih poslovanja nakon katastrofe analizom promjena u objavama na društvenim mrežama. Ovaj pristup uspješno je primijenjen na regije pogođene potresima i uraganima, omogućavajući procjenu oporavka u stvarnom vremenu.

Upravljanje prometom

Gužva i ograničeno parkiranje konstantni su izazovi u upravljanju prometom. Konvencionalne metode poput cjevastih senzora ili ručnog brojanja pružaju ograničene uvide u dinamiku prometa. Umjesto toga, sve više se koriste izvori podataka kao što su videoza-

pisi prometa uživo i uređaji za praćenje. Ti izvori pružaju cjelovite podatke o klasifikaciji vozila, brzini, smjeru i vremenu. Podaci se mogu koristiti za odabir bolje rute u stvarnom vremenu, kao i dugoročno za otkrivanje uzroka gužve i poboljšanja infrastrukture.

Novi izvori podataka uključuju i podatke o mobilnim telefonima, pametnim karticama i geokodiranim objavama na društvenim mrežama. Podaci mobilnih telefona, s većim dosegom i demografskim prikazom od tradicionalnih i skupljih istraživanja, mogu pomoći u postizanju ravnopravnog pristupa javnom prijevozu. Na temelju navika stanovnika o kupovini i putovanju na posao mogu se optimizirati rute javnog prijevoza za četvrti s niskim vlasništvom automobila i implementirati sigurnosne mjere za pješake. Analiza ruta kamiona može se provesti kako bi se otkrili utjecaji buke i onečišćenja na određene četvrti i ceste. [30]

3.2 Alati i metode za pronalaženje vrijednosti podataka

Ključan korak u analizi podataka je odabir modela za obradu. Bez obzira radi li se o grupiranju, klasifikaciji ili pronalaženju veza, odabrani model treba zadovoljiti ciljeve istraživanja. Potrebno je uzeti u obzir i mogućnosti organizacije i pojedinaca koji provode istraživanje, kako bi se osiguralo da odabrani model odgovara njihovim resursima. Osim toga, važno je razmotriti skalabilnost odabranog modela, posebno ako se radi s velikim skupom podataka ili se očekuje rast količine podataka u budućnosti. Da bi model mogao pravilno obraditi podatke, oni moraju biti prethodno obrađeni, a sam model na kraju je potrebno testirati na manjem uzorku provjerenih podataka.

Strojno učenje predstavlja podskup umjetne inteligencije usredotočen na algoritme koji mogu „učiti" iz podataka bez eksplicitnog programiranja. [32] Takvi algoritmi donose odluke ili predviđaju ishode na temelju uzoraka ulaznih podataka, često koristeći statističke metode. Primjenjuju se na razne probleme gdje je teško naći eksplicitne algoritme za efikasno rješavanje problema, što pogotovo dolazi do izražaja kod velikih količina podataka. Strojno učenje može se podijeliti na nadzirano, polunadzirano i nenadzirano učenje, ovisno o razini ljudskog usmjeravanja prema cilju.

Grupiranje podataka

Grupiranje podataka pripada nenadziranim metodama strojnog učenja. Cilj je pronalaženje skrivenih struktura unutar neoznačenih podataka grupiranjem sličnih objekata. Zato se često koristi radi preliminarnog istraživanja skupa podataka i kao uvod u klasifikaciju. Jedan od najčešćih algoritama koji se koristi za grupiranje je k-sredina.

Algoritam k-sredina

Algoritam k-sredina dijeli objekte s n atributa na k disjunktnih grupa prema udaljenosti svakog objekta od središta svake grupe. Središte grupe određuje se kao aritmetička sredina n-dimenzionalnih vektora atributa unutar te grupe. Neke primjene ovog algoritma su u obradi slika, grupiranju živih bića prema njihovim karakteristikama i odjeljivanju kupaca prema sličnim navikama. Odabir broja grupa uvelike utječe na točnost algoritma i predstavlja dodatan problem optimizacije. [39] Koraci algoritma su sljedeći:

1. Odabire se broj k i pretpostavlja se k prvotnih središta.
2. Računa se udaljenost svake točke koja predstavlja objekt od svih središta te se pridružuju grupi s najbližim središtem.
3. Sada se računa stvarno središte svake grupe prema aritmetičkoj sredini.
4. Koraci 2 i 3 se dalje ponavljaju dok algoritam ne konvergira.

Regresijska analiza

Regresijska analiza je statistička metoda za otkrivanje ovisnosti zavisne varijable o jednoj ili više nezavisnih varijabli. Koristi se za predviđanje nedostajućih podataka na temelju prošlih mjerenja i određivanje konstanti koje opisuju zavisnost između varijabli. Ovisnost se može zapisati pomoću funkcije $y = f(x) + e$, pri čemu je x neovisna varijabla, y zavisna, $f(x)$ funkcija ovisnosti i e greška (slučajna varijabla s normalnom distribucijom i očekivanjem 0). Cilj je pronaći funkciju ovisnosti uz smanjenje greške.

Linearna regresija pretpostavlja da je riječ o linearnoj ovisnosti varijable ili podatka, tj. $f(x) = ax + b$. Ako takva ovisnost nije odmah prisutna, moguća je transformacija podataka, primjerice logaritamskom funkcijom. Koeficijenti a i b često se računaju metodom najmanjih kvadrata. Linearna regresija koristi se za predviđanje neprekidnih varijabli, poput cijena. S druge strane, **logistička regresija** predviđa vjerojatnost ishoda i pripadnosti kategorijama. Funkcija vjerojatnosti je $P(y) = \frac{1}{1 + e^{-y}}$, pri čemu y pretpostavlja linearnu ovisnost o varijabli x. [39]

Klasifikacija i prediktivna analiza

Za razliku od grupiranja podataka gdje još ne znamo kako su podaci podijeljeni, u klasifikacijskom strojnom učenju postoje već označeni podaci podijeljeni u dvije ili više kategorija. Modeli se zatim treniraju na označenim podacima i predviđaju kategorije za neoznačene podatke. Osim klasifikacije, moguće je predviđati i buduće događaje i uzorke ponašanja na temelju podataka za treniranje, kao i novih spoznaja.

Iako već spomenuti regresijski modeli pružaju rješenja za klasifikaciju i predviđanje, naprednije metode strojnog učenja daju preciznije rezultate i mogu uhvatiti više nijansi među podacima. Neki od najčešće korištenih algoritama u istraživanju su sljedeći:

- **Stroj s potpornim vektorima** predstavlja svestrani model za klasifikaciju podataka prema više značajki. Cilj algoritma je pronaći optimalnu hiperravninu u višedimenzionalnom prostoru radi razdvajanja točaka koje predstavljaju podatke s maksimalnim marginama. To se postiže pronalaženjem potpornih vektora, točaka najbližih hiperravnini. Jezgra predstavlja funkciju za transformaciju originalnih podataka u točke prostora određenog promatranim značajkama podataka. Efikasnost algoritma ovisi o odabiru jezgre. [11]

- **Naivni Bayesov klasifikator** je klasifikacijski model temeljen na Bayesovom teoremu. Glavna pretpostavka ovog modela je međusobna nezavisnost promatranih značajki, koje su većinom kategorijske varijable ili pretvorene u kategorijske. Prednosti modela su brzina i jednostavnost implementacije, kao i potreba manjeg skupa podataka za treniranje. S druge strane, model može pogrešno klasificirati podatke ako dođe do trovanja podataka. [39]

- **Nasumična šuma** kombinira više stabala odlučivanja za zajedničku odluku klasifikacije. Koriste se razne metode za nasumično stvaranje nekoreliranih stabala radi točnije klasifikacije. Stabla odlučivanja koriste jednostavne odluke u čvorovima (često binarne) za podjelu podataka. Počevši od korijena, podaci se usmjeravaju prema listu koji označava kategoriju kojoj podatak pripada. [39]

- **Neuronske mreže** sastoje se od više povezanih slojeva „neurona" koji imitiraju odlučivanje u ljudskom mozgu. Podaci ulaze kroz ulazni sloj, prolaze kroz izračune u jednom ili više skrivenih slojeva i na kraju proizvode podatke za izlazni sloj. Neuroni preuzimaju ulaz iz prethodnog sloja, primjenjuju težine i pomake te rezultat prolazi kroz aktivacijsku funkciju. To proizvodi transformirani izlaz koji se šalje sljedećim neuronima. Težine i pomaci se prilagođavaju s vremenom kako bi se minimizirale pogreške predviđanja. Neuronske mreže zahtijevaju veći skup podataka

za treniranje, kao i snažna računala (često se koriste grafičke kartice), no pružaju moćan i precizan alat za razne zadatke. [17]

Obrada prirodnog jezika

Obrada prirodnog jezika objedinjuje jezikoslovlje i umjetnu inteligenciju za računalnu analizu ljudskog jezika. Može se podijeliti na generaciju i razumijevanje jezika, što se dalje može podijeliti na fonologiju, morfologiju, leksičku analizu, sintaktičku analizu, semantičku analizu, analizu diskursa i pragmatičku analizu teksta.

Primjene su raznolike, počevši sa strojnim prevođenjem teksta s jednog jezika na drugi uz očuvanje značenja i gramatike. Klasifikacija teksta prema kategorijama pomaže pri filtriranju neželjene pošte i drugih sadržaja. Najveću i najsvestraniju korist ima izvlačenje informacija iz teksta, poput ključnih pojmova ili sentimenta. U zadnje vrijeme popularni su i sustavi dijaloga, odnosno interakcija čovjeka i računala, gdje je uz razumijevanje bitna i generacija jezika. [27]

Pretprocesiranje

Prvi korak pri obradi prirodnog jezika je pretprocesiranje podataka radi poboljšanja točnosti modela. Prije svega, podaci prikupljeni s interneta mogu sadržavati dodatne informacije poput linkova i hashtagova koje je potrebno maknuti ili barem odvojiti iz teksta. Daljnje čišćenje podataka uključuje ispravljanje pogrešno napisanih riječi i zamjena kratica cijelim riječima. Uklanjaju se i znakovi koje model ne može obraditi, što je pogotovo slučaj kada su podaci iz više jezika ili se koriste kompliciraniji emotikoni.

Ako je bitno pronaći ključne pojmove, bez obzira na kontekst rečenice, koristi se lematizacija. To je postupak pretvorbe riječi u njihov korijenski oblik pomoću rječnika. Ukoliko rječnik nije dostupan, isti postupak može se provesti koristeći heurističke metode. Također se pretpostavlja da kratke riječi poput veznika ne pridonose informacijama koje se traže, pa se one uklanjaju. Korisno može biti i prepoznavanje vlastitih imenica i njihovo izvlačenje za detaljniju analizu.

Slijedi tokenizacija, postupak podjele teksta u riječi i fraze. Riječi su često predstavljene numeričkim oznakama zbog zahtijeva modela za obradu. Ovaj proces je jednostavan u jezicima s jasnim granicama među riječima, no postaje veći problem u jezicima gdje to nije slučaj. Svakoj riječi dalje se pridodaju leksičke i sintaktičke informacije analizom originalne rečenice. Na kraju se računaju dodatne, često numeričke, značajke za potrebe modela, primjerice učestalost riječi u cijelom tekstu.

Modeliranje tema

Modeliranje tema koristi se za otkrivanje nepoznatih tematskih uzoraka unutar teksta, klasificiranje dokumenata prema pronađenim temama te organizaciju dokumenata za lakše pretraživanje. Latentna Dirichletova alokacija (eng. *Latent Dirichlet Allocation* - LDA) često se upotrebljava za modeliranje tema. Tema u LDA modelu definira se kao distribucija riječi, a dokumenti kao distribucije tema, s pretpostavkom da će se u dokumentima sa sličnim temama koristiti slične riječi. Važnu ulogu u modeliranju ima Dirichletova distribucija.

Analiza sentimenta

Analiza sentimenta je proces obrade prirodnog jezika u svrhu klasifikacije teksta prema subjektivnim značajkama. Najčešća klasifikacija je u tri kategorije: pozitivan, neutralan i negativan sentiment. Neki od izvora podataka koji se koriste za analizu su razne stranice za kritike, kritike dostupne na online trgovinama poput Amazona, blogovi i društevene mreže poput Facebooka i Twittera. U pogledu velikih količina podataka i nestrukturiranih podataka sve više se koriste modeli strojnog učenja zbog nedostataka tradicionalnih metoda.

Algoritam temeljen na **ključnim riječima** pretražuje tekst u potrazi za unaprijed definiranim pozitivnim i negativnim riječima kao što su „sretan", „radostan", „ravnodušan" ili „tužan". Na temelju prisutnosti ovih ključnih riječi, algoritam dodjeljuje oznaku sentimenta tekstu, primjerice „pozitivan", „negativan" ili „neutralan". Ovakav pristup analizi sentimenta je vrlo jednostavan i dolazi s ograničenjima. Teško je točno klasificirati negacije ključnih riječi, kao i suptilnije iskaze osjećaja.

Klasifikacija na temelju **leksikona** uključuje izradu lista riječi označenih njihovim pozitivnim ili negativnim polaritetima. Ove liste djeluju kao rječnici sentimenta ili leksikoni. Algoritam izračunava ukupnu ocjenu sentimenta za dani tekst prema polaritetima riječi sadržanih u tekstu. Prednost ovog pristupa je što ne zahtijeva treniranje modela, a ipak pruža bolju procjenu od samog pretraživanja ključnih riječi. Učinkovitost ovog pristupa može biti ograničena pri radu s nestrukturiranim podacima iz izvora poput društvenih medija, gdje je jezik neformalan i dinamičan. Također, ova metoda ima problema s interpretacijom sentimenta u kontekstu i generalizacijom rječnika. [41]

Analiza sentimenta **strojnim učenjem** započinje odabirom značajki prema kojima će se svaki tekst ocjenjivati. Važno je odabrati značajke koje pomažu jasno razlučiti različite sentimente za precizniju klasifikaciju. Često se tekst prvo podijeli u različite riječi ili fraze i prati se njihova frekvencija. Daljnje značajke uključuju riječi koje iskazuju određeni sentiment ili mišljenje i značajke specifične za oblik podataka. Primjerice, podaci s društvenih

mreža često sadrže razne dodatne atribute poput *hashtagova* ili lajkova te specifične načine izražavanja poput kratica poznatih izraza i emotikona.

Neki od algoritama strojnog učenja koji se upotrebljavaju za analizu sentimenta su naivni Bayesov klasifikator, stroj s potpornim vektorima, stabla odlučivanja i k-najbližih susjeda. Koriste se i umjetne neuronske mreže te genetski algoritmi i metode ansambla poput slučajnih šuma. Među njima se ističu naivni Bayesov klasifikator zbog jednostavnosti implementacije i brzine te stroj s potpornim vektorima radi analize velikog broja značajki u relativno kratkom vremenu. Još veća točnost može se postići neuronskim mrežama koje zahtijevaju više podataka i vremena za treniranje.

Osim tekstualne analize, strojno učenje omogućava i analizu drugih medija. Stroj s potpornim vektorima je, među ostalim metodama, korišten za analizu vizualnih i auditornih podataka. Sami auditorni podaci često se prepisuju u tekst automatskom generacijom radi tekstualne obrade podataka. Vizualna analiza donosi prednost prepoznavanja izraza lica, za što se mogu upotrijebiti modeli poput konvolucijskih neuronskim mreža.

Platforme za analizu velikih podataka

Projekt **Apache Hadoop** razvija softver otvorenog koda za pouzdanu, skalabilnu i distribuiranu obradu velike količine podataka. Sastoji se od više komponenta poput HDFS-a (Hadoop Distributed File System) za pohranu podataka i MapReduce algoritma za obradu. Povezani Apache projekti za analizu podataka uključuju **Apache Hive** (skladištenje podataka i *ad hoc* upiti), **Apache Pig** (paralelno računanje) i **Apache Spark** (podržava širok raspon operacija, od ETL-a do strojnog učenja). Spark također podržava različite programske jezike, poput Pythona, Jave i R-a, i njihove module za obradu i vizualizaciju podataka.

Za obradu tokova podataka specijaliziraju se okviri i platforme poput **Apache Flinka** i **Apache Kafke**. Ključni dijelovi Flinka uključuju sposobnost rukovanja različitim tipovima tokova (ograničeni i neograničeni, u stvarnom vremenu i zabilježeni), učinkovito upravljanje stanjem aplikacije, te podršku obradi vremena događaja sa značajkama poput oznaka vremena i rukovanja zakašnjelim podacima. Apache Kafka je distribuirana platforma za procesiranje i pohranu događaja u tokovima podataka, a pruža i mogućnost obrade događaja u stvarnom vremenu ili retrospektivno. [2]

Izvan Apache organizacije, postoje i druge značajne tvrtke koje nude sofisticirane alate za analizu i obradu podataka, primjerice **SAS**. SAS je poznat po svojim moćnim alatima za poslovnu analizu i upravljanje podacima. Njihova platforma omogućava korisnicima analizu, vizualizaciju i modeliranje podataka te implementaciju različitih statističkih metoda i umjetne inteligencije za obradu podataka. Zadnjih godina, posebno se ističe SAS

Viya, analitička platforma temeljena na oblaku. [10]

Tehnologije temeljene na oblaku pružaju fleksibilnu infrastrukturu za analizu velike količine podataka i često se integriraju s postojećim analitičkim alatima ili pružaju vlastite. Najveće platforme za računanje u oblaku su AWS (Amazon Web Services), Microsoft Azure i Google Cloud. **Qubole** je analitička platforma temeljena na oblaku (AWS i Google Cloud) koja pruža mogućnosti obrade podataka pomoću strojnog učenja i *ad hoc* upita, a uključuje i analizu tokova podataka te mehanizam upravljanja podacima. [9]

Poglavlje 4

Promjenjivost podataka

Promjenjivost velikih podataka odnosi se na promjene i varijacije prisutne u podacima. Nastaje iz različitih uzroka poput raznolikih izvora i intrinzičnih promjena podataka kroz vrijeme. Razlike u oblicima i količini prikupljenih podataka, promjena kvalitete i promjena u distribuciji neki su od primjera raznolikosti koja može nastati. Razumijevanje i učinkovito upravljanje tim promjenama, zajedno s prilagodbom modela korištenih za analizu, ključni su za iskorištavanje potencijala velikih podataka.

4.1 Primjeri promjenjivosti u podacima

Promjenjivost vremenskih prilika

Vremenske prilike su izrazito promjenjive, što predstavlja izazov pri njihovom praćenju i predviđanju. Modeli korišteni za predviđanje obuhvaćaju širok raspon događaja, od lokalnih promjena do globalnih prognoza. Te promjene mogu se dogoditi iznenada, čak i u roku od nekoliko minuta, poput iznenadnih naleta vjetra na malim područjima. Danas se za praćenje promjena koriste brojni senzori kako bi se zabilježile čak i najmanje promjene, osiguravajući točnije vremenske prognoze i ranija upozorenja na ekstremne uvjete. Promatraju se i dulji vremenski okviri u kojima se mogu dogoditi promjene duljeg utjecaja, uključujući godišnja doba i periode od više godina.

Okoliš i geografski položaj također će utjecati na prikupljene podatke pa je potrebno precizno pratiti i lokaciju promjene. Posebne pojave, poput El Niña, dolaze svakih nekoliko godina i unose dodatne promjene u sustav, poput vlažnog i kišovitog vremena u inače suhim područjima Južne Amerike. [15] Uzroci promjena uvelike se razlikuju i bitno ih je razlikovati za pravilnu analizu. Vanjski uzroci uključuju prirodne fenomene poput promjena u Sunčevom zračenju, vulkanske aktivnosti i tektonskih procesa, kao i ljudski

utjecaj poput emisije stakleničkih plinova. Unutarnja promjenjivost proizlazi iz složenih interakcija unutar samog klimatskog sustava i odvija se tijekom kraćih ili duljih perioda. Za analizu i predviđanje ovih promjena koriste se modeli različite složenosti, od jednostavnih konceptualnih do složenih globalnih modela. [43]

Analiza vremenskih nizova

Analiza vremenskih nizova je metoda analize nizova kronološki uređenih vrijednosti zabilježenih u redovitim vremenskim intervalima. Podaci se mogu bilježiti u određenim trenucima ili tijekom duljeg intervala zbrajanjem individualnih vrijednosti. Koristi se za pronalaženje informacija o razvoju pojava i njihovim međusobnim ovisnostima te predviđanje budućih događaja na temelju povijesnih podataka. Osim za predviđanje vremenskih prilika, česta je upotreba i u financijskom sektora za predviđanje tržišta.

Analiziraju se trendovi, sezonske i cikličke komponente, a slučajna odstupanja rješavaju se postupkom izglađivanja. Prilikom izglađivanja mogu se koristiti metode izračuna prosjeka ili metode eksponencijalnog izglađivanja. Eksponencijalno izglađivanje dodjeljuje eksponencijalno rastuće težine starijim opažanjima. Najčešće korišteni modeli za analizu podataka uključuju Box-Jenkins ARIMA model za analizu jedne vremenski ovisne varijable i multivarijatni model za analizu više varijabli. [8]

Promjenjivost jezika i izražavanja

Jezik predstavlja dinamičan oblik komunikacije koji je stalno podložan promjenama. Uzroci promjena su razni, od kulturnih promjena do razvoja tehnologije i društva. Geografska odvojenost zajednica također pridonosi raznolikosti oblika jezika. Mogu se promijeniti razni dijelovi jezika, poput vokabulara, gramatike i značenja riječi. Novi izrazi stvaraju se kao posljedica važnih događaja i promjena u društvu.

Internet i društvene mreže potaknuli su velike promjene u današnjem govoru. Riječi kao što su „tweet”, „selfie” ili „hashtag” postale su uobičajene u mnogim jezicima, uključujući i hrvatski, kako bi opisale nove koncepte povezane s društvenim mrežama. Razni izrazi, često na engleskom, koji se često koriste dobili su svoje kratice radi bržeg tipkanja prvenstveno na mobitelima, što se kasnije još više proširilo putem interneta. Emotikoni se često koriste u svrhu izražavanja emocija kroz tekst, od jednostavnih poput „:)” do malih slika s posebnim kodiranjem.

Osim kroz vrijeme, značenje riječi mijenja se i ovisno o kontekstu u kojemu su izrečene. Potrebno je pažljivo analizirati cijelu izjavu i društvenu okolinu iz koje dolazi. Često je potrebno razumijevanje konteksta kako bi se otkrilo dublje značenje ili nijanse

poruke. Namjera autora i odabrani način izražavanja, primjerice sarkazam, također mogu iskriviti značenje. Model za analizu jezika mora se trenirati na novim podacima kada se pojave ili podacima za odabranu domenu ako je potrebna analiza specifičnog područja.

Promjenjivost sentimenta kroz vrijeme

Mnogi faktori mogu utjecati na promjenjivost sentimenta javnosti tijekom raznih događaja ili o raznim temama. Tvrtkama je bitno pratiti nagle promjene kako bi kontrolirali mišljenje o svojim proizvodima i ugledu. Primjerice, nagla promjena iz pozitivnog u negativni sentiment ukazuje na neočekivane probleme s novim ili nedavno promijenjenim proizvodom. Tijekom javnih događaja promjene u sentimentu ukazuju na moguće probleme ili dijelove koji su se pogotovo svidjeli publici. Tijekom kriza, analiza sentimenta kroz vrijeme pruža uvid u učinkovitost odgovora institucija.

Osim naglih promjena i kratkoročnih događaja, dugoročna analiza promjena može ukazati na trendove u društvu i ekonomiji na koje je potrebno obratiti pozornost. Na primjer, stalno rastući negativni sentiment prema određenoj industriji može ukazivati na potrebu za restrukturiranjem ili promjenom poslovnog pristupa. S druge strane, konstantno pozitivan sentiment prema određenom proizvodu neke tvrtke može poticati druge tvrtke da razviju slične proizvode ili usluge kako bi zadovoljile potrebe tržišta.

4.2 Prilagodba modela

Uz promjene podataka, modeli se često moraju prilagoditi kako bi ostali učinkoviti. To je pogotovo bitno u području financijske prijevare, gdje se stalno pojavljuju nove strategije prijevare. [44] Tradicionalne metode koje su se držale samo prethodno definiranih pravila i radile na većinom strukturiranim podacima danas su neučinkovite. Zbog toga se detekcija prijevare okrenula strojnom učenju.

Korišteni su modeli poput naivnog Bayesovog klasifikatora, stroja s potpornim vektorima, logističke regresije i nasumične šume. Značajke na kojima se modeli treniraju uključuju korisničke profile i povijest transakcija. Modeli dubokog učenja, poput neuronskih mreža, sami pronalaze bitne značajke te su korisni na velikim i raznolikim skupovima podataka, zbog čega su postali popularni zadnjih godina. No, još uvijek se pojavljuje problem praćenja promjena u stvarnom vremenu.

Stacionarni modeli strojnog učenja su oni modeli koje se trenira jednom na ulaznim podacima i zatim koristi. Takvi modeli ovise o distribuciji podataka, tj. zahtijevaju da se novi podaci podvrgavaju sličnoj distribuciji. No, to nije uvijek slučaj i promjene su

često nepredvidive. Zbog toga je potrebno ponovno trenirati modele na novim podacima ili razviti algoritme koji su sposobni prilagoditi se promjenama tijekom vremena.

Pomak koncepta

Pomak koncepta (eng. concept drift) odnosi se na promjenu distribucije podataka u nekom trenu tijekom toka podataka. Ovisno o brzini promjene, pomak se karakterizirati kao nagli ili postupni. Također, pomak se može dogoditi za stalno ili se različite distribucije mogu izmjenjivati tijekom vremena. Ako se dogodi promjena u vjerojatnosti podataka o kojima distribucija ovisi, govorimo o **virtualnom pomaku**, a ako se dogodi u samoj distribuciji, to je **pravi pomak**. Obje vrsta pomaka utječu na model na različite načine i trebaju se uzeti u obzir.

Za rješavanje pomaka često se koriste algoritmi „online učenja" koji kontinuirano ažuriraju model na temelju novih opažanja. Opažanja se procjenjuju jedno za drugim, a model se u trenu analize nekog novog podatka temelji na prethodnim podacima. Greška svakog predviđanja na temelju novih opažanja računa se odabranom funkcijom gubitka. Tako nastali modeli mogu se periodično procjenjivati testnim skupovima ili testirati prije ažuriranja. Testiranje prije ažuriranja uključuje računanje prosječne greške od početka, davanje veće pozornosti nedavnim greškama i procjenu točnosti na posljednjih k promatranja.

Osim učenja iz svakog opažanja, koristi se i učenje iz više opažanja prikupljenih u nekom vremenskom prozoru ili odabranog uzorka. Prijašnji podaci također se mogu zanemariti u potpunosti prilikom razvoja novog modela ili im se utjecaj umanjuje. Sam model može se razviti bez utjecaja prošlih modela ili se može samo prilagoditi novim podacima, kao što je riječ u online učenju.

Osim periodične prilagodbe modela, koriste se i funkcije za detektiranje pomaka nakon čega započinje prilagodba. Prate se promjene u odabranim mjerama kako bi se odlučilo je li došlo do pomaka i na temelju toga dostupne su dodatne informacije pri odabiru načina prilagodbe. Izgradnja potpuno novog modela korisna je ako se model mora nositi s naglim i sveobuhvatnim promjenama, pogotovo ako je potreban model već otprije u memoriji. No, postoji opasnost odbacivanja važnih informacija, pa se koristi i pristup lokalne prilagodbe u slučaju manjih promjena.

Metode ansambla kombiniraju više modela i objedinjuju rezultate u zajedničku procjenu. Prilagodba ovakvih modela može doći promjenom pravila kombiniranja, dok bazični modeli ostaju isti. S druge strane, sami bazični modeli mogu se mijenjati već opisanim metodama. Postoji i opcija korištenja većeg broja modela, pri čemu su samo najefikasniji i najviše prilagođeni podacima odabrani za aktivni dio ansambla u svakom trenutku. [28]

Learn++ NSE (non-stationary environment) algoritam primjer je algoritma za promjenjive okoline temeljenog na metodi ansambla. Algoritam pretpostavlja da skupovi podataka za treniranje dolaze u određenim intervalima i za svaki skup stvara novi bazični model. Svakom baznom modelu dodjeljuje se ocjena pogreške na temelju izvedbe, pri čemu veći utjecaj imaju novije procjene. Ukupan rezultat algoritma je objedinjeni rezultat svih modela, gdje neki modeli imaju veći utjecaj ovisno o procjeni, a procjena cijelog modela računa se na najnovijem skupu za treniranje. [12]

4.3 Povezivanje i integracija podataka

Skupovi velikih podataka često dolaze iz raznih izvora čije strukture ne moraju biti skladne ili uopće postojati. Razlike mogu biti i u tipovima podataka ili brzini kojom nastaju, a problem predstavlja i kasnije dodavanje novih izvora. Sustavi za upravljanje podacima moraju pravilno integrirati svaki novi izvor te uskladiti nove podatke s postojećima. Time se održava kvaliteta i pouzdanost podataka koji se obrađuju.

Usklađivanje shema

Shema baze podataka opisuje logičku strukturu baze, uključujući tipove podataka, entitete i njihove veze. Usklađivanje shema odnosi se na pronalaženje elemenata koji predstavljaju iste podatke u dvije baze i njihovo uparivanje. Ovaj zadatak može biti izazovan ako isti atributi imaju različita imena ili ista imena u bazama predstavljaju različite atribute. Isti entiteti mogu imati drugačije atribute ili više atributa u jednoj bazi može predstavljati samo jedan u drugoj.

Tradicionalno usklađivanje shema obavlja se ručno i zahtjevno je pa se pojavila potreba za istraživanjem automatiziranih metoda. U tim metodama svaki promatrani par atributa smatra se kandidatom za uparivanje. Zatim se koriste funkcije uparivanja koje im na temelju kriterija sličnosti dodjeljuju ocjenu od 0 do 1. Mogući kriteriji su razni, od sličnosti naziva i tipa podataka do kardinaliteta i čak samih podataka. **COMA** (COmbining Matching Algorithms) je primjer heurističkog sustava koji koristi matrice izračunate mjere sličnosti za procjenu uparivanja.

Nedostatak heurističkih metoda je u potrebi namještanja da bi se dobili dobri rezultati. [37] Zbog toga se razvoj metoda okrenuo prema strojnom učenju, smatrajući problem uparivanja kao problem klasifikacije između parova prihvaćenih za uparivanje i onih koji nisu. Značajke mogu biti bazirane na specifikaciji, primjerice tip podataka i duljina polja, ili specifične za tip podataka, kao što su ekstremi u numeričkim atributima ili posebni znakovi u poljima znakova. [38]

Razrješavanje entiteta

Zbog različitih zapisa istih podataka u različitim izvorima, pogotovo u slučaju velikih podataka, dolazi do mogućnosti duplikacije. Na primjer, prilikom spajanja raznih kliničkih zapisa iz jedne ili više bolnica za istog pacijenta, ti zapisi mogu postati odvojeni pod dva različita unosa pacijenta u sustav. Razrješavanje entiteta (eng. entity resolution) odnosi se na spajanje zapisa istih osoba ili drugih imenovanih entiteta radi uklanjanja duplikata.

Da bi se otkrilo koji zapisi su duplikati, potrebno je usporediti svaki sa svakim, što predstavlja veliki izazov kada se radi o velikim količinama podataka. **MapReduce** omogućava paralelnu i distribuiranu usporedbu svih mogućih parova. Kako bi se proces dodatno ubrzao, podaci se prvo dijele u blokove sličnih podataka među kojima je vjerojatnije da će se naći duplikati. `Map` funkcija tada uparuje ključ svakog bloka i entitet, a `reduce` funkcija spaja iste entitete unutar blokova. [24]

Spajanje podataka

Spajanje podataka (eng. data fusion) uključuje kombiniranje više izvora podataka kako bi se dobili kvalitetniji i relevantniji podaci. U okruženju Interneta stvari koristi se za profinjenje podataka iz mnoštva, često jeftinih, senzora. Sustavi za spajanje podataka često se dijele prema Dasarathyjevoj klasifikaciji. Sastoji se od pet kategorija, pri čemu ulazni podaci mogu biti sirovi podaci koji se pretvaraju u kvalitetnije sirove podatke ili značajke entiteta, značajke koje se pretvaraju u relevantnije značajke ili skup odluka te same odluke koje se poboljšavaju procesom spajanja. [20]

Iako centralizirane arhitekture pružaju najtočnije rezultate u teoriji, često se zbog količine podataka koje je potrebno obraditi koriste decentralizirane i distribuirane arhitekture ili kombinacija obje vrste, zvana **hijerarhijska arhitektura**. **Asocijacija podataka** nastoji svakom senzoru pridružiti njegova mjerenja kroz vrijeme radi smanjena šuma. Za rješavanje problema asocijacije koriste se razne metode poput k-sredina, različitih probabilističkih modela i višestrukog testiranja hipoteza.

Procjena stanja nastoji otkriti stanje cilja u pokretu za koji ne postoje direktna mjerenja na temelji obližnjih opažanja. Za pronalaženje vrijednosti vektora stanja, s atributima kao položaj, brzina i veličina, koriste se metode poput Kalmanovog i drugih filtra ili konzistentnosti kovarijance. Na kraju je potrebno donijeti odluku na temelju prikupljenog znanja o promatranjima. Neke od korištenih metoda temelje se na Bayesovom teoremu, pri čemu se nesigurnost među podacima predstavlja uvjetnim vjerojatnostima.

Poglavlje 5

Analiza svojstava velikih podataka koristeći podatke s društvenih mreža

Twitter ili, kako se odnedavno zove, X je popularna društvena mreža namijenjena za slanje kratkih poruka i objava (eng. microblogging). Poruke, zvane „tweetovi", sadrže najviše 280 znakova koje korisnik unosi i šalje na Twitterov poslužitelj koji dalje prosljeđuje poruku listi drugih korisnika, odnosno pratitelja autora poruke. Korisnici mogu komunicirati i putem spominjanja korisničkog imena nakon znaka „@" ili putem komentarnja objava. Osim toga, „hashtagovi", teme koje su označene znakom „#" na početku, služe za uključivanje u veće rasprave koje mogu zahvatiti korisnike iz cijelog svijeta. Omogućeno je i pretraživanje tweetova prema pojmovima ili upotrebom hashtagova.

Slanje prvog tweeta 21. ožujka 2006. godine može se smatrati početkom Twittera. Nakon toga je uslijedio brzi rast broja korisnika, pomognut rastom utjecaja slavnih osoba i političkim kampanjama. Tijekom iranskih predsjedničkih izbora 2009. godine cenzura tradicionalnih medija utemeljila je Twitter kao platformu za objavu vijesti. Pokazao se i kao platforma za povezivanje žrtava i donatora tijekom katastrofalnog potresa koji je pogodio Haiti 2010. godine. [16]

Svestranost tema, broj i rasprostranjenost korisnika te donedavno lak pristup podacima razlozi su zbog kojih se upravo Twitter koristio kao izvor podataka za mnoga istraživanja. Za ograničeni pristup podacima potrebno je bilo napraviti korisnički račun te zatražiti inženjerski pristup uz osobni ključ i lozinku. Akademski pristup nudio je dodatne privilegije, poput detaljnijeg pretraživanja i većeg broja tweetova koji su se mogli skinuti, uz dulji proces odobravanja. Pristup je otežan nedavnim promjenama, no mnogi prikupljeni skupovi još uvijek postoje u internetskim arhivama, poput UNT i GWU digitalnih arhiva korištenih u ovom radu.

5.1 Tweet objekt

Tweet je osnovni objekt unutar Twittera koji sadrži sve informacije o nekoj objavi. Zapisan je u JSON formatu i sadrži više atributa jednostavnih i složenih tipova podataka. Složenim tipovima pripadaju objekti koji sadrže informacije o korisnicima ili drugim tweetovima. U sljedećoj tablici dani su atributi važni za daljnju analizu.

Naziv	Tip	Opis
created_at	String	Vrijeme nastanka tweeta u UTC vremenskoj zoni.
id	Int64	Jedinstveni identifikator tweeta.
full_text	String	Tekst zapisan UTF-8 kodiranjem.
entities	Entities	Objekt koji sadrži hashtagove („#tema”), spominjanja („@korisničko_ime”), posebne simbole, poveznice i multimedijske sadržaje prisutne u tweetu.
user	User	Objekt koji sadrži informacije o korisničkom profilu autora tweeta.
retweeted_status	Tweet	Atribut koji se pojavljuje samo ako je tweet ponovna objava drugog tweeta. Sadrži originalni tweet.
is_quote_status	Boolean	Varijabla koja je istinita samo ako tweet citira drugi tweet.
retweet_count	Int	Broj ponovnih objava tweeta.
lang	String	Detektirani jezik unutar „full_text” atributa.
quoted_status	Tweet	Atribut koji se pojavljuje samo ako tweet citira drugi tweet. Sadrži originalni tweet.

Tablica 5.1: Atributi tweeta

Kategorije tweetova

Uz vlastite objave tweetova, Twitter pruža mogućnost korištenja tweetova drugih korisnika. Zbog složenosti usklađivanja analize takvih objava s vlastitima, dalje će se smatrati zasebnim kategorijama:

- **Ponovna objava** tweeta (eng. retweet) odnosi se na dupliciranje tweeta drugog autora kako bi se pojavio na listi vlastitih tweetova. Svaka ponovna objava u potpunosti prenosi sadržaj originalnog tweeta, uključujući i oznaku autora. Kako bi se ponovna objava razlikovala od vlastitih, u gornjem lijevom kutu nalazi se prepoznatljiva ikona, a unutar objekta tekst će na početku imati dodana slova „RT”. Također, pojavit će se atribut `retweeted_status` s originalnim tweetom.

- **Citiranje** tweeta (eng. quote tweet) razlikuje se od ponovne objave jer omogućuje komentiranje originalnog tweeta. Tekst komentara postaje novi fokus tweeta, dok je originalni tweet prikazan ispod u smanjenom obliku. Umjesto oznake „RT", atribut `is_quote_status` označava radi li se o citiranom tweetu. Tekst komentara smatra se tekstom citiranog tweeta, a original je zapisan pod `quoted_status`.

5.2 Odabir skupova podataka

Politika Twittera omogućuje dijeljenje prikupljenih skupova tweetova u istraživačke svrhe, no bez ograničenja smiju se dijeliti samo identifikatori prikupljenih tweetova. Iako to još uvijek zahtijeva prikupljanje podataka pomoću programskog sučelja, moguće je iskoristiti napredno pretraživanje i velike količine već filtriranih tweetova. Veći nedostatak je u tome što se tweetovi i korisnički računi mogu izbrisati prije nego što se ponovno skinu, pa se dio podataka izgubi. U ovom radu odabrana su dva skupa iz arhive Sveučilišta George Washington (GWU) i jedan iz arhive Sveučilišta Sjevernog Teksasa (UNT).

Twitter se pokazao kao platforma za dijeljenje vijesti pa su mnoge i prikupljene za analizu. Jedan od takvih skupova sadrži više od 160 milijuna tweetova prikupljenih od preko 9000 različitih korisničkih računa koji pripadaju medijima za prenošenje vijesti i tijekom duljeg vremenskog perioda. Zbog velikog broja tweetova, skup (identifikatora) je podijeljen na 4 dijela te se dalje za analizu koristi zadnji dio od oko 10 milijuna tweetova. Skup se nalazi na stranicama arhive GWU, zajedno sa sljedećim skupom.

XXIII. Zimske olimpijske igre održane su 2018. godine u gradu PyeongChangu u Južnoj Koreji. Igre su bile značajne zbog napetosti između Sjeverne i Južne Koreje u tom razdoblju. Neke države izrazile su zabrinutost oko sigurnosti i prijetile da neće nastupiti. Na kraju je pronađen kompromis te su dvije Koreje imale i zajednički tim u jednoj disciplini. Također, ovo su bile prve Olimpijske igre nakon teških sankcija ruskim atletičarima zbog dopinga. [1] Skup tweetova prikupljen na temu ovih Olimpijskih igara sadrži skoro 14 milijuna identifikatora.

Uragan Harvey bio je katastrofalan uragan 4. kategorije koji je pogodio Texas i Louisianu u kolovozu 2017. godine. Oštećeno je oko 200 tisuća domova, a preko milijun ljudi moralo je potražiti sklonište. Procjenjuje se da je uzrokovao materijalnu štetu od 125 milijardi dolara, čime prestiže sve oluje osim Katrine. [6] Arhiva UNT sadrži skup od 7 milijuna (identifikatora) tweetova nastalih tijekom odvijanja uragana.

Tweetovi o Olimpijskim igrama i uraganu većinom su osobni doživljaji autora ili opažanja o trenutnim događajima i kretanju uragana. Takvi podaci se teško mogu pro-

vjeriti i često su subjektivni. Zbog toga su prikladni za analizu sentimenta, no provjera istinitosti je otežana. S druge strane, skup vijesti lako se provjerava pretragom izvora na internetu jer će vijesti objavljene na društvenim mrežama vjerojatno imati dulji članak na nekoj stranici. Vrijednost analize sentimenta vijesti je u promatranju reakcije publike, no odgovori na tweetove nisu dostupni u ovom skupu osim ako i sami ne spominju vijesti i pripadaju nekom od odabranih autora. Zbog ovih razlika skup s vijestima će se koristiti za provjeru vjerodostojnosti, dok će se preostala dva koristiti za analizu sentimenta.

API i hidriranje tweetova

Tweetovi se prikupljaju preko programskog sučelja uz jedinstveni ključ i lozinku danu korisničkom računu koji to zatraži. Tijekom prikupljanja podataka najjednostavniji pristup je omogućavao ograničenu pretragu i skidanje 500000 tweetova mjesečno. Ta ograničenja mogla su se zaobići sa skupom već filtriranih identifikatora i procesom hidriranja. Hidriranje tweeta uzima identifikator nekog tweeta i skida tweet kojemu identifikator pripada. Hydrator je jednostavna aplikacija koja omogućuje hidriranje uz osobni ključ i lozinku. Prima tekstualnu datoteku identifikatora i zapisuje tweetove u JSON ili CSV datoteku. Jedino ograničenje je bila brzina skidanja, tj. koliko tweetova se može skinuti svakih 15 minuta.

Tweetovi se ne mogu skinuti ako su uklonjeni, ako je korisnički profil autora uklonjen (ili suspendiran) te ako su promijenjene postavke privatnosti. Nakon skidanja tweetova u 2. i 3. mjesecu ove godine, skup Harvey zadržao je samo 58% tweetova, a skup Olimpijske igre 55%. Skup Vijesti, zbog naknadnog čišćenja skupa, izgubio je samo 3% tweetova.

Važno je napomenuti da je ranije ove godine promijenjeno programsko sučelje, odnosno pristup Twitteru za skidanje i objavljivanje tweetova. Direktor Twittera naveo je veliki broj botova kao razlog promjena. Besplatni pristup i prva plaćena opcija više ne pružaju mogućnost skidanja tweetova, a i sve osim komercijalnog pristupa ima veća ograničenja od ranije najjednostavnijeg pristupa. Zbog ovih promjena postupak hidracije više nije praktičan.

5.3 Pretprocesiranje

Python je odabran kao programski jezik za obradu tweetova zbog razvijenih alata za obradu prirodnog jezika. Prije svega, NLTK (Natural Language Toolkit) svestrani je alat otvorenog koda koji objedinjuje razne tehnike obrade. Omogućava i klasifikaciju teksta strojnim učenjem. Drugi alati uključuju autocorrect, za automatsko ispravljanje teksta, i VADER,

za analizu sentimenta. Podaci su prije obrade učitani u pandas dataframe unutar Jupyter Notebook aplikacije.

Za analizu su se koristili samo tweetovi na engleskom jeziku. Razlog tomu je treniranje i upotreba modela za analizu na jedinstvenom skupu radi poboljšanja kvalitete, kao i upotreba već pripremljenih alata za engleski jezik. Također, engleski jezik je odabran jer ga za komunikaciju koristi najveći broj korisnika. Svaki skup podataka je filtriran pomoću atributa `lang` i oznake „en". Skup Harvey sadrži 68.7% tweetova na engleskom, Olimpijske igre 55.4%, a Vijesti 64.49%.

Slijedi podjela tweetova na obične, ponovne objave i citirane tweetove. Izdvojeni su tweetovi čiji tekst počinje s „RT" i čija je varijabla `is_quote_status` istinita te kojima su atributi `retweeted_status` ili `quoted_status` neprazni. Pojedini tweetovi izdvojeni su pomoću kriterija `retweeted_status` atributa, ali nisu imali oznaku „RT". Pokazalo se da takvi tweetovi pripadaju suspendiranim korisnicima i sadrže samo poruku o suspenziji pa su uklonjeni. Također su odbačeni tweetovi koji su ispunjavali samo jedan od kriterija za prepoznavanje citiranog tweeta.

Skup	Ukupno (na eng.)	Tweetovi	Ponovne objave	Citati	Odbačeno
Harvey	2808333	14.46%	76.82%	2.5%	6.21%
OI	4208374	17.02%	71.1%	4.64%	7.24%
Vijesti	6685101	84.63%	13.69%	0.72%	0.96%

Tablica 5.2: Podjela skupova podataka po kategorijama tweetova (na eng. jeziku)

Atributi koji nisu bitni za obradu su uklonjeni, pri čemu skup ponovnih objava zadržava `retweeted_status`, a skup citata `quoted_status`. Tekst svakog tweeta se čisti od hashtagova, spominjanja i poveznica upotrebom regularnih izraza. Slijedi uklanjanje interpunkcijskih znakova i stop-riječi (beznačajnih riječi koje se često koriste) te ispravljanje pogrešno napisanih riječi. Na kraju, riječi se pretvaraju u korijenski oblik i svaki tekst se dijeli u zasebne riječi postupkom tokenizacije.

5.4 Istinitost podataka

Detekcija lažnih vijesti i glasina ima iznimnu važnost u današnjem informacijskom okruženju, pogotovo na društvenim mrežama gdje se informacije mogu brzo proširiti. Mnoge informacije sadrže osobna iskustva i mišljenja koja se ne mogu provjeriti. Zbog toga je potrebno provjeriti vjerodostojnost informacija, tj. razlučiti glasine od provjerenih vijesti. Sljedeći dio rada pokazuje provjeru istinitosti podataka na temelju takve klasifikacije, po uzoru

na [19]. Za analizu je odabran naivni Bayesov klasifikator uz neke značajke korištene u spomenutom radu.

Naivni Bayesov klasifikator

Naivni Bayesov klasifikator je model strojnog učenja koji prvenstveno služi za klasifikaciju podataka izračunom vjerojatnosti pripadanja nekoj kategoriji. Odlikuje se jednostavnom implementacijom, skalabilnošću i brzinom ($O(n \cdot d)$, gdje je d broj značajki), te je iz tih razloga izabran za ovaj rad. Temelji se na sljedećem teoremu:

Teorem 5.4.1 (Bayesov teorem)**.** *Neka je H_i, $i = 1, 2, \ldots, n$ potpun sustav vjerovanja u vjerojatnosnom prostoru $(\Omega, \mathcal{F}, P)$, te neka je događaj $A \in \mathcal{F}$ t. d. je $P(A) > 0$ proizvoljan. Tada vrijedi:*

$$P(H_i|A) = \frac{P(H_i)P(A|H_i)}{\sum_{j=1}^{n} P(H_j)P(A|H_j)}. \tag{5.1}$$

Ako je A posljedica događaja ili u ovom slučaju skup vrijednosti značajki koje opisuju događaj ili podatak, a H_i oznaka klasifikacije, Bayesova formula (5.1) daje vjerojatnost klasifikacije uz dane vrijednosti značajki. „Naivna" pretpostavka klasifikatora je da su sve značajke međusobno nezavisne uz dani H_i. Neka je $A = (a_1, \ldots, a_n)$ vektor vrijednosti značajki koji opisuje neki podatak. Pod pretpostavkom nezavisnosti vrijedi $P(A|H_i) = \prod_{j=1}^{n} P(a_j|H_i)$. Formula se dalje pojednostavljuje uklanjanjem nazivnika budući da ostaje isti za sve H_i i nema utjecaja na relativne vjerojatnosti. Konačno, dobiva se sljedeća proporcionalnost:

$$P(H_i|A) \propto P(H_i) \cdot \prod_{j=1}^{n} P(a_j|H_i). \tag{5.2}$$

Nakon što se izračunaju vjerojatnosti pripadanja svakoj skupini, za klasifikaciju se uzima ona s najvećom vjerojatnosti.

Slijedi primjer korištenja klasifikatora za raspoznavanje glasina od provjerenih vijesti objavljenih na Twitteru. Tweetovi korisničkih računa raznih medija podijelit će se na **„glasine"** i **„provjerene vijesti"**. Glasinama se ne smatraju samo one same, već i osobna mišljenja, predviđanja (poput kretanja tržišta ili inovacija) i pokušaja drugačije interakcije s publikom (primjerice pitanja o tome kako su proveli dan). Provjerene vijesti su one kojima se može naći izvor i utvrditi istinitost, pri čemu je naglasak na sadržaju tweeta, a ne na člancima s kojima su povezani.

Značajke

Značajke podataka s društvenih mreža mogu se podijeliti na dvije skupine: značajke vezane uz autora i značajke vezane uz sadržaj. U idućim tablicama detaljno su opisane značajke korištene za klasifikaciju. Posebno treba napomenuti „provjerene" korisničke račune. U prošlosti je postojao sustav provjera kako bi se dokazalo da se stvarno radi o navedenoj osobi. Danas se taj status može dobiti uz pretplatu pa je ova značajka izgubila vrijednost. Budući da su tweetovi iz 2018. godine, o ovom slučaju ta oznaka ipak vrijedi.

Za svaki tweet prvo su izračunate vrijednosti značajki. Značajke vezane uz sentiment klasificirane su pomoću alata za analizu sentimenta predstavljenog u sljedećem poglavlju. Sve vrijednosti prikupljaju se u rječniku i predaju kao ulaz modelu za klasifikaciju. Korišten je naivni Bayesov klasifikator dostupan unutar NLTK-a koji prima rječnik značajki i njihovih vrijednosti za neki podatak te računa distribuciju vjerojatnosti po značajkama za danu kategoriju klasifikacije. Izračun vjerojatnosti temelji se na prethodnom treniranju klasifikatora na već označenom skupu podataka. Izlaz funkcije `classify` je kategorija klasifikacije s najvećom vjerojatnosti.

Značajka	Opis
Sentiment tweeta	Ukupni sentiment izražen u tekstu.
Broj negativnih riječi	Broj negativnih riječi u tekstu tweeta.
Broj pozitivnih riječi	Broj pozitivnih riječi u tekstu tweeta.
Tweet ima pozitivne emocije	Postoji li pozitivan dio teksta?
Tweet ima negativne emocije	Postoji li negativan dio teksta?
Vremenska razlika	Razlika vremena nastajanja tweeta i računa autora.
Omjer interpunkcije i broja riječi	Broj rečeničnih znakova / broj riječi u tekstu tweeta.
Ima li točki?	Postoji li točka na kraju izjave?
Broj riječi	Broj riječi u tekstu tweeta (bez poveznica, hashtagova i sl.).
Ima li URL?	Postoji li poveznica u tekstu tweeta?
Ima li hashtag?	Postoji li hashtag u tekstu tweeta?
Broj hashtagova	Broj hashtagova u tekstu tweeta.
Ima li upitnik?	Postoji li upitnik u tekstu tweeta?
Broj upitnika	Broj upitnika u tekstu tweeta.
Ima li uskličnika	Postoji li uskličnik u tekstu?
Broj uskličnika	Broj uskličnika u tekstu tweeta.
Je li tweet ponovno objavljen?	Ima li tweet zabilježenu ponovnu objavu.
Broj ponovnih objava	Broj svih ponovnih objava tweeta.

Tablica 5.3: Značajke vezane uz sadržaj

Značajka	Opis
Omjer broja pratitelja i starosti profila	Broj pratitelja / starost profila.
Omjer prijatelja i starosti profila	Broj prijatelja / starost profila.
Omjer broja statusa i starosti profila	Broj statusa / starost profila.
Broj favorita	Broj svih tweetova koji su se sviđali autoru.
Omjer pratitelja i prijatelja	Broj pratitelja / Broj prijatelja.
Broj lista	Broj popisa na kojima se autora nalazi.
Ima li opis?	Ima li autor uređeni opis profila?
Duljina opisa	Duljina uređenog opisa.
Duljina imena	Duljina imena koje se pokazuje drugim korisnicima.
Ima li URL?	Ima li profil poveznicu u opisu?
Je li račun provjeren?	Ima li račun oznaku da je autor provjeren?
Ima li bazičnu sliku profila?	Je li označen atribut `default_picture`?

Tablica 5.4: Značajke vezane uz autora

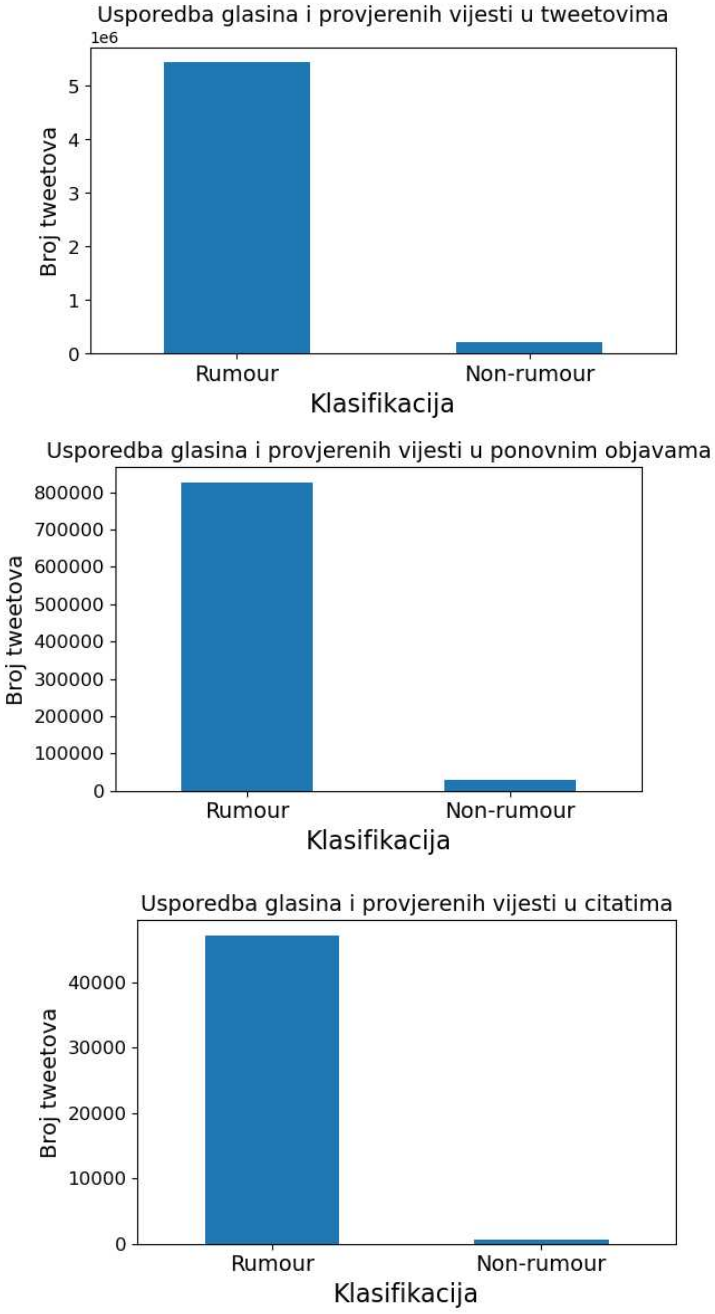
Model naivnog Bayesovog klasifikatora treniran je na oko 2000 ručno provjerenih podataka, pri čemu je 1400 poslužilo za treniranje, a 600 za testiranje i poboljšanje. Zadnji test pokazuje točnost od 69%. Najutjecajnijim značajkama pokazali su se prisutnost upitnika, broj uskličnika i upitnika, duljina uređenog opisa profila te omjer interpunkcije i broja riječi.

Rezultati

Analiza je odvojena po kategorijama tweetova. Obični tweetovi normalno su se obradili prema značajkama. Kod ponovnih objava bilo je potrebno izvući originalne tweetove te su se originalni tweet i autor uzimali kao subjekti značajki, dok je ponovna objava smatrana samo prenošenjem vijesti. S druge strane, kod citata je bitno gledati oba autora i obje objave pa je model uzimao kombinaciju značajki.

Rezultati pokazuju daleko više glasina nego provjerenih vijesti u sve 3 kategorije, što je iznenađujuće jer se očekuje da je ta podjela više uravnotežena. Osim samog skupa punog glasina, postoji više objašnjenja vezanih uz sam model. Kao prvo, uzorak za treniranje vrlo je malen naspram ostatka skupa koji se analizira, pa je sama priprema modela upitna. Nadalje, pregledom skupa podataka mogu se naći razni tweetovi gdje je vijest nečija izjava ili mišljenje. Mišljenje pripada glasinama, no vijest je u ovom slučaju to da je određena osoba nešto izjavila, pa se može smatrati provjerenom vijesti, što otežava klasifikaciju. Takvi tweetovi će najviše sličiti glasinama kroz obradu teksta, a već spomenute najutjecajnije značajke u modelu uglavnom se drže baš teksta tweeta. Dodatne značajke koje bi

mogle pomoći pri razlučivanju uključuju povijest korisničkog profila i analizu odgovora na tweet, no to nije dostupno u ovom skupu podataka.



Slika 5.1: Rezultati analize vjerodostojnosti po kategorijama

5.5 Vrijednost podataka

Analiza sentimenta ključna je za izvlačenje vrijednosti iz tekstualnih podataka, pogotovo s društvenih mreža. Pruža razumijevanje mišljenja javnosti o određenim temama ili događajima, što služi identificiranju potencijalnih problema ili prilika radi boljih i pravovremenih odluka. U slučaju upravljanja katastrofama može ukazati na nedovoljnu pripremljenost i loše odgovore, a prilikom odvijanja organiziranih događaja na nezadovoljstvo publike. Na pripremljenim skupovima podataka pokazat će se analiza sentimenta odabranih tema u svrhu procjene pripremljenosti i trenutnog odgovora na razvoj događaja. Sentiment o uraganu Harveyju promatrat će se kroz evakuaciju i dostupnost pomoći, a sentiment o Olimpijskim igrama preko elemenata koji trebaju privući pozornost posjetitelja i dostupnost smještaja. Za analizu i klasifikaciju sentimenta koristi se alat VADER.

VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) je alat za analizu sentimenta temeljen na rječniku i skupu pravila te je posebno obučen za analizu sentimenta na podacima s društvenih mreža. Otvorenog je koda i pod MIT licencom te je namijenjen za korištenje u različitim programskim jezicima, poput Jave, Pythona i R-a, i okolinama. Prema autorima projekta [13], vremenska složenost alata je otprilike $O(n)$.

Treniranje modela za prepoznavanje sentimenta na podacima s društvenih mreža uključivalo je elemente poput slenga povezanog s izražavanjem sentimenta, emotikona i izraza koji pojačavaju ili smanjuju intenzitet sentimenta, primjerice „jako" ili „nekako". Alat može prepoznati sentiment na cjelovitom tekstu ocjenjujući svaku riječ zasebno uz objedinjeni rezultat (eng. *compound*). Vrijednost zajedničkog rezultata nalazi se između -1 i 1, gdje je -1 najviše negativno, a 1 najviše pozitivno. Preporuča se mjeriti neutralni sentimenta uključivo između -0.05 i 0.05 te ostale dvije kategorije izvan tog intervala, pa se tako postupalo i u ovom radu. Također su preskočeni neki koraci pri pretprocesiranju zbog mogućnosti alata, a drugi su čak imali loš utjecaj na točnost procjene (npr. pretvorba riječi u korijenski oblik).

Odabir tema

Prilikom odabira tema na kojima će se temeljiti analiza sentimenta odabranih skupova uzeti su u obzir važni aspekti promatranih događaja, kao i prijašnja istraživanja. U slučaju uragana fokus je na neposrednom odgovoru nadležnih organizacija i šire javnosti na katastrofu, dok se Olimpijske igre promatraju s gledišta organizatora u svrhu poboljšanja iskustva posjetitelja. Svaku temu predstavljaju ključne riječi koje se uspoređuju s listama

riječi nastalima tokenizacijom svakog teksta. Cilj je bio odabrati ključne riječi koje dobro opisuju jednu temu i slabo ili nikako ostale.

Uragan Harvey

Analiza sentimenta o uraganu može pružiti važne uvide u pripremljenost odgovornih organizacija i vrijeme odgovora na opasnosti. Po uzoru na prijašnji rad [23] dalje se promatra reakcija stanovništva na prisutnost medicinskih potrepština i pružanje pomoći, evakuaciju i stanje skloništa te opskrbu vodom i hranom. Tweetovi povezani s ovim temama su izdvojeni su iz skupa podataka za analizu. Ključne riječi (na engleskom) korištene za pretragu su sljedeće:

- **Medicinske potrepštine i pomoć:** aid, care, bandage, medicine, drug, hospital, clinic, treatment, doctor, nurse, patient, disease, injury, wound, hurt, infection, illness.
- **Evakuacija i sklonište:** evacuate, transportation, bus, car, road, highway, traffic, stranded, leave, flee, vacate, move, abandon, depart, displace, desert, remain, stay, shelter, refuge, cot, bed, blanket.
- **Voda i hrana:** water, drink, bottle, food, meal, can, meat, soup, fruit, vegetable, cook, feed, hunger, donation.

Većina riječi dobivena je pretragom sinonima i antonima te stranica za doniranje u slučaju katastrofa. Ključne riječi su prije pretrage pretvorene u korijenski oblik kako bi se pokrili svi mogući oblici koji se pojavljuju u tweetovima.

Olimpijske igre

Organizatori Olimpijskih igara mogu analizirati reakciju javnosti kako bi ocijenili uspješnost događaja i bolje pripremili iduće događaje. Ceremonije otvaranja i zatvaranja posebno su važne za privlačenje pozornosti šire javnosti te služe predstavljanju domaćina. Puno novaca ulaže se u sportske objekte pa je njihov dojam bitan uz samu lokaciju. Na zadovoljstvo publike utječu i raspored događaja te cijene karata i smještaja. Ključne riječi za pretragu ovih pojmova su sljedeće:

- **Ceremonije:** ceremony, opening, closing, performance, torch, flame, lighting, parade, presentation, tradition, culture, artist, costume, music, act, coreography, production, fireworks.

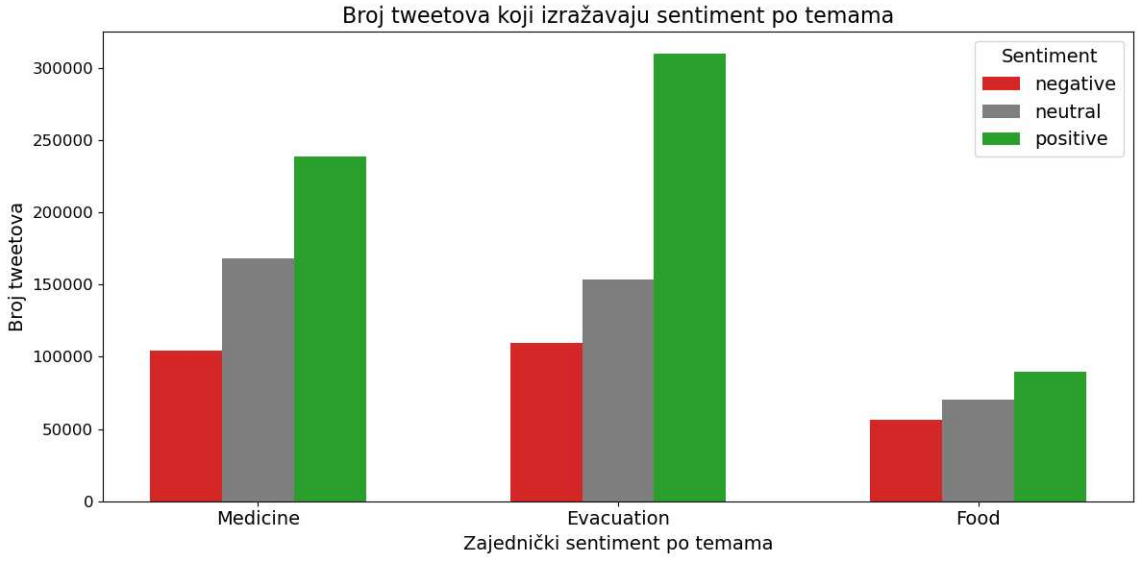
- **Infrastruktura:** infratructure, facility, village, stadium, park, centre, arena, venue, Alpensia, Yongpyong, Phoenix, Jeongseon, Kwandong.
- **Raspored i cijene:** ticket, price, expense, sale, purchase, buy, pay, affordable, booking, hotel, hostel, inn, resort, motel, room, accommodation, lodging, schedule, program, timetable, plan.

Riječi su dobivene pretragom sinomina i povezanih pojmova te uobičajenih događanja tijekom ceremonija. Preuzeti su i nazivi objekata u kojima su se igre održavale. Ključne riječi su ponovno pretvorene u korijenski oblik.

Rezultati

Tijekom analize sentimenta u ovom dijelu rada rezultati iz svih kategorija tweetova su se grupirali u zajedničku ocjenu. Budući da se ne gleda značenje teksta, svaka ponovna objava dodaje novu jedinicu sentimenta za mjeru. Pretpostavlja se da samo ponovno objavljivanje tweeta bez komentara znači da se novi autor slaže s izraženim sentimentom. S druge strane, kod citiranih tweetova promatra se sentiment komentara na originalni tweet.

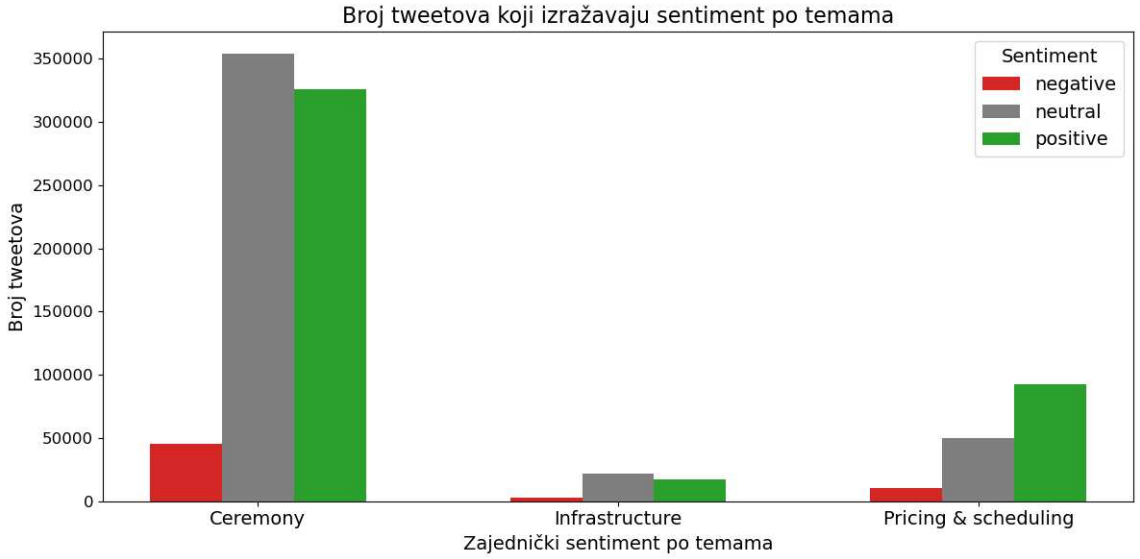
Uragan Harvey



Slika 5.2: Ukupni sentiment po odabranim temama o uraganu

Analiza sentimenta o uraganu Harveyju pokazuje većinom pozitivan ili neutralan sentiment, pri čemu je najviše tweetova pozitivno prema svakoj temi. To je pogotovo slučaj s tweetovima o evakuaciji i skloništu. Na temelju toga može se zaključiti da je odgovor na uragan bio zadovoljavajući te su pravovremeno stigla upozorenja i upute za evakuaciju. Pregledom stvarnih događaja može se vidjeti da, iako je uragan uzrokovao veliku materijalnu štetu, sigurnost u trenutku udara bila je ozbiljno shvaćena te su evakuirana mnoga područja. Nakon katastrofe došle su i mnoge donacije i druga pomoć pa je i tu prevagnuo pozitivan sentiment.

Olimpijske igre



Slika 5.3: Ukupni sentiment po odabranim temama o OI

Reakcija javnosti na ceremonije otvaranja i zatvaranja pokazuje najviše neutralnih tweetova, no skoro toliko i pozitivnih. Negativno ocijenjenih tweetova ima vrlo malo u odnosu na te dvije kategorije. Iz toga se može zaključiti da, iako nisu toliko oduševile javnost, ceremonije su zaslužile puno više pozitivnih reakcija nego negativnih. Dublje istraživanje moglo bi otkriti koju su točno dijelovi ceremonija privukli gledatelje, a koji nisu. Slična situacija je i s infrastrukturom, no broj tweetova je puno manji. Jedan od mogućih razloga je priroda samih ceremonija. To su spektakli koji su dostupni čitavoj publici, čak i ljudima koji nisu zainteresirani u sportski dio Igara.

Za razliku od prethodne dvije skupine, ocjena pristupačnosti pokazuje dosta pozitivnu reakciju, što da naslutiti da nije bilo većih problema s dolaskom i smještajem turista te ras-

poredom. Potrebno je napomenuti da su se analizirali samo tweetovi na engleskom jeziku pa su reakcije lokalnog stanovništva i turista koji ne koriste Twitter ili govore engleski izgubljene. Zaključno, igre se mogu smatrati turističkim uspjehom, no potrebno je analizirati i utjecaj na lokalno stanovništvo. Također, potrebna je detaljnija analiza ceremonija kako bi se ispravili nezadovoljavajući dijelovi za buduće Olimpijske igre.

5.6 Promjenjivost podataka

Zimske Olimpijske igre u Pyeongchangu popratile su dva kontroverzna razvoja događaja, zajednički nastup Sjeverne i Južne Koreje te sankcije prema ruskim atletičarima zbog sustavnog dopingiranja. Kako su se ovi događaji razvijali, sentiment javnost prema upletenim organizacijama i ljudima mijenjao se s novim otkrićima. Cilj analize podataka u ovoj sekciji je pratiti promjenjivost sentimenta kroz mjesec odvijanja Igara i pronaći uzroke promjena pregledom novinskih članaka (npr. The Guardian). Kako bi se izolirali tweetovi specifično o upletenim strankama, među temama navedenim u hashtagovima potražit će se najbitniji akteri, kao i riječi koje predstavljaju temu. Sentiment će u ovom dijelu analize biti predstavljen numerički, gdje 1 predstavlja pozitivan sentiment, 0 neutralan i -1 negativan. Podaci su grupirani po datumima, a za svaku grupu računa se prosječni sentiment.

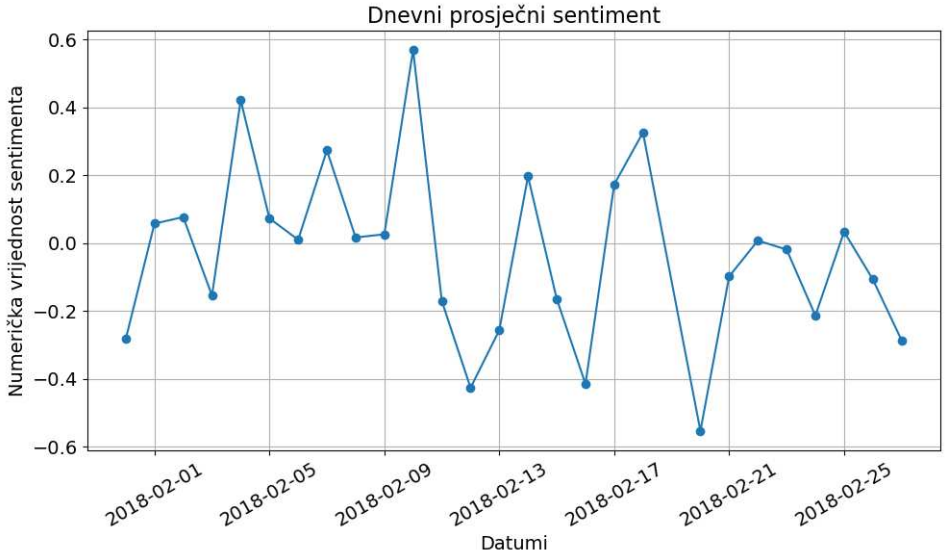
Napetosti između Sjeverne i Južne Koreje

Sukob Sjeverne i Južne Koreje eskalirao je 2017. godine testiranjem nuklearnog i drugog oružja u Sjevernoj Koreji. Zbog toga se pojavila zabrinutost oko sigurnosti natjecanja toliko blizu Sjevernoj Koreji, a neke reprezentacije su čak prijetile bojkotirati Igre. Dok su odnosi s drugim državama još uvijek bili napeti, južnokorejski predsjednik Moon Jae-In počeo je pregovore sa Sjevenom Korejom. Rezultat je bio smirivanje odnosa i zajednički nastup ma Olimpijskim igrama. Pojmovi za pretraživanje relevantnih tweetova bili su sljedeći:

- „NorthKorea” - jedna od upletenih država i glavni izvor napetosti.
- „PKR” - oznaka pod kojom Sjeverna Koreja nastupa na OI.
- „COR” - oznaka ujedinjenog nastupa obje Koreje.
- „KimJongUn” - vrhovni vođa Sjeverne Koreje.
- „MoonJaeIn” - predsjednik Južne Koreje.
- „UnifiedKorea” - poziv za ujedinjenje dviju država.

- „diplomacy” - važan dio Igara s predstavnicima sjevera i juga.
- „KimYoJong” - sestra vrhovnog vođe koja je posjetila Igre.

Rezultati



Slika 5.4: Sentiment prema ujedinjenju i zajedničkom nastupu Koreja

Sentiment je na početku bio pozitivan, vjerojatno zbog vijesti o poboljšanju odnosa i zajedničkog nastupa političara. Rast sentimenta 4. veljače mogao bi se pripisati objavi ujedinjenog tima za ženski hokej koji će nastupati pod zajedničkom oznakom COR. Ovo je bio konkretni korak prema zajedničkom nastupu na Olimpijskim igrama, kao i simbolički korak prema ujedinjenju. 7. veljače objavljeno je da će Kim Yo-jong prisutstvovati ceremoniji otvaranja. Njezino sudjelovanje smatralo se pozitivnim signalom o odnosima između dviju političkih strana. Najveće uzbuđenje je uoči prve igre hokejaškog tima 10. veljače.

Odmah nakon toga, sentiment naglo pada, vjerojatno zbog teškog poraza tima. Još niža vrijednost zabilježena je 12. veljače nakon drugog jednako teškog poraza. Sentiment je nastavio rasti i padati oko idućih nastupa tima i daljnjih poraza, pri čemu je poraz za zadnje mjesto 20. veljače uzrokovao najnegativniji sentiment. Tim je najavljen kao simbol ujedinjenja i boljih odnosa pa je osim razočaranja u uspjeh domaćina na natjecanju potaknuo skepticizam prema političkim obećanjima i propagandi o ujedinjenju. Zbog toga je mišljenje javnosti o Sjevernoj Koreji i ujedinjenju nakon natjecanja ponovno postalo negativno. Potrebno je naglasiti da se radi o sentimentu internacionalne javnosti te bi analiza reakcije u Južnoj Koreji mogla dati drugačiji rezultat.

Ruski doping skandal

Otkrivanje sustavnog dopingiranja ruskih atletičara uzrokovalo je burne reakcije javnosti i zahtjev za teškim sankcijama ruskih natjecatelja na Olimpijskim igrama. Prve slutnje pojavile su se 2010. godine preko zaposlenika Ruske agencije za anti-doping. Uslijedile su daljnje optužbe atletičara te je 2014. godine objavljen dokumentarni film „Tajna dopinga: kako Rusija stvara svoje pobjednike". Film je privukao pozornost javnosti i otvorena je službena istraga u aktivnosti odgovornih organizacija u Rusiji. Nakon opsežne istrage, Međunarodni olimpijski odbor uveo je službene sankcije 2017. godine, uključujući poništavanje 13 medalja. Ruskom olimpijskom odboru zabranjeno je sudjelovati u Olimpijskim igrama 2018. godine, no atletičarima koji su mogli dokazati svoju nevinost ipak je dopušteno sudjelovanje pod nadzorom Međunarodnog olimpijskog odbora. [5] Odabrani pojmovi za pretragu su bili sljedeći:

- „Russia", „doping" - riječi koje opisuju temu i uvode Rusiju kao jednog od aktera događaja.
- „IOC" - engleski akronim za Međunarodni olimpijski odbor.
- „WADA" - engleski akronim Svjetske antidopinške agencije.
- „CAS" - engleski akronim Sportskog arbitražnog suda na kojem se Rusija žalila na sankcije Olimpijskog odbora.
- „OAR" - oznaka za „olimpijske natjecatelje iz Rusije". Skupina je osnovana kao kompromis Olimpijskog odbora i Rusije omogućujući atletičarima koji mogu dokazati nevinost nastup pod olimpijskom zastavom.

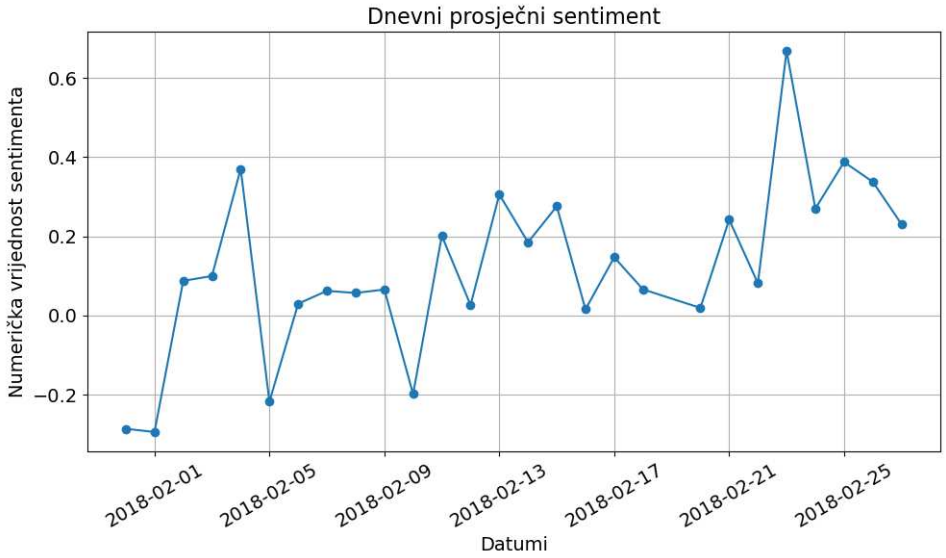
Rezultati

Pregledom vremenskog niza podataka odmah se nameće 1. veljače kao najniža točka prosječnog sentimenta. Pregledom vijesti otkriva se da je na taj dan CAS odbacio doživotne olimpijske zabrane 28 ruskih atletičara. Šira javnost, već upoznata s optužbama, dokazima i sankcija, negativno je reagirala na oslobađanje netom prije Olimpijskih igara. Međunarodni olimpijski odbor našao se pod pritiskom jer je već prije kritiziran zbog neaktivnosti u slučaju dopinga.

Odnosi dviju strana pogoršali su se odlukom Olimpijskog odbora da 13 oslobođenih atletičara ipak ne može nastupati. Kritizirana je odluka CAS-a i tražilo se objašnjenje razloga za uklanjanje zabrana. Te izjave su do javnosti došle 5. veljače, koji na grafu ponovno

zadire u negativne vrijednosti. Sentiment ostaje jednoličan dok ponovno ne postane negativan. Razlog tomu je vjerojatno kritika sankcija ruskog atletičara koji je upravo osvojio brončanu medalju i tako bio u središtu pozornosti.

Uz manje oscilacije sentiment prema Rusiji i navedenim organizacijama ostaje pozitivan, no pravi pomak dogodio se 23. veljače. To je ujedno i najveće odstupanje vrijednosti sentimenta od nule, i to u pozitivnom smjeru. Razlog tomu je senzacionalna pobjeda petnaestogodišnje Aline Zagitove u umjetničkom klizanju, čime je OAR osvojio prvu zlatnu medalju. Nakon tog uzbuđenja, reakcije javnosti se smiruju do kraja natjecanja. No, sentiment prema akterima ostaje pozitivan, suprotno od javnog mišljenja na početku mjeseca. Unatoč skandalu i daljnjim kontroverzama poput novih dokaza dopinga i nakon oslobađanja od sankcija, uzbudljiva natjecanja i osvajanje medalja prevagnuli su u korist uključenih stranki.



Slika 5.5: Sentiment prema ruskim atletičarima i sankcijama

Poglavlje 6

Zaključak

Veliki podaci predstavljaju ogromne količine podataka koje se prikupljaju iz različitih izvora, uključujući uređaje koje svakodnevno koristimo, društvene mreže, medicinske slike i senzore. Karakteriziraju se svojim svojstvima, od kojih su tri stečena kroz osnovna svojstva: istinitost (eng. *veracity*), vrijednost (eng. *value*) i promjenjivost (eng. *variablity*). Istinitost podataka odnosi se na kvalitetu i vjerodostojnost podataka, vrijednost na korisne uvide koji se pronalaze analizom podataka, a promjenjivost na promjene u podacima, njihovoj obradi i izvorima. Svrha ovog rada bila je pregled i analiza stečenih svojstava velikih podataka.

Podaci za obradu odabrani su s Twittera i donose posebne izazove i karakteristike. Odabrani su skupovi podataka o uraganu, Olimpijskim igrama i vijestima zbog različitih područja analize. Podaci su preuzeti s internetskih arhiva i uz postupak hidracije. Korišteni su samo tweetovi na engleskom jeziku i podijeljeni su na kategorije prema obliku radi lakše analize, a pretprocesiranje je još uključivalo i različite elemente obrade prirodnog jezika.

Istinitost podataka analizirala se kroz vjerodostojnost vijesti objavljenih na Twitteru, pri čemu su tweetovi podijeljeni na glasine i provjerene vijesti. Za klasifikaciju je korišten naivni Bayesov klasifikator treniran na ručno provjerenim podacima, a značajke prema kojima je provedena klasifikacija objedinjavale su analizu teksta i drugih zabilježenih podataka o tweetu i autoru. Rezultati analize pokazuju da postoji znatno više glasina nego provjerenih vijesti u svim kategorijama tweetova. Budući da je tolika razlika malo vjerojatna, nameće se zaključak da je model nedostatan za analizu odabranog skupa podataka.

Mali uzorak za treniranje modela, odabir značajki te prisutnost tweetova čije vrijednosti značajki mogu pomutiti klasifikaciju mogući su uzroci pogrešne klasifikacije. Također, sama klasifikacija je bila binarna, u glasine i provjerene vijesti, no već u samom

opisu onoga što se smatra glasinom, odnosno informacijom koja se ne može provjeriti, vidljive su mnoge nijanse koji bi trebalo dublje istražiti. Pogrešan rezultat možda bi se ispravio detaljnijom podjelom gdje bi podaci koji su izazivali poteškoće bili u zasebnoj kategoriji.

Vrijednost podataka istražena je analizom sentimenta po odabranim temama kroz dva događaja: uragan Harvey i Olimpijske igre. Cilj analize bio je procijeniti reakciju javnosti na ove događaje te ocijeniti pripremljenost i trenutni odgovor na njih. Ključne riječi vezane uz svaku temu korištene su za pronalaženje tweetova povezanih uz odabrane teme. U slučaju uragana Harvey, analiza sentimenta pokazala je većinom pozitivan ili neutralan sentiment, sugerirajući da je odgovor na uragan bio uglavnom zadovoljavajući te da su pravovremeno stigla upozorenja i upute za evakuaciju. Također, veliki broj pozitivnih tweetova ukazuje na solidarnost i donacije neposredno nakon katastrofe. Ovakva analiza slaže se sa stvarnim događajima pa se pokazala kao dobar pristup.

Kada su u pitanju Olimpijske igre, ceremonije otvaranja i zatvaranja te infrastruktura su većinom ocijenjene neutralno ili pozitivno, dok je negativnih tweetova bilo vrlo malo. Uz daleko najveći broj tweetova o toj temi, izgleda da su ceremonije privukle pozornost gledatelja i bile uglavnom prihvaćene, no količina neutralnih tweetova ukazuje da bi se pojedini dijelovi mogli poboljšati za bolju reakciju publike. Pristupačnost igrama također je ocijenjena pozitivno, pokazujući da nije bilo većih problema s dolaskom i smještajem turista te rasporedom događanja.

Promjenjivost podataka praćena je kroz promjene u sentimentu tijekom dva kontroverzna razvoja događaja na Olimpijskim igrama u Pyeongchangu: odnosi Sjeverne i Južne Koreje te ruski doping skandal. Promjene u sentimentu su promatrane oscilacijama u dnevnom prosječnom sentimentu, a razlozi promjena tražili su se u individualnim događajima zabilježenima u novinskim člancima. Analiza je bila usredotočena na temu obrade te odgovorne osobe i organizacije, pa su povezani tweetovi traženi putem hashtagova.

Analiza pokazuje da su na sentiment utjecale objave važnih događaja i rezultati natjecanja. Napetosti između Sjeverne i Južne Koreje čine se privremeno razriješenima objavom zajedničkom nastupa i posjeta političara, s početnim optimizmom koji je postupno zamijenila razočaranost zbog loših rezultata zajedničkog tima. Ruski doping skandal izazvao je većinom negativan sentiment na početku, no senzacionalna pobjeda ruske natjecateljice na kraju je rezultirala pozitivnim sentimentom prema akterima. Ovi rezultati ukazuju na kompleksnost percepcije javnosti tijekom velikih sportskih događaja te na važnost sportskih uspjeha u stvaranju pozitivnog sentimenta.

6.1 Nedostaci i poboljšanja korištenih metoda

Dani primjeri obrade sva tri svojstva podataka od interesa pružili su pregled jednostavnih tehnika i tehnologija za obradu. Prije svega sama količina podataka bila je ograničena brzinom obrade i skidanja podataka. Oslanjajući se na internetske arhive prikupljanje je olakšano, no izbor podataka na arhivama se pokazao dosta jednoličan, s puno skupova o izborima i pandemiji. Više resursa i bolji pristup sirovim podacima obogatili bi istraživanje s novim temama. Python pruža odlične alate za obradu prirodnog jezika, ali je spor jezik i zauzima puno memorije. Promjena programskog jezika i okruženja omogućila bi obradu veće količine podataka, a tu je i Pyspark koju omogućava programiranje u Pythonu uz Apache Spark.

Prilikom pretprocesiranja korišteni su većinom gotovi alati iz NLTK-a i drugih repozitorija na githubu. Ti alati su dobro i relativno brzo radili što su trebali, no pristup podacima je zbog njihove široke upotrebe bio dosta općenit. Upotreba specijaliziranih alata ili izgradnja novih poboljšala bi obradu u nekim područjima. VADER, korišten za analizu sentimenta, pruža mogućnost ručne prilagodbe prema korištenim podacima, ali to nije bilo potrebno u ovom radu. Drugi primjeri poboljšanih alata uključuju specijalizirano automatsko ispravljanje i prilagodbu alata specifičnom slengu koji se pojavljuje u podacima.

Uvedene su kategorije tweetova zbog različitih načina obrade svake grupe. Za pravu automatizaciju obrade potrebno bi bilo osmisliti sustav koji prepoznaje svaku kategoriju i primjenjuje odgovarajuće tehnike prema pravilima. To je već donekle postignuto na primjeru analize sentimenta, gdje je svaka kategorija doprinijela ukupnoj ocjeni sentimenta na svoj način. Također, gledao se samo sentiment teksta bez dodatnih značajki, poput multimedijskih podataka i interakcije s drugim korisnicima, koje vrijedi istražiti u svrhu poboljšanja analize. Analiza podataka iz više jezika može pružiti dodatne uvide u globalno mišljenje o temi, primjerice tijekom obrade Olimpijskih igara, s posebnim naglaskom na ruski i korejski jezik.

Provjera vjerodostojnosti vijesti oslanjala se na Bayesov klasifikator. Iako vrlo koristan i brz, to je jednostavni model podložan trovanju modela. Kompleksniji modeli i duboko učenje nogu pružiti precizniju procjenu. Treniranje i testiranje modela na većem skupu podataka također može poboljšati točnost modela, no tu se javlja mana ljudskog nadzora i količine posla koju to predstavlja. Zbog toga su se neka istraživanja okrenula *crowdsourcingu*, iako tu postoji problem kvalitete.

Zaključno, prikazana analiza pruža jednostavni pregled postojećih metoda i alata za obradu velikih podataka, s puno mogućih poboljšanja i smjerova istraživanja, od kojih su mnogi opisani u prijašnjim poglavljima.

Bibliografija

[1] *2018 Winter Olympics*. https://en.wikipedia.org/wiki/2018_Winter_Olympics, posjećena 6. 9. 2023.

[2] *The Apache Software Foundation*. https://www.apache.org/, posjećena 6. 9. 2023.

[3] *Big Data in Algorithmic Trading*. https://medium.com/analytics-vidhya/big-data-in-algorithmic-trading-bd0bb1f9dfca, posjećena 6. 9. 2023.

[4] *Big Data in Retail: Common Benefits and 7 Real-Life Examples*. https://www.talend.com/resources/smart-retailing/, posjećena 6. 9. 2023.

[5] *Doping in Russia*. https://en.wikipedia.org/wiki/Doping_in_Russia, posjećena 6. 9. 2023.

[6] *Hurricane Harvey*. https://disasterphilanthropy.org/disasters/hurricane-harvey/, posjećena 6. 9. 2023.

[7] *IBM: Topics*. https://www.ibm.com/topics, posjećena 6. 9. 2023.

[8] *Introduction to Time Series Analysis*. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm, posjećena 6. 9. 2023.

[9] *Qubole*. https://www.qubole.com/, posjećena 6. 9. 2023.

[10] *SAS Products & Solutions*. https://www.sas.com/en_us/software/all-products.html, posjećena 6. 9. 2023.

[11] *Support Vector Machine (SVM) Algorithm*. https://www.geeksforgeeks.org/support-vector-machine-algorithm/, posjećena 6. 9. 2023.

[12] *Tech: Adaptive ML*. https://medium.com/analytics-vidhya/adaptive-machine-learning-f4fba2f50bc1, posjećena 6. 9. 2023.

[13] *vaderSentiment*. https://github.com/cjhutto/vaderSentiment, posjećena 6. 9. 2023.

[14] *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025*. https://www.statista.com/statistics/871513/worldwide-data-created/, posjećena 6. 9.2023.

[15] *El Niño*. Hrvatska enciklopedija, mrežno izdanje, 2021. https://www.enciklopedija.hr/natuknica.aspx?id=17758, posjećena 6. 9. 2023.

[16] *X*. Encyclopedia Britannica, 2023. https://www.britannica.com/topic/Twitter, posjećena 6. 9. 2023.

[17] Arnx, A.: *First neural network for beginners explained (with code)*. https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf, posjećena 6. 9. 2023.

[18] Assiri, F.: *Methods for Assessing, Predicting, and Improving Data Veracity: A survey*. ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 9:5–30, 2020.

[19] Azer, M., M. Taha, H. Zayed i M. Gadallah: *Credibility Detection on Twitter News Using Machine Learning Approach*. International Journal of Intelligent Systems and Applications, 13:1–10, lipanj 2021.

[20] Castanedo, F.: *A Review of Data Fusion Techniques*. The Scientific World Journal, 2013:704504, siječanj 2013.

[21] Chui, M., M. Collins i M. Patel: *IoT value set to accelerate through 2030: Where and how to capture it*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/iot-value-set-to-accelerate-through-2030-where-and-how-to-capture-it, posjećena 6. 9. 2023.

[22] Dai, W., I. Wardlaw, Y. Cui, K. Mehdi, Y. Li i J. Long: *Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking*. U Latifi, S. (urednik): *Information Technology: New Generations*, stranice 439–450, Cham, 2016. Springer International Publishing.

[23] Dong, Z., L. Meng, L. Christenson i L. Fulton: *Social media information sharing for natural disaster response*. Natural Hazards, 107, srpanj 2021.

[24] El-Ghafar, R., M. Gheith, A. El-Bastawissy i E. Nasr: *Record linkage approaches in big data: A state of art study*. U *2017 13th International Computer Engineering Conference (ICENCO)*, stranice 224–230, prosinac 2017.

[25] Hilal, W., S. A. Gadsden i J. Yawney: *Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances*. Expert Systems with Applications, 193:116429, 2022.

[26] Jiao, Z., H. Ji, J. Yan i X. Qi: *Application of big data and artificial intelligence in epidemic surveillance and containment*. Intelligent Medicine, 3(1):36–43, 2023.

[27] Khurana, D., A. Koli, K. Khatter *et al.*: *Natural language processing: state of the art, current trends and challenges*. Multimedia Tools and Applications, 82:3713–3744, 2023.

[28] Loeffel, P. X.: *Adaptive machine learning algorithms for data streams subject to concept drifts*. Disertacija, prosinac 2017.

[29] Moore, S.: *How to Create a Business Case for Data Quality Improvement*. `https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement`, posjećena 6. 9. 2023.

[30] Murzyna, V.: *Innovative Uses for Big Data in Traffic Management*. `https://blog.goodvisionlive.com/innovative-uses-for-big-data-in-traffic-management`, posjećena 2023-09-06.

[31] Ogie, R. I., S. James, A. Moore, T. Dilworth, M. Amirghasemi i J. Whittaker: *Social media use in disaster recovery: A systematic literature review*. International Journal of Disaster Risk Reduction, 70:102783, 2022.

[32] Ongsulee, P., V. Chotchaung, E. Bamrungsi i T. Rodcheewit: *Big Data, Predictive Analytics and Machine Learning*. U *2018 16th International Conference on ICT and Knowledge Engineering (ICT&KE)*, stranice 1–6, 2018.

[33] Oshikawa, R., J. Qian i W. Y. Wang: *A Survey on Natural Language Processing for Fake News Detection*. U *Proceedings of the Twelfth Language Resources and Evaluation Conference*, stranice 6086–6093, Marseille, France, svibanj 2020. European Language Resources Association.

[34] Prohorchik, K.: *Big data governance: roles, frameworks, and a case study*. `https://www.itransition.com/blog/big-data-governance`, posjećena 6. 9. 2023.

[35] Ramasamy, A. i S. Chowdhury: *Big Data Quality Dimensions: A Systematic Literature Review*. svibanj 2020.

[36] Ridzuan, F. i W. M. N. Wan Zainon: *A Review on Data Cleansing Methods for Big Data*. Procedia Computer Science, 161:731–738, 2019.

[37] Rodrigues, D. i A.d. Silva: *A Study on Machine Learning Techniques for the Schema Matching Network Problem*. Journal of the Brazilian Computer Society, 27, 2021.

[38] Sahay, T., A. Mehta i S. Jadon: *Schema Matching using Machine Learning*. U *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, stranice 359–366, 2020.

[39] Services, EMC Education: *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley, 2015.

[40] Seth, S.: *Basics of Algorithmic Trading: Concepts and Examples*. https://www.investopedia.com/articles/active-trading/101014/basics-algorithmic-trading-concepts-and-examples.asp, posjećena 6. 9. 2023.

[41] Shayaa, S., N. I. Jaafar, S. Bahri, A. Sulaiman, P. Seuk Wai, Y. Wai Chung, A. Z. Piprani i M. A. Al-Garadi: *Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges*. IEEE Access, 6:37807–37827, 2018.

[42] Suwinski, P., C. Ong, M. H. T. Ling, Y. M. Poh, A. M. Khan i H. S. Ong: *Advancing Personalized Medicine Through the Application of Whole Exome Sequencing and Big Data Analytics*. 10:49, 2019.

[43] von der Heydt, A. S., P. Ashwin, C. D. Camp, M. Crucifix, H. A. Dijkstra, P. Ditlevsen i T. M. Lenton: *Quantification and interpretation of the climate variability record*. Global and Planetary Change, 197:103399, 2021.

[44] Zhu, X., X. Ao, Z. Qin, Y. Chang, Y. Liu, Q. He i J. Li: *Intelligent financial fraud detection practices in post-pandemic era*. The Innovation, 2(4):100176, 2021.

Sažetak

U ovom radu opisuju se svojstva velikih podataka, s dubljim uvidom u istinitost, vrijednost i promjenjivost. Dan je kratak pregled svih svojstava i neki primjeri izvora velikih podataka. Istinitost podataka promatra se kroz različita svojstva kvalitete podataka, a posebno su izdvojene poteškoće obrade jezika i ljudskog izražavanja. Dane su neke poslovne metode za poboljšanje i očuvanje kvalitete, kao i druge automatizirane metode.

Detaljno su opisani razni primjeri izvlačenja vrijednosti iz velikih podataka. Predstavljeni su alati i metode za obradu podataka, uključujući grupiranje podataka, regresijsku i prediktivnu analizu te klasifikaciju, uz naglasak na strojno učenje. Posebno se istražuje obrada prirodnog jezika i njene komponente. Navode se i neke platforme za analizu velikih podataka.

Promjenjivost je opisana primjerima promjenjivih podataka i modela. Dublje su istraženi prilagodljivi modeli na temelju pomaka koncepta. Navedeni su problemi povezivanja i integracije podataka iz novih izvora uz primjere algoritama za automatsko rješavanje tih problema.

U praktičnom dijelu na primjeru podataka s društvenim mreža pokazuju se neke metode obrade. Detaljno je opisan oblik podataka, kao i prikupljanje i pretprocesiranje. Istinitost, odnosno vjerodostojnost podataka ispituje se naivnim Bayesovim klasifikatorom, uz podjelu na glasine i provjerljive vijesti. Izvlačenje vrijednosti pokazuje se analizom sentimenta javnosti po temama koje pripadaju promatranim događajima, a promjenjivost pronalaženjem uzroka promjena u sentimentu prema odabranim temama, organizacijama i osobama.

Ključne riječi: veliki podaci, istinitost, kvaliteta podataka, promjenjivost, vrijednost, društvene mreže, obrada prirodnog jezika, analiza sentimenta, strojno učenje

Summary

In this paper the properties of big data are described, with a deeper insight into veracity, value, and variability. A brief overview of all properties and some examples of big data sources are provided. Data veracity is examined through various data quality attributes, and a particular focus is given to the challenges of analysing language and human expression. Some business methods for improving and preserving data quality are mentioned, as well as other automated methods.

Various examples of extracting value from big data are detailed. Tools and methods for data processing, including data clustering, regression and predictive analysis, and classification, are presented, with an emphasis on machine learning. Natural language processing and its components are specifically explored. Some platforms for big data analysis are also mentioned.

Variability is described with examples of variable data and models. Adaptive models based on concept drift are further explored. Data linkage and integration issues from new sources are discussed, along with examples of algorithms for automatic problem-solving.

In the practical part, some data processing methods are demonstrated using social media data as an example. Data format, collection, and preprocessing are described in detail. Veracity, or rather credibility of data is examined using a naive Bayes classifier, categorizing the data into rumours and verifiable news. Value extraction is demonstrated by analyzing public sentiment on topics related to observed events, while variability is explored by finding the causes of sentiment changes related to selected topics, organizations, and individuals.

Keywords: big data, veracity, data quality, variability, value, social media, natural language processing, sentiment analysis, machine learning

Životopis

Rođena sam 22. prosinca 1998. godine u Zagrebu. Pohađala sam Osnovnu školu Augusta Šenoe od 2005. do 2013. godine te V. gimnaziju (prirodoslovno-matematički smjer) u Zagrebu od 2013. do 2017. godine. Iste godine upisala sam preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu. Nakon završetka preddiplomskog studija 2021. godine upisala sam diplomski sveučilišni studij Računarstvo i matematika.