SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Mašta ili sjećanje? Otkrivanje razlika u kognitivnom doživljaju putem podatkovne znanosti

Fani Sentinella-Jerbić
Voditelj: Jurica Babić

Zagreb, svibanj 2022.

Sadržaj

1. Uvod i motivacija

Različite sfere znanosti i filozofije kroz prošlost su izbjegavale istraživanje problematike kognitivnih procesa radi njihove neopipljive i nevidljive prirode [10]. Ipak, razvojem različitih grana znanosti ova se tematika polako približila akademskoj zajednici kao vrijedna istraživanja. Maštanje i sjećanje, kao dvije osnovne i učestale radnje ljudskog uma, zadobile su pažnju različitih grana znanosti.

Za razumijevanje kognitivnih procesa uključenih u sjećanje i maštanje najčešće se koriste pojmovi epizodičkog i semantičkog pamćenja preuzetih iz psihologije. Epizodičko pamćenje vezano je za osobna iskustva čovjeka, dok je sematičko vezano za opće znanje o svijetu i načinu na koji stvari funkcioniraju [11]. Ovisno o tome mašta li osoba ili se prisjeća pretpostavlja se da će se koristiti semantičkim odnosno epizodičkim pamćenjem u većoj mjeri. Ovo vrlo intuitivno shvaćanje razlike između tih osnovnih kognitivnih radnji središnja je ideja mnogih dosadašnjih radova stručnjaka na istoj tematici.

Specifično, za proučavanje razlika između maštanja i sjećanja korisni su se pokazali tekstovi u kojima se prepričava neki događaj u obliku kratke priče. Za analizu takvih priča je dosad korištena složena obrada prirodnog jezika pomoću neuronskih jezičnih modela (engl. *neural language models*) kao metode istraživanja razlika u kognitivnim procesima sjećanja i zamišljanja. U sklopu ovog seminara pokušati će se postići slična saznanja o razlikama, ali pomoću jednostavnijih metoda podatkovne znanosti. Predstavit će se srodni rad na problemu, vlastito istraživanje, rezultati obavljene podatkovne analize i diskusija o postignutim rezultatima.

2. Povezano istraživanje

Različiti pristupi istraživanja kognitivnih procesa razlikuju se po promatranoj razini organizacije živćanog sustava, po korištenim tehnikama te po fokusnom aspektu kognicije [2]. U dosadašnjem radu za istraživanje kognitivnih procesa maštanja i sjećanja korišteni su razni pristupi poput funkcionalne magnetske rezonancije [1], mjerenja otkucaja srca i galvanske reakcije kože [4] te obrade prirodnog jezika [8]. Sukladno dostupnosti i mogućnostima dostupne opreme te ciljnoj složenosti tehnika podatkovne znanosti odabrana metoda za analizu u ovom je radu obrada prirodnog jezika. Korišten je postojeći skup podataka pod nazivom *Hippocorpus*. Detalji o korištenom skupu podataka i dosad ostvarenom radu na njemu detaljno su opisani u nastavku.

2.1. Skup podataka *Hippocorpus*

Inicijalno sakupljen u svrhu rada [8], *Hippocorpus* jest skup podataka od 6854 priča na engleskom jeziku. Ostvaren je putem platforme za masovnu podršku (engl. *crowdsourcing*) Amazon Mechanical Turk (MTurk). Originalno sakupljanje podataka provedeno je kroz tri faze:

1. **Stvarne priče** – u prvoj fazi od ispitanika su sakupljene priče od 15 do 25 rečenica o stvarnom događaju iz njihovog života u proteklih šest mjeseci, sažetci istih te proteklo vrijeme od događaja;

2. **Izmišljene priče** – u drugoj fazi ispitanicima su nasumično dodijeljeni sažetci iz prve faze na temelju kojih su također trebali napisati priču od 15 do 25 rečenica o dodijeljenom događaju;

3. **Prepričane priče** – u zadnjoj fazi kontaktirani su ispitanici iz prve faze nakon dva do tri mjeseca da ponovno napišu priču o prethodno ispričanom događaju.

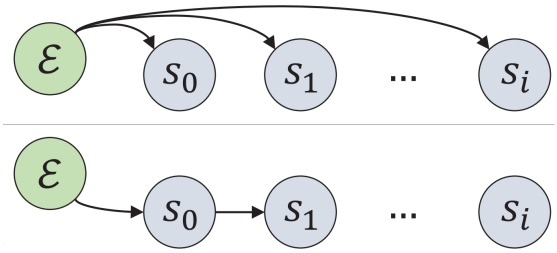
Osim samih priča sakupljene su dodatne informacije o subjektivnoj važnosti događaja, sličnosti događaja vlastitom životu, osobnosti ispitanika i učestalosti prisjećanja događaja u obliku kratke ankete. Skup podataka naknadno je dodatno nadograđen u svrhu rada [9] anotiranjem priča. Također putem MTurk platforme, unutar pojedine nasumično dodijeljene priče ispitanici su označavali mjesta novih događaja unutar priče te veličinu i očekivanost novih događaja u priči.

2.2. Dosadašnji rad na *Hippocorpusu*

U sklopu već navedenih radova [8] i [9] primarno su proučene razlike između izmišljenih i stvarnih priča kroz mjere tijeka narativa, količine konkretnih događaja i količine generalnog znanja. Ove i dodatne leksičke mjere opisane su u nastavku.

2.2.1. Mjere razlikovanja izmišljenih i stvarnih priča

Tijek narativa (engl. *narrative flow*) mjera je koja ilustrira linearnost slijeda rečenica u kontekstu priče. Izračunava se pomoću jezičnih modela lanca (engl. *chain*) i vreće (engl. *bag*). Modeli se razlikuju u načinu na koji objašnjavaju povezanost rečenica s glavnom temom priče ε predstavljajući dvije krajnosti kao što je moguće vidjeti na slici 2.1. Model lanca pretpostavlja da priča nastaje tako da iz glavne teme proizlaze rečenice koje onda linearno slijede jedna iz druge. Model vreće naprotiv pretpostavlja da je svaka rečenica neovisno proizašla iz glavne teme priče.



Slika 2.1: Konceptualni prikaz vjerojatnosnih modela vreće (gore) i modela lanca (dolje) ε je glavna tema priče, s_i je rečenica reda i. Preuzeto iz [8].

Uzevši u obzir predstavljene modele, tijek narativa izračunava se izrazom 2.1 odnosno kao logaritam negativne razlike vjerojatnosti rečenica dobivenih modelom vreće i modelom lanca normalizirane duljinom rečenice, pri čemu se vjerojatnost rečenice s_i u

kontekstu C računa kao zbroj logaritama vjerojatnosti pojedinih riječi rečenice w_t u danom kontekstu C: $\log p(s \mid C) = \sum_t \log p(w_t \mid C, w_{0:t-1})$.

$$\Delta l(s_i) = -\frac{1}{|s_i|}[\log p(s_i \mid \varepsilon) - \log p(s_i \mid \varepsilon, s_{1:i-1})] \tag{2.1}$$

Količina konkretnih događaja i količina izraženog generalnog znanja. Ove mjere direktno zrcale razliku između epizodičkog i semantičkog pamćenja u kontekstu teksta. Svaka komponenta procijenjena je zasebno:

1. **Konkretni događaji** su oni za koje su se autori izjasnili da su se dogodili, najčešće izraženi korištenjem prošlih glagolskih vremena primjerice "pojeo sam bananu" u kontrastu s hipotetskim događajima poput "pojeo bih bananu". Prepoznavanje konkretnih događaja ostvareno je modelom temeljenim na Google-ovom BERTu[3], transformacijskom modelu dubokog učenja za izgradnju kontekstualnih reprezentacija riječi (engl. *word embbedings*).

2. **Generalno znanje** mjereno je mapiranjem dijelova rečenica s grafom zdravorazumskog znanja ATOMIC [7] i služilo je za aproksimaciju semantičkog pamćenja uključenog u pričanje priče.

Leksičke mjere. O pričama su sakupljene dodatne informacije pomoću LIWC programa za analizu teksta. Izračunati su postotci riječi u danom tekstu koje spadaju u neku od 80 psiholoških, jezičnih ili tematskih kategorija.

2.2.2. Otkrivene razlike između izmišljenih i stvarnih priča

Na temelju spomenutih mjera radovima [8] i [9] otkrivene su sljedeće statistički značajne razlike. Na temelju tijeka narativa utvrđeno je da **izmišljene priče u prosjeku linearnije teku u odnosu na stvarne priče**. Na temelju njihovog konteksta, rečenice iz zamišljenih priča lakše je predviditi nego rečenice iz stvarnih priča. Osim toga **u stvarnim pričama detektirano je više konkretnih događaja**. S druge strane **u izmišljenim pričama detektirano je više iskazanog generalnog znanja**. Ovo reflektira intuitivno shvaćanje razlika između epizodičkog i semantičkog pamćenja predstavljenog u [11]. Pomoću leksičkih značajki otkriveno je da stvarne priče sadrže više analitičkih i konkretnih obilježja, dok izmišljene priče uzrokuju više kognitivnih procesa te fokusiranije su na osjećaje i osobne riječi.

2.2.3. Otkriveni odnosi stvarnih priča i naknadno prepričanih istih

Kod usporedbe stvarnih priča i naknadno prepričanih istih utvrđen je efekt narativizacije – **prepričane priče poprimaju karakteristike izmišljenih priča**; teku linearnije, sadrže manje konkretnih događaja i više generalnog znanja. Otkrivena je i poveznica da s većom učestalošću prisjećanja stvarne priče, raste i linearnost njenog prepričavanja, a pada broj konkretnih događaja.

3. Metode i postupci istraživanja

Za analizu razlika među stvarnim, izmišljenim i prepričanim pričama korišten je isti skup zadataka *Hippocorpus* predstavljen u 2. poglavlju. Cilj je bio doći do sličnih saznanja kao u predstavljenim radovima [8] i [9], ali pomoću jednostavnijih metoda podatkovne znanosti i obrade prirodnog jezika. Za to su odabrane metode inferencijalne statistike: statističko testiranje hipoteza i strojno učenje. Ove metode korištene su na ručno izgrađenim značajkama iz samih tekstova priča.

3.1. Odabir i izgradnja značajki

Za analizu je bilo potrebno izabrati određene postojeće značajke priča iz originalnog skupa podataka te izgraditi neke nove jednostavne značajke na temelju izabranih. U nastavku je detaljnije objašnjen postupak i motivacija iza učinjenog.

3.1.1. Odabir značajki iz skupa podataka

Hippocorpus sadrži razne već spomenute metapodatke o pričama poput važnosti, učestalosti prisjećanja i slično. Za analizu su ipak korišteni samo tekstovi i tipovi priča kako bi se postigli što općenitiji zaključci. Osim toga neki od metapodataka bili bi previše trivijalni pokazatelji o kojoj vrsti priče se radi. Primjerice, tek izmišljene priče sigurno imaju manju važnost od stvarnih i prepričanih. U sljedećoj tablici demonstriran je prikaz reduciranog skupa podataka koji je u nastavku korišten za analizu.

Tablica 3.1: Reducirani originalni skup podataka

priča	vrsta priče
Concerts are my most favorite thing, and my boyfriend knew it...	izmišljena
The day started perfectly, with a great drive up to Denver for...	stvarna
Me and my girlfriend had gone to the Los Angeles Zoo...	prepričana

3.1.2. Izgradnja novih značajki

Za izgradnju novih značajki bilo je potrebno zagrabiti u područje obrade prirodnog jezika (engl. *natural language processing*). Budući da je fokus bio na što jednostavnijem pristupu, izgrađene su uglavnom leksičke značajke. Leksička analiza odnosi se na proučavanje tekstova na razini samih riječi uz pomoć leksikona. Primjer takve značajke bio bi broj pojava određene riječi u tekstu. Distribucija novoizgrađenih značajki prikazana je na slici 3.1 pomoću histograma s pripadnom gustoćom vjerojatnosti dobivenom metodom procjene gustoće zrna (engl. *kernel density estimation*). Opis pojedinih značajki dostupan je u nastavku.

Broj znakova priče, broj riječi u priči i broj rečenica u priči

Ove iznimno jednostavne značajke osmišljene su s pretpostavkom da bi stvarne priče mogle biti dulje od izmišljenih.

Broj jedinstvenih riječi u priči, udio jedinstvenih riječi u priči i leksička raznolikost

Ove značajke osmišljene su kao jednostavan pokazatelj predvidljivosti teksta (analogno mjeri tijeka narativa). Što je više jedinstvenih riječi, moglo bi biti teže predvidjeti sljedeću. Leksička raznolikost je značajka koja se gradi pomoću 3.1 gdje w predstavlja riječ, $freq$ predstavlja broj pojavljivanja određene riječi i len označava broj riječi priče. Ova mjera je izgrađena inspirirano člankom [5].

$$H(story) = -\sum_{w \in story} \frac{freq(w)}{len(story)} \log_2 \frac{freq(w)}{len(story)} \tag{3.1}$$

Broj *neznačajnih* riječi u priči i udio *neznačajnih* riječi

Neznačajne riječi (engl. *stop words*) smatraju se riječi koje ne donose puno semantičke informacije pri analizi teksta poput veznika, prijedloga ili čestica.

Broj osobnih riječi u priči

Ova leksička značajka mjeri broj riječi koje se odnose na autora priče. Konkretno, brojana su pojavljivanja riječi *I*, *me* i *my*. Sudeći prema radovima [8] i [9], očekuje se da će izmišljene priče imati veće vrijednosti ove značajke.

Broj imenovanih entiteta

Imenovani entiteti su objekti iz stvarnog života poput organizacija, vremena, mjesta, osoba, aktivnosti, proizvoda i slično koje označavamo nekim imenom primjerice Fakultet elektrotehnike i računarstva, Plitvička jezera i Charles Darwin. Ova značajka zamišljena je kao aproksimacija konkretnih događaja. Naime, ako podrazumijevamo da osoba koja izmišlja priču se više koristi semantičkim pamćenjem nego epizodičkim, možemo pretpostaviti da će imati manje referenci na konkretne stvarne objekte, a više o općenitim i očekivanim događajima. Ova značajka nije leksičke prirode, ali postoji velik broj prethodno konstruiranih gotovih rješenja. U ovom slučaju korišten je predtrenirani model biblioteke *spaCy*[1] za Python.

Polarnost priče

Ova značajka predstavlja razinu pozitivnog ili negativnog sentimenta priče i njena vrijednost leži u intervalu [-1, 1]. Korištena je implementacija biblioteke *textblob*[2] za Python.

Subjektivnost priče

Ova značajka predstavlja razinu izraženih osobnih stavova i uvjerenja te njena vrijednost leži u intervalu [0,1]. Također je korištena implementacija biblioteke *textblob*[2] za Python.

Tablica 3.2: Transformirani skup podataka s izgrađenim značajkama za daljnju analizu

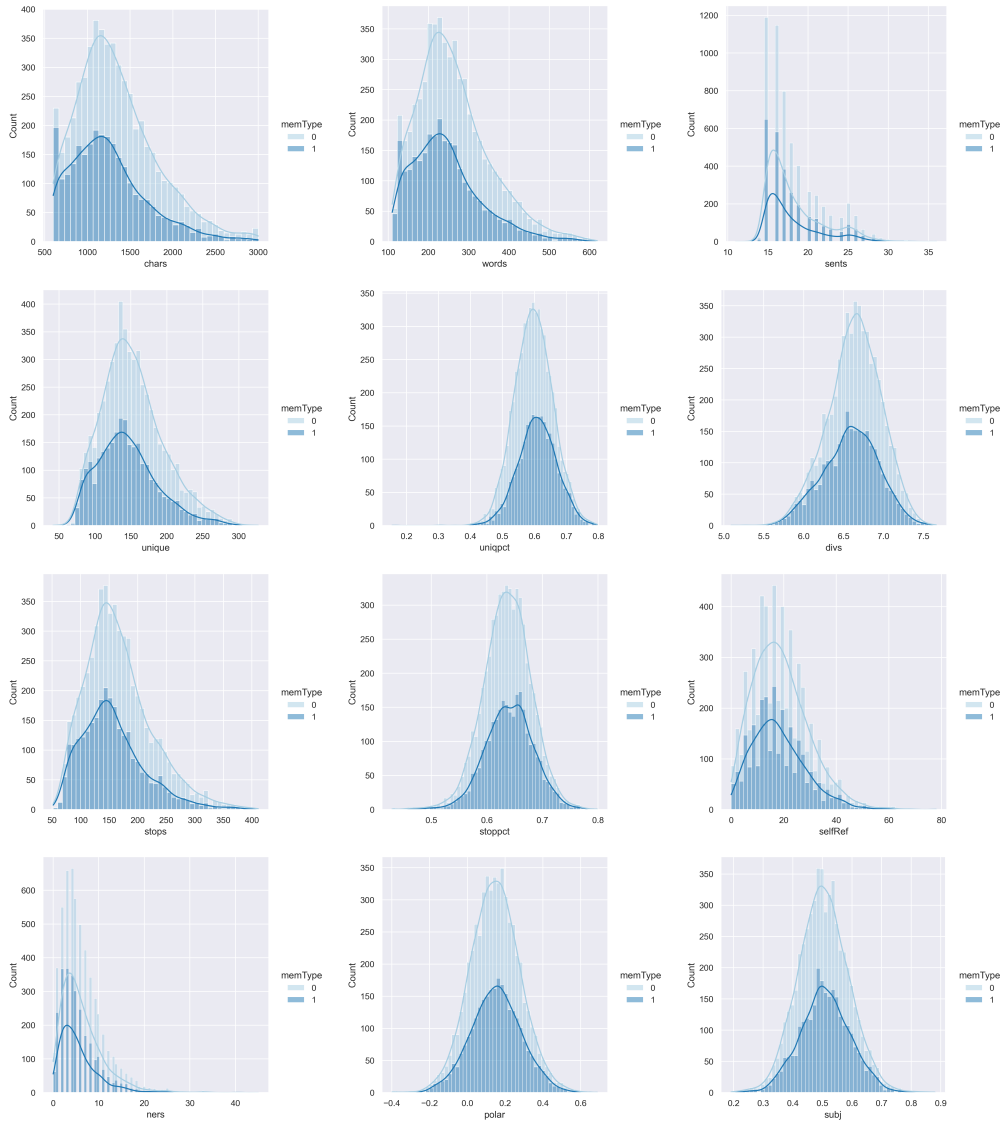
znakovi	riječi	rečenice	jedin.	udio jedin.	leks. var	neznač.	udio neznač.	osobne	im. entiteti	polar	subj	vrsta priče
1059	203	15	127	0.62562	6.5248	133	0.65517	18	2	0.30856	0.576111	stvarna
902	183	16	108	0.59017	6.23708	122	0.66667	12	4	0.485859	0.654444	izmišljena
1518	299	26	158	0.52843	6.57134	181	0.60535	16	10	0.19762	0.44405	prepričana

3.2. Statistički testovi

Statistički testovi čest su način za objašnjavanje pojava ili fenomena uočenih nad jednom ili više populacija. Danas korišteni statistički testovi najčešće koriste pristup testiranja hipoteza nad uzorcima populacije. Za ovaj pristup definiraju se nulta i alternativna hipoteza. Formuliraju se na način da alternativna hipoteza reprezentira

[1] `https://spacy.io/`

[2] Biblioteka *textblob* – `https://textblob.readthedocs.io/en/dev/`



Slika 3.1: Distribucija novoizgrađenih značajki pomoću histograma i procjene gustoće; 0 – stvarna priča, 1 – izmišljena priča

pitanje koje se želi odgovoriti ili teoriju koja se želi testirati, a nulta hipoteza reprezentira suprotnu tvrdnju [12]. Zatim se na temelju uzorka populacije utvrđuje jedan od dva moguća ishoda testiranja:

Odbacivanje nulte hipoteze. P-vrijednost je veća od razine značajnosti α – na testiranom uzorku postoji dovoljno dokaza da je moguće odbaciti tvrdnju H0.
Nemogućnost odbacivanja nulte hipoteze. P-vrijednost je manja od razine značajnosti α – na testiranom uzorku ne postoji dovoljno dokaza da bi se odbacila hipoteza H0.

Čest ilustarativni primjer koji se koristi jest pretpostavka nevinosti iz američkog sudskog sustava: *nevin dok se ne dokaže suprotno* [12]. Hipoteze su u tom slučaju prikazane s 3.2. Nulta hipoteza predstavlja *status quo* te se smatra istinitom dok se ne pokaže dovoljno dokaza koji ju opovrgavaju.

$$\begin{aligned} \mathbf{H0} &: optuženik\ je\ nevin \\ \mathbf{H1} &: optuženik\ je\ kriv \end{aligned} \tag{3.2}$$

Ovakav prikaz hipoteza za testiranje koristiti će se i u nastavku pri demonstraciji korištenih testova za pokušaj pronalaska statistički značajnih razlika među izmišljenim i stvarnim pričama.

3.2.1. Normalnost

Za velik dio statističkih testova potrebna je pretpostavka normalnosti razdiobe podataka. Postoji niz različitih testova za normalnost ovisno o dostupnim pretpostavkama o srednjoj vrijednosti i varijanci podataka. Ipak osnovna struktura im je ista, s nultom hipotezom da su podaci normalno distribuirani. Ovo je prikazano pomoću 3.3, radi konzistentnosti s ostatkom prikazanih testova.

$$\mathbf{H0} : uzorak\ podataka\ je\ normalno\ distribuiran \tag{3.3}$$

Za testiranje normalnosti izgrađenih značajki korišteni su Shapiro–Wilk i D'Agostinov K^2 testovi iz biblioteke *statsmodels*[3] za Python.

[3] Biblioteka *statsmodels* – `https://www.statsmodels.org/`

3.2.2. Homogenost varijanci

Osim pretpostavke o normalnosti, često se pri usporedbi dvaju (ili više) uzoraka podataka koristi pretpostavka homogenosti varijanci dvaju populacija. Formalno, nulta hipoteza jest da su varijance jednake, a alternativna da nisu, prikazano pomoću 3.4.

$$\begin{aligned} \mathbf{H0} &: \sigma_1^2 = \sigma_2^2 \\ \mathbf{H1} &: \sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{3.4}$$

U ovom slučaju korišteni su Bartlettov i Leveneov test homogenosti varijance, također iz biblioteke *statsmodels*[3] za Python.

3.2.3. Studentov t-test na dva uzorka

Ovaj test koristi se za određivanje jesu li dvije srednje vrijednosti populacije jednake [12]. Ovisno jesu li poznate varijance populacija i jesu li jednake u obje populacije, postoje različite inačice testa. U ovom istaživanju korištene su inačice kada su varijance populacije nepoznate te kada su varijance populacija jednake i kada nisu. Postavljene hipoteze za ovaj test prikazane su pomoću 3.5.

$$\begin{aligned} \mathbf{H0} &: \mu_1 = \mu_2 \\ \mathbf{H1} &: \mu_1 \neq \mu_2 \end{aligned} \tag{3.5}$$

Ovaj test korišten je nad izgrađenim značajkama u pokušaju dokazivanja statistički značajnih razlika među stvarnim i izmišljenim pričama. Iako se za ovaj test generalno podrazumijeva pretpostavka da je uzorak normalno distribuiran, ovaj postupak je robustan na nepoštivanje te pretpostavke pa je svejedno korišten čak i kada su testovi normalnosti pokazali da podaci značajki nisu normalno distribuirani.

3.3. Strojno učenje

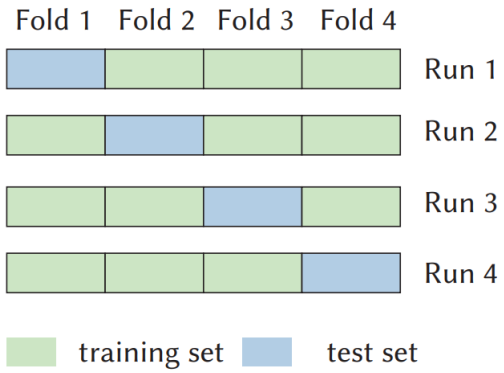
Cilj ovog dijela obrade skupa podataka jest pokazati prethodno spomenut efekt narativizacije. Ideja je utrenirati model strojnog učenja da na temelju priče klasificira je li ona stvarna ili izmišljena. Nakon uspješnog treniranja modela, testirati ga na prepričanim pričama. Ukoliko model odluči da je prepričana priča izmišljena, možemo zaključiti da je došlo do narativizacije sjećanja, tj. da opisani događaj autoru sve više pripada semantičkom sjećanju nego epizodičkom.

3.3.1. Treniranje klasifikatora

Treniranje modela strojnog učenja za klasifikaciju između stvarnih i izmišljenih priča obavljeno je nad prethodno objašnjenim transformiranim skupom podataka uz dodatno skaliranje oduzimanjem srednje vrijednosti i djeljnjem varijancom. Isprobani su algoritmi strojnog učenja logistička regresija, slučajna šuma, stroj potpornih vektora, stablo odluke i k-susjeda. Za one koje je to vremenski resurs dopuštao obavljen je pronalazak optimalne kombinacije hiperparametara metodom pretraživanja po rešetci (engl. *grid search*).

3.3.2. Odabir klasifikatora

Klasifikatori su međusobno uspoređivani k-strukom unakrsnom provjerom (engl. k-*folded cross-validation*)[13] pri čemu se skup primjera dijeli u k particija koje se nazivaju preklopi. Postupak je ilustriran slikom 3.2. Klasifikatori se uče k puta na $(k-1)$ preklopa (označenih zelenom bojom), a ispituju na k-tom preklopu (označen plavom bojom). Mjera uspješnosti klasifikatora se zatim uprosječuje među vrijednostima dobivenim iz svih ponavljanja.



Slika 3.2: Ilustracija metode k-struke unakrsne provjere. Preuzeto iz [6].

4. Rezultati i diskusija

U ovom poglavlju predstavljeni su rezultati postignuti predstavljenim metodama: statističkim testovima i strojnim učenjem.

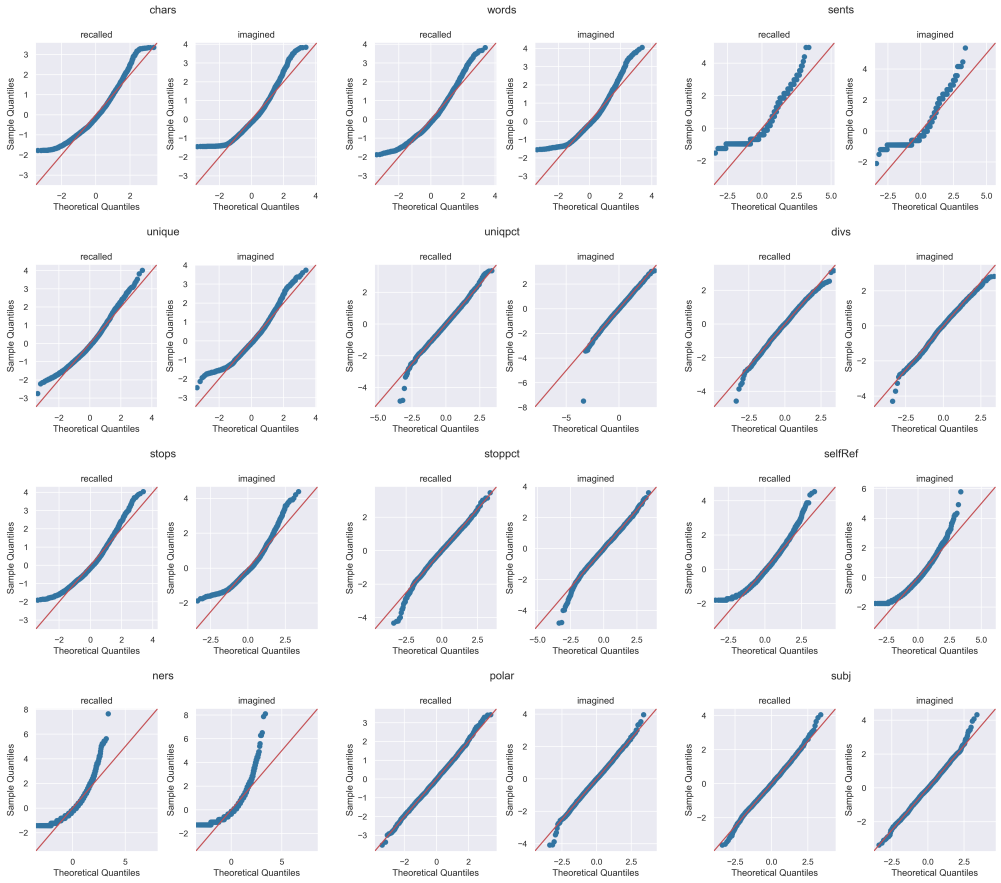
4.1. Statistički testovi

4.1.1. Normalnost

Testovi normalnosti s razinom značajnosti $\alpha = 0.05$ pokazali su za većinu novih značajki da nisu normalno distribuirane. Isto je moguće zaključiti i na temelju histograma iz prošlog poglavlja. Normalno distribuirana se pokazala samo polarnost priča. Za provjeru normalnosti također su korišteni Q-Q grafovi koji prikazuju usporedbu dvaju distribucija po kvantilima. Ako oba kvantila u usporedbi dolaze iz iste distribucije, točke bi trebale ležati na crvenoj liniji prikazanoj na grafu. Na slici 4.1 moguće je vidjeti da podaci ugalavnom ne prate normalnu distribuciju.

4.1.2. Homogenost varijanci

Testovi homogenosti varijanci s razinom značajnosti $\alpha = 0.05$ pokazali su miješane rezultate i prikazani su u tablici 4.1. Bartlettov test češće je pokazao da je varijanca značajki jednaka za izmišljene i stvarne priče, dok je Leveneov test češće pokazao da nisu jednake. Ovaj rezultat je zanimljiv s obzirom na to da se Bartlettov test generalno smatra strožim.



Slika 4.1: QQ grafovi uzoraka novoizgrađenih značajki

Tablica 4.1: Rezultati testova homogenosti varijanci

značajka	test	p-vrijednost	odbacivanje H0
broj znakova	Bartlett	0.15532	ne
	Levene	0.02878	da
broj riječi	Bartlett	0.13271	ne
	Levene	0.02705	da
broj rečenica	Bartlett	0.00394	da
	Levene	0.00698	da
broj jedinstvenih riječi	Bartlett	0.65766	ne
	Levene	0.94979	ne
udio jedinstvenih riječi	Bartlett	0.01252	da
	Levene	0.04265	da
leksička raznolikost	Bartlett	9.69321e-10	da
	Levene	2.09856e-09	da
broj *neznačajnih* riječi	Bartlett	0.04231	da
	Levene	0.01046	da
udio *neznačajnih* riječi	Bartlett	0.00593	da
	Levene	0.00742	da
osobne riječi	Bartlett	0.05765	ne
	Levene	0.01120	da
imenovani entiteti	Bartlett	3.56041e-11	da
	Levene	4.39898e-08	da
polarnost	Bartlett	0.00356	da
	Levene	0.03075	da
subjektivnost	Bartlett	0.66397	ne
	Levene	0.91825	ne

4.1.3. Studentov t-test na dva uzorka

Studentov t-test za izgrađene značajke s razinom značajnosti $\alpha = 0.05$ pokazao je da postoje značajne razlike u srednjim vrijednostima značajki stvarnih i izmišljenih priča.

Tablica 4.2: Rezultati t-testova dvaju uzoraka

značajka	varijance uzoraka	p-vrijednost	odbacivanje H0
broj znakova	jednake	2.61646e-46	da
broj riječi	jednake	4.13384e-44	da
broj rečenica	različite	4.82241e-05	da
broj jedinstvenih riječi	jednake	1.58477e-34	da
udio jedinstvenih riječi	različite	5.74908e-47	da
leksička raznolikost	različite	1.40730e-25	da
broj *neznačajnih* riječi	različite	2.47623e-36	da
udio *neznačajnih* riječi	različite	1.91675e-11	da
osobne riječi	jednake	1.23125e-05	da
imenovani entiteti	različite	2.21610e-32	da
polarnost	različite	0.00226	da
subjektivnost	jednake	2.09257e-08	da

4.2. Strojno učenje

U tablici 4.3 prikazani su rezultati petostruke unakrsne provjere korištenih modela strojnog učenja.

Tablica 4.3: Rezultati petostruke unakrsne provjere

	preklop					
algoritam	1.	2.	3.	4.	5.	srednja točnost
logistička regresija	0.63028	0.60606	0.61336	0.63239	0.64775	0.62597
slučajne šume	0.57904	0.57696	0.55524	0.56571	0.59559	0.57450
stabla odluke	0.56450	0.53946	0.58297	0.56585	0.55058	0.56067
stroj potpornih vektora	0.53673	0.55608	0.53846	0.53239	0.58686	0.55010
k-susjeda	0.51619	0.53443	0.52165	0.50355	0.55253	0.52567

Logistička regresija pokazala se najboljim klasifikatorom. Ona je uzeta za zadnji korak - testiranje nad prepričanim pričama. Dobivena je srednja vrijednost klasifikacija od **0.41622** nad prepričanim skupom.

Dodatno, provedena je procjena dobrote značajki za klasifikaciju pomoću analize varijance s pripadnom f-vrijednošću. Najznačajnije su se pokazale značajke broj znakova i udio jedinstvenih riječi.

4.3. Diskusija

Statistički testovi pokazali su razlike među stvarnim i izmišljenim pričama. Ipak, međusobne interakcije među značajkama su donjele još više nove informacije koje su strojnim učenjem detektirane. Srednja vrijednost klasifikacija prepričanih priča od 0.42911, odnosno točnost s 0.57089 pokazuje da je dio prepričanih priča klasificiran kao izmišljen. Uzevši u obzir srednju točnost modela od 0.61030 možemo zaključiti da smanjenje točnosti znači da je dio prepričanih priča doista dosegao narativizaciju. Ako razmotrimo jednostavnost konstruiranih značajki ovo je iznenađujući rezultat uzevši u obzir kompleksnost ljudske spoznaje. Posljednje, najznačajnije značajke pokazale su se broj znakova i udio jedinstvenih riječi. Potencijalna interpretacija bila bi da je maštanje zahtijevniji proces od sjećanja pa je teže osmisliti veću količinu teksta, a jedinstvene riječi moguće je shvatiti kao procjenu predvidljivosti teksta, a u prethodnom radu već je pomoću tijeka narativa pokazano da su izmišljene priče predvidljivije od stvarnih [8][9].

5. Zaključak

Za istraživanje razlika u kognitivnim procesima maštanja i sjećanja proučeni su postojeći pristupi i analiziran je postojeći skup priča na engleskom jeziku pomoću relativno jednostavih statističkih metoda testiranja hipoteza i strojnog učenja. Transformiran je originalni skup priča i stvorene su relevantne leksičke značajke. Pokazano je da postoje značajne razlike u opisivanju stvarnih i izmišljenih priča na razini jedne leksičke značajke te na razini više leksičkih značajki. Daljnji rad na ovoj tematici mogao bi koristiti uparene statističke testove nad izmišljenim i stvarnim pričama te istražiti razlike u sintaktičkim značajkama. Iznenađujuće je što su otkrivene tolike razlike običnim leksičkim značajkama s obzirom na veliku složenost ljudske spoznaje. Ovo pokazuje potencijal podatkovne znanosti u području proučavanja ljudskog uma koji će se uz malo sreće u budućnosti dalje razvijati.

6. Literatura

[1] Mathias Benedek, Till Schües, Roger E Beaty, Emanuel Jauk, Karl Koschutnig, Andreas Fink, i Aljoscha C Neubauer. To create or to recall original ideas: Brain processes associated with the imagination of novel object uses. *Cortex*, 99: 93–102, 2018.

[2] José Luis Bermúdez. *Cognitive Science: An Introduction to the Science of the Mind*. Cambridge University Press, 3 izdanju, Siječanj 2020. doi: 10.1017/9781108339216.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, i Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Paul S Foster, Daniel G Webster, i John Williamson. The psychophysiological differentiation of actual, imagined, and recollected mirth. *Imagination, Cognition and Personality*, 22(2):163–180, 2002.

[5] Jekaterina Novikova, Aparna Balagopalan, Ksenia Shkaruta, i Frank Rudzicz. Lexical features are more vulnerable, syntactic features have more predictive power. *arXiv preprint arXiv:1910.00065*, 2019.

[6] Fabian Pedregosa-Izquierdo. *Feature extraction and supervised learning on fMRI: from practice to theory*. Doktorska disertacija, Université Pierre et Marie Curie-Paris VI, 2015.

[7] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, i Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. U *Proceedings of the AAAI Conference on Artificial Intelligence*, svezak 33, stranice 3027–3035, 2019.

[8] Maarten Sap, Eric Horvitz, Yejin Choi, Noah A Smith, i James W Pennebaker.

Recollection versus Imagination: Exploring Human Memory and Cognition via Neural Language Models. U *Association for Computational Linguistics*, 2020.

[9] Maarten Sap, Anna Jafarpour, Yejin Choi, Noah A Smith, James W Pennebaker, i Eric Horvitz. Computational Lens on Cognition: Study of Autobiographical Versus Imagined Stories With Large-Scale Language Models. *arXiv preprint arXiv:2201.02662*, 2022.

[10] Paul Thagard, Yamini Chauhan, i Brian Duignan. cognitive science. *Encyclopædia Britannica*, Mar 2009. URL `https://www.britannica.com/science/cognitive-science`.

[11] Endel Tulving i Wayne Donaldson. *Organization of Memory*, stranice 381–403. NY: Academic Press, 1972.

[12] Ronald E Walpole, Raymond H Myers, Sharon L Myers, i Keying Ye. *Probability and Statistics for Engineers and Scientists*, svezak 5. Macmillan New York, 1993.

[13] Jan Šnajder. Vrednovanje modela. U *Strojno učenje 1 – predavanja*. Fakultet elektrotehike i računarstva, Sveučilište u Zagrebu, 2021.