

”Physics vs Chemistry vs Biology”

Sanja Nedeljković

June 19, 2023

Seminarski rad u okviru kursa Istraživanje podataka 1 na Matematičkom fakultetu.

Sadržaj

1 Uvod	2
1.1 Analiza skupa podataka	2
1.2 Identifikacija elemenata van granica	3
1.3 Rad sa nedostajućim vrednostima	4
1.4 Preprocesiranje	5
1.4.1 Priprema za klasifikaciju	6
1.4.2 Priprema za klasterovanje	7
1.4.3 Priprema za pravila pridruživanja	7
2 Klasifikacija	8
2.1 Naivni Bajes	8
2.1.1 Procena performansi podešavanjem hiperparametra alfa	8
2.1.2 Primena na test skup	9
2.2 Neuronske mreže	11
2.2.1 Izbor optimalne arhitekture neuronske mreže	11
2.2.2 Osnovni model	11
2.2.3 Regularizacija smanjenjem broja neurona	13
2.2.4 Duboka neuronska mreža	13
2.2.5 L2 regularizacija	14
2.2.6 Dropout regularizacija	14
2.2.7 Kombinovanje L2 i Dropout	15
2.2.8 Upoređivanje i primena na test skup	16
3 Klasterovanje	17
3.1 Algoritam K-sredina	17
3.1.1 Klasterovanje u tri grupe i prikaz pravih klasa	17
3.1.2 Klasterovanje normalizovanih podataka u tri grupe	20
3.1.3 Traženje optimalnog broja klastera na redukovanim podacima	20
3.1.4 Traženje optimalnog broja klastera na originalnim podacima	22
3.2 Hjерархијско klasterovanje	23
3.2.1 Dodatna transformacija podataka	23
3.2.2 Odabir broja klastera na osnovu analize dendograma	23
3.2.3 Vizuelizacija rezultata	24
4 Pravila pridruživanja	26
4.1 Primena apriori algoritma na ceo skup podataka	27
4.2 Primena apriori algoritma na posebne klase	28
5 Zaključak	30

1 Uvod

Ovaj rad se fokusira na bazu podataka [Physics vs Chemistry vs Biology](#) koji je posebno dizajniran za zadatke klasifikacije u oblasti obrade prirodnog jezika. Sprovedena je detaljna analiza podataka, kako bi se stekao uvid u karakteristike ovog skupa. Demonstrirana je primena različitih tehniki, uključujući klasifikaciju pomoću Naivnog Bajesa i neuronskih mreža, klasterovanje pomoću K-sredina i hijerarhijskih metoda, i pronalaženje pravila pridruživanja pomoću Apriori algoritma.

1.1 Analiza skupa podataka

Skup podataka je podeljen na trening i test. Ukupno postoji 10281 instanci, od kojih 8695 pripada trening skupu, a 1586 test skupu. U nastavku, sva analiza je vršena nad trening skupom.

Id	Comment	Topic
0 0x840	A few things. You might have negative- frequency dependent selection going on where the least common phenotype, reflected by genotype, is going to have an advantage in the environment. For instance, if a prey animal such as a vole were to have a light and a dark phenotype, a predator might recognize the more common phenotype as food. So if the light voles are more common, foxes may be keeping a closer eye out for light phenotypic voles, recognising them as good prey. This would reduce the light causing alleles due to increased predation and the dark genotypes would increase their proportion of the population until this scenario is reversed. This cycle continues perpetually. \n\nHowever, this is unlikely to be strictly yearly as it usually takes more time than a year for an entire populations allele frequencies to change enough to make a large enough difference to alter fitness. \n\nMore likely on a "year to year" basis, the population is experiencing fluctuating selection where al...	Biology
1 0xbf0	Is it so hard to believe that there exist particulars out that that we can't detect with anything we've invented so far.\n\nI mean look how long it took humans to find out a way to detect radiation.	Physics
2 0x1dfc	There are bees	Biology
3 0xc7e	I'm a medication technician. And that's alot of drugs on your liver. You probably won't die immediately you'll be fine. Take care of your self tho that's definitely not good for your body	Biology
4 0xbba	Cesium is such a pretty metal.	Chemistry

Figure 1: Prikaz prvih pet instanci.

U bazi postoje tri atributa:

- **Id** - Id korisnika koji je ostavio komentar. Atribut nije relevantan za analizu te će biti uklonjen.
- **Comment** - Sirov oblik komentara sa društvene mreže Reddit.
- **Topic** - Kategorički atribut koji sadrži vrednosti klasa: 'Biology', 'Chemistry', 'Physics'.

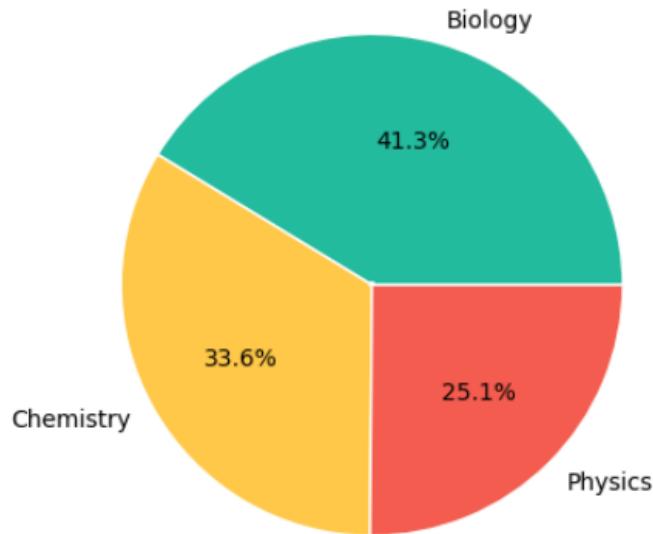


Figure 2: Raspodela po klasama.

1.2 Identifikacija elemenata van granica

Nije očigledno šta bi bili elementi van granica. Međutim, reči koje se nalaze u veoma dugačkim komentarima mogu višestruko da uticu na mogućnost generalizacije algoritama. Ukoliko se javljaju samo u timinstancama, biće označene kao značajne za tu klasu, iako se ne zna da li su domenske reci niti da li se mogu naći u skupu predviđenom za testiranje. Dodatno, povećavaju sam rečnik što može dovesti do preprilagodavanja.

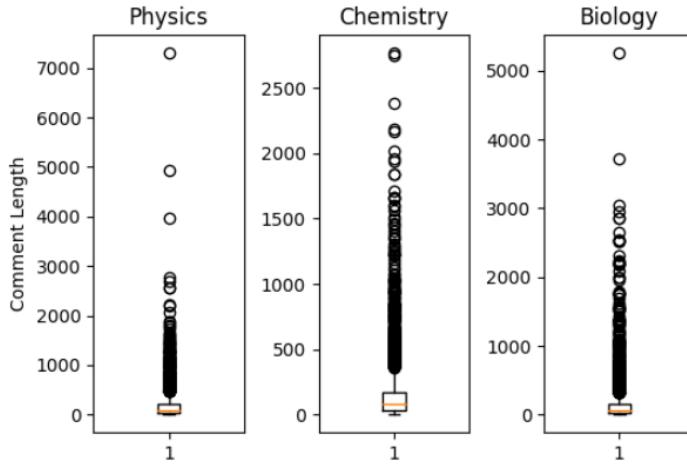
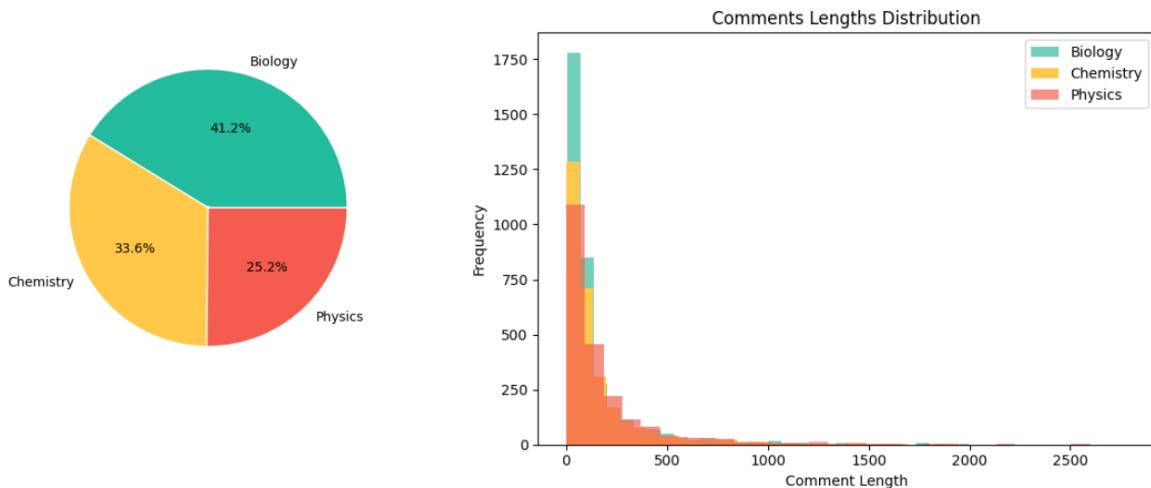


Figure 3: Boxplot dijagram dužine komentara po klasama.

Jasni elementi van granica bi bile dužine veće od 3000, 2500, 3000, za fiziku, hemiju, biologiju, tim redosledom. Za klasu hemija veoma mali broj instanci je veći i od 2000, te je granica spuštena. U slučaju klase biologija postoji relativno veliki broj podataka u segmentu [2000, 3000]. Međutim kako u trening skupu ta klasa čini 41% svih instanci i je granica pomerena na 2000 radi bolje balansiranosti i uniformnije raspodele. Fizika ostaje nepromjenjena jer ima najmanju zastupljenost u treningu sa 25.1% ukupnih podataka, te se ovde izbacuju samo jasni outlajeri.



(a) Raspodela po klasama nakon uklanjanja elemenata.

(b) Raspodela dužina komentara nakon uklanjanja elemenata.

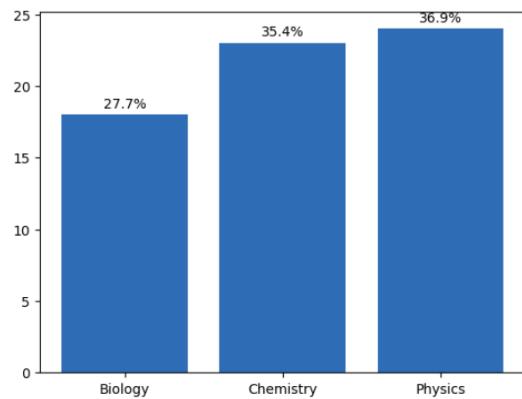
1.3 Rad sa nedostajućim vrednostima

```
data.isna().any().any()
```

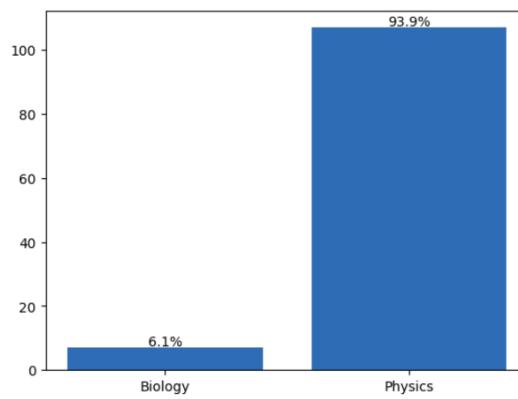
```
False
```

Ustanovljeno je da nema pristunih null ili NaN vrednosti. Ipak, imajući u vidu da je ovo neprocesiran skup podataka sačinjen od komentara sa društvene mreže, neke druge reči ili izrazi bi mogli biti indikatori nedostajuće vrednosti.

U okviru Reddit platforme, termini "[removed]" i "[deleted]" označavaju uklanjanje i brisanje komentara. Dodatno, prisustvo "[removed]" označava da su moderatori uklonili odgovarajući komentar, što često implicira da je originalni komentar sadržao uvredljiv sadržaj, kontroverzna mišljenja ili je prekršio smernice zajednice u kojoj je pokrenuta diskusija. Sa druge strane, kada je prisutan termin "[deleted]", to sugerije da je korisnik dobrovoljno odlučio ga ukloni. U oba slučaja, pojava ovih termina je izolovana, bez bilo kakvih dodatnih znakova ili reči u komentaru.



(a) Broj [deleted] instanci po klasama.



(b) Broj [removed] instanci po klasama.

Postoji ukupno 179 nedostajućih vrednosti, od kojih 65 predstavlja obrisane komentare, a 114 uklonjene. Skup obrisanih komentara pokazuje pretežno ujednačenu distribuciju po klasama, uz blagu naklonjenost ka fizici. Konkretno, sastoji se od 27,7% biologije, 35,4% hemije i 36,9% fizike.

Sa druge strane, skup uklonjenih komentara ima malo interesantniji obrazac sa značajno različitim raspodelama po klasama. Konkretno, biologija cini 6.1% podataka, dok fizika dominira sa znatnom većinom od 93.3%. Primetno, nijedan uklonjen komentar se ne odnosi na kategoriju hemije. Ovo implicira da su diskusije vezane za biologiju manje sklene ukljanjuju u poređenju sa fizikom. Dodatno, velika zastupljenost fizike ukazuje na veću verovatnoću kontroverznih diskusija u tom domenu, koje mogu dovesti do preterano neprimernog izražavanja ili kršenja smernica zajednice. Nasuprot tome, odusustvo komentara u kategoriji hemije može ukazati na manju učestalost polarizujućih tema u toj oblasti.

Iako je verovatno da je ovakva raspodela indikovana količinom kontroverznih tema ili mišljenja u sklopu svake od ove tri oblasti, ne treba isključiti mogućnost da su različite smernice u svakoj od zajednica, i da moderatori nemaju isti stepen prihvatljivosti mišljenja ili neprimernog sadržaja.

Sve navede instance se uklanjaju iz trening skupa.

1.4 Pretpresiranje

Kao sto je već napomenuto, ovi komentari su sirovi. Neophodno je uraditi pretpresiranje koje ima za cilj da poboljša kvalitet reprezentacije skupa pre primene TF-IDF matrice. Primjenjuje se sledeći niz koraka na svaki od komentara:

- 1. Mala slova:** Ceo tekst je konvertovan u mala slova da bi se obezbedila neosetljivost na mala i velika slova. Ne treba da se 'physics' i 'Physics' posmatraju kao dva različita tokena.

'PhYSics' → 'physics'

- 2. Filtriranje reči:** Zadržavaju se samo slova, zamenjujući belinama brojeve, simbole i znakove interpunkcije.

'physics!' → 'physics'

'chemistry, bio, physics . ' → 'chemistry bio physics '

- 3. Uzastopno ponavljanje:** Reči koje sadrže slovo ponovljeno više od dva puta uzastopno biće eliminisane. Ove reči su šum jer je u engleskom jeziku najveći broj uzastopnog ponavljanja karaktera jednak upravo dva.

'physicsss' → "

- 4. Tokenizacija:** Deljenje rečenice na pojedinačne reci ili lekseme, koristeći razmak kao separator.

'sentence about physics' → 'physics', 'about', 'physics'

- 5. Stop words:** Skup reči koji se veoma često koristi u jeziku, kao sto su predlozi i zamenice se uklanjaju jer uglavnom dodaju malo kontekstualnih informacija. Kao i kod prethodnog koraka, predstavljaju šum.

'there are bees' → 'bees'

- 6. Lematizacija:** Skraćivanje reči do njihovog korena ili osnovnog oblika u rečniku. Motivacija je da različiti i izvedeni oblici iste reči budu jedan token.

'run' → 'run'

'running' → 'run'

- 7. Minimalna dužina:** Odbacivanje tokena manjih od 3 karaktera, jer verovatno nisu smisleni i značajni.

'to' → "

Napomena: u praksi se često uklanjaju linkovi. Ovde to nije urađeno, jer je moguće da su prosleđene domenski specifične stvari, kao što su naučni članci, koji bi mogli u svom nazivu da sadrže bitne termine.

	Comment	Processed Comments
0	A few things. You might have negative- frequency dependent selection going on where the least common phenotype, reflected by genotype, is going to have an advantage in the environment. For instance, if a prey animal such as a vole were to have a light and a dark phenotype, a predator might recognize the more common phenotype as food. So if the light voles are more common, foxes may be keeping a closer eye out for light phenotypic voles, recognising them as good prey. This would reduce the light causing alleles due to increased predation and the dark genotypes would increase their proportion of the population until this scenario is reversed. This cycle continues perpetually. \n\nHowever, this is unlikely to be strictly yearly as it usually takes more time than a year for an entire populations allele frequencies to change enough to make a large enough difference to alter fitness. \n\nMore likely on a "year to year" basis, the population is experiencing fluctuating selection where al...	thing negative frequency dependent selection common phenotype reflected genotype advantage environment instance prey animal vole dark phenotype predator recognize common phenotype food vole common fox keeping closer eye phenotypic vole recognising prey reduce causing allele due increased predation dark genotype increase proportion population scenario reversed cycle continues perpetually however unlikely strictly yearly take entire population allele frequency large difference alter fitness nmore basis population experiencing fluctuating selection alternating condition environment favor genotype perhaps plant specie living area flooded phenotype population plant dryer wet flooding dry genotype fitness leading offspring therefore dry allele population flooded year wet liking phenotype propagate wet gene
1	Is it so hard to believe that there exist particulars out that that we can't detect with anything we've invented so far.\n\nI mean look how long it took humans to find out a way to detect radiation.	exist particular detect invented took human detect radiation
2	There are bees	bee
3	I'm a medication technician. And that's alot of drugs on your liver. You probably won't die immediately you'll be fine. Take care of your self tho that's definitely not good for your body	medication technician alot drug liver die immediately fine care self tho
4	Cesium is such a pretty metal.	cesium metal
5	I meant that the question itself is unclear.	meant unclear
6	Shove it up your ass and see what happens	shove happens
7	??? I mean it has some butter, but besides that it's sugar, baking soda, and peanuts, so yeah...	butter besides sugar baking soda peanut
8	https://t.me/joinchat/3gEILHLMCxhNGIO	joinchat gelllumcxhngi
9	Well, that's just the thing. You can't really induce an immune response against yourself by introducing a protein that's already there.	induce immune response introducing protein

Figure 6: Uporedni prikaz komentara pre i nakon preprocesiranja.

1.4.1 Priprema za klasifikaciju

Da bi se omogućila primenu algoritama koji obično rade sa numeričkim podacima, izvršena je transformacija podataka iz stringovne u numeričku reprezentaciju koristeći TF-IDF (*Term Frequency-Inverse Document Frequency*) ¹ matricu. TF IDF matrica svakom tokenu dodeljuje vrednost, i obuhvata važnost termina u odnosu na dokument (komentar) u kome se pojavlja, kao i njegovu relevantnost u širem kontekstu čitavog skupa podataka. Zbog efikasnijeg izvršavanja i prevelike količine reči u rečniku, ograničen je broj atributa na 8000.

Dodatno, enkodirana je ciljna promenljiva Topic.

[‘Biology’, ‘Chemistry’, ‘Physics’]

↓

[0, 1, 2]

	aaronson	aat	abdomen	abduction	ability	abomination	abrasive	abroad	absence	absent	...	zerg	zero	zhu	zika	zinc	zone	zoom	ztsi	zurich	0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

Figure 7: Prikaz prvih 10 transformisanih instanci.

Na isti način se obrađuju podaci u test skupu, uz napomenu da se za transformaciju u TF IDF matricu koristi vectorizer naučen na trening podacima da ne bi došlo do prilagođavanja nepoznatim podacima.

1.4.2 Priprema za klasterovanje

Za potrebe klasterovanja, vrši se spajanje trening i test skupa u jedan skup podataka i uklanjanje ciljne promenljive. Preprocesiranje komentara ostaje isto, a TF IDF matrica sada uči iz celog skupa.

	abdomen	ability	abomination	abortion	abrasive	absence	absent	absolute	absolutely	absorb	...	youtube	youtuber	yup	zebra	zeolite	zero	zika	zinc
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 8: Prikaz prvih 10 transformisanih instanci.

1.4.3 Priprema za pravila pridruživanja

Za određivanje pravila pridruživanja pogodnija je reprezentacija pomoću binarne vreće reci gde 1 označava da se reč pojavljuje u dатој instanci, a 0 suprotno. S obzirom da je i za ovaj zadatak potreban ceo skup podataka, obrađuje se dobijena reprezentacija podataka za klasterovanje, tako što sve vrednosti veće od 0.0 u matrici zamenjuju sa 1.

	abdomen	ability	abomination	abortion	abrasive	absence	absent	absolute	absolutely	absorb	...	youtube	youtuber	yup	zebra	zeolite	zero	zika	zinc
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0

Figure 9: Prikaz prvih 5 transformisanih instanci.

Dodatno posebno su napravljeni rečnici za svaku od klasa, radi kasnije detaljnije analize.

2 Klasifikacija

2.1 Naivni Bajes

Naivni Bajes (konkretnije, Multinomijalni Naivni Bajes)² je pogodan model za dati problem zbog karakteristike efikasnog izvršavanja nad skupovima podataka velikih razmara. Iako jednostavan, u praksi pokazuje dobre rezultate za klasifikaciju dokumenata, višeklasna predviđanja i rada nad TF IDF matricom.

2.1.1 Procena performansi podešavanjem hiperparametra alfa

Jedan ključni hiperparametar u ovom algoritmu je α , koji kontroliše izglađivanje (*smoothing*) primenjeno na procene verovatnoće. U praktičnom smislu, on utiče na osetljivost modela na specifične obrasce i raspodelu podataka u trening skupu. Izbor vrednosti alfa može da ima značajan uticaj na mogućnost generalizacije modela.

Da bi se procenilo kako se Bajesov model ponaša sa različitim α vrednostima, kao prvi korak neophodno je podeliti trening podatke na trening i validacioni skup, zadržavajući u oba raspodelu početnog skupa. Validacioni skup se koristi za testiranje, kako ne bi došlo do prilagođavanja nepoznatim podacima. Ispitivanje je vršeno za vrednosti [0.0, 0.1, 0.5, 1.0, 2.0], redom, ne menjajući ostale parametere, osim u slučaju $\alpha = 0.0$, gde je neophodno postaviti parametar `force_alpha=True` kako bi vrednost bila zaista nula, a ne bliska nuli. Efikasnost modela se procenjuje analizom narednih metrika³: matrica konfuzije, f1, preciznost, tačnost, odziv.

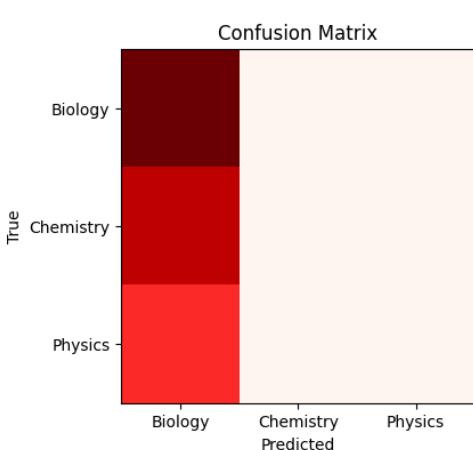
$$\begin{bmatrix} 710 & 0 & 0 \\ 578 & 0 & 0 \\ 410 & 0 & 0 \end{bmatrix}$$

Figure 10:
Matrica konfuzije

	Preciznost	Odziv	F1-mera
0	0.42	1.00	0.59
1	0.00	0.00	0.00
2	0.00	0.00	0.00
Tačnost	-	-	0.42
Makro sredina	0.14	0.33	0.20
Težinska sredina	0.17	0.42	0.25

Figure 11: Metrike klasifikacije

Figure 12: Dobijeni rezultati za $\alpha = 0.0$



Confusion Matrix

Posebno, za slučaj $\alpha = 0.0$, svi podaci su klasifikovani kao najbrojnija klasa i dobijena je tačnost od 42% (što ga čini boljim od nasumičnog modela za 8.67%), međutim, preciznost, odziv i f1 su jako niski, sa vrednostima 0.14%, 0.33%, 0.20%, redom. Model se u potpunosti prilagodio trening skupu i nema mogućnost generalizacije. Ovo je bilo očekivano, jer je kod NLP problema potrebno naći malo suptilnije obrasce u podacima.

Figure 13: Ilustrovana matrica konfuzije za $\alpha = 0.0$

Table 1: Izveštaj o klasifikaciji

α	Preciznost	Odziv	F1-mera	Tačnost
0.1	0.74	0.74	0.73	0.74
0.5	0.75	0.74	0.74	0.74
1.0	0.75	0.73	0.73	0.73
2.0	0.75	0.71	0.70	0.71

Za preostale vrednosti parametra, model ispoljava slične performanse za svaku od metrika, varirajući od 0.68% do 0.75%. Ipak, uočava se da je najbolji rezultat ostvaren za $\alpha = 0.5$, te je ova vrednost korišćena za testiranje.

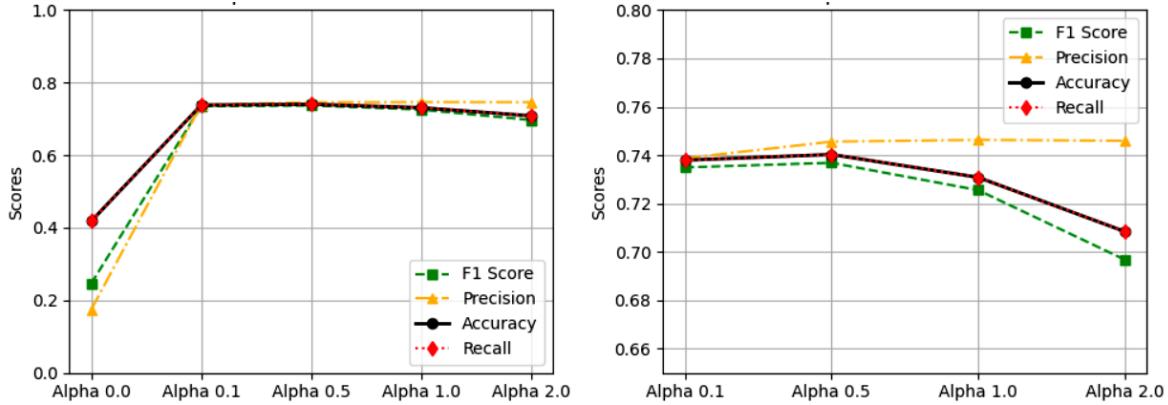


Figure 14: Izveštaj o klasifikaciji svih modela (levo) i *smooth* modela (desno)

2.1.2 Primena na test skup

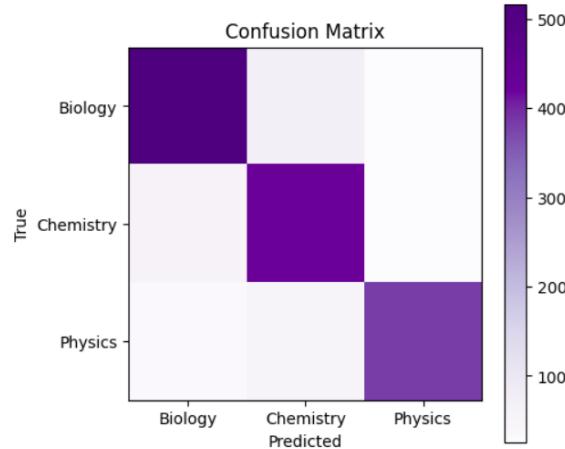


Figure 15: Ilustrovana matrica konfuzije na test skupu

Na test skupu su dobijeni zadovoljavajući rezultati. Preciznost je 84%, dok su odziv, F1 i tačnost po 83% što je prihvatljivo za ovaj tip izazova.

$$\begin{bmatrix} 516 & 73 & 25 \\ 57 & 424 & 25 \\ 32 & 51 & 383 \end{bmatrix}$$

Figure 16:
Matrica konfuzije

	Preciznost	Odziv	F1-mera
0	0.85	0.84	0.85
1	0.77	0.84	0.80
2	0.88	0.82	0.85
Tačnost	-	-	0.83
Makro sredina	0.84	0.83	0.83
Težinska sredina	0.84	0.83	0.83

Figure 17: Metrike klasifikacije
Figure 18: Dobijeni rezultati na test skupu

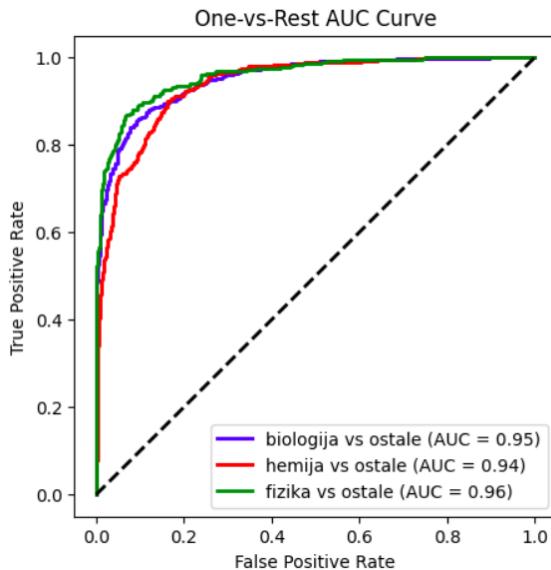


Figure 19: AUC kriva



Figure 20: Kriva učenja

U svrhu demonstracije koliko dobro model razlikuje svaku od klase od ostalih korišćena je AUC kriva. Veoma visoke vrednosti, veće od 0.94, ukazuju na to da model ima jake diskriminatorene sposobnosti: efektivno pravi razlike između pozitivne i preostalih klasa.

Idealno, funkcija traninga i funkcija validacije se ponašaju slično i stabilizuju se nakon nekog vremena. U konkretnom primeru, može da se zaključi da je model preprilagođen i ima ograničenu mogućnost generalizacije. Međutim, imajući u vidu prirodu problema i broj instanci, ovakvo ponašanje nije neočekivano. Jedno od potencijalnih rešenja bi bilo povećanje dostupnih podataka prikupljanjem dodatnih instanci.

2.2 Neuronske mreže

Neuronska mreža (konkretnije, Višeslojni perceptron ili *Multilayer Perceptron*)⁴ predstavlja drugi izbor za ovaj problem zbog svoje sposobnosti da uočava složene obrasce, karakteristike i odnose unutar podataka.

2.2.1 Izbor optimalne arhitekture neuronske mreže

Kada je u pitanju izgradnja efektivnog modela, u okviru ovog algoritma je od najveće važnosti izbor arhitekture, koja određuje dubinu, širinu, stepen regularizacije i funkcije aktivacije mreže.

U narednim odeljcima biće na različitim arhitekturama predstavljena ispitivanja metrika performansi koje uključuju funkcije gubitka (*loss*) i tačnost. Sva analiza, donošenje zaključaka i testiranje vrši se na trening i validacionom skupu.

Postavka modela: U kontekstu višeklasne klasifikacije, često se koriste sledeća podešavanja:

1. Funkcija cilja = '`sparse_categorical_crossentropy`': pogodna u slučaju da je ciljna klasa kodirana celim brojevima⁵.
2. Funkcija aktivacije poslednjeg sloja = '`softmax`': prikladna jer obezbeđuje da se predviđene vrednosti sabiraju na 1⁶.
3. Poslednji (izlazni) sloj neuronske mreže ima 3 neurona: po jedan za svaku od klase koje se predviđaju.
4. Optimizator = '`adam`' (široka primena).
5. Aktivaciona funkcija = '`relu`' (široka primena).

Iako podesivi, ovi parametri ostaju isti za svaki od modela, a fokus je na uticaju širine, dubine i posebnih tehnika regularizacije.

2.2.2 Osnovni model

Kao polazna tačka se koristi osnovni model, čija se implementacija sastoji od ulaznog sloja, unutrašnjeg sloja od 64 perceptronova i izlaznog sloja. Glavna svrha je uspostavljanje merila performansi. Pored toga, služi i za ispitivanje efekta različitih vrednosti hiperparametra *batch* na tacnost klasifikacije, u cilju pronaalaženja optimalne veličine grupe(*batch size*) za kasnije razvijanje modela.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
first_layer (Dense)	(None, 64)	512,064
predictions (Dense)	(None, 3)	195
Total params:	512,259	
Trainable params:	512,259	
Non-trainable params:	0	

Na osnovu slike 21, u sva tri slučaja tačnost na trening skupu posle svega 5 epoha dostiže približno 90%, dok na validacionom skupu ne prelazi 72.5% i nakon istog vremena ispoljava blagi pad. Ovakav odnos tačnost nije poželjan i jasan je indikator prekomernog prilagodjavanja. Tu prepostavku dodatno potvrđuje rast funkcije gubitka na validacionom skupu. Od sve tri veličine, najbolje performanse se mogu primetiti u slučaju *batch=128*, usled najkasnijeg i najblažeg uspona funkcije gubitka i najmanje razlike tačnosti na skupovima.

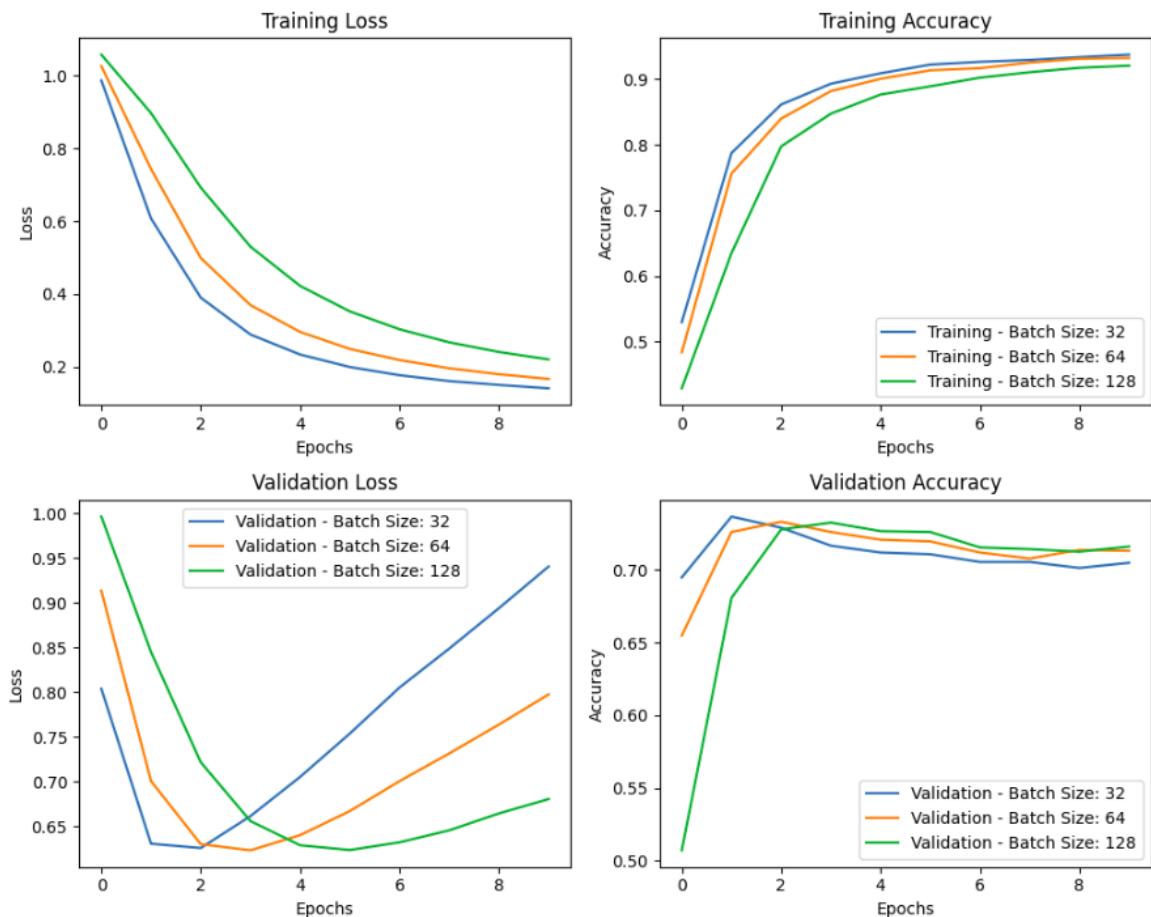


Figure 21: Tačnost i funkcija gubitka na trening i validacionim skupovima

2.2.3 Regularizacija smanjenjem broja neurona

Da bi se ublažila preterana osetljivost na trening podatke, neophodno je smanjiti ukupnu složenost modela koristeći različite oblike regularizacije. Početna tehnika regularizacije uključuje smanjenje broja neurona u unutrašnjem sloju mreže sa 64 na 32, čime se dobija jednostavniji model i veća otpornost na šum u podacima.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
first_layer (Dense)	(None, 32)	256,032
predictions (Dense)	(None, 3)	99
Total params:	256,131	
Trainable params:	256,131	
Non-trainable params:	0	

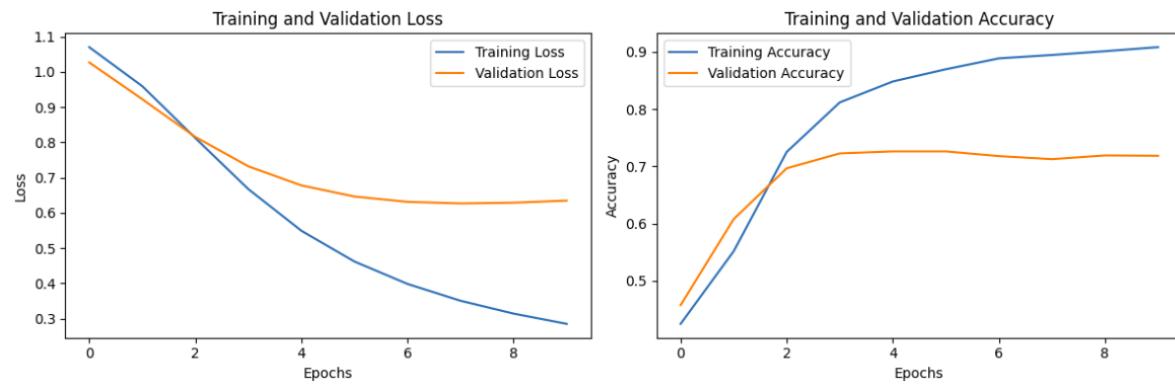


Figure 22: Bolje obe metrike.

2.2.4 Duboka neuronska mreža

Duboku neuronsku mrežu karakteriše postojanje više od jednog unutrašnjeg sloja. Svrha uključivanja složenije arhitekture je da potvrdi pretpostavku o preprilagođavanju.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
first_layer (Dense)	(None, 64)	512,064
second_layer (Dense)	(None, 32)	2,080
third_layer (Dense)	(None, 32)	1,056
predictions (Dense)	(None, 3)	99
Total params:	515,299	
Trainable params:	515,299	
Non-trainable params:	0	

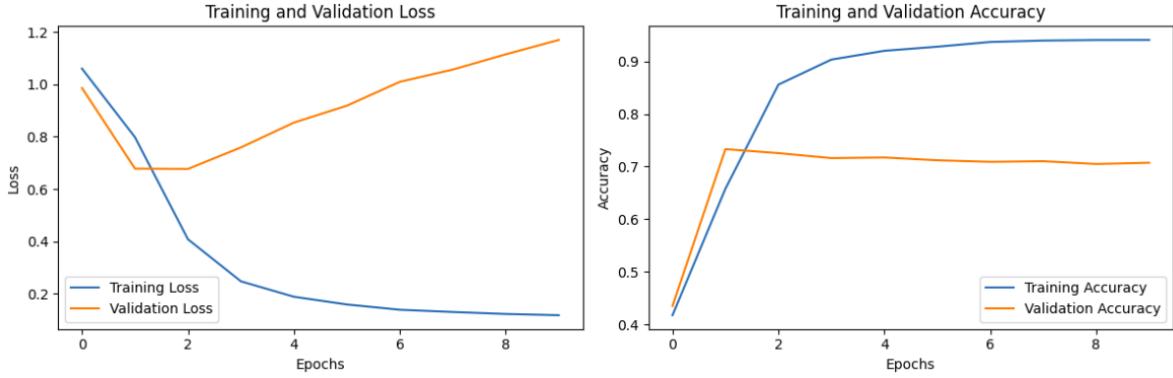


Figure 23: Još veći stepen preprilagođavanja.

2.2.5 L2 regularizacija

L2 regularizacija (takođe poznata kao Ridžova regularizacija)⁷ funkcioniše dodavanjem "penala" ili kazne funkciji gubitka tokom treninga. Kažnjavaju se velike vrednosti težina, te se na taj način model podstiče da što ravnomernije rasporedi vrednosti, umesto da se u velikoj meri oslanja na neki podskup karakteristika. Ovom modifikacijom se pored tačnosti optimizuju i težine na granama. λ je parametar koji kontroliše jačinu regularizacije: za veće vrednosti lambda, jači je efekat. U ovom slučaju je postavljen na 0.01, što se smatra slabom vrednošću.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
L2_regularized_layer (Dense)	(None, 64)	512,064
predictions (Dense)	(None, 3)	195
Total params:	512,259	
Trainable params:	512,259	
Non-trainable params:	0	

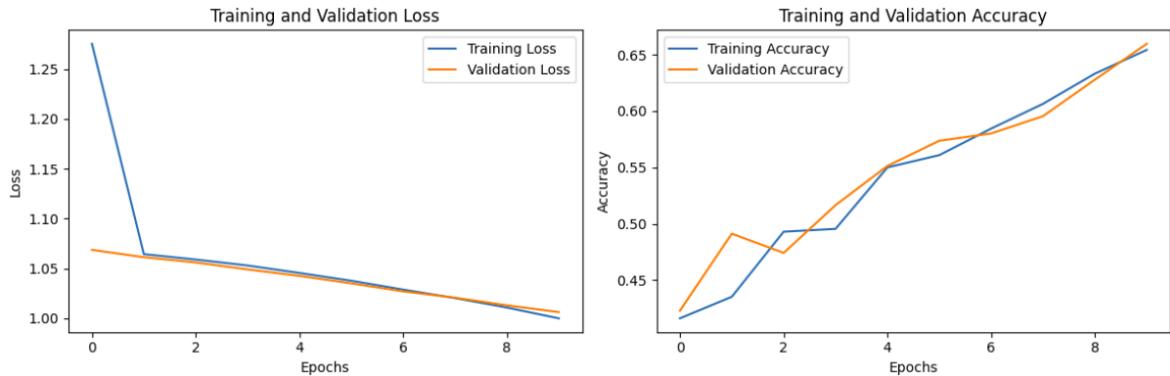


Figure 24: Sporija konvergencija.

2.2.6 Dropout regularizacija

Dropout regularizacija⁸ podrazumeva, tokom treninga, nasumično blokiranje ili "izbacivanje" određenog procenta perceptronu u datom sloju mreže postavljajući ih na nulu. Postiže se efekat sličan kao kod L2 regularizacije: mreža se manje oslanja na specifične neurone. Postavljena vrednost hiperparametra je 0.5, što znači da će u svakoj epohi biti nasumično izbačeno 50% neurona u unutrašnjem sloju na koji se dropout odnosi.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
first_layer (Dense)	(None, 64)	512,064
dropout_layer (Dropout)	(None, 64)	0
predictions (Dense)	(None, 3)	195
Total params:	512,259	
Trainable params:	512,259	
Non-trainable params:	0	

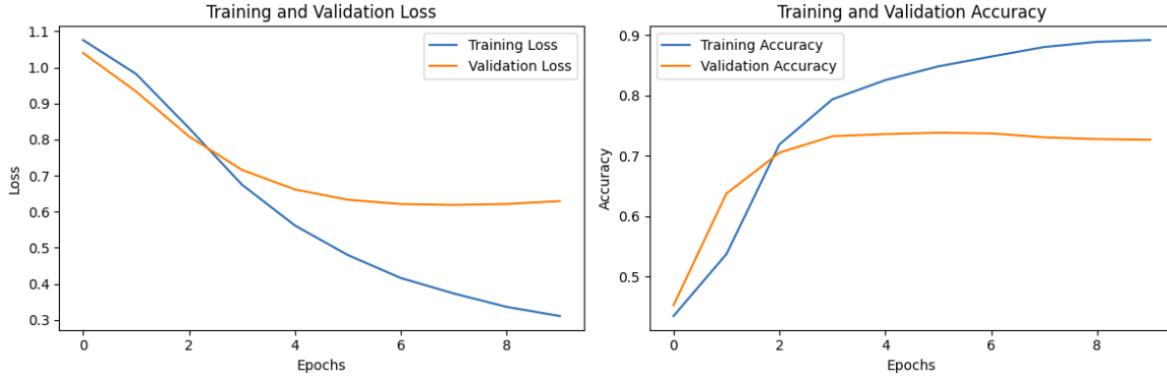


Figure 25: Bolje metrike, ponašanje slično kao kod modela sa smanjenim brojem neurona.

2.2.7 Kombinovanje L2 i Dropout

Kombinovanjem ove dve metode dobija sa poboljšana stabilnost i dobra generalizacija. Međutim, brzina konvergencije je znatno sporija, i model možda neće uhvatiti dovoljno kompleksne obrasce.

Layer (type)	Output Shape	Param #
input (InputLayer)	(None, 8000)	0
L2_regularized_layer (Dense)	(None, 64)	512,064
dropout_layer (Dropout)	(None, 64)	0
predictions (Dense)	(None, 3)	195
Total params:	512,259	
Trainable params:	512,259	
Non-trainable params:	0	

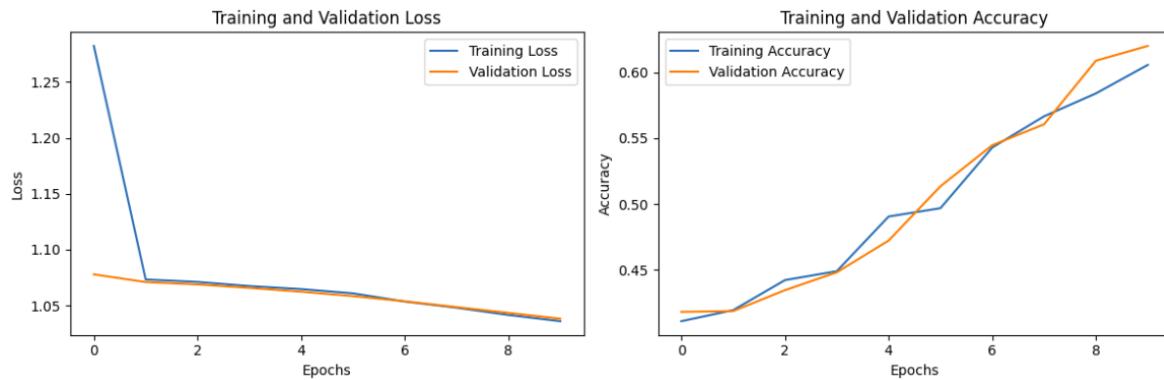


Figure 26: Veoma spora konvergencija.

2.2.8 Upoređivanje i primena na test skup

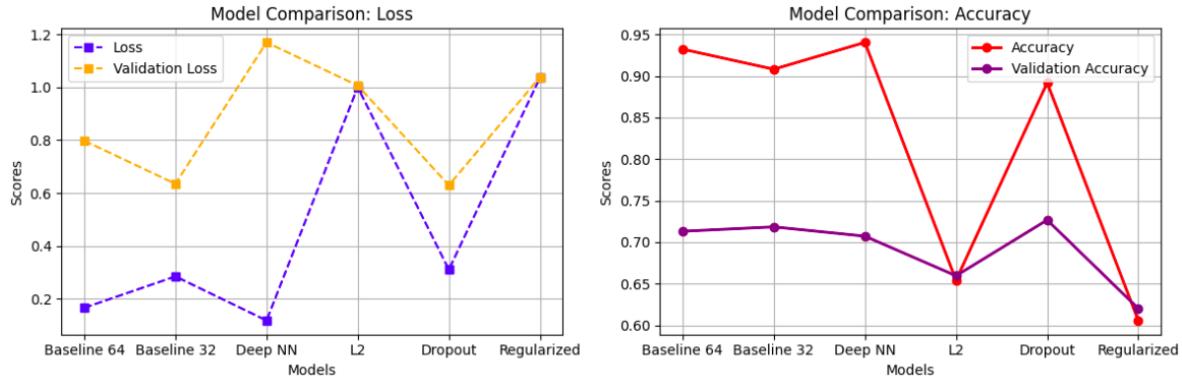


Figure 27: Tačnost i funkcija cilja na trening i validacionom skupu

Nakon analize grafikona za šest posmatranih arhitektura, očigledno je da dropout model pokazuje najveću tačnost i najmanji gubitak na validacionom skupu, što ga cini najboljim. Osnovni model sa 32 neurona se nalazi na drugom mestu sa jako bliskim metrikama, dok je duboka mreža pokazala najslabije performanse. Imajući u vidu ova zapažanja, evidentno je da bi potencijalni model za testiranje bio onaj koji uključuje dropout i smanjenje broja neurona.

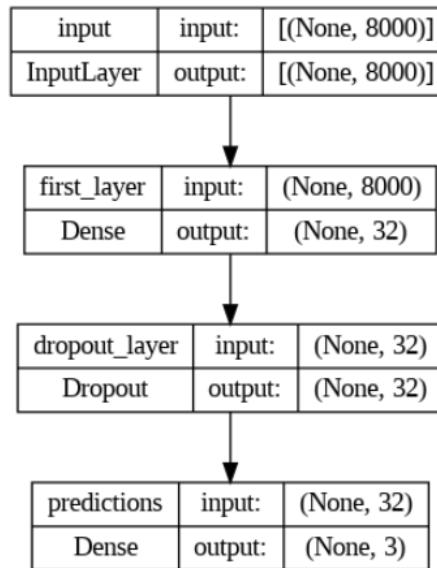


Figure 28: Arhitektura modela za primenu nad test skupom.

```

50/50 - 0s - loss: 0.4742 - accuracy: 0.8380 - 118ms/epoch - 2ms/step
Test loss: 0.474152535200119
Test accuracy: 0.8379571437835693
  
```

Dobijena je vrednost od 0.476 za funkciju gubitka i 0.8436 za tačnost. Neuronska mreža nadmašuje Naivni Bajes u pogledu tačnosti, pokazujući rezultat bolji za 1.36%.

3 Klasterovanje

3.1 Algoritam K-sredina

Algoritam K-sredina (K-means)⁹ korišćen je za grupisnje podataka u klastere na osnovu njihove sličnosti.

3.1.1 Klasterovanje u tri grupe i prikaz pravih klasa

Uzimajući u obzir da skup podataka sadrži tri različite klase, inicijalno se algoritam primenjuje sa unapred određenim brojem klastera postavljenim na 3. Izbor se zasniva na pretpostavci da će svaka klasa biti predstavljena pojedinačnim klasterom. Cilj je da se proceni u kojoj meri algoritam hvata osnovnu strukturu klasa koja je prisutna u skupu.

```
Cluster 0
phd magic youtube degree cream ice mushroom chemistry physic look

Cluster 1
effect research year sense solution science electron matter make thing

Cluster 2
sidebar hell physicsmemes mod username question explain stop check please
```

Figure 29: Dobijeni klasteri.

Na slici 29 prikazno je prvih deset reči iz svakog klastera. Iako je na osnovu nekih termina moguće identifikovati klase, prisustvo šuma je veoma uočljivo i nisu toliko jasne granice. Ipak, naslućuje se da je 0 - biologija, 1 - hemija, 2 - fizika.

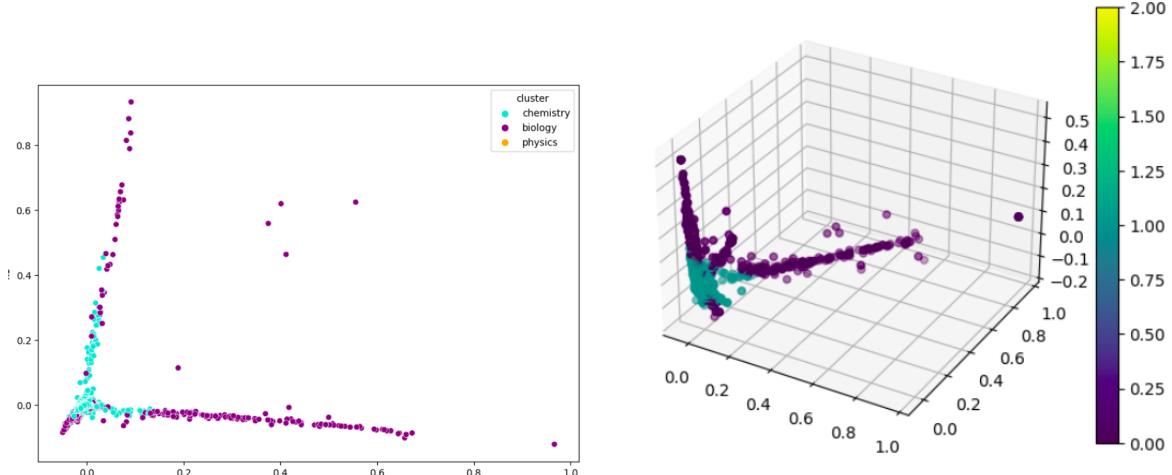


Figure 30: k-sredina + PCA

Redukovanjem dimenzije primenom PCA algoritma¹⁰ predstavljen je grupisan skup podataka u 2D i 3D prostoru. Međutim, bitno je napomenuti da zbog značajno niske količine opisane varijanse ($\tilde{0.02\%}$) ove vizuelizacije opisuju podatke mnogo jednostavnijim oblikom nego što oni zapravo jesu u prostoru. Na prikazima je primetno odusustvo klastera fizike, sto znači da ga čini da ga čini mali broj instanci.

Na slici 31 se nalazi prikaz pomoću tehnike t-SNE¹¹. Glavna motivacija iza korišćenja ovog algoritma je njegova sposobnost da bolje očuva strukturu nelinearnih i kompleksnih obrazaca u poređenju sa PCA. Konkretnije, postavlja akcenat na očuvanje lokalne strukture, što znači da je veća verovatnoća da će obližnje tačke u visokodimenzionalnom prostoru biti postavljene jedna blizu druge u niskodimenzionalnoj vizualizaciji. Na ovaj način je dobijen precizniji prikaz osnovne strukture.

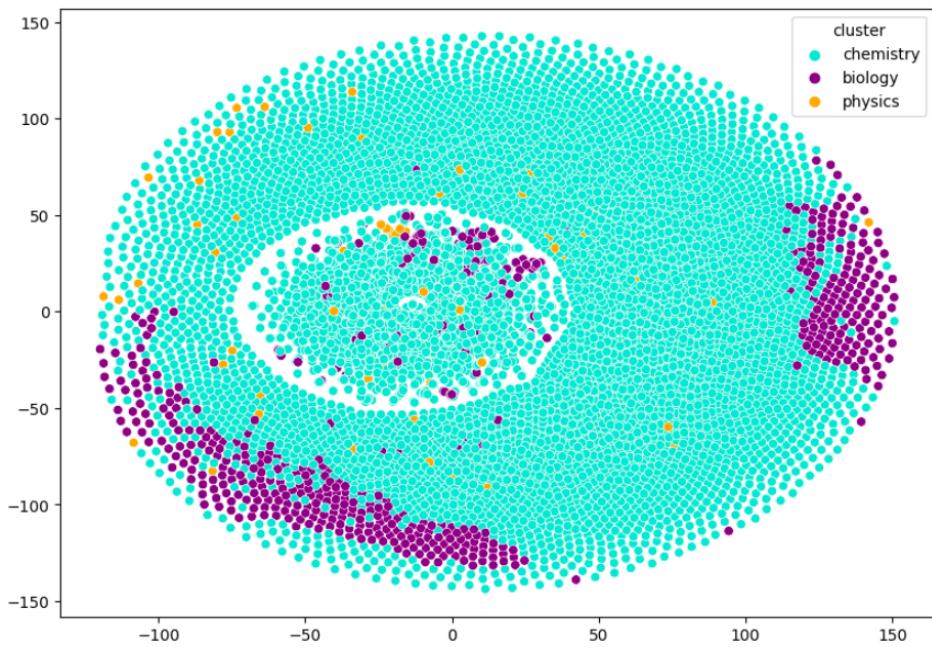


Figure 31: k-sredina + tSNE

Latentna semantička analiza (LSA)¹² je još jedna tehnika smanjenja dimenzionalnosti koja se pokazuje posebno korisnom za dokumenta. Ovaj algoritam u redukovanim prostoru preslikava dokumente sličnih značenja bliže jedan drugom, i na taj način obuhvata semantičku strukturu podataka. U nastavku će biti korišćen LSA za potrebe redukcije i vizuelizacije podataka(uz istu napomenu za varijansu kao kod PCA). Na sledećoj stranici je uporedni prikaz instanci u dodaljenim klasterima i originalnim klasama.

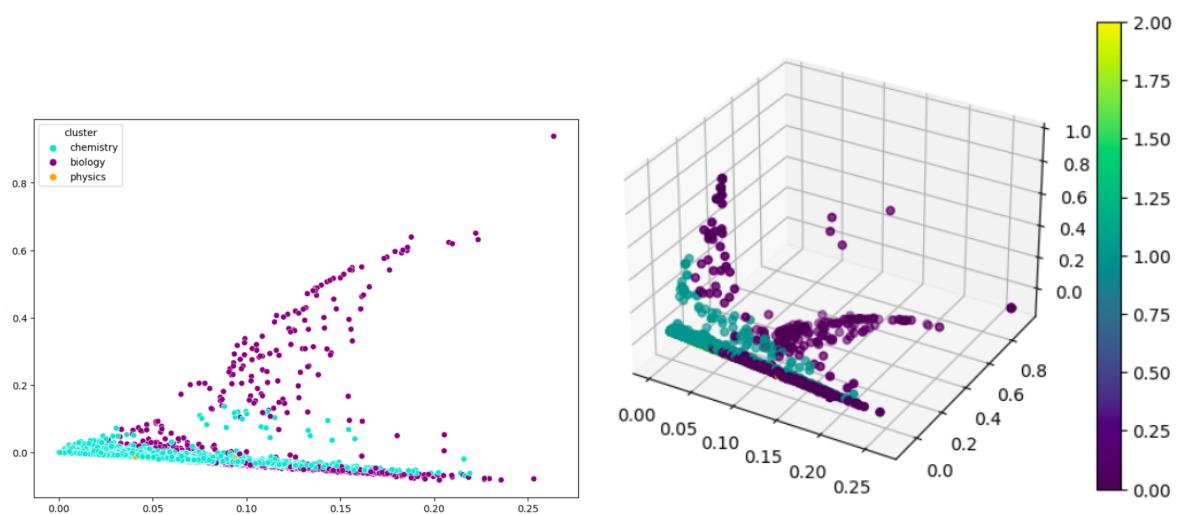


Figure 32: k-sredina + LSA

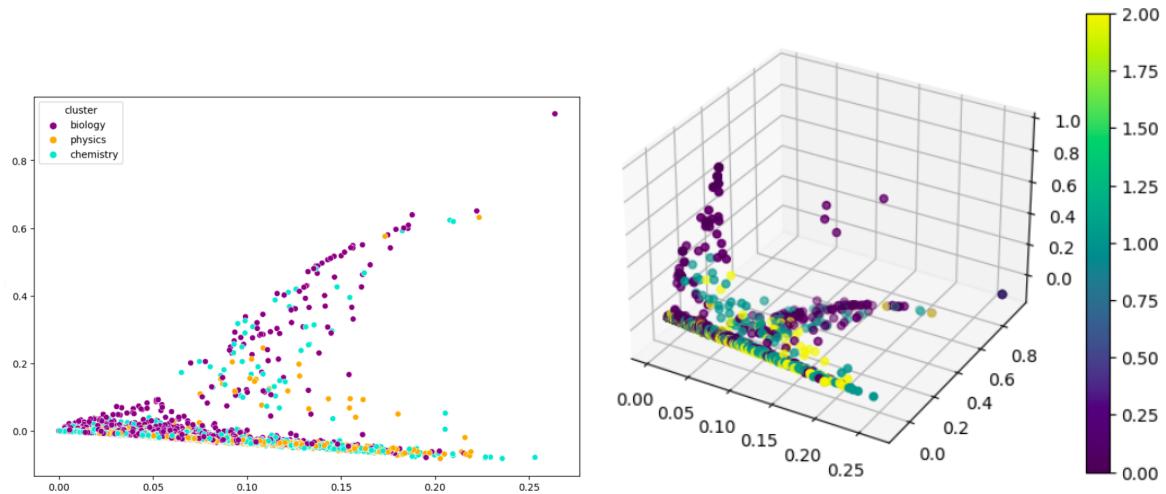


Figure 33: originalne klase + LSA

Upoređivanjem rezultata K-sredina sa originalnim podacima, čini se da algoritam uspeva da uhvati neki osnovni oblik podataka koji se odnose na biologiju i hemiju. Instance klase fizika u originalnim podacima ispoljavaju veoma nepravilan i kompleksan oblik, te su većinski grupisane u jednu od dve prethodno navedene grupe. Međutim, ako se malo preciznije pogleda raspodela klasa po ovim grupama, ne primećuje se nikakva korelacija(slika 34).

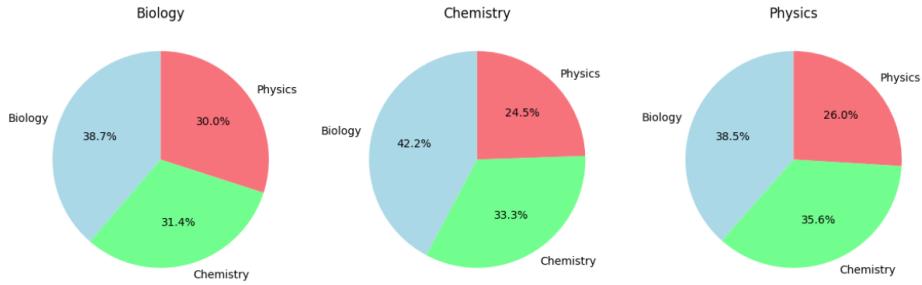


Figure 34: Raspodela klasa po klasterima.

3.1.2 Klasterovanje normalizovanih podataka u tri grupe

```

Cluster 0
message github compose sujal wikisummarizer wikisummarizerbot optout opt wikipedia wiki

Cluster 1
state reaction level particle mass chemistry thing make physic look

Cluster 2
thy snewd ebooks beep boop shakespeare fordo ncommands shakespeareinsult swapp

```

Figure 35: Dobijeni klasteri.

Na slici 35 prikazano je prvih deset reči iz svakog klastera dobijenog nad normalizovanim podacima. U ovom slučaju, samo grupa '1' ima termine koji su u nekoj vezi sa originalnim klasama, a preostale dve predstavljaju šum. Moguće je iskoristiti ovo saznanje za dodatno čišćenje trening skupa podataka prilikom pripreme za klasifikaciju (ponoviti ceo postupak na trening skupu i izbaciti instance grupisane kao '0' ili '2').

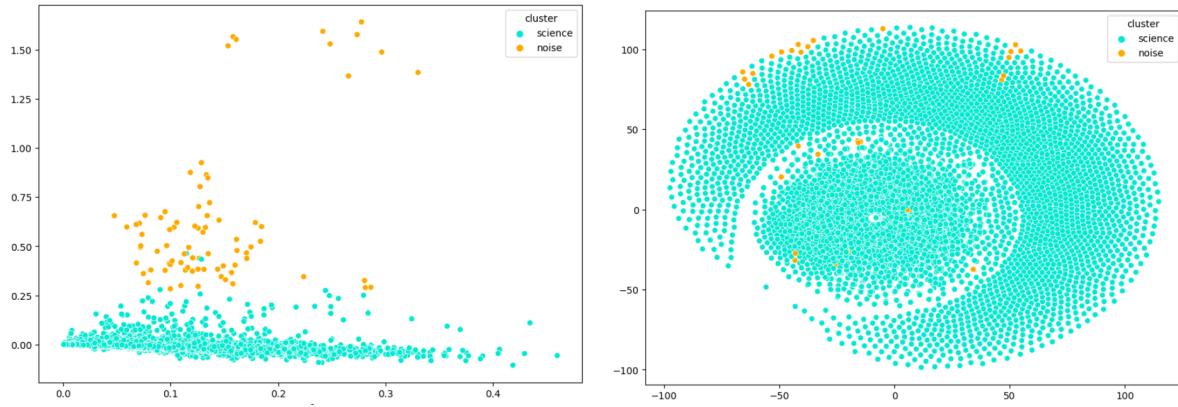
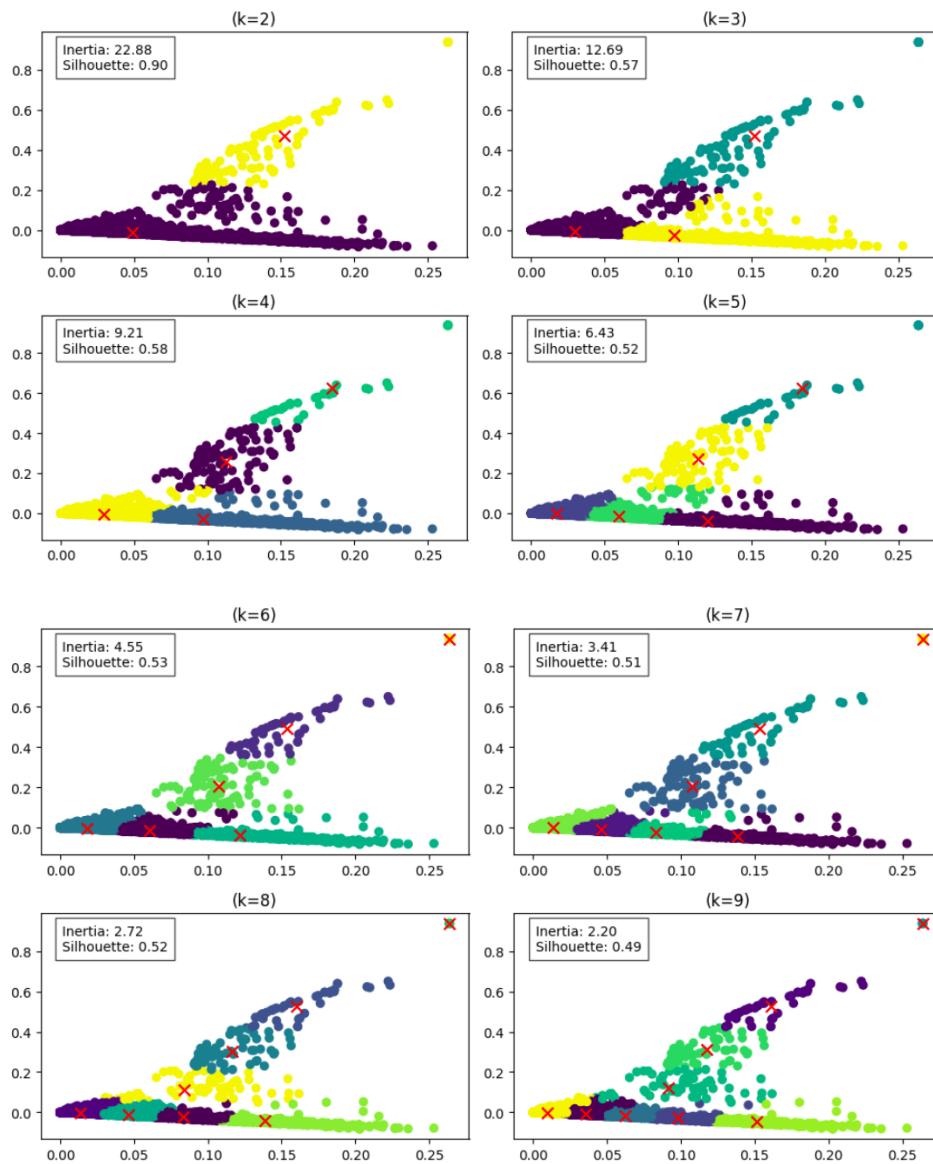


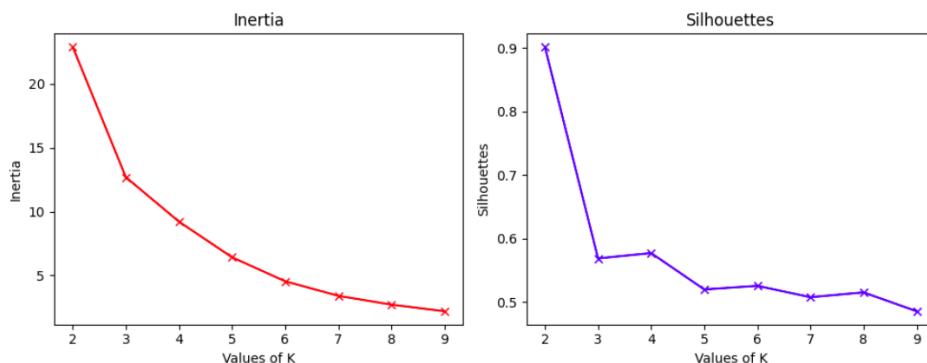
Figure 36: k-sredina na normalizovanim podacima: LSA(levo), tSNE(desno)

3.1.3 Traženje optimalnog broja klastera na redukovanim podacima

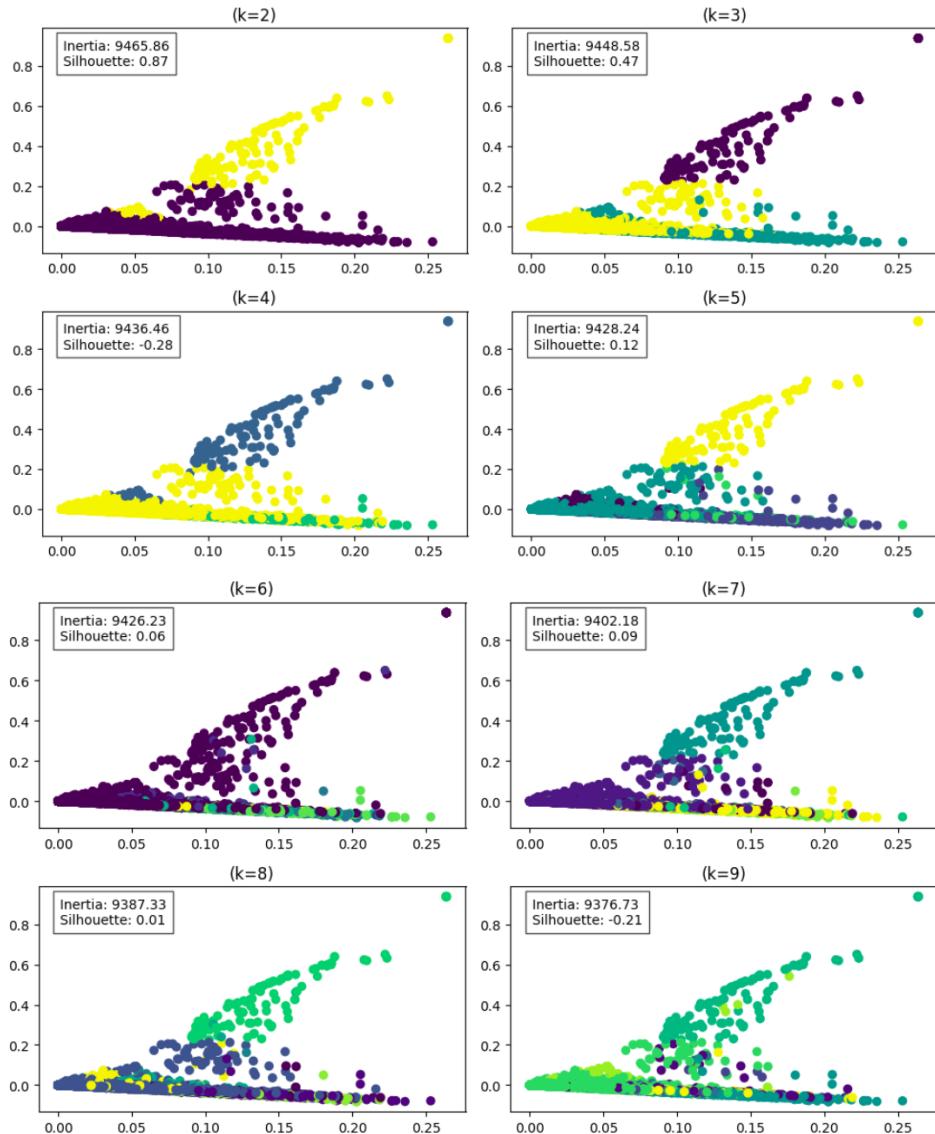
Podaci su redukovani na dve dimenzije. Za klasterovanje je korišćena naprednija verzija algoritma, K-sredina++, koja ravnomernije raspoređuje centroide. Dodatno, parametar `n_init` koji označava koliko puta će K-sredina biti pokrenut sa različitim inicijalizacijama, je postavljen na 10.



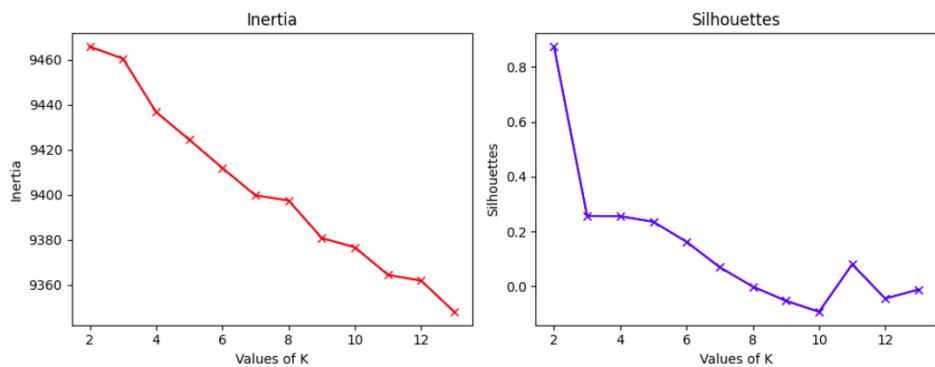
Određivanje optimalnog broja klastera može varirati u zavisnosti od metrike evaluacije koja se razmatra. Na osnovu analize grafika inercije i primene metode lakta, može se zaključiti da je optimalan broj klastera $k=3$. Iako sa povećanjem broja k grafik nije približno linearan, primetno je usporavanje smanjenja inercije. Sa druge strane, ubedljivo najbolja vrednost siluete je za $k=2$. Ukoliko uzmemo u obzir obe ove metrike, u proseku su optimalne vrednosti 2, 3 ili 4 klastera.



3.1.4 Traženje optimalnog broja klastera na originalnim podacima



Grafik inercije nije od nekog značaja, dok je na osnovu rezultata siluete očigledno da su klasteri najizraženiji za $k=2$.



3.2 Hijerarhijsko klasterovanje

Zbog prokletstva dimenzionalnosti koje je prisutno u hijerarhijskim algoritmima, neophodno je primeniti dodatne transformacije na podatke pre samog procesa klasterovanja.

3.2.1 Dodatna transformacija podataka

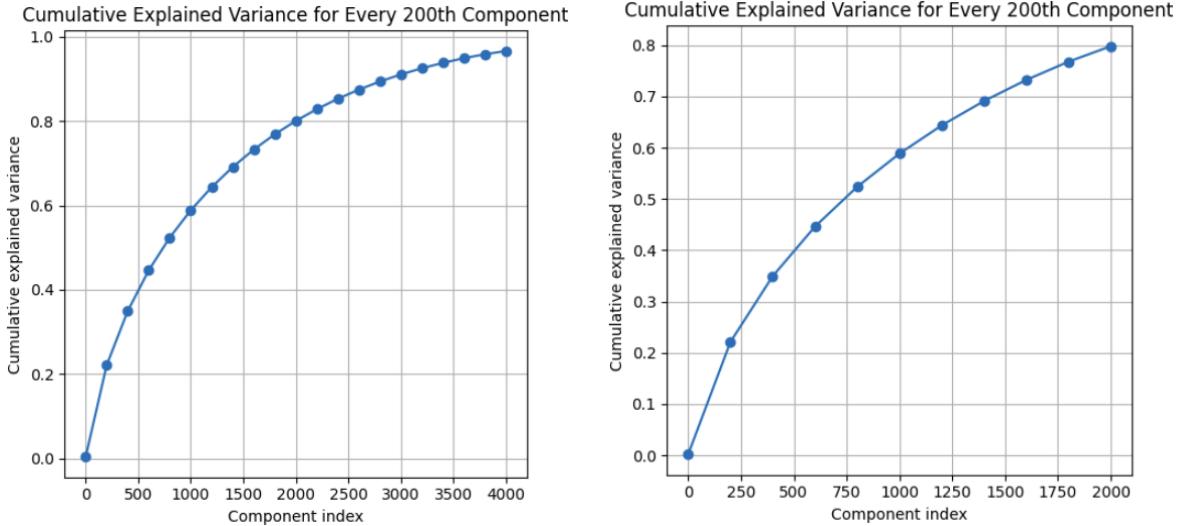


Figure 37: Ukupan udeo objasnjene varijanse nakon primene PCA (levo) i LSA (desno)

Transformacija bez značajnih gubitaka se ostvaruje u slučaju kada je opisana varijansa veća od 90%. Stoga, poželjno bi bilo raditi sa barem 3000 atributa. Međutim, kako podaci imaju puno šuma, i smanjenje dimenzionalnosti u što većoj količini je veoma značajno, moguce je smanjiti broj atributa na 2000 i time zadržati 80% ukupne varijanse originalnih podataka. Za konačnu redukciju koristimo LSA.

3.2.2 Odabir broja klastera na osnovu analize dendograma

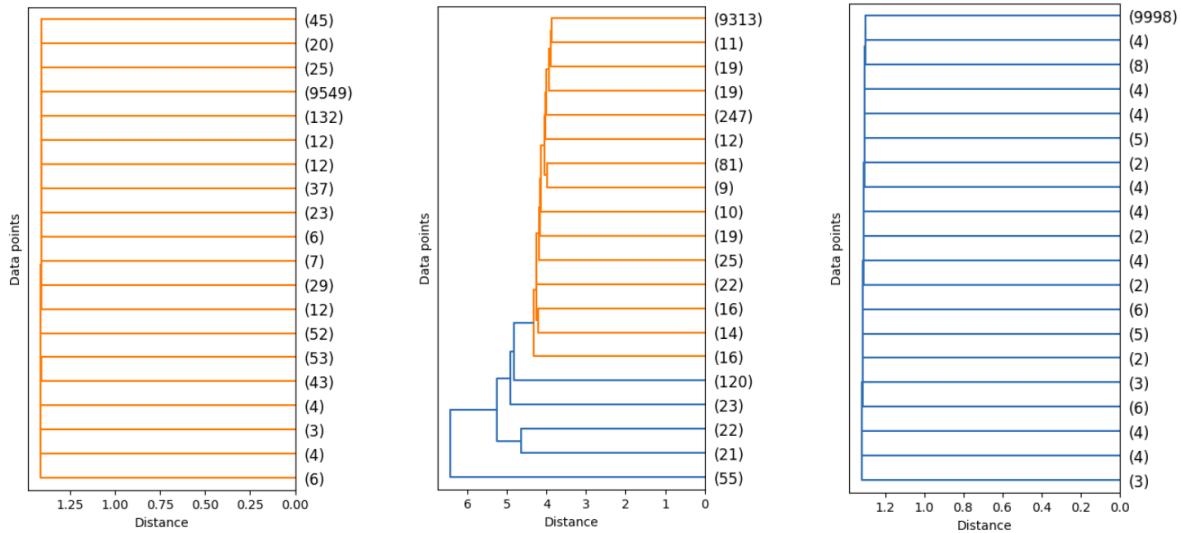


Figure 38: S leva na desno: potpuna metoda, metoda minimalne varijanse, prosečna metoda

Na slici 39 je prikaz dendograma¹³ koji su generisani na osnovu tri različite metode povezivanja¹⁴: potpuna, minimalne varijanse, prosečna. Na svakom je predstavljeno spajanje samo poslednjih 20 klastera. U sva tri slučaja su rastojanja između klastera jako mala. Očigledno je da potpuna i prosečna metoda pokazuju loše performanse i praktično grupišu sve podatke u jedan klanster. Nasuprot njima, metoda minimalne varijanse prikazuje nešto bolje rezultate sa uočljivijim grupama. Moguće je prepoznati 5 dovoljno udaljenih klastera, s tim da se u jednom nalazi više od 95% svih instanci.

Podrazumevano, algoritam povezivanja koristi Euklidsko rastojanje kao meru udaljenosti. Potencijalno, bolja opcija je korišćenje kosinusnog rastojanja, koje generalno dalje poboljšane performanse kada se radi sa tekstualnim podacima. Da bi ovo bilo moguće, izračunata je kosinusna sličnost izmedju instanci. Dobijene vrednosti sličnosti se pretvaraju u kosinusne udaljenosti oduzimanjem od 1.

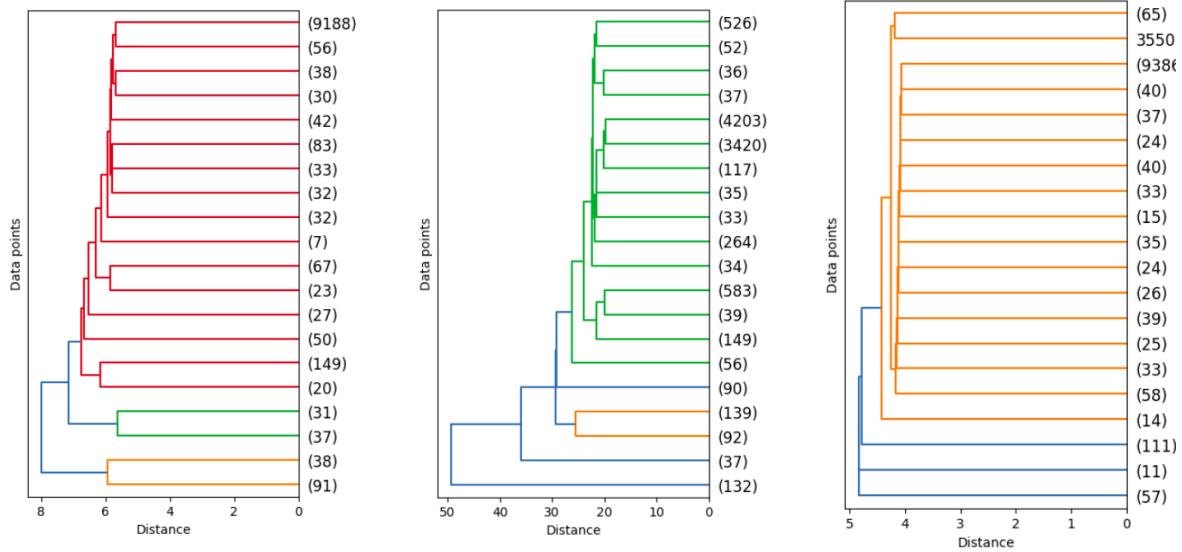


Figure 39: S leva na desno: potpuna metoda, metoda minimalne varijanse, prosečna metoda

Primetne su nešto bolje performanse algoritma na transformisanim podacima. Ponovo, metoda minimalne varijanse ispoljava najbolje rezultate. Prepostavka je da je zbog visine spajanja optimalan broj klastera jednak 3.

3.2.3 Vizuelizacija rezultata

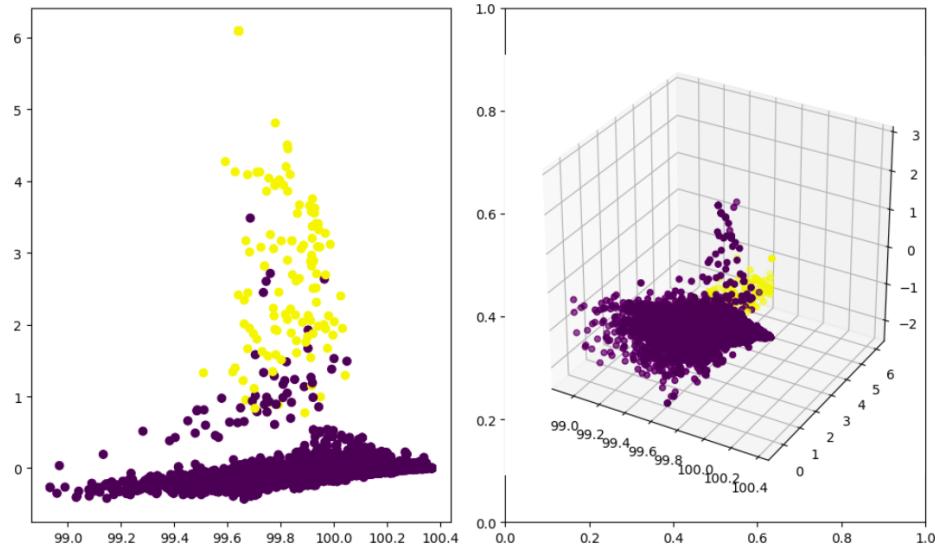


Figure 40: Silueta: 0.23384467312257823

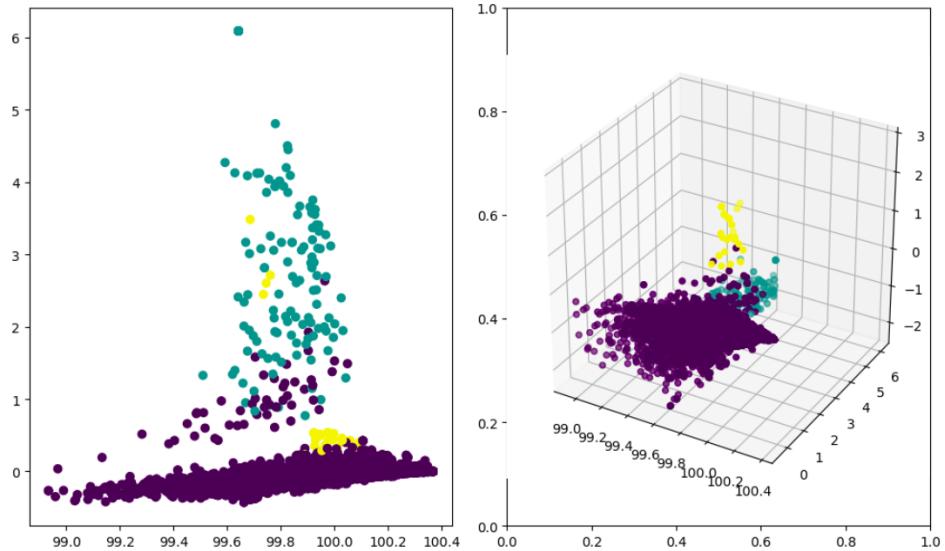


Figure 41: Silueta: 0.23573741415504804

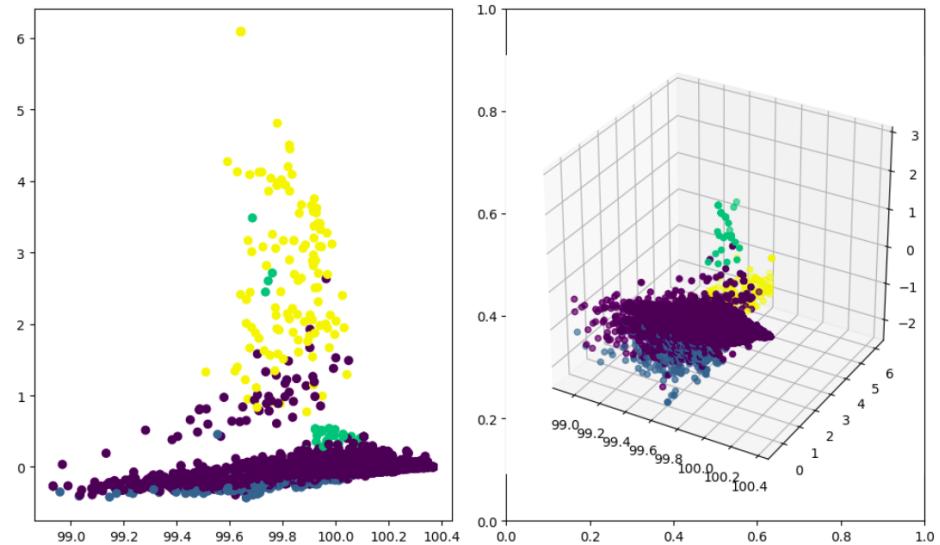


Figure 42: Silueta: 0.10192102786479605

Na osnovu vrednosti silueta, najbolje performanse se dostižu za $k = 3$. Čak i uz dodatna prilagođavanja reprezentacije podataka za ovaj algoritam, rezultati su lošiji u odnosu na K-sredina. Pretpostavka je da je ovo usled veće osetljivosti aglomerativnih metoda na šum.

4 Pravila pridruživanja

Apriori algoritam¹⁵ se koristi za traženje interesantnih obrazaca koji postoje u skupovima podataka. Kada se primeni na dokumenta, Apriori može pružiti uvid u istovremenu pojavu reči ili postojanje određenih fraza. Na ovaj način možemo identifikovati uobičajene obrasce između reči ili fraza koje se pojavljuju zajedno češće nego što bi se slučajno očekivalo.

Za primeni algoritma i dalje analize korišćen je SPSS Modeler¹⁶. Izvršeno je konvertovanje vrednosti kolona iz neprekidnih u binarne, kako bi reprezentacija bila pogodna za Apriori. Dijagram toka ispod ilustruje tok posla implementiran u SPSS Modeler, prikazujući korak preprocesiranja praćen Apriori algoritmom primjenjenim na ceo skup podataka i pojedinačne podskupove kategorija.

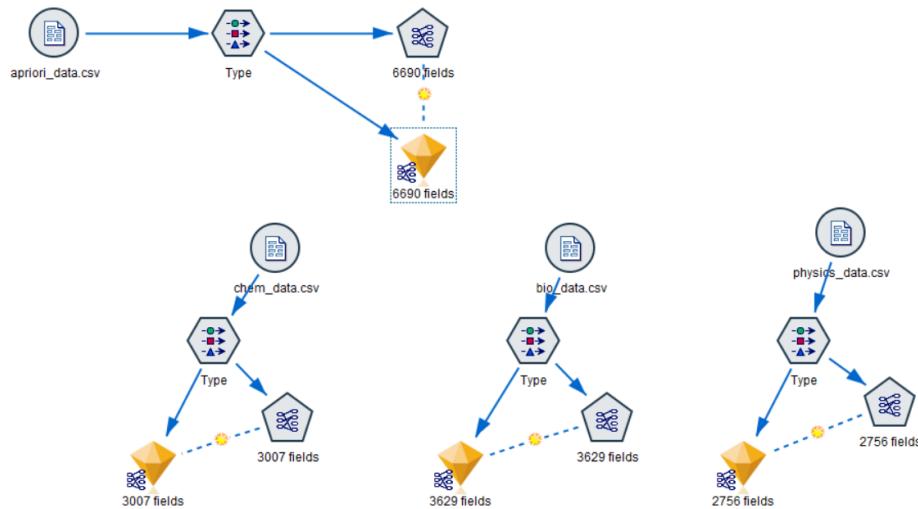


Figure 43: Dijagram toka

Zbog načina preprocesiranja podataka, očekivano je da skup podataka sadrži specifične obrasce i da ne postoje reči ili fraze koje se nalaze u velikom broju instanci. Zbog toga, podrška se postavlja na svega 1%. Fokus je na pronalaženju jakih veza između termina.

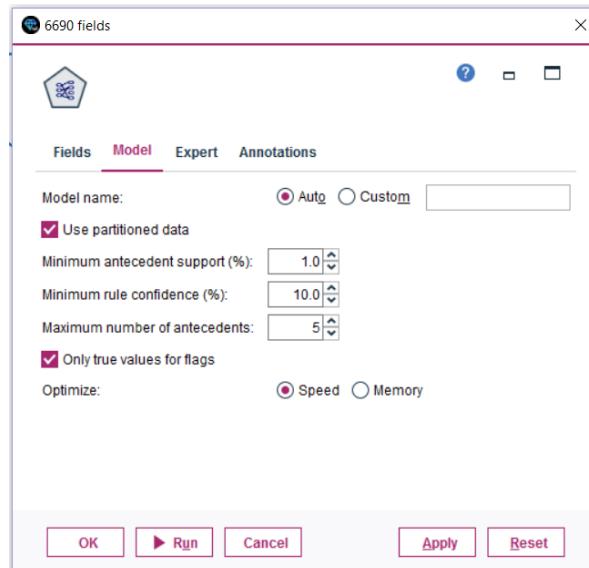


Figure 44: Konfiguracija

4.1 Primena apriori algoritma na ceo skup podataka

Consequent	Antecedent	Support %	Confidence %	Lift
compose	reddit	1.032	11.538	96.865
utm	reddit	1.032	10.577	96.865
vaccination	covid	1.042	10.476	81.182
mrna	vaccine	1.072	32.407	64.014
comment	reddit	1.032	37.5	58.119
mechanic	quantum	1.082	25.688	57.507
salary	phd	1.092	10.0	55.967
classical	quantum	1.082	11.009	48.22
immunity	vaccine	1.072	14.815	46.639
message	reddit	1.032	14.423	45.406
vaccinated	covid	1.042	12.381	43.009
relativity	general	1.032	10.577	42.621
travel	speed	1.082	24.771	39.609
journal	data	1.142	12.174	38.325
oxidation	state	1.112	12.5	38.159
vaccine	covid	1.042	40.0	37.311
covid	vaccine	1.072	38.889	37.311
cat	quantum	1.082	11.009	36.969
heart	covid	1.042	16.19	36.245
bullet	speed	1.082	18.349	36.244

Figure 45: Prvih 20 pravila sortiranih prema liftu

Consequent	Antecedent	Support %	Confidence %	Lift
thing	make	2.015	12.808	3.762
thing	done	1.122	12.389	3.639
thing	higher	1.132	12.281	3.607
thing	experience	1.221	12.195	3.582
thing	ever	1.142	12.174	3.576
physic	basically	1.062	10.28	3.523
thing	nit	1.261	11.811	3.469
thing	certain	1.013	11.765	3.455
thing	matter	1.876	11.64	3.419
thing	word	1.211	11.475	3.37
thing	low	1.221	11.382	3.343
thing	specie	1.062	11.215	3.294
thing	research	1.886	11.053	3.246
thing	science	1.906	10.938	3.212
thing	world	1.102	10.811	3.175
thing	non	1.102	10.811	3.175
thing	based	1.201	10.744	3.155
thing	time	1.112	10.714	3.147
thing	big	1.39	10.714	3.147
thing	particle	1.499	10.596	3.112
thing	general	1.032	10.577	3.106

Figure 46: Poslednjih 20 pravila sortiranih prema liftu

Agloritam je pronašao 356 pravila pridruživanja. Iako se sva pravila mogu smatrati interesantnim jer imaju lift znatno veći od 1, očigledno je da neka ne nose značajne informacije jer im se u telu nalaze veoma opšti izrazi kao što je *thing*.

4.2 Primena apriori algoritma na posebne klase

Consequent	Antecedent	Support %	Confidence %	Lift
spectrometry	mass	1.148	10.256	87.103
office	hour	1.148	20.513	87.103
rubbing	alcohol	1.089	13.514	76.509
tailor	job	1.178	12.5	70.771
helium	oxygen	1.266	16.279	69.125
wick	flame	1.178	10.0	67.94
manager	job	1.178	10.0	67.94
culture	phd chemistry	1.178	10.0	67.94
isopropyl	alcohol	1.089	29.73	67.328
hemoglobin	oxygen	1.266	11.628	65.833
octet	rule	1.119	18.421	62.576
peak	show	1.207	14.634	62.14
lattice	crystal	1.384	12.766	61.951
sanitize	alcohol	1.089	10.811	61.207
upon	come	1.089	10.811	61.207
valence	electron	1.649	21.429	60.661
reacts	chlorine	1.001	20.588	58.282
ocl	mass	1.148	10.256	58.068
dioxide	carbon	1.472	32.0	57.213
hydrochloric	chlorine	1.001	11.765	57.092

Figure 47: Pravila pridruživanja na skupu hemije

Dobijen je jako lep uvid u zavisnosti termina u domenu hemije. Originalni skup podataka označen ovom klasom nema prisustvo šuma izazvano opštim izrazima.

Consequent	Antecedent	Support %	Confidence %	Lift
pitched	higher	1.081	11.111	92.467
teneral	adult	1.081	13.333	92.467
ncommands	word	1.106	15.217	90.457
fordo	word	1.106	15.217	90.457
swapp	word	1.106	15.217	90.457
shakespeare	word	1.106	15.217	90.457
shakespeareinsult	word	1.106	15.217	90.457
thy	word	1.106	15.217	90.457
cream	ice	1.057	88.636	89.955
creationevolution	reddit	1.154	12.5	86.688
compose	reddit	1.154	12.5	86.688
compromised	immune	1.178	10.204	84.918
meal	food	1.202	10.0	83.22
opt	reddit	1.154	12.5	74.304
reindeer	drink	1.033	13.953	72.576
model	nit	1.033	13.953	72.576
bull	shit	1.226	11.765	69.933
funding	scientific	1.033	11.628	69.12
beep	reddit	1.154	14.583	67.424
boop	reddit	1.154	14.583	67.424
handled	bug	1.178	12.245	63.689
integrated	immune	1.178	12.245	63.689
web	spider	1.009	16.667	63.045
integrated	mRNA	1.226	11.765	61.191
vaccination	vaccine	1.009	19.048	60.967
covid				
tea	drink	1.033	11.628	60.48
strawberry	ice	1.057	43.182	59.893

Figure 48: Pravila pridruživanja na skupu biologije

Mogu se videti neke asocijacije reči u oblasti biologije, ali u ovom primeru ima dosta neinteresantnih reči.

Sort by: Lift 5960 of 5960

Consequent	Antecedent	Support %	Confidence %	Lift
alternatively	calculation	1.033	15.385	96.769
redirected	calculation	1.033	19.231	96.769
forum	calculation	1.033	19.231	96.769
assisting	calculation	1.033	19.231	96.769
predicting	value	1.033	11.538	96.769
distinct	open	1.033	11.538	96.769
distinct	towards	1.033	11.538	96.769
witness	travel	1.033	11.538	96.769
witness	speed	1.033	11.538	96.769
alternatively	asking	1.073	14.815	93.185
hgdk	name	1.073	11.111	93.185
high efficiency	name	1.073	14.815	93.185
redirected	asking	1.073	18.519	93.185
forum	asking	1.073	18.519	93.185
assisting	asking	1.073	18.519	93.185
iossmf	name	1.073	18.519	93.185
app	name	1.073	18.519	93.185
inflaton	field	1.073	11.111	93.185
worthwhile	done	1.073	11.111	93.185
captured	capture	1.073	14.815	93.185
flue	capture	1.073	14.815	93.185
adsorbent	capture	1.073	18.519	93.185
thermodynamic	capture	1.073	18.519	93.185
renewables	capture	1.073	29.63	93.185
div	equal	1.153	10.345	86.759
symmetric	equal	1.153	10.345	86.759
counterintuitive	observation	1.153	10.345	86.759
cart	kinetic	1.153	20.69	86.759

Figure 49: Pravila pridruživanja na skupu fizike

Zaključak je isti kao i u prethodnom primeru.

5 Zaključak

Kod pristupa ovakvom zadatku ključno je imati razumevanje porekla, značenja i konteksta podataka. Pored odabira pogodne reprezentacije, sam korak pretprocesiranja će imati najznačajniji uticaj na rezultate klasifikacije, za šta je ovaj skup podataka i namenjen. Zbog toga, pored standardnih metoda, možda bi trebalo razmotriti i primenu klasterovanja i pravila pridruživanja u koraku pretprocesiranja zbog njihove mogućnosti detekcije šuma.