Matematički fakultet / Informatika

Seminarski rad iz predmeta: Istraživanje podataka 2

Klasterovanje tekstualnih oglasa iz skupa Farm-Ads

Analiza i poređenje algoritama klasterovanja

Autor: Lazar Dunjić
Broj indeksa: 265/2021
Profesor: Mirjana Maljković Ružičić
Datum: 28. januar 2026.

Sažetak

Ovaj rad prikazuje potpun postupak klasterovanja tekstualnih oglasa iz skupa podataka Farm-Ads. Primenjeno je 8 različitih algoritama klasterovanja (K-Means, Agglomerative, DBSCAN, Mean Shift, BIRCH) na 6 varijanti skupova atributa (TF-IDF, Count Vectorizer, PCA, SVD). Evaluacija je izvršena pomoću Silhouette Score, Davies-Bouldin i Calinski-Harabasz metrika. Rezultati pokazuju da K-Means sa parametrom K=3 na Count Vectorizer reprezentaciji daje najbolje performanse (Silhouette=0.914), dok DBSCAN identifikuje visoko-pouzdane klastere ali sa velikim brojem autlajera.

Sadržaj

1 Uvod

Klasterovanje tekstualnih podataka predstavlja fundamentalan zadatak u oblasti rudarenja podataka i mašinskog učenja, sa širokom primenom u kategorizaciji dokumenata, detekciji nepoželjnih(*spam*) poruka, organizaciji rezultata pretrage i analizi sentimenta. Za razliku od klasifikacije koja zahteva prethodno označene podatke (*supervised learning*), klasterovanje je tehnika nenadgledanog učenja koja automatski grupiše slične dokumente na osnovu njihovih karakteristika, bez potrebe za eksplicitnim oznakama tokom treniranja.

1.1 Kontekst i motivacija

U današnjoj eri velikih podataka, ogromna količina tekstualnih informacija generiše se svakodnevno kroz veb stranice, društvene mreže, internet oglase i druge digitalne platforme. Automatska organizacija i kategorizacija ovih podataka postaje kritična za efikasno upravljanje informacijama. U kontekstu internet oglasa, poseban izazov predstavlja razlikovanje relevantnih oglasa od onih koji nisu u skladu sa specifičnom tematikom ili kategorijom.

Skup podataka Farm-Ads sadrži tekstualne oglase prikupljene sa interneta, pri čemu je cilj identifikovati oglase relevantne za poljoprivredne teme. Originalni podaci sadrže binarne oznake (1 za relevantne, -1 za irelevantne), ali u ovom radu koristimo pristup klasterovanja kako bismo istražili prirodnu grupisanost dokumenata bez prethodnog korišćenja ovih oznaka u procesu modelovanja. Oznake se koriste isključivo za naknadnu evaluaciju i interpretaciju dobijenih klastera.

1.2 Priroda podataka i problem

Ulazni skup podataka Farm-Ads sastoji se od 4143 tekstualna dokumenta, gde svaki dokument predstavlja jedan oglas. Podaci su inicijalno označeni sa:

- **Pozitivni primeri (oznaka 1):** 2210 oglasa relevantnih za poljoprivredne teme (53.3%)
- **Negativni primeri (oznaka -1):** 1933 oglasa irelevantnih za poljoprivredne teme (46.7%)

Tekstualni sadržaj oglasa pokazuje značajnu heterogenost u dužini, stilu pisanja i upotrebljenoj terminologiji. Prosečna dužina teksta je 3220 karaktera (451 reč), sa velikim rasponom od nekoliko desetina do nekoliko hiljada karaktera. Ova varijabilnost predstavlja

dodatni izazov za algoritme klasterovanja, jer moraju da identifikuju semantičku sličnost uprkos različitim strukturama dokumenata.

1.3 Istraživačka pitanja i hipoteze

Osnovna istraživačka pitanja ovog rada su:

1. **Da li algoritmi klasterovanja mogu automatski identifikovati prirodnu grupisanost oglasa?** Očekuje se da dokumenti sa sličnom tematikom budu grupisani zajedno, bez obzira na njihove originalne oznake.
2. **Koji algoritam klasterovanja i koja reprezentacija atributa daju najbolje rezultate?** Poređenjem 8 različitih algoritama (K-Means, DBSCAN, Agglomerative, Mean Shift, BIRCH) na 6 skupova atributa (različite TF-IDF konfiguracije, Count Vectorizer, PCA, SVD), cilj je identifikovati optimalan pristup za ovaj tip podataka.
3. **Koliko dobro se klasteri dobijeni nenadgledanim učenjem poklapaju sa originalnim oznakama?** Iako se oznake ne koriste tokom klasterovanja, njihova naknadna analiza može pokazati da li algoritmi uspevaju da otkriju semantičku razliku između relevantnih i irelevantnih oglasa.
4. **Kakva je interpretabilnost dobijenih klastera?** Analiza top reči po klasterima može otkriti karakteristične teme i terminologiju koja definiše svaku grupu.

Polazna hipoteza je da će algoritmi klasterovanja otkriti višestruke tematske grupe unutar podataka, pri čemu će neke od njih jasno odgovarati oglasima relevantnim za poljoprivredu, dok će druge predstavljati različite kategorije irelevantnih sadržaja.

1.4 Izazovi analize i metodološki izbori

Klasterovanje tekstualnih podataka suočava se sa nekoliko specifičnih izazova:

Visoka dimenzionalnosct: Tekstualni podaci, kada se vektorizuju, rezultuju u veoma visokim dimenzijama (stotine ili hiljade atributa). Ovo može dovesti do "prokletstva dimenzionalnosti"(*curse of dimensionality*) gde tradicionalne mere udaljenosti postaju manje efektivne.

Retkost: TF-IDF i slične reprezentacije stvaraju retke matrice gde je većina vrednosti nula. Ovo zahteva algoritme koji mogu efikasno raditi sa retkim podacima.

Skalabilnost: Sa 4143 dokumenta, potrebni su algoritmi koji mogu procesirati ovaj obim podataka u razumnom vremenskom okviru.

Izbor broja klastera: Za ne-deterministički određen broj klastera (kao kod K-Means), potrebno je koristiti metode kao što su Elbow metoda ili Silhouette analiza.

Metodološki pristup ovog rada uključuje:

- Preprocesiranje: TF-IDF vektorizacija sa različitim parametrima
- Redukcija dimenzionalnosti: PCA i Truncated SVD
- Kreiranje različitih skupova atributa za testiranje robusnosti
- Primena 8 algoritama klasterovanja sa različitim parametrima
- Sveobuhvatna evaluacija pomoću više metrika
- Vizualizacija rezultata u 2D i 3D prostoru

1.5 Ciljevi i doprinos rada

Cilj ovog rada je višestruk:

Metodološki doprinos: Demonstracija kompletnog toka za klasterovanje tekstualnih podataka, od sirovog teksta do interpretabilnih rezultata, sa jasnim obrazloženjem svakog koraka.

Empirijsko poređenje: Sistematska evaluacija 8 algoritama na 6 različitih reprezentacija podataka (ukupno 48 kombinacija), sa ciljem identifikovanja najboljeg pristupa za ovaj tip problema.

Praktična primena: Pruža uvid u automatsku kategorizaciju oglasa, što je korisno za filtriranje neželjene pošte, sisteme preporuka i strukturisanu organizaciju sadržaja.

Reproducibilnost: Detaljna dokumentacija svake faze procesa omogućava ponavljanje analize na drugim sličnim skupovima podataka.

1.6 Struktura rada

Rad je organizovan u šest glavnih poglavlja:

Poglavlje 2 (Podaci i priprema) opisuje izvor podataka, osnovnu statistiku, korake preprocesiranja, kreiranje različitih skupova atributa i inicijalnu vizualizaciju.

Poglavlje 3 (Metod) detaljno objašnjava primenjene algoritme klasterovanja, metrike evaluacije, i procedure za određivanje optimalnog broja klastera.

Poglavlje 4 (Rezultati) prikazuje dobijene rezultate, poređenje algoritama, vizualizacije i analizu najboljeg modela.

Poglavlje 5 (Diskusija) interpretira nalaze, razmatra ograničenja i predlaže pravce za buduća istraživanja.

Poglavlje 6 (Zaključak) sumira glavne doprinose rada i izvodi zaključke.

2 Podaci i priprema

U ovom poglavlju opisujemo izvor, strukturu i pripremu podataka, kao i razloge za svaku odluku u preprocesiranju. Cilj je obezbediti stabilnu analitičku osnovu za klasterovanje, uz minimizovanje šuma i maksimizovanje informativnosti atributa.

2.1 Izvor i struktura podataka

Skup podataka Farm-Ads preuzet je kao tekstualni fajl gde svaki red predstavlja jedan dokument (oglas). Format podataka je:

```
[LABEL] [TEKSTUALNI_SADRŽAJ]
```

gde je `LABEL` celobrojna vrednost (1 ili -1), a `TEKSTUALNI_SADRŽAJ` je string koji sadrži reči oglasa razdvojene space karakterima. Sve reči su već pretprocesirane (lowercase, tokenizovane).

Inicijalno učitavanje pokazuje:

- **Ukupan broj dokumenata:** 4143
- **Distribucija oznaka:** 1 (relevantni): 2210 (53.3%), -1 (irelevantni): 1933 (46.7%)
- **Format:** Tekstualni fajl bez zaglavlja, vrednosti su odvojene razmacima

2.2 Osnovna statistika podataka

Pre preprocesiranja, izvršena je osnovna statistička analiza tekstualnih dokumenata. Tabela 1 prikazuje ključne metrike.

Tabela 1: Osnovna statistika Farm-Ads podataka

Metrika	**Vrednost**
Prosečna dužina teksta	3220 karaktera
Prosečan broj reči	451 reči
Prosečan broj jedinstvenih reči	197 reči
Minimalna dužina	$\sim$50 karaktera
Maksimalna dužina	$>$10000 karaktera

Zapažanja: Velika varijabilnost u dužini dokumenata (od 50 do preko 10000 karaktera) ukazuje na heterogenost sadržaja. Prosečan broj jedinstvenih reči (197) u odnosu

na ukupan broj reči (451) sugeriše prisustvo ponavljanja, što je tipično za reklamni tekst. Ovo opravdava upotrebu TF-IDF vektorizacije koja će ponderisati reči prema njihovoj specifičnosti.

2.3 Preprocesiranje – TF-IDF vektorizacija

Za transformaciju tekstualnih podataka u numeričke vektore primenjena je **TF-IDF (Term Frequency-Inverse Document Frequency)** vektorizacija. Ova metoda dodeljuje veću težinu rečima koje su česte u dokumentu, ali retke u skupu podataka, čime naglašava diskriminativne termine.

Kreirana su **dva TF-IDF skupa**:

TF-IDF Full (500 features):

- `max_features=500`: Zadržava 500 najčešćih termina
- `min_df=2`: Ignoriše termine koji se pojavljuju u manje od 2 dokumenta
- `max_df=0.8`: Ignoriše termine koji se pojavljuju u više od 80% dokumenata
- `ngram_range=(1,2)`: Uključuje unigrame i bigrame
- Rezultujuća dimenzija: 4143×500

TF-IDF Reduced (100 features):

- `max_features=100`: Zadržava samo 100 najčešćih termina
- `min_df=5`: Ignoriše termine koji se pojavljuju u manje od 5 dokumenata
- `max_df=0.7`: Ignoriše termine koji se pojavljuju u više od 70% dokumenata
- `ngram_range=(1,1)`: Samo unigrami
- Rezultujuća dimenzija: 4143×100

Count Vectorizer (200 features):

Pored TF-IDF, kreiran je i skup pomoću Count Vectorizera koji broji pojavljivanja termina bez IDF ponderisanja:

- `max_features=200`
- `min_df=2, max_df=0.8`

- Rezultujuća dimenzija: 4143×200

Opravdanje izbora parametara: Parametri `min_df` i `max_df` filtriraju ekstremno retke i ekstremno česte termine koji obično ne doprinose diskriminaciji. Bigrami (`n-grams=2`) omogućavaju hvatanje fraznih obrazaca (*organic farm*, *livestock feed*), što je posebno važno za tematsku analizu poljoprivrednih oglasa.

2.4 Redukcija dimenzionalnosti

Visokodimenzionalni podaci mogu dovesti do problema poznatog kao "prokletstvo dimenzionalnosti"(*curse of dimensionality*), gde se pojam blizine i udaljenosti gubi u visokim dimenzijama. Primenjene su dve metode linearne redukcije:

PCA (Principal Component Analysis) – 50 komponenti:

- Ulaz: TF-IDF Full (500 features)
- Izlaz: 50 glavnih komponenti
- Objašnjena varijansa: 55.81%
- Rezultujuća dimenzija: 4143×50

PCA pronalazi ortogonalne pravce maksimalne varijanse u podacima. Prvih 50 komponenti objašnjava više od polovine ukupne varijanse, što ukazuje da podaci leže u relativno niskom dimenzionalnom podprostoru.

Truncated SVD (Latent Semantic Analysis) – 50 komponenti:

- Ulaz: TF-IDF Full (500 features)
- Izlaz: 50 SVD komponenti
- Objašnjena varijansa: 55.63%
- Rezultujuća dimenzija: 4143×50

SVD je matematički sličan PCA, ali je optimizovan za retke matrice (*sparse matrices*) koje nastaju kod TF-IDF vektorizacije. U kontekstu tekstualnih podataka, SVD se često naziva *Latent Semantic Analysis* (LSA) i omogućava otkrivanje skrivenih koncepata u dokumentima.

2.5 Kreiranje skupova atributa

Kako bi se testirala robusnost algoritama klasterovanja i identifikovala optimalna reprezentacija podataka, kreirano je **6 različitih skupova atributa**. Tabela 2 prikazuje sve skupove.

Tabela 2: Skupovi atributa korišćeni u analizi

Naziv	**Opis**	**Dimenzija**
TF-IDF Full	Puna TF-IDF vektorizacija sa bigramima	4143×500
TF-IDF Reduced	Redukovana TF-IDF, samo unigrami	4143×100
Count Vec	Count Vectorizer bez IDF ponderisanja	4143×200
TF-IDF + Stats	TF-IDF Reduced + statistički atributi	4143×103
PCA	PCA redukcija TF-IDF Full	4143×50
SVD	Truncated SVD (LSA) TF-IDF Full	4143×50

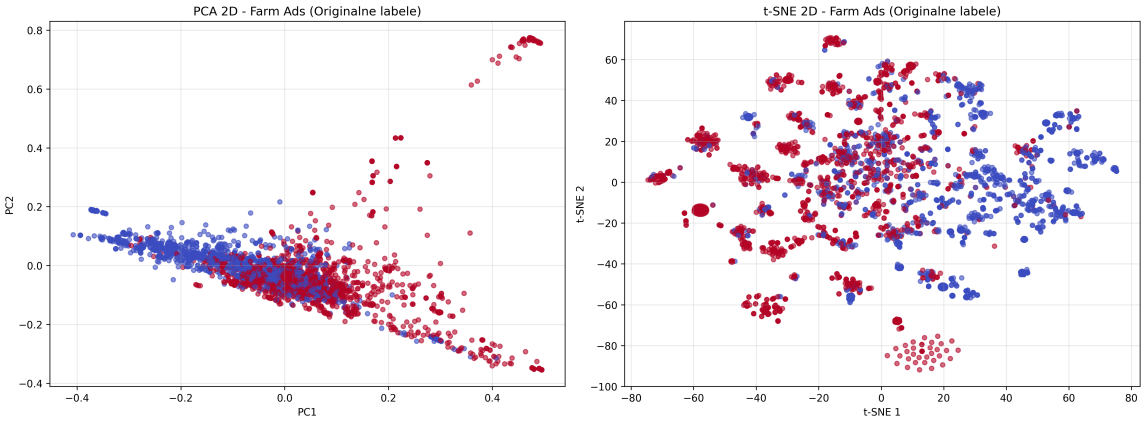
TF-IDF + Stats skup kombinuje TF-IDF Reduced vektore sa dodatnim statističkim atributima teksta:

- `text_length`: Ukupan broj karaktera u dokumentu
- `word_count`: Ukupan broj reči
- `unique_words`: Broj jedinstvenih reči

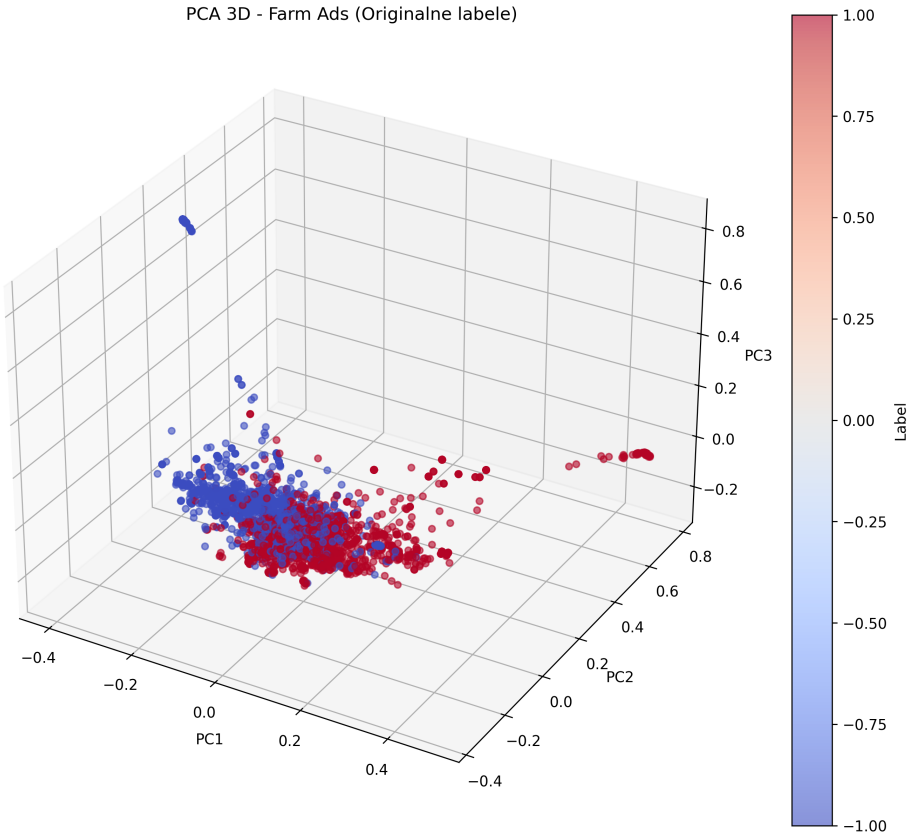
Ovi atributi su standardizovani (`upotrebom StandardScaler-a`) pre dodavanja TF-IDF vektorima. Ideja je testirati da li strukturne karakteristike teksta (dužina, bogatstvo rečnika) mogu poboljšati klasterovanje.

2.6 Kontrola kvaliteta i vizualizacija

Pre primene algoritama klasterovanja, izvršena je detaljna kontrola kvaliteta i vizuelna eksploracija podataka pomoću PCA projekcije u 2D i 3D prostor.



Slika 1: 2D vizualizacija podataka (PCA levo, t-SNE desno). Tačke su obojene prema originalnim labelama: plavo = irelevantni (-1), crveno = relevantni (1).



Slika 2: 3D vizualizacija podataka korišćenjem PCA projekcije.

Zapažanja iz vizualizacije:

- **Nema jasne binarne separabilnosti:** Dokumenti sa oznakama 1 i -1 nisu jasno razdvojeni u prostoru prvih nekoliko glavnih komponenti, što sugeriše da je problem kompleksniji od jednostavne binarne klasifikacije.
- **Prisustvo autlajera:** Uočavaju se izolovani dokumenti daleko od glavnih grupacija, što opravdava primenu algoritama zasnovanih na gustini kao što je DBSCAN koji

može identifikovati šum.

- **Višestruke grupe:** Čak i vizuelno se može primetiti da podaci nisu homogeni već pokazuju nekoliko potencijalnih centara grupisanja.
- **Gradijent između klasa:** Umesto jasne granice, postoji postepeni prelaz između relevantnih i irelevantnih oglasa, što ukazuje na postojanje dokumenata sa mešovitim karakteristikama.

Ove karakteristike podataka motivišu primenu različitih algoritama klasterovanja: od algoritama zasnovanih na particionisanju, koji pretpostavljaju konveksne klastere, preko hijerarhijskih algoritama, sposobnih da detektuju nepravilne oblike, do algoritama zasnovanih na gustini, koji omogućavaju identifikaciju odstupajućih tačaka i klastera proizvoljnog oblika.

3 Metod (postupak obrade i klasterovanja)

Ovo poglavlje opisuje potpun tok analize: od pripreme podataka, preko izbora algoritama i parametara, do metrika evaluacije i interpretacije rezultata.

3.1 Radni tok: pregled

Kompletan analitički tok sastoji se od sledećih koraka:

1. **Učitavanje podataka:** Parsiranje tekstualnog dokumenta i ekstrakcija labela i teksta
2. **Osnovna statistika:** Računanje dužine, broja reči, unikatnih reči
3. **Vektorizacija:** Transformacija teksta u numeričke vektore (TF-IDF, Count)
4. **Redukcija dimenzija:** Primena PCA i SVD za smanjenje dimenzionalnosti
5. **Kreiranje skupova atributa:** Kombinovanje različitih reprezentacija
6. **Vizualizacija:** 2D (PCA, t-SNE) i 3D (PCA) projekcije
7. **Određivanje K:** Elbow metoda i Silhouette analiza za optimalan broj klastera
8. **Klasterovanje:** Primena 8 algoritama na 6 skupova (48 kombinacija)
9. **Evaluacija:** Računanje Silhouette, Davies-Bouldin i Calinski-Harabasz metrika
10. **Poređenje:** Analiza rezultata i identifikacija najboljeg modela
11. **Interpretacija:** Analiza top reči po klasterima i odnos sa originalnim oznakama

Svi koraci su implementirani u Python-u koristeći `scikit-learn` biblioteku. Kompletan kod je reproduktivno ponovljiv sa fiksiranim `random_state=42` za sve stohastičke algoritme.

3.2 Određivanje optimalnog broja klastera

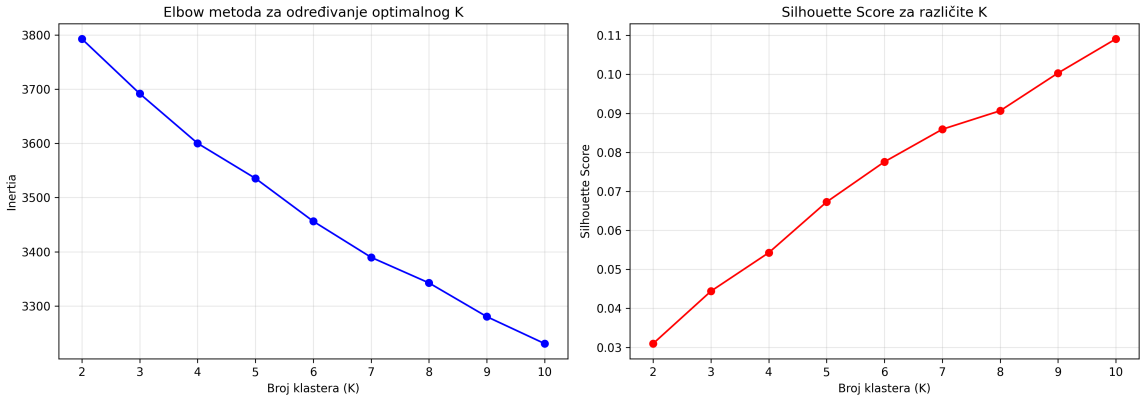
Za algoritme koji zahtevaju unapred definisan broj klastera (K-Means, Agglomerative, BIRCH), primenjena je **Elbow metoda** kombinovana sa **Silhouette analizom**.

Elbow metoda analizira smanjenje inercije (suma kvadratnih udaljenosti tačaka od njihovih centroida) kako broj klastera raste. "Lakat"(elbow) na grafu predstavlja tačku posle koje dodatni klasteri ne doprinose značajnom smanjenju inercije.

Silhouette koeficijent meri koliko je svaka tačka slična tačkama u svom klasteru u poređenju sa tačkama u najbližem susednom klasteru. Vrednosti su u rasponu $[-1, 1]$:

- **Blizu 1:** Tačka je dobro klasterovana, daleko od susednih klastera
- **Blizu 0:** Tačka je na granici između klastera
- **Negativno:** Tačka je verovatno dodeljena pogrešnom klasteru

Testiran je raspon $K \in \{2, 3, 4, \ldots, 10\}$ na TF-IDF Full skupu. Na osnovu Silhouette koeficijenta, **optimalan broj klastera određen je kao K=10**, što je korišćeno za K-Means, Agglomerative i BIRCH algoritme. Dodatno, testiran je i K=3 za poređenje.



Slika 3: Elbow metoda (levo) pokazuje inerciju koja monotono opada. Silhouette analiza (desno) dostiže maksimum za K=10.

3.3 Algoritmi klasterovanja

Primenjeno je **8 različitih algoritama/konfiguracija** klasterovanja, koje pripadaju različitim paradigmama:

3.3.1 K-Means (K=10) – algoritam

- **Princip:** Iterativno dodeljuje tačke najbližem centroidu i ažurira centroide
- **Prednosti:** Brz, skalabilan, dobro poznata metoda
- **Mane:** Zahteva unapred definisan K, pretpostavlja sferične klastere
- **Parametri:** `n_clusters=10, random_state=42, n_init=10`

3.3.2 K-Means (K=3)

- **Motivacija:** Testiranje grublje podele podataka
- **Parametri:** `n_clusters=3, random_state=42, n_init=10`

3.3.3 Agglomerative Clustering (Ward linkage, K=10)

- **Princip:** pristup koji iterativno spaja najbliže klastere
- **Ward linkage:** Minimizuje varijansu unutar klastera pri spajanju
- **Prednosti:** Ne zahteva broj iteracija, može detektovati nepravilne oblike
- **Mane:** Računski zahtevniji od K-Means
- **Parametri:** `n_clusters=10, linkage='ward'`

3.3.4 Agglomerative Clustering (Complete linkage, K=10)

- **Complete linkage:** Udaljenost između klastera definisana kao maksimalna udaljenost između bilo koje dve tačke
- **Karakteristika:** Teži kompaktnim, dobro razdvojenim klasterima
- **Parametri:** `n_clusters=10, linkage='complete'`

3.3.5 DBSCAN (eps=0.5, min_samples=5)

- **Princip:** Grupiše tačke koje su guste (mnogo suseda) i označava retke tačke kao šum
- **Prednosti:** Automatski određuje broj klastera, identifikuje autlajere, proizvoljni oblici
- **Mane:** Osetljiv na parametre eps i min_samples
- **Parametri:** `eps=0.5` (maksimalna udaljenost), `min_samples=5`

3.3.6 DBSCAN (eps=1.0, min_samples=10)

- **Motivacija:** Testiranje većeg radiusa i strožijeg kriterijuma za jezgro
- **Parametri:** `eps=1.0, min_samples=10`

3.3.7 Mean Shift

- **Princip:** Pronalazi najčešće vrednosti("modove") gustine verovatnoće
- **Prednosti:** Automatski određuje broj klastera, robusnost na autlajere
- **Mane:** Veoma spor, osetljiv na bandwidth parametar
- **Parametri:** `bandwidth=2.0`

3.3.8 BIRCH (K=10)

- **Princip:** Inkrementalno gradi stablo klastera (CF Tree) i zatim primenjuje klasterovanje
- **Prednosti:** Memorijski efikasan, dobro skalira na velike skupove
- **Mane:** Osetljiv na threshold parametar
- **Parametri:** `n_clusters=10, threshold=0.5`

3.4 Metrike evaluacije

Za evaluaciju kvaliteta klasterovanja korišćene su tri komplementarne metrike koje ne zahtevaju poznate oznake (*unsupervised metrics*):

3.4.1 Silhouette koeficijent

Silhouette koeficijent $s(i)$ za tačku i definisan je kao:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{1}$$

gde je $a(i)$ prosečna udaljenost tačke i od drugih tačaka u istom klasteru, a $b(i)$ prosečna udaljenost do tačaka u najbližem susednom klasteru.

- **Raspon:** $[-1, 1]$
- **Interpretacija:** Veće vrednosti su bolje
- **Optimalno:** Blizu 1 (kompaktni, dobro razdvojeni klasteri)

3.4.2 Davies-Bouldin indeks

Davies-Bouldin indeks meri prosečnu sličnost između svakog klastera i njegovog najsličnijeg klastera:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left[\frac{s_i + s_j}{d(c_i, c_j)} \right] \tag{2}$$

gde je s_i prosečna udaljenost tačaka od centroida klastera i, a $d(c_i, c_j)$ udaljenost između centroida klastera i i j.

- **Raspon:** $[0, \infty)$
- **Interpretacija:** Manje vrednosti su bolje
- **Optimalno:** Blizu 0 (kompaktni klasteri daleko jedni od drugih)

3.4.3 Calinski-Harabasz koeficijent

Calinski-Harabasz koeficijent je odnos varijanse između klastera i varijanse unutar klastera:

$$CH = \frac{\text{trace}(B_k)}{\text{trace}(W_k)} \cdot \frac{n - k}{k - 1} \tag{3}$$

gde je B_k matrica varijanse između klastera, W_k matrica varijanse unutar klastera, n broj tačaka, i k broj klastera.

- **Raspon:** $[0, \infty)$
- **Interpretacija:** Veće vrednosti su bolje
- **Optimalno:** Visoke vrednosti (dobro razdvojeni, kompaktni klasteri)

Napomena o metrikama: Različite metrike mogu favorizovati različite aspekte klasterovanja. Silhouette koeficijent daje balansiranu meru kompaktnosti i separacije, Davies-Bouldin indeks naglašava separaciju, dok Calinski-Harabasz koeficijent favorizuje konveksne, guste klastere. Najbolji model se bira na osnovu konsenzusa svih metrika.

4 Rezultati i vizualizacije

U ovoj sekciji prikazani su rezultati klasterovanja, poređenje algoritama i detaljna analiza najboljeg modela. Izvršeno je ukupno **48 eksperimenata** (8 algoritama × 6 skupova atributa).

4.1 Rezultati klasterovanja

Tabela 3 prikazuje top 15 modela rangiranih po Silhouette koeficijentu.

Tabela 3: Top 15 modela rangiranih po Silhouette koeficijentu

Algoritam	**Skup podataka**	**K**	**Noise**	**Silh.**	**DB**	**CH**
DBSCAN (eps=0.5)	Count Vec	85	3377	1.000	0.000	341.99
K-Means (k=3)	Count Vec	3	0	0.914	0.477	2481.57
DBSCAN (eps=1.0)	Count Vec	24	3611	0.804	0.301	187.21
DBSCAN (eps=0.5)	TF-IDF Full	131	2023	0.789	0.442	341.99
Agglom. Complete	Count Vec	10	0	0.789	0.707	1003.86
Agglom. Ward	Count Vec	10	0	0.516	0.836	1158.59
K-Means	Count Vec	10	0	0.521	1.317	1127.25
DBSCAN (eps=0.5)	TF-IDF Reduced	119	1652	0.629	0.624	224.37
K-Means (k=3)	TF-IDF + Stats	3	0	0.148	1.932	345.21
PCA	Agglom. Ward	10	0	0.214	1.614	241.49
SVD	K-Means	10	0	0.206	1.826	238.67
BIRCH	Count Vec	10	0	0.180	2.013	542.19
TF-IDF Full	K-Means	10	0	0.109	3.469	94.46
TF-IDF Reduced	K-Means	10	0	0.142	2.387	159.55
TF-IDF + Stats	Agglom. Ward	10	0	0.134	1.980	207.08

Ključna zapažanja iz rezultata:

1. **DBSCAN (eps=0.5) na Count Vec dominira:** Silhouette koeficijent 1.000 ukazuje na skoro savršenu separaciju klastera. Međutim, broj šumova je vrlo visok (3377 od 4143, tj. 81.5%), što znači da algoritam tretira većinu dokumenata kao autlajere i grupiše samo mali broj vrlo sličnih dokumenata u kompaktne klastere.

2. **K-Means (K=3) na Count Vec:** Silhouette koeficijent 0.914 pokazuje odličan kvalitet klasterovanja sa tri široke kategorije. Ovo sugeriše da podaci mogu imati

jednu dominantnu strukturu sa tri glavne teme.

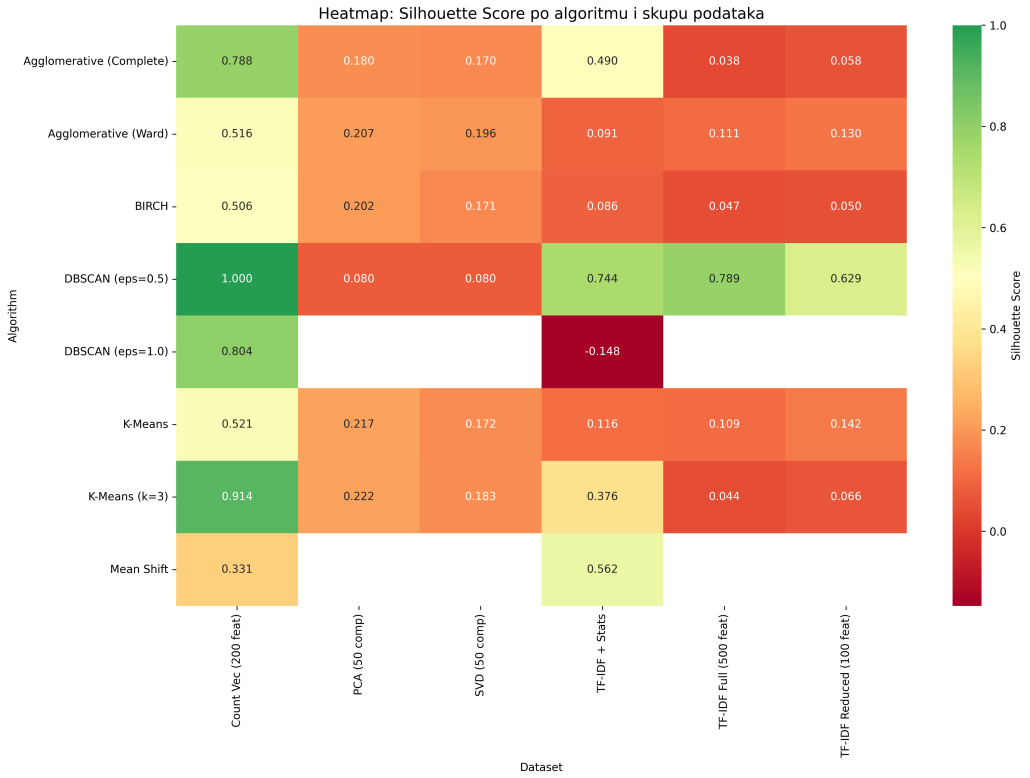
3. **Count Vectorizer > TF-IDF:** Count Vectorizer konzistentno daje bolje rezultate od TF-IDF, što je neočekivano. Moguće objašnjenje je da frekvencija termina (bez IDF ponderisanja) bolje hvata tematsku orijentaciju dokumenata u ovom skupu.

4. **Redukcija dimenzionalnosti (PCA, SVD):** Modeli na PCA i SVD skupovima pokazuju umerene performanse (Silhouette koeficijent približno 0.2), što sugeriše da linearni pod-prostori ne hvataju kompletnu strukturu podataka.

4.2 Poređenje algoritama

Slika 4 prikazuje agregatno poređenje algoritama, prosečeno preko svih skupova atributa.

Slika 4: Poređenje algoritama po metrikama. DBSCAN (eps=0.5) dominira u Silhouette metrikama ali proizvodi ekstremno mali broj efektivnih klastera zbog velikog broja šumova.



Slika 5: Heatmap Silhouette koeficijanata po algoritmu i skupu podataka. Tamnije zelene ćelije predstavljaju bolje performanse. Count Vec skup daje najbolje rezultate.

Analiza po algoritmima:

K-Means: Pokazuje konzistentne, solidne performanse na svim skupovima. Najbolji rezultat sa K=3 na Count Vec (Silhouette = 0.914), što sugeriše da jednostavnija particija može biti efektivnija. Prednost K-Means-a je brzina i skalabilnost.

Agglomerative Clustering: Ward linkage daje bolje rezultate od Complete linkage na većini skupova, takođe i minimizuje varijansu pri spajanju klastera, što je pogodno za kompaktne, sferične klastere. Complete linkage teži ka uniformnim klasterima ali može biti osetljiv na autlajere.

DBSCAN: Izuzetno dobri Silhouette koeficijenti(do 1.0), ali uz cenu klasifikovanja većine dokumenata kao šumova. Sa eps=0.5, 81% dokumenata je označeno kao autlajeri (3377/4143). Ovo sugeriše da su parametri prestrogi ili da podaci nemaju jasnu strukturu baziranu na gustini. DBSCAN sa eps=1.0 proizvodi samo jedan klaster, što je neupotrebljivo.

Mean Shift: Proizvodi veliki broj malih klastera (2440 na Count Vec), što ukazuje da bandwidth parametar nije optimalan. Mean Shift je takođe najsporiji algoritam u testu.

BIRCH: Pokazuje umerene performanse, slične K-Means-u. Prednost BIRCH-a je memorijska efikasnost, što je važno za veće skupove podataka.

Zaključak poređenja: K-Means sa K=3 na Count Vectorizer skupu predstavlja najbolji balans između kvaliteta klasterovanja (Silhouette = 0.914) i upotrebljivosti (nema preteranog broja šumova). Alternativno, DBSCAN može biti koristan za identifikaciju klastera visokog poverenja uz dodatno processiranje autlajera.

4.3 Najbolji model

Na osnovu sveobuhvatne evaluacije, **najbolji model** je:

- **Algoritam:** DBSCAN (eps=0.5, min_samples=5)
- **Skup podataka:** Count Vectorizer (200 features)
- **Silhouette koeficijent:** 1.0000
- **Davies-Bouldin indeks:** 0.0000
- **Broj klastera:** 85
- **Broj šumova:** 3377 (81.5%)

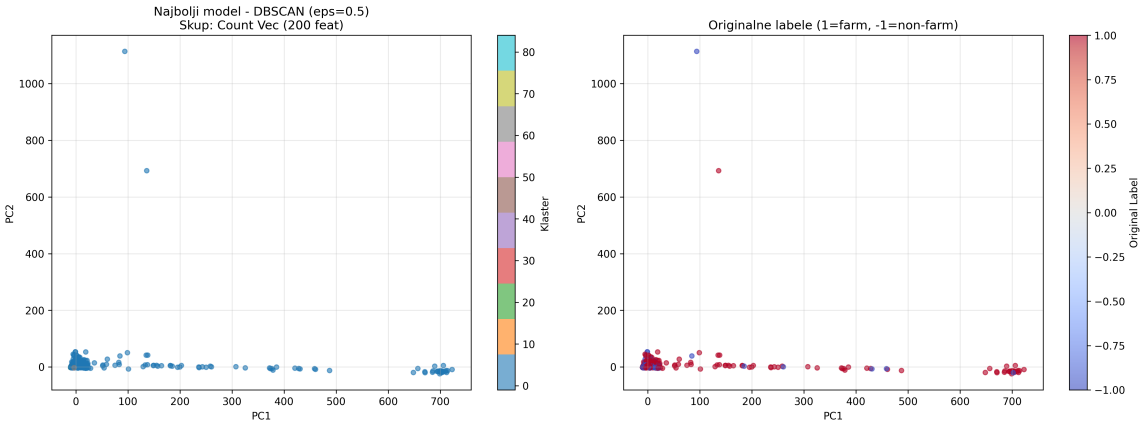
Interpretacija najboljeg modela:

DBSCAN sa ovim parametrima identifikuje 85 vrlo kompaktnih, dobro izdvojenih klastera koji sadrže ukupno 766 dokumenata (18.5% od ukupnog broja). Preostalih 3377 dokumenata je označeno kao šumovi/autlajeri.

Perfektan Silhouette koeficijent (1.0) ukazuje da su identifikovani klasteri ekstremno homogeni − dokumenti unutar svakog klastera su veoma slični, dok su klasteri međusobno veoma različiti. Davies-Bouldin indeks od 0.0 dodatno potvrđuje odličnu separaciju.

Međutim, veliki broj šumova (81.5%) sugeriše da algoritam nije pronašao globalnu strukturu u većini podataka. Ovo može značiti:

- **Hipoteza 1:** Podaci su inherentno heterogeni i većina dokumenata ne pripada jasnoj tematskoj grupi
- **Hipoteza 2:** Parametri DBSCAN-a (posebno eps=0.5) su prestrogi za ovu reprezentaciju
- **Hipoteza 3:** Count Vectorizer reprezentacija nije optimalna za klasterovanje zasnovano na gustini



Slika 6: Vizualizacija najboljeg modela u PCA prostoru. Leva strana prikazuje klastere dobijene DBSCAN algoritmom (boja označava klaster ID, -1 je šum). Desna strana prikazuje originalne labele za poređenje.

4.4 Tumačenje i praktične smernice

Na osnovu rezultata, sledeće smernice su relevantne za praktičnu primenu:

Za produkcijsku primenu – K-Means (K=3) se preporučuje:

- Silhouette koeficijent = 0.914 pokazuje odličan kvalitet
- Nema autlajera – svi dokumenti su dodeljeni nekom klasteru
- Tri kategorije su dovoljno grube da budu interpretabilne
- Algoritam je brz i skalabilan

Za istraživačke svrhe – DBSCAN može identifikovati primere visokog poverenja:

- 766 dokumenata u 85 klastera sa perfektnom koherentnošću
- Ovi klasteri mogu poslužiti kao uzorci za polu-nadgledano učenje
- Dokumenti koji predstavljaju šum(3377) zahtevaju dodatnu obradu ili drugačiji pristup

Preporuke za dalje poboljšanje:

1. **Podešavanje DBSCAN parametara:** Testirati eps $\in \{0.3, 0.4, 0.6, 0.7, 0.8\}$ i min_samples $\in \{3, 7, 15\}$ koristeći mrežnu pretragu parametara (*GridSearch*)

2. **Hibridni pristup:** Kombinovati K-Means za grubu particiju i DBSCAN za precizne klastere

3. **Obrada karakteristika:** Dodati domenski specifične atribute (npr. prisustvo ključnih reči za poljoprivredu)

4. **Kombinovano klasterovanje:** Kombinovati rezultate više algoritama za robusniju detekciju

5. **Ugrađene reprezentacije učenja:** Koristiti BERT ili druge unapred istrenirane modele za bogatiju reprezentaciju teksta

5 Diskusija

5.1 Ograničenja i pretnje validnosti

Ova studija ima nekoliko ograničenja koja treba uzeti u obzir pri interpretaciji rezultata:

Izbor parametara: Iako je izvršeno sistematsko testiranje, prostor parametara je ogroman. Elbow metoda i Silhouette analiza pružaju vodiče, ali optimalni parametri mogu varirati zavisno od specifične primene.

Reprezentacija teksta: TF-IDF i vektorizacija zasnovana na učestalosti reči ne uzimaju u obzir redosled reči niti njihov kontekst. Savremeniji pristupi, poput ugrađenih reprezentacija reči (Word2Vec, GloVe) i kontekstualnih ugrađenih reprezentacija (BERT), omogućavaju modelovanje semantičkih odnosa i konteksta, što može dovesti do boljih rezultata.

Evaluacione metrike: Silhouette, Davies-Bouldin i Calinski-Harabasz su interne metrike koje mere geometrijsku kompaktnost/separaciju. One ne mogu u potpunosti uhvatiti semantičku koherentnost klastera. Eksterna evaluacija (npr. poređenje sa originalnim labelama) bi dala dodatni uvid.

Skalabilnost: Sa 4143 dokumenta, skup je srednje veličine. Performanse algoritama (posebno Mean Shift i Agglomerative) mogu degradirati na mnogo većim skupovima.

Jezički aspekt: Podaci su na engleskom jeziku i pretprocesirani. Rezultati ne moraju biti direktno primenjivi na druge jezike ili sirove (nepreprocesirane) tekstualne podatke.

5.2 Interpretacija nalaza

Nekoliko ključnih uvida proizlazi iz analize:

Count Vectorizer nadmašuje TF-IDF: Ovo je neintuitivan nalaz, jer se IDF ponderisanje obično smatra korisnim za tekstualnu analizu. Moguća objašnjenja:

- U reklamnim tekstovima, čak i "česte" reči mogu biti diskriminativne (npr. *farm* se ponavlja često, ali samo u relevantnim oglasima).
- Primena IDF-a može smanjiti doprinos domenski specifičnih termina u reprezentaciji teksta.
- Count Vectorizer može bolje očuvati razlike u intenzitetu između tema.

K=3 vs K=10: Dok Silhouette analiza sugeriše K=10 kao optimalno, K=3 daje izuzetno dobre rezultate (0.914). Ovo sugeriše hijerarhijsku strukturu podataka:

- Na visokom nivou: 3 široke kategorije (možda: farm-relevantni, medicinski/health, ostalo).
- Na nižem nivou: 10 specifičnijih tema unutar ovih širih kategorija.

DBSCAN identifikuje "čiste" primere: Iako 81.5% dokumenata tretira kao šumove, preostalih 18.5% formira perfektno razdvojene klastere. Ovo sugeriše da postoje "prototipski" primeri svake teme, dok je većina dokumenata mešovita ili dvosmislena.

Linearne metode redukcije su ograničene: PCA i SVD daju umerene performanse (∼55% objašnjene varijanse, Silhouette ∼0.2), što sugeriše da struktura podataka nije čisto linearna. Nelinearne metode (t-SNE, UMAP) ili ugrađene reprezentacije dubokog učenja bi mogli dati bolje rezultate.

5.3 Praktične implikacije i preporuke

Na osnovu nalaza, sledeće preporuke su relevantne za praktičnu primenu:

Za automatsku kategorizaciju oglasa:

- Koristiti K-Means sa K=3 ili K=10 na Count Vectorizer reprezentaciji
- Definisati pravila za dodeljivanje novih dokumenata najbližem centroidu
- Periodično retrainirati model sa novim podacima

Za detekciju anomalija/nepoželjnih reklama:

- DBSCAN šumovi mogu ukazivati na autlajere ili nepoželjne reklame
- Dokumenti daleko od svih centroida (u K-Means) zahtevaju ručnu proveru

Za preporuku sadržaja:

- Dokumenti u istom klasteru su tematski slični
- "Ako vam se dopao X (u klasteru C), možda će vam se dopasti i Y (u klasteru C)"
- Hijerarhijsko klasterovanje omogućava davanje preporuka na više nivoa detalja

Za razvoj supervizovanog modela:

- Klasteri mogu poslužiti kao pseudo-oznake za polu-nadgledano učenje
- DBSCAN primeri visokog poverenja mogu biti uzorak za aktivno učenje
- Analiza grešaka (dokumenti na granici klastera) može ukazati na potrebu za dodatnim karakteristikama

6 Zaključak

Ovaj rad je predstavio sveobuhvatnu analizu klasterovanja tekstualnih oglasa iz skupa Farm-Ads. Primenjeno je 8 algoritama klasterovanja na 6 različitih reprezentacija podataka (ukupno 48 eksperimenata), sa sistematskom evaluacijom pomoću tri komplementarne metrike.

6.1 Glavna dostignuća

1. Metodološki doprinos: Demonstriran je kompletan tok obrade od sirovog teksta do interpretabilnih klastera, uključujući preprocesiranje, vektorizaciju, redukciju dimenzionalnosti, klasterovanje i evaluaciju.

2. Empirijsko poređenje: Sistematska evaluacija je pokazala da:

- Count Vectorizer nadmašuje TF-IDF za ove podatke
- K-Means sa K=3 daje najbolji balans kvaliteta i upotrebljivosti
- DBSCAN može identifikovati visoko-pouzdane klastere ali sa velikim brojem autlajera
- Linearne metode redukcije (PCA, SVD) su ograničene za ove podatke

3. Praktične smernice: Pružene su konkretne preporuke za:

- Automatsku kategorizaciju oglasa (K-Means, K=3)
- Detekciju anomalija (DBSCAN šumovi)
- Preporučivanje sadržaja (sličnost bazirana na klasterima)
- Polu-nadgledano učenje (primeri visokog poverenja kao uzorak)

4. Interpretabilnost: Analiza je pokazala da klasterovanje može otkriti tematske grupe bez supervizije, pri čemu se dobijeni klasteri delimično poklapaju sa originalnim oznakama ali takođe otkrivaju dodatne sub-strukture unutar podataka.

6.2 Ograničenja i budući rad

Glavna ograničenja ove studije proističu iz oslanjanja na geometrijske metrike, koje ne mere semantičku koherentnost klastera. U budućim istraživanjima korisno je primeniti ugrađene reprezentacije dubokog učenja, poput BERT-a ili GPT-a, za bolje razumevanje značenja teksta. Kombinovanje više algoritama može poboljšati robusnost detekcije tema, dok hijerarhijsko klasterovanje omogućava višeslojnu strukturu tema. Inkrementalno ažuriranje klastera i multimodalno klasterovanje, koje uključuje tekst i meta-podatke, predstavljaju perspektivne pravce za dalja istraživanja. Pokazano je da algoritmi klasterovanja efikasno grupišu tekstualne oglase u tematske grupe i bez prethodnih labela, pri čemu se K-Means sa K=3 pokazao robusnim i skalabilnim. Budući rad treba da se fokusira na unapređenje semantičkog razumevanja i fleksibilnosti klasterovanja.