# Mətnlərin Müəlliflərə Görə Klasterləşdirilməsində Dərin Öyrənmə Modellərinin Tətbiqi

Leyla Məmmədova
İnformasiya Texnologiyaları İnstitutu, Bakı, Azərbaycan
*mammedova-1998@list.ru*

***Xülasə*— Son illərdə kibercinayətkarlar anonimliyi qorumaq məqsədilə e-poçtlar, sosial media və digər onlayn mühitlərdə mətn əsaslı ünsiyyət vasitələrindən geniş istifadə edirlər. Bu səbəbdən mətn müəllifinin müəyyən edilməsi, xüsusilə onların klasterləşdirilməsi, mühüm tədqiqat problemlərindən biridir. Tədqiqat işində dərin klasterləşdirmə modelləri müəllif klasterləşdirmə probleminə tətbiq edilmiş, onların bu sahədəki effektivliyi təhlil edilmişdir. Əldə edilən nəticələr dərin klasterləşdirmə modelləri müəllif klasterləşdirmə probleminin həllində perspektivli istiqamət olduğunu göstərir və rəqəmsal kriminalistika sahəsində praktik tətbiq potensialına malik olduğunu göstərir.**

***Açar sözlər*— *rəqəmsal kriminalistika; müəllif klasterləşdirmə; dərin klasterləşdirmə.***

## I. GİRİŞ

Rəqəmsal kriminalistikanın əsas istiqamətlərindən biri olan müəlliflik analizi (authorship analysis) anonim mətnlərin (məsələn, e-poçtlar, sosial media paylaşımları, təhdid və ya şantaj xarakterli mesajlar) müəllifinin müəyyənləşdirilməsində geniş tətbiq olunur [1]. Müəlliflik analizi mətnlərin müəllifini müəyyən etməyə yönəlmiş tədqiqat sahəsidir. Bu sahə daxilində müəllifin yoxlanılması (author verification), müəllifin identifikasiyası (authorship attribution) və müəllif klasterləşdirməsi (authorship clustering) kimi əsas istiqamətlər mövcuddur [2].

Praktik tətbiqdə çox vaxt məlumatlar sinifsiz (unlabeled) olur, bu da müəllimsiz (unsupervised) yanaşmanın tətbiqini zəruri edir. Müəllif klasterləşdirməsi problemində verilmiş sənədlər toplusu oxşarlıq əsasında qruplaşdırılır. Məqsəd eyni müəllif tərəfindən yazılmış sənədləri eyni klasterdə, fərqli müəlliflərə aid sənədləri isə digər klasterlərdə toplamaqdır [3].

İlk dəfə 2012-ci ildə CLEF (Conference and Labs of the Evaluation Forum) konfransının PAN (Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection) laboratoriyası çərçivəsində müəlliflik analizi məsələsinin bir hissəsi kimi təqdim olunan [2] müəllif klasterləşdirməsi, sonrakı illərdə aparılan tədqiqatlarla aktual elmi istiqamətə çevrilmişdir.

Bu tədqiqat işində müəllif klasterləşdirməsi probleminə baxılmışdır. Belə ki, dərin klasterləşdirmə modelləri müxtəlif sahələrdə uğurla tətbiq olunsa da, onların müəllif klasterləşdirilməsi problemlərinə tətbiqi məhdud xarakter daşıyır. Əksər hallarda bu modellər müəllif üslubunun spesifik xüsusiyyətlərini nəzərə almır. Bu məqalədə müəllifin yazı üslubu dərin klasterləşdirmə modelləri vasitəsilə analiz edilmiş və oxşar yazı tərzinə malik mətnlər eyni klasterlərdə qruplaşdırılmışdır. İşin strukturu aşağıdakı kimidir: ikinci bölmədə tətbiq olunan dərin klasterləşdirmə modelləri və istifadə olunan qiymətləndirmə göstəriciləri təqdim olunur; üçüncü bölmədə eksperimentlərdə istifadə edilən verilənlər bazası, əldə olunan nəticələr və onların vizual təqdimatı göstərilir; sonda isə tədqiqatın nəticələri ümumiləşdirilərək gələcək istiqamətlər müzakirə edilir.

## II. DƏRİN KLASTERLƏŞDİRMƏ MODELLƏRİ

Dərin klasterləşdirmə, verilənlərdən dərin neyron şəbəkələri vasitəsilə xüsusiyyətlərin öyrənilməsinə (feature learning) əsaslanan və sürətlə inkişaf edən bir sahədir. Bu sahədə mühüm irəliləyişlərdən biri xüsusiyyət öyrənmə və klasterləşdirmə mərhələlərini eyni model daxilində birləşdirən (simultaneous deep clustering) yanaşmaların ön plana çıxmasıdır. Bu yanaşma xüsusiyyətlərin öyrənilməsi və klasterləşdirmə proseslərinin eyni vaxtda optimallaşdırılmasını təmin edir. Aşağıda bu yanaşmanın əsasını təşkil edən Dərin Daxili Klasterləşmə (Deep Embedded Clustering) və Təkmilləşdirilmiş Dərin Daxili Klasterləşmə (Improved Deep Embedded Clustering) modelləri haqqında ətraflı məlumat təqdim olunur:

### A *Dərin Daxili Klasterləşmə (DEC-Deep Embedded Clustering)*

Dərin Daxili Klasterləşmə modeli Xie və həmkarları tərəfindən təklif edilmişdir [4]. Model ilkin mərhələdə avtokodlayıcının (autoencoder) əvvəlcədən öyrədilməsi ilə başlayır, daha sonra isə dekodlayıcı (decoder) hissəsi çıxarılaraq yalnız kodlayıcı (encoder) strukturu saxlanılır və verilənlərin ilkin fəzadan gizli xüsusiyyət fəzasına çevrilməsi üçün istifadə olunur. Klaster mərkəzlərinin başlanğıc qiymətlərini müəyyən etmək məqsədilə verilənlər gizli ilkin klaster mərkəzləri formalaşdırılır. Daha sonra model bu ilkin mərkəzlər əsasında iterativ şəkildə optimallaşdırılır. DEC modelində hər bir nöqtənin klasterlərə mənsubiyyəti Student t-paylanması əsasında hesablanır və bu paylanma $Q$ ilə işarə olunur. Daha sonra $Q$ paylanmasına əsasən köməkçi hədəf paylanması $P$ formalaşdırılır. $P$ paylanması daha yüksək ehtimal qiymətlərinə malik nöqtələrin təsirini artıraraq klasterlərin daha sıx və aydın strukturda formalaşmasına şərait yaradır.

DEC modelinin əsas yeniliyi klasterləşdirmə itkisinin tətbiqidir [5]. Bu itki funksiyası Kullback–Leibler divergensiyası əsasında müəyyən edilir və aşağıdakı kimi ifadə olunur:

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} \, log \frac{p_{ij}}{q_{ij}}, \qquad (1)$$

DOI: 10.25045/forensic.gov.2026.23

$$q_{ij} = \frac{(1+||z_i-\mu_j||^2)^{-1}}{\sum_j (1+||z_i-\mu_j||^2)^{-1}}, \quad (2)$$

Burada $q_{ij}$ , t-paylanması (Student's t-distribution) ilə ölçülən, $z_i$ nöqtəsi ilə $\mu_j$ klaster mərkəzi arasındakı oxşarlıq dərəcəsidir [6].

$p_{ij}$ isə aşağıdakı kimi ifadə olunan hədəf paylanmasıdır (target distribution):

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})}, \quad (3)$$

Klaster mərkəzlərinin ilkin qiymətlərini əldə etmək məqsədilə verilənlər əvvəlcədən öyrədilmiş Dərin Neyron Şəbəkəsi (Deep Neural Network) vasitəsilə gizli fəzaya proyeksiya olunur. Sonra isə gizli Z fəzasında k-means alqoritmi icra edilərək $\{\mu_j\}_{j=1}^{k}$ başlanğıc klaster mərkəzləri hesablanır.

Beləliklə, bu proses nəticəsində eyni klasterə aid nöqtələr bir-birinə yaxınlaşır, fərqli klasterlərə aid nöqtələr isə bir-birindən uzaqlaşır

### *B. Təkmilləşdirilmiş Dərin Daxili Klasterləşmə (IDEC-Improved Deep Embedded Clustering)*

2017-ci ildə Xifeng Guo və həmkarları DEC modelinin təkmilləşdirilmiş versiyası olan Təkmilləşdirilmiş Dərin Daxili Klasterləşmə modelini təklif etmişdir [5].

İlkin mərhələdə DEC-də olduğu kimi IDEC modelində də avtokodlayıcının əvvəlcədən öyrədilməsi ilə başlayır. Lakin DEC modelindən fərqli olaraq, IDEC-də kodlayıcı ilə yanaşı dekodlayıcı hissəsi də saxlanılır və təlim prosesi boyunca istifadə olunur. Bu yanaşma gizli fəzanın klasterləşdirməyə uyğun strukturda formalaşmasını və giriş verilənlərinin rekonstruksiyasının qorunmasını təmin edir.

IDEC modelinin ümumi itki funksiyası aşağıdakı kimi hesablanır:

$$L = L_{KL} + \gamma L_{rec} \quad (4)$$

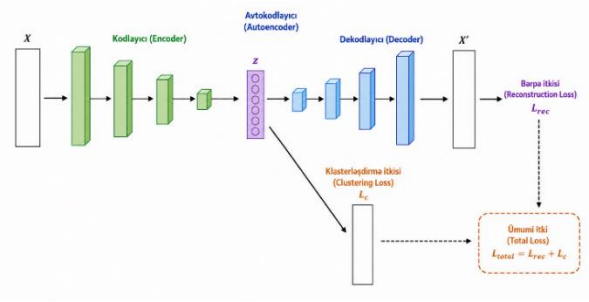
burada $L_{KL}$ klasterləşdirmə itkisini, $L_{rec}$ isə rekonstruksiya itkisini göstərir, $\gamma$ isə gizli fəzanın təhrif olunmasını tənzimləyən hiperparametrdir. $\gamma = 1$ və $L_{rec} \equiv 0$ olduqda bu ifadə DEC modelinin məqsəd funksiyasına uyğun gəlir. Praktik olaraq isə rekonstruksiya itkisini qorumaq və gizli fəzanın həddindən artıq təhrif olunmasının qarşısını almaq üçün $\gamma$adətən 1-dən kiçik seçilir.

Klasterləşdirmə itkisi – DEC modelində olduğu kimi, Kullback–Leibler divergensiyası əsasında hesablanır və daha aydın və ayrılmış klasterlərin formalaşmasını təmin edir.

Rekonstruksiya itkisi - avtokodlayıcının çıxışı ilə giriş verilənləri arasındakı fərqi minimuma endirməyə yönəldilmişdir. Adətən bu itki Orta Kvadratik Xəta (Mean Squared Error) vasitəsilə hesablanır.:

$$L_{rec} = \sum_{i=1}^{n} \left\| x_i - g_{w'}(z_i) \right\|_2^2 \quad (5)$$

Burada $x_i$ giriş, $z_i = f_w(x_i) \in Z$ giriş verilənin kodlayıcı ($f_w$) vasitəsilə söz vektorunun yaradılması əməliyyatından sonra əldə olunan uyğun təsviri, $g_{w'}(z_i)$ isə dekodlayıcı vasitəsilə yenidən əldə olunmuş verilənlər. $w'$ dekodlayıcının çəki parametrlərini, $\|\cdot\|_2^2$-Evklid normasının kvadratını bildirir.




Şəkil 1. IDEC modelinin strukturu

### *C. Qiymətləndirmə İndeksləri*

Dərin klasterləşdirmə modellərinin effektivliyi bu sahədə geniş tətbiq olunan Dəqiqlik (Accuracy), Normallaşdırılmış Qarşılıqlı İnformasiya (Normalized Mutual Information) və Korreksiya olunmuş Rand İndeksi (Adjusted Rand Index) metrikaları vasitəsilə qiymətləndirilir.

Dəqiqlik ($Acc$) - indeksi aşağıdakı düsturla hesablanır [7]:

$$Acc = \max_m \frac{\sum_{i=1}^{N} 1\{y_i = m(\tilde{y}_i)\}}{N} \quad (6)$$

Burada $N$ nöqtələrin ümumi sayını, $Y = \{y_i | 1 \le i \le N\}$və $\tilde{Y} = \{\tilde{y}_i | 1 \le i \le N\}$isə müvafiq olaraq $i$ -ci nöqtə üçün həqiqi (ground-truth) və proqnozlaşdırılmış sinifləri ifadə edir. Həqiqi etiketlərin məqsədi, alqoritmin verilənləri əvvəlcədən təyin olunmuş siniflərə nə dərəcədə düzgün qruplaşdırdığını qiymətləndirməkdir. Klasterləşdirmə nəticələrinin doğruluğunu yoxlamaq üçün modelin çıxışları birbaşa bu həqiqi etiketlərlə müqayisə edilir. $m$ isə proqnozlaşdırılmış siniflərlə həqiqi siniflər arasında bütün mümkün bir-birinə uyğunlaşdırmaları ifadə edir. $1\{\bullet\}$ifadəsi isə indikator funksiyası olub, şərt doğru olduqda 1, əks halda isə 0 qiymətini alır.

Normallaşdırılmış Qarşılıqlı İnformasiya ($NMI$) - aşağıdakı düsturla hesablanır [8]:

$$NMI(\tilde{Y}, Y) = \frac{I(\tilde{Y};Y)}{\frac{1}{2}[H(\tilde{Y})+H(Y)]}, \quad (7)$$

$Y$ -nin entropiyası $H(Y)$ olaraq göstərilir, $I(\tilde{Y};Y)$ isə $\tilde{Y}$ ilə Yarasında qarşılıqlı informasiyanı (mutual information) ifadə edir.

Korreksiya olunmuş Rand indeksi($ARI$) [9]- Rand İndeks (RI) əsasında formalaşdırılmışdır və klasterləşdirmə nəticələrini cüt-cüt (pairwise) qərarlar toplusu kimi qiymətləndirir.

$$RI = \frac{TP + TN}{C_N^2}, \quad (8)$$

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)}, \quad (9)$$

Burada $TP$ və $TN$ müvafiq olaraq doğru sinifləndirilmiş düzgün halların sayı (true positive) və yanlış halların (true negative) sayını ifadə edir, $C_N^2$ isə mümkün cütlərin sayını göstərir və $\frac{N(N-1)}{2}$kimi hesablanır. $ARI$ isə (9) düsturuna əsasən hesablanır.

$Acc$ və $NMI$ göstəriciləri $[0;1]$ intervalında dəyişir, $ARI$ isə$[-1;1]$ intervalında qiymətlər alır. Hər 3 metrika üzrə yüksək qiymətlər daha yaxşı klasterləşdirmə nəticələrini göstərir.

## III. EKSPERİMENT

Bu bölmədə DEC və IDEC dərin klasterləşdirmə modellərinin müəllif klasterləşdirməsi problemində performansının qiymətləndirilməsi məqsədilə aparılan eksperimentlər və əldə edilən nəticələr təqdim olunur. Həmçinin verilənlərin xüsusiyyətləri, nəticələrin müqayisəli təhlili və onların vizual analizi göstərilir.

### *A. Verilənlər bazası*

DEC və IDEC modellərinin müəllif klasterləşdirməsi sahəsində effektivliyini qiymətləndirmək məqsədilə tədqiqatda iki fərqli xarakterə malik verilənlər bazasından istifadə edilmişdir. Bu verilənlər bazalarının seçilməsində əsas məqsəd modellərin həm formal, həm də qeyri-formal mətnlər üzərində performansını təhlil etməkdir.

Enron - rəqəmsal kriminalistikada (digital forensics) geniş istifadə edilən verilənlər bazasıdır və Enron korporasiyasının 150-dən çox əməkdaşına məxsus təxminən 500000 daxili e-poçt yazışmasını əhatə edir. Bu tədqiqatda hesablama effektivliyini təmin etmək məqsədilə verilənlər bazasının ilk 200000 e-poçtu istifadə edilmişdir. Daha sonra 10 müəllif seçilmiş və hər müəllif üçün 50 e-poçt filtrasiya edilərək ümumilikdə 500 sənəddən ibarət alt verilənlər bazası formalaşdırılmışdır.

Reuters C50 - Reuters C50 (və ya RCV1-v2 alt-çoxluğu) müəllif identifikasiyası və stilometrik analiz sahəsində geniş istifadə olunur. Bu verilənlər bazası Reuters agentliyinin jurnalistləri tərəfindən yazılmış rəsmi xəbər məqalələrini əhatə edir. Verilənlər bazası 50 fərqli müəllif tərəfindən yazılmış 5000 məqalədən ibarətdir ki, hər bir müəllif üçün tam olaraq 100 məqalə təqdim olunur. Tədqiqatda ilk 10 müəllif və hər birindən 30 məqalə olmaqla, ümumilikdə 300 sənəd seçilmişdir.

Tədqiqatda istifadə olunan verilənlər bazaları haqqında məlumat Cədvəl 1-də əks olunmuşdur.

CƏDVƏL I. ENRON VƏ REUTERS C50 VERİLƏNLƏR BAZALARI HAQQINDA MƏLUMAT

| **Verilənlər bazası** | ***Enron*** | ***Reuters C50*** |
|---|---|---|
| Müəllif Sayı | 10 | 10 |
| Sənəd Sayı | 500 | 300 |
| Mətn növü | Qeyri-rəsmi e-poçt | Rəsmi məqalə |

### *B. Parametrlərin təyini*

Hər iki verilənlər bazası üçün eksperimentlərdə mətn xüsusiyyətlərinin çıxarılması mərhələsində hibrid stilometrik yanaşmadan istifadə olunmuşdur. Bu məqsədlə söz və simvol səviyyəsində TF-IDF xüsusiyyətləri çıxarılmış, daha sonra bu xüsusiyyətlər birləşdirilərək ölçünün azaldılması üçün TruncatedSVD tətbiq olunmuş və nəticələr normallaşdırılmışdır. Model arxitekturası tam əlaqəli çoxlaylı perseptron (Multi-Layer Perceptron) əsasında qurulmuş, kodlayıcı və dekodlayıcı hissələri simmetrik strukturda təşkil edilmiş və bütün gizli laylarda ReLU aktivasiya funksiyasından istifadə olunmuşdur. Avtokodlayıcı modeli əvvəlcədən öyrətmə mərhələsində 150 iterasiya ərzində Adam optimizatoru ilə təlim edilmişdir. Daha sonra DEC və IDEC modelləri üçün birgə təlim mərhələsində SGD optimizatoru istifadə olunmuş və model 2000 iterasiya ərzində öyrədilmişdir.

### *C. Nəticələr və Analiz*

Dərin klasterləşdirmə modellərinin müəllif klasterləşdirilməsi problemində effektivliyini qiymətləndirmək məqsədilə DEC və IDEC modelləri həm sərbəst üslublu (Enron), həm də rəsmi-akademik (C50) dil xüsusiyyətlərinə malik verilənlər bazaları üzərində sınaqdan keçirilmişdir. Modellərin performansı üç əsas metrika — Dəqiqlik ($Acc$), Normallaşdırılmış Qarşılıqlı İnformasiya ($NMI$) və Korreksiya olunmuş Rand indeks ($ARI$) vasitəsilə qiymətləndirilmişdir. Aparılan eksperimentlərin müqayisəli nəticələri Cədvəl II və Cədvəl III-də təqdim olunmuşdur. Burada K parametri verilənlər bazasındakı potensial müəllif qruplarının sayını ifadə edir.

Cədvəl II - Enron verilənlər bazası üzrə əldə olunan nəticələr göstərir ki, klaster sayının artması ilkin mərhələdə modellərin performansına müsbət təsir göstərmişdir. Xüsusilə K=11 olduqda həm DEC, həm də IDEC modelləri ən yüksək nəticələri əldə etmişdir. DEC modeli üçün ən yüksək Acc göstəricisi 0.7700, NMI göstəricisi 0.7216 və ARI göstəricisi 0.5955 olmuşdur. IDEC modeli isə eyni verilənlər bazasında daha yüksək nəticələr göstərmişdir. Bu üstünlük IDEC modelində rekonstruksiya itkisinin nəzərə alınması nəticəsində gizli fəzanın strukturunun daha yaxşı qorunması ilə izah olunur.

CƏDVƏL II. ENRON VERİLƏNLƏR BAZASI ÜZRƏ DƏRİN KLASTERLƏŞDİRMƏ NƏTİCƏLƏRİ

| **Indeks / Model** \ **K** | | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Acc** | *DEC* | 0.1940 | 0.2700 | 0.3560 | 0.4280 | 0.4580 | 0.5520 | 0.6560 | 0.7360 | 0.7400 | **0.7700** | 0.7460 | 0.6760 | 0.6980 | 0.6820 |
| | *IDEC* | 0.1880 | 0.2880 | 0.3740 | 0.4220 | 0.5040 | 0.5820 | 0.6380 | 0.6940 | 0.7020 | **0.8200** | 0.7220 | 0.7140 | 0.6820 | 0.6680 |
| **NMI** | *DEC* | 0.2527 | 0.3283 | 0.3857 | 0.5183 | 0.4842 | 0.5688 | 0.6307 | 0.6857 | 0.6802 | 0.6803 | 0.6984 | 0.6832 | **0.7216** | 0.6983 |
| | *IDEC* | 0.1720 | 0.4159 | 0.4221 | 0.4943 | 0.5282 | 0.6153 | 0.6260 | 0.6527 | 0.6553 | **0.7351** | 0.6968 | 0.7082 | 0.6938 | 0.7168 |
| **ARI** | *DEC* | 0.1218 | 0.1774 | 0.2495 | 0.3280 | 0.2903 | 0.4111 | 0.4814 | 0.5724 | 0.5750 | **0.5955** | 0.5836 | 0.5350 | 0.5844 | 0.5850 |
| | *IDEC* | 0.0809 | 0.2203 | 0.2690 | 0.3030 | 0.3741 | 0.4587 | 0.4766 | 0.5379 | 0.5202 | **0.6544** | 0.5825 | 0.5976 | 0.5632 | 0.5871 |

CƏDVƏL III. C50 VERİLƏNLƏR BAZASI ÜZRƏ DƏRİN KLASTERLƏŞDİRMƏ NƏTİCƏLƏRİ

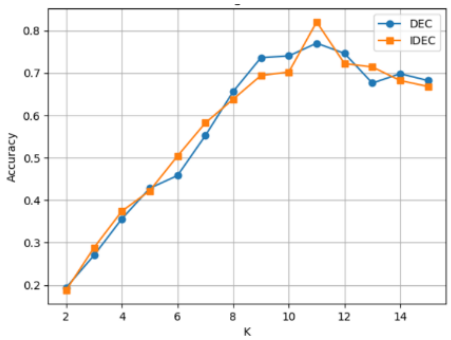
| Indeks / Model | | K=2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Acc** | *DEC* | 0.2000 | 0.2733 | 0.3667 | 0.4067 | 0.4533 | 0.5067 | 0.5033 | 0.5333 | **0.5833** | 0.4167 | 0.5467 | 0.5367 | 0.4833 | 0.5067 |
| | *IDEC* | 0.2000 | 0.2933 | 0.3833 | 0.4433 | 0.5300 | 0.4733 | 0.4333 | **0.6500** | 0.5833 | 0.4367 | 0.4967 | 0.5633 | 0.5400 | 0.5133 |
| **NMI** | *DEC* | 0.2595 | 0.2464 | 0.4196 | 0.4256 | 0.4394 | 0.5295 | 0.4801 | 0.5336 | 0.6002 | 0.4858 | **0.6167** | 0.5497 | 0.5853 | 0.5692 |
| | *IDEC* | 0.2143 | 0.3035 | 0.4795 | 0.5157 | 0.5882 | 0.4652 | 0.4179 | **0.6577** | 0.5738 | 0.4369 | 0.5633 | 0.6165 | 0.5976 | 0.5843 |
| **ARI** | *DEC* | 0.0935 | 0.1061 | 0.2222 | 0.2194 | 0.2626 | 0.3194 | 0.2876 | 0.3384 | **0.3764** | 0.2616 | 0.3749 | 0.3648 | 0.3302 | 0.3315 |
| | *IDEC* | 0.0913 | 0.1415 | 0.2866 | 0.2962 | 0.4066 | 0.2761 | 0.2176 | **0.4763** | 0.3662 | 0.2192 | 0.3291 | 0.3803 | 0.3490 | 0.3436 |

C50 verilənlər bazası üzrə nəticələr göstərir ki, IDEC modeli əksər klaster saylarında DEC modeli ilə müqayisədə daha yüksək performans nümayiş etdirmişdir. Xüsusilə K=9 olduqda IDEC modeli ən yüksək nəticələri əldə etmiş və Acc, NMI, ARI göstəriciləri müvafiq olaraq 0.6500, 0.6577 və 0.4763 təşkil etmişdir. DEC modeli üçün isə ən yüksək ACC göstəricisi K=10-da 0.5833, ən yüksək NMI göstəricisi K=12-də 0.6167 və ən yüksək ARI göstəricisi isə K=10-da 0.3764 olmuşdur.

Ümumilikdə, hər iki verilənlər bazası üzrə əldə olunan nəticələr göstərir ki, IDEC modeli müəllif klasterləşdirmə problemində daha effektiv model kimi qiymətləndirilə bilər. Bu model klasterləşdirmə itkisini və rekonstruksi ya itkisini eyni vaxtda optimallaşdırdığı üçün daha məzmunlu xüsusiyyətlərin formalaşmasına şərait yaradır. Nəticədə, onun rəqəmsal kriminalistika kimi praktik tətbiq sahələrində istifadəsi məqsədəuyğun hesab olunur.

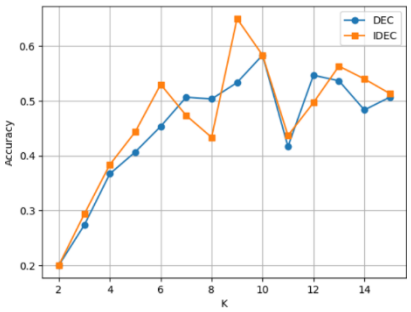
## D. Vizuallaşdırma

Şəkil 2 və Şəkil 3-də uyğun olaraq Enron və C50 verilənlər bazaları üçün müəllif klasterləşdirməsi nəticələrinin vizuallaşdırılması təqdim olunmuşdur. Təqdim olunan qrafiklər, klaster sayından (K) asılı olaraq dəqiqlik (Accuracy) indeksinin dəyişməsini əks etdirir.




Şəkil 2. Enron verilənlər bazası üzrə klaster sayından asılı olaraq dəqiqlik indeksinin dəyişməsi.

Hər iki şəkildə əldə olunmuş qrafiklər əsasında modellərin performansı klaster sayından asılı olaraq ümumiləşdirilmiş şəkildə qiymətləndirilə bilər. Ümumilikdə müşahidə olunur ki, klaster sayının artması ilkin mərhələdə dəqiqliyin yüksəlməsinə səbəb olur, lakin müəyyən həddən sonra bu artım əksinə azalır. Bundan əlavə, IDEC modelinin əksər hallarda DEC modelindən daha yüksək nəticə göstərdiyi aydın görünür.




Şəkil 3. C50 verilənlər bazası üzrə klaster sayından asılı olaraq dəqiqlik indeksinin dəyişməsi.

Şəkil 2 (Enron dataseti) üzrə analiz göstərir ki, klasterlərin sayı K=2-dən K=11-ə qədər artdıqca hər iki modelin dəqiqliyi yüksəlir. K=11-də hər iki model maksimum dəqiqliyə çatır. Bu nöqtədən sonra klaster sayının artırılması nəticələrin azalmasına səbəb olur. IDEC modeli bir çox hallarda üstün olsa da, ümumilikdə hər iki model oxşar nəticə nümayiş etdirir.

Şəkil 3 (C50 dataseti) üzrə analiz göstərir ki, qrafikdə kəskin artım və azalmalar mövcuddur. Bu, C50 datasetindəki mətnlərin bir-birinə daha yaxın yazı üslublarına malik olduğunu və modellərin klaster sərhədlərini müəyyən etməkdə çətinlik çəkdiyini göstərir. Ən yüksək performans IDEC modeli tərəfindən K=9 nöqtəsində qeydə alınmışdır. K=10-dən sonra hər iki modeldə performansın azalması müşahidə olunur və bu klasterlərin həddindən artıq bölünməsi ilə izah edilə bilər. Ümumilikdə, optimal klaster sayının K=9 kimi qiymətləndirilə bilər.

## NƏTİCƏ

Bu tədqiqat işində müəllif klasterləşdirməsi üçün eyni anda yenilənən - yəni xüsusiyyətlərin öyrənilməsi (feature learning) modulu ilə klasterləşmə modulunun birgə (jointly) optimallaşdırıldığı - dərin klasterləşdirmə modelləri tətbiq edilmişdir. Beləliklə, DEC və IDEC modellərinin praktik effektivliyi və müəllif üslublarını müəyyən etmə imkanı araşdırılmışdır.

Eksperimental təhlillər göstərir ki, birgə optimallaşdırma (joint optimization) prinsipli modellər müəllif klasterləşdirilmə məsələsində yüksək effektivlik nümayiş etdirmişdir. Bu modellərin əsas üstünlüyü, xüsusiyyətlərin təsviri (feature representation) ilə klasterləşdirmə parametrlərinin birgə optimallaşdırılmasıdır ki, bu da yüksək ölçülü məlumatlarda daha dəqiq qruplaşdırmaya imkan verir.

Xüsusilə IDEC modeli müəllif klasterləşdirilməsi məsələsində üstün nəticələr göstərmişdir. Modelin arxitekturasında yer alan rekonstruksiya itkisinin klasterləşmə itkisi ilə paralel optimallaşdırılması, müəlliflərin yazı üslubunu ifadə edən detalların itməsinin qarşısını almışdır. Buna baxmayaraq, sahənin gələcək inkişafı üçün həlli vacib olan bir neçə fundamental problem hələ də aktual olaraq qalır. Bunlara aşağıdakılar daxildir:

- Müəlliflik analizi sahəsində aparılan tədqiqatların əsasən ingilis dili üzərində cəmlənməsi və digər dillərin kifayət qədər araşdırılmaması mövcud yanaşmaların müxtəlif dil mühitlərində tətbiqində ciddi problem kimi çıxış edir. Bu da fərqli dil xüsusiyyətlərini nəzərə alan daha genişmiqyaslı tədqiqatların aparılmasını zəruri edir.

- Hazırda ən vacib problemlərdən biri mətnin hansı hissəsinin insana, hansı hissəsinin maşına aid olduğunu müəyyən etməkdir. Bu tip müəlliflik (human-AI co-authorship), ənənəvi müəllif analizi metodlarını səmərəsiz edir. Çünki mövcud alqoritmlər adətən mətnin tək bir müəllifə aid olduğunu fərz edir. Böyük Dil Modelinin (BDM) mətn redaktəsi və ya insanın süni intellekt mətnlərini modifikasiya etməsi, orijinal müəlliflik izlərini təhrif edir. Bu vəziyyət, müəllifliyin etibarlı şəkildə təyin olunmasını çətinləşdirir.

- Süni intellekt tərəfindən yaradılan məzmunun həcminin artması mətnin müəllifinin insan, yoxsa BDM olduğunu müəyyənləşdirməyi aktual bir problemə çevirmişdir. [10]. Müasir rəqəmsal mühitdə süni intellektin insan yazısına xas olan sintaktik və semantik xüsusiyyətləri yüksək dəqiqliklə təqlid etməsi, mətnin insan tərəfindən yazılıb-yazılmadığının müəyyən edilməsini (AI detection) kritik bir məsələyə çevirmişdir. Burada əsas hədəf, süni intellekt tərəfindən yaradılan kontenti aşkar edərək akademik dürüstlüyü və informasiya etibarlılığını qorumaqdır.

- Huang və Chen [11] tərəfindən təqdim edilən müasir tədqiqat istiqamətlərindən biri BDM-lərin müəllif kimi identifikasiyasıdır. Bu istiqamət, mətnin hansı spesifik Böyük Dil Modeli (məsələn, GPT-4, Llama və ya Claude) tərəfindən yaradıldığını müəyyən etməyi nəzərdə tutur. Sözügedən yanaşma, hər bir modelin özünəməxsus "rəqəmsal üslubunu" və ya linqvistik xüsusiyyətlərini analiz edərək, mətnin məhz hansı arxitektura tərəfindən generasiya edildiyini müəyyənləşdirməkdir. Bu tədqiqat sahəsi hüquqi məsuliyyətin təyin olunması və rəqəmsal kriminalistika baxımından müstəsna əhəmiyyət kəsb edir.

Tədqiqat çərçivəsində tətbiq edilən bu modellər müəlliflik analizi prosesinin dərindən analiz edilməsinə imkan yaradaraq, rəqəmsal kriminalistika sahəsində daha dəqiq və obyektiv qərarların qəbul edilməsini təmin edəcəkdir.

ƏDƏBİYYAT

[1] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi, "Mining writeprints from anonymous e-mails for forensic investigation," Digit. Investig., vol. 7, nos. 1–2, pp. 56–64, 2010.

[2] H. Gómez-Adorno, Y. Aleman, D. V. Ayala, M. A. Sanchez-Perez, D. Pinto, and G. Sidorov, "Author clustering using hierarchical clustering analysis," in CLEF (Working Notes), Sep. 2017.

[3] R. M. Aliguliyev, "Performance evaluation of density-based clustering methods," Information Sciences, vol. 179, no. 20, pp. 3583–3602, 2009.

[4] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in Proc. Int. Conf. Mach. Learn. (ICML), 2016, pp. 478–487.

[5] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in Proc. IJCAI, 2017, pp. 1753–1759.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

[7] D. Zhang, Y. Sun, B. Eriksson, and L. Balzano, "Deep unsupervised clustering using mixture of autoencoders," arXiv preprint arXiv:1712.07788, 2017.

[8] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," IEEE Trans. Neural Netw., vol. 20, no. 2, pp. 189–201, 2009.

[9] S. Zhou, H. Xu, Z. Zheng, J. Chen, Z. Li, J. Bu, X. Wang, B. Zhang, D. Wu, C. Wang, and M. Ester, "A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions," ACM Comput. Surv., vol. 57, no. 3, pp. 1–38, 2024.

[10] N. Habib, T. Adewumi, M. Liwicki, and E. Barney, "Trends and challenges in authorship analysis: A review of ML, DL, and LLM approaches," arXiv preprint arXiv:2505.15422, 2025.

[11] B. Huang, C. Chen, and K. Shu, "Authorship attribution in the era of LLMs: Problems, methodologies, and challenges," ACM SIGKDD Explorations Newsl., vol. 26, no. 2, pp. 21–43, 2025.

## Investigation of Authorship Identification Problems Using Text Mining Technologies

**Leyla Mammadova**
Institute of Information Technology, Baku, Azerbaijan
*mammedova-1998@list.ru*

**Abstract**- In recent years, cybercriminals have extensively utilized text-based communication tools across emails, social media, and other online environments to maintain anonymity. For this reason, authorship identification, particularly authorship clustering, stands out as a crucial research problem. In this study, deep clustering models were applied to the authorship clustering problem, and their effectiveness in this field was analyzed. The obtained results demonstrate that deep clustering models serve as a promising direction for solving the authorship clustering problem and possess significant potential for practical application in the field of digital forensics.

**Keywords**- digital forensics; authorship clustering; deep clustering.

# II SEKSİYA

# Rəqəmsal kriminalistikanın elmi-texnoloji problemləri