TELEGRAM KANALLARINI MAVZUGA KO‘RA KLASTERLASHDA TF-IDF VA SENTENCE-BERT YONDASHUVLARINING SOLISHTIRMA TAHLILI

Babomurodov Ozod Jurayevich
Doctor of Science (DSc)
Jizzax shaxar sun’iy intellekt yo’nalishi bo’yicha hokim o’rinbosari.
Qo’yliyeva Feruzaxon Alisher qizi
Toshkent Davlat Agrar Universitet assisenti.

https://doi.org/10.5281/zenodo.20166346

Annotatsiya. *Mazkur tadqiqotda Telegram kanallaridagi matnlarni mavzuga ko‘ra avtomatik guruhlash jarayonida ikki xil vektorlashtirish yondashuvi — TF-IDF va Sentence-BERT —ning samaradorligi solishtirildi. Dastlab kanal xabarlari tozalanib, standart shaklga keltirildi. TF-IDF statistik xususiyatlarga asoslangan yuqori o‘lchamli vektorlarni yaratdi, Sentence-BERT esa qisqa Telegram xabarlarining semantik mazmunini chuqur aks ettiruvchi kontekstual embeddinglar hosil qildi. Har ikki yondashuvda K-Means algoritmi qo‘llanib, natijalar siluet ko‘rsatkichi, Davies–Bouldin indeksi va qo‘lda semantik tahlil orqali baholandi.*

Tadqiqot natijalariga ko‘ra, semantik jihatdan izchil va mavzuviy jihatdan bir xil klasterlar shakllantirishda Sentence-BERT TF-IDFga nisbatan ancha ustun ekanligi isbotlandi.

Kalit so‘zlar: *TF-IDF, Sentence-BERT, klasterlash, Telegram yozishmalari, semantik tahlil, ijtimoiy tarmoq monitoringi.*

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ПОДХОДОВ TF-IDF И SENTENCE-BERT ПРИ КЛАСТЕРИЗАЦИИ TELEGRAM-КАНАЛОВ ПО ТЕМАТИКЕ

Аннотация. *В данном исследовании сравнивается эффективность двух подходов к векторизации текстов — TF-IDF и Sentence-BERT — при автоматической тематической кластеризации сообщений Telegram-каналов. На первом этапе сообщения каналов были очищены и приведены к стандартному формату. TF-IDF формировал высокоразмерные векторы на основе статистических характеристик текста, тогда как Sentence-BERT создавал контекстуальные embedding-векторы, глубже отражающие семантическое содержание коротких Telegram-сообщений. Для обеих моделей векторизации применялся алгоритм кластеризации K-Means. Полученные результаты оценивались с использованием коэффициента силуэта, индекса Davies–Bouldin, а также ручного семантического анализа. Результаты исследования показали, что Sentence-BERT значительно превосходит TF-IDF при формировании семантически согласованных и тематически однородных кластеров.*

Ключевые слова: *TF-IDF, Sentence-BERT, кластеризация, Telegram-сообщения, семантический анализ, мониторинг социальных сетей.*

COMPARATIVE ANALYSIS OF TF-IDF AND SENTENCE-BERT APPROACHES FOR TOPIC-BASED CLUSTERING OF TELEGRAM CHANNELS

Abstract. *This study compares the effectiveness of two text vectorization approaches — TF-IDF and Sentence-BERT — in the automatic topic-based clustering of Telegram channel messages. At the initial stage, channel messages were cleaned and transformed into a standardized format. TF-IDF generated high-dimensional vectors based on statistical text features, while Sentence-BERT produced contextual embeddings that more accurately captured the semantic meaning of short Telegram messages. The K-Means clustering algorithm was applied to both vectorization approaches. The obtained results were evaluated using the silhouette coefficient, Davies–Bouldin index, and manual semantic analysis.*

The findings demonstrated that Sentence-BERT significantly outperforms TF-IDF in creating semantically consistent and topically coherent clusters.

Keywords: *TF-IDF, Sentence-BERT, clustering, Telegram messages, semantic analysis, social network monitoring.*

Kirish. Telegram ijtimoiy tarmog‘i O‘zbekistonda axborot almashinuvi, jamoaviy kommunikatsiya va turli ko‘rinishdagi kontentlarni tarqatishning eng faol platformalaridan biriga aylangan. So‘nggi yillarda Telegram kanallari, guruhlari va botlari sonining keskin ortishi natijasida foydalanuvchilar tomonidan yaratilayotgan matnli ma’lumotlar hajmi ham sezilarli darajada ko‘paydi. Mazkur platformada yangiliklar, reklama, siyosiy fikrlar, ijtimoiy muhokamalar, shaxsiy yozishmalar va turli mavzudagi kontentlar tezkor ravishda tarqalmoqda.

Shu bilan birga, noto‘g‘ri axborot, manipulyativ xabarlar, spam, nafrat nutqi hamda yoshlar ongiga salbiy ta’sir ko‘rsatishi mumkin bo‘lgan yozishmalar soni ham ortib bormoqda.

Natijada Telegramdagi matnli kontentni avtomatik tahlil qilish, monitoring qilish va xavfli axborotni erta aniqlash masalasi dolzarb ilmiy-amaliy vazifalardan biriga aylandi.

Ijtimoiy tarmoq matnlarini avtomatik tahlil qilishda asosiy bosqichlardan biri — matnlarni vektor ko‘rinishga o‘tkazish jarayonidir. Matnlarni klasterlash va klassifikatsiya qilish sifatiga aynan vektorlashtirish usuli katta ta’sir ko‘rsatadi.

An’anaviy TF-IDF (Term Frequency – Inverse Document Frequency) usuli matndagi so‘zlarning uchrash chastotasi va statistik og‘irligiga asoslanadi. Ushbu yondashuv katta hajmdagi rasmiy matnlar uchun samarali bo‘lsa-da, qisqa va norasmiy ijtimoiy tarmoq yozishmalarida semantik ma’noni to‘liq ifodalay olmaydi. Ayniqsa Telegramdagi qisqa gaplar, slanglar, shevalar, emoji va qisqartmalar TF-IDF modelida yetarlicha kontekst hosil qilmaydi.

So‘nggi yillarda chuqur o‘rganish (deep learning) asosidagi transformer modellarining rivojlanishi matnlarni semantik jihatdan chuqurroq tahlil qilish imkonini yaratdi. Sentence-BERT (SBERT) modeli jumlalarni kontekstual embeddinglar orqali ifodalab, ularning ma’no jihatdan o‘xshashligini aniqlashda yuqori samaradorlik ko‘rsatmoqda. Ushbu model nafaqat so‘z chastotasini, balki gapning umumiy mazmuni va kontekstini ham hisobga oladi. Natijada o‘xshash mazmundagi yozishmalar bir-biriga yaqin vektor maydonida joylashadi va klasterlash sifati sezilarli yaxshilanadi.

Telegram muhitining o‘ziga xos jihati shundaki, foydalanuvchilar ko‘pincha qisqa, grammatik jihatdan to‘liq bo‘lmagan, hissiyotga boy va norasmiy uslubdagi yozishmalardan foydalanadi. Bu esa an’anaviy statistik yondashuvlar uchun murakkab vaziyatni yuzaga keltiradi.

Shu sababli TF-IDF va Sentence-BERT usullarini aynan Telegram yozishmalari asosida taqqoslash ilmiy jihatdan muhim hisoblanadi. Ushbu taqqoslash orqali semantik vektorlashtirishning klasterlash sifatiga ta’siri, xavfli kontentni aniqlashdagi ustunliklari hamda katta hajmdagi ijtimoiy tarmoq ma’lumotlarini qayta ishlashdagi samaradorligi baholanadi.

Mazkur tadqiqotda Telegramdan yig‘ilgan o‘zbek tilidagi matnlar asosida TF-IDF va Sentence-BERT yondashuvlari yordamida klasterlash tajribalari o‘tkazildi. Tadqiqotning asosiy maqsadi — qisqa ijtimoiy tarmoq yozishmalarini semantik jihatdan samarali guruhlash, xavfli va xavfsiz kontentni avtomatik ajratish hamda keyingi bosqichdagi transformer modellarini qayta o‘qitish uchun sifatli dataset yaratishdan iborat. Olingan natijalar ijtimoiy tarmoqlarda zararli axborotni erta aniqlash va monitoring qilish tizimlarini ishlab chiqishda amaliy ahamiyatga ega hisoblanadi.

Asosiy qism. Telegram kanallaridan olingan matnlar asosida mavzuga ko‘ra avtomatik klasterlash jarayonida TF-IDF va Sentence-BERT vektorlashtirish yondashuvlarining samaradorligini taqqoslash, ularning afzallik va cheklovlarini aniqlash hamda real monitoring tizimlari uchun eng optimal variantni belgilash mazkur tadqiqotning asosiy vazifalaridan biri hisoblanadi. Tadqiqotda ijtimoiy tarmoqlardagi qisqa va norasmiy yozishmalarning semantik tuzilishini aniqlash, o‘xshash mazmundagi matnlarni bir guruhga jamlash hamda xavfli yoki manipulyativ kontentni erta aniqlash imkoniyatlari tahlil qilindi.

Telegram muhitida foydalanuvchilar ko‘pincha qisqartmalar, slang birliklar, shevaga oid ifodalar, emojilar va grammatik jihatdan to‘liq bo‘lmagan gaplardan foydalanadi. Shu sababli bunday yozishmalarni klassik statistik metodlar yordamida tahlil qilish qiyin hisoblanadi.

Tadqiqotda aynan semantik jihatdan yaqin bo‘lgan yozishmalarni klasterlash sifati, turli vektorlashtirish usullarining qisqa matnlardagi barqarorligi va monitoring tizimlarida qo‘llash samaradorligi baholandi.

Metodlar. Tadqiqot uchun Telegramning turli tematik kanallaridan yig‘ilgan matnlar yagona korpusga birlashtirildi va oldindan qayta ishlash bosqichidan o‘tkazildi. Dataset tarkibiga yangiliklar, ijtimoiy muhokamalar, texnologiya, siyosat, reklama va umumiy muloqot xarakteridagi yozishmalar kiritildi. Matnlarni tozalash jarayonida URL manzillar, emoji, reklama naqshlari, texnik bot xabarlari, takroriy yozuvlar, ortiqcha belgilar hamda o‘zbek tilida keng uchraydigan stop-so‘zlar olib tashlandi. Shuningdek, matnlar yagona formatga keltirilib, barcha belgilar pastki registrga o‘tkazildi. Semantik mazmunga ega bo‘lmagan juda qisqa xabarlar, faqat stiker yoki emoji asosidagi yozishmalar filtrdan chiqarildi.

Keyingi bosqichda matnlar ikki xil yondashuv — TF-IDF va Sentence-BERT yordamida vektorlashtirildi. TF-IDF (Term Frequency – Inverse Document Frequency) usulida 1-gram va 2-gram asosidagi statistik vaznlar hisoblandi hamda matnlar yuqori o‘lchamli sparce vektorlar shaklida ifodalandi. O‘lchamning haddan tashqari oshib ketishini oldini olish maqsadida maksimal xususiyatlar soniga cheklov qo‘yildi va vektorlar L2 normalizatsiya orqali standartlashtirildi. Ushbu yondashuv so‘zlarning uchrash chastotasini samarali aks ettirsa-da, matnning umumiy konteksti va semantik ma’nosini chuqur ifodalab bera olmaydi. Ayniqsa sinonimlar, qisqartmalar yoki turli shakldagi bir xil mazmundagi gaplarni farqlashda TF-IDF cheklangan natija ko‘rsatadi.

Semantik va kontekstual tahlil sifatini oshirish maqsadida Sentence-BERT (SBERT) modeli qo‘llanildi. Ushbu model transformer arxitekturasi asosida ishlab chiqilgan bo‘lib, matnlarni 384 o‘lchamli zich embeddinglar ko‘rinishida ifodalaydi. SBERT modeli matnning umumiy ma’nosini hisobga olib, semantik jihatdan yaqin yozishmalarni vektor maydonida bir-biriga yaqin joylashtiradi.

Bu ayniqsa Telegramdagi qisqa, norasmiy va emotsional yozishmalarni tahlil qilishda muhim ustunlik beradi. Sinonimlar, yashirin ma’no, kontekstual bog‘liqlik va bir xil mazmunni turli shaklda ifodalovchi yozishmalar SBERT yordamida ancha aniq aniqlanadi. Hosil qilingan embeddinglar cosine similarity asosida baholash uchun tayyorlandi.

Har ikkala vektorlashtirish yondashuvi uchun K-Means klasterlash algoritmi qo‘llanildi.

Tajriba jarayonida k = 4 va k = 5 qiymatlar sinovdan o‘tkazildi hamda klasterlarning mazmuniy barqarorligi o‘zaro taqqoslandi. Algoritmning iteratsiyalar soni 300 etib belgilandi, boshlang‘ich markazlarni tanlashda esa k-means++ usuli qo‘llanildi. Ushbu yondashuv lokal minimumga tushib qolish ehtimolini kamaytirib, klasterlash natijalarining barqarorligini oshirishga xizmat qildi.

Natijalar sifatini baholash uchun Silhouette Score va Davies–Bouldin Index ko‘rsatkichlaridan foydalanildi. Silhouette Score klaster ichidagi yaqinlik va klasterlararo farqni baholash imkonini berdi. Ushbu ko‘rsatkichning yuqori qiymati klasterlash sifatining yaxshi ekanligini bildiradi. Davies–Bouldin indeksi esa klasterlararo o‘xshashlik darajasini aniqlash uchun ishlatildi; indeks qiymati qancha kichik bo‘lsa, klasterlash sifati shuncha yaxshi hisoblanadi. Bundan tashqari, klasterlarning mazmuniy tozaligi va mavzular bo‘yicha ajralish darajasi qo‘lda ekspert tahlili orqali ham tekshirildi.

Tajriba natijalari TF-IDF va Sentence-BERT yondashuvlari orasidagi sezilarli farqlarni ko‘rsatdi. TF-IDF tez ishlashi va kam resurs talab qilishi bilan ajralib tursa-da, qisqa Telegram yozishmalarida semantik aniqlik bo‘yicha SBERT’dan past natija ko‘rsatdi. Sentence-BERT esa mazmuniy jihatdan yaqin yozishmalarni aniqroq klasterlashga muvaffaq bo‘ldi va ayniqsa xavfli yoki manipulyativ kontentlarni bir guruhga jamlashda yuqori samaradorlik namoyish etdi. Bu esa real monitoring tizimlarida kontekstual embeddinglardan foydalanish samaraliroq ekanligini ko‘rsatadi.

Natijalar

O‘tkazilgan tajribalar natijalari TF-IDF va Sentence-BERT yondashuvlari orasida klasterlash sifati bo‘yicha sezilarli farqlar mavjudligini ko‘rsatdi. TF-IDF usuli uzun va nisbatan rasmiy matnlardan tashkil topgan kanallarda, ayniqsa tahliliy maqolalar va keng izohli postlarda mavzularni aniq ajratishga muvaffaq bo‘ldi. So‘z chastotasi va statistik vaznlarga asoslangan ushbu yondashuv texnik yoki aniq terminologiyaga ega matnlarda yaxshi natija berdi. Masalan, iqtisodiyot, sport yoki texnologiyaga oid uzun postlarda muhim kalit so‘zlar yuqori vazn olgani sababli klasterlar o‘zaro yaxshi farqlandi.

Biroq qisqa Telegram postlari, reklama xabarlari, emojilar bilan boyitilgan yozishmalar va norasmiy uslubdagi gaplarda TF-IDF semantik yaqinlikni yetarlicha aniq baholay olmadi.

Ayniqsa turli mavzudagi kanallarda bir xil reklama yoki ommabop iboralarning tez-tez uchrashi klasterlash sifatiga salbiy ta’sir ko‘rsatdi. Masalan, “aksiya boshlandi”, “chegirma bor”, “sovg‘a yutib ol”, “hoziroq ulaning” kabi iboralar turli tematik kanallarda takrorlangani sababli TF-IDF ushbu yozishmalarni semantik jihatdan bog‘liq deb baholab, bir xil klasterga joylashtirdi. Natijada klaster ichida mavzu jihatidan aralash holatlar kuzatildi.

Shuningdek, TF-IDF sinonim birliklarni, shevaga oid so‘zlarni va kontekstual jihatdan o‘xshash bo‘lgan, ammo turli so‘zlar bilan yozilgan gaplarni bir-biriga yaqinlashtirishda qiyinchilikka duch keldi. Bu ayniqsa qisqa yozishmalarda yaqqol namoyon bo‘ldi, chunki qisqa matnlarda statistik ma’lumotlar hajmi kam bo‘lgani sababli model yetarli semantik bog‘liqlik hosil qila olmadi. Natijada ayrim klasterlarda mavzularning chalkashishi va ichki bir xillikning pasayishi kuzatildi.

Sentence-BERT yondashuvi esa qisqa matnlar, sinonimlar, emojilar va og‘zaki nutqqa yaqin yozishmalarda ham semantik ma’noni ancha yaxshi saqlab qoldi. Transformer asosidagi embeddinglar matnning umumiy mazmunini hisobga olgani sababli, turli shaklda yozilgan, ammo ma’nosi bir xil bo‘lgan xabarlar bir klasterga muvaffaqiyatli birlashtirildi. Masalan, “telefon arzonlashdi”, “smartfon uchun chegirma”, “mobil qurilmalar aksiyasi” kabi turli shakldagi reklama yozishmalari bir xil tematik klasterda jamlandi.

SBERT modeli xavfsizlik, iqtisodiyot, texnologiya, reklama va umumiy muloqotga oid yozishmalarni TF-IDF ga nisbatan sezilarli darajada aniqroq ajrata oldi. Ayniqsa xavfli yoki manipulyativ kontentni o‘z ichiga olgan xabarlar alohida klasterlarda barqaror shakllandi.

Bu esa semantik embeddinglarning ijtimoiy tarmoqlardagi monitoring tizimlari uchun samaraliroq ekanligini ko‘rsatdi.

Natijalarni miqdoriy baholashda Silhouette Score va Davies–Bouldin Index ko‘rsatkichlari qo‘llanildi. Tajribalar davomida Sentence-BERT asosidagi klasterlashda Silhouette Score qiymatlari TF-IDF ga nisbatan yuqoriroq bo‘ldi, bu esa klaster ichidagi yozishmalar o‘zaro yaqin va klasterlar bir-biridan yaxshi ajralganligini ko‘rsatadi.

Davies–Bouldin indeksida ham SBERT yaxshiroq natijalarni namoyish etdi, ya’ni klasterlararo chalkashlik darajasi past bo‘ldi.

Qo‘lda o‘tkazilgan ekspert tahlili ham ushbu natijalarni tasdiqladi. TF-IDF klasterlarida mavzu jihatidan aralash yozishmalar ko‘proq uchragan bo‘lsa, SBERT klasterlarida ichki mavzu birligi ancha yuqori bo‘ldi. Ayniqsa qisqa Telegram yozishmalarida kontekstual ma’noni hisobga olish klasterlash sifatiga sezilarli ijobiy ta’sir ko‘rsatdi. Shu sababli Sentence-BERT real vaqt monitoring tizimlari, xavfli kontentni aniqlash va ijtimoiy tarmoqlardagi yozishmalarni semantik tahlil qilish vazifalari uchun yanada istiqbolli yondashuv sifatida baholandi.

Xulosa

O‘tkazilgan tajribalar natijalari Telegram ijtimoiy tarmog‘idagi qisqa, norasmiy va kontekstga boy yozishmalarni avtomatik klasterlash jarayonida Sentence-BERT yondashuvi TF-IDF usuliga nisbatan sezilarli ustunlikka ega ekanligini ko‘rsatdi. TF-IDF statistik jihatdan so‘z chastotasiga asoslanganligi sababli uzun va rasmiy matnlarda yetarlicha samarali ishlagan bo‘lsa-da, qisqa Telegram xabarlarida semantik bog‘liqlikni to‘liq aks ettira olmadi. Ayniqsa reklama xarakteridagi iboralar, sinonim birliklar, emojilar va og‘zaki nutqqa yaqin yozishmalarda ushbu usulning cheklovlari yaqqol namoyon bo‘ldi.

Sentence-BERT modeli esa transformer arxitekturasi asosida kontekstual embeddinglar hosil qilgani sababli qisqa yozishmalarning ichki ma’nosini samarali saqlab qoldi. Model turli shaklda yozilgan, biroq mazmun jihatidan o‘xshash bo‘lgan matnlarni muvaffaqiyatli tarzda bir klasterga birlashtirdi. Bu esa klasterlarning mavzu jihatidan yanada toza, barqaror va semantik jihatdan aniq shakllanishiga olib keldi. Tajriba natijalarida Silhouette Score ko‘rsatkichlarining yuqoriligi va Davies–Bouldin indeksining pastligi ham SBERT yondashuvining ustunligini tasdiqladi.

Tadqiqot natijalari real vaqt rejimida ishlovchi axborot monitoring tizimlari uchun muhim amaliy ahamiyatga ega. Xususan, Telegram va boshqa ijtimoiy tarmoqlarda kontentni mavzular bo‘yicha avtomatik saralash, manipulyativ yoki xavfli yozishmalarni erta aniqlash, spam va reklama kontentlarini filtrlash kabi vazifalarda Sentence-BERT asosidagi vektorlashtirish samaraliroq natija berishi aniqlandi. Ushbu yondashuv katta hajmdagi matnli ma’lumotlarni semantik jihatdan tez va aniq tahlil qilish imkonini yaratadi.

Shuningdek, tadqiqot davomida klasterlash algoritmlarining samaradorligi vektorlashtirish sifatiga bevosita bog‘liq ekani kuzatildi. Shu sababli zamonaviy monitoring va kontent tahlili tizimlarida kontekstual embeddinglardan foydalanish muhim ahamiyat kasb etadi.

Ayniqsa o‘zbek tilidagi ijtimoiy tarmoq yozishmalarini qayta ishlashda transformer modellarining qo‘llanilishi semantik tahlil sifatini sezilarli darajada oshiradi.

Kelgusidagi tadqiqotlarda ushbu yondashuvni multimodal ma’lumotlar — matn va tasvirlarni birgalikda tahlil qilish orqali kengaytirish rejalashtirilmoqda. Bundan tashqari, HDBSCAN kabi zichlikka asoslangan klasterlash algoritmlarini qo‘llash, shuningdek mahalliy o‘zbek tilidagi korpuslarda maxsus o‘qitilgan transformer modellaridan foydalanish klasterlash aniqligini yanada oshirishi mumkin.

Shu bilan birga, real vaqt monitoring tizimlarida xavfli kontentni avtomatik aniqlash va risk darajasini baholash uchun semantik embeddinglar asosidagi yondashuvlarni chuqurlashtirish istiqbolli yo‘nalishlardan biri hisoblanadi.

Adabiyotlar

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3982–3992.
2. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 6000–6010.
4. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.
5. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (3rd ed.). Pearson.
6. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., et al. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*, 8440–8451.
8. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
9. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.
10. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
11. Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the NLP4SocialMedia Workshop*, 1–10.
12. Abdullayev, S., Mirzakhalilov, M., & Yusupov, M. (2023). BERTbek: A pretrained language model for Uzbek. *arXiv preprint arXiv:2306.00602*.
13. Qo’yliyeva, F. A., & Babomurodov, O. J. (2026). Detection of Risk Levels in Social Network Messages Using K-Means Clustering and Threshold-Based Analysis. *Digital Transformation and Artificial Intelligence*, 4(1), 45–53.
14. Qo’yliyeva, F. A. (2025). O‘zbek tilidagi toksik xabarlar uchun maxsus mini korpus yaratish va uning asosida klassifikatsiya modeli qurish. *Geoaxborot texnologiyalarini takomillashtirish masalalari: innovatsiyalar, barqaror rivojlanish va raqamli transformatsiya*, 112–118.