KLASTERLASH ASOSIDA IJTIMOIY TARMOQ YOZISHMALARINING XAVF DARAJASINI TAHLIL QILISH YONDASHUVI (K=4)

Babomurodov Ozod Jurayevich[1], Qo’yliyeva Feruzaxon Alisher qizi[2]

[1]Jizzax shaxridagi Qozon federal universiteti filiali ijrochi direktori, DSc, professor,
[2]Toshkent davlat agrar universiteti asissenti,
E-mail: qoyliyevaferuza@gmail.com

K E Y W O R D S

K-means, klasterlash, xavfli kontent, ijtimoiy tarmoq, matnni tahlil qilish, semantik guruhlanish, BERT

A B S T R A C T

Ushbu maqolada ijtimoiy tarmoqlarda, xususan Telegram platformasida tarqalgan matnli yozishmalarni xavf darajasiga ko‘ra avtomatik guruhlash masalasi ko‘rib chiqiladi. Tadqiqotning asosiy maqsadi K-means klasterlash algoritmining k = 4 parametri asosida ishlash samaradorligini baholash hamda turli hajmdagi datasetlarda klasterlash sifati va barqarorligini aniqlashdan iborat. Tadqiqot jarayonida 1000, 5000, 10000 va 20000 ta yozishmadan iborat to‘rtta turli hajmdagi ma’lumotlar to‘plami shakllantirildi va ular asosida klasterlar sifati, semantik chegaralar hamda ichki uyg‘unlik darajasi tahlil qilindi. Olingan natijalar kichik hajmdagi datasetlarda klasterlash natijalari yetarli darajada barqaror bo‘lmasligini, o‘rtacha hajmdagi ma’lumotlar bazasida esa klasterlar shakllanish jarayonida ekanini ko‘rsatdi. 10000 ta yozishmadan iborat datasetda k = 4 parametri bilan eng toza, aniq va barqaror semantik guruhlanish hosil bo‘lgani aniqlandi. 20000 ta yozishmadan iborat datasetda esa semantik variativlikning ortishi klasterlar barqarorligining pasayishiga olib keldi. Tadqiqot natijalari K-means algoritmi asosida olingan klasterlash natijalari keyingi bosqichda transformer modellarini (XLM-R, mBERT, BERTbek) o‘qitish uchun samarali belgilash bazasini yaratishda muhim ahamiyatga ega ekanini ko‘rsatadi. Ushbu yondashuv ijtimoiy tarmoqlarda xavfli kontentni avtomatik aniqlash va monitoring qilish tizimlarini takomillashtirishda amaliy ahamiyat kasb etadi.

1. Kirish

So‘nggi yillarda ijtimoiy tarmoqlar axborot almashinuvida eng asosiy kommunikatsiya maydonlaridan biriga aylandi. Ushbu platformalarda tarqaladigan matnli xabarlar qisqa vaqt ichida katta auditoriyaga yetib borishi bilan ajralib turadi. Biroq mazkur tezkorlik va ommaviylik axborot xavfsizligi nuqtai nazaridan jiddiy tahdidlarni yuzaga keltirmoqda. Xususan, ijtimoiy tarmoqlarda tahdidli ifodalar, zo‘ravonlikni targ‘ib qiluvchi chaqiriqlar, manipulyativ yoki yolg‘on ma’lumotlar, shuningdek nafrat uyg‘otuvchi va salbiy provokatsion mazmundagi yozishmalar keng tarqalmoqda.

Bunday xavfli kontent jamiyat barqarorligiga, shaxslararo munosabatlarga hamda axborot muhitining ishonchliligiga salbiy ta’sir ko‘rsatadi. Shu sababli matnli xabarlarni xavf darajasiga ko‘ra avtomatik tarzda aniqlash va guruhlash masalasi dolzarb ilmiy-amaliy vazifalardan biri hisoblanadi. Ayniqsa, katta hajmdagi yozishmalarni qo‘lda tahlil qilishning amalda imkonsizligi ushbu jarayonlarni avtomatlashtirish zaruratini yanada oshirmoqda.

Mazkur vazifani samarali hal etishda mashinaviy o‘rganish algoritmlaridan foydalanish muhim ahamiyatga ega. Ular orasida K-means klasterlash algoritmi soddaligi, hisoblash tezligi va turli turdagi matnli ma’lumotlarga moslashuvchanligi bilan ajralib turadi. K-means algoritmi yozishmalarni oldindan belgilangan klasterlar soniga ajratish orqali semantik jihatdan o‘xshash matnlarni bitta guruhga jamlash imkonini beradi.

Ushbu maqolada matnli yozishmalarni xavf darajasiga ko‘ra segmentatsiya qilishda K-means algoritmining *k = 4* parametri tanlandi. Mazkur parametr xavfsiz, neytral, o‘rta xavf va yuqori xavf kabi to‘rtta asosiy semantik guruhni shakllantirish imkonini beradi. Bunday guruhlash keyingi bosqichlarda yanada murakkab va aniq ishlovchi klassifikator modellarini o‘qitish учун ishonchli asos vazifasini bajaradi.

© O.J. Babomurodov, F.A. Qo’yliyeva

Shu tariqa, ijtimoiy tarmoqlardagi matnli yozishmalarni xavf darajasi bo‘yicha avtomatik guruhlash masalasi nafaqat ilmiy tadqiqot nuqtai nazaridan, balki amaliy axborot xavfsizligi tizimlarini rivojlantirishda ham muhim ahamiyat kasb etadi.

2. Tadqiqot metodologiyasi

Mazkur tadqiqot ijtimoiy tarmoqlarda tarqaladigan matnli yozishmalarni xavf darajasi bo‘yicha avtomatik guruhlash samaradorligini baholashga qaratilgan bo‘lib, tadqiqot jarayonida ma’lumotlarni yig‘ish, oldindan qayta ishlash, klasterlash va klaster sifatini tahlil qilish bosqichlari izchil amalga oshirildi.

2.1. Datasetlar

Tadqiqotda turli hajmdagi ma’lumotlar bazalarining klasterlash natijalariga ta’sirini aniqlash maqsadida to‘rtta alohida datasetdan foydalanildi. Jumladan, 1000, 5000, 10000 va 20000 ta yozishmadan iborat ma’lumotlar to‘plamlari shakllantirildi. Ushbu yondashuv K-means algoritmining barqarorligi va semantik guruhlash sifati ma’lumotlar hajmiga qanchalik bog‘liq ekanini aniqlash imkonini berdi.

Matnli yozishmalar Telegram ijtimoiy tarmog‘idagi ommaviy kanallardan yig‘ildi. Dataset tarkibi real axborot muhitini aks ettiruvchi xususiyatlarga ega bo‘lib, unda kirill va lotin yozuvlarining aralash qo‘llanilishi, sleng va qisqartmalar, emotsiyalar hamda turli darajadagi xavfli ifodalar mavjud. Bunday xilma-xillik tadqiqot natijalarining amaliy ahamiyatini oshirib, klasterlash algoritmining real sharoitlarda ishlash samaradorligini baholashga xizmat qildi.

Yig‘ilgan matnlar oldindan qayta ishlash bosqichidan o‘tkazilib, ortiqcha belgilar, takrorlanishlar va semantik jihatdan ahamiyatsiz elementlar tozalandi. Natijada tahlil uchun mos va bir xilda qayta ishlangan matnlar majmuasi hosil qilindi.

2.2. Klasterlash algoritmi

Tadqiqotda matnli yozishmalarni xavf darajasi bo‘yicha segmentatsiya qilish uchun K-means klasterlash algoritmi qo‘llanildi. Klasterlar soni $k = 4$ qilib belgilandi, bu esa yozishmalarni xavfsiz, neytral, o‘rta xavf va yuqori xavf toifalariga ajratish imkonini berdi. Ushbu tanlov xavf darajasini bosqichma-bosqich ifodalash va semantik tafovutlarni aniqroq ajratib ko‘rsatish maqsadida amalga oshirildi.

K-means algoritmi matnli yozishmalarni vektor ko‘rinishiga o‘tkazilgandan so‘ng qo‘llanildi. Klasterlash jarayonida har bir yozishma o‘ziga eng yaqin klaster markaziga biriktirildi va iterativ tarzda klaster markazlari yangilandi. Natijada semantik jihatdan o‘xshash mazmunga ega yozishmalar bitta klaster doirasida jamlandi.

Har bir klaster ichidagi yozishmalar alohida tahlil qilinib, ulardagi agressiv ifodalar, manipulyativ mazmun, tahdid va zo‘ravonlik belgilarining mavjudligi baholandi. Ushbu semantik tahlil klasterlarning mazmuniy mosligini aniqlash hamda ularni xavf darajasi bo‘yicha to‘g‘ri talqin qilish imkonini berdi.

2.3. Klaster sifatini baholash

Klasterlash natijalarining sifatini baholash jarayoni бир necha mezonlar asosida amalga oshirildi. Avvalo, klasterlar orasidagi semantik chegaralarning aniqligi tahlil qilindi. Har bir klaster ichidagi yozishmalarning mazmuniy uyg‘unligi va bir-biriga o‘xshashlik darajasi baholandi, shuningdek turli klasterlar orasidagi farqlanish darajasi o‘rganildi.

Bundan tashqari, klasterlarning barqarorligi turli hajmdagi datasetlarda taqqoslab ko‘rildi. Klasterlash natijalarining keyingi bosqichlarda transformer modellarini o‘qitish uchun qanchalik mos ekani alohida mezon sifatida qaraldi. Ya’ni, klasterlar tomonidan taqdim etilgan semantik guruhlar qo‘lda belgilash jarayonini soddalashtirishi va sifatli o‘quv ma’lumotlarini shakllantirishga xizmat qilishi baholandi.

Mazkur baholash mezonлари asosida K-means алгоритмининг $k = 4$ параметри turli hajmdagi ma’lumotlar bazalarida sinovdan o‘tkazildi va eng barqaror hamda semantik jihatdan toza natijalar qaysi hajmda kuzatilishi aniqlandi.

© O.J. Babomurodov, F.A. Qo’yliyeva

ISSN: 3030-3346

3. Natijalar va tahlil

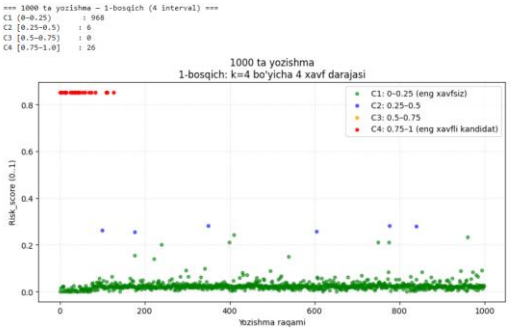
3.1. 1000 ta yozishma (k = 4). 1000 ta yozishmadan iborat kichik hajmdagi datasetda K-means algoritmining *k = 4* parametri bilan olingan klasterlash natijalari semantik jihatdan yetarli darajada barqaror bo‘lmadi. Ma’lumotlar hajmining kichikligi sababli klasterlar orasidagi chegaralar aniq shakllanmaganligi kuzatildi. Natijada semantik jihatdan bir-biriga yaqin bo‘lgan yozishmalar turli klasterlarga ajralib ketgan yoki aksincha, mazmunan farqli matnlar bitta klaster doirasida jamlangan.

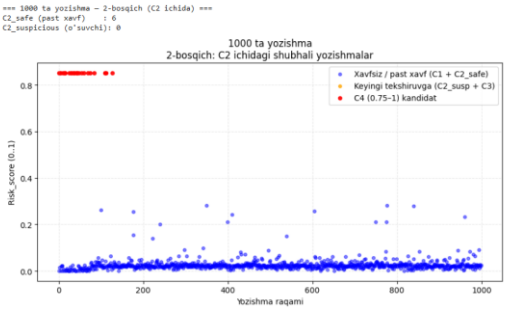
Xususan, “o‘rta xavf” va “yuqori xavf” toifalariga mansub yozishmalar o‘rtasida aniq semantik farqlanish hosil bo‘lmadi. Ushbu ikki guruh mazmun jihatidan bir-biriga juda yaqinlashib ketgan bo‘lib, klaster markazlari orasidagi masofa yetarlicha katta bo‘lmagan. Bu holat xavf darajasini bosqichma-bosqich aniqlash jarayonini murakkablashtirdi.

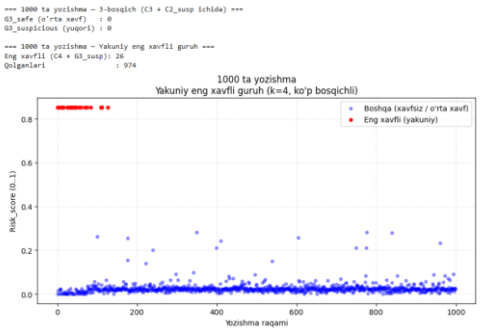
Shuningdek, ayrim aniq tahdidli yoki agressiv ifodalarga ega matnlarning noto‘g‘ri klasterlarga joylashtirilgani qayd etildi. Masalan, ochiq zo‘ravonlik yoki tahdid belgilariga ega yozishmalar ba’zi holatlarda “neytral” yoki “o‘rta xavf” guruhiga tushib qolgan. Bu esa klasterlash algoritmining kichik hajmdagi ma’lumotlar bilan ishlashda semantik signallarni to‘liq ajrata olmasligini ko‘rsatadi.

Klasterlash natijalarining umumiy barqarorligi past darajada bo‘lib, iteratsiyalar davomida klaster markazlarining sezilarli siljishi kuzatildi. Bu holat klasteratsiyaning notekis va ishonchsiz ekanini anglatadi hamda keyingi bosqichlarda ushbu natijalardan to‘g‘ridan-to‘g‘ri o‘quv ma’lumotlari sifatida foydalanishni cheklaydi.

Xulosa qilib aytganda, 1000 ta yozishmadan iborat dataset K-means algoritmining *k = 4* parametri bilan xavf darajasiga ko‘ra semantik jihatdan aniq va barqaror klasterlar hosil qilish uchun yetarli emas. Kam hajmdagi ma’lumotlar klasterlashda noaniqlikka olib kelib, klasterlar sifati va barqarorligini sezilarli darajada pasaytiradi.







3.2. 5000 ta yozishma (k = 4). Dataset hajmining 5000 ta yozishmagacha oshirilishi K-means algoritmining *k = 4* parametri bilan olingan klasterlash natijalarida sezilarli ijobiy o‘zgarishlarni keltirib chiqardi. Avvalgi kichik hajmli dataset bilan solishtirganda klasterlar orasidagi semantik chegaralar ancha ravshanlashgani kuzatildi. Natijada to‘rt guruh — xavfsiz, neytral, o‘rta xavf va yuqori xavf toifalari ichidagi mazmuniy farqlar aniqroq ifodalana boshladi.

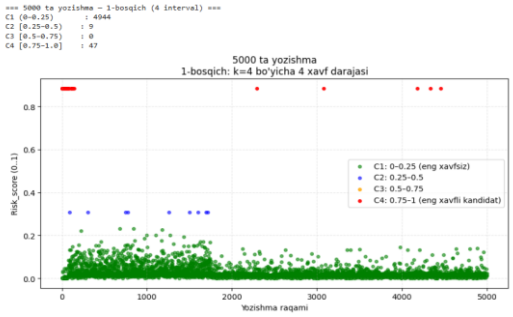
Mazkur hajmdagi ma’lumotlar bazasida yuqori xavfli xabarlar alohida klaster sifatida shakllana boshlagani alohida e’tiborga loyiqdir. Tahdid, zo‘ravonlik yoki agressiv ifodalar mavjud bo‘lgan yozishmalar ko‘pchilik holatlarda bitta klaster doirasida jamlanib, ushbu guruhning semantik yaxlitligini oshirdi. Bu esa xavfli kontentni aniqlashda K-means algoritmining samaradorligi ortib borayotganini ko‘rsatadi.

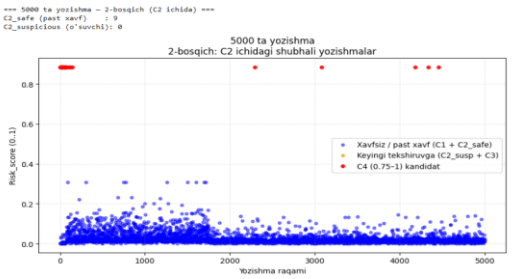
(cc) BY © O.J. Babomurodov, F.A. Qo’yliyeva

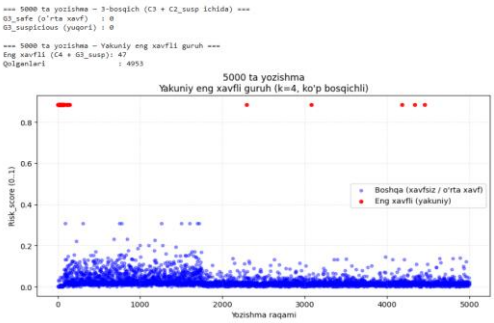
Shu bilan birga, neytral va past xavf darajasiga mansub yozishmalar o‘rtasida hali ham ma’lum darajada aralashuv saqlanib qoldi. Ba’zi holatlarda bahsli yoki kontekstga bog‘liq mazmunga ega matnlar neytral va o‘rta xavf klasterlari orasida joylashib, aniq chegaralanmadi. Bu holat xavf darajasi past bo‘lgan yozishmalar semantik jihatdan bir-biriga juda yaqin ekanini va ularni to‘liq ajratish uchun yanada katta hajmdagi ma’lumotlar zarurligini ko‘rsatadi.

Umuman olganda, 5000 ta yozishmadan iborat dataset K-means algoritmining *k = 4* parametri билан klasterlash jarayonida o‘tish bosqichi sifatida baholandi. Klasterlar shakllanish jarayonida bo‘lib, aniqlik darajasi o‘rtacha bo‘lsada, u keyingi bosqichlarda yanada barqaror va semantik jihatdan toza klasterlashga erishish imkoniyatini ko‘rsatib berdi.

Xulosa qilib aytganda, o‘rtacha hajmdagi ma’lumotlar bazasi klasterlash sifatini sezilarli darajada yaxshilaydi, biroq to‘liq barqaror va yuqori aniqlikka ega natijalarga erishish uchun yanada katta hajmdagi dataset talab etiladi.







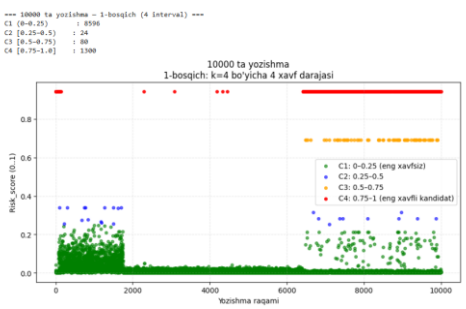
3.3. 10000 ta yozishma (k = 4). 10000 ta yozishmadan iborat datasetda K-means algoritmining *k = 4* parametri bilan amalga oshirilgan klasterlash jarayoni tadqiqot davomida eng yuqori sifat ko‘rsatkichlarini namoyon etdi. Mazkur hajmda klasterlar semantik jihatdan toza, aniq va barqaror shakllanib, xavf darajasi bo‘yicha guruhlar orasidagi chegaralar ravshan va izchil ko‘rinish oldi.

Xususan, “yuqori xavf” klasteri alohida e’tiborga loyiq bo‘lib, u mazmunan aniq ajralgan holda shakllandi. Ushbu klaster doirasida agressiv, nafratli, tahdidli va zo‘ravonlikka undovchi ifodalarga ega yozishmalar bir joyga jamlandi. Bu holat K-means algoritmining ushbu hajmdagi ma’lumotlar bazasida xavfli kontentni samarali aniqlay olishini yaqqol ko‘rsatdi.

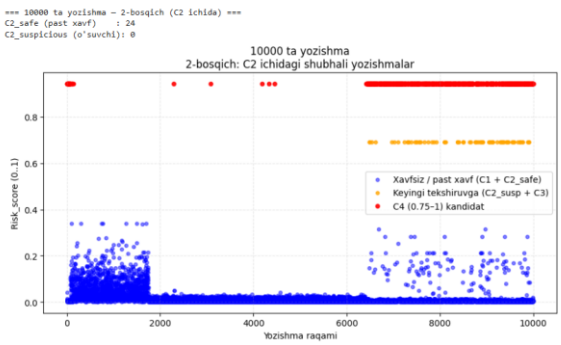
Shuningdek, “neytral” va “o‘rta xavf” toifalari o‘rtasidagi semantik farqlar ham to‘liq shakllandi. Avvalgi kichik hajmdagi datasetlarda kuzatilgan aralashuv holatlari bu bosqichda deyarli bartaraf etildi. Har bir klaster o‘ziga xos mazmuniy belgilar asosida aniq tavsiflanib, ichki uyg‘unlik darajasi yuqori bo‘ldi.

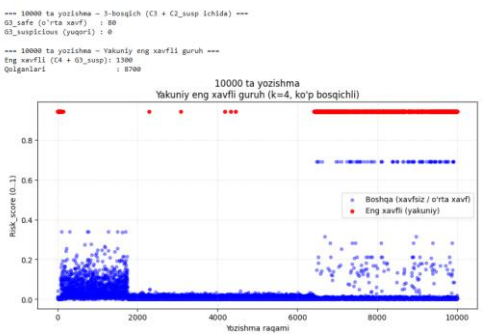
Mazkur klasterlash natijalari keyingi bosqichda Transformer asosidagi BERT modellarini o‘qitish uchun sifatli va ishonchli belgilash bazasini shakllantirish imkonini berdi. Klasterlar tomonidan taqdim etilgan semantik tozalik va barqarorlik qo‘lda belgilash jarayonini sezilarli darajada soddalashtirib, o‘quv ma’lumotlarining sifatini oshirdi.

Xulosa qilib aytganda, 10000 ta yozishmadan iborat dataset K-means algoritmining *k = 4* parametri bilan ishlashda eng optimal variant sifatida baholandi. Ushbu hajm aniq klasterlar hosil bo‘lishini, maksimal barqarorlikni va keyingi klassifikator modellarini o‘qitish uchun eng qulay sharoitni ta’minladi.



(cc) BY © O.J. Babomurodov, F.A. Qo’yliyeva



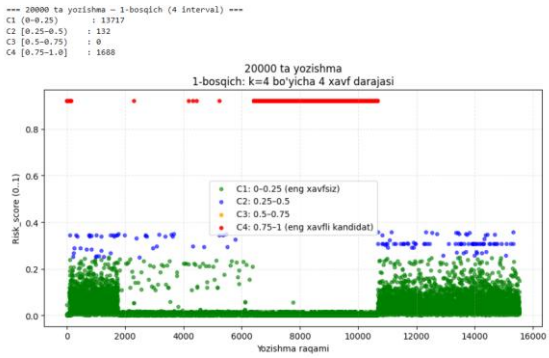


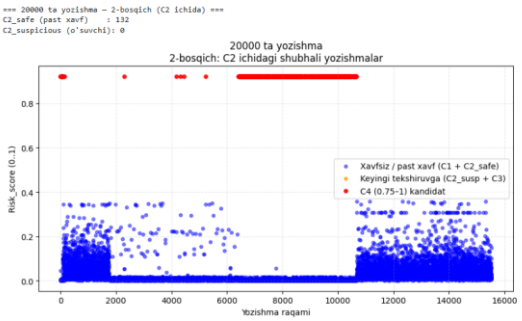
3.4. 20000 ta yozishma (k = 4). 20000 ta yozishmadan iborat datasetda K-means algoritmining *k = 4* parametri bilan amalga oshirilgan klasterlash jarayoni avvalgi bosqichlar bilan solishtirilganda ayrim murakkabliklarni yuzaga keltirdi. Ma’lumotlar hajmining keskin ortishi natijasida semantik guruhlar sezilarli darajada kengayib, klasterlar ichida qo‘shimcha ichki bo‘linishlar paydo bo‘lgani kuzatildi. Bu holat klasterlar o‘rtasidagi aniq chegaralarning saqlanishini qiyinlashtirdi.

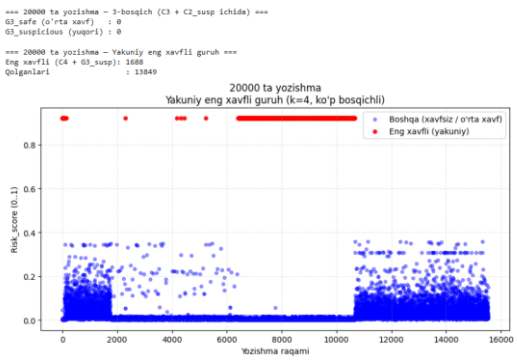
Shuningdek, yuqori xavfli ifodalarga ega yozishmalar tarkibida mazmuniy variativlikning ortishi qayd etildi. Turli shakldagi tahdid, agressiya va zo‘ravonlik ifodalari klaster markazlarining bir-biridan uzoqlashishiga olib keldi, bu esa umumiy klasterlash sifatiga salbiy ta’sir ko‘rsatdi. Natijada klasterlar ichidagi semantik yaxlitlik avvalgi optimal bosqichga nisbatan pasaydi.

Umuman olganda, klasterlash natijalarining umumiy barqarorligi 20000 ta yozishmadan iborat datasetda kamaygani aniqlandi. Ma’lumotlar hajmining ortishi semantik aniqlikni avtomatik ravishda oshirmasligi, aksincha, ortiqcha variativlik tufayli klasterlash jarayonini murakkablashtirishi mumkinligi ushbu bosqichda yaqqol namoyon bo‘ldi.

Xulosa qilib aytganda, juda katta hajmdagi ma’lumotlar bazasi K-means algoritmining *k = 4* parametri bilan ishlashda ortiqcha semantik variativlikni yuzaga keltirib, klasterlar barqarorligining pasayishiga olib keladi. Shu sababli, xavf darajasi bo‘yicha aniq va barqaror semantik guruhlash uchun 20000 ta yozishmadan iborat dataset optimal variant sifatida baholanmadi.







4. Muhokama

O‘tkazilgan tajribalar K-means algoritmining k = 4 parametri bilan ishlash samaradorligi dataset hajmi va undagi semantik xilma-xillikka bevosita bog‘liq ekanini ko‘rsatdi. Kichik hajmdagi, ya’ni 1000 ta yozishmadan iborat datasetda klasterlash natijalari noaniq bo‘lib, semantik jihatdan o‘xshash yozishmalar turli klasterlarga ajralib ketdi va barqarorlik past bo‘ldi.

(cc) BY © O.J. Babomurodov, F.A. Qo’yliyeva

5000 ta yozishmadan iborat datasetda klasterlash sifati yaxshilangan bo‘lsa-da, ayrim guruhlar, xususan “neytral” va “o‘rta xavf” klasterlari o‘rtasida semantik noaniqliklar saqlanib qoldi. Bu holat modelning o‘rtacha aniqlik darajasida ishlayotganini ko‘rsatdi.

10000 ta yozishmadan iborat datasetda k = 4 parametri eng yuqori samaradorlikni namoyon etdi. Klasterlar semantik jihatdan toza, aniq va barqaror shakllanib, ayniqsa “yuqori xavf” klasteri ravshan ajralib chiqdi. Natijada ushbu hajm keyingi Transformer modellarini o‘qitish uchun eng maqbul belgilash bazasi sifatida tanlandi.

20000 ta yozishmadan iborat datasetda esa semantik variativlikning ortishi klasterlash barqarorligining pasayishiga olib keldi. Bu holat ma’lumotlar hajmini oshirish har doim ham klasterlash sifatini yaxshilamasligini ko‘rsatadi.

Xulosa

O‘tkazilgan tadqiqot natijalari ijtimoiy tarmoqlarda tarqalgan matnli yozishmalarni xavf darajasi bo‘yicha avtomatik guruhlashda K-means algoritmining k = 4 parametri samaradorligi ma’lumotlar bazasi hajmiga bevosita bog‘liq ekanini ko‘rsatdi. Kichik hajmdagi, ya’ni 1000 ta yozishmadan iborat datasetda klasterlash natijalari noaniq bo‘lib, barqaror semantik chegaralar hosil bo‘lmadi. 5000 ta yozishmadan iborat datasetda klasterlash sifati yaxshilangan bo‘lsa-da, ayrim guruhlar o‘rtasida semantik noaniqliklar saqlanib qoldi.

10000 ta yozishmadan iborat datasetda k = 4 parametri eng barqaror va aniq klasterlash natijalarini taqdim etdi. Ushbu hajmda klasterlar semantik jihatdan toza va ravshan shakllanib, xavf darajasi bo‘yicha guruhlar o‘rtasida aniq chegaralar kuzatildi. 20000 ta yozishmadan iborat datasetda esa semantik variativlikning ortishi klasterlash barqarorligining pasayishiga olib keldi.

Shu sababli, ijtimoiy tarmoqlardagi yozishmalarni xavf darajasi bo‘yicha avtomatik belgilash jarayonida K-means algoritmi uchun k = 4 parametri bilan 10000 ta yozishmadan iborat dataset eng optimal hajm sifatida aniqlandi. Ushbu natijalar keyingi bosqichda Transformer modellarini o‘qitish uchun samarali belgilash bazasini shakllantirishda muhim amaliy ahamiyatga ega.

Adabiyotlar ro‘yxati

1. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. — Berkeley, 1967. — P. 281–297.
2. Aggarwal C. C., Zhai C. Mining Text Data. — New York: Springer, 2012. — 543 p.
3. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. — 3rd ed. — San Francisco: Morgan Kaufmann, 2011. — 703 p.
4. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // arXiv preprint. — 2013. — arXiv:1301.3781.
5. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT. — Minneapolis, 2019. — P. 4171–4186.
6. Conneau A., Khandelwal K., et al. Unsupervised cross-lingual representation learning at scale // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). — Seattle, 2020. — P. 8440–8451.
7. Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow. — Birmingham: Packt Publishing, 2019. — 622 p.
8. Srivastava A., Sahami M. Text Mining: Classification, Clustering, and Applications. — Boca Raton: Chapman & Hall/CRC, 2009. — 344 p.
9. Uzbek Language BERT (BERTbek) Model Description. HuggingFace Model Card. — 2023. — URL: https://huggingface.co (murojaat sanasi: 2025.01.15).

(cc) BY © O.J. Babomurodov, F.A. Qo’yliyeva