

IJTIMOY TARMOQ MATNLI XUJJATLARINI TAQSIMLASH YONDASHUVLARIDAN OLINGAN NATIJALAR SOLISHTIRMASI

Babomurodov Ozod Jurayevich

DSc, professor, Jizzax shaxridagi Qozon federal universiteti filiali ijrochi direktori.

babomurodovozod@gmail.com

Qo'ziliyeva Feruzaxon Alisher qizi

Toshkent davlat agrar universiteti asissenti.

qoyliyevaferuza@gmail.com

<https://doi.org/10.5281/zenodo.18762973>

Annotatsiya. Ushbu maqolada ijtimoiy tarmoqlardagi matnli hujjatlarni tahlil qilish va tasniflash bo'yicha mavjud yondashuvlar natijalari solishtirma tahlil qilinadi. Tadqiqotlarda VKontakte, Facebook, Twitter va Instagram platformalarida qo'llanilgan klasterlash, tasniflash va sentiment tahlil usullari o'rganildi. K-Means, SVM, Naive Bayes, KNN, LSTM, OLS va ansambl modellari asosida olingan natijalar aniqlik darajasi bo'yicha taqqoslandi. Shuningdek, TF-IDF, Bag-of-Words, N-gram, POS-tagging va stemming kabi matnni qayta ishlash usullari samaradorligi ko'rib chiqildi. Tadqiqot natijalari shuni ko'rsatadiki, katta hajmdagi ma'lumotlar va gibrid modellardan foydalanish aniqlikni sezilarli oshiradi hamda ijtimoiy media monitoringi va xavfli kontentni erta aniqlash tizimlarini yaratishda muhim ahamiyat kasb etadi.

Kalit so'zlar: ijtimoiy tarmoqlar, klasterlash, K-Means, SVM, sentiment tahlili, TF-IDF, ansambl model, matn tasnifi.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ ПОДХОДОВ К РАСПРЕДЕЛЕНИЮ ТЕКСТОВЫХ ДОКУМЕНТОВ СОЦИАЛЬНЫХ СЕТЕЙ

Аннотация. В статье представлен сравнительный анализ подходов к распределению и классификации текстовых документов социальных сетей. Рассмотрены методы кластеризации, классификации и анализа тональности, применяемые в социальных сетях VKontakte, Facebook, Twitter и Instagram. Проведено сравнение результатов, полученных с использованием алгоритмов K-Means, SVM, Naive Bayes, KNN, LSTM, OLS и ансамблевых моделей. Также проанализирована эффективность методов предварительной обработки текста, таких как TF-IDF, Bag-of-Words, N-gram, POS-tagging и stemming. Результаты исследования показывают, что использование больших массивов данных и гибридных моделей существенно повышает точность анализа и имеет практическое значение для систем мониторинга социальных медиа и раннего выявления потенциально опасного контента.

Ключевые слова: социальные сети, кластеризация, K-Means, SVM, анализ тональности, TF-IDF, ансамблевая модель, классификация текста.

COMPARATIVE ANALYSIS OF RESULTS OBTAINED FROM APPROACHES TO SOCIAL NETWORK TEXT DOCUMENT CLASSIFICATION

Annotation. This paper presents a comparative analysis of approaches for classifying and distributing textual documents in social networks. Clustering, classification, and sentiment analysis methods applied to VKontakte, Facebook, Twitter, and Instagram platforms are examined. The performance of algorithms such as K-Means, SVM, Naive Bayes, KNN, LSTM, OLS, and ensemble models is compared. Additionally, the effectiveness of text preprocessing techniques including TF-IDF, Bag-of-Words, N-gram, POS-tagging, and stemming is analyzed.

The results demonstrate that the use of large-scale datasets and hybrid models significantly improves analytical accuracy and plays an important role in developing social media monitoring systems and early detection mechanisms for potentially harmful content.

Keywords: social networks, clustering, K-Means, SVM, sentiment analysis, TF-IDF, ensemble model, text classification.

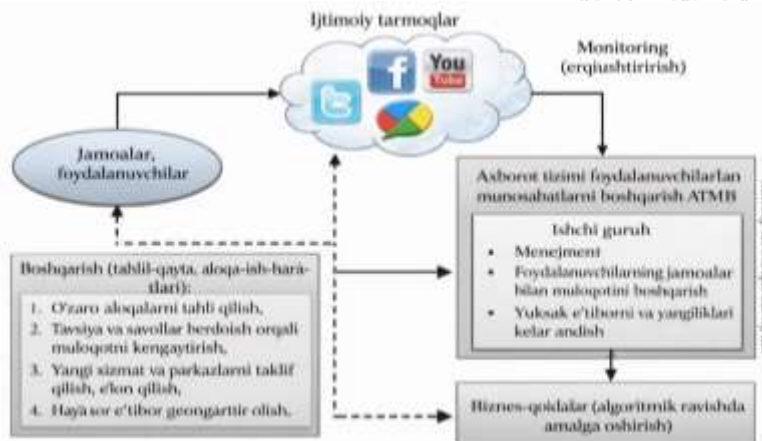
KIRISH

So‘nggi yillarda ijtimoiy tarmoqlar global axborot almashinuvi va ijtimoiy kommunikatsiyaning asosiy platformasiga aylandi. Foydalanuvchilar tomonidan yaratilayotgan katta hajmdagi matnli ma‘lumotlar jamiyatdagi ijtimoiy, iqtisodiy va siyosiy jarayonlarni o‘rganishda muhim manba hisoblanadi. Shu bilan birga, noto‘g‘ri, manipulyativ yoki xavfli kontentning tez tarqalishi ijtimoiy barqarorlikka salbiy ta‘sir ko‘rsatishi mumkin.

Ijtimoiy media matnlarini avtomatik tahlil qilish, klasterlash va tasniflash usullari mazkur muammolarni hal etishda muhim vosita hisoblanadi. Zamonaviy tadqiqotlarda mashinaviy o‘rganish va sun‘iy intellekt algoritmlaridan keng foydalanilmoqda. Ushbu maqolada turli ijtimoiy tarmoqlarda qo‘llanilgan yondashuvlar solishtirilib, ularning samaradorligi baholanadi.

ASOSIY QISM

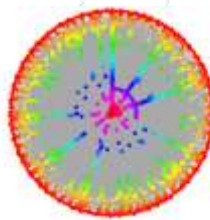
Ijtimoiy tarmoqlarni tahlil qilish yondashuvlari asosan sotsiologik so‘rov usullari va statistik tahlilga bo‘linadi. Ijtimoiy tarmoqlar virtual muhitda guruh va jamoalarning o‘zaro aloqasini ta‘minlab, xabarlar va izohlar mazmunini tahlil qilish orqali jamiyatdagi dolzarb muammolarni o‘rganish imkonini beradi [1](1-rasm).



1-rasm. Ijtimoiy tarmoqlarda foydalanuvchilar bilan munoosabatlarni boshqarish jarayoni sxemasi

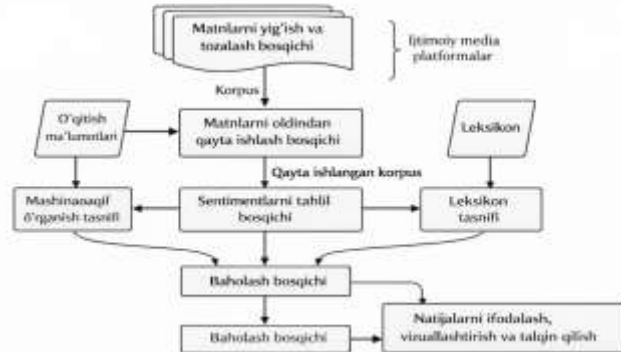
VKontakte ijtimoiy tarmog‘ida yozilgan kommentariyalarni tahlil qilish masalasi o‘rganilgan bo‘lib, kommentariyalarni ajratish uchun gibrid klasterlash usuli qo‘llanilgan [2].

Ma‘lumotlar ijtimoiy tarmoqdan yig‘ilib, tahlil va hisob-kitoblar R dasturlash muhitida amalga oshirilgan (2-rasm).

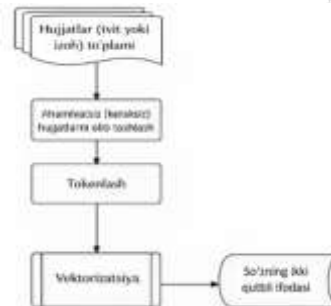


2-rasm. Gibrid klasterlash usuli natijasining ko‘rinishi.

Postlar ijtimoiy-iqtisodiy mavzular bo'yicha VK API orqali yig'ilib, dastlabki qidiruv natijalari mavzuga to'liq mos kelmagani sababli, ma'lumotlarni tozalash uchun matn tasnifi qo'llanilganda yahshi natija qayd etilganligini ko'rish mumkin. Bu yerda tasniflash Python muhitida SVM klassifikatori yordamida amalga oshirilgan. Postlar ijobiy, salbiy va mavzuga aloqador bo'lmagan sinflarga ajratilib, jami 6200 dan ortiq postlar to'plangan [1,3](3,4-rasm).



3-rasm. Taklif qilinayotgan frameworkning arxitekturası.



4-rasm. Ma'lumotlarga dastlabki ishlov berish jarayoni

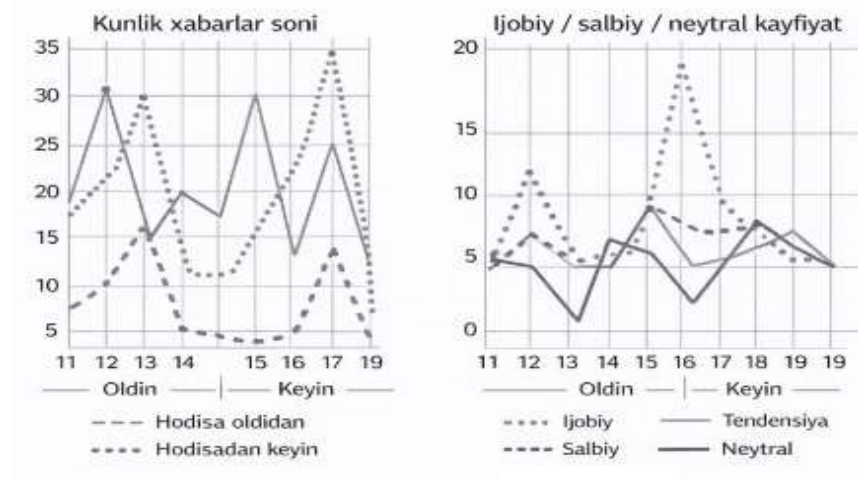
SVM modeli samarali ishlashi uchun matnlar oldindan qayta ishlangan: so'zlar pymorphy2 yordamida ildiz shakliga keltirilib, tinish belgilari va jargonlar tozalangan.

Xususiyatlar tf-idf asosida ajratilib, takrorlanuvchi so'zlarning ta'siri kamaytirilgan.

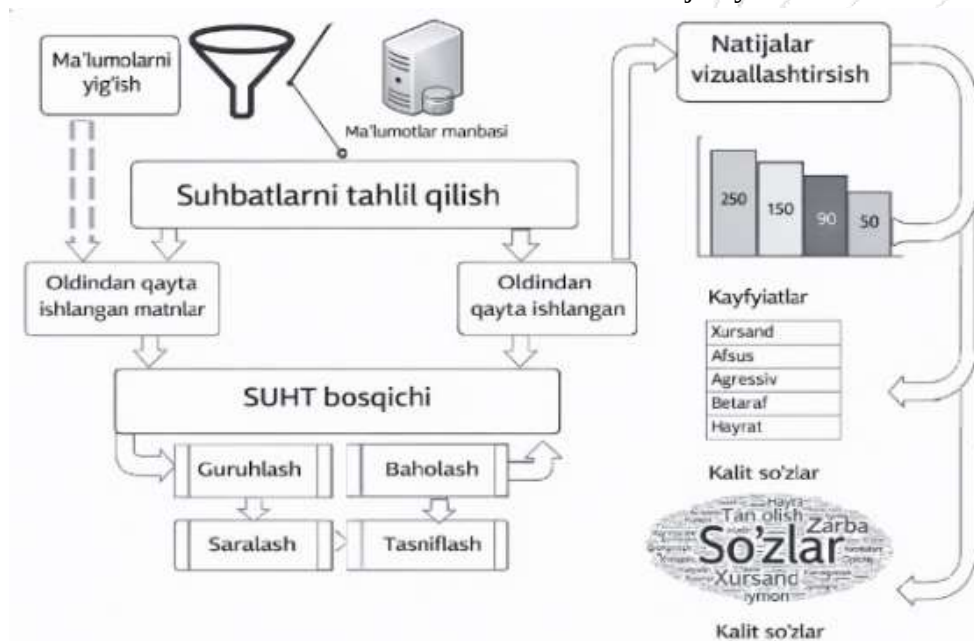
Klassifikatori ishlatish orqali 84-86% natija qayd etilganligini ko'rish mumkin [1,4].

VKontakte ijtimoiy tarmog'idagi foydalanuvchi sahifalarini tahlil qilishda ma'lumotlar SNAP tayyor ombori, VK API va Python yordamida olingan. Tadqiqotda 40 mingga yaqin foydalanuvchi sahifasi va ularning shaxsiy ma'lumotlari yig'ilgan. Klasterlash uchun Kohonen o'z-o'zini tashkil qiluvchi tarmoqlaridan foydalanilgan. Klasterlashdan maqsad foydalanuvchilarning bank xizmatlari va kredit olish imkoniyatlarini baholash bo'lgan [5].

[6]-ishda Strukturali bo'lmagan matnli ma'lumotlarni vektorlash va kontekstli munosabatlarni tahlil qilishda yana shu VKontakte ijtimoiy tarmog'idagi yangiliklar va ularga yozilgan izohlar tanlangan bo'lib, Python muhitida ma'lumotlarni yig'ish, filtrlash va kontekstni tahlil qilish modullaridan iborat dasturiy ta'minoti yordamida 2018-yil avgust-sentabr oylarida 11 ming post va 190 ming izoh yig'ilgan va ishlov amalga oshirilgan, bu holatda ham ijobiy natija olingan (5,6-rasm).



5-rasm. Xabarlarni tahlillash jarayoni



6-rasm. 11-19.08.18 vaqt oralig'ida 77580ta izoh, 3358ta postdan olingan natija

VKontakte ijtimoiy tarmog'idagi postlar, izohlar va sahifa ma'lumotlari kalit so'zlar asosida daraxtli klasterlash usuli bilan mavzularga ajratilgan. Tahlil sport, IT-texnologiyalar, musiqa, biznes va boshqa mavzulardagi izohlarni qamrab olgan. Tadqiqotda aniqlik 0,00018%ga teng

Ijtimoiy tarmoqlarda samarali reklama olib borish uchun ma'lumotlarni intellektual tahlil qilish masalasini statistik, vizual va retrospektiv tahlil usullaridan foydalanilgan. Kontentga auditoriyaning munosabatini aniqlash uchun hissiyotlar lug'ati hamda Bayes va KNN klassifikatorlari qo'llanilgan. Obunachilar portretini shakllantirishda k-means klasterlash usulidan, foydalanuvchilar qiziqishlarini aniqlashda esa Apriori assotsiativ qoidalaridan foydalanilgan. Natijalarga ko'ra, eng yaxshi ko'rsatkich tayanch vektorlar usulida qayd etilgan bo'lib, uning aniqligi 82%ni tashkil etgan. Eng past natija esa his-tuyg'ular lug'atiga asoslangan usulda kuzatilib, 71%ga teng bo'lgan.

Facebook. 2004-yil Facebook ijtimoiy tarmog'i paydo bo'lgan, o'sha yildan boshlab insonlarning katta qismining ijtimoiy va ijtimoiy-psixologik tajribasi orta boshlagan [4].

Odamlar Facebookda uchta xatti-harakat orqali suhbatlashadilar: yoqtirish, izoh bildirish va almashish(ulashishlar) [8].

Tadqiqot tahlilida brendX, brendY va brendZ kompaniyalarining sovun, krem va dezodorant kabi mahsulotlariga oid foydalanuvchi izohlari o'rganilgan. Kommentariyalarni tahlil qilish jarayoni matnlarni yig'ish va tozalash, dastlabki ishlov berish, tahlil va baholash bosqichlarida amalga oshirilgan.

Matnlarga dastlabki ishlov berishda bag-of-words usuli qo'llanilib, to'xtash so'zlar va raqamlar olib tashlangan. Xususiyatlarni ajratish uchun tf-idf yondashuvidan foydalanilgan.

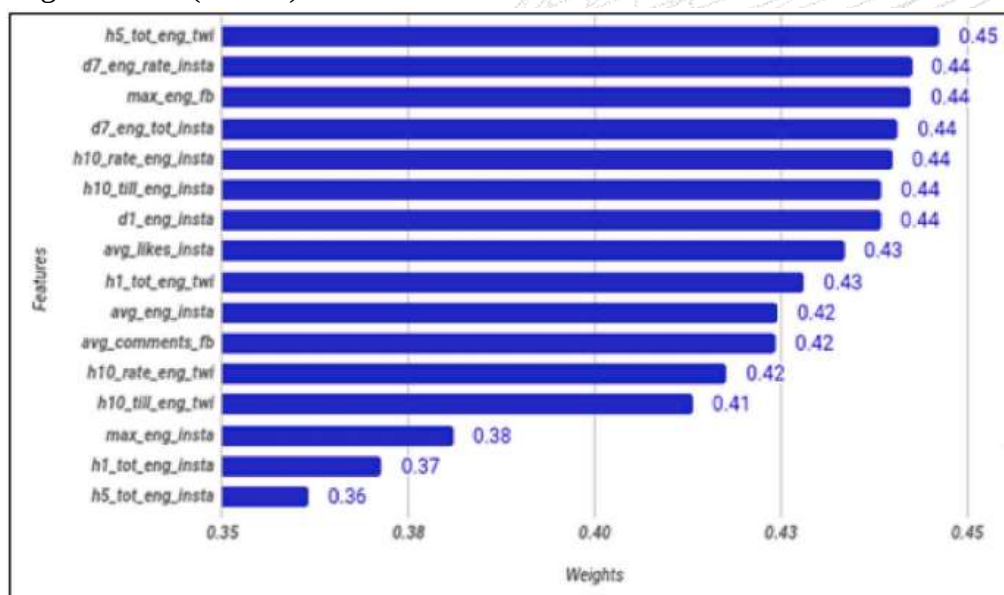
Kommentariyalarni tahlil qilishda Naive Bayes usuli qo'llanilgan. Matnni tahlillash uchun Naive baes klasterlash usulidan foydalanilgan.

Ma'lumotlarni olish jarayoni Twitter, Instagram va Facebook ijtimoiy tarmoqlaridagi mashhur sportchilar va aktyorlar sahifalariga yozilgan izohlarni yig'ishdan boshlangan. Turli platformalardan API orqali ma'lumotlar olinib, JSON formatda saqlangan. Jami 1000 ga yaqin mashhur shaxslar haqidagi ma'lumotlar to'plangan.

Ma'lumotlarni tahlil qilishda obunachilar soni, postlardagi faollik, yoqtirishlar, izohlar va ulashishlar kabi ko'rsatkichlar hisobga olingan. Turli platformalardan olingan ma'lumotlar moslashtirilib, yagona identifikatorlar asosida birlashtirilgan. Ma'lumotlar turli vaqt oraliqlarida yig'ilib, 90 kun davomida to'plangan ijtimoiy media atributlari asosida tahlil qilingan.

Tahlil jarayonida auditoriya qamrovi, jalb qilish darajasi, soatlik va kunlik faollik, shuningdek, auditoriya kayfiyati va post joylashtirish tezligi kabi ko'rsatkichlar aniqlangan.

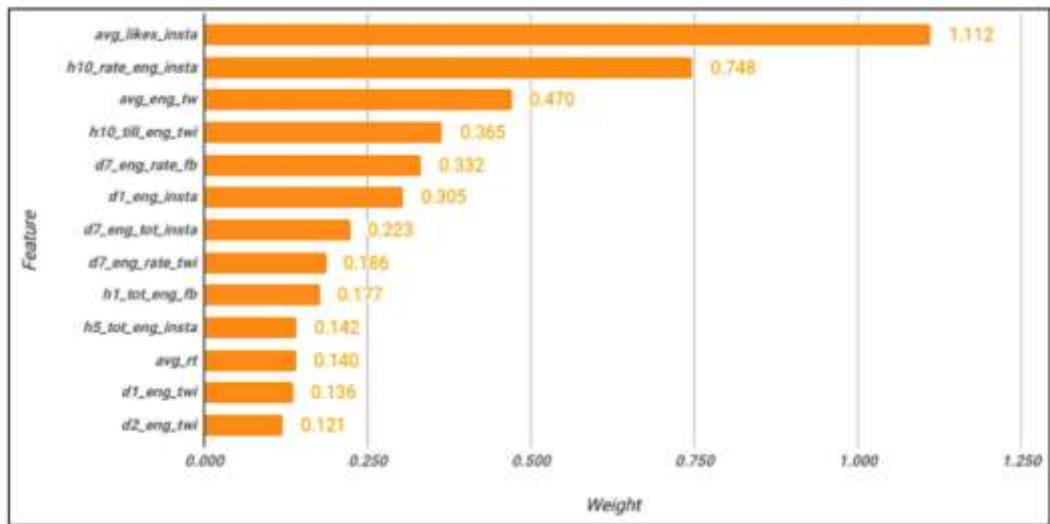
KNN modeli eng oddiy va keng qo'llaniladigan algoritmlardan biri bo'lib, o'xshashlikni masofa funksiyalari asosida baholaydi. U parametrik bo'lmagan usul sifatida statistik tahlil va tasvirlarni aniqlashda qo'llaniladi. Regressiyada KNN eng yaqin qo'shnilarni topish orqali bashoratni amalga oshiradi (7-rasm).



7-rasm. KNN usuli yordamida olingan ahamiyatli belgilar

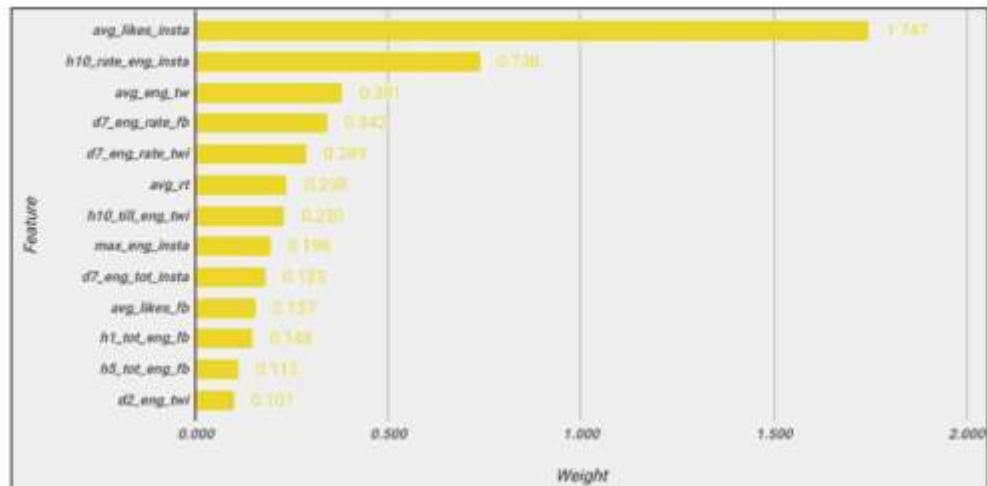
Lasso modeli parametrlarni qisqartirish va o'zgaruvchilarni avtomatik tanlash uchun qo'llaniladi. U ayniqsa parametrlar soni ko'p bo'lgan hollarda samarali bo'lib, ahamiyatsiz o'zgaruvchilarni nolga tenglashtirish orqali modeldan chiqarib yuboradi. Bu hisob-kitoblarni kamaytirish va vaqt sarfini qisqartirishga yordam beradi.

SVM usuli klassifikatsiya va regressiyada qo'llaniladi. SVR model epsilon-insensitive yo'qotish funksiyasi orqali barqaror natija beradi. Model kam o'quv nuqtalari bilan samarali ishlab, chiziqli yadro va grid search yordamida optimal parametrlar tanlangan (8-rasm).



8-rasm. SVR usuli yordamida aniqlangan ahamiyatli belgilar

OLS(Ordinary Least Squares)-model kuzatilgan va bashorat qilingan qiymatlar o'rtasidagi kvadratik farqlarning yig'indisi minimallashtiriladigan tarzda tanlangan. Modeli bir nechta taxminlarga asoslanadi (9-rasm).



9-rasm. OLS usuli yordamida aniqlangan ahamiyatli belgilar

Modellarni qo'llab olingan xulosa shundan iboratki, ushbu modellar asosida marketologlarga ansambl modeli taqdim etiladi. Ushbu model ijtimoiy ta'sir indeksini yaratishning eng yuqori aniqligini ta'minlay oladi - 93,7%. Marketologlar qaysi ijtimoiy media platformasi ma'lum bir mashhur kishini targ'ib qilish uchun eng yaxshi ekanligini aniqlashlari mumkin.

Instagram- Instagram ijtimoiy tarmog'i xalqaro internet muxiti bo'lib, unda foydalanuvchi tomonidan kiritilgan rasm yoki videolarni almashish va o'z akkaunti orqali ularni tarqatish uchun mo'ljallangan platforma[8]. [10]-tadqiqotda Oman moda sanoatidagi kichik biznes va chakana sotuvchilarning Instagram sahifalari tahlil qilingan. 12.12.2012 dan 07.10.2020 yilgacha bo'lgan davrda 530 366 ta xabar to'plangan va faol bo'lmagan akkauntlar chiqarib tashlangan. Klasterlar sonini tanlash uchun Elbow curve usuli qo'llanilib, 2-4 ta klaster oralig'i maqbul deb topilgan. Keyin postlar k-means algoritmi bilan 3 guruhga ajratilgan: 1-guruhda 527 921 ta post, 2-guruhda 62 ta post, 3-guruhda 2 380 ta post. Klasterlashda post turi, video ko'rishlar, layklar, sharhlar, sana (kun-oy-yil) va ishtirok etish darajasi hisobga olingan.

Natijada 527 921 ta postdan 401 731 tasi to'g'ri klasterlanib, 75.75% aniqlikka erishilgan.

Twitter – Twitter ijtimoiy tarmog'i qisqa xabarlar, g'oyalar va yeslatmalar bilan almashish xizmati. Ushbu keng ko'lamli Internet-loyiha ma'lumot manbai hisoblanadi[8]. [11]-ishda Twitter ijtimoiy tarmog'idan ma'lumot olish va tahlil qilish usuli taklif etilgan. IP-blokirovkani oldini olish uchun har bir so'rov uchun alohida akkauntlar qo'llanilgan va freymvork ko'p oqimli yuklashni qo'llab-quvvatlagan. Freymvork Twitter (1 soatda 500 profil), Facebook (1 soatda 3000 profil) va Hunch (1 soatda 100 profil) tarmoqlarida sinovdan o'tkazilgan.

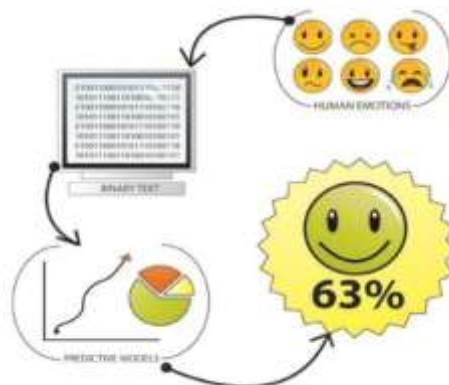
O'qitish bosqichida onlayn passiv-agressiv algoritmi asosida tasniflash modeli qurilgan.

Klasterlashda xabar matnlari va foydalanuvchi profili ma'lumotlaridan foydalanilgan.

Tajriba ingliz tilidagi Twitter foydalanuvchilari bo'yicha olib borilib, jins, yosh (20–40, 40+), oilaviy holat, siyosiy va diniy qarashlar asosida klasterlash amalga oshirilgan, aniqlik darajasi 74-89%gachani tashkil etgan.

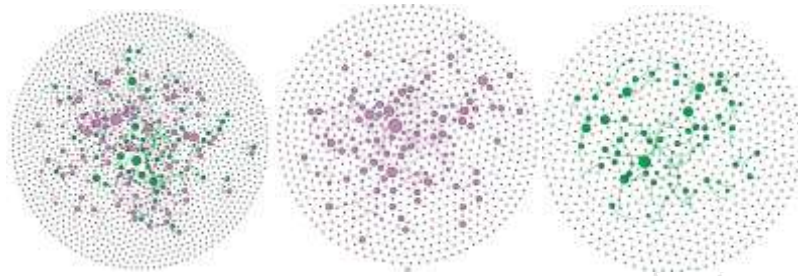
VKontakte, Twitter va Instagram ijtimoiy tarmoqlaridagi foydalanuvchilarni tahlil qilish va bir shaxsga tegishli profillarni aniqlash har bir guruhda har bir ijtimoiy tarmoqdan bittadan ortiq profil bo'lmasligi va barcha proffillar bir shaxsga tegishli bo'lishi shart qilib olingan holda Apache Spark va Spark Streaming hamda RabbitMQ xabar brokeri yordamida amalga oshirilgan [12]. Ma'lumotlar network_id va user_id belgilari asosida qayta ishlanib, RDD shakllantirilgan va turli ijtimoiy tarmoqlardagi proffillar o'zaro bog'langan. Olingan natijalar MongoDB NoSQL ma'lumotlar bazasida saqlangan. Tizimning qayta ishlash tezligi sekundiga taxminan 120–130 ta profilni tashkil qilgan.

2018-yilgi futbol bo'yicha Jahon chempionati davrida Twitter ijtimoiy tarmog'ida arab tilida yozilgan tvitlar tahlil qilinganda, ma'lumotlar o'nta turli hashtag orqali tasodifiy yig'ilib, keraksiz ma'lumotlar olib tashlangan va faqat matn hamda emojilar saqlab qolingan [12] va his-tuyg'ular tahlili amalga oshirilgan (10-rasm).



10-rasm. Ijtimoiy tarmoqlardan his-tuyg'ularni tahlil qilish

LSTM(Long short-term memory Uzoq qisqa muddatli xotira) modeli. Xissiyotlatni tahlil qilish uchun sun'iy neyron tarmoqlarining LSTM modelidan foydalanilgan. Garchi LSTM modeli kiritilgan kontekst ketma-ketligini qayta ishlashda yaxshi ishlasada, modelning ba'zi kamchiliklari ham bor, u kiritish yo'nalishini hisobga olmaydi va matn ketma-ketligi kontekstini faqat bitta yo'nalishda qayta ishlay oladi (11-rasm).

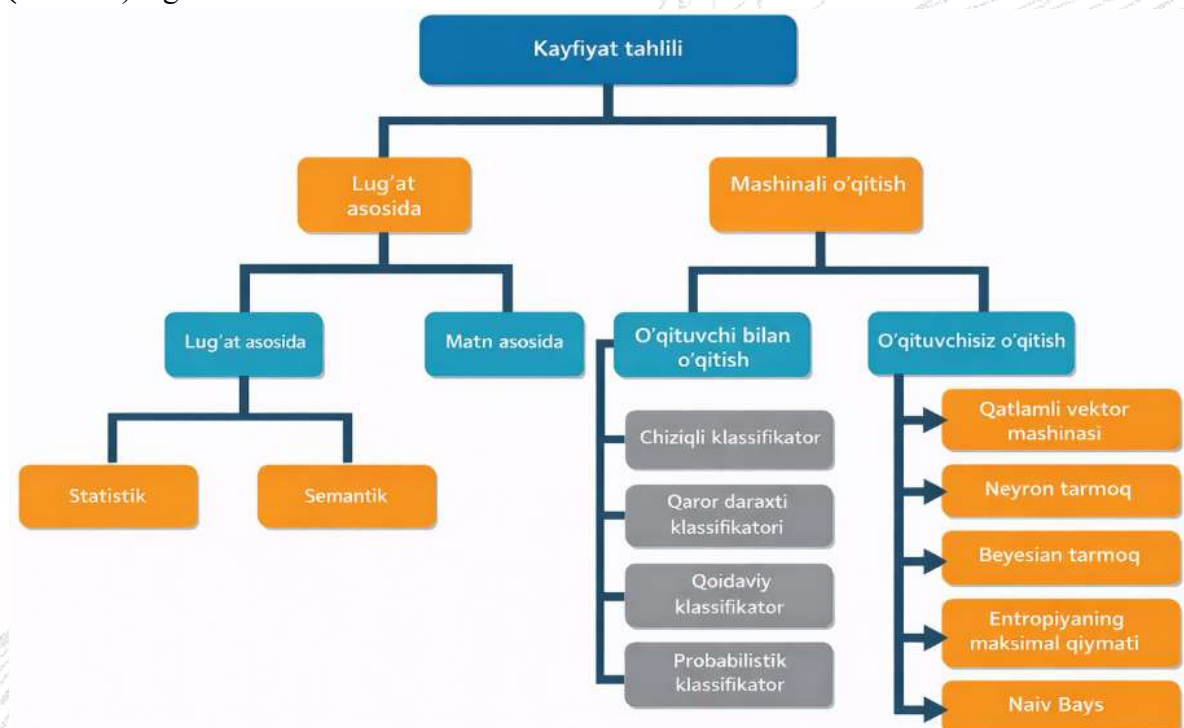


11-rasm. Ma'lumotlarni to'g'ri va yolg'onlarni ajratib olish

Bu yerda berilgan postga qo'yilgan smayliklarga qarab uni to'g'ri yoki yolg'on ma'lumotligini aniqlashdagi xatolik soni 0,01%ga teng, postga yozilgan izohlarga qarab esa xatolik 0.05%ni tashkil etgan.

Twitterda yozilgan xabarlarni xis-tuyg'ularini aniqlashda smayliklarlar orqali uzatilayotgan xabarga berilayotgan xis-tuyg'ularni aniqlash oson, biroq sharxga qarab, yaxshi yoki yomon fikrligini aniqlash masala yechimi aniqligini oshiradi [12]. Bu yerda amal qayta ishlashdan boshlangan bo'lib, N-gram, POS-tagging, stemming, stop-word, conjunction va negation handling usullari qo'llanilgan. Fikrlar yaxshi, yomon va neytral toifalarga ajratilgan. Klassifikatsiyalashda Naive Bayes, SVM va neyron tarmoqlar ishlatilib, eng yuqori aniqlik aynan shu usullarda kuzatilgan.

O'tkazilgan tahlillar asosida his-tuyg'ular tahlilini amalga oshirishning umumlashgan tarkibi (12-rasm)dagidek bo'ladi.



12-rasm. Kayfiyat tahlili metodlari diagrammasi

XYJOCA

O'rganilgan tadqiqotlar shuni ko'rsatadiki, ijtimoiy tarmoqlardagi matnli hujjatlarni tahlil qilishda mashinaviy o'rganish algoritmlari yuqori samaradorlik beradi. K-Means klasterlash usuli foydalanuvchilarni segmentlashda samarali bo'lsa, SVM va ansambl modellari tasniflashda yuqori aniqlik ko'rsatgan. LSTM modeli kontekstni hisobga olishda afzalliklarga ega bo'lsa-da, hisoblash murakkabligi yuqori.

Eng yuqori aniqlik ansambl modelda (93,7%) kuzatilgan, bu esa bir nechta algoritmlarni birlashtirish samarali ekanligini ko'rsatadi. Katta hajmdagi ma'lumotlar va to'g'ri tanlangan xususiyatlar (TF-IDF, N-gram va boshqalar) natijaning aniqligini oshiradi.

Mazkur natijalar ijtimoiy media monitoringi, marketing tahlili hamda xavfli kontentni erta aniqlash tizimlarini ishlab chiqishda muhim amaliy ahamiyatga ega.

ADABIYOTLAR RO'YXATI

1. Smith J., Brown T. Social Media Text Mining Techniques. – Journal of Data Science, 2018.
2. Ivanov P. Hybrid Clustering in VKontakte Analysis. – Information Technologies, 2019.
3. Kumar A., Singh R. Sentiment Analysis using SVM and Naive Bayes. – IEEE Access, 2020.
4. Boyd D., Ellison N. Social Network Sites: Definition and History. – Journal of Computer-Mediated Communication, 2007.
5. Chen L. Deep Learning for Social Media Analytics. – Springer, 2021.
6. VKontakte Data Analysis using SVM and TF-IDF Methods. – Applied Informatics, 2019.
7. Kohonen T. Self-Organizing Maps. – Springer, 2001.
8. Facebook User Interaction Patterns Study. – Social Media Research, 2017.
9. Twitter Data Mining Framework. – ACM Digital Library, 2018.
10. Instagram Clustering using K-Means. – International Journal of Computer Applications, 2020.
11. Apache Spark for Social Media Profiling. – Big Data Analytics, 2019.
12. LSTM-based Sentiment Analysis Model. – Neural Networks Journal, 2020.