DOI 10.52209/1609-1825_2023_4_360 ӘОЖ 004.93'14

Фэйк-аккаунттарды анықтау мәселесінде машиналық оқытуды қолдану

[1]**МУСИРАЛИЕВА Шынар Жеңісбекқызы,** *ф.-м.ғ.к., кафедра меңгерушісі, mussiraliyevash@gmail.com,*
[1]***АЗАМАТОВА Динара Тілеубердіқызы,** *магистрант, dinarazamatova@gmail.com,*
[1]**БАЙСПАЙ Гүлшат Болатқызы,** *постдокторант, gulshat.bgb2@gmail.com,*
[1]*«Әл-Фараби атындағы Қазақ ұлттық университеті» КеАҚ, Қазақстан, Алматы, әл-Фараби даңғылы, 71,*
**автор-корреспондент.*

Аңдатпа. *Қазіргі уақытта әлеуметтік желілердің қарқынды өсуіне байланысты әр түрлі қауіптердің пайда болу мүмкіндігі артып келеді, олардың бірі фэйк-аккаунттар. Әлеуметтік желіде фэйк-аккаунтты анықтау мәселесінде машиналық оқытуды қолдану қарастырылды. Жұмыстың негізгі мақсаты – әлеуметтік желілердегі фэйк-аккаунттарды тиімді анықтау үшін машиналық оқыту әдістері мен үлгілерін зерттеу және әзірлеу. Бұл зерттеу бопсалау, шантаж, аңдау, жалған қайыр сұрау, адамды қорлау, жеке ақпаратқа рұқсатсыз қол жеткізуді, пайдаланушының ақпаратын заңсыз қолдануды болдырмау секілді мәселелердің алдын алуына мүмкіндік береді. Зерттеу нысаны – фэйк-аккаунттар. Зерттеу жұмысында қолданылған әдістер: масштабтау, машиналық оқыту, атрибуттарды таңдау, модельді бағалау. Зерттеу жұмысының пәндік аймағы – Instagram. Зерттеудің ерекшелігі – фэйк-аккаунттарды анықтаудың әртүрлі, кең таралған әдістерін қолдана отырып, ең тиімдісін анықтау. Зерттеудің теориялық және практикалық маңызы – нақты пайдаланушыларға еліктейтін, фэйк-аккаунттарды анықтау арқылы спам жіберу, жеке деректерді жинау, қауіпті бағдарламаны тарату, онлайн рейтингтерді бұрмалау және т.б. зиянды әрекеттерден әлеуметтік желіні қауіпсіз қылдыру, сонымен қоса әлеуметтік желідегі қорғаныс және қауіпсіздікке арналған қосымшаларды дамытуға көмектесу. Фэйк-аккаунтты анықтау мақсатында машиналық оқыту классификаторлары тиімді қолданылды. Мақалада келесі модельдер пайдаланылады: шешім ағашының классификаторы; KNN жіктеуіші; логистикалық регрессия классификаторы; кездейсоқ орман классификаторы; қолдау векторлық машинасы (SVM) классификаторы. Дәлдік (accuracy), ROC қисығы, жіктеу есебі сияқты әртүрлі көрсеткіштерді қолдана отырып, модельдердің өнімділігін бағалау жүргізілді. Decision Tree, Random Forest және KNN сияқты кейбір модельдер бастапқы тестілеуде де, қайта оқытудан кейін де AUC-ROC, accuracy, precision, recall және F1-score көрсеткіштерінің тұрақты және жоғары мәндерін көрсетті. Logistic Regression және SVM сияқты кейбір модельдер таңдалған атрибуттерге байланысты әр түрлі тиімділік деңгейлерін көрсетеді. Бұл модельді оқыту кезінде функцияларды дұрыс таңдаудың маңыздылығын көрсетеді және барлық функциялар барлық модельдер үшін бірдей пайдалы емес екенін көрсетеді. Зерттеу нәтижесінде машиналық оқыту алгоритмдерін фэйк-аккаунтты анықтау мақсатында қолдану тиімді екеніне көз жеткізе отырып, зерттеу барысында жіктеу алгоритмдерін салыстыра келе, жоғары дәлдікке қол жеткізген алгоритм анықталды.*

Кілт сөздер: *машиналық оқыту, фэйк-аккаунт, ақпараттық қауіпсіздік, пайдаланушы профилі, әлеуметтік желілер, жалған тіркелгілерді анықтау, Instagram, Decision Tree, KNN, Logistic Regression, SVM, Random Forest.*

Кіріспе

Әлеуметтік желілердің белсенді дамуы біздің қазіргі ақпараттық қоғамымыз үшін өте маңызды. Олар жедел байланысқа, ақпарат алмасуға және виртуалды қауымдастықтарды құруға мүмкіндік беретін коммуникация процесінің ажырамас бөлігіне айналды. Дегенмен, қазіргі уақытта әлеуметтік желілердің қарқынды өсуіне байланысты әр түрлі қауіптердің пайда болу мүмкіндігі артып келеді, олардың бірі фэйк-аккаунттар.

Бұл шынайы пайдаланушыларды жасыру мақсатында немесе ұрланған сәйкестендіру ақпараты негізінде құрылған – фэйк-аккаунттар, әлеуметтік желідегі танымалдылық пен әсер ұғымдарын бұрмалайтын және бұл өз кезегінде экономикаға, саясатқа және қоғамға зиянды әсер етуі мүмкін аккаунт. Олардың желілерде таралуы мен белсенділігі ауыр зардаптарға әкеледі, соның ішінде пайдаланушылардың сенімін жоғалту, құпиялылықты бұзу, киберқылмыстың өсуі және ұлттық қауіпсіздікке қауіп төндіреді. Бұл аккаунттарды табу қиын болуы мүмкін, өйткені оларды жасаушылар әртүрлі маскировка әдістерін қолдана отырып, оларды нақты пайдаланушыларға

мүмкіндігінше жақындатуға тырысады. Осыған байланысты фэйк-аккаунттарды анықтау міндеті барған сайын өзекті және маңызды болып келеді. Фэйк-аккаунттарды анықтаудың дәстүрлі әдістері мен тәсілдері жеткілікті тиімді емес, өйткені шабуылдаушылар назардан тыс қалу үшін өз әдістері мен айла-амалдарын үнемі жетілдіріп отырады. Осы қиындықтарды ескере отырып, машиналық оқыту жалған тіркелгілерді анықтауға және әлеуметтік медиа қауіпсіздігін жақсартуға көмектесетін қуатты құрал болып табылады. Машиналық оқыту алгоритмдерінің көмегімен аккаунттардың мінез-құлық және құрылымдық сипаттамаларын автоматты түрде талдауға, жасырын үлгілер мен ауытқуларды анықтауға болады, бұл фэйк-аккаунттарды анықтау тиімділігін арттыруға мүмкіндік береді.

Жұмыс аясында масштабтау, машиналық оқыту, атрибуттарды таңдау, модельді бағалау әдістері қолданылады. Масштабтау әдістері үлкен көлемдегі мәліметтерді өңдеу және олармен тиімді жұмыс істеу үшін қолданылады. Шешімдер ағашы, KNN, логистикалық регрессия, кездейсоқ орман, SVM сияқты машиналық оқыту әдістері фэйк-аккаунттарды анықтау мақсатында жіктеу алгоритмдерін оқыту үшін пайдаланылады. Атрибуттарды талдау әдістері функциялардың маңыздылығын анықтау үшін пайдаланылады. Модельді бағалау әдістері оқытылған үлгілердің тиімділігі мен сапасын бағалау үшін әртүрлі әдістер қолданылады.

Фэйк-аккаунттарды анықтауға арналған оңтайлы атрибуттарды талдауды және таңдауды қамтиды. Функцияларды талдау әдістерін қолдану классификация нәтижесіне әсер ететін маңызды ақпаратты белгілерін анықтауға мүмкіндік береді. Жұмыс сонымен қатар оқытылған үлгілердің өнімділігін бағалауды қамтиды. Сонымен қоса, атрибуттардың маңыздылығын зерттегеннен кейін, сол атрибуттар негізінде қайта оқыту жүргізе отырып, үлгінің өнімділігінің өсуі де бұл жұмыстың ғылыми жаңалығының бірі болып табылады. Машиналық оқытуды пайдалана отырып, фэйк-аккаунттарды анықтау саласындағы бұл жұмыс осы аспектілердің барлығы ғылыми жаңалықты білдіреді және осы зерттеу саласының дамуына ықпал етеді.

Әдебиеттерге шолу

Kate Bojkov «2023 жылғы маркетингке арналған 29 Instagram статистикасы» мақаласы [1] Instagram маркетингіне қатысты соңғы статистикаға толық шолу жасайды. Автор бүкіл әлем бойынша 1 миллиардтан астам белсенді пайдаланушысы бар Instagram әлеуметтік медиа платформасы ретіндегі үздіксіз өсуі мен танымалдылығын атап көрсетеді. CNBC есебі бойынша, АҚШ-тағы жалған «танымал тұлғалар» Instagram әлеуметтік желісінде фэйк ізбасарларды сатып алуға жыл сайын 1,3 млрд доллар жұмсайтынын анықтады. Сондықтан осындай әлеуметтік платформада сенімділікті қолдауды маңызды.

Boshmaf Yazan және т.б. OSN-де жалған тіркелгілерді анықтау осы платформалардың тұтастығы мен қауіпсіздігін сақтау үшін өте маңызды екенін атап өтеді [2]. Фэйк-аккаунтты анықтаудың тәсілдері, ең алдымен, тіркелгі жасау уақыты, тіркелгі әрекеті және әлеуметтік графикті талдау сияқты метрикаларға негізделген. Дегенмен, бұл тәсілдер шынайы пайдаланушылардың мінез-құлқына еліктейтін және анықтаудан жалтаратын күрделі жалған тіркелгілерді анықтауда тиімді болмауы мүмкін. Бұл мәселені шешу үшін авторлар Integro деп аталатын «құрбандықты болжауға» негізделген тәсілді ұсынады, ол фэйк-аккаунттар көбінесе «нақты құрбандарға» бағытталған фактіні қолданады. Авторлар Integro өнімділігін Twitter және Weibo секілді әлеуметтік желілерден нақты деректер жиынтығы арқылы бағалап, оның өнімділігін бар тәсілдермен салыстырған.

Cresci Stefano және т.б. «Сатылатын даңқ: Twitter желісіндегі жалған ізбасарларды тиімді анықтау» мақаласы әлеуметтік желілерге әсер етушілердің көбеюіне және осы кеңістіктегі алаяқтық әлеуетіне байланысты маңызды мәселе болып табылатын жалған Twitter жазылушыларын анықтау әдістемесін ұсынады [3]. Авторлар желі құрылымын, әрекет үлгілерін және лингвистикалық белгілерді қоса алғанда, метадеректер тіркесімін пайдаланатын машиналық оқытуға негізделген тәсілді ұсынады. Зерттеу авторлардың қолданыстағы әдістермен шынайы немесе жалған деп анықталған Twitter аккаунттарының үлкен деректер жинағын жинауынан басталады. Содан кейін олар осы тіркелгілерден әртүрлі метадеректерді, соның ішінде жазылушылар мен достар санын, тіркелгі жасын және твиттер жиілігі мен уақытын шығарады. Бұған қоса, олар автоматтандыруды немесе басқа күдікті әрекетті көрсетуі мүмкін лингвистикалық маркерлер үшін твиттер мәтінін талдайды. Осы метадеректерді пайдалана отырып, авторлар тіркелгілерді шынайы немесе жалған деп жіктеу үшін машиналық оқыту үлгісін үйретеді. Олар ең тиімді тәсілді анықтау үшін әртүрлі метадеректер комбинацияларымен және машиналық оқыту алгоритмдерімен тәжірибе жасайды. Бұл зерттеудің қызықты аспектісі – тиімділікке назар аудару. Авторлар мәселенің ауқымы әрбір есептік жазбаны қолмен қарап шығуды мүмкін емес ететінін мойындайды, сондықтан олар тез және оңай есептелетін атрибуттарге басымдық береді. Бұл олардың көзқарасын салыстырмалы түрде қысқа уақыт ішінде үлкен деректер жиынына қолдануға мүмкіндік береді.

Daniel Arp және т.б. «Компьютерлік қауіпсіздікте машиналық оқытудың орындалатын және орындалмайтын тұстары» мақаласы компьютерлік қауіпсіздік мәселелеріне машиналық оқыту әдістерін қолданудың ең жақсы тәжірибелеріне толық шолу жасайды [4]. Авторлар жалпы қателіктерді атап көрсетеді және оларды болдырмау туралы нұсқаулар береді. Авторлар дұрыс

емес нәтижелерге әкелуі мүмкін тым қарапайым немесе толық емес деректер жиынын пайдаланудан сақтандырады. Олар сондай-ақ шешілетін мәселені мұқият анықтаудың және өнімділікті бағалау үшін сәйкес көрсеткіштерді таңдаудың маңыздылығын атап көрсетеді. Зерттеудің тағы бір негізгі аспектісі – функцияларды таңдау және функцияларды жобалауды талқылау. Авторлар қарастырылып отырған мәселеге сәйкес келетін және оңай есептелетін функцияларды таңдаудың маңыздылығын атап көрсетеді. Олар сондай-ақ өнімділікті жақсарту үшін функцияларды түрлендіру және біріктіру әдістерін талқылайды. Авторлар сонымен қатар машиналық оқыту алгоритмдерін дұрыс баптаудың және артық орнатуды болдырмаудың маңыздылығын атап көрсетеді. Олар өнімділікті бағалау және оңтайландыру үшін кросс-валидацияны және басқа әдістерді пайдалануды талқылайды. Соңында, зерттеу компьютерлік қауіпсіздік үшін машиналық оқытудағы этикалық ойларды талқылаумен аяқталады. Авторлар біржақтылықтан аулақ болу және машиналық оқытуды пайдалану жеке адамдарға немесе топтарға зиян келтірмейтінін қамтамасыз етудің маңыздылығын атап көрсетеді. Жалпы алғанда, «Компьютерлік қауіпсіздікте машиналық оқытудың орындалатын және орындалмайтын тұстары» компьютерлік қауіпсіздік мәселелеріне машиналық оқытуды қолданғысы келетін тәжірибешілер үшін құнды нұсқаулық береді. Проблемалық аймақты түсінуге, сәйкес деректер көздерін таңдауға және жалпы қателіктерді болдырмауға баса назар аудару пайдалы. Сонымен қатар, этикалық ойларды талқылау осы салада машиналық оқытуды жауапты және мұқият пайдалану қажеттілігін көрсетеді.

Thomas K., Grier C. және т.б. спам тіркелгілерінің сипаттамаларын және олар тарататын мазмұн түрлерін анықтауға бағытталған [5]. Сондай-ақ, олар Twitter-дің спамды анықтау алгоритмдерінің тиімділігін зерттейді. Авторлар уақытша тоқтатылған Twitter тіркелгілерінің деректер жинағын жинап, тіркелгі белсенділігі мен мазмұнына қарай оларды спам немесе спам емес деп жіктеді. Содан кейін олар спам тіркелгілерінің сипаттамаларын, соның ішінде олар бөлісетін мазмұн түрлерін, олардың жазылушыларының санын және олардың белсенділік үлгілерін талдады. Тұтастай алғанда, «Ретроспективада тоқтатылған тіркелгілер: Twitter спамының талдауы» Twitter-дегі спам тіркелгілерінің сипаттамалары және Twitter-дің спамды анықтау алгоритмдерінің тиімділігі туралы құнды түсінік береді. Спамды анықтау алгоритмдерін жақсарту және әлеуметтік медиа платформаларындағы спаммен күресу бойынша үздіксіз күш-жігердің маңыздылығын зерттеу көрсетеді.

Khaled Sarah және т.б. «Әлеуметтік желілердегі жалған аккаунттарды анықтау» мақаласында әлеуметтік медиа платформаларындағы жалған аккаунттарды анықтау мәселесіне шолу жасауға және осы мәселені шешу үшін ұсынылған әртүрлі тәсілдерді талқылауға бағытталған [6]. Мақала жалған аккаунттар түсінігін және әлеуметтік медиа платформаларында табуға болатын жалған аккаунттардың әртүрлі түрлерін анықтаудан басталады. Содан кейін авторлар жалған тіркелгілерді анықтау үшін ұсынылған әртүрлі әдістерді сипаттайды, соның ішінде мазмұнға негізделген, әлеуметтік графикаға негізделген және мінез-құлыққа негізделген тәсілдер. Олар, сондай-ақ, осы әдістердің шектеулері мен қиындықтарын талқылайды және жақсырақ анықтау дәлдігі үшін бірнеше тәсілдерді біріктірудің маңыздылығын көрсетеді.

Әдістер мен құралдар

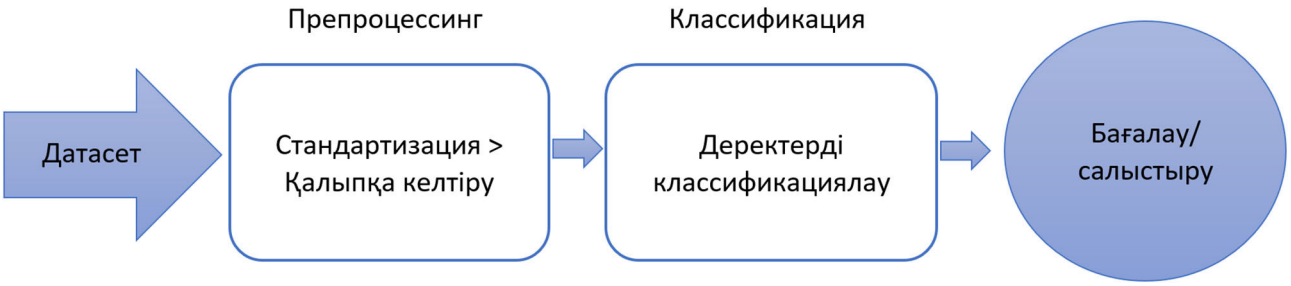
Әлеуметтік желідегі фэйк-аккаунтты анықтауға ұсынылған әдіс 4 негізгі қадамнан тұрады, яғни: деректер жиынтығын жинау, деректерді алдын ала өңдеу кезеңінен өткізу, препроцессингтен өткізілген деректерді классификациялау, бағалау және салыстыру кезеңі. Берілген зерттеу жұмысында фэйк-аккаунтты анықтау процесінің өту кезеңдері 1-суретте көрсетілген.

Фэйк-аккаунтты анықтау мақсатында деректер жиынтығы

Жұмыстың бірінші кезеңі – аккаунт туралы ақпаратты жинау. Бұл зерттеуде дайын белгілер қойылған датасет қолданылды [7]. Бұл деректер жиынтығы Kaggle ашық ресурсынан алынды. Осы жұмыста қолданылатын мәліметтер жиынтығын автор Bardiya Bakhshandeh жасаған және анықтаған. Автор спам/фэйк-аккаунттарды анықтау үшін деректердің әр данасын мұқият зерттеді. Деректер жиынтығы 2019 жылдың 15-19 наурызы аралығында краулердің көмегімен жиналды. Бұл процесс фэйк-аккаунттарды анықтау үшін пайдалы болуы мүмкін әртүрлі аккаунттар атрибуттары мен сипаттамаларын жинады.

Алайда, кейбір аккаунттар фэйк деп қате анықталғанымен, мақала авторлары деректердің әрбір данасын мұқият зерттеу арқылы фэйк-аккаунтты анықтауда жоғары дәлдікті қамтамасыз етуге тырысқанын атап өткен жөн. Авторлар өз жобасында Instagram-дағы фэйк және спамдық аккаунттарды анықтау мәселесін шешу мақсатында осы деректер жиынтығын машиналық оқыту модельдерін әзірлеу және оқыту үшін пайдаланды. Деректер модельдерді оқыту және олардың өнімділігін тексеру үшін, сондай-ақ жалған аккаунттар мен әлеуметтік медиа спамымен күресу үшін қолданылатын әдістерді зерттеу үшін пайдаланылуы мүмкін.

Зерттеу жұмысында бұл датасет Instagram әлеуметтік желісінде машиналық оқытуды қолдану арқылы фейк және спам аккаунттарды анықтау үшін қолданылады. Берілген датасет 2-ге бөлінген: train және test. Train деректер жинағы 576 жолдан × 12 бағаннан, ал test деректер жинағы 120 жолдан × 12 бағаннан тұрады. Бұл датасет келесідей ме-



1-сурет – Фэйк-аккаунтты анықтау процесінің кезеңдері

тадеректерден құралған: профиль суреті, пайдаланушы атының саны/ұзындығы, толық атындағы сөздер, толық атауының ұзындығы/сан кездесуі, аты = = пайдаланушы аты, сипаттама ұзындығы, сыртқы URL, жеке аккаунт растығына тексеру, #хабарламалар, #аккаунтқа подписка жасаған жазылушылар саны, #аккаунттың жазылушылардың саны, фэйк белгісі.

Бұл датасет келесідей метадеректерден құралған:

1. profile pic. Пайдаланушының профиль суреті бар немесе жоқ.
2. nums/length username. Пайдаланушы атындағы сандық таңбалар санының оның ұзындығына қатынасы.
3. fullname words. Сөз лексемаларындағы толық аты.
4. nums/length fullname. Толық атаудағы сандық таңбалар санының оның ұзындығына қатынасы.
5. name==username. Пайдаланушы аты мен толық аты сөзбе-сөз сәйкес келе ме.
6. description length. Пайдаланушының сипаттама ұзындығы (био).
7. external URL. Сыртқы URL бар немесе жоқ.
8. private. Жеке немесе жоқ.
9. #posts. Хабарламалар саны.
10. #followers. Ізбасарлар саны (жазылушылар).
11. #follows. Жазылымдар саны.
12. fake. Белгі.

Деректерді алдын ала өңдеу

Машиналық оқытудағы деректерді алдын ала өңдеу модельді оқытудан бұрын орындалатын маңызды қадам болып табылады. Оның мақсаты – деректерді модельге ыңғайлы етіп дайындау және оның тиімділігін арттыру. Препроцессинг кезеңінде атрибуттарды (функцияларды) масштабтау таңдалды, өйткені деректер жиынындағы атрибут мәндерінің шкалалары әртүрлі.

Бұл зерттеуде алдын ала өңдеудің екі түрі қолданылды:

1. Қалыпқа келтіру (Normalization: зерттеуде `sklearn` кітапханасының `preprocessing` модулінен `normalizer` нысаны қолданылды. Нормализация атрибут мәндері масштабталатын және белгілі бір ауқымда болатындай деректерді түрлендіруді білдіреді. Бұл жағдайда атрибут мәндерінің стандарттауын және сәйкестігін қамтамасыз ету үшін деректер «Normalizer» көмегімен қалыпқа келтірілді.

2. Стандарттау (Standardization): зерттеуде `sklearn` кітапханасының `preprocessing` модулінен `StandardScaler`нысаны қолданылды. Стандарттау сонымен қатар деректерді масштабтауды білдіреді, бірақ бұл жағдайда белгілердің мәндері олардың орташа мәні 0 және стандартты ауытқуы 1 болатындай етіп түрлендірілді. Бұл оқу үлгісіндегі әрбір белгінің орташа және стандартты ауытқуын есептейтін, содан кейін train және test деректеріне түрлендіруді қолданатын `StandardScaler` нысанын пайдалану арқылы жасалды. Төменде 1-кестеде таңдалған алдын ала өңдеу түрлерінің артықшылықтары мен кемшіліктері көрсетілген.

Машиналық оқытудың жіктеу алгоритмдерін қолдану

Деректерді препроцессингтен өткізгеннен кейін, машиналық оқыту модельдерін таңдау кезеңіне өтуге болады. Бұл жұмыста келесі модельдер пайдаланылды:

- Шешім ағашының классификаторы;
- KNN жіктеуіші;
- Логистикалық регрессия классификаторы;
- Кездейсоқ орман классификаторы;
- Қолдау векторлық машинасы (SVM) классификаторы.

Фэйк-аккаунттарды анықтау мәселесіне арналған үлгілерді таңдау деректер сипаттамаларын, деректер жинағының өлшемін, қолжетімді ресурстарды және үлгі дәлдігі мен түсіндірмелілігіне қойылатын талаптарды қоса алғанда, әртүрлі факторларға байланысты. Ұсынылған модельдердің әрқайсысының өзіндік артықшылықтары мен шектеулері бар, модельді таңдаудың тиімділігі мәселенің және деректердің ерекшеліктеріне байланысты болады. Таңдалған модельдердің артықшылығы мен кемшіліктері 2-кестеде көрсетілген.

Оқытылған машиналық оқыту моделінің сапасын бағалау үшін әртүрлі көрсеткіштер мен әдістерді қолдануға болады. Бұл зерттеуде келесідей әдістер қолданылды:

- ROC талдау қисығы: сезімталдық (True Positive Rate) мен ерекшелік (False positive Rate) арасындағы сәйкестікті ескере отырып, үлгі өнімділігін бағалауға мүмкіндік береді. Қисық диаграмманың жоғарғы сол жақ бұрышына неғұрлым жақын болса, модель соғұрлым жақсы болады.

- Жіктеу есебі: Жіктеу есебінде әр класс үшін дәлдік (precision), recall (толықтығы), F1-ұпайы (F1-score) және support сияқты әртүрлі көрсеткіштер бар. Бұл көрсеткіштер жалған және нақты тіркелгілерді болжаудағы үлгі өнімділігі туралы ақпаратты береді.

1-кесте – Таңдалған препроцессинг түрлерінің артықшылығы мен кемшіліктері

Препроцессинг түрі	Артықшылықтары	Кемшіліктері
Нормализация	- Нормалау функция мәндерін белгілі бір диапазонға таратады, бұл атрибуттар арасында сәйкестікті қамтамасыз етеді. - Белгілердің мәндерінде әртүрлі шкалалар немесе өлшем бірліктері болған кезде пайдалы.	- Нормалау деректердегі шектен тыс мәндерге сезімтал болуы мүмкін, себебі ол ең төменгі және ең үлкен атрибут мәндеріне негізделген. - Деректерді тарату туралы ақпараттың жоғалуына әкелуі мүмкін, әсіресе деректерде экстремалды мәндер болса.
Стандартизация	- Стандарттау атрибут мәндерін орташа мәнге 0 және стандартты ауытқуға 1 әкеледі, бұл деректер қалыпты түрде таратылады деген болжамға негізделген үлгілер үшін пайдалы. - Нормализациямен салыстырғанда деректердегі шектен тыс мәндердің әсері аз.	- Сирек деректермен жұмыс істейтін үлгілер немесе атрибуттар арасындағы қашықтықты немесе корреляция коэффициенттерін пайдаланатын үлгілер сияқты бастапқы деректер масштабының сақталуын талап ететін үлгілер үшін жарамсыз болуы мүмкін.

2-кесте – Моделдердің артықшылығы мен кемшілігі

Модельдің аты	Артықшылығы	Кемшілігі
Decision tree	- түсіндіру мүмкіндігі: шешім ағашы шешім қабылдау процесінің нақты графикалық көрінісін береді; - сандық және категориялық деректерді өңдеу; - автоматты атрибутты таңдау; - масштабтау және оқу жылдамдығы.	- шулы деректерге шамадан тыс бейімделу үрдісі; - күрделі сызықтық емес қатынастарды модельдеу қабілетсіздігі.
Random Forest	- жақсы жалпылау және болжау дәлдігі; - сандық және категориялық атрибуттарды өңдей алады; - шамадан тыс орнатуға және деректер шуына төзімді; - ерекшеліктердің маңыздылығын бағалай білу.	- кейбір басқа модельдермен салыстырғанда көбірек есептеу ресурстары мен оқу уақытын қажет етеді; - нәтижелерді түсіндіру қиын болуы мүмкін.
Logistic Regression	- үлгінің қарапайымдылығы мен түсіндірмелілігі; - ақпараттық белгілер болған жағдайда тиімді болуы мүмкін; - деректер сызықты түрде бөлінетін жағдайларда жақсы жұмыс істей алады.	- атрибуттар мен мақсатты айнымалы арасындағы қатынастар сызықты емес болса, тиімді болмауы мүмкін; - шамадан тыс көрсеткіштерге және үлгі болжамдарының бұзылуына сезімтал болуы мүмкін.
Support Vector Machine	- күрделі бөлу геометриясын жақсы меңгеру; - сандық және категориялық атрибуттарды өңдей алады; - ядро функцияларын пайдалану сызықтық емес тәуелділіктерді модельдеуге мүмкіндік береді.	- дұрыс ядро таңдауды және гиперпараметрді баптауды талап етеді; - үлкен деректер жинақтары үшін есептеу қымбат болуы мүмкін.
K-Nearest Neighbors	- іске асыру және түсіну жеңілдігі; - функция кеңістігіндегі ұқсас нүктелердің топтарын анықтай алады; - жақын объектілерде ұқсас белгілер болса, жақсы жұмыс істей алады.	- k-параметрін таңдауға сезімталдық; - жақсы таңдалған және масштабталған атрибуттарды қажет етеді; - үлкен деректер жинақтары үшін есептеу қымбат болуы мүмкін.

Үлгінің өнімділігін бағалау

Модель үйретілгеннен кейін оның өнімділігі бұрын көрмеген жаңа деректер жинақтарында сынау мақсатында test деректер жинағын қажет етеді, таңдалған датасет өзінде train және test датасетін қамтығандықтан тексеруді test датасетінде жүргізсек болады. Моделіміз оқу деректер жинағында (train) оқытылғандықтан, келесі қадам test деректер жиынындағы белгілерді болжау үшін осы үлгіні пайдалану, содан кейін болжанған белгілерді test деректер жиынының шынайы белгілерімен салыстыру. Test датасетінің құрылымы 2-суретте көрсетілген.

Test_data датасетін препроцессинг кезеңінен өткізіп алғаннан кейін, бұл датасет өзінде класс белгілерін қамтығандықтан, оны test_features, test_labels деп бөліп алдық. test_features класс белгілерін қамтитын соңғы бағанды қоспағанда, test деректер жиынындағы барлық атрибут бағандарын қамтиды. test_labels класс белгілерін қамтитын test деректер жинағының соңғы бағанын ғана қамтиды. Бағалау әдісі ретінде алғашында дәлдік көрсеткіші тексерілді: бұл жіктеу мәселелеріне арналған ең көп тараған көрсеткіштердің бірі. Ол мысалдардың жалпы санына қатысты дұрыс жіктелген мысалдардың үлесін өлшейді. test_date дәлдігін бағалау үшін score() функциясын қолданылды және жүзеге асу коды 3-суретте көрсетілген.

Модель нәтижелеріне көз жүгіртетін болсақ, Decision Tree, KNN, Random Forest жақсы өнімділік көрсеткенін, ал Logistic Regression және SVM едәуір аз өнімділік көрсетті, осыған орай атрибуттардың маңыздылығын таңдау бойынша қайта оқытуда осы модельдердің өнімділігін көтеруге негізделеді.

Атрибуттардың маңыздылығын талдау фэйк-аккаунттарды болжауда қандай сипатта-

In [104]:
```
import pandas as pd
test_data = pd.read_csv("test.csv")
test_data
```

Out[104]:

	profile pic	nums/length username	fullname words	nums/length fullname	name==username	description length	external URL	private	#posts	#followers	#follows	fake
0	1	0.33	1	0.33	1	30	0	1	35	488	604	0
1	1	0.00	5	0.00	0	64	0	1	3	35	6	0
2	1	0.00	2	0.00	0	82	0	1	319	328	668	0
3	1	0.00	1	0.00	0	143	0	1	273	14890	7369	0
4	1	0.50	1	0.00	0	76	0	1	6	225	356	0
...	...	...	...	...	...	...	...	...	...	...	...	...
115	1	0.29	1	0.00	0	0	0	0	13	114	811	1
116	1	0.40	1	0.00	0	0	0	0	4	150	164	1
117	1	0.00	2	0.00	0	0	0	0	3	833	3572	1
118	0	0.17	1	0.00	0	0	0	0	1	219	1695	1
119	1	0.44	1	0.00	0	0	0	0	3	39	68	1

120 rows × 12 columns

2-сурет – «Test» датасеті

In [131]:
```
predicted_labels_clf = clf.predict(test_data)
predicted_labels_knn = knn.predict(test_data)
predicted_labels_lr = lr.predict(test_data)
predicted_labels_rfm = rfm.predict(test_data)
predicted_labels_clf_svm = clf_svm.predict(test_data)
true_labels = test_labels  # Здесь test_labels - истинные метки классов для тестового датасета

accuracy1 = accuracy_score(true_labels, predicted_labels_clf)
accuracy2 = accuracy_score(true_labels, predicted_labels_knn)
accuracy3 = accuracy_score(true_labels, predicted_labels_lr)
accuracy4 = accuracy_score(true_labels, predicted_labels_rfm)
accuracy5 = accuracy_score(true_labels, predicted_labels_clf_svm)

print("Accuracy Decision Tree:", accuracy1)
print("Accuracy KNN:", accuracy2)
print("Accuracy Logistic Regression:", accuracy3)
print("Accuracy Random Forest:", accuracy4)
print("Accuracy SVM:", accuracy5)
```

```
Accuracy Decision Tree: 1.0
Accuracy KNN: 0.85
Accuracy Logistic Regression: 0.5166666666666667
Accuracy Random Forest: 1.0
Accuracy SVM: 0.5
```

3-сурет – «Test» датасетіне score() функциясын қолдану

малардың маңызды рөл атқаратынын анықтауға көмектеседі. Бұл аккаунттың жазылушылар саны, белсенділік және фэйк және нақты аккаунттар арасында әр түрлі болуы мүмкін басқа белгілер сияқты атрибуттарды қамтуы мүмкін. Бұл жұмыста функцияның маңыздылығын бағалау үшін 3 модель негізінде есептелінді: Random Forest, Decision Tree, Logistic Regression. Өкінішке орай, KNN және SVM функцияның маңыздылығын анықтау үшін бұл модельдерді қолдана алмадық, себебі бұл модельдер атрибуттарды таңдауға арналған кірістірілген механизмді қамтамасыз етпейді және өз модельдерінде бастапқы деректер жиынындағы барлық атрибуттарды пайдаланады.

Қолданылған үлгілердің әрқайсысы атрибут маңыздылығын бағалаудың әртүрлі жолдарын қамтамасыз етеді:

- Шешімдер ағашы: функциялардың маңыздылығын анықтау үшін әр түйіндегі деректерді бөлу үшін қолданылатын функцияларды пайдалануға болады. Функция деректерді бөлу үшін неғұрлым көп қолданылса немесе ағашта неғұрлым жоғары болса, соғұрлым ол маңызды болып саналады.

- Кездейсоқ орман: атрибуттардың маңыздылығын ағаштардың бүкіл ансамблі бойынша деректерді бөлу сапасын жақсартуға қосқан үлесі негізінде бағалауға болады. Атрибут маңыздылығының көрсеткіші атрибуттың ағаштардағы деректерді бөлу үшін пайдаланылған кезде модель болжамдарындағы белгісіздікті қаншалықты азайтатынына негізделген.

- SVM: функциялардың маңыздылығын кластар арасындағы бөлу шекарасын анықтауға қосқан үлесі негізінде бағалауға болады. Бөлу шекарасын анықтауда шешуші рөл атқаратын және ең жақсы классификацияны қамтамасыз ететін белгілер маңызды болып саналады.

Маңызды атрибуттарды таңдау арқылы қайтадан оқыту нәтижелерін өзара салыстыру.

Келесі Logistic Regression моделіне жүргізілген салыстырулар нәтижелері 3-кестеде көрсетілген.

Келесі SVM моделіне жүргізілген салыстырулар нәтижелері 4-кестеде көрсетілген.

Салыстырудың барлық кезеңінен кейін қайта даярлаудан кейін келесі қорытындылар жасауға болады:

1. Decision Tree, Random Forest және KNN сияқты кейбір модельдер бастапқы тестілеуде де, қайта оқытудан кейін де AUC-ROC, accuracy, precision, recall және F1-score көрсеткіштерінің тұрақты және жоғары мәндерін көрсетті. Бұл олардың таңдалған функциялардың өзгеруіне қарамастан деректерді жақсы жалпылау және болжау қабілетін көрсетеді.

2. Logistic Regression моделі таңдалған атрибуттарды пайдалана отырып, қайта оқытудан кейін көрсеткіштердің айтарлықтай жақсарғанын көрсетті. Бұл таңдалған функциялар оның тиімділігіне және дәлірек болжау жасау қабілетіне айтарлықтай әсер ететінін көрсетеді.

3. SVM моделі сонымен қатар таңдалған атрибуттармен қайта оқытудан кейін көрсеткіштердің жақсарғанын көрсетті. Бұл таңдалған атрибуттар модельге сыныптарды жақсырақ бөлуге және оның жіктеу қабілетін жақсартуға көмектескенін көрсетеді.

4. Logistic Regression және SVM сияқты кейбір модельдер таңдалған атрибуттарға байланысты әр түрлі тиімділік деңгейлерін көрсетеді. Бұл модельді оқыту кезінде функцияларды дұрыс таңдаудың маңыздылығын көрсетеді және барлық функциялар барлық модельдер үшін бірдей пай-

3-кесте – Логистикалық регрессия моделы бойынша оқытылған кейін және таңдалған функциялар арқылы оқытылған кейінгі көрсеткіштері

Модель	AUC-ROC	accuracy	precision	recall	F1-score
Logistic Regression	0.52	0.52	0.75	0.52	0.37
Rfm негізінде таңдалған функциялар (Logistic Regression)	0.95	0.91	0.93	0.91	0.91
Dt негізінде таңдалған функциялар (Logistic Regression)	0.86	0.86	0.86	0.86	0.86
Lr негізінде таңдалған функциялар (Logistic Regression)	0.85	0.86	0.88	0.85	0.86

4-кесте – SVM моделі бойынша оқытылған кейін және таңдалған функциялар арқылы оқытылған кейінгі көрсеткіштері

Модель	AUC-ROC	accuracy	precision	recall	F1-score
SVM	0.50	0.50	0.25	0.50	0.33
Rfm негізінде таңдалған функциялар (SVM)	0.89	0.47	0.73	0.52	0.35
Dt негізінде таңдалған функциялар (SVM)	0.52	0.47	0.73	0.52	0.35
Lr негізінде таңдалған функциялар (SVM)	0.86	0.87	0.89	0.86	0.87

далы емес екенін көрсетеді.

Тұтастай алғанда, таңдалған атрибуттарды пайдалана отырып, қайта оқыту кейбір жағдайларда модельдердің тиімділігін жақсартуға әкелді және оңтайлы нәтижелерге қол жеткізу үшін атрибуттарды дұрыс таңдаудың маңыздылығын атап өтті.

Қорытынды

Әлеуметтік желілердегі фэйк-аккаунттарды анықтау мәселесінде машиналық оқыту әдістердің қолданылуы зерттелді және қарастырылды.

Зерттеу жұмысында машиналық оқытуды қолдана отырып, Instagram әлеуметтік желісіндегі фэйк және фэйк емес аккаунттарды анықтау бойынша зерттеулер жүргізілді. Ол үшін фэйк және фэйк емес аккаунттар туралы мәліметтер бар дайын деректер жиынтығы қолданылды. Машиналық оқыту үлгілерінің тиімділігін қамтамасыз ету үшін объектілерді масштабтауды қоса алғанда, деректерді алдын ала өңдеу процесі қолданылды.

Зерттеу барысында әртүрлі машиналық оқыту алгоритмдері қарастырылды, соның ішінде Decision Tree, KNN, Logistic Regression, Random Forest және SVM. Дәлдік (accuracy), ROC қисығы, жіктеу есебі сияқты әртүрлі көрсеткіштерді қолдана отырып, модельдердің өнімділігін бағалау жүргізілді.

Бұл зерттеудің маңызды үлесі – атрибуттардың маңыздылығын және олардың фэйк-аккаунттарды анықтау процесіне әсерін талдау. Функциялардың маңыздылығын талдау қандай пайдаланушы атрибуттары фэйк-аккаунттарды анықтау процесіне көбірек әсер ететінін анықтауға мүмкіндік берді. Test датасетінде оқытылған модельдің өнімділігін тексеруден кейінгі нәтижелерді алғаннан кейін Logistic Regression және SVM модельдерінің төмен нәтиже көрсетуінің себебін табу мақсатында әрбір атрибут үшін атрибуттың маңыздылығын анықтадық. Белгілердің маңыздылығын талдау Random Forest, Decision Tree және Logistic Regression модельдерін қолдану арқылы жүргізілді. Бұл фэйк және фэйк емес аккаунттарды болжауға әсер ететін ең маңызды белгілерді анықтауға мүмкіндік берді. Таңдалған белгілерге негізделген модельдерді қайта оқытудан кейін кейбір модельдер, соның ішінде Decision Tree, Random Forest және KNN, қайта оқытуға дейін де, одан кейін де тұрақты және жоғары өнімділікті көрсетті. Сонымен қатар, Logistic Regression және SVM модельдері қайта оқытудан кейін өнімділіктің айтарлықтай жақсарғанын көрсетті. Таңдалған модельдер мен әдістердің шектеулері мен атрибуттары бар екенін ескеру маңызды. Осы нәтижелерге сүйене отырып, таңдалған функциялар модельді жақсартуда тиімді болды және болжамдардың тиімділігі мен дәлдігіне ықпал етті деген қорытынды жасауға болады.

Қорытындылай келе, фэйк-аккаунттарды анықтау мәселесінде машиналық оқыту әдістерін пайдалану үлкен әлеуетке ие және әлеуметтік медиа тіркелгілерін анықтаудың тиімді және перспективалы әдісі екенін растайды. Фэйк-аккаунттардың анықталған негізгі сипаттамалары анықтау модельдері мен алгоритмдерін нақтылауға, олардың тиімділігі мен дәлдігін арттыруға көмектеседі. Бұл тәсіл фэйк-аккаунттарды анықтаудың күрделі және жетілдірілген әдістерін әзірлеуге жаңа мүмкіндіктер ашады.

Берілген зерттеу Қазақстан Республикасы Ғылым және жоғары білім министрлігінің Ғылым комитеті қаржыландыратын «Қазақ тіліндегі кибер экстремизмнің идеологиялық бағыттарын жасанды интеллект әдістері көмегімен мультикласификациялау» жобасы аясында орындалды (грант AP19676342, жоба жетекшісі Мусиралиева Ш.Ж.).

ӘДЕБИЕТТЕР ТІЗІМІ

1. Кейт Бойкова. 29 Статистика Instagram по маркетингу в 2023 г. // [Электронный ресурс] / URL: https://embedsocial.com/blog/instagram-statistics/
2. Boshmaf Yazan, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. «Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs». [Текст] / In NDSS, vol. 15, pp. 8-11. – 2015.
3. Cresci Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, Maurizio Tesconi «Fame for sale: Efficient detection of fake Twitter followers». [Текст] / Decision Support Systems 80, 2015.
4. Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallarok, Konrad Rieck «Dos and Don'ts of Machine Learning in Computer Security» [Текст] / USENIX Security Symposium 2022, 2022.
5. Thomas K., Grier C., Song D., & Paxson, V. «Suspended accounts in retrospect: an analysis of twitter spam». [Текст] / In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement (pp. 243-258). 2011.
6. Khaled Sarah, El-Tazi Neamat, Mokhtar Hoda «Detecting Fake Accounts on Social Media». [Текст] / IEEE International Conference, 2018.
7. Dataset // [Elektronnyi-resurs] / URL: https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts?select=train.csv

Применение машинного обучения в задаче обнаружения фейковых аккаунтов

[1]***МУСИРАЛИЕВА Шынар Женисбековна,*** *к.ф.-м.н., зав. кафедрой, mussiraliyevash@gmail.com,*
[1]****АЗАМАТОВА Динара Тилеубердиевна,*** *магистрант, dinarazamatova@gmail.com,*
[1]***БАЙСПАЙ Гүлшат Болатқызы,*** *постдокторант, gulshat.bgb2@gmail.com,*
[1]НАО «Казахский национальный университет имени аль-Фараби», Казахстан, Алматы, пр. аль-Фараби, 71,
**автор-корреспондент.*

Аннотация. *В настоящее время из-за быстрого роста социальных сетей увеличивается вероятность возникновения различных угроз, одна из которых – фейковые аккаунты. Фейк-аккаунт – это учетная запись, созданная на основе поддельной или украденной идентификационной информации, которая искажает концепции популярности и влияния в социальных сетях, что, в свою очередь, может иметь пагубные последствия для экономики, политики и общества. Рассмотрено использование машинного обучения в вопросе выявления фейковых аккаунтов в социальной сети. Основная цель работы – исследование и разработка методов и моделей машинного обучения для эффективного выявления фейковых аккаунтов в социальных сетях. Это позволяет избежать таких проблем, как вымогательство, шантаж, преследование, поддельные просьбы, оскорбление человека, несанкционированный доступ к личной информации, незаконное использование информации пользователя посредством исследования. Объект исследования – фейковые аккаунты. Методы, используемые в исследовательской работе: методы масштабирования, методы машинного обучения, методы выбора признаков, методы оценки моделей. Предметная область исследовательской работы – Instagram. Уникальностью исследования является определение наиболее надежного способа обнаружения фейковых аккаунтов с помощью различных, широко распространенных методов. Теоретическая и практическая значимость исследования заключается в выявлении фейковых аккаунтов, которые выдают себя за реальных пользователей, рассылают спам, собирают персональные данные, распространяют вредоносное ПО, фальсифицируют онлайн-рейтинги и т.д., в защите социальной сети от вредоносных действий, а также в разработке приложений для защиты и безопасности социальных сетей. Для выявления фейковой учетной записи эффективно использовались классификаторы машинного обучения: классификатор дерева решений; классификатор KNN; классификатор логистической регрессии; классификатор случайного леса; классификатор метода опорных векторов (SVM). Оценка производительности моделей проводилась с использованием различных показателей, таких как точность, ROC-кривая, классификационный балл. Некоторые модели, такие как Decision Tree, Random Forest и KNN, показали стабильные и высокие значения AUC-ROC, точности, прецизионности, полноты и F1-показателя как при первоначальном тестировании, так и после переобучения. Это подчеркивает важность выбора правильных функций при обучении модели и показывает, что не все функции одинаково полезны для всех моделей. В результате исследования, убедившись в эффективности использования алгоритмов машинного обучения с целью выявления фейковых аккаунтов, сравнив алгоритмы классификации в ходе исследования, мы определили алгоритм, достигший высокой точности.*

Ключевые слова: *машинное обучение, фэйк аккаунт, информационная безопасность, профиль пользователя, социальные сети, обнаружение поддельных учетных записей, Instagram, Decision Tree, KNN, Logistic Regression, SVM, Random Forest.*

Using Machine Learning to Identify a Fake Account

[1]***MUSSIRALIYEVA Shynar,*** *Cand. of Phys. and Math. Sci., Head of Department, mussiraliyevash@gmail.com,*
[1]****AZAMATOVA Dinara,*** *Master Student, dinarazamatova@gmail.com,*
[1]***BAISPAY Gulshat,*** *Postdoctoral Student, gulshat.bgb2@gmail.com,*
[1]NPJSC «Al-Farabi Kazakh National University», Kazakhstan, Almaty, Al-Farabi Avenue, 71,
**corresponding author.*

Abstract. *Nowadays, due to the rapid growth of social networks, the possibility of various threats is increasing, one of which is fake accounts. The use of machine learning in the issue of identifying fake accounts on a social network is considered. The main goal of the work is to research and develop methods and models of machine learning for the effective detection of fake accounts in social networks. This avoids problems such as extortion, blackmail, harassment, fake requests, insulting a person, unauthorized access to personal information, illegal use of user information through research. The object of the study is fake accounts. Methods used in the research work: scaling methods, machine learning methods, feature selection methods, model evaluation methods. The subject area of the research work is Instagram. The uniqueness of the study is to determine the most reliable way to detect fake accounts using various, widely used methods. Machine learning classifiers were effectively used to detect a fake account: decision tree classifier; classifier KNN; logistic regression classifier; random forest classifier; support vector machine (SVM) classifier. The performance of the models was evaluated using various indicators, such as accuracy, ROC-curve, classification score. Some models, such as Decision Tree, Random Forest, and KNN, have shown stable and high AUC-ROC, Accuracy, Precision, Recall, and F1 scores both during initial testing and after retraining. This highlights the importance of choosing the right features when training a model and shows that not all features are equally useful for all models. As a result of the study, after verifying the*

effectiveness of using machine learning algorithms to identify fake accounts, comparing classification algorithms during the study, we identified an algorithm that achieved high accuracy.

Keywords: *machine learning, fake account, information security, user profile, social networks, fake account detection, Instagram, Decision Tree, KNN, Logistic Regression, SVM, Random Forest.*

REFERENCES

1. Kate Bojkov. 29 Instagram Statistics For Marketing in 2023 // [Elektronnyi resurs] / URL: https://embedsocial.com/blog/instagram-statistics/
2. Boshmaf Yazan, Dionysios Logothetis, Georgos Siganos, Jorge Lería, Jose Lorenzo, Matei Ripeanu, and Konstantin Beznosov. «Integro: Leveraging Victim Prediction for Robust Fake Account Detection in OSNs». [Tekst] / In NDSS, vol. 15, pp. 8-11. – 2015.
3. Cresci Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, Maurizio Tesconi «Fame for sale: Efficient detection of fake Twitter followers». [Tekst] / Decision Support Systems 80, 2015.
4. Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallarok, Konrad Rieck «Dos and Don'ts of Machine Learning in Computer Security». [Tekst] / USENIX Security Symposium 2022, 2022.
5. Thomas K., Grier C., Song D., & Paxson, V. «Suspended accounts in retrospect: an analysis of twitter spam». [Tekst] / In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement (pp. 243-258). 2011.
6. Khaled Sarah, El-Tazi Neamat, Mokhtar Hoda «Detecting Fake Accounts on Social Media». [Tekst] / IEEE International Conference, 2018.
7. Dataset // [Elektronnyi-resurs] / URL: https://www.kaggle.com/datasets/free4ever1/instagram-fake-spammer-genuine-accounts?select=train.csv