DOI 10.52209/1609-1825_2023_3_453 ӘОЖ 351.86

Web-ресурстардағы экстремисттік мәліметтерді анықтауға арналған машиналық әдістерді оқыту және сынау үшін қазақ тіліндегі мәтіндер корпусын құру

[1]**МУСИРАЛИЕВА Шынар Женисбековна,** *ф.-м.ғ.к., кафедра меңгерушісі, mussiraliyevash@gmail.com,*
[1]***БАЙСПАЙ Гүлшат Болатқызы,** *постдокторант, gulshat.bgb2@gmail.com,*
[1]**БОЛАТБЕК Милана Асланбекқызы,** *PhD, bolatbek.milana@gmail.com,*
[1]**САҒЫНАЙ Мөлдір,** *докторант, sagynaymoldir11@gmail.com,*
[2]**ТЕРЕЙКОВСКИЙ Игорь Анатольевич,** *т.ғ.д., профессор, terejkowski@ukr.net,*
[1]*«Әл-Фараби атындағы Қазақ ұлттық университеті» КеАҚ, Қазақстан, Алматы, әл-Фараби даңғылы, 71,*
[2]*Игорь Сикорский атындағы Киев политехникалық институты, Украина, Киев, Берестейский даңғылы, 37,*
**автор-корреспондент.*

Аңдатпа. *Мақалада web-ресурстардағы экстремистік мәліметтерді анықтаудың семантикалық моделін құруға қажетті қазақ тілінде корпус құру мәселесі қарастырылады. Зерттеу нәтижесінде бұрын жасалған талдаушының көмегімен «ВКонтакте» әлеуметтік желісінің жабық топтарынан, «YouTube» комментарийлерінен және жаңалықтар сайттарынан мәтіндік деректер жиналды. Мәтіндерді екі сыныпқа «экстремистік» және «бейтарап» жіктеу жүргізілді, әрі қарай модельдерді оқыту үшін корпус құрылды. Корпусқа енгізілген экстремистік сипаттағы хабарламалардың жалпы саны 1200-ге жуық, ал корпустағы жалпы сөздер саны шамамен 140 000 сөзді құрайды. Экстремистік және бейтарап сипаттағы мәтіндердің таралуы, сөз бұлты арқылы корпусқа талдау жүргізілді. Деректерді талдау үшін pandas, numpy, matplotlib, plotly, bokeh, cufflinks, spacy, googletrans пакеттері бар Python 3.7 бағдарламалау тілі есептеу және визуализацияның негізгі кітапханалары ретінде қолданылды. Экстремистік мазмұндағы мәтіндердің ерекшеліктері анықталды.*

Кілт сөздер: *корпус, экстремисттік мәліметтерді анықтау, машиналық әдістерді оқыту, қазақ тіліндегі корпус, web-ресурстар, әлеуметтік жүйелер, ВКонтакте, парсер, семантикалық модель, сөздер бұлты.*

Кіріспе

Қазіргі таңда ақпараттық-коммуникациялық Интернет желісі адамзат өмірінің ажырамас бөлігіне айналды. Адамдар «Твиттер», «ВКонтакте», «Facebook» және т.с.с. әлеуметтік желілерді жаһандық байланыс орнату, пікір алмасу, білім алу және т.б. мақсаттарда пайдалануда. Экстремистік ұйымдар жаңа ақпараттық технологияларды қолдана отырып желіде топқа жаңа мүшелер тарту, экстремистік іс-әрекеттерді жоспарлау мен орындау, оқыту жұмыстарын жүргізу, әлеуметке қауіпті әрекеттерді басқару мен координациялауда жасырын ақпарат алмасу, экстремистік іс-әрекеттерді орындау үшін қаржыландыру көздерін іздеу, қолданушыларға, соның ішінде жастарға мақсатты түрде идеологиялық насихат жүргізу мақсатында жабық сайттар құру және т.б. әрекеттерді ұтымды орындауда [1].

Экстремизм идеяларының таралуына қарсы тұру үшін қазіргі уақытта құқық қорғау органдары қылмыстық заңнаманың «Экстремизмге қарсы іс-қимыл туралы Қазақстан Республикасының 2005 жылғы 18 ақпандағы № 31 Заңы» [2] нормасы, «Қазақстан Республикасында діни экстремизм мен терроризмге қарсы іс-қимыл жөніндегі 2018-2022 жылдарға арналған мемлекеттік бағдарламаны бекіту туралы», Қазақстан Республикасы Үкіметінің 2018 жылғы 15 наурыздағы № 124 қаулысы [3] пайдаланылады. Аталған мәселелерді Қазақстан Республикасының ұлттық қауіпсіздігіне төнетін қатер ретінде қарастыруға болады. Интернеттегі экстремистік іс-әрекеттерге қарсы іс-қимыл саласындағы ахуал күрделі болып қала береді, бұл ғылыми зерттеулер жүргізуді және экстремизмнің кез келген көріністерін анықтауға, алдын алуға және жолын кесуге бағытталған тиімді және уақытылы шараларды кешенін жүзеге асыруды қажет етеді.

Зерттеулерге шолу

Google, Facebook және Twitter алпауыттары интернеттегі террористік мазмұнды жыл-

дам анықтап, жою үшін жасанды интеллект (AI) технологиясын қолдануға уәде берді [4]. IBM-де жоғарыда аталған әлеуметтік желілердегі барлық деректерді талдай алатын Watson әзірлемесі бар [5]. Ресейде Платонның IT-авторы әлеуметтік желілерді бақылау және қауіп-қатерлерді болжау жүйесін құруда [6]. Ғалымдар web-ресурстардағы экстремистік бағыттарды анықтау үшін машиналық және терең оқыту әдістерін тиімді пайдалануда [7]. Соңғы уақытта web-ресурстарда неміс [8], орыс [9], араб [10] тілдерінде жазылған экстремистік мазмұнды қамтитын мәтіндерді анықтауға арналған әр түрлі жүйелер құрылу үстінде. [11] жұмыста әлеуметтік желі мен микроблогтардан анықталған экстремистік мәтіндердің лингвистикалық ерекшеліктерін анықтауға қатысты зерттеулер жүргізілген.

Қазақ тілі үшін қылмыстық сипаттағы мәтіндерді анықтауға және талдауға арналған бірқатар ғылыми зерттеулер жүргізілген. Атап айтатын болсақ, Шәріпбай А.Ә. бастауымен «Жасанды интеллект» ғылыми-зерттеу институтының ғалымдарының [12]. [13] жұмыста террористік қауіптерге байланысты қазақ тіліндегі мәтіндердің көңіл-күйін талдау үшін сөздікті қолдана отырып, ережеге негізделген әдісті сипаттайды. Жоғарыда келтірілген тұжырымдарды ескере отырып, web-ресурстардағы қазақ тіліндегі экстремистік бағыттағы мәтіндерді анықтауға арналған модельдерді құру тапсырмасы аса өзекті деген қорытындыға келуге болады.

Зерттеу әдістері мен нәтижелері

Экстремистік бағытты (ЭБ) анықтау үшін web-контентті жинауға және талдауға арналған бағдарламалық жабдықтама құрастыру

Мәліметтерді алуға дайындық кезеңі. Web-ресурстардағы экстремистік мәтіндерді анықтау мақсатында семантикалық талдау модельдерін құру жұмысын жүргізу үшін ең алдымен үлкен көлемдегі корпус қажет болады. Аталған корпус мәтіндері машиналық оқыту алгоритмдерін үйрету және тестілеу кезеңдерінде қолданылады. Бастапқыда ашық қол жетімді ресурстардағы экстремисттік мәтіндер іздестірілді. Қазақ тіліндегі экстремистік мәтіндер негізінен жаңалық порталдарындағы, «YouTube», «ВКонтакте» әлеуметтік желілеріндегі комментарийлер арасынан табылды. Алайда мұндай әдіспен табылған экстремистік мәтіндер машиналық оқыту жүйелерін үйрету үшін жеткіліксіз болды. Ашық дереккөздерден шамамен 4400 сөзді қамтитын 400-ге жуық қазақ тіліндегі экстремистік мазмұндағы мәтіндер табылып, корпусқа енгізілді.

Экстремистік мәтіндерді қамтитын корпусты кеңейту үшін қазақ тілді аудитория арасында белсенді қолданылатын әлеуметтік желі түрін таңдау тапсырмасы қойылды. «ВКонтакте», «Facebook», «Twitter» әлеуметтік желілеріндегі мәтіндерге талдау жасалып, нәтижесінде «ВКонтакте» әлеуметтік желісі таңдалды. Келесі кезекте парсер модулінің кірісіне мәтіндері жүктелуі керек болатын топтардың тізімдерін беру қажет. Әлеуметтік желідегі парсинг жасалатын топтар тізімі ашық дереккөздердегі экстремистік мәтіндерге TF-IDF әдісін қолдану арқылы анықталған кілттік сөздер көмегімен құрылды. Атап айтатын болсақ, әлеуметтік желі топтарын анықтау үшін «соғыс», «жиһад», «джихад», «шам», «сирия» және т.б. сөздер қолданылды. «Регион» өрісінде «Казахстан» таңдалды. Зерттеу жұмысында корпусқа экстремистік мәтіндермен қатар бейтарап мазмұндағы мәтіндерді жинау да жоспарланды.

Іздеу нәтижесінде Қазақстан Республикасының территориясында таратуға тыйым салынған 76 топ анықталды. Алайда кей топтарға мүлдем парсинг жасау мүмкін болмады. Ал кей топтардың ішінде экстремистік мазмұндағы мәтіндер табылмады. Діни мазмұндағы топтар да жоғарыда сипатталған әдіс бойынша анықталды. Іздеу нәтижесінде 433 топ табылды. Анықталған топтар ішінен мәтіні жоқ, тек суреттерді қамтитын, мәтіні басқа тілдегі, атаулары қайталанған топтар өшірілді. Нәтижесінде корпусқа енгізу мақсатында 265 топ таңдалды. Жалпы лексикадағы топтар да жоғарыдағы әдіс бойынша анықталды. «ВКонтакте» әлеуметтік желісіндегі 100 топ таңдалды.

Келесі кезекте Қазақстан Республикасында орын алған экстремистік іс-әрекеттер жайлы жаңалық мақалалары жинақталды. Жаңалықтар порталдарға кіріп, «жиһад», «Сирия», «экстремизм» және т.б. сөздерді енгізу арқылы іздестірілді.

Мәліметтерді жинау. Ақпаратты экстремистік санатқа жатқызбас бұрын, «қауіп» критерийін анықтау қажет. Шешімдердің бірі – кілт сөздер жиынтығын анықтау. Әзірленген бағдарламалық кешенде ақпарат түрлерін анықтаудың дәл осы әдісі қолданылды. Анықтау үшін «ВКонтакте» әлеуметтік желісіндегі ақпаратты талдау үшін қолданылатын кілт сөздер жиынтығы жасалды. Бағдарламалық кешен мәтінде көрсетілген кілт сөздердің болуы немесе болмауы негізінде бұл мәтін одан әрі зерттеу үшін жарамды деген қорытынды жасайды. 1-суретте хабарламалар туралы деректерді жинау, талдау және жіктеудің сұлбасы көрсетілген.

Максималды өнімділікке қол жеткізу үшін, ақпарат көздерінен (API) ақпарат алудың кіріктірілген әдістерін қолдану қажет. Егер мұндай әдістер болмаса, HTTP сұраныстары арқылы қажетті ақпаратты алу қажет.

Бағдарламалық кешен үш бөлек модульден тұрады:

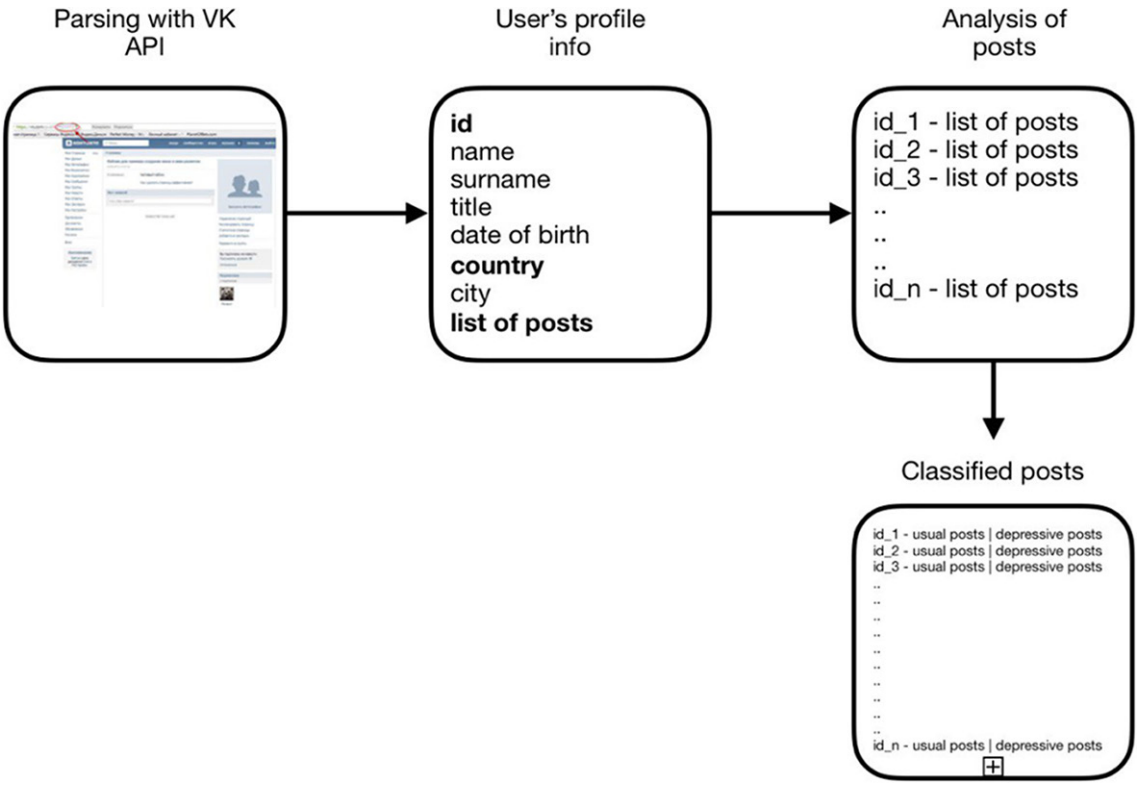
1) Ақпарат жинау модулі – ашық көздерден ақпаратты қабылдауға және оны әрі қарай өңдеуге беруге жауап береді; «ВКонтакте» әлеуметтік желісінен деректерді талдау үшін Python бағдарламалау тілі қолданылды. Ресми VK API қолдана отырып, қазақстандық профильдер ішінара талданды, сондай-ақ деректер Solr дерекқорында сақталды.

2) Кілт сөздерді іздеу модулі – ақпараттың үлкен көлемінің ішінен кілт сөздерді іздеуге жауап береді.

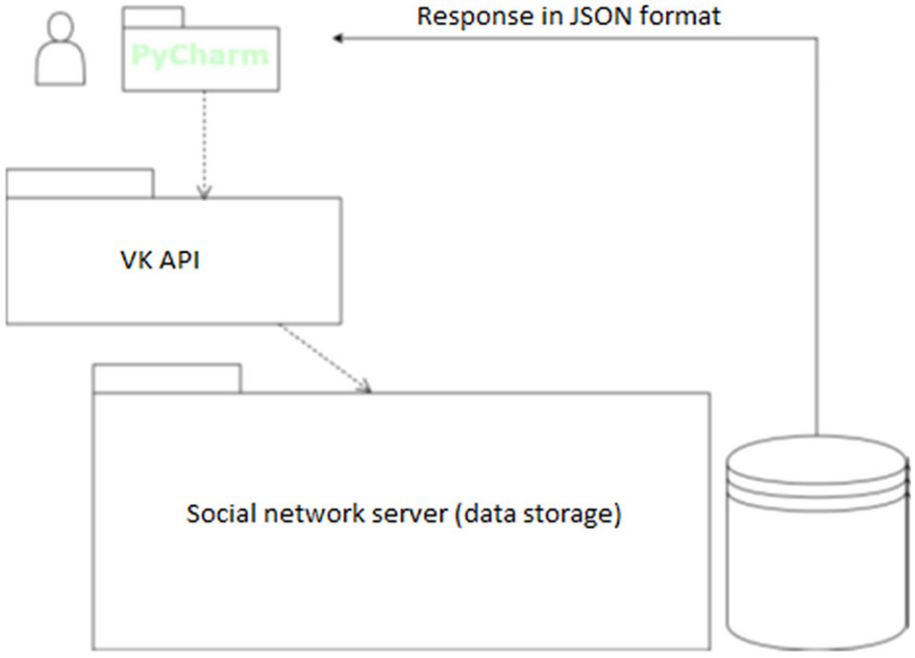
3) Құжаттарды саралау модулі – ақпараттың қауіпті екендігін анықтауға жауап береді. Құжаттарды қауіптілік дәрежесі бойынша саралау үшін Long Short Term Memory (LSTM) терең оқыту алгоритмі қолданылды.

Деректерді жинау үшін Python 3.7 қолданылады [14]. 2-суретте деректерді жинау процесінің сұлбасы көрсетілген.

Деректерді талдау үшін pandas, numpy, matplotlib, plotly, bokeh, cufflinks, spacy, googletrans пакеттері бар Python 3.7 бағдарламалау тілі есеп-



1-сурет – Деректерді жинау, талдау және жіктеу сұлбасы



2-сурет – Деректерді жинау сұлбасы

теу және визуализацияның негізгі кітапханалары ретінде қолданылды. Тиісті мәтіндерді іздеу үшін экстремизммен байланысты кілт сөздер анықталды. Мысалы, «кафир», «өлтіру», «жару» және т.б. бұл кілт сөздер әлеуметтік желілердегі экстремистік посттарды табуға көмектеседі. Деректердің аннотациясы үшін «ВКонтакте» әлеуметтік желісінен экстремистік идеялардың мәтіндері жиналды және олардың дұрыс таңбаланғанына көз жеткізу үшін барлық хабарламалар қолмен тексерілді.

Мәліметтерді жинау үшін парсер құрастыру.

Берілген зерттеу үшін үлкен көлемдегі деректер қажет. Қазақ тілінде дайын экстремистік корпус болмағандықтан, «ВКонтакте», «YouTube», «Твиттер» және елдің жаңалықтар порталдары сияқты ашық ақпарат көздерінен мәлімет жинақтау үшін әр түрлі дереккөздерден ақпарат жинайтын парсер жазылды. Парсердің нәтижесін көрсету үшін шағын web-қосымша құрылды. 3-суретте әртүрлі ашық көздерден деректерді жинауға арналған парсер бағдарламасы көрсетілген.

Нәтижесінде, әзірленген парсер көмегімен экстремистік бағыттағы мәліметтер жиналды. Жіктеу модельдерін одан әрі оқыту үшін корпус әзірленді, ол «экстремистік» және «бейтарап» бағыт ретіндегі екі класстан тұрады.

Web-ресурстардағы экстремистік мәліметтерді анықтаудың семантикалық моделін құруға қажетті корпус құру. Корпусқа енгізілген экстремистік сипаттағы хабарламалардың жалпы саны 1200-ге жуық, ал корпустағы жалпы сөздер саны шамамен 140000 сөзді құрайды. Құрастырылған корпус .csv форматындағы құжат түрінде сақталған. Экстремистік мәтіндерді қамтитын құжат 4 бағаннан тұрады: хабарламаның реттік нөмірі, хабарлама, хабарламаның препроцессингтен өтіп, тазартылған нұсқасы және қай класқа тиісті жазба екендігін білдіретін «1/0» атрибуты (4-сурет).

4-суретте келтірілген қазақ тіліндегі экстремистік мәтіндерге талдау жасалып, оларға тән кейбір ерекшеліктер анықталды:

1) Қазақ тілінің төл әріптерінің кирилл әріптерімен алмастырылуы (мысалы, «соғысқа» сөзінің орнына «согыска», «өлтіру» сөзінің орнына «олтиру» және т.б. деп жазу).

2) Биграммалардың жиі кездесуі (мысалы, «жиһад жасау», «Шамға бару», «көпірлерді қыру» және т.б.).

3) Араб тіліндегі діни терминдердің жиі кездесуі (мысалы, «фард кефайр», «даулатуль ислам» және т.б.).

Корпус талдауы.

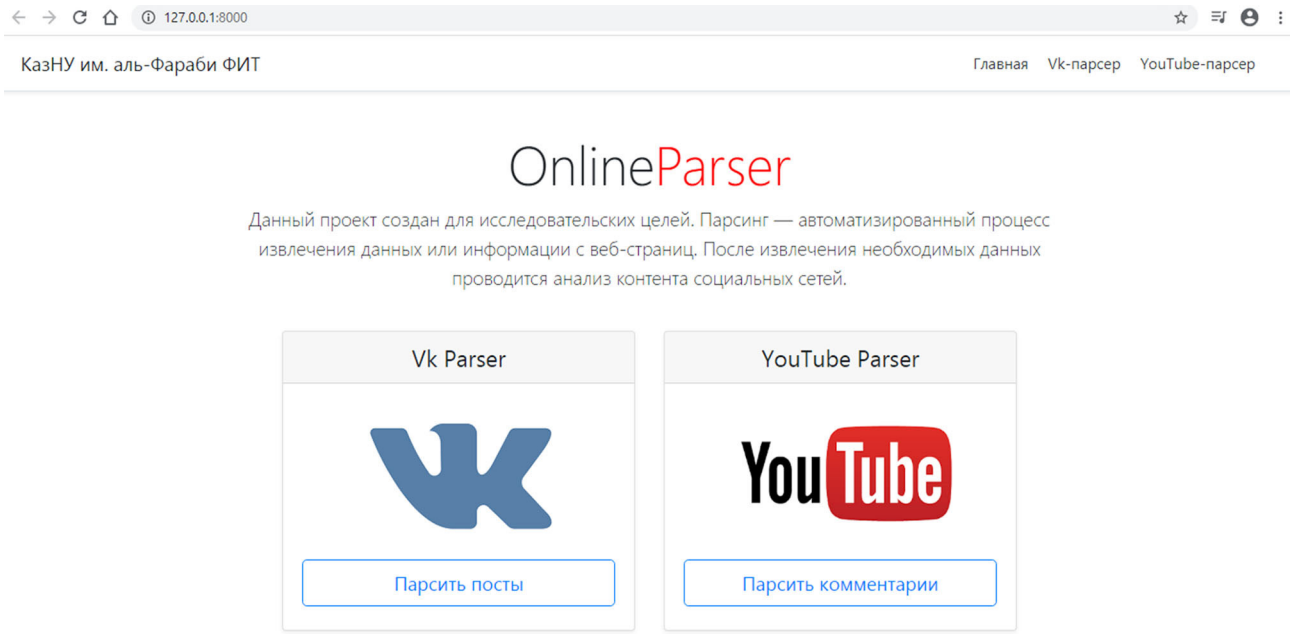
5-суретте деректер жиынтығындағы экстремистік және бейтарап мәтіндердің үлестірімі көрсетілген.

Сөздер бұлты. Деректерді визуалды түрде көрсету үшін сөз бұлттары қолданылды. Ықтимал экстремистік идеялары бар пайдаланушылардың хабарламалары 6-суретте бөлек көрсетілген.

Экстремистік посттар пайдаланушылардың экстремистік ойларына тікелей нұсқау бере отырып, «қарсы» (қарсы), «әскери» (әскери) сияқты сөздерді жиі қолданады.

Қорытынды

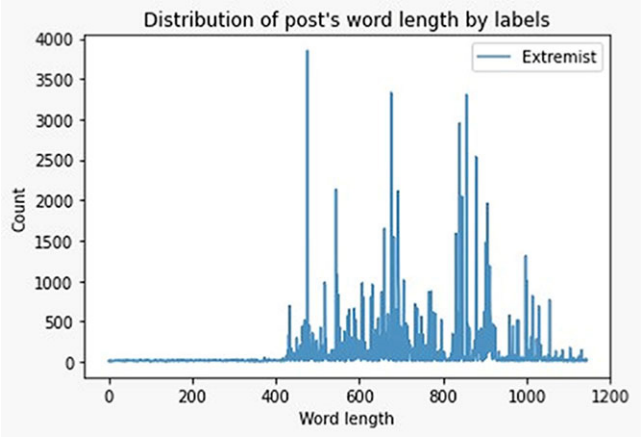
Осы мақаланың нәтижесінде а) бұрын жасалған талдаушының көмегімен «ВКонтакте» әлеуметтік желісінің жабық топтарынан, «YouTube» комментарийлерінен және жаңалықтар сайттарынан мәтіндік деректер жиналды; б) мәтіндерді



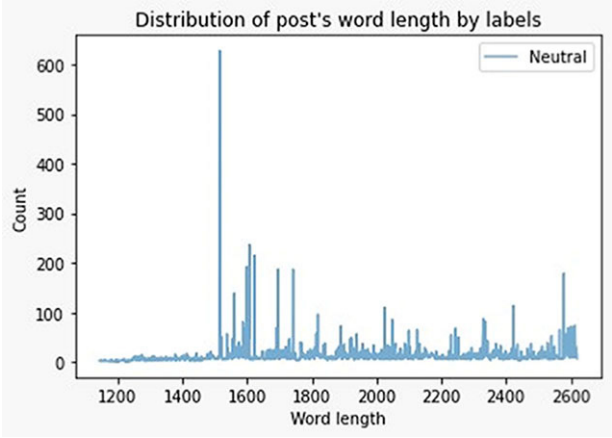
3-сурет – Зерттеу жұмысы барысында құрастырылған парсер бағдарламалары

1.	Лива Туввар Сирия АлЛажатта # IslamState-қа қарсы ұрыс бастайды https://t.co/Fm1iTwk0Z8	1
2.	террористік иттер бүлікшілерге айналады! #Turkey #PKK #TwitterKurds https: // t…	1
3.	Әл-Хайрдың Газваты 13 арнайы аймақ. Олардан кейін олар Алеппеге қайтып оралады.	1
4.	Сирия-дағы # Ресей | n 'басып кіру' туралы алаңдамаудың кіруі мүмкін емес. Меккедегі бір жұма туралы «Қорғаныс джихад» деп жарияланады ...	1
5.	Ядролық соғыс соғысы келе жатыр !!! https://t.co/v73GmveabI	1
6.	Давлатул ислам! Баақия! Ансарул Халифа - Оңтүстік-Шығыс Азиядағы барлық мужахидтер топтарын жақсы көремін деп сенемін. https://t.co/KaVLDresnk	1
7.	Орыстар, сириялықтар мен ирандықтар Бағдадта әскери үйлестіру камерасын құруда http://t.co/HYabzUFqZC	1
8.	@WilayatNinawa сізге xpose-ді жауларға тапсырма аласыз!	1
9.	Фаллужаның #IS мандаты Фаллужаның солтүстік-шығысында (12,5) мм қару-жарақпен Сафави әскеріне тиесілі Хамвейде #IRAQ.	1
10.	URGENT | #HOMS IS сириялық режимді күшейтеді, 4 сарбазды өлтіруге мүмкіндік береді (Хомс-Пальмира)	1
11.	# Анбар ИСМ басқарылды 2 өлтіру 23 ирак армиясы, Аль-Багдади базасында d әскери жолмен 5 күш жұмсалады	1
12.	Бұл жол # Сириядан емес .. # Фаллужадан #Iraq | мен армияны бомбалау. #ISIS http://t.co/Qte40bIUE1	1
13.	Жаңа таңертең Тел-наам ауданында Нусайристе 10 тонна бомба (халаб) бар.	1
14.	Бұл твит (жүктеме #RISia #ISIS бомбалау емес) төңкерушілерге арналады .. Жаңа	1

4-сурет – Экстремистік сипаттағы хабарламалар мысалы

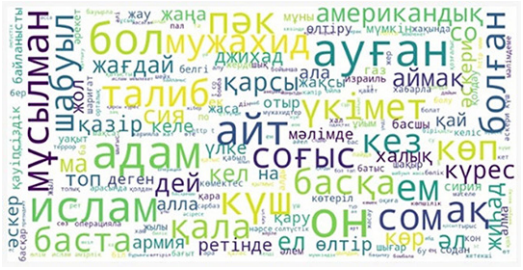


а) Экстремистік хабарламалардың үлестірім графигі



б) Бейтарап мәтіндердің үлестірім графигі

5-сурет – Корпустағы экстремистік және бейтарап мәтіндердің ұзындықтары бойынша үлестірім кестесі



а) Экстремизмге байланысты хабарламалардың сөздер бұлты

б) Экстремизммен байланыссыз хабарламалардың сөздер бұлты

6-сурет – Экстремистік және бейтарап мәтіндер корпусын визуализациялау

екі сыныпқа «экстремистік» және «бейтарап» жіктеу жүргізілді; в) әрі қарай модельдерді оқыту үшін корпус құрылды; г) экстремистік және бейтарап сипаттағы мәтіндердің таралуы, сөз бұлты арқылы корпусқа талдау жүргізілді және экстремистік мазмұндағы мәтіндердің ерекшеліктері анықталды.

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. Gaikwad M., Ahirrao S., Phansalkar S., Kotecha K. Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools // IEEE Access. – 2021. – Vol. 9. – Pp. 48364-48404.
2. Қазақстан Республикасының № 31 Заңы. Экстремизмге қарсы іс-қимыл туралы: 2005 жылдың 18 ақпанында бекітілген. https://adilet.zan.kz/kaz/docs/Z050000031_
3. Қазақстан Республикасы Үкіметінің № 124 қаулысы. Қазақстан Республикасында діни экстремизм мен терроризмге қарсы іс-қимыл жөніндегі 2018-2022 жылдарға арналған мемлекеттік бағдарламаны бекіту туралы: 2018 жылдың 15 наурызында бекітілген. https://adilet.zan.kz/kaz/archive/docs/P1800000124/15.03.2018
4. Addressing the abuse of tech to spread terrorist and extremist content. https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c. 20.06.2021.
5. IBM Watson is AI for smarter business. https://www.ibm.com/watson. 20.02.2022.
6. IT-соавтор «Платона» создает систему мониторинга соцсетей и предсказания угроз. https://www.vedomosti.ru/technology/articles/2016/06/17/645694-it-soavtor-platona-sozdaet-sistemu-monitoringa-sotssetei-predskazaniya-ugroz. 18.09.2021.
7. Soliman G.M.A., Abou-El-Enien T.H.M. Terrorism prediction using artificial neural network // Revue d'Intelligence Artificielle. – 2019. – Vol. 33, no. 2. – Pp. 81-87.
8. Koehler D. Recent Trends in German Right-Wing Violence and Terrorism: What Are the Contextual Factors behind 'Hive Terrorism'? // Perspectives on Terrorism. – 2018. – Vol. 12, no. 6. – Pp. 72-88.
9. Deviatkin D., Smirnov I., Solovyev F., Suvorova M., Chepovskiy A. Extremist Text Detection In Social Web // Multi Conference on Computer Science and Information Systems, MCCSIS 2019. – Porto, 2019. – Pp. 344-350.
10. AlGhamdi M.A., Khan M.A. Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages // Arabian Journal for Science and Engineering. – 2020. – Vol. 45, no. 8. – Pp. 6021-6032 (2020).
11. Ul Rehman Z., Abbas S., Khan M.A., Mustafa G., Fayyaz H., Hanif M., Saeed M.A. Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning // Computers, Materials and Continua. – 2020. – Vol. 66, no. 2. – Pp. 1075-1090.
12. Aktayeva A., Niyazova R., Muradilova G., Makatov Y., Kusainova U. Cognitive Computing Cybersecurity: Social Network Analysis // Communications in Computer and Information Science. – 2020. – Vol. 1140. – Pp. 28-43.
13. Bekmanova G., Yelibayeva G., Aubakirova S., Dyussupova N., Sharipbay A., Nyazova R. Methods for Analyzing Polarity of the Kazakh Texts Related to the Terrorist Threats // Computational Science and Its Applications – ICCSA 2019 – 19th International Conference. – Saint Petersburg: Springer, 2019. – Pp. 717-730.
14. Python 3.7.0. https://www.python.org/downloads/release/python-370/. 03.03.2022.

Формирование текстового корпуса на казахском языке для обучения и тестирования машинных методов выявления экстремистской информации в веб-ресурсах

*[1]**МУСИРАЛИЕВА Шынар Женисбековна,** к.ф.-м.н., зав. кафедрой, mussiraliyevash@gmail.com,*
*[1]***БАЙСПАЙ Гүлшат Болатқызы,** постдокторант, gulshat.bgb2@gmail.com,*
*[1]**БОЛАТБЕК Милана Асланбекқызы,** PhD, bolatbek.milana@gmail.com,*
*[1]**САҒЫНАЙ Мөлдір,** докторант, sagynaymoldir11@gmail.com,*
*[2]**ТЕРЕЙКОВСКИЙ Игорь Анатольевич,** д.т.н., профессор, terejkowski@ukr.net,*
[1]НАО «Казахский национальный университет имени аль-Фараби», Казахстан, Алматы, пр. аль-Фараби, 71,
[2]Киевский политехнический институт имени Игоря Сикорского, Украина, Киев, пр. Берестейский, 37,
**автор-корреспондент.*

***Аннотация.** Рассматривается вопрос создания корпуса на казахском языке, который необходим для создания семантической модели выявления экстремистской информации на веб-ресурсах. В результате исследования были собраны текстовые данные из закрытых групп социальной сети «ВКонтакте», комментариев «YouTube» и новостных сайтов с помощью ранее разработанной аналитики. Тексты были классифицированы на два класса: «экстремистские» и «нейтральные». Был создан корпус для обучения моделей. Общее количество экстремистских сообщений, включенных в корпус, составляет около 1200, а общее количество слов в корпусе составляет около 140 000 слов. Проведено распределение экстремистских и нейтральных текстов, анализ корпуса осуществлялся с использованием облака слов. В качестве основных библиотек расчета и визуализации для анализа данных использовался язык программирования Python 3.7 с пакетами pandas, numpy,*

matplotlib, plotly, bokeh, cufflinks, spacy, googletrans. Выявлены признаки текстов экстремистского содержания.

Ключевые слова: *корпус, выявление экстремистских текстов, методы машинного обучение, корпус на казахском языке, веб-ресурсы, социальные системы, ВКонтакте, парсер, семантическая модель, облако слов.*

Formation of a Text Corpora in the Kazakh Language for Training and Testing Machine Methods for Detecting Extremist Information in Web Resources

[1]***MUSSIRALIYEVA Shynar,*** *Cand. of Phys. and Math. Sci., Head of Department, mussiraliyevash@gmail.com,*
[1]****BAISPAY Gulshat,*** *Postdoctoral Fellow, gulshat.bgb2@gmail.com,*
[1]***BOLATBEK Milana,*** *PhD, bolatbek.milana@gmail.com,*
[1]***SAGYNAY Moldir,*** *Doctoral Student, sagynaymoldir11@gmail.com,*
[2]***TEREIKOVSKYI Ihor,*** *Dr. of Tech. Sci., Professor, terejkowski@ukr.net,*
[1]*NPJSC «Al-Farabi Kazakh National University», Kazakhstan, Almaty, Al-Farabi Avenue, 71,*
[2]*Igor Sikorsky Kyiv Polytechnic Institute, Ukraine, Kyiv, Beresteisky Avenue, 37,*
**corresponding author.*

Abstract. *The article deals with the issue of creating a corpus in the Kazakh language, which is necessary to create a semantic model for identifying extremist information on web resources. As a result of the study, text data was collected from closed groups of the VKontakte social network, YouTube comments and news sites using previously developed analytics. The texts were classified into two classes: «extremist» and «neutral». A corpus for further training of models was created. The total number of extremist messages included in the corpus is about 1,200, and the total number of words in the corpus is about 140,000 words. The distribution of extremist and neutral texts was carried out, the analysis of the corpus was carried out using a word cloud. The Python 3.7 programming language with pandas, numpy, matplotlib, plotly, bokeh, cufflinks, spacy, googletrans packages was used as the main calculation and visualization libraries for data analysis. Signs of texts of extremist content were identified.*

Keywords: *corpus, detection of extremist texts, machine learning methods, corpus in Kazakh language, web resources, social networks, VKontakte, parser, semantic model, word cloud.*

REFERENCES

1. Gaikwad M., Ahirrao S., Phansalkar S., Kotecha K. Online Extremism Detection: A Systematic Literature Review with Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools // IEEE Access. – 2021. – Vol. 9. – Pp. 48364-48404.
2. Kazakhstan Respublikasynyn No. 31 Zany. Extremism karsy is-kymyl turaly: 2005 zhyldyn 18 akpanynda bekitilgen. https://adilet.zan.kz/kaz/docs/Z050000031_
3. Kazakhstan Republics Ykimetinin No. 124 kaulysy. Kazakhstan Respublikasynda day extremism men terrorism karsy is-qimyl zhonindegi 2018-2022 https://adilet.zan.kz/kaz/archive/docs/P1800000124/03/15/2018
4. Addressing the abuse of tech to spread terrorist and extremist content. https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c. 06/20/2022.
5. IBM Watson is AI for smarter business. https://www.ibm.com/watson. 02/20/2023.
6. IT co-author of Platon creates a system for monitoring social networks and predicting threats. https://www.vedomosti.ru/technology/articles/2016/06/17/645694-it-soavtor-platona-sozdaet-sistemu-monitoringa-sotssetei-predskazaniya-ugroz. 09/18/2021.
7. Soliman G.M.A., Abou-El-Enien T.H.M. Terrorism prediction using artificial neural network // Revue d'Intelligence Artificielle. – 2019. – Vol. 33, no. 2. – Pp. 81-87.
8. Koehler D. Recent Trends in German Right-Wing Violence and Terrorism: What Are the Contextual Factors behind 'Hive Terrorism'? // Perspectives on Terrorism. – 2018. – Vol. 12, no. 6. – Pp. 72-88.
9. Deviatkin D., Smirnov I., Solovyev F., Suvorova M., Chepovskiy A. Extremist Text Detection In Social Web // Multi Conference on Computer Science and Information Systems, MCCSIS 2019. – Porto, 2019. – Pp. 344-350.
10. AlGhamdi M.A., Khan M.A. Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages // Arabian Journal for Science and Engineering. – 2020. – Vol. 45, no. 8. – Pp. 6021-6032 (2020).
11. Ul Rehman Z., Abbas S., Khan M.A., Mustafa G., Fayyaz H., Hanif M., Saeed M.A. Understanding the language of ISIS: An empirical approach to detect radical content on twitter using machine learning // Computers, Materials and Continua. – 2020. – Vol. 66, no. 2. – Pp. 1075-1090.
12. Aktayeva A., Niyazova R., Muradilova G., Makatov Y., Kusainova U. Cognitive Computing Cybersecurity: Social Network Analysis // Communications in Computer and Information Science. – 2020. – Vol. 1140. – Pp. 28-43.
13. Bekmanova G., Yelibayeva G., Aubakirova S., Dyussupova N., Sharipbay A., Nyazova R. Methods for Analyzing Polarity of the Kazakh Texts Related to the Terrorist Threats // Computational Science and Its Applications – ICCSA 2019 – 19th International Conference. – Saint Petersburg: Springer, 2019. – Pp. 717-730.
14. Python 3.7.0. https://www.python.org/downloads/release/python-370/. 03/03/2022.