ҚАЗАҚСТАН РЕСПУБЛИКАСЫ БІЛІМ ЖӘНЕ ҒЫЛЫМ МИНИСТРЛІГІ

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РЕСПУБЛИКИ КАЗАХСТАН

ҚазҰТЗУ ХАБАРШЫСЫ

ВЕСТНИК КазНИТУ

VESTNIK KazNRTU

№ 6 (142)

АЛМАТЫ 2020 НОЯБРЬ

ISSN *2709-4766* (Online)
ISSN 2709-4758 (Print)

Учредитель:

Казахский национальный исследовательский технический университет
имени К.И. Сатпаева

Регистрация:

Министерство культуры, информации и общественного согласия
Республики Казахстан № 951 – Ж "25" 11. 1999 г.

Основан в августе 1994 г. Выходит 6 раз в год

Адрес редакции:

г. Алматы, ул. Сатпаева, 22,
каб. 607, тел. 292-63-46
Nina. Fedorovna. 52 @ mail.ru

Өздеріңіз білетіндей, өткізгіштігі бойынша біртекті емес бірнеше қабаттарға суды бірлесіп айдау кен орындарының тез сулануына әкеледі, ал майданның жылжуын жеделдету мұнай мен судың өткізгіштігі жоғары қабаттарға ығысуы болып табылады. Бұл жұмыста бір уақытта-бөлек айдау технологиясының тиімділігіне талдау жасалды, сонымен қатар жер қабатының мұнай өнімділігін арттыруға мүмкіндік беретін шаралар кешені бойынша ұсыныстар берілді.

Кілт сөздер. Қабаттық қысымды, пакер, қабаттардың коллекторлық қасиеттері, қабаттың өткізгіштігін сақтау.

ОӘЖ 351.76

Sh. Mussiraliyeva, M. Bolatbek, B. Ziyat
(al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail:mussiraliyevash@gmail.com)

INCREASING THE ACCURACY OF CLASSIFICATION OF EXTREMIST TEXTS THROUGH STEMMING ALGORITHM

Abstract. Currently, various extremist organizations are actively using social networks for their activities. Therefore, it is important to create programs that require the implementation of a set of effective measures aimed at identifying, preventing and combating extremism. In this paper, the authors propose to use a stemming algorithm for corpus texts designed to increase the accuracy of the classification of extremist texts in the Kazakh language.

Keywords: extremism detection, social media, text classification, stemming algorithm, cybersecurity.

Ш.Ж. Мусиралиева, М. Болатбек, Б. Зият
(Әл-Фараби атындағы Қазақ ұлттық университеті
e-mail:mussiraliyevash@gmail.com, bolatbek.milana@gmail.com, ziyat.bekbol@gmail.com)

СТЕММИНГ АЛГОРИТМІ АРҚЫЛЫ ЭКСТРЕМИСТІК МӘТІНДЕРДІ ЖІКТЕУ ДӘЛДІГІН АРТТЫРУ

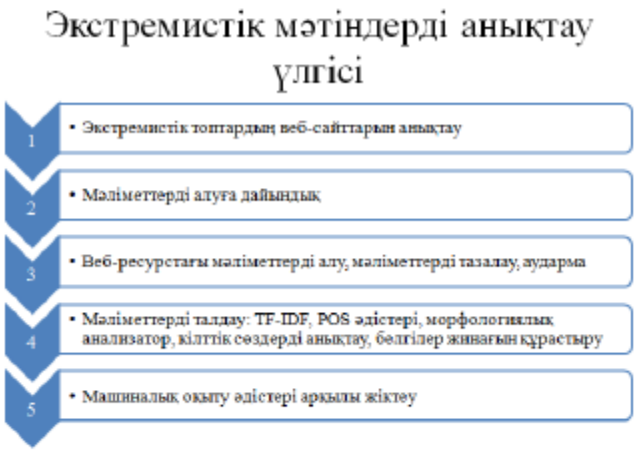
Аннотация. Қазіргі таңда әр түрлі экстремистік ұйымдар әлеуметтік желілерді өз қызметтері үшін белсенді пайдалануда. Сол себепті экстремизм көріністерін анықтауға, алдын алуға және жолын кесуге бағытталған тиімді шаралар кешенін іске асыруды қажет ететін бағдарламаларды құру өзекті болып табылады. Бұл жұмыста авторлар қазақ тіліндегі экстремистік мәтіндерді жіктеу дәлдігін арттыру мақсатында құрастырылған корпус мәтіндеріне стемминг алгоритмін қолдануды ұсынады.

Кілтсөздер: экстремизмді анықтау, әлеуметтік желі, мәтінді жіктеу, стемминг алгоритмі, киберқауіпсізік.

Кіріспе

Қазіргі таңда халықаралық ақпараттық-коммуникациялық Интернет желісі экстремистік материалдарды таратуда белсенді қолданылады. Бұл жаһандық саяси процестің негізгі қатысушыларының бірі ретінде Қазақстан Республикасы үшін өте маңызды болып табылады. Ғаламтор алпауыттары Google, Facebook және Twitter ланкестік мазмұнды табу және жою үшін жасанды интеллект технологиясын қолдануда. IBM-де әлеуметтік желілердегі барлық деректерді талдайтын Watson бағдарламасы бар. Ресейде Платонның ақпараттық серіктес авторы әлеуметтік желілерді бақылау және қауіптерді болжау жүйесін құрастырды. Германия үкіметі террористік шабуылдардан кейін Интернетте террористермен күресу үшін ZITiS деп аталатын жаңа киберқауіпсіздік бөлімшесін құру туралы жариялады. Қазақстанда мұндай жүйе жоқ. Сол себепті экстремизм көріністерін анықтауға, алдын алуға және жолын кесуге бағытталған тиімді шаралар кешенін іске асыруды қажет ететін бағдарламаларды құру өзекті болып табылады.

Берілген мақала веб-ресурстардағы экстремистік мәтіндерді анықтау мақсатында семантикалық үлгілер құруға қатысты зерттеу жұмысының бір бөлігі болып табылады. Бұл жұмыста авторлар қазақ тіліндегі экстремистік мәтіндерді жіктеу дәлдігін арттыру мақсатында құрастырылған корпус мәтіндеріне стемминг алгоритмін қолдануды ұсынады. Зерттеу жұмысының алдыңғы кезеңдерінде ашық ресурстардағы мәтіндер жинақталып, TF-IDF әдісі арқылы кілттік сөздер анықталған және құрастырылған корпус бойынша машиналық оқыту әдістері арқылы кіріс мәтінді экстремистік және бейтарап санаттарға жіктеу жүргізілген. 1-суретте веб-ресурстардағы экстремистік мәтінді анықтау алгоритмі келтірілген.



1-сурет. Веб-ресурстардағы экстремистік мәтінді анықтау алгоритмі

Берілген мақала мәліметтерді талдау кезеңіндегі алгоритмдердің бірі — стеммингке, яғни кіріс мәтіндегі сөздердің негіздері мен қосымшаларын автоматты түрде анықтау және ажырату тапсырмасына арналады. Лингвистикада сөздің негізі дегеніміз — түбірге жұрнақ жалғану арқылы жаңа лексикалық номинативтік мағына білдіретін сөздің кіші бөлшегі, сондай-ақ құрамы әрі қарай бөлшектеуге келмейтін жалаң түбірлер де сөздің негізі болады: *ат, от, тас* [1].

Стеммингалгоритмін іске асыруға арналған бірнеше әдістер бар және олардың көбісі бастапқы тілге тәуелді болып келеді. Кіріс мәтінге талдау жасау барысында стеммингалгоритмінің орындалуы жіктеу дәлдігін айтарлықтай жоғарылатуға септігін тигізуі мүмкін, мысалы стеммингалгоритмі орындалмаған жағдайда жіктеу жүйесі мәтіндегі "соғысқа", "соғыстың", "соғыста", "соғыс" сөздерін әр түрлі сөз ретінде таниды, ал аталған сөздердің қосымшаларын қарастырмай, стеммингалгоритмін орындап, сөз негіздерін алатын болсақ, онда қосымшалардың барлығы алынып тасталатындықтан, негізі бір сөздердің барлығы бір сөз ретінде анықталатын болады. Сондай-ақ, стеммингалгоритмін орындамаған жағдайда мәліметтер қорына сөздіктегі сөздердің барлық морфологиялық нұсқасын енгізу қажет болады, ал бұл өте үлкен жадыны қажет етеді және сәйкесінше жүйенің жұмысын айтарлықтай баяулатады. Бұл жұмыста қазақ тіліндегі мәтінге стеммингалгоритмін орындау әдісі көрсетіледі, атап айтатын болсақ, стеммингалгоритмін орындау арқылы экстремистік мәтіндерді анықтау дәлдігін арттыруға қадам жасалады.

Әдебиеттерге шолу

Мәтіндегі сөздердің негізін табуға арналған бірнеше дайын алгоритмдер бар. Солардың бірі - 1980 жылы ағылшын тілі үшін ұсынылған Портер стеммері. Мартин Портер жариялаған стеммингалгоритмі [2] сөздер негіздерінің қорын пайдаланбайды, оның орнына сөздердің аяқталу ережелері мен суффикстерді тізбектеп қолдану арқылы жұмыс істейді. Бұл алгоритм сөздердің жұрнақтарын анықтай алады және префикстерге аса мән бермейді. Портер алгоритмі бес сатыдан тұрады және әр қадамда сөз префикстерін алып тастауға арналған арнайы ережелер бар.

Стохастикалық алгоритм сөздің негізін ықтималдықпен анықтаумен байланысты. Бұл алгоритм ықтималдық модель құрады және түбірлік, флективтік формалардың сәйкестік кестесі арқылы оқытылады. Бұл модель әдетте жалғау мен жұрнақтардың кесілуі мен лемматизация алгоритмдерінде қолданылатын, өзінің сипаты бойынша ережелерге ұқсас күрделі лингвистикалық ережелер түрінде ұсынылған.

2000 жылға дейін стеммингалгоритміне арналған жұмыстардың басым көпшілігі ағылшын тіліне арналды, себебі Интернет желісіндегі ақпараттың 60% пайызы ағылшын тілінде жазылады. Алайда 2000-жылдардан кейін басқа тілдер үшін де стеммингалгоритмін орындау қажеттілігі туындай бастады. [3] жұмыста авторлар серб тіліне арналған стеммингалгоритмін құрастырған, нәтижесінде алынған бағдарлама сентимент талдау жүйесінде қолданылады. Аталған стеммер жаңа ережелер құрастыруда екі әдіс қолданылған. Алдымен серб тіліндегі сөз түрлері мен олардың түрленулері туралы грамматикалар зерттелсе, екінші әдісте дұрыс табылмаған сөздерді қолмен

анықтап, аталған сөз түрлері үшін жаңа ереже құруға қатысты болып табылады.

[4] жұмыста стеммер бағдарламаларында жиі кездесетін негізгі мәселелерді минимизациялауға арналған екі әдіс ұсынылады. Алғашқы әдіс - Extended-Light деп аталатын, жеңіл стеммер, оның мақсаты жеткіліксіз деңгей мәселесінің әсер етуін тежеу, оған префикстар жұрнақтарды, соның ішінде әдетте етістіктерге жалғанатын аффикстерді енгізу арқылы қол жеткізіледі. Extended-Light стеммерін араб тіліндегі кез келген мәтінге қолдануға болады. Екінші әдіс - лингвистикалық стеммер, бұл әдіс араб тіліндегі сөздердің бірнеше ережеге сәйкес әр түрлі үлгі бойынша өрнектелуіне негізделеді, яғни ұсынылып отырған лингвистикалық стеммер кіріс сөздердің сөз таптарын дұрыс анықтау үшін және сәйкесінше қандай техниканы қолдану керектігін анықтау үшін аталған үлгілерді қолданады.

[5] жұмыста малай тіліне арналған стеммер бағдарламасы құрылады. Малай тіліне арналған стеммерлерде қате көп кездеседі деп есептеледі, себебі олар стемминг алгоритмі барысында жалған нәтиже қайтаратын онлайн сөздікке тәуелді болып келеді. Бұл жұмыста 9512 сөзді қамтитын офлайн сөздік қолданылады. Алгоритм әр кіріс сөзді жоғарыдағы сөздіктенек негіз ретінде іздейді, табылмаған жағдайда сөзді өңдеуге кіріседі. Келесі 5 үлгі бойынша іздеу жүргізіледі: негіз-қосымша-жұрнақ, негіз-көптік жалғау, негіз-инфикс, негіз-префикс және негіз-жұрнақ. Құрастырылған бағдарлама префикс, жұрнақ және инфиксті жоғары дәлдікпен анықтайды.

[6] жұмыста телугу (дравидий, Үндістанда қолданылатын тіл) тіліндегі мәтіндерді жіктеуде стеммерлердің тигізетін әсеріне талдау жасалады. Телугу корпусына стеммерсіз және бірнеше стеммер әдістерін пайдалану арқылы тәжірибе жүргізілген. Әр түрлі жеті санаттағы 1150 құжатқа талдау жасалған. Жіктеу бағдарламаларын бағалау үшін KNN әдісі қолданылған. Нәтижелер жіктеу дәлдігінің айтарлықтай артқанын көрсетеді. Авторлар қорыта келе, стемминг алгоритмінің телугу тілінде жіктеу барысында аса маңызды екендігін айтады.

Ұсынылатын әдіс

Стемминг үлгісін оқыту өзгертілген нысандарды енгізу және үлгі ережелерінің ішкі жиынтығына сәйкес түбірлік нысанды генерациялау арқылы орындалады, ең тиісті ережелерді немесе ережелердің бірізділігін қолдануға, сондай-ақ сөз негіздерін таңдауға байланысты шешімдер нәтижелік дұрыс сөздің ең жоғары ықтималдығы болуы негізінде қолданылады.

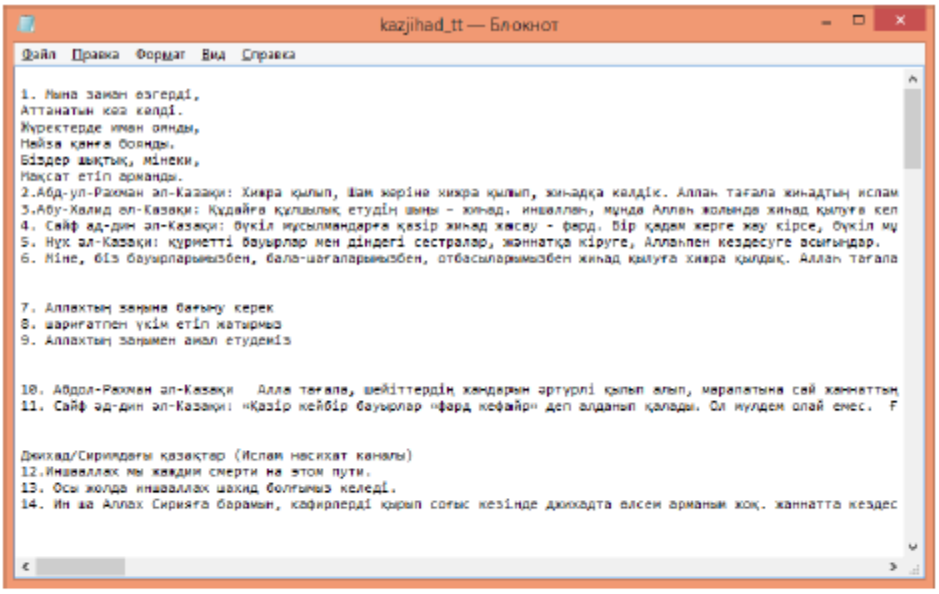
Алайда жоғарыда айтылған және басқа да әдістер қазақ тіліне арналып жасалмаған. Қазақ әліпбиінде кириллицадан бөлек арнайы 9 өзіндік әріп ә, ғ, қ, ң, ө, ұ, ү, h, і пайдаланылады. Осы жағдайларды ескере отырып, өз алгортимізді ұсындық. Ұсынылған алгоритм бойынша жұрнақ пен жалған кіріс сөздің соңынан деректер қорына сұраныс жасау арқылы ізделеді. Деректер қорында бірнеше сөз табылған жағдайда олардың ішіндегі ең ұзын сөз қайтарылады, ал сәйкестік болмаса, берілген сөздің басынан бастап деректер қорынан іздестіріледі. Деректер қорынынан табылмаған жағдайда сөздің өзі қайтарылады

Қазақ тіліндегі түбір сөздерді деректер қорына пайдалану үшін SQLite қолданылды. Сөзді алдын ала құрылған кестедегі түбір сөздермен салыстырылады. Егер де сәйкестік болса chekWord() әдісі “True”, яғни дұрыс деген нәтиже қайтарылады. Ал сәйкестік табылмаған жағдайда “False”, яғни бұрыс нәтижесі шығады.

Қазақ тілінде кездесетін жұрнақ пен жалғауларды біріктіре отырып, пайда болған қосымшаларды ұзындығы бойынша файлдарға бөлінді [7]. Қазақ тіліндегі кейбір қосымшалар сөздің түбірін өзгертеді. Түбір сөзге қосымшалар жалғанған кезде түбірегі “б” әрібі “п” әрібіне, “қ” әрібі “ғ” әрібіне және “г” әрібі “к” әрібіне ауысады. Мысалы “кітап” деген сөзге “ы” жалғауы жалғанған кезде “кітабы” деген сөз пайда болады. Осы жағдайды ескере отырып, түбір сөзді дұрыс табу үшін арнайы checkKazChar() әдісіні қолданылады. Жұрнақ не жалғау алынған сөзді деректер қорында бар немесе жоқтығы тексеріледі. Егер деректер қорынан сөз табылмаған жағдайда сөздің соңғы әрібі checkKazChar() әдісі арқылы тексеріледі. Сонда “кітаб” деген сөз “кітап” сөзіне ауыстырылады. chekWord() әдісі арқылы сөздің деректер қорында бар не жоқтығы анықталады.

Нәтиже

Мысал ретінде 4400 сөзден тұратын мәтінді қарастырдық (2-сурет). Мәтінді input.txt файлына саламыз. Арнайы жазылған бағдарлама кіріс мәтіндегі сөздерге жоғарыда сипатталғандай стемминг алгоритмін қолданады, алдымен мәтіндегі тыныс белгілері, арнайы таңбалар өшіріледі. Кейінгі қадамда мәтіндегі әрбір сөздің жалғауы жоғарыда сипатталған стемминг алгоритмі бойынша өшіріліп, тек негіздері қалдырылады.



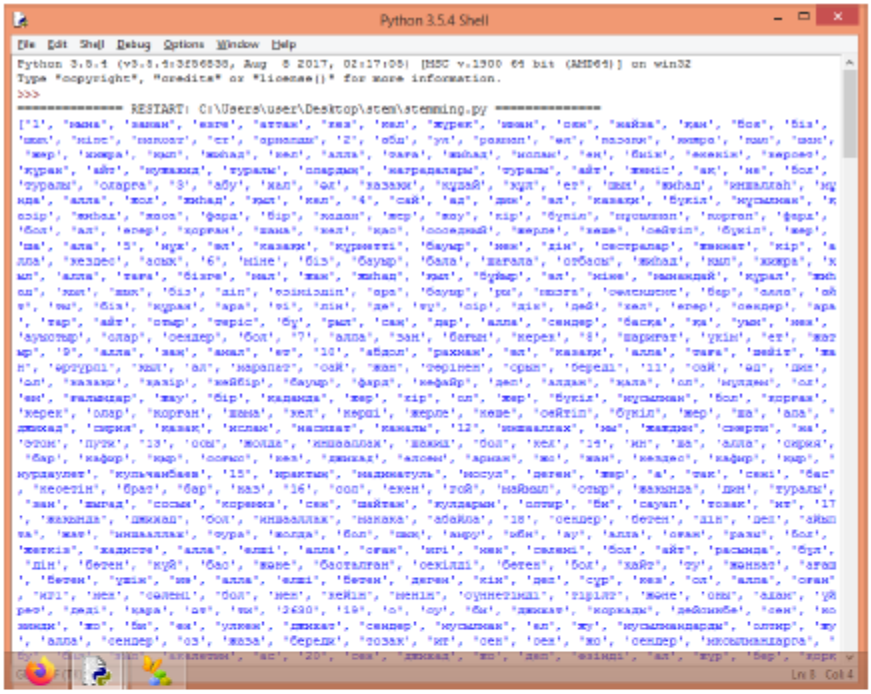
2-сурет. Кіріс мәтін

Кіріс мәтіндегі сөздердің негіздерін анықтау үшін қазақ тіліндегі 24 мың және 76 мыңға жуық сөз негіздерін қамтитын екі деректер қоры пайдаланылды (3-сурет). Экстремистік мәтіндерге тән кілттік сөздер де кестеге енгізілді.

id	kaz	type
Click here to define a filter		
9928	норма	noun
9929	вахабис	noun
9930	жихад	noun
9931	өлтір	verb
9932	жиһад	noun
9933	соғыс	noun
9934	қыр	verb
9935	жихат	noun
9936	джихад	noun
9937	джихат	noun
9938	соғыс	noun
9939	шайқас	noun
9940	шайкас	noun
9941	кафир	noun
9942	кафир	noun

3-сурет. 76000 сөз негізінен тұратын кесте

Кіріс мәтіндегі сөздердің жоғарыда сипатталған стемминг алгоритмі бойынша анықталған негіздері 4-суретте келтірілген.



4-сурет. Бағдарлама нәтижесі

Сондай-ақ, бағдарлама нәтижесі result.csv файлына да жазылып отырады (5-сурет). Аталған құжат екі бағанды қамтиды: бастапқы кіріс сөз және стемминг алгоритмін қолдану арқылы анық талған сөз негізі.

	Кіріс сөз	Сөз негізі
4	мына	мына
5	заман	заман
6	өзгерді	өзге
7	аттанатын	аттан
8	кез	кез
9	келді	кеп
10	жүректерде	жүрек
11	иман	иман
12	оянды	оян
13	найза	найза
14	қанға	қан
15	боянды	боя
16	біздер	біз
17	шықтық	шық
18	мінеки	міне

28	хикра	хикра
29	қыпып	қып
30	шам	шам
31	жерінe	жер
32	хикра	хикра
33	қыпып	қып
34	жиһадқа	жиһад
35	келдік	кел
36	аллаһ	алла
37	тағала	таға
38	жиһадтың	жиһад
39	исламның	ислам
40	ең	ең
41	бигі	бик
42	екенін	екенін
43	көрсетті	көрсет

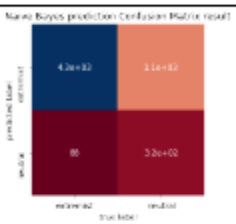
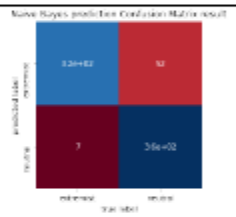
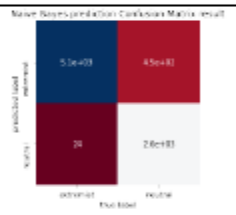
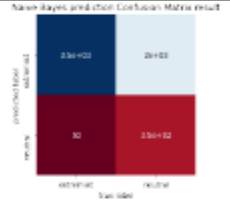
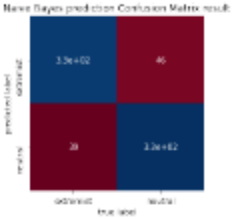
5-сурет. Бағдарлама нәтижесі жазылатын result.csv файлы

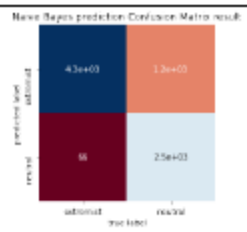
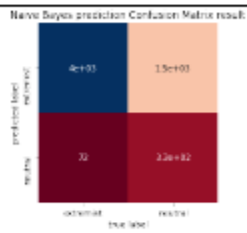
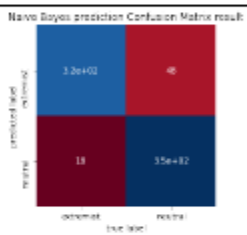
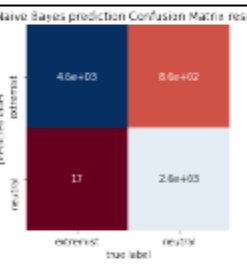
Сөздің түбірін дұрыс табу дәлдігі деректер қорындағы сөздерге тікелей байланысты. Деректер қорында сөздердің аз болуы дәлдікті төмендетеді. Сонымен қатар сөз түрлендіруші жалғаулар сөздің түбірін өзгертіп жіберетінін де ескерген жөн. Мәселен "халық" деген сөзге "ның" жалғауы жалғанғанда "халқының" деген сөзге айналады. Осы себепті сөз негіздерін қамтитын деректер қорын діни,

экстремистік сипаттағы кілттік сөздермен толықтырып отыру қажет. Ұсынылған стемминг алгоритмі бойынша машиналық оқыту әдістері арқылы қазақ тіліндегі мәтіндерді экстремистік және бейтарап топтарға жіктеу тапсырмасы орындалды.

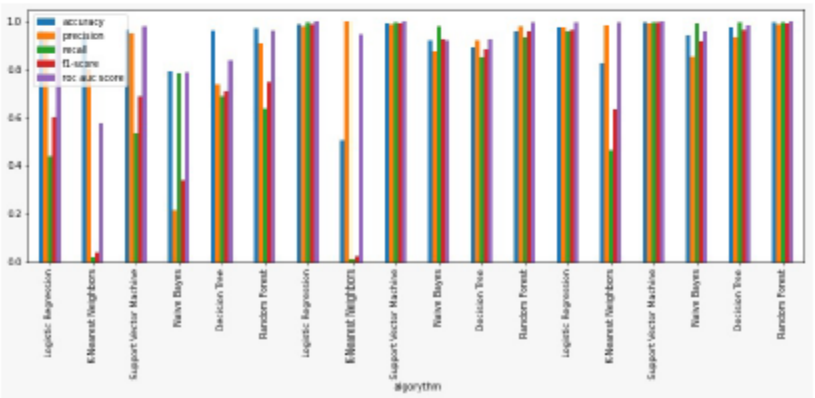
Құрастырылған корпустағы бейтарап және экстремистік мәтіндер үлесі біркелкі болмағандықтан oversampling және undresampling атты мәтіндерді теңестіру әдістері қолданылды. Кіріс мәтінді бейтарап және экстремистік класстарға жіктеу үшін LR – Logistic Regression, K-NN – K-Nearest Neighbors, SVM – Support Vector Machines, NB – Naïve Bayes, DT – Decision Tree, RF – Random Forest машиналық оқыту әдістері қолданылды. Жіктеу нәтижесі 1-кестеде келтірілген.

1-кесте. Кіріс мәтінді стемминг алгоритмі арқылы және стеммингсіз жіктеу нәтижелері

Мәтін түрі	Мәтінді теңестіру әдісі	Класстар саны	Алгоритм нәтижелері	Дәлсіздік матрицасы
Стеммингке дейінгі мәтін	Теңестірусіз	1 – 1857 0 - 27479	LR: 0.96 K-NN: 0.93 SVM: 0.97 NB: 0.79 DT: 0.96 RF: 0.97	
	Undersampling	1 – 1857 0 - 1857	LR: 0.98 K-NN: 0.51 SVM: 0.99 NB: 0.92 DT: 0.88 RF: 0.95	
	Oversampling	1 – 12999 0 - 27479	LR: 0.98 K-NN: 0.82 SVM: 0.99 NB: 0.94 DT: 0.97 RF: 0.99	
24000 сөздік ДҚ	Теңестірусіз	1 – 1857 0 - 27479	LR: 0.96 K-NN: 0.94 SVM: 0.97 NB: 0.65 DT: 0.96 RF: 0.97	
	Undersampling	1 – 1857 0 - 1857	LR: 0.91 K-NN: 0.54 SVM: 0.92 NB: 0.89 DT: 0.80 RF: 0.88	

	Oversampling	1 – 12999 0 - 27479	LR: 0.95 K-NN: 0.84 SVM: 0.98 NB: 0.85 DT: 0.95 RF: 0.97	
76 мың сөздік ДҚ	Теңестірусіз	1 – 1857 0 - 27479	LR: 0.96 K-NN: 0.93 SVM: 0.97 NB: 0.74 DT: 0.96 RF: 0.97	
	Undersampling	1 – 1857 0 - 1857	LR: 0.98 K-NN: 0.50 SVM: 0.98 NB: 0.91 DT: 0.90 RF: 0.95	
	Oversampling	1 – 12999 0 - 27479	LR: 0.97 K-NN: 0.85 SVM: 0.99 NB: 0.89 DT: 0.98 RF: 0.99	

Стеммингг алгоритмін қолдану барысында мәтінді жіктеу көрсеткіштерінің жоғарылағандығын 6-суреттен көруге болады.



6-сурет. Кіріс мәтінді стеммингг алгоритмі арқылы және стеммингсіз жіктеу нәтижелері

Қорытынды

Берілген мақала веб-ресурстардағы экстремистік мәтіндерді анықтау мақсатында семантикалық үлгілер құруға қатысты зерттеу жұмысының бір бөлігі болып табылады. Бұл жұмыста авторлар қазақ тіліндегі экстремистік мәтіндерді жіктеу дәлдігін арттыру мақсатында құрастырылған корпус мәтіндеріне стемминг алгоритмін қолдануды ұсынады. Стемминг алгоритмін қолданған жағдайда мәтінді жіктеу көрсеткіштерінің артқаны байқалады. Алдынғы зерттеу жұмыстарында деректер қорындағы сөз негіздерін көбейту арқылы ол жерде кездеспейтін, түбірі өзгерген сөздердің негізін дұрыс таба отырып, экстремистік мәтіндерді анықтау дәлдігін арттыру жоспарлануда. Деректер қорына діни, экстремистік мазмұндағы сөз негіздерін енгізу, қазақ тілінің төл әріптерін кирилл әріптерімен алмастыру заңдылықтарын анықтау тапсырмасы қойылды.

Берілген мақала Қазақстан Республикасының цифрлық даму, инновациялар және аэроғарыш өнеркәсібі министрлігінің тапсырысы бойынша ғарыштық қызмет және ақпараттық қауіпсіздік саласындағы қолданбалы ғылыми зерттеулер бағытындағы "Мәтіндегі экстремистік бағытты анықтау үшін веб-ресурстардағы семантикалық талдау модельдерін, алгоритмдерін құрастыру және кибер-криминалистика құрал-жабдықтарын әзірлеу" жобасы аясында жазылды, ЖТН AP06851248.

ӘДЕБИЕТТЕР

[1] Электрондық ресурс: https://kk.wikipedia.org/wiki/ [Қаралған күні: 02.03.2020]

[2] M. E. Porter and J. Leo, "Competitive strategy: Techniques for analysing industries and competitors Porter, Michael E. Free Press (Macmillan), New York, 396 p., 17.95," Industrial Marketing Management, vol.11, no.4, pp.318–319, 1982.

[3] N.Milošević, Stemmer for Serbian language , Natural language processing, 1209.4471. https://arxiv.org/ftp/arxiv/papers/1209/1209.4471.pdf

[4] M.Mohammad, A.S.Aldeen, M.E.Zidan, R.E.Ahmed, Y.Eltigani, Developing Two Different Novel Techniques for Arabic Text Stemming, Intelligent Information Management, 2019, 11, 1-23.

[5] R.U.Khan, F.S.Mohamad, M.I.UlHaq, Sh.A.Zadi Adruce, Ph.N.Anding, S.N.Khan, A.Y.Saleh Al-Hababi, Malay Language Stemmer, International Journal For Research In Emerging Science And Technology, V.4, Issue-12, Dec-2017

[6] N.Swapna, P.Subhashini, B.Padmaja Rani, Impact of Stemming on Telugu Text Classification, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019 https://www.ijrte.org/wp-content/uploads/papers/v8i2/B2338078219.pdf

[7] U.Tukeyev, D.Rakhimova, A.Turganbayeva, D. Amirova, B.Abduali, A. Karibayeva, Lexicon-free stemming for Kazakh language information retrieval, The IEEE 12th International Conference on Application of Information and Communication Technologies / AICT 2018,95-98

Повышение точности классификации экстремистических текстов с помощью стемминг алгоритма

Мусиралиева Ш.Ж., Болатбек М., Зият Б.

Повышение точности классификации экстремистских текстов с помощью алгоритма стемминга

Резюме. В настоящее время различные экстремистские организации активно используют социальные сети для своей деятельности. Поэтому важно создавать программы, требующие реализации комплекса эффективных мер, направленных на выявление, предотвращение и борьбу с экстремизмом. В данной статье авторы предлагают использовать алгоритм поиска корней для текстов в корпусе, предназначенный для повышения точности классификации экстремистских текстов на казахском языке.

Ключевые слова: выявление экстремизма, социальные сети, классификация текстов, стемминг алгоритм, кибербезопасность.

УДК 532.536

D. Bolysbek[1] A.B. Kuldzhabekov [2], A.A.[2] Kudaikulov
([1]KazNU named after al-Farabi, [2] Satbayev University
E-mail: bolysbek.darezhat@gmail.com)

REVIEW OF METHODS FOR PROCESSING STRUCTURAL AND DYNAMIC PROCESSES OCCURRING IN A POROUS MEDIUM ON A PORE SCALE

Abstract. With the increased availability and ease of use of digital imaging and image analysis software, these technologies are becoming the standard method for studying the properties of porous geological materials. However, the dependence of the image-based methods on the user means that one must be careful when interpreting this data as absolute values. Therefore, the scientific community needs a standardized workflow that researchers from all institutions around the world can follow. While this need is widely recognized, its implementation is hampered by the fact that the quality of the images obtained (and therefore the method of analysis) depends on many different factors, for example, the hardware