ҒТАХР 89.01.94
https://doi.org/10.56132/2791-3368-2026-1-65-71-82

Р.Б. Абдрахманов[1], Т.С.Картбаев[2], А.Б. Тоқтарова[3]
[1]International University of Tourism and Hospitality, Түркістан, Қазақстан
[2]Kazakh National Women's Teacher Training University, Алматы, Қазақстан
[3]Auezov University, Шымкент, Қазақстан
E-mail: kartbayev.t@qyzpu.edu.kz

BERT МОДЕЛІ МЕН ЗЕЙІН МЕХАНИЗМДЕРІ НЕГІЗІНДЕ ҚАЗАҚ ТІЛІНДЕГІ БЕЙӘДЕП СӨЗДЕРДІ АНЫҚТАУ

Аңдатпа. Бұл зерттеу мақалада қазақ тіліндегі әлеуметтік желілердегі бейәдеп сөздер мен агрессивті пікірлерді автоматты түрде анықтау әдістері қарастырылады. Зерттеудің өзектілігі желідегі бейәдеп тілді пікірлер көлемінің артуымен және ресурстары шектеулі тілдер үшін тиімді автоматтандырылған шешімдердің аздығымен байланысты. Қазақ тілі күрделі морфологиясымен және контекстік ерекшелігімен сипатталады, бұл дәстүрлі табиғи тілді өңдеу әдістерін қолдануды қиындық туындрады. Ұсынылып отырған зерттеу жұмысында көп басты (multihead) зейін механизмі мен конволюциялық нейрондық желілерді біріктіре отырып, BERT трансформатор архитектурасына негізделген гибридті модель ұсынылады. Бұл архитектура бейәдеп тілді текстерді, соның ішінде курсанттар мен офицерлер талқылайтын онлайн пікірлердегі ашық және жасырын түрлерін тиімді анықтауға мүмкіндік береді. Модельді оқыту және тестілеу үшін сарапшылар аннотациялаған қазақ тіліндегі пікірлердің корпусы құрастырылды, оған курсанттар, командирлер және жеке құрам арасындағы пікірлер де енгізілді.

Аңдатпа сапасы Флейсстің Каппа коэффициентін пайдаланып бағаланды, ол 0,6649 берді, жалпы келісім деңгейі 91,67%, бұл деректердің жоғары сенімділігін көрсетеді. Тәжірибелік нәтижелер ұсынылған модельдің дәлдік, F1-ұпай және ROC-AUC бойынша базалық әдістерден (Naive Bayes, LSTM, FastText) асып түсетінін көрсетті. Нәтижесінде алынған ROC-AUC мәні 0,95 болды.

Зерттеу нәтижелері қазақ тіліндегі ғадауат тілді сөздерді анықтауға арналған назар аудару механизмі бар трансформаторлық модельдердің тиімділігін растайды және оларды автоматтандырылған мазмұнды модерациялау және ақпараттық қауіпсіздікті бақылау жүйелерін әзірлеуде пайдалануға болады.

Түйінді сөздер: Bert моделі, бейәдеп сөздер, зейін механизімі, әлеуметтік желі, онлайн контент, қазақ тілді сөздер корпусы, Files Kappa.

Кіріспе.

Сандық технологиялар мен әлеуметтік желі мен медианың қарқынды дамуымен желіні пайдаланушылар жасаған онлайн контенттің көлемі үнемі артып келеді. Сандық байланыстың ақпаратқа әр уақытта және кез келген жерде қолжетімділігі және өзара баланнысу мүмкіндіктерінің кеңеюі сияқты оң аспектілерімен қатар, бейәдеп тілді сөздерінің, кемсітушілік мағынаға ие пікірлердің және ғадауат тілді сөздерінің таралуы артып келеді [1]. Бұл құбылыстар пайдаланушылардың психологиялық әл-жағдайына, қоғамдық пікірге және ақпараттық қауіпсіздікке теріс әсер етеді, әсіресе курсанттар, әскери қызметшілер және әскери бөлімдердің ақпараттық кеңістігінде тәртіп пен этикалық нормаларды сақтауға қауіп төндіруі мүмкін, бұл әлеуметтік ортада бейәдеп сөздерді анықтауға арналған автоматтандырылған жүйелерді әзірлеуді маңызды етеді [2].

Бейәдеп тілді сөздер - күрделі тілдік және әлеуметтік құбылыс, ол агрессияның ашық түрлерін де, жасырын, контекстке тәуелді дұшпандық пен кемсіту белгілері бар мәтіндерді де қамтиды. Пікірлерді сүзгіден өткізудің және қолмен модерациялаудың дәстүрлі әдістері деректердің масштабына, бағалаудың субъективтілігіне және тілдік формалардың жоғары динамикалық сипатына байланысты тиімсіз. Әсіресе әскери тәртіп, бұйрық және командалық құрылым сақталуы тиіс ортада ақпараттық бақылаудың тиімді тетіктері қажет. Сондықтан машиналық және терең оқытуға негізделген деректерді дайындау және табиғи тілді өңдеу әдістерін қолдану ерекше маңызға ие. Соңғы жылдары трансформатор архитектураларын, әсіресе мәтінді жіктеуде, сөздердегі «сезімді» талдауда және семантикалық тәуелділікті анықтауда жоғары өнімділік көрсеткен Transformers-тен екі бағытты кодтаушы көріністері (BERT) моделін пайдалануға айтарлықтай көңіл бөлінуде [3-5]. Зейін механизмдерін пайдалану сөйлемдердегі сөздердің контекстін қарастыруға және жасырын семантикалық қатынастарды анықтауға мүмкіндік береді, бұл бейәдеп тілді сөздерді анықтауда жоғары маңыздылыққа ие. Дегенмен, қазіргі зерттеулердің көпшілігі ағылшын немесе қытай тілдері сияқты жоғары ресурстық тілдерге бағытталған, ал қазақ тілі ресурстары аз тілдер қатарына еніп, терең зерттеу мен тәжірибені талап етеді.

Қазақ тілі агглютинативті морфологиямен, сөздік қорының бай жүйесімен және жоғары контексттік өзгергіштігімен сипатталады, бұл мәтінді машииналық оқыту алгоритмдерінде автоматты түрде өңдеуді қиындатады және классикалық машиналық оқыту алгоритмдерінің тиімділігін төмендетеді [6]. Ірі көлемді таңбаланған корпустар мен стандартталған лингвистикалық ресурстардың болмауы сенімді бейәдеп тілді сөздерді анықтау модельдерін құруда біршама кедергілер келтіреді. Сондықтан, қазақ тілінің құрылымдық және семантикалық ерекшеліктерін ескеретін мамандандырылған тәсілдерді әзірлеу қажет.

Бұл зерттеудің мақсаты – зейін механизмі бар BERT трансформатор архитектурасына негізделген қазақ тілді әлеуметтік медиадағы бейәдеп

сөздерді автоматты түрде анықтау моделінің тиімділігін әзірлеу және бағалау. Зерттеу деректерді аннотациялау сапасына, сараптамалық келісімді талдауға және ұсынылған модельді негізгі машиналық оқыту әдістерімен салыстырмалы бағалауға бағытталған. Алынған нәтижелер автоматты түрде модерациялау жүйелерінің сенімділігін арттыруға бағытталған және әлеуметтік желі кеңістігін бақылау мен ақпараттық қауіпсіздікте қолданылуға ұсынылады.

Әдебиеттік шолу.

Соңғы жылдары цифрлық кеңістіктегі бейәдеп тілді сөздерді автоматты түрде анықтау міндеті табиғи тілді өңдеу және деректерді өңдеуде қиыншылықтарға әкелуде [7]. Әлеуметтік желілерде пайдаланушылар жазған, жариялаған пікірлердің өсуі агрессивті, кемсітушілік мазмұндағы және ғадауат тілді сөздерді тиімді анықтауға қабілетті автоматтандырылған модерация әдістерін әзірлеуді қажет етті. Академиялық ғылыми зерттеулерде бейәдеп сөздер жоғары контекстік тәуелділікпен, семантикалық түсінікпен және тілдегі ерекшеліктерімен сипатталатын мәтіндік мазмұнның ерекше түрі болып саналады.

Бұл саладағы алғашқы зерттеулер қолмен ерекшеліктерді табуда (n-грамм, TF-IDF, лексикалық және синтаксистік сипаттамалар) біріктірілген Naive Bayes, Support Vector Machine, Logistic Regression және Random Forest сияқты классикалық машиналық оқыту әдістеріне сүйенді [8]. Салыстырмалы қарапайымдылығы мен түсіндірілуіне қарамастан, бұл тәсілдер бейәдеп тілді сөздерінің жасырын түрлерімен және күрделі контекстік құрылымдармен жұмыс барысында тиімділіктерінің шектеулі екендігін көрсетті.

Терең оқыту әдістерінің дамуымен нейрондық желілерге, әсіресе конволюциялық нейрондық желілерге (CNN) және қайталанатын нейрондық желілерге (RNN, LSTM) айтарлықтай көңіл бөлінді. CNN жергілікті үлгілер мен мәтіннің негізгі үзінділерін анықтауда жоғары тиімділік көрсетті, ал LSTM тізбекті тәуелділіктерді модельдеу және контекстті ескеру үшін қолайлырақ болып шықты [9]. Дегенмен, бұл модельдер әдетте айтарлықтай мөлшерде белгіленген деректерді қажет етеді және полисемия мен ұзақ контексттік тәуелділіктерді тиімді түрде шеше алмайды.

Трансформатор архитектураларының пайда болуы мәтінді жіктеу тапсырмаларын әзірлеудегі маңызды қадам болды. BERT моделі және оның модификациялары сезім талдауында, уыттылықты анықтауда және бейәдеп тілді сөздерді анықтауда өнімділіктің айтарлықтай жақсарғанын көрсетті. Өзіне назар аудару механизмдерін пайдалану сөз контекстін екі бағытты қарастыруға мүмкіндік береді, бұл әсіресе контекстке және прагматикалық мағынаға байланысты бейәдеп тілді сөздерді талдау кезінде маңызды. Бірнеше зерттеулер BERT дәлдік, F1-ұпай және ROC-AUC тұрғысынан дәстүрлі нейрондық архитектуралардан асып түсетінін көрсетті [10].

Зерттеудің жеке бағыты трансформаторлық ендірулерді CNN немесе LSTM қабаттарымен біріктіретін гибридті модельдерді пайдалануды

қамтиды. Бұл архитектуралар BERT контексттік көріністерінің артықшылықтарын нейрондық желілердің мәтіннің құрылымдық ерекшеліктерін анықтау мүмкіндігімен біріктіреді. Нәтижелер гибридті модельдердің бейәдеп тілді сөздер мен саркастикалық пікірлердің жасырын түрлерін анықтауда әсіресе тиімді екенін көрсетеді.

Сонымен қатар, қолданыстағы жұмыстардың көп бөлігі жоғары ресурсты тілдерге бағытталған, ал төмен ресурсты тілдер әлі де толық зерттелмеген. Қазақ тілі агглютинативті морфологиямен, бай сөзжасамдық жүйемен және кең көлемді аннотацияланған корпустардың болмауымен сипатталады, бұл стандартты NLP тәсілдерін қолдануды айтарлықтай қиындатады [11]. Сонымен қатар, әдебиеттерде деректерді аннотациялау сапасының және аннотаторлардың сәйкестігін бағалаудың маңыздылығы атап өтіледі, себебі аннотация қателіктері оқытылған модельдердің сенімділігіне тікелей әсер етеді.

Осылайша, қолданыстағы зерттеулерді талдау тілдік ерекшеліктер мен аннотация сапасын ескеретін төмен ресурсты тілдер үшін мамандандырылған бейәдеп тілді сөздерді анықтау модельдерін әзірлеу қажеттілігін көрсетеді. Осыған байланысты трансформатор архитектураларын назар аудару механизмдерімен және аннотаторлардың сәйкестігін талдаумен бірге пайдалану одан әрі зерттеулер үшін перспективалы бағыт болып көрінеді.

Материалдар мен тәсілдер.

Бұл зерттеу терең оқыту әдістерін қолдана отырып, қазақ тіліндегі пайдаланушылар жасаған мазмұндағы бейәдеп тілді сөзді автоматты түрде анықтау моделінің тиімділігін әзірлеуге және бағалауға бағытталған. Зерттеу әдіснамасына деректерді жинау және алдын ала өңдеу, мәтіндік аннотация, белгілеу сапасын бағалау, модель құру және эксперименттік валидация кіреді.

Деректерді жинау және алдын ала өңдеу.

Пайдаланылған бастапқы деректер әлеуметтік желілерден және ашық қолжетімді онлайн платформалардан API және веб-скрапинг арқылы жиналған мәтіндік түсініктемелерден тұрды. Нәтижесінде алынған корпус бейтарап және ықтимал агрессивті мәлімдемелерді қамтитын пайдаланушы хабарламаларын қамтыды. Алдын ала өңдеу кезеңі келесі процедураларды қамтыды: HTML тегтері мен арнайы таңбаларды алып тастау, мәтінді қалыпқа келтіру, кіші әріптерге түрлендіру және қосымша бос орындар мен ақпараттық емес белгілерді алып тастау. Деректер құрылымдалып, оқыту, валидация және тест жинақтарына бөлінді.

Деректерді аннотациялау процесі.

Мәтін аннотациясы бірнеше сарапшы аннотаторларды қамтитын адамдық тәсілді қолдану арқылы орындалды [12].

$$k = \frac{P - P_e}{1 - P_e}. \quad (1)$$

Бұл жерде:

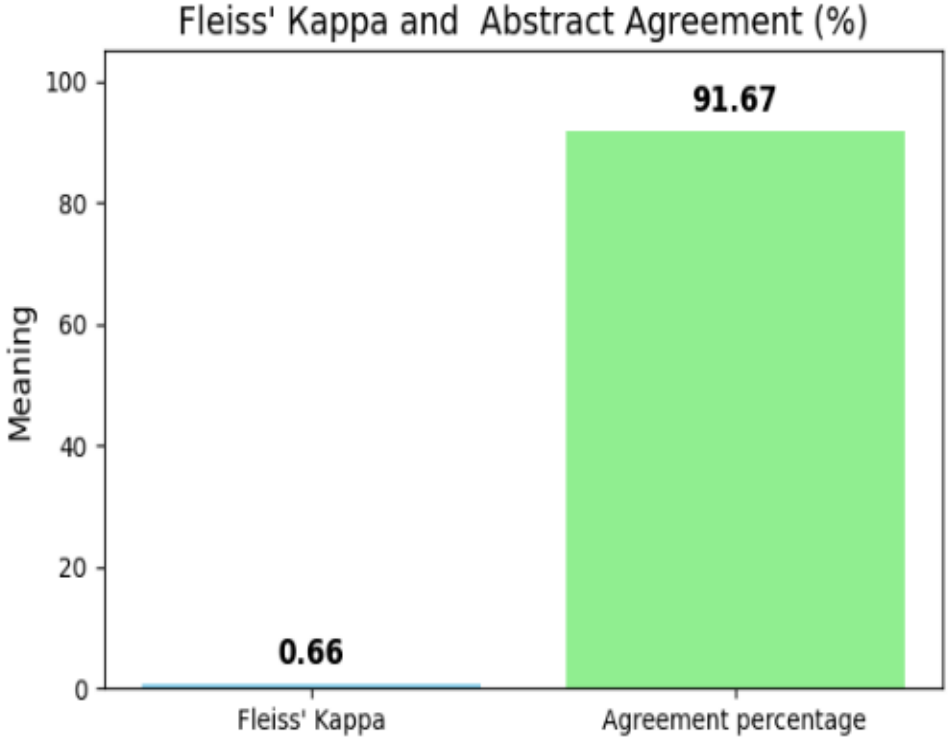
$\underline{P}$- нақты келісім үлесі;

$\underline{P_e}$ - кездейсоқтық келісімінің күтілетін үлесі.

Әрбір мәтін фрагменті бейәдеп тілді сөздің болуын немесе болмауын анықтау үшін алдын ала әзірленген критерийлерге сәйкес бірнеше аннотаторлармен тәуелсіз аннотацияланды. Бұл тәсіл аннотацияның сенімділігін арттырды және субъективтілікті азайтты.

Аннотация сапасын бағалау.

Аннотатор келісімін бағалау үшін бірнеше аннотациялар кезінде сарапшылар арасындағы келісім деңгейін сандық бағалау үшін Флейсстің Каппа коэффициенті пайдаланылды [13]. Сонымен қатар, жалпы келісімнің пайызы есептелді.



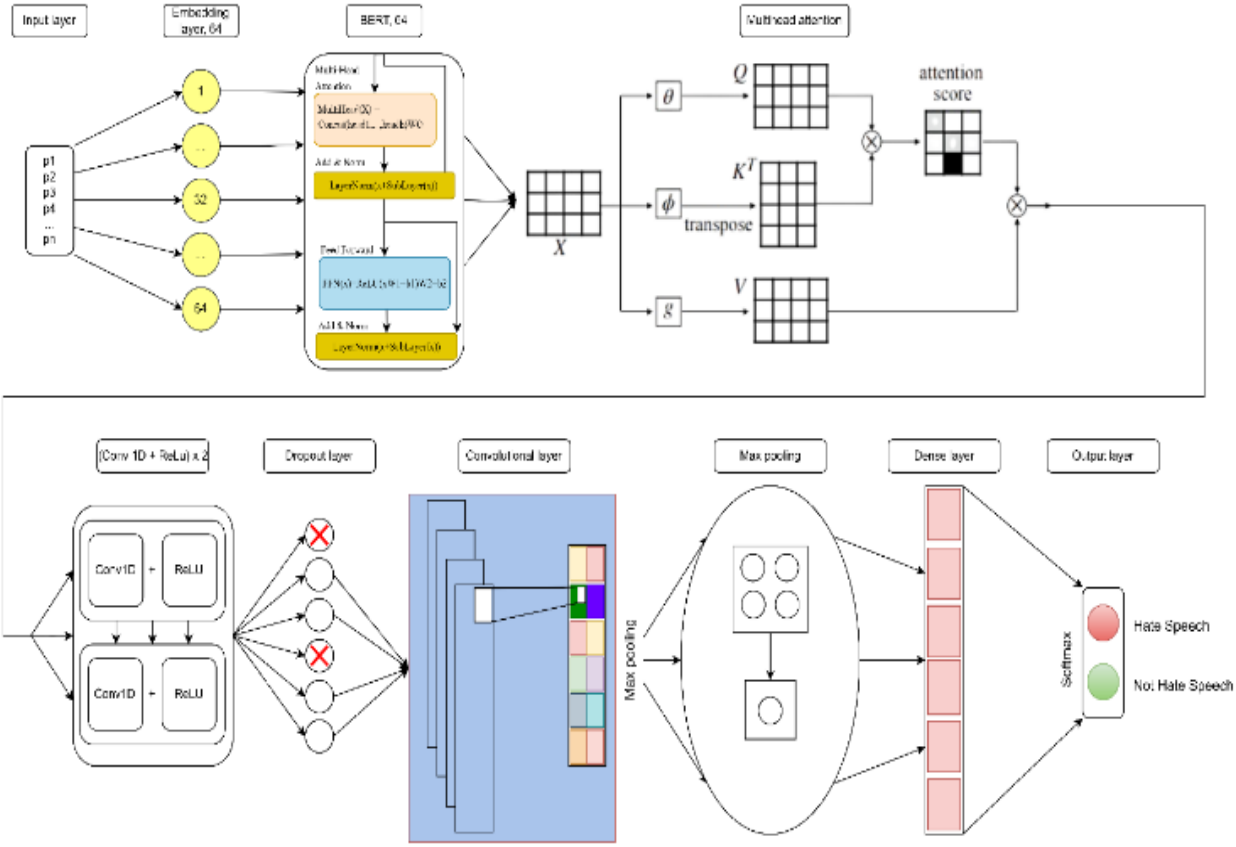
1 сурет - Флейсс каппасы және жалпы келісім пайызы

Есептеулер Флейсстің Каппасының 0,6649-ға тең екенін және жалпы келісім пайызы 91,67%-ға жеткенін көрсетті. Бұл деректер жиынтығының айтарлықтай сенімділігін және аннотацияның жоғары сапасын білдіреді (1-сурет).

Алынған мәндер аннотация сенімділігінің жоғары деңгейін көрсетті, бұл алынған корпусты машиналық оқыту моделін құру үшін жоғары сапалы оқыту деректер жиынтығы ретінде пайдалануға мүмкіндік берді.

Модель архитектурасы.

Бұл зерттеу зейін механизмі мен конволюциялық нейрондық желілерді (CNN) біріктіре отырып, BERT трансформатор архитектурасына негізделген гибридті модельді әзірленді.



2 сурет – Ұсынылған гибридті модель

Алдын ала дайындалған BERT моделі мәтіндердің контекстке сезімтал векторлық көріністерін алу үшін пайдаланылды. Шығыс ендірмелері жергілікті ерекшеліктерді алу және кейіннен жіктеу үшін конволюциялық қабаттарға енгізілді. Бұл архитектура мәтіндегі жаһандық контекстті де, жергілікті тілдік үлгілерді де тиімді қарастыруға мүмкіндік береді.

Эксперименттік орнату және бағалау метрикалары.

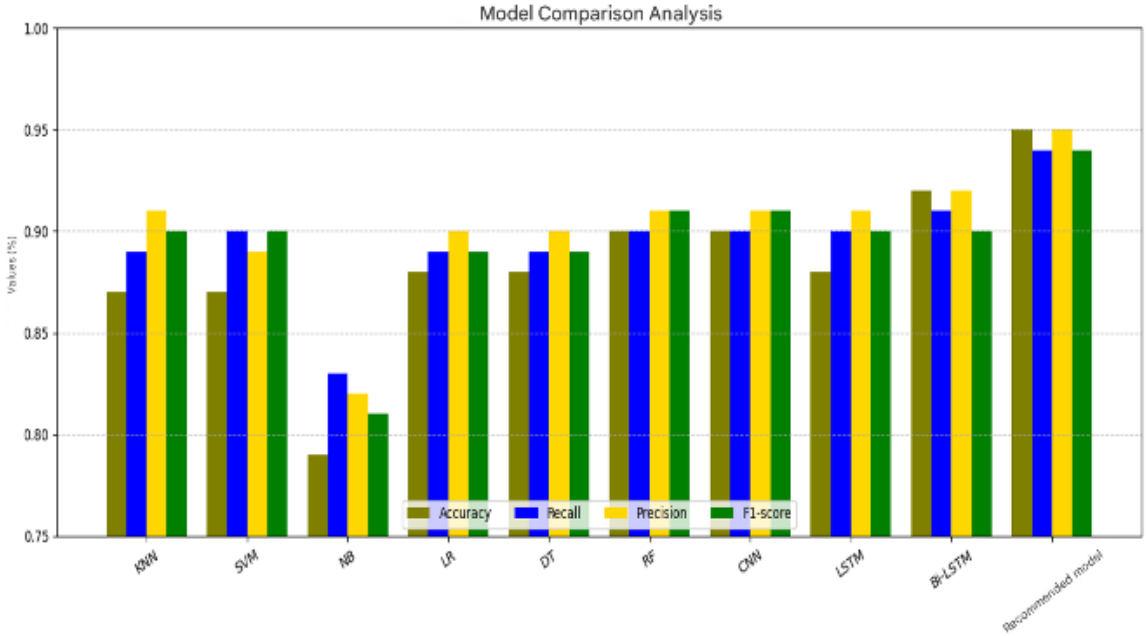
Модель стандартты оңтайландыру алгоритмдерін және екілік жіктеу мәселелері үшін жоғалту функциясын пайдаланып оқытылды. Ұсынылған тәсілдің тиімділігін бағалау үшін дәлдік, дәлдік, еске түсіру, F1-ұпай және ROC-AUC метрикалары пайдаланылды. Нәтижелер базалық машиналық оқыту модельдерімен және жеке нейрондық архитектуралармен салыстырылды, бұл ұсынылған әдістің артықшылықтарын объективті бағалауға мүмкіндік берді.

Нәтижелер мен талқылау.

Зерттеудің эксперименттік бөлігі ұсынылған гибридті LSTM-CNN моделінің тиімділігін бағалауға бағытталған, ол қазақ тіліндегі әлеуметтік желі қолданушылары жазған, жариялаған бейәдеп тілді сөздерді автоматты түрде анықтау үшін BERT трансформаторлық ендірмелерін пайдаланады. Бағалау модельді оқыту процесінде пайдаланылмаған тест жиынтығында жүргізілді, бұл оның жалпылау қабілетін объективті бағалауға мүмкіндік берді.

Аннотация сапасын бағалау нәтижелері.

Модельді оқытудан бұрын мәтіндік деректерді аннотациялау сапасы бағаланды. Есептелген Флейсстің Каппа коэффициенті 0,66 мәнін берді, бұл аннотаторлар арасындағы келісімнің жоғары деңгейін көрсетеді. Жалпы келісім пайызы 91,67% құрады, бұл аннотацияның сенімділігін және алынған корпустың жоғары сапалы оқыту деректер жиынтығы ретінде жарамдылығын көрсетеді. Салыстырмалы модельді талдау



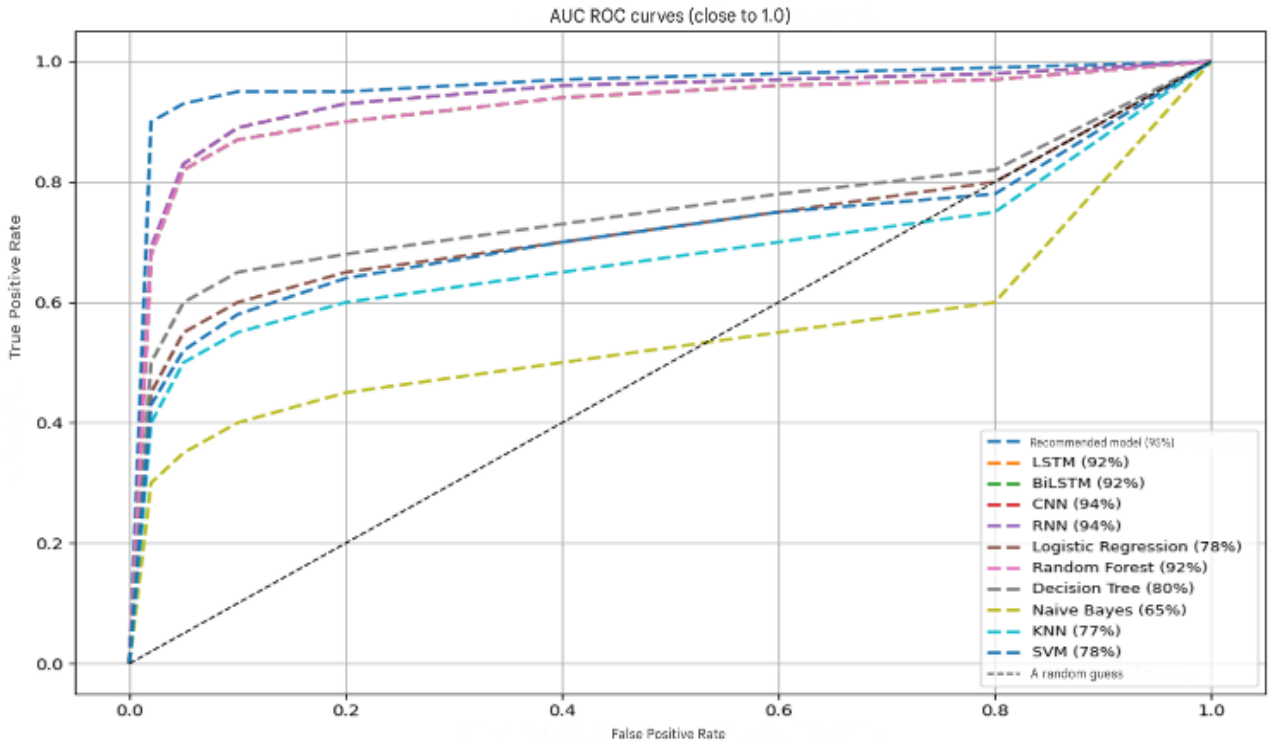
3 сурет - Жіктеу көрсеткіштерін салыстырмалы талдау

Салыстырмалы бағалау үшін классикалық машиналық оқыту алгоритмдері (Naive Bayes, Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Kongres), сондай-ақ жеке нейрондық архитектуралар (CNN, LSTM) пайдаланылды. Барлық модельдер бірдей бағалау көрсеткіштерін пайдаланып, бірдей деректер жиынтығында оқытылды және тестілеуден өткізілді.

Эксперименттік нәтижелер ұсынылған BERT негізіндегі модельдің интеграцияланған назар аудару механизмдері мен конволюциялық қабаттары бар модельдің барлық негізгі көрсеткіштер бойынша бастапқы әдістерден асып түсетінін көрсетті (сурет 3). Атап айтқанда, гибридті модель жіктеу дәлдігі бойынша 93,2%, дәлдік бойынша 91,5%, еске түсіру бойынша 94,0% және F1 ұпайы бойынша 92,7% көрсетті. Салыстыру үшін классикалық алгоритмдер F1 ұпайының төмен мәндерін көрсетті, 78%-дан 86%-ға дейін.

ROC қисығы және AUC талдауы

Қосымша бағалау ROC қисығы мен қисық астындағы ауданды (AUC) пайдаланып жүргізілді. Нәтижесінде алынған 0,95 AUC мәні модельдің әртүрлі шектердегі «жек көрушілік сөз» және «бейтарап мазмұн» кластарын ажыратудың жоғары қабілетін көрсетеді.



4 сурет - AUC-ROC қисықтары: ұсынылған модельдің базалық алгоритмдермен салыстырмалы тиімділігі

ROC қисығын талдау сонымен қатар модельдің сыныптық теңгерімсіздікке беріктігін көрсетті, бұл нақты әлемдегі әлеуметтік деректермен жұмыс істеген кезде маңызды фактор болып табылады.

Эксперименттік нәтижелер заманауи трансформатор архитектураларымен бірге жоғары сапалы аннотацияны қолдану қазақ тілінде жек көрушілік сөзді автоматты түрде анықтау тиімділігін айтарлықтай жақсарта алатынын растайды. Алынған нәтижелер ұсынылған модельдің автоматтандырылған мазмұнды модерациялау және сандық кеңістікті бақылау үшін практикалық қолданылуын көрсетеді.

Қорытынды.

Зерттеу барысында алынған нәтижелер қазақ тіліндегі пайдаланушылар жасаған мазмұндағы жеккөрушілік сөзді автоматты түрде анықтау үшін ұсынылған модельдің жоғары тиімділігін растайды. BERT трансформатор моделін назар аудару механизмдерімен және конволюциялық нейрондық желілермен біріктіретін гибридті архитектураны пайдалану мәтіннің жаһандық контекстін де, жергілікті тілдік үлгілерді де ескеруге мүмкіндік берді. Бұл әсіресе агрессияның жасырын түрлерімен, контекстке тәуелділікпен және семантикалық түсініксіздікпен сипатталатын жеккөрушілік сөзді талдау кезінде маңызды.

Салыстырмалы талдау ұсынылған модель барлық негізгі сапа көрсеткіштері бойынша классикалық машиналық оқыту әдістері мен жеке нейрондық архитектуралардан асып түсетінін көрсетті. Жоғары дәлдік,

дәлдік, еске түсіру және F1 ұпай мәндері модельдің агрессивті мазмұнды дұрыс анықтау ғана емес, сонымен қатар жалған оң нәтижелер санын азайту мүмкіндігін көрсетеді. ROC қисығын талдау және 0,95 AUC мәні модельдің нақты әлемдегі әлеуметтік медиа деректерімен жұмыс істеген кезде жиі кездесетін мәселе болып табылатын сыныптық теңгерімсіздікке төзімділігін одан әрі растайды.

Аннотация сапасының модельдің жұмысына әсері ерекше назар аударуға тұрарлық. Флейсстің Каппа коэффициентімен (0,66) расталған аннотатораралық келісімнің жоғары деңгейі және жалпы келісімнің жоғары пайызы берік оқыту корпусының қалыптасуын қамтамасыз етті. Бұл деректер шуын азайтты және модельдің жалпылау мүмкіндігін арттырды. Осылайша, зерттеу нәтижелері аннотация сапасының автоматты мәтінді талдау жүйелерін құруда, әсіресе ресурстары шектеулі тілдер үшін курсанттарды даярлау кезіндегі тәртіп секілді маңызды фактор екенін растайды.

Қол жеткізілген нәтижелерге қарамастан, зерттеудің белгілі бір шектеулері бар. Біріншіден, пайдаланылатын деректер корпусы белгілі бір әлеуметтік медиа платформаларымен шектелген, бұл басқа салаларға ауыстырылған кезде модельдің жалпылау мүмкіндігін төмендетуі мүмкін. Екіншіден, трансформаторлық модельдердің есептеу күрделілігі айтарлықтай аппараттық ресурстарды қажет етеді, бұл оларды нақты уақыт жүйелерінде қосымша оңтайландырусыз пайдалануды шектеуі мүмкін.

Қорытындылай келе, ұсынылған тәсіл қазақ тіліндегі жеккөрушілік сөздерді автоматты түрде анықтау мәселесінің тиімді шешімін білдіреді. Жоғары сапалы аннотацияларды және заманауи терең оқыту архитектураларын пайдалану жіктеудің дәлдігі мен сенімділігін айтарлықтай жақсартады. Алынған нәтижелер автоматтандырылған мазмұнды модерациялау жүйелерін, сандық кеңістікті бақылауды және ақпараттық қауіпсіздікті дамыту үшін маңызды практикалық салдарға ие. Бұл тәсіл курсанттар, сарбаздар және офицерлер қолданатын ақпараттық жүйелердегі қауіпсіз коммуникацияны қамтамасыз етуге де ықпал етеді. Болашақтағы перспективалы зерттеу салаларына деректер жиынтығын кеңейту, келесі буын трансформациялық модельдерін енгізу, диалог контекстін ескеру және модельді көптілді және мәдениетаралық ортаға бейімдеу кіреді

Алғыс білдіру. Бұл жұмыс Қазақстан Республикасы Ғылым және жоғары білім министрлігі қаржыландырған «Жасанды интеллектті қолдана отырып, әлеуметтік желілерде жастар арасында кибербуллингті автоматты түрде анықтау» атты ғылыми жоба аясында орындалды. Грант № IRN AP23488900.

ӘДЕБИЕТТЕР

1. C. Van Hee, G. Jacobs, C. Emmery, B. Desmet, E. Lefever and B.Verhoeven, «Automatic detection of cyberbullying in social media text.» PloS one 13.10 (2018): e0203794.

2. G. Valle-Cano, «SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles» Expert Systems with Applications 216 (2023): 119446.

3. S. Mombekova, S. Akhmetova, G. Shaimerdenova, S. Alisheva, E. Mussirepova, A. Kydyrbekova and N. Torebay, «Eco-efficient industrial processes: Leveraging ai-powered management for reduced environmental footprint.» E3S Web of Conferences. Vol. 614. EDP Sciences, 2025.

4. K. Kumari, J.P. Singh, Y.K. Dwivedi and N.P. Rana, «Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach» Soft computing 24 (2020): 11059-11070.

5. H. Elzayady, M.S. Mohamed, K.M. Badran, and G.I. Salama, «A hybrid approach based on personality traits for hate speech detection in Arabic social media» International Journal of Electrical and Computer Engineering 13.2 (2023): 1979.

6. Z. Makhanova, G. Beissenova, A. Madiyarova, M. Chazhabayeva, G. Mambetaliyev, M. Suimenova and A. Baiburin, «A Deep Residual Network Designed for Detecting Cracks in Buildings of Historical Significance.» International Journal of Advanced Computer Science & Applications 15.5 (2024).

7. J.L.Wang, L.A. Jackson, J. Gaskin and H.Z. Wang»The effects of Social Networking Site (SNS) use on college students' friendship and well-being» Computers in Human Behavior 37 (2014): 229-236.

8. A. Toktarova, A. Abushakhma, E. Adylbekova, A. Manapova, B. Kaldarova and Y.Atayev, «Offensive language identification in low resource languages using bidirectional long-short-term memory network.» International Journal of Advanced Computer Science and Applications 14.6 (2023).

9. M. Rezvan, S. Shekarpour, L. Balasuriya and K. Thirunarayan, «A quality type-aware annotated corpus and lexicon for harassment research» Proceedings of the 10th acm conference on web science. 2018.

10. B.A. Talpur and O. Declan, «Cyberbullying severity detection: A machine learning approach» PloS one 15.10 (2020): e0240924.

11. A. Toktarova, D. Sultan, and Zh. Azhibekova, «Review of Machine Learning Models in Cyberbullying Detection Problem» 2024 IEEE 4th International Conference on Smart Information Systems and Technologies (SIST). IEEE, 2024.

12. C. Comito, F. Agostino, and C. Pizzuti, «Word embedding based clustering to detect topics in social media» IEEE/WIC/ACM International Conference on Web Intelligence. 2019.

Rustam Abdrakhmanov, candidate of technical sciences, associate professor, International University of Tourism And Hospitality, Turkestan, Kazakhstan

Timur Kartbayev, PhD, professor, Digital Officer, Kazakh National Women's Teacher Training University, Almaty, Kazakhstan

Aigerim Toktarova, PhD, senior lecturer, Auezov University, Shymkent, Kazakhstan

DETECTION OF OFFENSIVE CONTENT USING THE BERT MODEL AND MECHANISMS OF MONITORING

Abstract. This research paper considers methods for automatic detection of profanity and aggressive comments in Kazakh social networks. The relevance of the study is due to the increasing volume of profanity comments online and the lack of effective automated solutions for resource-limited languages. The Kazakh language is characterized by complex morphology and contextual specificity, which makes it difficult to use traditional natural language processing methods.

The proposed study introduces a hybrid model based on the BERT transformer architecture, combining multihead attention and convolutional neural networks. This architecture enables the effective detection of offensive texts, including both explicit and subtle forms of comments exchanged among cadets and officers online. To train and test the model, a corpus of Kazakh-language comments annotated by experts was constructed, including opinions from cadets, commanders, and personnel.

The quality of the annotation was assessed using Fleiss's Kappa coefficient, which gave 0.6649, with an overall agreement level of 91.67%, which indicates high reliability of the data. Experimental results showed that the proposed model outperforms the baseline methods (Naive Bayes, LSTM, FastText) in terms of accuracy, F1-score, and ROC-AUC. The resulting ROC-AUC value was 0.95.

The research results confirm the effectiveness of transformer models with an attention mechanism for detecting profanity in the Kazakh language and can be used in the development of automated content moderation and information security monitoring systems.

Keywords: Berta model, obscene words, attention mechanism, social network, online content, corpus of the Kazakh language, Files Kappa.

Рустам Абдрахманов, к.т.н, доцент, International University of Tourism And Hospitality, Туркестан, Казахстан

Тимур Картбаев, PhD, профессор, Digital Officer, Kazakh National Women's Teacher Training University, Алматы, Казахстан

Айгерим Токтарова, PhD, старший преподаватель, Auezov University, Шымкент, Казахстан

ОБНАРУЖЕНИЕ ОСКОРБИТЕЛЬНОГО КОНТЕНТА С ИСПОЛЬЗОВАНИЕМ МОДЕЛИ BERT И МЕХАНИЗМОВ ВНИМАНИЯ

Аннотация. В данной исследовательской работе рассматриваются методы автоматического обнаружения нецензурной лексики и агрессивных комментариев в казахских социальных сетях. Актуальность исследования обусловлена растущим объемом нецензурных комментариев в интернете и отсутствием эффективных автоматизированных решений для языков с ограниченными ресурсами. Казахский язык характеризуется сложной морфологией и контекстной спецификой, что затрудняет использование традиционных методов обработки естественного языка.

В представленной исследовательской работе предложена гибридная модель на основе архитектуры трансформера BERT, сочетающая многоголовочный механизм внимания и сверточные нейронные сети. Эта архитектура позволяет эффективно выявлять нецензурные тексты, включая как явные, так и скрытые формы комментариев, обсуждаемых курсантами и офицерами онлайн. Для обучения и тестирования модели был создан корпус казахскоязычных комментариев, аннотированных экспертами, который включает мнения курсантов, командиров и личного состава.

Качество аннотации оценивалось с помощью коэффициента Каппа Флейсса, который составил 0,6649, с общим уровнем согласованности 91,67%, что указывает на высокую надежность данных. Результаты экспериментов показали, что предложенная модель превосходит базовые методы (наивный байесовский классификатор, LSTM, FastText) по точности, F1-мере и ROC-AUC. Полученное значение ROC-AUC составило 0,95.

Результаты исследования подтверждают эффективность трансформерных моделей с механизмом внимания для обнаружения ненормативной лексики в казахском языке и могут быть использованы при разработке автоматизированных систем модерации контента и мониторинга информационной безопасности.

Ключевые слова: модель Берта, нецензурные слова, механизм внимания, социальная сеть, онлайн-контент, корпус казахского языка, Files Kappa.

Қабылданған күні: 2026 жылғы 26 қантар
Рецензиядан өткен күні: 2026 жылғы 06 наурыз
Мақұлданған күні: 2026 жылғы 12 наурыз