әл-Фараби атындағы Қазақ ұлттық университеті

ӘОЖ 004.056 Қолжазба құқығында

САҒЫНАЙ МӨЛДІР

Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған тереңдетіп оқыту үлгілерін жасау

8D06301 – Ақпараттық қауіпсіздік жүйелері

Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші:
Болатбек Милана Асланбекқызы,
PhD, ассистент профессор,
әл-Фараби атындағы Қазақ ұлттық
университеті, Алматы, Қазақстан
Шетелдік ғылыми кеңесші:
Миан Мухаммад Васим Икбал,
PhD, қауымдастырылған
профессор, Пәкістан

Қазақстан Республикасы
Алматы, 2026

МАЗМҰНЫ

НОРМАТИВТІ СІЛТЕМЕЛЕР 3
ҚЫСҚАРТУЛАР МЕН БЕЛГІЛЕУЛЕР 4
КІРІСПЕ 5
1 Қазақ тіліндегі деструктивті веб-контентті анықтаудың теориялық және әдістемелік негіздері 12
1.1 Деструктивті веб-контент ұғымы, оның ақпараттық қауіпсіздікке әсері және қазақ тіліндегі деструктивті веб-контенттің ерекшеліктері 12
1.2 Деструктивті веб-контентті анықтауда қолданылатын табиғи тілді өңдеу, дәстүрлі машиналық оқыту және терең оқыту әдістері 16
1.3 Қазақ тіліндегі деструктивті веб-контентті классификациялаудың ғылыми мәселелері және зерттеу есебінің қойылымы 22
1.4 1-тарау бойынша қорытынды 26
2 Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған әдісті әзірлеу ... 28
2.1 Қазақ тіліндегі деструктивті веб-контент корпусын құру және аннотациялау әдістемесі 28
2.2 Қазақ тіліндегі деструктивті веб-контенттің морфологиялық және онтологиялық білім моделін әзірлеу 37
2.3 Контекстік, морфологиялық және семантикалық белгілердің интеграцияланған моделін әзірлеу 44
2.4 2-тарау бойынша қорытынды 48
3Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған тереңдетіп оқыту үлгілерін жасау 49
3.1 Базалық және кеңейтілген классификациялау модельдерінің салыстырмалы талдауы 50
3.2 KazBERT негізіндегі морфологиялық және онтологиялық модульдерді әзірлеу 54
3.3 Контекстік, морфологиялық және онтологиялық белгілерді біріктірудің математикалық моделі 58
3.4 Ұсынылған гибридті классификациялау архитектурасы 61
3.5 3-тарау бойынша қорытынды 67
4 Ұсынылған әдістердің эксперименттік зерттелуі және олардың тиімділігін бағалау 68
4.1 Эксперименттік орта және бағалау әдістемесі 68
4.2 Ұсынылған модельді базалық әдістермен салыстыру және бағалау 70
4.3 Ұсынылған гибридті архитектура компоненттерінің тиімділігін талдау 73
4.4 Қателерді талдау (Error Analysis) 74
4.5 Ұсынылған модель нәтижелерінің статистикалық маңыздылығын бағалау 78
4.6 Ұсынылған модельдің тиімділігін талқылау және оның практикалық қолданысы 81
4.7 4-тарау бойынша қорытынды 85
ҚОРЫТЫНДЫ 86
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ 89
ҚОСЫМША А 98

НОРМАТИВТІ СІЛТЕМЕЛЕР

Берілген диссертацияда келесі стандарттарға сәйкес сілтемелер қолданылды: Қазақстан Республикасы МЖМБС 5.04.034-2011. «Қазақстан Республикасы Мемлекеттік жалпыға білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура». Негізгі ережелер ҚР БҒМ бекітілген. 17.06.2011 ж. №261. Астана, 2011 ж.

«Диссертацияны безендіру нұсқаулығы», Қазақстан Республикасы БҒМ ЖАК 28 қыркүйек, 2004 жыл. №377-3 ж.

ГОСТ 7.32-2001. Ғылыми зерттеу жұмысының есебі. Безендіру ережелері мен құрылымы.

ГОСТ 7.1-2003. Библиографиялық таспа. Библиографиялық сипаттау.

ҚЫСҚАРТУЛАР МЕН БЕЛГІЛЕУЛЕР

АҚ – Ақпараттық қауіпсіздік
AI – Artificial Intelligence
BERT – Bidirectional Encoder Representations from Transformers
BiGRU – Bidirectional Gated Recurrent Unit
BiLSTM – Bidirectional Long Short-Term Memory
CLS – Classification Token
CNN – Convolutional Neural Network
CSV – Comma-Separated Values
DL – Deep Learning
FC – Fully Connected Layer
FN – False Negative – жалған теріс болжамдар саны
FP – False Positive – жалған оң болжамдар саны
GAT – Graph Attention Network
GCN – Graph Convolutional Network
GNN – Graph Neural Network
GPU – Graphics Processing Unit
GRU – Gated Recurrent Unit
HTML – HyperText Markup Language
KazBERT – Kazakh BERT
LSTM – Long Short-Term Memory
Macro-F1 – Macro-averaged F1 Score
MCC – Matthews Correlation Coefficient
MFA – Morphological Feature Augmentation
ML – Machine Learning
MLM – Masked Language Modeling
NLP – Natural Language Processing
PR-AUC – Area Under the Precision–Recall Curve
ReLU – Rectified Linear Unit
RNN – Recurrent Neural Network
ROC-AUC – Area Under the Receiver Operating Characteristic Curve
SEP – Separator Token
SupCon – Supervised Contrastive Learning
SVM – Support Vector Machine
TF-IDF – Term Frequency–Inverse Document Frequency
TN – True Negative
TP – True Positive
UPOS – Universal Part-of-Speech Tags

КІРІСПЕ

Ақпараттық-коммуникациялық технологиялар дамыған сайын ақпараттық кеңістіктегі қауіптердің сипаты да өзгерді. Бұрын зиянды контент шектеулі ақпарат көздері арқылы таралса, қазіргі цифрлық ортада ол әлеуметтік желілер, мессенджерлер, бейнеплатформалар және ашық интернет платформалары арқылы қысқа уақытта кең аудиторияға жетеді. Сондықтан деструктивті мәтіндерді дер кезінде анықтау бүгінгі күні тек техникалық мәселе емес, қоғамдық қауіпсіздік пен ақпараттық тұрақтылықты сақтаудың маңызды шартына айналды [1].

Деструктивті веб-контенттің қауіптілігі оның тек ашық агрессия немесе тікелей қорлау түрінде ғана көрінбейтіндігінде болып табылады. Мұндай контент кей жағдайда сарказм, ирония, астарлы сөз, жанама ишара немесе контекстке тәуелді мағына арқылы беріледі. Сондықтан деструктивті мәтінді автоматты түрде тану қарапайым сөздік сүзгілеу деңгейінен әлдеқайда күрделі міндет болып табылады. Бір ғана сөз әртүрлі контексте әртүрлі мағына беретіндіктен бір жағдайда ол бейтарап ақпараттық сипатта қолданылса, екінші жағдайда қорлау, кемсіту немесе өшпенділік мағынасын күшейтуі мүмкін [2-5].

Әлеуметтік платформалардағы веб-контент көлемі үздіксіз өсіп, күн сайын жаңарып отырады. Бұл жағдай деструктивті веб-контентті қолмен бақылауды күрделендіреді, себебі қауіпті жазбалар қысқа уақыт ішінде өңделіп үлгермей таралып кетуі мүмкін [6]. Мұндай жағдайда модератордың тәжірибесіне ғана сүйену жеткіліксіз, оның орнына деструктивті веб-контентті алдын ала байқап, мазмұнына қарай өз бетінше жіктей алатын құралдар қажет. Мұндай контент адамның эмоционалдық күйі мен мінез-құлқына әсер етіп қана қоймай, қоғам ішіндегі әлеуметтік шиеленісті де тереңдетуі мүмкін [7]. Дәл осы себептен оны уақтылы анықтап, алдын алу бүгінгі таңдағы аса маңызды міндеттердің біріне айналды.

Заңнама тұрғысынан алғанда, Қазақстанның коммуникация мен қоғамдық қауіпсіздік саласындағы құжаттары интернет кеңістігінде таралатын заңсыз, қауіпті және әлеуметтік тұрғыдан зиянды ақпаратты анықтау мен шектеу мәселелерін де қамтиды [8-10]. Бірақ қазіргі веб-контенттің көлемі мен жаңару жылдамдығы мұндай бақылауды тек қолмен жүргізуге мүмкіндік бермейді. Сондықтан деструктивті веб-контентті алдын ала анықтайтын, оны сараптамалық қарауға ұсынатын және ақпараттық қауіпсіздік шараларын жеделдетуге көмектесетін интеллектуалды жүйелерді әзірлеу ғылыми тұрғыдан да, практикалық тұрғыдан да маңызды.

Соңғы жылдары деструктивті контентті анықтау үшін машиналық оқыту және терең оқыту әдістері кеңінен қолданылып келеді. Дәстүрлі тәсілдер Naive Bayes, SVM сияқты әдістер мәтіндегі сөздердің жиілігі мен таралуына негізделіп, бастапқы классификациялау міндеттерінде қолданылды. Сонымен қатар CNN, LSTM, GRU және Transformer архитектуралары мәтіннің күрделірек құрылымын өңдеуге мүмкіндік берді [11-14]. Әсіресе BERT негізіндегі тілдік модельдер мәтіндегі сөздердің контекстік мағынасын тереңірек бейнелеу қабілетіне

байланысты табиғи тілді өңдеу міндеттерінде жоғары нәтиже көрсетіп келеді [14].

Қазіргі зерттеулердің көбі ағылшын, орыс, араб, неміс сияқты ресурсы мол тілдерге арналған. Ал қазақ тіліндегі деструктивті веб-контентті анықтау әлі толық зерттеле қойған жоқ [15-18]. Мұның себебі қазақ тілінің агглютинативті болуымен, сөздер сан түрлі морфологиялық формада келуімен, ал әлеуметтік желідегі жазба тілі тіпті бөлек болуымен түсіндіріледі. Интернеттегі қазақша мәтіндерде қате жазылған сөздер, қысқартулар, аралас тіл, транслитерация, сленг, астарлы қолданыс жиі ұшырасады. Аталған себептер бәрі мәтінді автоматты өңдеудің сапасына тікелей әсер етеді.

Қазақ тіліндегі деструктивті веб-контентті анықтауда контекстік модельдер сөз мағынасын оның мәтіндегі қолданысына қарай ажыратуға мүмкіндік береді. Алайда тек контекстік ақпаратқа сүйену жеткіліксіз, өйткені зиянды мағына әрдайым сөздің тура мағынасынан анық байқала бермейді. Кей жағдайда ол сөздің сөйлемдегі қызметі, грамматикалық тұлғасы, айтылымның бағыты және белгілі бір мағыналық санатпен байланысы арқылы көрінеді. Мысалы, «қорлау», «шабуыл», «ұлттық үстемдік» сөздері бір мәтінде деструктивті веб-контент белгісі ретінде қолданылса, басқа мәтінде сол әрекеттерді сынау немесе жоққа шығару мағынасында берілуі мүмкін [15]. Бұйрық райдағы етістіктер де үндеуді білдіруі ықтимал, бірақ олардың нақты мәні сөйлемнің жалпы контексіне байланысты анықталады. Сол себепті модель сөздің контекстік мағынасымен қатар мәтіндегі грамматикалық құрылымды, терістеуді және мағыналық бағытты да ескеруі қажет. Қолданыстағы модельдердің бір бөлігінде мұндай ақпарат классификациялау үдерісіне тікелей енгізілмейді немесе толық көлемде пайдаланылмайды [16-17].

Сондықтан қазақ тіліндегі деструктивті веб-контентті классификациялауда контекстік модельдерді морфологиялық белгілермен және онтологияға сүйенген семантикалық сипаттармен толықтыру қажет. Морфологиялық белгілер сөздердің грамматикалық құрылымын көрсетеді, ал онтологиялық семантикалық белгілер мәтіндегі терминдердің деструктивті контенттің түрі, ашық берілуі және нысанасы сияқты кластармен қалай байланысатынын тануға көмектеседі. Сонда мәтін жай сөздер қатары емес, тілдік те, мағыналық та қабаттары бар күрделі құрылым ретінде қаралады.

Деструктивті веб-контентті анықтауға арналған деректер жиынтықтары мен модельдердің көбі ресурсы мол тілдерге арналып жасалған. Сол себепті оларды қазақ тіліне сол күйінде көшіру көп жағдайда дұрыс нәтиже бермейді. Қазақ тілінің морфологиялық жүйесі, сөзжасамы, ұлттық-мәдени аясы және веб-мәтіндердегі бейресми қолданыстар басқа тілдерден едәуір бөлек. Сондықтан қазақ тіліне лайықталған, оның лингвистикалық табиғатын ескеретін арнайы модельдер керек. Бұл ғылыми жағынан да, қолданбалы жағынан да маңызды [18].

Берілген диссертациялық жұмыстың құндылығы қазақ тіліндегі деструктивті веб-контентті жоғары дәлдікпен классификациялайтын гибридті модель жасау қажеттілігімен түсіндіріледі. Зерттеу барысында әлеуметтік желілерден қазақ тіліндегі мәтіндер жиналып, деструктивті веб-контент корпусы

құрылады. Мәтіндер бейтарап, буллинг, нацизм және зорлық-зомбылық экстремизмі деген кластарға бөлініп қарастырылады. Бұл корпус кейін қазақ тіліндегі деструктивті веб-контентті автоматты түрде табуға және модельдерді өзара салыстыруға тірек болады.

Қазақ тіліндегі деструктивті веб-контентті классификациялау кезінде KazBERT арқылы алынған контекстік белгілер MFA модулінің морфологиялық сипаттарымен және онтологияға негізделген семантикалық белгілермен толықтырылады. Мәтін құрамындағы сөздер мен олардың өзара тізбектік байланысын ескеру үшін модель архитектурасына BiLSTM қабаты қосылады. Нәтижесінде классификациялау барысында сөздердің контекстік мағынасымен қатар олардың морфологиялық ерекшеліктері мен деструктивті веб-контентке қатысты семантикалық байланыстары да бірге ескеріледі. Алынған нәтижелер қазақ тіліндегі ақпараттық қауіпсіздік міндеттеріне арналған интеллектуалды талдау жүйелерін әзірлеуге қолданылуы мүмкін.

Зерттеу тақырыбының өзектілігі. Қазіргі цифрлық ортада деструктивті веб-контентті автоматты түрде анықтау ақпараттық қауіпсіздік пен қоғамдық тұрақтылық үшін қажет ғылыми және практикалық бағытқа айналды. Интернетте таралатын кибербуллинг, өшпенділікке құрылған пікірлер, агрессиялық үндеулер және құқыққа қайшы контент тек жеке адамға зиян келтіріп қоймайды, қоғамдағы әлеуметтік қатынастарға да әсер етеді. Сол үшін мұндай контентті уақытында байқап, ажырата алатын интеллектуалды жүйелерді жасау бүгін өте қажет міндеттердің бірі.

Зерттеудің өзектілігі заңнамалық талап, технологиялық әлсіз тұстар және лингвистикалық қиындықтар сияқты үш бағыттан көрінеді.

Заңнамалық қырына келсек, Қазақстан Республикасының Қылмыстық кодексінің 174-бабында [8] әлеуметтік, ұлттық, этностық, нәсілдік немесе діни алауыздықты қоздыруға байланысты әрекеттердің құқықтық қауіптілігі нақты көрсетілген. Бұдан бөлек, «Байланыс туралы» Қазақстан Республикасының заңында байланыс және интернет кеңістігінде құқыққа қайшы ақпараттың таралуын реттейтін нормалар бар. Демек, деструктивті веб-контентті дер кезінде анықтау тек техникалық мәселе емес, ұлттық ақпараттық қауіпсіздік пен құқықтық тәртіп үшін де қажет шарттардың бірі [9].

Технологиялық тұрғыдан алғанда көп зерттеу жұмыстары ағылшын, орыс және т. б. ресурсттық тілдерге қатысты жүргізіледі. Қазақ тіліндегі деструктивті веб-контентке қатысты дайын классификаторлар, жеткілікті аннотацияланған корпус немесе белгілі бір міндеттерге бағытталған жүйелер жоқ. Мәтіндерді автоматтандырылған түрде талдау үшін жаңа әдістер мен модельдер құру қажет екенін білдіреді.

Лингвистикалық жақтан мәселе одан да күрделірек. Қазақ тілі агглютинативті, морфологиясы күрделі тілдердің бірі. Сөздің мағынасы түбірге жалғанған қосымшалар арқылы өзгеріп отырады, ал сөйлемнің реңкі көбіне контекстке, семантикалық байланысқа тәуелді. Оның үстіне деструктивті веб-контент көп жағдайда тура айтылмайды: астармен беріледі, эмоциялық реңкке жасырынады, кейде жанама тілдік құрылымдармен жеткізіледі. Сондықтан қазақ тіліндегі веб-контентті талдайтын модель тек контексті емес, морфологиялық

және семантикалық ерекшеліктерді де қатар ескеруі керек.

Сол себепті берілген диссертациялық жұмыс тақырыбының өзектілігі деструктивті интернет-контенттің қоғамға қауіп төндіретіндігімен, оны құқықтық тұрғыдан дер кезінде анықтау қажеттілігімен, қазақ тіліне арналған технологиялық шешімдер жеткіліксіздігімен түсіндіріледі. Аталған жағдайлар қазақ тіліндегі деструктивті веб-контентті анықтауға арналған жетілдірілген модель жасаудың ғылыми да, практикалық та мәні бар екенін көрсетеді.

Диссертациялық жұмыстың мақсаты. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін тереңдетіп оқыту модельдерін кешенді зерттеу және құру.

Зерттеудің міндеттері:

1. Әлеуметтік желі платформаларынан қазақ тіліндегі деструктивті мәтіндердің деректер жинағын құру;
2. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін мәтінді ендіру әдістері мен классикалық машиналық оқыту және терең оқыту әдістерін зерттеу;
3. Қазақ тілі үшін морфологиялық белгілер модулін әзірлеу;
4. Қазақ тіліндегі деструктивті веб-контентті классификациялауда онтологияға негізделген семантикалық белгілерді әзірлеу;
5. Қазақ тілінің ерекшеліктерін ескеретін деструктивті веб-контенттің түрлері мен белгілерін классификациялайтын гибридті архитектураны құру;
6. Әзірленген модельді эксперименталды ортада сынау.

Зерттеу объектісі: Әлеуметтік желілер мен веб-ресурстардағы қазақ тіліндегі деструктивті мәтіндік контент.

Зерттеу пәні. Қазақ тіліндегі деструктивті веб-контентті көпкласты классификациялауға арналған контекстік, морфологиялық және онтологияға негізделген семантикалық белгілерді біріктіру әдістері мен модельдері.

Зерттеу әдістері. Машиналық және терең оқыту әдістері, мәтінді классификациялау әдістері, табиғи тілді өңдеу әдістері, әлеуметтік желілерді талдау әдістері, онтологиялық классификациялау әдістері.

Диссертациялық жұмыстың ғылыми жаңалығы:

1. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін деструктивті веб-контенттің көп классты аннотацияланған корпус құрылды.
2. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін қазақ тіліндегі деструктивті веб-контенттің түрін, ашықтығын және нысанасын анықтайтын онтологияға негізделген семантикалық белгілер әзірленді.
3. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін контексттік эмбеддингтерді, қазақ тілінің морфологиялық белгілерін және онтологияға негізделген семантикалық белгілерді біріктіретін гибридті классификациялық архитектура әзірленді.

Қорғауға шығарылған негізгі тұжырымдама.

1. Қазақ тіліндегі деструктивті веб-контентті көпкласты классификациялау міндетін шешу үшін деструктивті веб-контент корпусы әзірленіп, контенттің түрін, ашықтығын және нысанасын сипаттайтын онтологияға негізделген семантикалық белгілерді KazBERT эмбеддингтерімен және морфологиялық

белгілермен біріктіру әдісі ұсынылды, ұсынылған әдісті қолдану базалық KazBERT-пен салыстырғанда Accuracy көрсеткішін 88,015%-дан 90,492%-ға, ал Macro F1-score көрсеткішін 88,086%-дан 90,567%-ға дейін арттыруға мүмкіндік берді.

2. Қазақ тіліндегі деструктивті веб-контентті көпкластық классификациялауға арналған KazBERT эмбеддингтері, морфологиялық белгілер, онтологияға негізделген семантикалық белгілер және BiLSTM негізіндегі тізбектік талдау қабаты біріктірілген гибридті архитектура әзірленіп, базалық KazBERT-пен салыстырғанда Accuracy көрсеткіші 88,015%-дан 91,045%-ға, ал Macro F1-score көрсеткіші 88,086%-дан 91,103%-ға дейін арттырылып, Accuracy бойынша салыстырмалы қате классификациялау үлесі 25,28%-ға төмендетілді.

Зерттеу нәтижесінде қазақ тіліндегі деструктивті веб-контенттің аннотацияланған деректер корпусы қалыптастырылып, қазақ тілінің морфологиялық ерекшеліктері мен деструктивті контент санаттары арасындағы семантикалық байланыстарды ескеретін KazBERT, морфологиялық ерекшеліктерді кеңейту модулі (MFA) және онтологияға негізделген семантикалық белгілер негізіндегі гибридті архитектура ұсынылды. Ұсынылған тәсіл мәтіннің контекстік, грамматикалық және семантикалық ерекшеліктерін кешенді түрде модельдеуге мүмкіндік беріп, аз ресурсты агглютинативті тілдерге арналған деструктивті контентті автоматты анықтау жүйелерін дамытуға ғылыми негіз қалайды. Сонымен қатар зерттеу нәтижелерін Қазақстан Республикасының ақпараттық қауіпсіздік саласындағы нормативтік-құқықтық талаптарын іске асыру мақсатында әлеуметтік желілер мен онлайн платформалардағы қазақ тіліндегі деструктивті контентті автоматты мониторингтеу жүйелерінде, контентті модерациялау құралдарында, білім беру платформаларында, ақпараттық-талдамалық жүйелерде және қауіпті веб-контентті ерте анықтауға арналған интеллектуалды шешімдерде пайдалануға болады.

Диссертациялық жұмыс нәтижелерін апробациялау. Ұсынылған диссертациялық жұмыстың зерттеу нәтижелері әл-Фараби атындағы Қазақ Ұлттық университеті Ақпараттық технологиялар және жасанды интеллект факультетінің, Киберқауіпсіздік және криптология кафедрасынның ғылыми семинарларында бірнеше рет тыңдалып, талқыланды. Сонымен қатар, халықаралық және отандық ғылыми конференцияларда баяндама жасалып, ғылыми қауымдастық тұрғысынан оң бағаланды.

Ғылыми зерттеу жұмысының нәтижелерінің негізділігі мен сенімділік дәрежесі. Диссертациялық жұмыстың нәтижелері 10 ғылыми басылымда жарияланды. 1 мақала Scopus базасында индекстелетін басылымда, 3 мақала Scopus базасында индекстелетін конференцияларда, 2 мақала ҚР БҒМ Білім және ғылым саласындағы бақылау комитеті ұсынған журналдарында жарық көрді. 4 мақала халықаралық конференциялар материалдарында жарияланып, ҚР авторлық құқықпен қорғалатын объектілерге құқықтарды мемлекеттік тізілімге мәліметтерді енгізу туралы 1 куәлік алынды.

Қазақстан Республикасы Ғылым және жоғары білім министрлігінің

Ғылым және жоғары білім саласындағы сапаны қамтамасыз ету комитеті ұсынатын журналдардағы мақалалар:

1. Болатбек М.А., Сағынай М., Мусиралиева Ш.Ж. Қазақ тіліндегі деструктивті веб-контентті анықтау үшін машиналық оқыту әдістерін қолдану //Известия НАН РК. Серия физико-математическая. – 2024. – №. 4. – С. 99–111. https://doi.org/10.32014/2024.2518-1726.310

2. Сағынай М., Болатбек М. Қазақ тіліндегі деструктивті веб-мазмұнды жіктеудің морфологиялық және семантикалық тұрғыдан жетілдірілген моделі //Вестник КазАТК. – 2026. – Т. 144. – №. 3. – С. 151-166.

Скопуста индекстелетін журналдардағы ғылыми мақалалар:

1. Bolatbek M, Sagynay M, Mussiraliyeva S, Yeltay Z. 2025. Detection of offensive content in the Kazakh language using machine learning and deep learning approaches. PeerJ Computer Science 11:e3027 https://doi.org/10.7717/peerj-cs.3027 (Indexed in Web of Science: Q2, Indexed in SCOPUS, SJR 2024 равен 0.719, процентиль 80)

Скопуста индекстелетін конференциялардағы ғылыми мақалалар:

1. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Yeltay Zhastay, Machine learning methods for identifying destructive texts in the Kazakh language 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology (ADMIT 2024) https://doi.org/10.1145/3701100.3701121.

2. Milana Bolatbek, Moldir Sagynay, Shynar Mussiraliyeva, Kymbat Baisylbayeva, Zhastay Yeltay. Kazakh language dataset for hate speech detection on social media text, 2024 9th International Conference on Computational Intelligence and Applications (ICCIA 2024) 10.1109/ICCIA62557.2024.10719327.

3. Moldir S., BolatbekM., Mussiraliyeva S., & Baisylbayeva, K. (2025, March). Natural Language Processing and Machine Learning for Detecting Harmful Speech in Kazakh Social Media. In 2025 11th International Conference on Computing and Artificial Intelligence (ICCAI) (pp. 347-351). IEEE 10.1109/ICCAI66501.2025.00060

Халықаралық конференция материалындағы мақалалар:

1. Сағынай М., Болатбек М. Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған терең оқыту үлгілерін жасау, Материалы международной конференции студентов и молодых учёных «Фараби әлемі»

2. Болатбек, М. Сағынай М., Ш. Мусиралиева Веб ресурстағы зиянды веб контентті анықтау мәселесі және оның өзектілігі Материалы IX Международной научно-практической конференции «Информатика и прикладная математика», посвященной 90-летию Казахского национального университета имени аль-Фараби

3. Сағынай М., Болатбек М., Қазақ тіліндегі қауіпті веб-контентті анықтау және жіктеу үшін машиналық оқыту модельдерін құру Фараби әлемі 2025, 265 бет.

4. Bolatbek M., Mussiraliyeva S., Sagynay M. Machine Learning-Based Hate Speech Detection in the Kazakh Language // Research in Language. – 2025. – Vol. 23. – P. 311–319. – DOI: https://doi.org/10.18778/1731-7533.23.19.

Авторлық құқықпен қорғалатын объектілерге құқықтардың

мемлекеттік тізіліміне мәліметтерді енгізу туралы куәліктер:

1. Болатбек М.А., Сағынай М. «Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған BERT негізіндегі терең оқыту модельдерінің тиімділігін Supervised Contrastive Learning (SupCon) әдісі арқылы арттыру», №69171, күні: 27.03.2026 ж.

Жұмыстың көлемі мен құрылымы. Диссертациялық жұмыс кіріспеден, 4 бөлімнен, қорытынды, әдебиеттер тізімі және қосымшалардан тұрады. Диссертацияның толық көлемі: 98 бет машиналық мәтіні, оның ішіне 26 сурет, 21 кесте, 107 пайдаланылған әдебиеттер тізімі және 1 қосымша кіреді.

1 Қазақ тіліндегі деструктивті веб-контентті анықтаудың теориялық және әдістемелік негіздері

Бірінші тарауда деструктивті веб-контент ұғымы мен оның таралу ерекшеліктері сипатталып, қазақ тіліндегі деструктивті веб-контентті автоматты түрде классификациялаудың теориялық және әдістемелік негіздері талданады. Дәстүрлі машиналық оқыту және терең оқыту әдістерінің мүмкіндіктері мен шектеулері салыстырылып, зерттеу есебі тұжырымдалады.

1.1 Деструктивті веб-контент ұғымы, оның ақпараттық қауіпсіздікке әсері және қазақ тіліндегі деструктивті веб-контенттің ерекшеліктері

Интернет пен әлеуметтік желілер ақпаратты жариялау және тарату мүмкіндіктерін кеңейтті. Пайдаланушылар хабарламаларды өзара бөлісіп, өз парақшаларында немесе басқа платформаларда қайта жариялайды, осылайша ақпарат бастапқы аудиториядан тыс таралады. Жеке тұлғаға, әлеуметтік топқа немесе қоғамға зиян келтіруі мүмкін контент те осы жолмен таралады. Зерттеуде деструктивті веб-контент ретінде интернетте жарияланатын, жеке тұлғаны не әлеуметтік топты кемсітетін, өшпенділікті қоздыратын немесе зорлық-зомбылыққа шақыратын ақпарат түсіндіріледі.

Деструктивті веб-контенттің ақпараттық қауіпсіздікке төндіретін қатері оның адамның көзқарасы мен мінез-құлқына тигізуі мүмкін теріс ықпалына байланысты болады. Ақпараттық ортадағы зиянды ықпалдың бір мысалын жалған ақпарат жөніндегі зерттеулерден көруге болады. Жалған мәлімет адамның оқиға туралы түсінігін бұрмалауы мүмкін, ал оның пайымдауға әсері сол мәлімет теріске шығарылғаннан кейін де сақталуы ықтимал [19]. Өшпенділікке негізделген онлайн контенттің қаупі басқа қырынан көрінеді. [20] жұмыста белгілі бір әлеуметтік жағдайларда мұндай контенттің нақты өмірдегі зорлық-зомбылықпен байланысы көрсетілген. Бұл байланысты барлық жағдайға ортақ заңдылық ретінде емес, зерттеу жүргізілген әлеуметтік орта шегінде қарастырған жөн.

Контенттің таралу сипатына пайдаланушылар арасындағы байланыс та әсер етеді. Әлеуметтік желіде көзқарастары ұқсас пайдаланушылар бір ортаға бірігіп, өз ұстанымдарына сәйкес келетін ақпаратпен жиі алмасуы мүмкін. Соның нәтижесінде сол аудиторияның пікіріне жақын хабарламалар бірнеше рет қайта бөлісіледі [21]. Алайда мұндай өзара әрекеттесу сипаты платформаның құрылымы мен пайдаланушылар арасындағы байланыс ерекшеліктеріне қарай өзгеріп отырады.

Зерттеу жұмыстарында мәтіннің тілдік сипаты мен оның қайта таралу жиілігінің арақатынасы да қарастырылады. Моральдық және саяси мәселелерге қатысты хабарламаларды талданатын [22] жұмыста моральдық бағалау мен эмоцияны қатар білдіретін сөздер кездесетін мәтіндердің көп жағдайда қайта жарияланатыны көрсетілген. Бұл тұжырымды деструктивті веб-контенттің барлық түріне тікелей қолдануға болмайды, өйткені ол нақты тақырыптағы

хабарламаларды талдау нәтижесіне негізделген, дегенмен мұндай деректер контентті зерттеу кезінде оның мағынасымен ғана шектелмей, эмоциялық реңкі мен таралатын әлеуметтік ортасын да ескеру қажеттігін көрсетеді.

Деструктивті веб-контенттің негізгі таралу арналары, қолданылатын тетіктері және ықтимал зияны кесте 1.1-де салыстырылған.

Кесте 1.1 - Деструктивті веб-контенттің таралу арналары мен ықтимал зияны

Таралу арнасы	**Таралу механизмі**	**Ықтимал зиян**
Әлеуметтік желілер мен ұсыныс алгоритмдері	Контент лайктар, бөлісулер және алгоритмдік ұсыну арқылы таралады. Ұсыныс алгоритмдері эмоциялық тұрғыдан әсерлі материалдардың таралуын күшейтуі мүмкін.	Жылдам әрі жаппай таралу, эхо-камералардың қалыптасуы, қайталап ұсыну арқылы радикалды көзқарастардың күшеюі.
Мессенджерлер мен жабық топтар	Контент шифрланған чаттар мен жеке топтардың ішінде айналып, негізінен сыртқы модерациядан тыс қалады.	Осал пайдаланушылар арасында байқалмай таралу, мемлекеттік бақылаудың қиындығы, мақсатты психологиялық манипуляция.
Тікелей эфир платформалары	Бейне және аудиоматериалдар қысқа клиптер немесе тікелей эфир арқылы кең аудиторияға тез жетеді.	Күшті эмоциялық әсер, зорлық-зомбылық пен қауіпті мінез-құлықты қалыпты құбылыс ретінде қабылдату, әсіресе кәмелетке толмағандар арасындағы еліктеу.
Анонимді форумдар мен тіркелмеген веб-сайттар	Контент автор немесе платформа иесі туралы мәліметтерді шектеулі көлемде көрсете отырып не мүлдем көрсетпей орналастырылады.	Өзіне зиян келтіру немесе радикалды идеология төңірегінде жабық қауымдастықтардың қалыптасуы, құқық қорғау органдары үшін контенттің дереккөзін анықтаудың қиындығы.
Бот желілері мен үйлестірілген жалған аккаунттар	Автоматтандырылған немесе жалған аккаунттар бірдей контентті қайта жариялап, көпшіліктің жаппай қолдауы бар деген әсер қалыптастырады.	Қоғамдық ақпараттық кеңістіктің бұрмалануы, шеткі көзқарастардың жасанды күшеюі, онлайн пікірталасқа деген сенімнің әлсіреуі.
Шекарааралық және диаспоралық арналар	Контент шетелде өндіріліп немесе орналастырылып, кейін интернеттің ұлттық сегментінде қайта таратылады.	Ұлттық шектеу механизмдерін айналып өту, күрделі юрисдикция, тұрақты экстремистерді тарту.

Контентке қол жеткізу мүмкіндігі де платформа мен топтың ұйымдастырылуына байланысты. Ашық жарияланымдар кең аудиторияға қолжетімді болса, жабық топтардағы хабарламаларды көру үшін сол топқа қосылу талап етіледі. Мессенджерлерде ашық және жабық арналардың екеуі де кездеседі, сондықтан оларды бірыңғай жабық ақпараттық орта ретінде сипаттау дұрыс емес. Мысалы, Telegram-дағы кей топтарды зерттеу олардың арасында

идеологиялық және ұлттық белгілер бойынша қалыптасқан желілік байланыстар бар екенін көрсетеді [23]. Сондықтан мұндай платформалардағы деструктивті веб-контентті талдағанда арналар мен топтардың ашықтық деңгейін, аудитория құрамын және олардың өзара желілік байланыстарын жеке ескерген жөн.

Пайдаланушылардың өзара ақпарат алмасуымен қатар контентті автоматтандырылған аккаунттар арқылы үйлестіріп тарату тетіктері де қолданылады. Бот желілері бір хабарламаны бірнеше аккаунттан қайталап жариялайды. Осылайша оның ақпараттық ортада көріну жиілігі жасанды түрде артады, ал мұндай белсенділік белгілі бір пікірді көптеген адам қолдап отырғандай әсер туғызып, пайдаланушылардың қоғамдық ұстаным туралы түсінігін бұрмалауы мүмкін [24].

Кесте 1.2-де ақпараттық қауіпсіздікке әсер етуші үш арнаның бірге жұмыс істейтіні көрсетіледі, олар бір уақытта қоғамның үш деңгейінде әрекет етеді. Жеке деңгейдегі салдары психологиялық зиян мен жеке құндылықтардың бұрмалануы, топтық деңгейдегі салдары әлеуметтік поляризация және жеке адамдар мен институттар арасындағы сенімнің жоғалуымен түсіндіріледі.

Кесте 1.2 - Деструктивті веб-контенттің негізгі ықпал ету бағыттары

Ықпал ету бағыты	Ғылыми сипаттамасы	Деструктивті контенттің көрініу формалары
Когнитивтік ықпал	Ақпаратты қабылдау мен интерпретациялау үдерісін бұрмалау арқылы пайдаланушының ақпарат туралы түсінігіне әсер ету	Жалған ақпарат, манипуляциялық ақпарат, адастырушы хабарламалар
Эмоциялық ықпал	Пайдаланушыда қорқыныш, үрей, өшпенділік немесе басқа да жағымсыз эмоциялық реакцияларды қалыптастыру	Өшпенділік сөздері, қорқыныш тудыратын контент, қорлау, қоқан-лоқы
Мінез-құлықтық ықпал	Пайдаланушының әрекетіне әсер ету, белгілі бір мінез-құлық үлгісін қалыптастыру немесе белгілі бір әрекетке ынталандыру	Зорлық-зомбылыққа шақыру, экстремистік үндеулер, агрессивті әрекетке итермелеу
Ақпараттық ықпал	Ақпаратты бұрмалау, контекстен үзіп ұсыну немесе жалған мәліметтерді тарату арқылы ақпараттық ортаның контентіне әсер ету	Фейк жаңалықтар, жалған деректер, ақпараттық манипуляция
Әлеуметтік-коммуникациялық ықпал	Контенттің әлеуметтік желілердегі таралуы мен пайдаланушылар арасындағы байланыстар арқылы кең аудиторияға жеткізілуі	Репосттар, үйлестірілген ақпараттық науқандар, боттар, желілік таралу

Деструктивті веб-контентті модерациялау мәселесі ұлттық ақпараттық қауіпсіздікпен де байланысты. Қазақстан Республикасының Президенті интернет-платформалардың мұндай контентті таратқаны үшін жауапкершілігін заңнамалық тұрғыдан бекіту мәселесін көтерді. Мемлекет басшысы жалған ақпарат пен ақпараттық шабуылдардың қоғамдық пікірге ықпалын азаматтардың қауіпсіздігі мен қоғамдық тұрақтылығына төнетін қатерлермен байланыстырды.

Деструктивті веб-контенттің қоғамдық сенім мен әлеуметтік тұрақтылыққа ықпалы ғылыми зерттеулерде де қарастырылған [25]. Әлеуметтік желілердегі мәтіндерді бастапқы деңгейде талдауда автоматты анықтау жүйелерін қолдану осы тұрғыдан өзекті. Олар мәтін құрамындағы деструктивті веб-контент белгілерін анықтап, хабарламаларды тиісті кластар бойынша бөлуге арналған. Мұндай өңдеу нәтижесі модерация кезінде қосымша сараптауды қажет ететін хабарламаларды алдын ала іріктеуге қолданылуы мүмкін.

Қазақ тіліндегі деструктивті веб-контентті автоматты анықтау әдістерін әзірлеу бірнеше деректік және тілдік ерекшелікпен байланысты. Солардың бірі – осы міндетке бейімделген сапалы аннотацияланған корпустардың жеткіліксіздігі. Қазақ тілінің агглютинативті құрылымы бір сөздің әртүрлі грамматикалық сипатта кездесуіне әкелетіндіктен, бұл ерекшелік мәтінді өңдеу барысында ескеріледі. Әлеуметтік желі мәтіндерінде бұған бейресми жазу үлгісі және бір хабарлама ішінде бірнеше тілдің қатар қолданылуы, яғни код-алмастыру құбылысы қосылады (Сурет 1.1). Осы ерекшеліктер қазақ тіліндегі әлеуметтік желі мәтіндерін тек әдеби тіл нормалары шегінде қарастыру жеткіліксіз екенін көрсетеді. Сондықтан модельдерді әзірлеу барысында нақты онлайн қолданысқа тән тілдік ерекшеліктерді де ескеру қажет [26].



Сурет 1.1 - Қазақ тілінің деструктивті веб-контенттегі орны

Қазақ тіліндегі әлеуметтік желі пікірлерінде емле нормасынан ауытқулар, қысқартулар және жаргондық қолданыстар жиі кездеседі. Кейбір сөздер транслитерациямен беріледі, сөйлемдерде тыныс белгілері де әрдайым әдеби тіл нормасына сәйкес қойыла бермейді. Мұндай жазылым ерекшеліктері сөздердің шекарасын дұрыс анықтауға және олардың морфологиялық тұлғасын тануға әсер етуі мүмкін. Қысқартулар мен жаргондық бірліктердің мағынасы көбіне олардың сөйлем ішіндегі қолданысына байланысты анықталатындықтан, семантикалық өңдеу кезінде контексті ескеру қажет.

1.2 Деструктивті веб-контентті анықтауда қолданылатын табиғи тілді өңдеу, дәстүрлі машиналық оқыту және терең оқыту әдістері

Деструктивті веб-контентті автоматты анықтау мәтіндерді белгілі бір кластар бойынша ажыратуға негізделеді. Бұл міндетті шешуде табиғи тілді өңдеу тәсілдері машиналық және терең оқыту әдістерімен бірге қолданылады. Қазақ тіліндегі әлеуметтік желі мәтіндерін автоматты өңдеуге арналған зерттеулерде деректерді дайындау мен модельді тілдің құрылымдық ерекшеліктеріне бейімдеу мәселелері қарастырылған [27]. Мұндай мәтіндермен жұмыс істеу арнайы аннотацияланған деректер жиынтығын қалыптастыруды, мәтінді алдын ала өңдеуді және қазақ тілінің ерекшеліктерін ескеретін модельдерді қолдануды талап етеді.

Деструктивті веб-контентті анықтауда Naive Bayes, Support Vector Machines, Logistic Regression және Random Forest сияқты дәстүрлі машиналық оқыту әдістері де қолданылады. Бұл модельдер мәтіннің өзімен тікелей емес, одан алынған сандық белгілермен жұмыс істейді: мәтін белгілер векторына түрлендіріліп, оның класы сол белгілер негізінде анықталады. Ал классификация нәтижесін түсіндіру деңгейі қолданылған белгілерге және таңдалған модельдің құрылымына байланысты болады. Бұл әдістер шағын көлемдегі деректер жиынтығында да белгілі бір деңгейде нәтиже бере алады.

Зерттеуде мәтінді сандық сипаттамалар түрінде ұсыну үшін TF-IDF әдісі пайдаланылды ((1.1) формула):

$$x_{ij} = TF(t_i, d_j) \times IDF(t_i) \tag{1.1}$$

Мұндағы,

t_i – термин,

d_j – құжат,

x_{ij} – d_j құжатындағы t_i терминінің TF-IDF салмағы;

TF(t_i, d_j) – терминнің осы құжаттағы жиілігін сипаттайтын шама;

IDF(t_i) – терминнің кері құжаттық жиілігі.

$DF(t_i)$ – термин кездесетін құжаттар саны. Терминнің кері құжаттық жиілігі (1.2) формуламен анықталады:

$$IDF(t_i) = log \frac{N}{DF(t_i)} \tag{1.2}$$

Мұндағы,

N – барлық құжат саны.

Naive Bayes әдісі белгілердің классқа қатысты шартты тәуелсіздігі туралы болжамға сүйенеді. Мәтін үшін апостериорлық ықтималдығы ең жоғары класс (1.3) формула бойынша таңдалады:

$$\hat{y} = argmax_c P(c) \prod_{i=1}^{n} P(x_i|c) \quad (1.3)$$

Мұндағы,

c – класс,

x_i – мәтіннің i-белгісінің мәні,

$P(c)$ – c класының априорлық ықтималдығы.

n – белгілер саны;

P(x_i | c) – c класы берілгендегі x_i белгі мәнінің шартты ықтималдығы;

$\hat{y}$ – болжанған класс.

Сызықтық Support Vector Machine әдісінің шешім функциясы (1.4) формуласы арқылы беріледі:

$$f(x) = w^T x + b \quad (1.4)$$

Екі класты сызықтық SVM-де класс таңбасы шешім функциясының таңбасына қарай анықталады ((1.5) формула). Бұл жағдайда кластар −1 және +1 түрінде кодталады:

$$\hat{y} = sign(w^T x + b) \quad (1.5)$$

Мұндағы,

w – салмақ векторы,

x – сипаттама векторы,

b – ығысу параметрі.

Бинарлық Logistic Regression моделі мәтіннің оң класқа жату ықтималдығын (1.6) формуламен бағалайды:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (1.6)$$

Мұндағы,

$P(y = 1 \mid x)$ – берілген мәтіннің деструктивті веб-контент класына жату ықтималдығы;

бұл бинарлық мысалда $y = 1$ – деструктивті веб-контент, $y = 0$ – бейтарап мәтін,

w – салмақ векторы,

b – ығысу параметрі. Нәтиже 0 мен 1 аралығындағы ықтималдық түрінде беріледі.

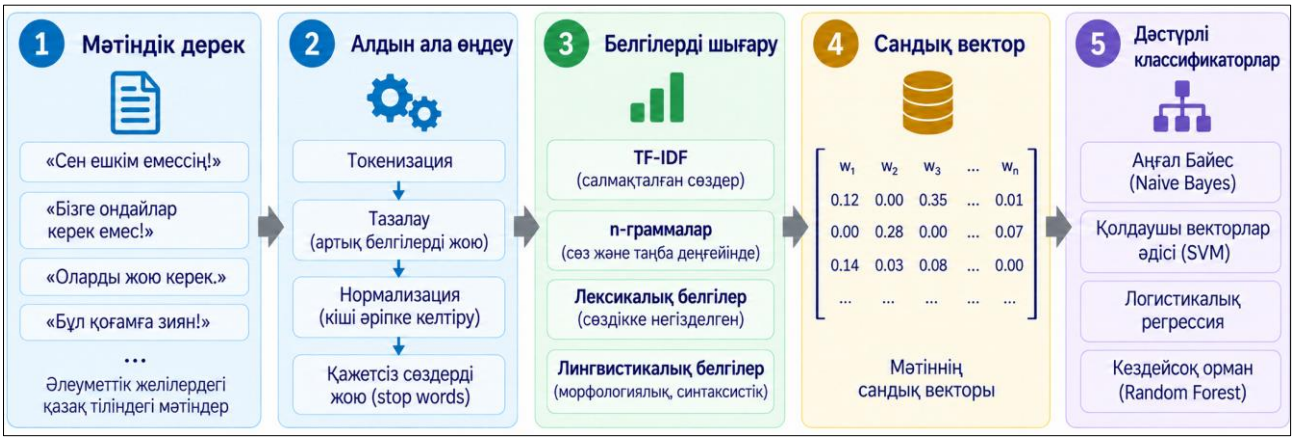
Мұнда x – мәтін белгілерінің векторы. (1.5) және (1.6) формулалары әдістердің екі класты жағдайдағы жұмысын түсіндіреді.

Random Forest әдісі бірнеше шешім ағашының нәтижесін біріктіреді. Классикалық дауыс беру нұсқасында болжам (1.7) формуласымен анықталады:

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_T(x)\} \quad (1.7)$$

Мұндағы,
$h_t(x)$ – t-ағаштың x мәтіні үшін болжаған класы;
T – ағаштар саны;
mode – ағаштар болжамдары арасында ең жиі кездесетін класс.

Сурет 1.2-де дәстүрлі машиналық оқыту негізіндегі мәтінді өңдеу құрылымы көрсетілген. Енгізілген мәтін алдымен алдын ала өңдеуден өтеді: токенизация, тазарту, нормалау және стоп-сөздерді жою. Одан кейін TF–IDF, n-граммалар, лексикалық және лингвистикалық белгілер негізінде сандық белгі векторы қалыптастырылады, ал алынған вектор дәстүрлі машиналық оқыту әдістерімен классификациялау үшін пайдаланылады.



Сурет 1.2 - Дәстүрлі машиналық оқыту негізіндегі мәтінді өңдеу құрылымы

Дәстүрлі машиналық оқыту әдістерінде контекстің қаншалықты ескерілетіні мәтінді сипаттау үшін таңдалған белгілерге байланысты. Мысалы, жеке сөздердің жиілігіне негізделген тәсілдер сөздердің сөйлем ішіндегі орналасу ретін сақтамайды. N-граммалар белгілі бір ұзындықтағы сөз тізбектерін қамту арқылы жергілікті байланыстарды ішінара ескергенімен, астарлы немесе бейнелі мағынаны әрдайым ажырата алмайды [28]. Соған қарамастан, бұл әдістер архитектурасының салыстырмалы түрде қарапайым болуымен, есептеу ресурстарына қойылатын талаптарының төмендігімен және алынған нәтижелерді интерпретациялау мүмкіндігімен ерекшеленеді. Алайда мәтіндегі ұзақ қашықтықтағы семантикалық байланыстарды, контекстке тәуелді мағыналарды және жасырын не ирониялық сипаттағы деструктивті мазмұнды анықтау кезінде олардың мүмкіндігі шектеулі болып қалады.

Терең оқыту модельдерінің ерекшелігі классификацияға қажет белгілерді оқыту деректерінен автоматты түрде қалыптастыруы болып табылады. Мұндай жағдайда модель алдын ала қолмен белгіленген сипаттармен шектелмей, мәтіннің семантикалық және құрылымдық ерекшеліктерін бейнелейтін ішкі

көріністерді үйренеді. Соның нәтижесінде белгілерді алдын ала қолмен іріктеуге тәуелділік төмендейді.

Мәтінді өңдеуде CNN, RNN және Transformer архитектурасына негізделген модельдер әртүрлі міндеттер үшін қолданылады. CNN мәтіндегі қысқа тіркестер мен жергілікті үлгілерді анықтауға бейім болса, рекуррентті модельдер сөздердің тізбектік байланысын өңдеуге бағытталған.

LSTM, GRU және BiLSTM осы рекуррентті модельдер қатарына жатады және сөздер арасындағы тәуелділіктерді ескеруге мүмкіндік береді. Дегенмен ұзақ тізбектерде маңызды ақпаратты сақтау және бір-бірінен алшақ орналасқан сөздердің байланысын қамту рекуррентті модельдер үшін қиындық туғызуы мүмкін [28]. Transformer архитектурасында мәтін элементтері арасындағы байланыстар назар аудару механизмі арқылы өңделеді. Сондықтан бұл архитектураның контексті ескеру тәсілі рекуррентті желілердің тізбектік өңдеуінен өзгеше.

Өшпенділік сөздер мен деструктивті веб-контентті анықтау оңай міндет емес. Бұл мәселенің лингвистикалық та, әлеуметтік те қыры бар [29]. Өшпенділік тілі ұлттық, этникалық, діни, нәсілдік, жыныстық немесе өзге де әлеуметтік белгілерге негізделіп, белгілі бір адамға немесе топқа бағытталуы мүмкін. Мұндай мәтіндер әртүрлі формада көрінеді: агрессия кейде ашық түрде білдірілсе, кейде жанама меңзеу немесе бағалаушылық сипаттағы тұспалдау арқылы жеткізіледі. Сондықтан жеке сөздер деңгейіндегі талдау автоматты классификациялау үшін жеткіліксіз болып табылады. Сенімді жіктегіш мәлімдеменің нысанасын, сөйлем құрылымын және мәтіннің жалпы мазмұндық бағытын да ескеруі тиіс. Деструктивті веб-контенттің кейбір түрлері жалған немесе адастырушы ақпаратпен де тығыз байланысты. Мұндай контентті анықтауға арналған зерттеулер, әдетте, бірнеше белгіге сүйенеді: мәтіннің мазмұны, дереккөздің сенімділік деңгейі және ақпараттың таралу үлгісі. Әлеуметтік желілердегі байланыстар да маңызды белгілердің бірі ретінде қарастырылады [30].

BERT архитектурасы пайда болғаннан бері деструктивті контентті анықтау саласындағы басым тәсіл біржола өзгерді. Дәстүрлі эмбеддинг тәсілдерінде әрбір сөзге тұрақты вектор беріліп, ол контекстке қарамастан өзгермей қалатын. BERT мұндай шектеуден арылды: сөздің мағынасы оны қоршаған сөздерге қарай динамикалық түрде есептеледі. Нәтижесінде бір сөздің әртүрлі сөйлемдегі әртүрлі мағынасын модель дәлірек ажырата алады.

Дегенмен бұл артықшылық қазақ тілі үшін автоматты түрде іске аса салмайды. Қазақ сөздері қосымшалар арқылы жиі түрленеді, ал сөздер арасындағы мағыналық байланыс та күрделі болып келеді. Сол себептен зерттеуде базалық BERT емес, оның қазақ тіліне бейімделген нұсқасы – KazBERT таңдалды. Ол мәтіндегі контекстік мағынаны тереңірек талдауға мүмкіндік береді. Дәл осы қасиет, KazBERT деструктивті веб-контентті автоматты классификациялау үшін қолайлы модель ретінде таңдауға түрткі болды [14]. Кесте 1.3-те BERT және KazBERT модельдерінің артықшылықтары мен кемшіліктерін салыстыру көрсетілген.

Кесте 1.3 - BERT және KazBERT модельдерінің артықшылықтары мен кемшіліктерін салыстыру

Салыстыру өлшемі	**BERT**	**KazBERT**
Модельдің мақсаты	Модель нұсқасына қарай ағылшын тіліндегі немесе көптілді тілдік міндеттерде қолданылады	Қазақ тіліндегі мәтіндерді өңдеуге бейімделген
Тілдік ерекшелік	Қазақ тілінің ерекшеліктерін арнайы ескермейді	Қазақ тілінің лексикалық және грамматикалық ерекшеліктеріне бейімделген
Оқыту корпусы	Ағылшын тіліндегі немесе бірнеше тілдегі мәтіндер; нақты құрам модель нұсқасына байланысты	Қазақ тіліндегі мәтіндер негізінде алдын ала оқытылған
Контексті модельдеу	Сөздің мағынасын контекст арқылы анықтайды	Қазақ тіліндегі контекстік байланыстарды модельдеуге бағытталған
Қазақ тілінің агглютинативтілігі	Арнайы бейімделмегендіктен, сөздердің көптеген морфологиялық формаларын өңдеуде шектеулер болуы мүмкін	Қазақ тілінің агглютинативті құрылымына бейімделуі модельдің қазақша мәтінді өңдеуіне мүмкіндік береді
Токенизация	Таңдалған BERT нұсқасының токенизаторы қолданылады	Қазақ тіліндегі сөз құрылымына бейімделген токенизация тиімдірек болуы мүмкін
Артықшылығы	Көптеген зерттеулерде кеңінен қолданылған; архитектурасы жақсы зерттелген; дайын құралдар мен модельдер көп	Қазақ тіліне бағытталған; қазақша мәтіннің контекстік және лингвистикалық ерекшеліктерін ескеруге бейімделген
Кемшілігі	Қазақ тіліндегі сирек сөздер мен морфологиялық формаларды өңдеу кезінде қосымша қиындықтар туындауы мүмкін	Қазақ тіліне тәуелді; басқа тілдерге қолдану мүмкіндігі шектеулі
Деректердің жеткіліктілігі	Ірі тілдер үшін үлкен көлемді оқыту корпустары бар	Қазақ тіліндегі сапалы деректер көлемі салыстырмалы түрде шектеулі
Деструктивті веб-контентті анықтау	Қазақ мәтіндеріне қосымша бейімдеуді қажет етеді	Қазақ тіліндегі деструктивті веб-контентті классификациялау міндетіне тікелей бейімдеуге қолайлы
Морфологиялық ақпарат	Морфологиялық белгілерді арнайы енгізбейді	Қазақ тілінің морфологиялық ерекшеліктерін модельдеуге ыңғайлы

Деструктивті веб-контентті анықтайтын модельдердің бір әлсіз тұсы олардың бейімделу қабілетінің шектеулі екендігінде. Бір ортада жоғары нәтиже көрсеткен модель екінші тілде немесе басқа платформада қате жіберуі әбден мүмкін [31].

Бұл әлеуметтік желідегі мәтіннің өзіндік табиғатымен түсіндіріледі, ол жерде орфографиялық қателер мен қысқартулар жиі кездеседі, эмоциялық боямасы басым сөздер мен бейресми қолданыстар да аз емес. Осының салдарынан мәтінді өңдеу процесі айтарлықтай күрделене түседі. Кесте 1.4-те қазақ тіліндегі деструктивті веб-контентті анықтауда дәстүрлі машиналық оқыту және терең оқыту әдістерінің салыстыруы келтірілген.

Кесте 1.4 - Қазақ тіліндегі деструктивті веб-контентті анықтауда дәстүрлі машиналық оқыту мен терең оқытуды салыстыру

Салыстыру өлшемі	**Дәстүрлі машиналық оқыту**	**Терең оқыту**
Негізгі әдістер	Naive Bayes, SVM, Logistic Regression, Random Forest	CNN, RNN, LSTM, BiLSTM, GRU, Transformer
Мәтінді ұсыну	TF-IDF, n-граммалар, қолмен таңдалған белгілер	Векторлық көріністерді модельдің өзі үйренеді
Белгілерді анықтау	Көбіне қолмен жасалған немесе алдын ала таңдалған белгілер	Жасырын мәтіндік белгілер автоматты түрде оқытылады
Семантикалық байланыстарды модельдеу	Қолданылатын белгілерге тәуелді	Контекстік көріністер арқылы семантикалық байланыстарды модельдей алады
Морфологияны ескеру	Морфологиялық ақпаратты қосымша белгілер арқылы енгізуге болады	Морфологиялық заңдылықтарды контекст арқылы жанама түрде үйрене алады
Контексті өңдеу	Сөздер арасындағы байланыстарды ескеру қолданылатын белгілерге тәуелді	LSTM/BiLSTM және Transformer ұзақ контексті модельдей алады
Қазақ тілінің агглютинативтілігі	Сөздердің көптеген формалары модель үшін қиындық туғызады	Контекстік және морфологиялық ерекшеліктерді модельдеуге икемдірек
Деректерге қойылатын талап	Салыстырмалы түрде аз деректермен жұмыс істей алады	Нөлден оқытуда дерек көлеміне талап жоғары; алдын ала оқытылған модельді бейімдеу белгіленген деректерге қажеттілікті азайтуы мүмкін
Есептеу ресурстары	Салыстырмалы түрде төмен	Жоғары GPU/есептеу ресурстарын қажет етуі мүмкін
Модельді түсіндіру	Алгоритмге байланысты; сызықтық модельдерді түсіндіру әдетте жеңілірек	Күрделірек
Деструктивті веб-контентті анықтау	Лексикалық үлгілерді анықтауда тиімді	Күрделі семантикалық және контекстік үлгілерді анықтауға мүмкіндік береді
Қазақ тіліне бейімдеу	Морфологиялық және лингвистикалық белгілерді бөлек енгізуге болады	KazBERT сияқты тілдік модельдер арқылы қазақ тілінің контекстік ерекшеліктерін пайдалануға болады

Шетел тілдері үшін дайындалған модельдерді қазақ тіліндегі деректерге өзгеріссіз қолдану әрдайым жеткілікті нәтиже бермейді. Себебі қазақ тілінің грамматикалық құрылымы мен семантикалық ерекшеліктері модельдің мәтінді қалай өңдейтініне әсер етеді [32]. Бұл талаптар деструктивті веб-контентті классификациялау кезінде де ескерілуі қажет.

Қазақ тіліндегі сөздерге әртүрлі қосымшалардың жалғануы олардың грамматикалық формасын және кей жағдайда сөздің мағынасын өзгертеді.

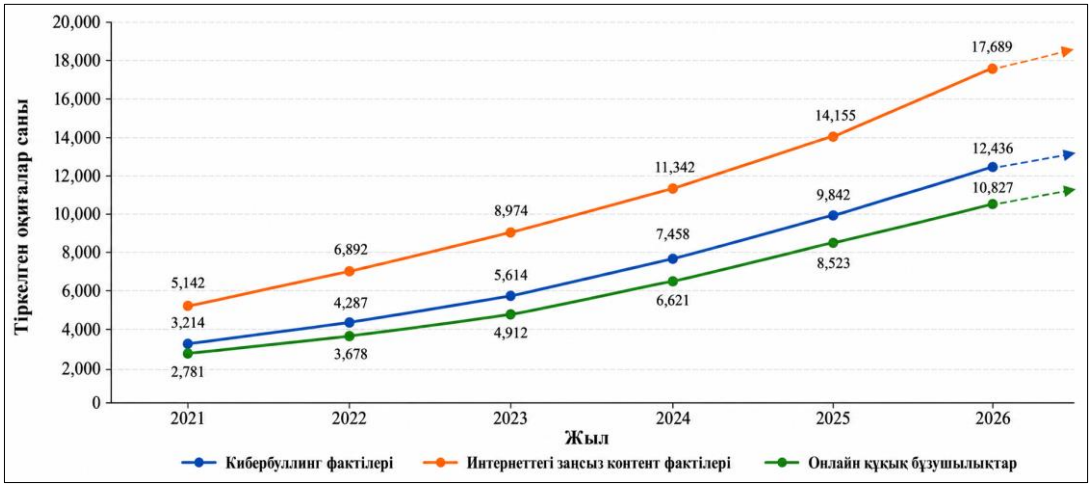
Сондықтан мәтінді өңдеу кезінде сөздің түбірін ғана емес, оның құрамындағы сөзжасамдық және грамматикалық қосымшаларды да ажырату маңызды. Морфологиялық талдау арқылы сөздің құрамы, оған жалғанған қосымшалар және грамматикалық сипаттары анықталады. Мұндай ақпаратты алудың бір тәсілі – ережеге негізделген морфологиялық анализаторларды қолдану [33].

Дәстүрлі машиналық оқыту әдістерінде классификация алдын ала қалыптастырылған белгілерге сүйенсе, терең оқыту модельдері мәтіннің семантикалық және құрылымдық сипаттарын оқыту деректерінен үйренеді. Қазақ тіліндегі сөз тұлғаларының көптүрлілігіне байланысты бұл модельдерді морфологиялық ақпаратпен толықтыру орынды. Сонда деструктивті веб-контентті классификациялау кезінде мәтіннің семантикалық сипаттарымен бірге сөздердің грамматикалық белгілері де ескеріледі.

1.3 Қазақ тіліндегі деструктивті веб-контентті классификациялаудың ғылыми мәселелері және зерттеу есебінің қойылымы

Әлеуметтік желілерде өшпенділік тілі, зорлық-зомбылыққа негізделген экстремизм және буллинг белгілері кездесетін веб-контентті автоматты анықтау ақпараттық қауіпсіздік саласында қарастырылатын міндеттердің қатарына жатады. Сурет 1.3-те деструктивті веб-контентке байланысты құқық бұзушылықтардың 2021-2026 жылдар аралығындағы динамикасы көрсетілген.

Қазақ тіліндегі мұндай мәтіндерді классификациялау кезінде мәселе тек тілдің құрылымдық ерекшеліктерімен шектелмейді. Деструктивті мағына көп жағдайда мәтіннің контексіне байланысты анықталатындықтан, модель осы ерекшелікті де ескеруі тиіс. Бұған осы міндетке арналған аннотацияланған деректердің шектеулілігі қосылып, модельді әзірлеу және оқыту үшін қолжетімді деректер көлемін шектейді.



Сурет 1.3 - Интернет кеңістігіндегі деструктивті веб-контентке байланысты құқық бұзушылықтардың өсу динамикасы (qamqor.gov.kz)

Қазақ тіліндегі деструктивті веб-контентті автоматты түрде классификациялау табиғи тілді өңдеу мен ақпараттық қауіпсіздік салалары

тоғысындағы міндеттердің бірі. Машиналық және терең оқыту әдістерін қазақ тіліндегі мәтіндерге қолдануда негізгі қиындықтар тілдің құрылымдық ерекшеліктері мен арнайы аннотацияланған деректердің жеткіліксіздігіне байланысты. Қазақ тілінің агглютинативті құрылымы бір сөздің әртүрлі морфологиялық тұлғада кездесуіне әкелетіндіктен, мәтінді өңдеу барысында осы ерекшелікті ескеру қажет [34]. Сондықтан қолданылатын әдістер қазақ тілінің морфологиялық құрылымына бейімделуі тиіс.

Қазақ тіліндегі деструктивті веб-контентті классификациялауға қатысты негізгі ғылыми мәселелер сурет 1.4-те жинақталған.



Сурет 1.4 - Қазақ тіліндегі деструктивті веб-контентті классификациялаудың ғылыми мәселелері

Қазақ тілінің агглютинативті құрылымына байланысты бір түбірге әртүрлі қосымшалар жалғанып, бірнеше сөз тұлғасы қалыптасады. Мұндай тұлғалардың бір бөлігі мәтінде сирек кездесетіндіктен, сөз деңгейіндегі белгілер бойынша тұрақты статистикалық сипаттама алу қиындайды. Бұл жағдайда деректердің сиректігі мәтін санының аздығын емес, сөз тұлғаларының әрқайсысына қатысты мысалдардың жеткіліксіздігін білдіреді. Осы қиындықты азайту үшін сөздерді морфемаларға бөлуге негізделген морфологиялық сегментация қарастырылады.

Морфемаларға бөлу деструктивті мағынаны өздігінен анықтай алмайды. Бір сөз бейтарап хабарламада да, қорлау немесе зорлық-зомбылыққа шақыру мәтінінде де қолданылуы мүмкін. Сондықтан классификациялау кезінде морфологиялық белгілермен бірге сөздердің контекстік мағынасы ескерілуі қажет. Қазақ тіліндегі деструктивті веб-контентті классификациялаудың негізгі ғылыми мәселелері кесте 1.5-те жинақталған.

Кесте 1.5 - Қазақ тіліндегі деструктивті веб-контентті классификациялаудың негізгі ғылыми мәселелері

Ғылыми мәселе	**Негізгі ерекшелігі**	**Классификацияға әсері**
Агглютинативті құрылым және морфологиялық өзгергіштік	Қазақ тілінде бір түбірге бірнеше қосымша жалғанып, сөздің көптеген морфологиялық формалары түзіледі. Сөздің морфологиялық құрылымын талдау кезінде фонологиялық және морфотактикалық ережелерді ескеру қажет [35].	Бір түбірдің әртүрлі морфологиялық формаларда кездесуі мәтіндік белгілердің тұрақты түрде алынуын қиындатады және деректердің сиректігін арттырады. Морфологиялық сегментация бұл мәселені азайтуға мүмкіндік береді.
Контекстің рөлі және көпмағыналылық	Бір сөз әртүрлі контексте әртүрлі мағына беруі мүмкін. Деструктивті веб-контент тек ашық агрессиямен шектелмей, мысқыл, астарлау, стереотип немесе жасырын қорқыту түрінде де беріледі [15, 36].	Жекелеген сөздерге негізделген классификация мәтіннің толық мағынасын әрдайым қамти бермейді. Сондықтан контекстік және семантикалық байланыстарды ескеру қажет.
Семантикалық жақындық және көпкласты классификация	Деструктивті контент санаттарының шекаралары бір-бірінен анық ажырамауы мүмкін. Классификация кезінде мәтіннің түрі, нысанасы және агрессия деңгейі сияқты белгілер маңызды [37].	Бір мәтін бірнеше деструктивті санатқа тән белгілерді қамтуы мүмкін. Сондықтан тек жалпы сипаттамаларға емес, нақты деструктивті ішкі санатты ажыратуға мүмкіндік беретін семантикалық белгілер қажет.
Бір мәтінде бірнеше деструктивті белгінің қатар кездесуі	Бір мәтінде бірнеше деструктивті белгілер қатар кездесуі мүмкін. Мысалы, этникалық кемсітушілік элементтері зорлық-зомбылыққа шақыру белгілерімен бірге берілуі ықтимал [38].	Бір мәтін бірнеше санаттың белгілерін қамтыған жағдайда, оны бір класқа жатқызудың нақты критерийлері қажет. Классификация кезінде мәтіндегі белгілердің өзара байланысы мен басым семантикалық мазмұны ескеріледі.
Ашық және жасырын деструктивті веб-контент	Деструктивті контент ашық түрде де, жанама немесе жасырын түрде де беріледі. Жасырын контентті анықтау үшін мәтіннің жалпы контекстін, нысанасын және мәтін бөліктері арасындағы семантикалық байланыстарды ескеру қажет [39].	Жасырын мағынаны тек сөздік немесе жеке лексикалық белгілер арқылы анықтау қиын. Сондықтан контекстік және семантикалық тәуелділіктерді модельдейтін әдістер қажет.

Берілген зерттеу үшін ашық интернет-ресурстардан жиналған қазақ тіліндегі мәтіндер пайдаланылады. Қойылатын есеп – әрбір мәтінді контенттіне қарай алдын ала белгіленген төрт кластың біріне жатқызу: буллинг, нацизм, зорлық-зомбылық экстремизмі немесе бейтарап мәтіндер. Бейтарап класқа зерттеуде қабылданған аннотациялау өлшемдері бойынша деструктивті

белгілері анықталмаған мәтіндер енгізіледі. Модельдің нәтижесі әр мәтінге тағайындалатын қорытынды класс белгісі түрінде беріледі.

Осы мақсатқа жету үшін төмендегі міндеттер қойылады.

1. Қазақ тіліндегі деструктивті және бейтарап мәтіндер корпусын қалыптастыру. Әлеуметтік желі платформаларынан зерттеу мақсатына сәйкес қазақ тіліндегі мәтіндер жиналады. Олар алдын ала өңделіп, белгіленген төрт класс бойынша аннотацияланады. Қалыптастырылған корпус модельдерді оқытуға, олардың классификациялау сапасын бағалауға және нәтижелерін өзара салыстыруға арналған ортақ деректер негізі ретінде пайдаланылады.

2. Мәтінді векторлық түрде ұсыну тәсілдерін және классификациялау әдістерін салыстырмалы талдау. Мәтіндік эмбеддингтерді қалыптастыру тәсілдері, дәстүрлі машиналық оқыту модельдері және терең оқыту архитектуралары қарастырылады. Олардың қазақ тіліндегі деструктивті веб-контентті классификациялаудағы мүмкіндіктері мен шектеулері салыстырылып, мәтінді векторлық түрде ұсыну тәсілі анықталады. Осы талдау кейінгі эксперименттерде салыстыруға қолданылатын базалық модельді таңдауға негіз болады.

3. Қазақ тіліне арналған морфологиялық белгілер модулін әзірлеу. Мәтіндегі сөздердің грамматикалық сипаттамаларын анықтау үшін арнайы бағдарламалық модуль әзірленеді. Морфологиялық белгілер құрамына сөз табы, шақ, жақ және рай енгізіледі. Бұл ақпарат классификациялау моделінде қосымша белгі ретінде пайдаланылады, ал оның модель нәтижесіне ықпалы эксперименттік салыстыру арқылы бағаланады.

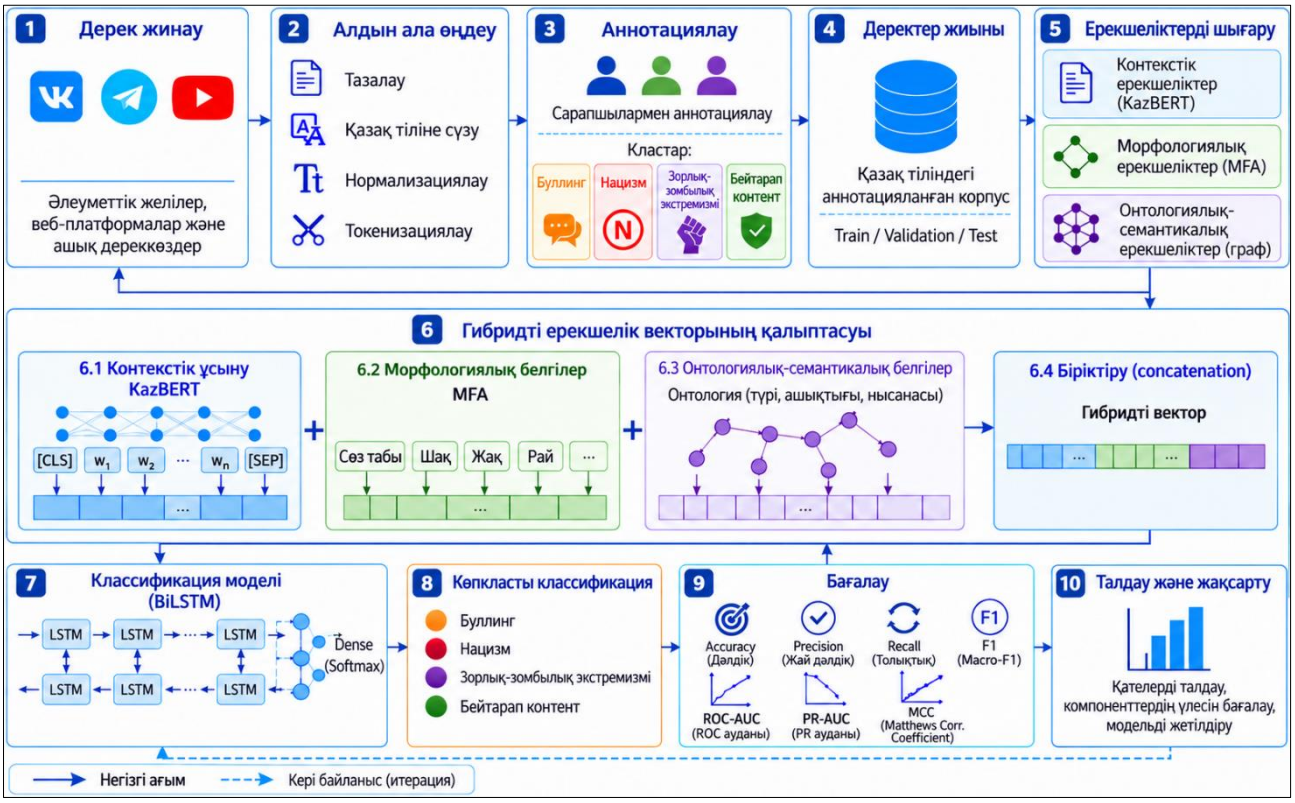
4. Онтологияға негізделген семантикалық белгілерді қалыптастыру. Деструктивті веб-контентке қатысты ұғымдар контенттің түрі, ашықтығы және нысанасы бойынша топтастырылады. Бұл өлшемдер мазмұнның қандай түрге жататынын, оның ашық немесе жасырын түрде берілуін және кімге бағытталғанын сипаттайды. Осы құрылым негізінде онтологиялық лексикон қалыптастырылып, ұғымдар арасындағы байланыстар онтологиялық графта көрсетіледі. Мәтіндегі терминдер тиісті онтологиялық ұғымдармен сәйкестендірілгеннен кейін, алынған сәйкестіктер семантикалық белгілерге түрлендіріліп, классификациялау моделіне қосылады.

5. Қазақ тілінің ерекшеліктерін ескеретін гибридті архитектура әзірлеу. KazBERT моделінен алынған контекстік көріністер морфологиялық белгілермен және онтологияға негізделген семантикалық белгілермен біріктіріледі. Гибридті архитектура осы ақпаратты бір модельде өңдеп, мәтіннің қорытынды класын анықтауға арналған. Мұнда контекстік көріністер мәтіндегі сөздердің қолданысын сипаттаса, қосымша белгілер олардың грамматикалық қасиеттері мен онтологиялық ұғымдарға қатысын ескеруге мүмкіндік береді.

6. Әзірленген модельдердің классификациялау сапасын эксперименттік тұрғыдан бағалау. Нәтижелерді өзара салыстыру үшін барлық модельдерге ортақ оқыту, валидациялау және тестілеу жиынтықтары пайдаланылады. Классификациялау сапасы Accuracy, Macro-F1, Precision, Recall, ROC-AUC, PR-AUC және MCC көрсеткіштерімен бағаланады. Гибридті модельдің нәтижелері

базалық KazBERT моделімен және зерттеуде қарастырылған басқа модельдердің нәтижелерімен салыстырылады.

Эксперименттерде архитектураның жекелеген компоненттерінің классификациялау сапасына қосқан үлесі анықталады. Сонымен қатар контекстік көріністерді морфологиялық және онтологияға негізделген семантикалық белгілермен толықтырудың мағыналық тұрғыдан жақын кластарды ажыратуға әсері талданады. Зерттеу есебін шешуге ұсынылатын әдістің негізгі кезеңдері сурет 1.5-те көрсетілген.



1.5 сурет - Қазақ тіліндегі деструктивті веб-контентті көпкластық классификациялаудың әдістемелік сұлбасы

Мәтіннің контекстік белгілері KazBERT арқылы алынады, морфологиялық сипаттамалары MFA модулінің көмегімен анықталады. Сонымен бірге деструктивті веб-контентке қатысты онтология негізінде семантикалық белгілер қалыптастырылады. Контекстік, морфологиялық және семантикалық белгілер бір сипаттамалық векторға біріктіріліп, BiLSTM архитектурасына беріледі. Соңғы кезеңде модельдің классификация сапасы таңдалған бағалау метрикалары бойынша анықталады.

1.4 1-тарау бойынша қорытынды

Бұл тарауда деструктивті веб-контенттің негізгі түрлері мен оның ақпараттық қауіпсіздікке ықпалы қарастырылып, мұндай мәтіндерді автоматты классификациялауда қолданылатын дәстүрлі машиналық және терең оқыту

әдістері салыстырылды. Талдау барысында аталған әдістердің мүмкіндіктерімен қатар қазақ тіліндегі мәтіндерге қолдану кезінде кездесетін шектеулері де анықталды.

Қазақ тіліндегі деструктивті веб-контентті өңдеуде тілдің агглютинативті құрылымы мен деструктивті мағынаның контекске тәуелділігі ерекше ескерілуі тиіс. Мұндай контент мәтінде ашық түрде ғана емес, жасырын түрде де берілуі мүмкін. Сонымен бірге арнайы аннотацияланған корпустардың шектеулілігі және семантикалық белгілерді анықтаудың күрделілігі қазақ тіліндегі мәтіндерді классификациялау кезінде шешуді қажет ететін мәселелер қатарына жатады.

Аталған ерекшеліктер зерттеудің мақсаты мен міндеттерін нақтылауға негіз болды. Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін мәтіннің контекстік көріністерін морфологиялық және онтологияға негізделген семантикалық белгілермен біріктіретін әдістемені әзірлеу зерттеудің негізгі бағыттарының бірі ретінде белгіленді.

2 Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған әдісті әзірлеу

Бұл тарауда қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған мәтіндік корпусты дайындау және одан қажетті белгілерді алу әдістемесі қарастырылады. Корпусты қалыптастыру мәтіндерді жинау, алдын ала өңдеу және аннотациялау кезеңдерінен тұрады. Дайындалған деректер негізінде морфологиялық және онтологиялық білім модельдері қолданылып, мәтіннен грамматикалық және семантикалық белгілерді алу тәсілдері сипатталады.

2.1 Қазақ тіліндегі деструктивті веб-контент корпусын құру және аннотациялау әдістемесі

Зерттеу үшін қазақ тіліндегі деструктивті және бейтарап мәтіндерден тұратын корпус қалыптастырылды. Материалдарды іріктеу кезінде олардың зерттеу тақырыбына мазмұндық сәйкестігі, мәтін сапасы және қарастырылатын кластарды қамтуы ескерілді [39-41].

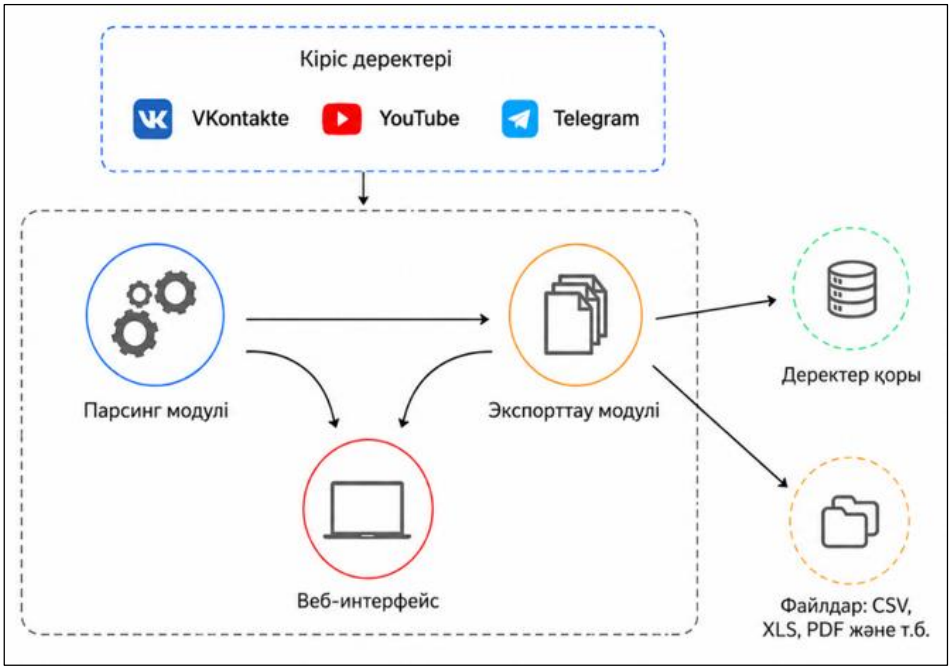
Мәтіндер қазақ тіліндегі ашық интернет-ресурстардан жиналды. Негізгі дереккөздер ретінде Telegram арналары, YouTube платформасындағы пікірлер және VKontakte желісіндегі талқылаулар алынды [42-43]. Бұл платформалардағы жазбаларда ресми емес тілдік қолданыстар, қысқа пікірлер және пайдаланушылар арасындағы еркін коммуникация үлгілері жиі кездеседі. Осындай материалдар қазақ тіліндегі онлайн қарым-қатынасқа тән мәтіндік ерекшеліктерді зерттеуге негіз болды [44].

Деректер арнайы парсинг модулі арқылы жиналды [45]. Барлық алынған материал корпус құрамына бірден енгізілген жоқ, қайталанған жазбалар, жарнамалық және спам хабарламалар, мазмұны жеткіліксіз мәтіндер, сондай-ақ зерттеу тақырыбына қатысы жоқ материалдар алынып тасталды. Сүзгілеуден өткен мәтіндер кейінгі алдын ала өңдеу мен аннотациялау кезеңіне жіберілді.

Корпусты қалыптастырудың техникалық кезеңдері сурет 2.1-де көрсетілген. Ашық ресурстардан мәтіндерді жинау және бастапқы өңдеу парсинг модулі арқылы орындалады, ал өңделген деректер экспорттау модулінің көмегімен құрылымдалған түрде сақталады. Осы өңдеуден өткен мәтіндер аннотациялау кезеңіне жіберіледі.

Парсер арқылы жиналған материалдарды бастапқы сұрыптау кезінде мазмұндық критерийлер қолданылды [45]. Олардың мақсаты мәтінге түпкілікті класс тағайындау емес, деструктивті веб-контент белгілері байқалатын материалдарды кейінгі аннотациялауға іріктеу болды. Мәтіннің нақты классы алдын ала сұрыптаудан кейін сараптамалық бағалау арқылы анықталды [46-47].

Деструктивті мәтіндерді бастапқы іріктеу кезінде қолданылған мазмұндық белгілер кесте 2.1-де берілген.



Сурет 2.1 - Қазақ тіліндегі веб-мәтіндерді жинау және корпусқа дайындау үдерісінің жалпы схемасы

Корпусқа деструктивті белгілері анықталмаған бейтарап мәтіндер де енгізілді. Бұл мәтіндер бейтарап класты қалыптастырып, оны деструктивті веб-контентке жататын басқа кластардан ажыратуға пайдаланылды.

Кесте 2.1 - Корпусқа мәтіндерді таңдау және аннотациялау критерийлері

№	Мәтін санаты	Белгісі	Мысал
1	Нацизм класы этникалық үстемдік идеяларын, ұлттық кемсітушілікті және нацистік идеологияны қолдауды білдіретін мәтіндерді қамтиды.	Шығу тегіне байланысты кемсіту, басқа ұлттың сөйлеу ерекшелігін немесе дәстүрін мазақ ету.	1.Біздің ұлт өзгелерден жоғары, сондықтан елді басқару құқығы тек бізге тиесілі. 2.Адамдардың құқықтары шыққан тегіне қарай бөлінуі керек, басқа ұлт өкілдеріне бізбен тең құқық беруге болмайды.
2	Зорлық-зомбылық экстремизмі – идеологиялық мақсаттарға жету үшін зорлық-зомбылықты қолдау, ақтау немесе оған шақырумен байланысты құбылыс. Веб-контентте ол күш қолдануды дәріптеу немесе зорлық әрекеттеріне үндеу арқылы көрінуі мүмкін.	Халықты үрейлендіру мақсатындағы террористік әрекеттер, интернетте «әділдік үшін күрес» деген желеумен зорлық-зомбылыққа шақыру; физикалық күшті немесе қоқан-лоқыны қолданатын радикалды қозғалыстар.	1.Біздің идеологиямызды қабылдамайтындарға күш қолдануды ақтаймын. 2. Саяси мақсатымызға жету үшін бейбіт адамдарға жасалған шабуылдарды қолдаймын.

Кесте 2.1 жалғасы

3	Буллинг – адамды қорлау, кемсіту немесе оқшаулау мақсатында оған психологиялық, эмоциялық не әлеуметтік қысым көрсетумен сипатталатын қасақана және жүйелі агрессивті мінез-құлық. Оның идеологиялық негізі болуы міндетті емес.	Қорлаушы жәбірленушіні саналы түрде таңдап, әрекеттерін жүйелі түрде қайталайды; ұжымнан шеттету, басқалардың көзінше мазақ ету, өсек тарату, ат қою, кибербуллинг (интернетте қорлау, беделге нұқсан келтіретін мәліметтерді жариялау).	1.Сенен бәрібір түк шықпайды, күн сайын сыныптың чатында сені мазақтаймыз. 2. Сені топқа қоспаймыз, басқаларға да сенімен сөйлеспеуді күнде айтып отырамыз.
4	Бейтарап – қорлау, кемсіту, этникалық үстемдікті насихаттау немесе зорлық-зомбылықты қолдау белгілері жоқ ақпараттық және күнделікті қарым-қатынас мәтіндері.	Ақпаратты хабарлау немесе оқиғаны сипаттау, адамға не топқа бағытталған агрессияның болмауы, деструктивті құбылыстарды қолдаусыз, танымдық немесе алдын алу мақсатында талқылау.	1. Ертең университетте ғылыми жобаларға арналған семинар өтеді. 2. Тарих сабағында нацизмнің пайда болуы мен оның зардаптары талқыланды.

Бұл кезеңде мәтіндерді тек кілт сөздер арқылы іріктеу жеткіліксіз болды, өйткені бір сөздің мағынасы оның қолданылған контексіне қарай өзгеруі мүмкін. Мысалы, «шабуыл», «қару», «қорлау», «ұлт» сөздері бір мәтінде деструктивті веб-контенттің белгісі ретінде кездессе, басқа мәтінде жаңалықтық немесе талдамалық бейтарап мазмұнда қолданылуы мүмкін. Сондықтан бастапқы іріктеу кезінде жеке сөздермен қатар мәтіннің жалпы семантикалық мазмұны мен олардың қандай мақсатта қолданылғаны да ескерілді [48].

Жиналған деректер кейін бірнеше кезең бойынша алдын ала өңделді. Алдымен мәтіндер техникалық элементтерден тазартылып, артық бос орындар, қайталанатын таңбалар, қажетсіз сілтемелер, HTML қалдықтары және мәтіндік талдауға қатысы жоқ басқа элементтер алынып тасталды [49]. Екінші кезеңде бір мәтіннің бірнеше рет қайталануы тексеріліп, қайталанатын жазбалар жойылды [27]. Үшінші кезеңде мәтіннің қазақ тіліне қатыстылығы қарастырылды. Өзге тілдегі немесе мағыналық талдауға жарамсыз мәтіндер корпус құрамынан өшірілді. Төртінші кезеңде мәтіндер құрылымдалған кестелік форматқа келтіріліп, кейінгі аннотациялау мен модель оқытуға дайындалды [47].

Алдын ала өңделген корпус қазақ тіліндегі деструктивті веб-контентті классификациялау моделін әзірлеудің эмпирикалық негізі ретінде пайдаланылды [50]. Төменде корпустың көлемі, кластар бойынша үлестірімі және лингвистикалық сипаттамалары берілген [51].

1. Корпустың жалпы статистикалық сипаттамасы

Корпусқа Telegram, VKontakte және YouTube платформаларынан парсер арқылы жиналған мәтіндер енгізілді [48]. Деректер CSV форматында сақталды. Корпус төрт кластан тұрады: буллинг, зорлық-зомбылық экстремизмі, нацизм және бейтарап мәтіндер [52].

Берілген зерттеуде расизмнің жеке класс ретінде қарастырылмауы жіктеу санаттарының шекарасын нақтылау қажеттілігімен негізделеді. Расизм мен нацизм дербес ұғымдар болғанымен, олардың мәтіндік көріністерінде шығу тегіне байланысты кемсіту, этникалық үстемдікті насихаттау және белгілі бір топқа өшпенділік білдіру сияқты белгілер ішінара қабаттасуы мүмкін. Мұндай қабаттасу әр мәтінге бір ғана класс тағайындалатын классификациялау міндетінде аннотациялау кезіндегі екіұштылыққа әкелуі ықтимал. Сондықтан зерттеу аясын шектеу және белгілеу критерийлерінің бірізділігін қамтамасыз ету мақсатында соңғы классификациялау схемасы бейтарап мәтіндер, буллинг, зорлық-зомбылық экстремизмі және нацизм кластарымен шектелді. Бұл ретте расизм мен нацизмнің ұғымдық айырмашылығы сақталып, расизмді дербес анықтау міндеті осы зерттеудің аясынан тыс қалдырылды.

Жиналған мәліметтердің үлестірімі төмендегідей:

– буллинг класында – 3234 мәтін;

– зорлық-зомбылық экстремизмі класында – 3088 мәтін;

– нацизм класына – 3074 мәтін;

– бейтарап класына – 2663 мәтін.

Жалпы корпус көлемі – 12 059 мәтін.

Корпустың көлемі мен әр кластағы мәтін саны семантикалық модельді оқыту үшін жеткілікті теңгерімді жиынтықты қамтамасыз етеді [53]. Датасет мөлшері және құрамы кесте 2.2-де көрсетілген.

Кесте 2.2 - Датасет мөлшері және құрамы

	label	**message**
290	violent	заң жүйесін толық жою керек сот ғимараттарын бомбалап ішіндегілерді тірідей өртеу керек себебі ол әділет емес зорлық…
…	…	…
297	violent	түрмелердегі заңдар жалған сондықтан түрмелерді жарып тұтқындарды босатып күзетшілерді өлтіру керек…
…	…	…
903	neutral	енді бірінші болып жазбаймын көрейік мені қашан еске алатындарыңды
…	…	…
1325	neutral	тауға шыққан сапар табиғатқа жақындата түсті…
…	…	…
4278	bullying	сен мұны түсіне алмайсың бактерия жазғаның дым түсініксіз...
…	…	…
4688	bullying	сен тыңдаған адамды шаршатасың себебі сөзің ұзақ мәнің қысқа
…	…	…
5616	nazism	біз кез келген жолмен ұлттық социалистік әлемді құрамыз...
…	…	…
12059	nazism	ұлттық қауіпсіздікке заңсыз иммигранттар қауіп төндіреді бірақ бұл ұлттық құзырет

2. Лингвистикалық ерекшеліктерді талдау

Корпусты лексикалық деңгейде талдау деструктивті мәтіндерге тән агрессия маркерлерінің, кемсітушілік лексемалардың, қорлау мақсатындағы

тілдік бірліктердің жиі қолданылатынын көрсетті. Буллинг бағытындағы мәтіндерде ирония, келемеж, әдейі төмендету, балағат сөздер жиі байқалады. Зорлық-зомбылық пен экстремистік дискурс көбіне қорқыту, күш көрсету, қоқан-лоқы сипатындағы етістіктер мен бұйрық рай формаларын пайдалануымен ерекшеленеді.

Нацизм класына жатқызылған мәтіндерде дәріптеу реңкі бар сөздер, қарсы топқа бағытталған кемсіту үлгілері және насихаттық құрылымдар кездеседі. Бейтарап мәтіндерде ақпараттық мазмұн басым болуы мүмкін, алайда эмоциялық реңктің болуы өздігінен деструктивтілік белгісі саналмайды [52-53].

3. Семантикалық-прагматикалық талдау

Корпус семантикалық тұрғыдан әр категорияның өзіндік мағыналық өрістерін айқын көрсетеді. Деструктивті санаттарда агрессия, өшпенділік, теріс бағалау, үстемдік көрсету, зорлық әрекетін ақтау сияқты мағыналық компоненттер басым [49, 54].

Прагматикалық талдауда кластардың мынадай коммуникативтік ерекшеліктері қарастырылды:

• буллинг мәтіндерінде адресатты қорлау немесе оның беделіне нұқсан келтіру мақсаты байқалады;

• зорлық-зомбылық мазмұны оқырманда үрей, қысым, қорқыныш сезімін тудыруы мүмкін;

• нацизм мәтіндері идеологиялық ықпал етуге және оқырманды белгілі бір ұстанымды қабылдауға итермелеуге бейім;

• бейтарап мәтіндерде деструктивті мақсат айқындалмайды; эмоциялық реңк болуы мүмкін [27, 50, 53].

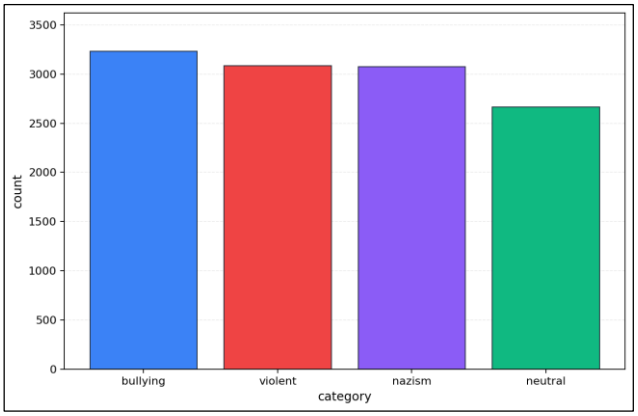
4. Корпустың семантикалық модель құруға жарамдылығы

Корпус қазақ тіліндегі деструктивті веб-контенттің лингвистикалық және мазмұндық ерекшеліктерін зерттеу үшін пайдаланылды. Оның көлемі, кластар бойынша үлестірімі және аннотация сапасы модельді оқыту мен бағалауды ұйымдастыру кезінде ескерілді [27, 53].

Корпус буллинг, нацизм және зорлық-зомбылық экстремизміне тән тілдік және мазмұндық белгілерді ажыратын модельдерді оқыту үшін қалыптастырылды [47-48, 54].

Корпустағы мәтіндердің құрылымдық ерекшеліктерін талдау үшін арнайы морфологиялық талдау құралы қолданылды. Сөйлемдер токендерге бөлініп, әр сөздің түбірі, морфологиялық категориясы және нормаланған нұсқасы анықталды. Осы мәліметтер кілттік сөздердің кездесу жиілігін есептеуге және кластарға тән ерекшеліктерді салыстыруға пайдаланылды. Сөздердің лемма деңгейінде түрлендірілген нұсқалары «short_word» өрісінде сақталып, кейінгі жиілік есептеулерінде қолданылды.

Әр класқа жататын мәтіндер алдын ала нормаланғаннан кейін, олардың құрамындағы сөздердің кездесу жиілігі Counter құралы арқылы есептелді. Жиілігі жоғары сөздер әр кластың лексикалық құрамын сипаттау үшін алынды. Мәтіндердің төрт класс бойынша үлестірімі сурет 2.2-те көрсетілген.



Сурет 2.2 - Корпус көлемінің төрт негізгі классқа бөліну диаграммасы

Әр класқа тән жиі кездесетін сөздер олардың мәтіндегі қолданылу контексімен бірге қарастырылды.

1. Bullying класына тән кілттік сөздер

Bullying класында «жетім», «ақымақ», «мүгедек», «мисыз» сөздері жиі кездеседі. Алайда бұл сөздердің өзі мәтінді бірден буллинг класына жатқызуға жеткіліксіз. Олардың мағынасы нақты контекстке байланысты анықталады: бір жағдайда сөз жеке тұлғаны кемсіту немесе қорлау үшін қолданылса, басқа жағдайда бейтарап не сипаттамалық мәнде берілуі мүмкін.

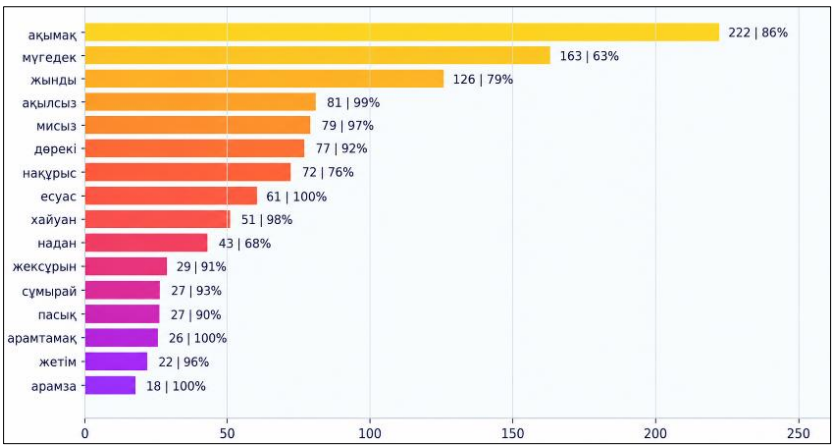
Bullying класына тән негізгі тілдік-мазмұндық белгілер:

– жеке адамға бағытталған кемсіту және қорлау, мысалы, «ақымақ» сөзінің тіл тигізу мәнінде қолданылуы;

– адамның қабілеті, дене ерекшелігі немесе әлеуметтік мәртебесі бойынша кемсіту;

– эмоциялық қысым мен қорқытуға қатысты тілдік құрылымдар.

Bullying_text класында таңдалған сөздердің негізгі тұлға бойынша жиіліктері сурет 2.3-те көрсетілген.



Сурет 2.3 - Bullying_text классындағы ең жиі кездесетін сөздер

2. Violent класына тән кілттік сөздер

Violent класында «шабуыл», «бомба», «тәлибан», «әскери», «соғыс» сөздері жиі кездеседі. Бұл лексика көбіне қарулы қақтығыс, күш қолдану және қару-жарақ тақырыптарымен байланысты. Алайда аталған сөздер жаңалықтық

немесе талдамалық мәтіндерде де қолданылуы мүмкін, сондықтан олардың деструктивті мағынасы нақты контекстке қарай анықталады.

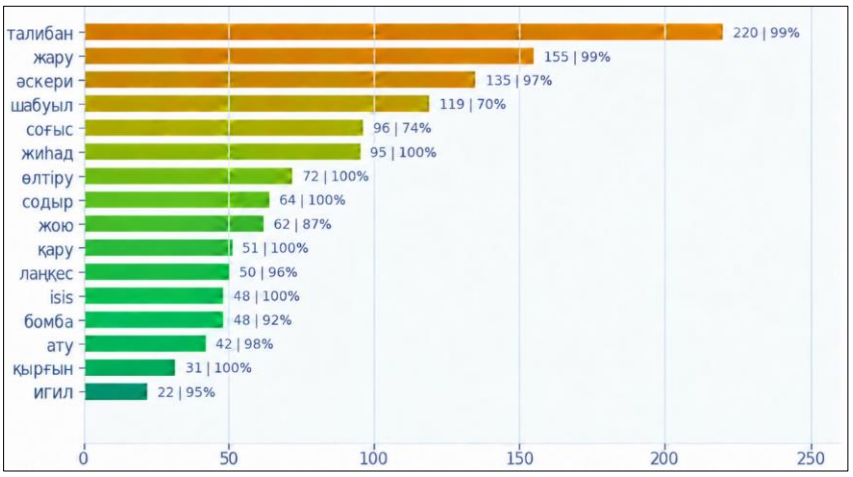
Violent класына тән негізгі тілдік-мазмұндық белгілер:

– қарулы қақтығыс пен күш қолдануға қатысты лексика («әскери», «шабуыл», «атыс»);

– ұйым атауларының деструктивті веб-контент контексінде қолданылуы, мысалы, «талибан»;

– зорлық-зомбылыққа шақыруды білдіретін сөздер мен тілдік құрылымдар.

Violent_text класында таңдалған сөздердің негізгі тұлға бойынша жиіліктері сурет 2.4-те көрсетілген.



Сурет 2.4 - Violent_text классындағы ең жиі кездесетін сөздер

3. Nazism класына тән кілттік сөздер

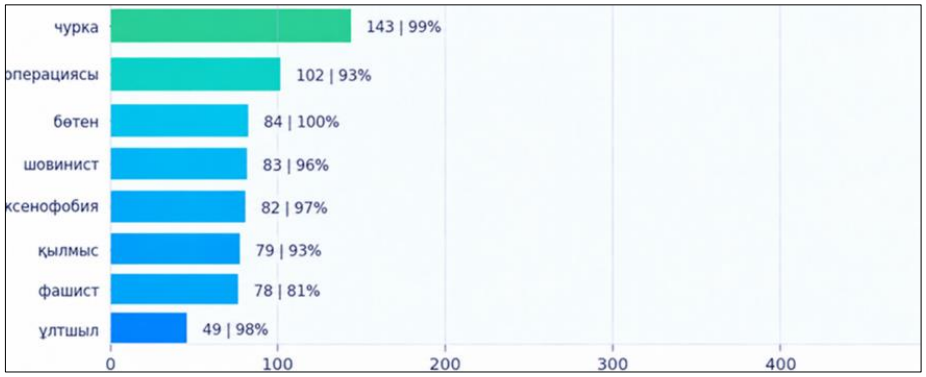
Бұл сөздердің бір бөлігі этникалық кемсіту немесе ксенофобиялық контентте қолданылуы мүмкін. Бірақ «бөтен», «операция» сөздерінің өзі нацизм белгісі болып саналмайды; олардың мағынасы мәтін контексінде анықталады.

Nazism класын сипаттайтын негізгі терминдер:

– шетелдіктер мен мигранттарды кемсітетін қолданыстар;

– этникалық үстемдік немесе тазалық идеясын білдіретін лексика.

Сурет 2.5-те Nazism_text класындағы таңдалған сөздердің негізгі тұлға бойынша жиіліктері бейнеленген.



Сурет 2.5 - Nazism_text классындағы ең жиі кездесетін сөздер

4. Neutral класына тән кілттік сөздер

Жиі кездесетін сөздер: «бір», «адам», «өмір», «жақсы».

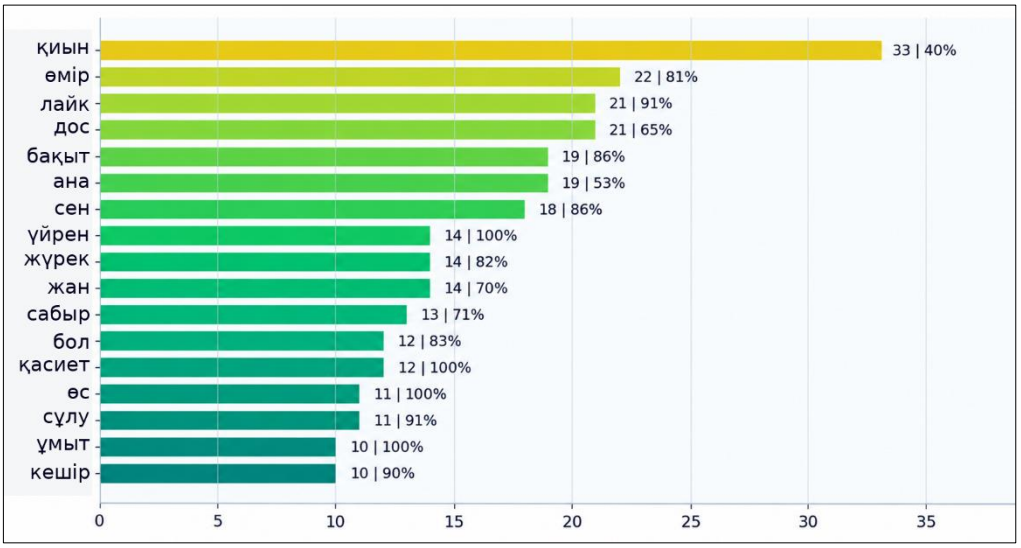
Бұл сөздер жалпы қолданыстағы лексикаға жатады. Олардың кездесуі мәтіннің бейтарап екенін өздігінен анықтамайды.

Neutral класын көрсететін ерекшеліктер:

− жалпы қолданыстағы сөздердің кездесуі;

− деструктивті мағынаның болмауы.

Сурет 2.6-да Neutral_text класындағы таңдалған сөздердің негізгі тұлға бойынша жиіліктері бейнеленген.



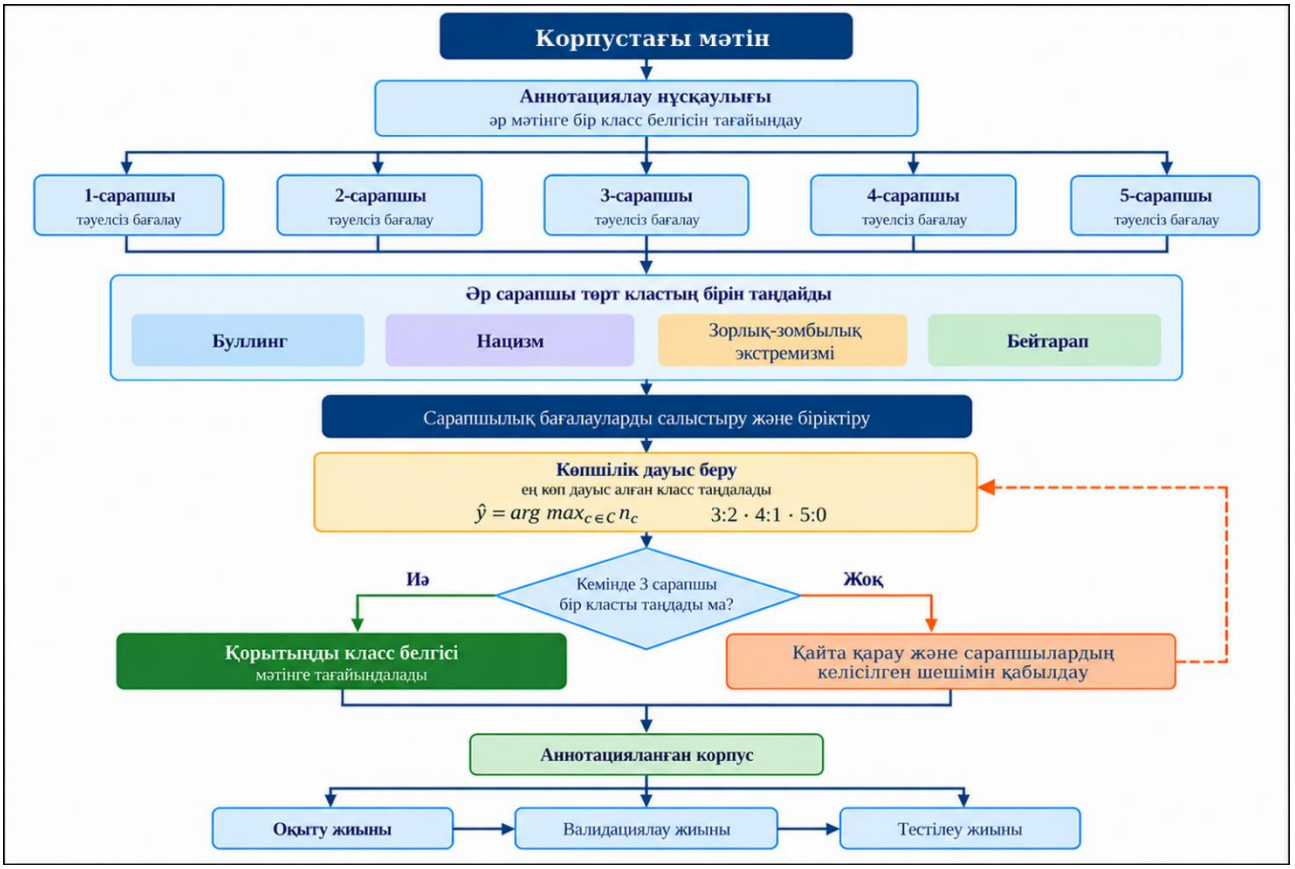
Сурет 2.6 - Neutral_text классындағы ең жиі кездесетін сөздер

Аннотациялау әдістемесі

Корпусты модельдерді оқытуға, валидациялауға және тестілеуге дайындау үшін әр мәтінге оның мазмұнын сипаттайтын класс белгісі тағайындалды [55]. Аннотациялау зерттеуде қабылданған төрт класс бойынша жүргізілді. Аннотациялау бірнеше кезеңмен жүргізілді. Алдымен сарапшылар мәтіндерді бір-бірінен тәуелсіз бағалады. Сарапшылардың бағалаулары салыстырылып, әр мәтінге қорытынды класс белгісі көпшілік дауыс беру тәсілі бойынша тағайындалды [34, 54]. Бұл тәсілде әр сарапшының бағасы жеке дауыс ретінде есептеледі, ал көп дауыс жинаған класс қорытынды белгі ретінде қабылданады. Мәтіндерді бес сарапшы бағалағандықтан, белгілі бір класты таңдау үшін кемінде үш сарапшының пікірі сәйкес келуі тиіс. Осыған сәйкес, дауыстар 3:2, 4:1 немесе 5:0 қатынасында бөлінген жағдайда көпшілік қолдаған класс мәтіннің қорытынды белгісі ретінде алынды [56-57]. Мұндай тәсіл бір сарапшының байқаусызда жіберген қатесінің нәтижеге әсерін азайтады, ал белгілеу ұжымдық шешімге негізделетіндіктен оның сапасы да артады.

Аннотациялау мен қорытынды класты анықтаудың жалпы барысы сурет 2.7-де берілген. Алдымен әр мәтінді бес сарапшы бір-бірінен тәуелсіз оқып, төрт кластың біріне жатқызады. Үш немесе одан көп сарапшының таңдауы сәйкес келсе, мәтінге көпшілік дауыс бойынша класс белгісі беріледі. Ал пікірлер бір

жерден шықпаған мәтіндер қайта қаралып, сарапшылар ортақ келісімге келгеннен кейін ғана белгіленеді. Осы кезеңдерден өткен корпус соңында оқыту, валидациялау және тестілеу жиындарына бөлінеді.



Сурет 2.7 - Мәтіндерді аннотациялау және қорытынды класс белгісін анықтау үдерісі

Формальды түрде қорытынды белгі (2.1) формуламен анықталды:

$$y_i = arg^{max}_{c \epsilon C} \sum_{j=1}^{5} I(y_i^{(j)} = c) \quad (2.1)$$

Мұндағы,

y_i – i-мәтіннің соңғы класс белгісі,

C – кластар жиыны,

$y_i^{(j)}$ – j -сарапшының i-мәтінге қойған белгісі,

$I(\cdot)$ – шарт орындалса 1, орындалмаса 0 мәнін қабылдайтын индикатор функциясы.

Бұл зерттеуде кластар жиыны (2.2) формулада берілді:

$$C = \{\text{бейтарап}, \text{буллинг}, \text{нацизм}, \text{зорлық} - \text{зомбылық экстремизмі}\} \quad (2.2)$$

Егер сарапшылардың белгілері бірмәнді шешім шығармайтын болса, яғни екі не одан да көп класс теңдей дауыс алған жағдайда, осындай мәтіндер даулы үлгі ретінде белгіленіп, қосымша талқылауға жолданады. Талқылау барысында

мәтіннің жалпы контексті, деструктивті мағынаның ашық немесе жасырын берілуі, мәтіндегі нысана, сөйлеу бағыты және аннотация нұсқаулығындағы класс анықтамалары ескерілді [58]. Түпкілікті шешім сарапшылардың ортақ ұстанымы мен қорытындысы негізінде бекітілді.

Аннотациялау кезінде жеке сөздермен қатар мәтіннің жалпы мағынасы мен контексі ескерілді [59].

Бес сарапшының аннотациялары арасындағы жалпы келісімді бағалау үшін кездейсоқ сәйкестікті ескеретін Fleiss' κ коэффициенті қолданылады [60]. Көрсеткіш әр мәтінге бес сарапшы берген бастапқы тәуелсіз белгілер бойынша, көпшілік дауыс беру мен келісу талқылауына дейін есептелуі тиіс. Fleiss' κ (2.3) формула бойынша анықталады:

$$k = \frac{\bar{p} - p_e}{1 - p_e} \tag{2.3}$$

Мұндағы,

$\bar{p}$ – барлық мәтіндер бойынша байқалған келісімнің орташа үлесі;

p_e – кластардың жалпы үлестері негізінде есептелетін кездейсоқ келісім ықтималдығы.

N – бес сарапшы толық бағалаған мәтіндер саны, n= 5 – сарапшылар саны, k = 4 – кластар саны, n_{ij} – і-мәтінді j-класқа жатқызған сарапшылар саны. Әр мәтін үшін $\sum j = 1^4$ n_{ij} = 5 шарты орындалады.

Әр мәтін бойынша келісім үлесі, кластың жалпы үлесі және жинақталған келісім шамалары (2.4) формулалар жүйесі бойынша есептеледі:

$$P_i = \frac{\sum_{j=1}^{4} n_{ij}(n_{ij}-1)}{20}$$
$$p_j = \frac{\sum_{i=1}^{N} n_{ij}}{5N} \tag{2.4}$$
$$\bar{P} = \frac{\sum_{i=1}^{N} P_i}{N}; \quad P_e = \sum_{j=1}^{4} {p_j}^2$$

Fleiss' κ бес сарапшының төрт класс бойынша келісімін біртұтас көрсеткішпен сипаттайды [61-62]. Деректерді жинау және өңдеу кезінде пайдаланушылардың құпиялығына қатысты талаптар сақталды. Корпусқа пайдаланушыны тікелей сәйкестендіруге болатын жеке мәліметтер енгізілген жоқ. Зерттеу барысында тек мәтіндердің мазмұны талданды.

2.2 Қазақ тіліндегі деструктивті веб-контенттің морфологиялық және онтологиялық білім моделін әзірлеу

Бұл бөлімде 1-тарауда сипатталған қазақ тілінің агглютинативті табиғатын негізге ала отырып, мәтіннен морфологиялық белгілерді шығаратын модель ұсынылады. Модельдің мақсаты – сөздердің грамматикалық сипаттамаларын классификация кезінде пайдаланылатын қосымша белгілерге айналдыру.

Зерттеуде морфологиялық білім моделі сөз табы, шақ, жақ және рай белгілері негізінде жасалды [63-66]. Аталған белгілер әрекеттің қай уақытта өтетінін, оны кім орындайтынын және сөйлеушінің сол әрекетке көзқарасын білдіреді. Оларды контекстік белгілермен ұштастыру мәтіндегі қорқыту, үндеу немесе әрекетке итермелеу мағынасын дәлірек талдауға көмектеседі. Морфологиясы күрделі әрі ресурсы аз тілдер үшін қосымша морфологиялық ақпаратты қолданудың тиімділігі [67] жұмыста зерттелген.

Алғашқы кезеңде септік, зат есімнің түрі, сөз табы, шақ, жақ, рай және лемма белгілерінен тұратын 339 өлшемді конфигурация сыналды. Мұның 300 өлшемін лемманың векторлық көрінісі, ал қалған 39 өлшемін грамматикалық белгілер құрады [54].

Морфологиялық белгілердің бастапқы конфигурациясы кесте 2.3-те келтірілген [54]. Морфологиялық модуль KazBERT берген контекстік ақпаратты грамматикалық белгілермен толықтыру үшін енгізілгендіктен, соңғы нұсқада лемманың векторлық көрінісі қолданылмады. Бұл лексикалық ақпараттың екі рет қайталанып кетпеуі үшін жасалды.

Кесте 2.3 - Жеті санат негізіндегі талдау нәтижесі

№	Морфологиялық белгі	Вектор
1	Септік	7
2	Зат есімнің түрі	2
3	Сөз табы	17
4	Шақ	4
5	Жақ	4
6	Рай	5
7	Лемма	300
	Жалпы	339

Кесте 2.4-те берілген конфигурацияға сөз табы, шақ, жақ және рай белгілері енеді [54]. Сөз табы арқылы сөздің қай лексика-грамматикалық топқа жататыны анықталса, шақ әрекеттің қашан болатынын көрсетеді. Жақ категориясы әрекеттің сөйлеушіге, тыңдаушыға не үшінші тұлғаға қатысты екенін айқындайды. Ал рай сөйлеушінің әрекетке көзқарасын, яғни оның хабарлау, бұйыру, қалау немесе шарт мәнінде айтылғанын білдіреді.

Кесте 2.4 - Төрт санат негізіндегі талдау нәтижесі

№	Морфологиялық белгі	Вектор
1	Сөз табы	17
2	Шақ	4
3	Жақ	4
4	Рай	5
	Жалпы	30

Морфологиялық белгілер кеңістігі 339 өлшемнен 30 өлшемге дейін қысқартылды. Бұл конфигурация ең жоғары классификациялау көрсеткішін берген нұсқа ретінде емес, KazBERT қалыптастыратын контекстік көріністерді грамматикалық ақпаратпен толықтыратын ықшам белгілер жиыны ретінде

таңдалды. Кейінгі кезеңде 30 өлшемді морфологиялық вектор KazBERT-тен алынған контекстік көріністпен біріктіріліп, KazBERT_MFA моделінің құрамында қолданылды.

Сөз табы белгісі 17 өлшеммен берілді. Олардың қатарында зат есім, сын есім, етістік, үстеу, сан есім, жалқы есім, одағай және шылау санаттары бар. Бұл белгілер әлеуметтік желі мәтіндерінің лексика-грамматикалық құрылымын сипаттайды. Зат есімдер мәтінде аталған нысандарды, етістіктер әрекеттерді, ал сын есімдер бағалауыш сипаттарды білдіруі мүмкін. Дегенмен олардың деструктивті мазмұнды анықтаудағы қызметі сөздің грамматикалық категориясымен ғана шектелмейді, өйткені нақты мағына мәтіндегі қолданыс контексіне байланысты өзгереді. Осы себепті морфологиялық белгілер KazBERT арқылы алынған контекстік көріністерге қосымша грамматикалық ақпарат береді. Морфологиялық модульдің бағдарламалық іске асырылуы 3.2-бөлімде сипатталған.

Шақ белгісі төрт санатпен берілді: өткен шақ, осы шақ, келер шақ және «шақ анықталмаған». Соңғы санат жеке грамматикалық шақты білдірмейді, ол шақ белгісі жоқ немесе морфологиялық талдау кезінде шақ мәні анықталмаған токендерді белгілеу үшін қолданылды. Осылайша шақ категориясы қолданылмайтын токендер де модельдің кіріс белгілерінде жеке көрсетілді.

Жақ белгісі үшін де төрт санат пайдаланылды: бірінші жақ, екінші жақ, үшінші жақ және белгісіз жақ. Жақ категориясы іс-әрекеттің сөйлеушіге, тыңдаушыға немесе үшінші тұлғаға қатынасын сипаттайды. Әлеуметтік желілердегі мәтіндерде бұл ақпарат нақты тұлғаға немесе басқа қатысушыға бағытталған тілдік құрылымдарды сипаттау кезінде қосымша грамматикалық белгі ретінде қолданылады. Сондықтан жақ белгісі модельге мәтіндегі субъектілік қатынас жөнінде қосымша ақпарат береді.

Рай белгісі бес өлшеммен берілді. Қазақ тіліндегі ашық рай, бұйрық рай, шартты рай және қалау рай санаттарымен қатар белгісіз рай категориясы қолданылды. Бұл категория жеке рай түрі болып саналмайды, ол рай белгісі жоқ немесе морфологиялық талдау кезінде райы анықталмаған токендерді белгілеуге арналған. Мұндай санаттың енгізілуі рай категориясының негізінен етістікке тән болуымен және мәтіндегі барлық токендерде рай белгісінің кездесе бермеуімен байланысты.

Қорытынды морфологиялық вектор төрт белгілер тобынан қалыптасты: сөз табы, шақ, жақ және рай. Сөз табына 17 өлшем, шаққа 4 өлшем, жаққа 4 өлшем және райға 5 өлшем бөлінді. Осы белгілер біріктірілген кезде морфологиялық вектордың жалпы өлшемі 30-ды құрады [54].

Морфологиялық модельді сипаттау үшін кіріс веб-мәтін X токендер тізбегі ретінде беріледі (2.5) формула бойынша:

$$X = \{w_i, w_2, \dots, w_n\} \tag{2.5}$$

Мұндағы,
X – кіріс мәтін,
w_i – мәтіндегі жеке токен,

n – мәтіндегі токендер саны.

Әрбір токеннің морфологиялық сипаттамасы сөз табы, шақ, жақ және рай белгілерінен құралады ((2.6) формула):

$$m_i = \{pos_i, tense_i, person_i, mood_i\} \quad (2.6)$$

Мұндағы,
pos_i– сөз табы,
$tense_i$ – шақ белгісі,
$person_i$ – жақ белгісі,
$mood_i$– рай белгісі.

Осы белгілер негізінде мәтіннің толық морфологиялық сипаттамасы (2.7) формулада қалыптасады:

$$M(X) = \{m_1, m_2, \dots, m_n\} \quad (2.7)$$

Мұндағы,
$M(X)$ – мәтіннің морфологиялық білім моделі.

Мәтін деңгейіндегі жинақталған морфологиялық вектор (2.8) формулада беріледі:

$$V_m = Pool(M(X)) \quad (2.8)$$

Мұндағы,
V_m– мәтіннің морфологиялық векторы,
$Pool$ – токендер бойынша алынған морфологиялық белгілерді біріктіру функциясы.

Морфологиялық белгілер мәтіннің грамматикалық құрылымын сипаттайды. Деструктивті веб-контентке қатысты терминдердің ұғымдармен семантикалық байланысын көрсету үшін онтологиялық білім моделі қолданылды. Онтологияға негізделген семантикалық белгілер граф құрылымы арқылы қалыптастырылды: терминдер, ішкі ұғымдар және жоғары деңгейдегі санаттар граф түйіндері ретінде беріліп, олардың арасындағы семантикалық байланыстар қырлармен көрсетілді.

Зерттеуде ұғымдар мен олардың иерархиялық байланыстары онтологиялық білім моделі негізінде жүйеленді. Модель үш өлшемді қамтиды: түрі, ашықтығы және нысанасы. «Түрі» деструктивті веб-контенттің мазмұндық бағытын сипаттайды, «ашықтығы» деструктивті мағынаның ашық немесе жасырын берілуін көрсетеді. Ал «нысанасы» мазмұнның жеке адамға немесе белгілі бір топқа бағытталуын білдіреді [68-71]. Осы өлшемдер арқылы терминдер жеке лексикалық бірлік ретінде ғана емес, олардың мазмұндық және семантикалық байланыстары тұрғысынан да сипатталды.

Онтологиялық графты өңдеуге графтық нейрондық желі (GNN) қолданылды. GNN түйіндердің векторлық көріністерін көршілес түйіндерден алынған ақпарат негізінде жаңартады [72]. [73] жұмыста GNN тәсілдерінің өшпенділік тілі (hate speech) мен зиянды контентті анықтауда контекстік,

семантикалық және байланыстық ақпаратты бірге өңдеу мүмкіндіктері қарастырылған. [54] жұмыста онтологиялық білімді Transformer негізіндегі модельдермен біріктірудің классификация сапасы мен түсіндірмелілікке ықпалы талданған.

Осы графтық құрылымды есептеу моделінде қолдану үшін онтологиялық модель формальды түрде (2.9) формулада беріледі:

$$G_o = \{V_o, E_o, R_o\} \tag{2.9}$$

Мұндағы,
G_o – онтологиялық граф,
V_o – онтологиялық ұғымдар жиыны,
E_o – ұғымдар арасындағы байланыстар жиыны,
R_o – байланыс түрлерінің жиыны.
Онтологиядағы ұғымдар жиыны (2.10) формулада беріледі:

$$V_o = \{v_0, v_1, \dots, v_{47}\} \tag{2.10}$$

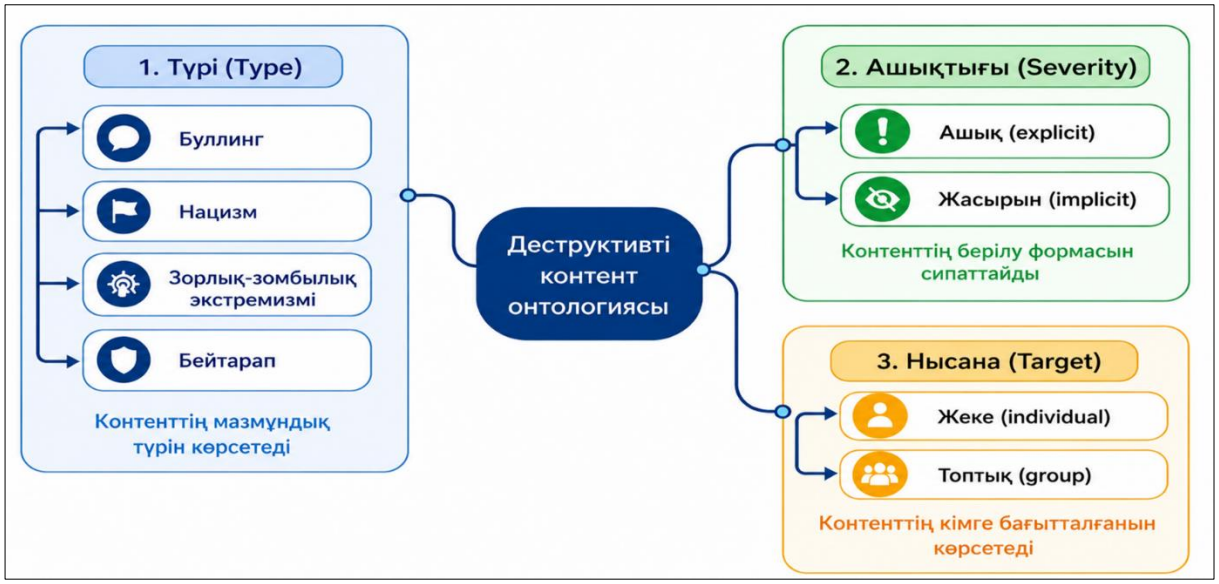
Мұндағы,
v_i – онтологиядағы жеке түйін.
Толық графта 48 түйін қарастырылғанымен, мәтіннің онтологиялық белгі векторы v_1-ден v_{47}-ге дейінгі 47 түйін бойынша қалыптастырылады, (2.11) формула бойынша көрсетілген:

$$x_{onto}(X) \in R^{47} \tag{2.11}$$

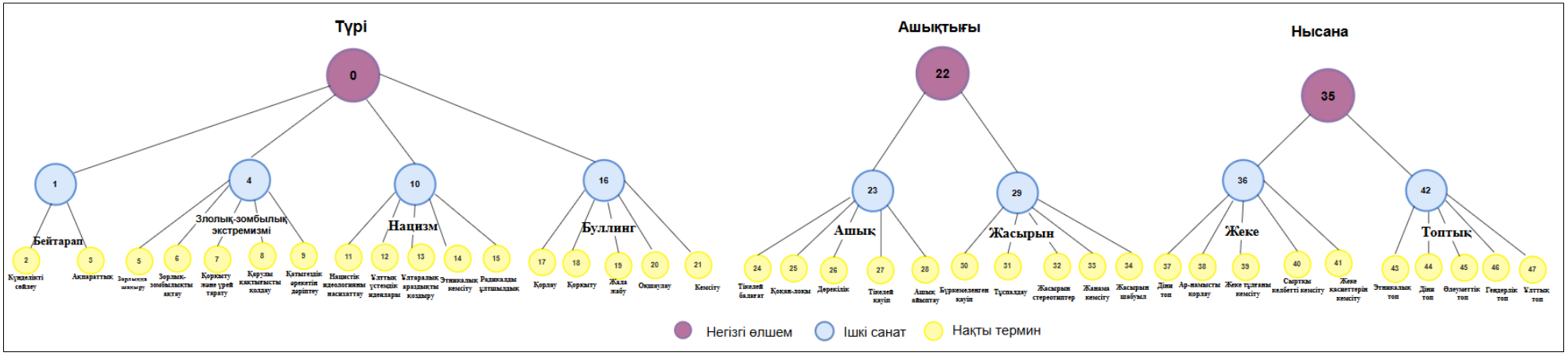
Ұғымдар арасындағы байланыс төмендегі (2.12) формулада беріледі:

$$e_{ij} = \{v_i, r_{ij}, v_j\} \tag{2.12}$$

Мұндағы,
v_i – бастапқы түйін,
r_{ij}– қатынас түрі,
v_j– соңғы түйін.
Бұл жұмыста негізгі байланыс түрі ретінде “ата-ұғым – ішкі ұғым” иерархиялық қатынасы қарастырылады. Онтологиялық модельдің құрылымы мен ұғымдар арасындағы иерархиялық байланыстар 2.8 және 2.9-суреттерде берілген. Сурет 2.8-де модельдің негізгі өлшемдері мен олардың ішкі санаттарын сипаттайды, ал сурет 2.9-да осы ұғымдардың өзара байланысын бейнелейтін онтологиялық графтың құрылымы берілген [68].



Сурет 2.8 - Онтологиялық модельдің негізгі өлшемдері мен санаттары



Сурет 2.9 - Қазақ тіліндегі деструктивті веб-контенттің онтологиялық граф құрылым

Графтағы байланыстар көршілестік матрицасы арқылы (2.13) формулада сипатталады:

$$A_{ij}\begin{cases}1, \text{егер } v_i \text{ және } v_j \text{ тікелей байланысты болса,}\\ 0, \text{егер тікелей байланыс болмаса.}\end{cases} \quad (2.13)$$

Мұндағы,

$A \in R^{48\times48}$, себебі толық онтологиялық графта v_0-ден v_{47}-ге дейінгі түйіндер бар [72, 74].

Мәтінді онтологиялық түйіндермен байланыстыру үшін лексикалық құралдар пайдаланылады. Әрбір термин нақты бір онтологиялық ұғымға (2.14) формула арқылы сәйкестендірілген:

$$\mu: L \to V_o^* \quad (2.14)$$

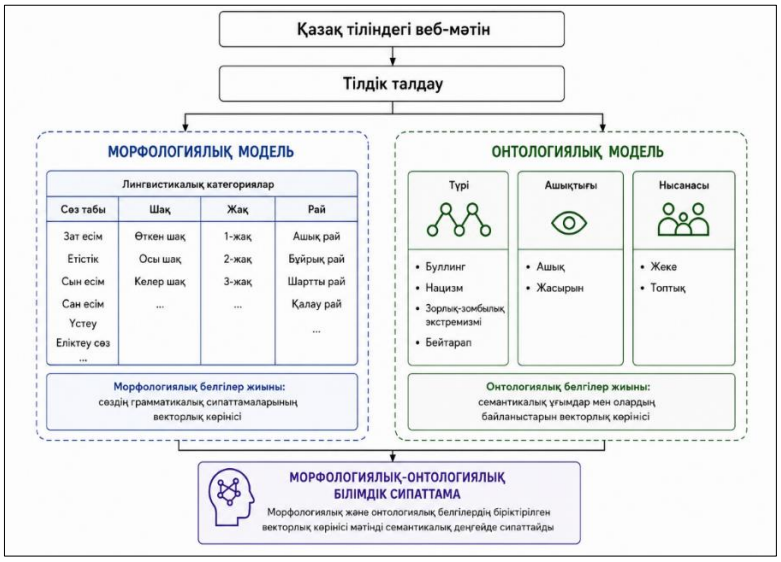
Мұндағы,

L – онтологиялық лексикондағы терминдер жиыны,

$V_o^* = \{v_0, v_1, \dots, v_{47}\}$, μ – терминді онтологиялық түйінмен байланыстыратын функция.

Онтологиялық білім моделі қазақ тіліндегі деструктивті веб-контентті мағыналық деңгейде сипаттау үшін пайдаланылды. Бұл белгілер KazBERT қалыптастыратын контекстік көріністер мен морфологиялық модуль арқылы анықталған грамматикалық сипаттарды семантикалық ақпаратпен толықтырды [75].

Морфологиялық және онтологиялық модельдер мәтіннің әртүрлі қырларын сипаттайды. Біріншісі сөздердің грамматикалық ерекшеліктерін анықтаса, екіншісі ұғымдар арасындағы семантикалық қатынастарды көрсетеді [76-78]. Осы модельдер негізінде белгілерді қалыптастыру кезеңдері сурет 2.10-да көрсетілген.



Сурет 2.10 - Қазақ тіліндегі деструктивті веб-контенттің морфологиялық және онтологиялық білімдік сипаттамасын қалыптастыру схемасы

Морфологиялық-онтологиялық білімдік сипаттама мәтінді сипаттау кезінде грамматикалық және онтологияға негізделген семантикалық ақпараттың қатар пайдаланылуын білдіреді.

2.3 Контекстік, морфологиялық және семантикалық белгілердің интеграцияланған моделін әзірлеу

Бұл бөлімде 2.2-бөлімде сипатталған морфологиялық және онтологияға негізделген семантикалық белгілерді KazBERT арқылы алынған контекстік көріністермен біріктіру тәсілі қарастырылады. Модельде қолданылатын үш белгі тобының қызметі кесте 2.5-те берілген.

Кесте 2.5 - Ұсынылған модельде қолданылатын белгілердің сипаттамасы

Белгі қабаты	**Негізгі қызметі**	**Модельдегі көрінісі**
Контекстік белгі	Мәтіннің жалпы контекстік және семантикалық мағынасын қалыптастыру	KazBERT моделі арқылы алынған контекстік эмбеддингтер
Морфологиялық белгі	Қазақ тілінің грамматикалық ерекшеліктерін сипаттау	Сөз табы, шақ, жақ, рай сияқты морфологиялық белгілер
Семантикалық белгі	Деструктивті ұғымдар арасындағы семантикалық байланыстарды көрсету	Онтологиялық лексикон, семантикалық категориялар және граф түйіндері

Мәтіннің контекстік белгілерін алу үшін KazBERT моделі қолданылды. Модель қалыптастыратын көріністер сөздердің мәтін ішіндегі өзара байланысын және олардың нақты контекске тәуелді мағынасын сипаттайды [79-80]. Кіріс мәтіннің контекстік көрінісі (2.15) формула бойынша анықталады:

$$C(X) = KazBERT(X) \quad (2.15)$$

Мұндағы,
X – кіріс мәтін,
$C(X)$ – мәтіннің контекстік көрінісі.

Морфологиялық қабат KazBERT арқылы алынған контекстік көріністерді қазақ тілінің грамматикалық ақпаратымен толықтырады. Қосымша белгілер ретінде 2.2-бөлімде сипатталған сөз табы, шақ, жақ және рай категориялары қолданылады [81]. Олардың жиынтығы (2.16) формула бойынша анықталады:

$$F_m = \{pos, tense, person, mood\} \quad (2.16)$$

Мұндағы,
F_m – морфологиялық белгілер топтарының жиыны.

Онтологияға негізделген семантикалық белгілер мәтіндегі сөздер мен сөз тіркестерінің онтологиялық ұғымдармен байланысын сипаттайды [82]. Олар онтологиялық лексикон мен граф құрылымы негізінде қалыптастырылып, үш негізгі онтологиялық өлшем бойынша беріледі [83]. Семантикалық белгілердің осы құрылымы (2.17) формула бойынша анықталады:

$$O(X) = \{type, openness, target\} \quad (2.17)$$

Мұндағы,

$type$ – мәтіннің мазмұндық түрі,

$openness$ – деструктивтіліктің ашық немесе жасырын берілуі,

$target$ – мәтіннің нысанасы.

Онтологиялық граф 0-ден 47-ге дейін нөмірленген 48 түйіннен тұрады. Оны құруда сараптамалық жүйелеу мен автоматты терминдік талдауды біріктіретін гибридтік тәсіл қолданылды. Пәндік саланың негізгі санаттары мен ішкі ұғымдары деструктивті веб-контенттің түрі, ашықтығы және нысанасы бойынша жүйеленіп, корпус мәтіндерінен алынған терминдер осы құрылымдағы тиісті ұғымдармен сәйкестендірілді.

Мәтіндегі терминдерді онтологиялық түйіндермен байланыстыру үшін 5000 бірегей терминнен тұратын лексикон пайдаланылды. Оған енгізілетін терминдерді анықтау алдында корпус мәтіндерінен алынған токендер өңделіп, қалыпқа келтірілді. Терминдерді таңдауда олардың жиілігі, мәтіндерде кездесу тұрақтылығы және деструктивті санаттармен мағыналық байланысы ескерілді.

Онтологиялық лексиконды қалыптастыру кезінде тест жиыны пайдаланылған жоқ. Терминдерді іріктеу және оларды онтологиялық санаттармен сәйкестендіру тек оқыту жиыны мен сыртқы сараптамалық білім негізінде жүргізілді. Тест жиынындағы мәтіндер мен олардың класс белгілері лексиконның құрамын толықтыруға немесе өзгертуге, сондай-ақ терминдердің онтологиялық санаттарға сәйкестігін қайта қарауға қолданылған жоқ.

Әр терминге негізгі онтологиялық түйін сәйкестендірілді. Егер термин бірнеше түйінмен мағыналық байланысқа түссе, оның нақты санаты мәтіндегі контекстік және категориялық байланыстар бойынша анықталады. Мысалы, «кемсіту» сөзі жалпы мағынада қорлауды білдіруі мүмкін.

Жеке тұлғаға бағытталған психологиялық қысым контексінде ол буллинг класына қатысты түйінмен байланыстырылса, ұлттық немесе этникалық белгіге қатысты қолданылған жағдайда нацизм тармағындағы ұғыммен сәйкестендіріледі.

Лексиконға енгізілген термин мәтінде кездескен кезде оған сәйкес онтологиялық түйін белсендіріледі. Егер термин төменгі деңгейдегі нақты ұғымға сәйкес келсе, сол түйінмен иерархиялық байланыстағы жоғары деңгейдегі санат та есепке алынады. Мысалы, «қорлау» термині анықталғанда, «Қорлау» түйінімен бірге оның жоғары деңгейдегі «Буллинг» санатымен байланысы да белгіленеді. Осы тәсіл мәтінді жеке сөздер жиыны ретінде ғана емес, онтологиялық ұғымдар мен олардың арасындағы байланыстар арқылы сипаттауға мүмкіндік береді.

Мәтіндегі сөз онтологиялық лексиконда кездеспеген жағдайда оған сәйкес онтологиялық белгі белсендірілмейді. Бұл мәтінді әрі қарай өңдеуді тоқтатпайды: KazBERT оның контекстік көрінісін қалыптастырады, ал морфологиялық модуль грамматикалық сипаттарын анықтайды. Онтологиялық лексикон осы компоненттерге қосымша семантикалық ақпарат береді, ал лексиконда жоқ терминдер модельдің басқа бөліктері арқылы өңделеді.

Оқыту және тест деректері арасындағы ақпараттың өтіп кету қаупін (train/test leakage) азайту үшін тест жиыны модель параметрлерін баптауға, қосымша оқытуға (fine-tuning) немесе модель конфигурациясын таңдауға пайдаланылмайды [84]. Ол тек қорытынды бағалау кезеңінде қолданылады. Бұл жағдайда онтологиялық құрылым пәндік саланың мағыналық санаттары мен олардың өзара байланыстарын сипаттайтын білім моделі ретінде қызмет етеді.

Модельдің классификациялау сапасы жеке тест жиынында бағаланады және алынған көрсеткіштер базалық KazBERT пен жетілдірілген модельдердің нәтижелерімен салыстырылады.

Онтологияға негізделген семантикалық белгілерді қалыптастыру кезеңдері кесте 2.6-да берілген.

Кесте 2.6 - Семантикалық белгілерді қалыптастыру логикасы

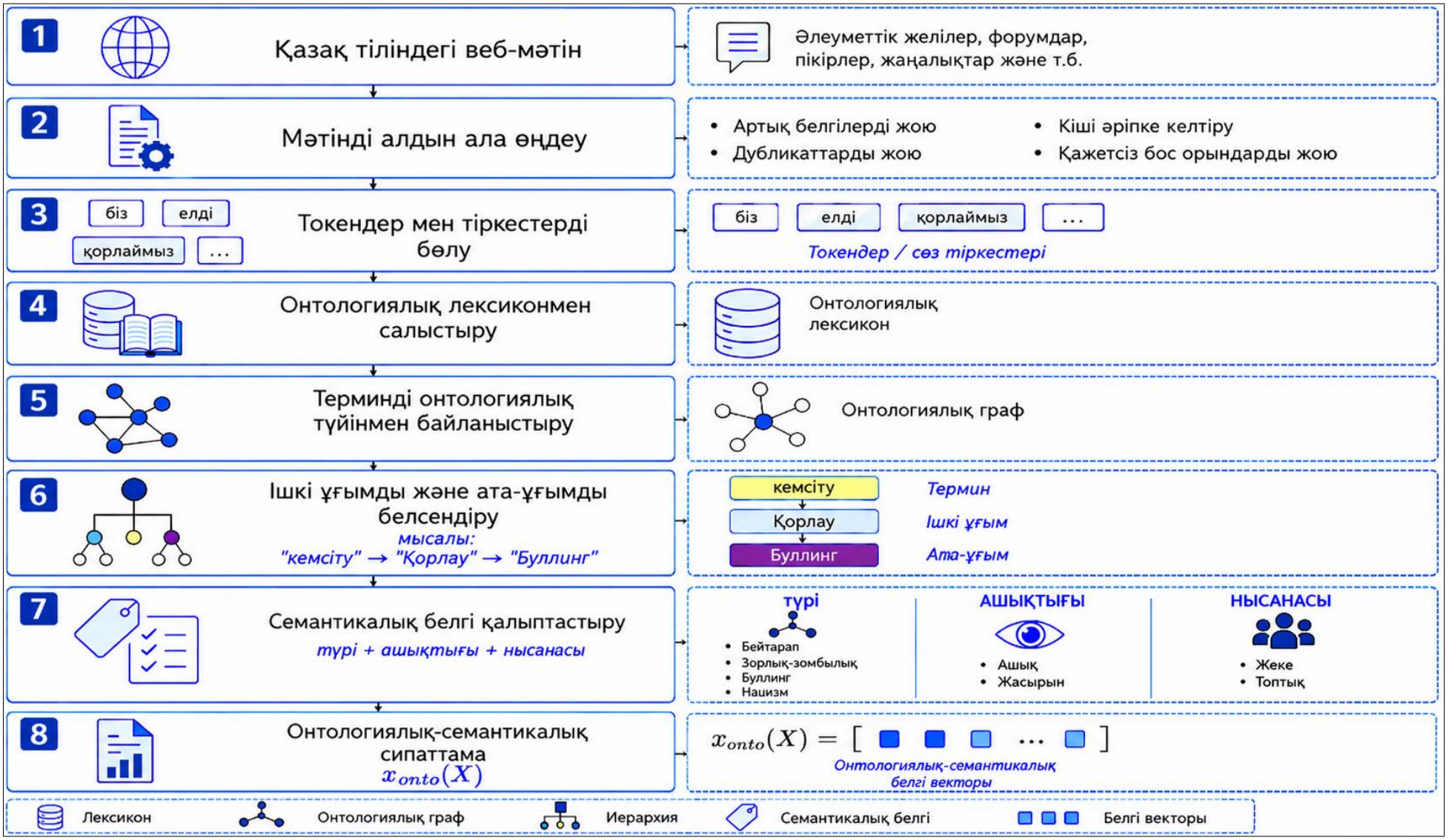
Кезең	**Сипаттамасы**
Терминді табу	Мәтіндегі маңызды сөздер мен тіркестер анықталады
Лексиконмен сәйкестендіру	Термин онтологиялық түйінмен байланысады
Категорияны анықтау	Терминнің қай деструктивті бағытқа жататыны белгіленеді
Семантикалық белгі алу	Мәтіннің түрі, ашықтығы және нысанасы сипатталады

Сурет 2.11-де мәтінді алдын ала өңдеуден бастап $x_{onto}(X)$ белгі векторын қалыптастыруға дейінгі үдеріс жинақталып көрсетілген. Суретте қазақ тіліндегі веб-мәтіннің онтологиялық-семантикалық сипаттамасын қалыптастыру кезеңдері көрсетілген. Алдын ала өңделген мәтін токендер мен сөз тіркестеріне бөлініп, онтологиялық лексиконмен салыстырылады. Анықталған терминдер графтың тиісті түйіндерімен байланыстырылып, ұғымдар иерархиясы негізінде мәтіннің түрі, ашықтығы және нысанасы бойынша белгілер қалыптастырылады. Нәтижесінде мәтіннің семантикалық ерекшеліктерін сипаттайтын онтологиялық-семантикалық белгі векторы алынады.

Осы байланысты қысқаша (2.18) формуладан көре аламыз:

$$I(X) = \{C(X), M(X), O(X)\} \tag{2.18}$$

Мұндағы,
$I(X)$ – мәтіннің интеграцияланған сипаттамасы,
$C(X)$ – контекстік белгі,
$M(X)$ – морфологиялық белгі,
$O(X)$ – семантикалық белгі.



Сурет 2.11 - Онтологиялық лексикон негізінде семантикалық белгілерді қалыптастыру үдерісі

(2.18) өрнегі бір мәтінді сипаттайтын контекстік, морфологиялық және онтологияға негізделген семантикалық белгілердің қатар қолданылуын көрсетеді. Бұл өрнек аталған белгілерді біріктіруге арналған нақты математикалық операцияны анықтамайды. Белгілерді біріктірудің есептеу моделі 3.3-бөлімде сипатталған.

2.4 2-тарау бойынша қорытынды

Бұл тарауда қазақ тіліндегі деструктивті веб-контент корпусын қалыптастыру мен аннотациялау әдістемесі, сондай-ақ модельде қолданылатын морфологиялық және онтологияға негізделген семантикалық белгілерді құру тәсілдері сипатталды. Морфологиялық сипаттама сөз табы, шақ, жақ және рай белгілерінен қалыптастырылды. Ал онтологиялық білім моделі деструктивті веб-контентке қатысты ұғымдарды түрі, ашықтығы және нысанасы бойынша жүйелеп, мәтіндегі терминдерді осы ұғымдармен байланыстыру үшін онтологиялық лексикон мен граф құрылымын пайдаланды.

Тарауда KazBERT қалыптастыратын контекстік көріністерді морфологиялық және онтологияға негізделген семантикалық белгілермен қатар қолданудың әдістемелік негізі де анықталды. Қалыптастырылған корпус пен белгілер кеңістігі келесі бөлімде қарастырылатын гибридтік классификациялау моделін әзірлеудің бастапқы негізін құрайды. Жекелеген белгілер топтарының классификациялау нәтижелеріне ықпалы кейінгі тәжірибелік бағалау барысында анықталады.

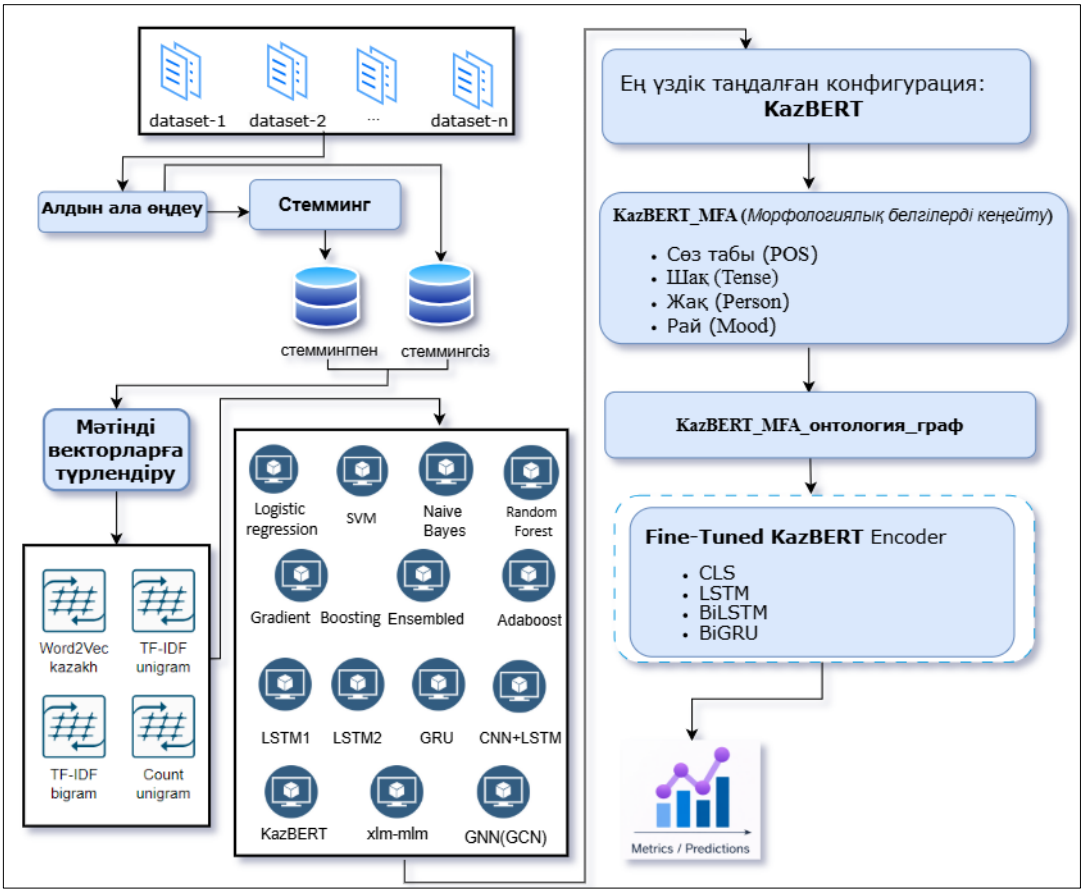
3 Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған тереңдетіп оқыту үлгілерін жасау

Бұл тарауда қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған гибридті әдіс сипатталады. Әдістің негізі ретінде KazBERT моделі таңдалып, оған морфологиялық мен онтологияға негізделген семантикалық белгілер және BiLSTM тізбектік өңдеу модулі қосылды.

Қосымша тілдік ақпарат классификация сапасын жақсартуы мүмкін деген болжам компоненттерді кезең-кезеңімен қосып салыстыру арқылы тексеріледі [85].

Зерттеудің жалпы барысы сурет 3.1-де көрсетілген. Екінші тарауда сипатталған корпус модельдерді құру мен салыстыруға арналған бастапқы деректер ретінде пайдаланылады. Бұл бөлімде модельдердің құрамы сипатталып, кейінгі бөлімдерде әртүрлі белгілерді біріктіру тәсілі қарастырылады.

Сурет 3.1-дің сол жақ бөлігінде деректерді дайындау, стемминг қолданылған және қолданылмаған нұсқаларды салыстыру, мәтінді сандық түрде ұсыну және бастапқы модельдерді бағалау кезеңдері берілген. TF–IDF, Word2Vec және сөздердің кездесу санын есептеуге негізделген тәсілдер эксперименттік конфигурацияға қарай әртүрлі модельдермен бірге қолданылады [86]. Сол себепті олардың барлығы KazBERT моделінің кірісіне бірдей түрде берілмейді.



Сурет 3.1 - Зерттеудің тұжырымдамалық схемасы

Сурет 3.1-дің оң жақ бөлігінде KazBERT негізіндегі модельге морфологиялық белгілер мен онтологиялық-графтық модульді қосу тәсілі берілген. Классификациялау кезінде [CLS] көрінісі тікелей қолданылатын нұсқа және тізбекті LSTM, BiLSTM немесе BiGRU арқылы қосымша өңдейтін нұсқалар салыстырылады.

3.1 Базалық және кеңейтілген классификациялау модельдерінің салыстырмалы талдауы

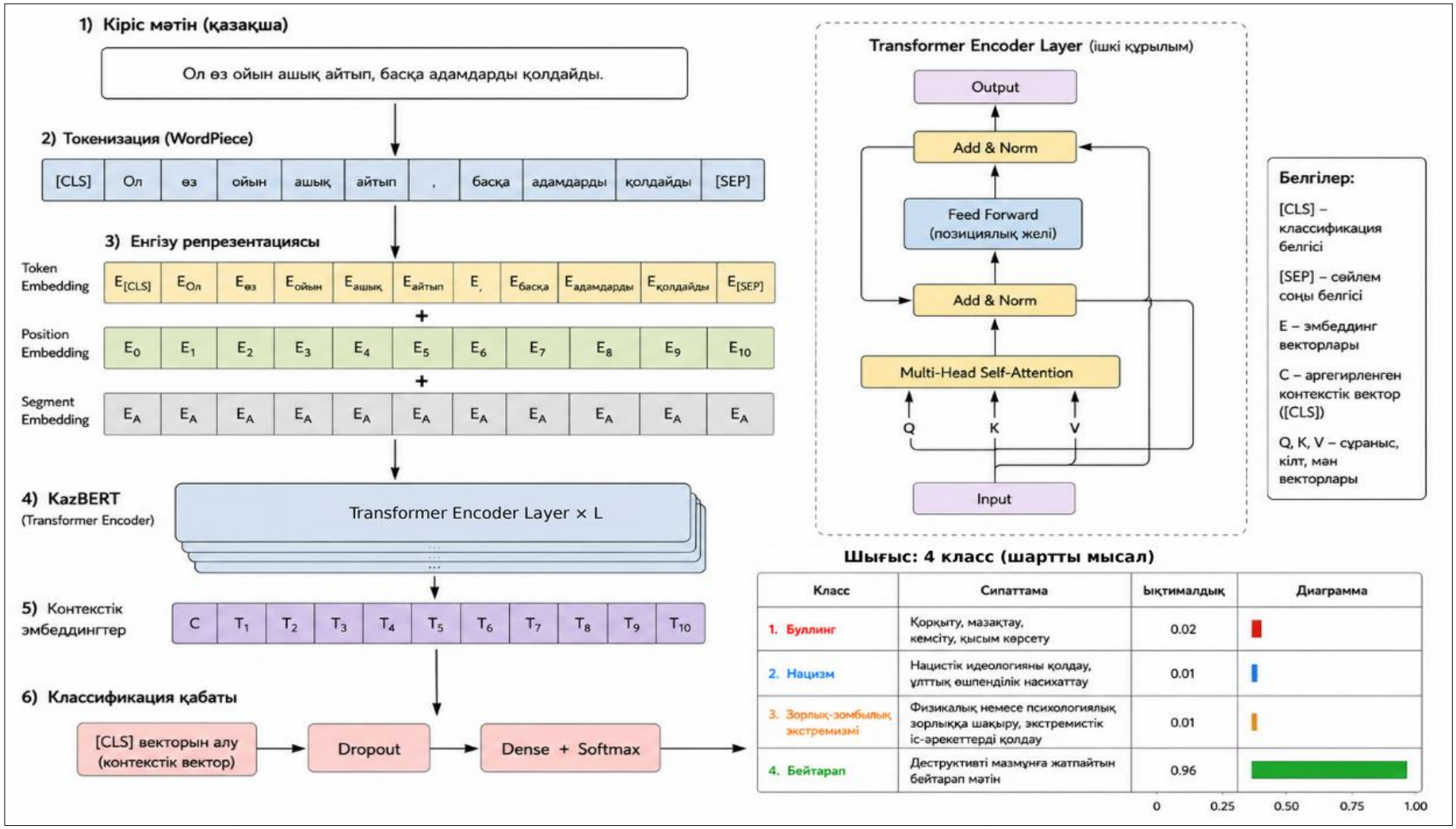
Салыстыруға бастапқы KazBERT моделі және оған қосымша компоненттер енгізілген үш нұсқа алынды. Компоненттерді модель құрамына біртіндеп қосу олардың классификация нәтижесіне қосатын үлесін жеке салыстыруға арналған.

KazBERT қазақ тіліндегі мәтіннің контекстік көріністерін қалыптастыратын базалық модель ретінде пайдаланылды. Ол BERT архитектурасының Transformer энкодеріне сүйенеді. Әр токеннің көрінісі оның алдындағы және кейінгі токендермен байланыста есептеледі. Сондықтан бір лексикалық бірліктің векторы қолданылған контекстіне қарай өзгереді. Бұл қасиет деструктивті веб-контентті классификациялауда ескеріледі: бір сөз қорлау мақсатында да, өзгенің сөзін дәйексөзге алу немесе оны теріске шығару үшін де қолданылуы мүмкін.

BERT-тің жалпы архитектурасы бірінші тарауда қарастырылғандықтан, мұнда оның модельге беретін шығысы нақтыланады. Энкодер өңдеген мәтін $H \in \mathbb{R}^{n \times d}$ түріндегі контекстік көріністер матрицасымен сипатталады. Мұндағы n – токендер саны, d – қолданылған модельдің жасырын өлшемі. Матрицаның әр жолы тиісті токеннің контекстік векторын білдіреді. Токендік көріністерден мәтінге ортақ вектор алу немесе оларды тізбек ретінде әрі қарай өңдеу таңдалған классификациялау нұсқасына байланысты [87-90].

Сурет 3.2-де [CLS] көрінісі арқылы классификациялау жолы берілген. Кіріс мәтін токенделіп, оған [CLS] және [SEP] арнайы токендері енгізіледі. Токендік, позициялық және сегменттік эмбеддингтер қосылып, Transformer энкодеріне беріледі.

Энкодердің соңғы қабатындағы [CLS] көрінісі dropout және толық байланысқан қабат арқылы өңделеді. Softmax функциясы төрт класс бойынша ықтималдықтарды есептейді. L энкодер қабаттарының санын білдіреді. Оның мәні қолданылған модельдің конфигурациясымен анықталады. Токендер көрнекілік үшін шартты түрде бөлінген, нақты бөліну токенизатор сөздігіне байланысты болады. Сөйлем мен төрт ықтималдық шығыс құрылымын түсіндіретін мысал ретінде берілген және модельдің сапасын бағалау нәтижесі болып саналмайды. Классификация сапасы белгіленген деректерде бөлек өлшенеді.



Сурет 3.2 - KazBERT негізінде қазақ тіліндегі веб-мәтінді классификациялау үдерісі

KazBERT_MFA нұсқасында контекстік ақпарат морфологиялық белгілермен толықтырылды. MFA модулі сөз табы, шақ, жақ және рай категорияларын қамтиды. Әр мәтін үшін 30 өлшемді морфологиялық вектор қолданылады. Бұл вектор грамматикалық белгілерді модельдің кірісіне береді, ал оның құрамы мен қалыптастыру тәсілі келесі бөлімде сипатталады.

Морфологиялық ақпарат сөздің грамматикалық тұлғасын мәтін контексімен бірге қарастыруға қолданылады. Мысалы, бұйрық райдағы етістік пен екінші жақ тұлғасы үндеудің кімге бағытталғанын және сөйлемнің коммуникативтік сипатын нақтылауы мүмкін. Бірақ бұйрық рай бейтарап нұсқауларда да кездесетіндіктен, жеке морфологиялық белгі мәтінді өздігінен деструктивті классқа жатқызбайды. Ол KazBERT қалыптастыратын контекстік көріністі толықтыратын грамматикалық ақпарат ретінде пайдаланылады.

Модельдің келесі нұсқасында контекстік және морфологиялық белгілер онтологияға негізделген семантикалық ақпаратпен толықтырылды. KazBERT_MFA_онтология_граф моделі осы үш белгі тобын бірге қолданады.

Мәтінде лексиконға енгізілген термин анықталған жағдайда, ол тиісті онтологиялық түйінмен сәйкестендіріледі. Осы сәйкестік негізінде модельге берілетін семантикалық белгілер қалыптастырылады.

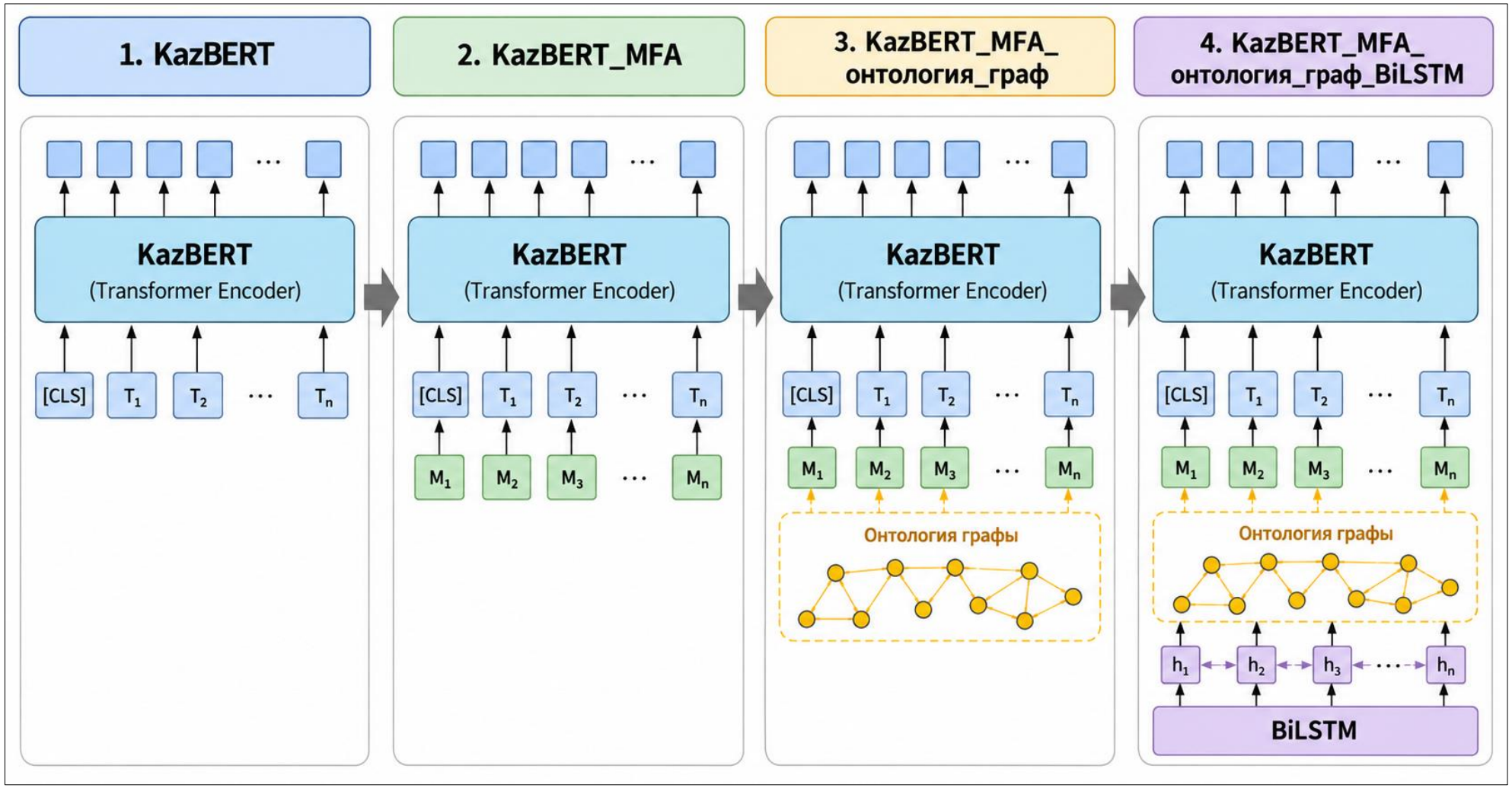
Бұл жерде терминді ұғыммен байланыстыру мен мәтінге түпкілікті класс тағайындау ажыратылады: онтологиялық сәйкестік модельге қосымша белгі береді, ал класс болжамы контекстік, морфологиялық және онтологияға негізделген семантикалық белгілерді бірге өңдеу нәтижесінде алынады.

KazBERT_MFA_онтология_граф_BiLSTM нұсқасы токендер тізбегін қосымша өңдеуімен ерекшеленеді. BiLSTM токендер тізбегін алға және кері бағытта өңдеп, екі бағыт бойынша алынған жасырын көріністерді біріктіреді. Сондықтан оның кірісінде токендердің бастапқы реті мен әр токенге сәйкес көріністер сақталуы қажет. Бұл ерекшелік мәтін деңгейінде жинақталған бір векторды токендер тізбегін сипаттайтын векторлық көріністерден ажыратуды талап етеді.

KazBERT те токеннің алдындағы және кейінгі контексті ескереді. Сондықтан BiLSTM-нің қосымша пайдасын тек екі бағытты өңдеуге сүйеніп негіздеу жеткіліксіз. Оның осы міндеттегі үлесі BiLSTM бар және жоқ нұсқаларды салыстыру арқылы анықталады. Сол сияқты жасырын деструктивті мағыналарды танудағы артықшылық туралы тұжырым осындай мәтіндер бойынша алынған нәтижелерге сүйенуі тиіс.

Сурет 3.3-те төрт модель нұсқасының құрамы салыстырылады. Солдан оңға қарай қосымша белгі көздері мен тізбектік өңдеу модулі енгізілген. Сурет компоненттер құрамының кеңеюін көрсетеді, блоктардың тік орналасуы есептеу амалдарының толық ретін білдірмейді. Нақты дерек ағыны, морфологиялық векторды қосу деңгейі және BiLSTM кірісі архитектураның кейінгі сипаттамасында жеке беріледі.

3.1-кестеде базалық және кеңейтілген модельдердің құрамы мен салыстыру мақсаты берілген.



Сурет 3.3 - Классификациялау модельдерінің кезең-кезеңімен кеңею схемасы

Кесте 3.1 - Базалық және кеңейтілген модельдердің құрамы мен салыстыру мақсаты

Модель	**Қолданылатын белгілер**	**Негізгі қызметі**	**Салыстырудағы мақсат**
KazBERT	Контекстік белгілер	Токендердің контекстік көріністерін қалыптастыру	Қосымша модульдерсіз базалық деңгейді анықтау
KazBERT_MFA	Контекстік және морфологиялық белгілер	30 өлшемді морфологиялық векторды қосу	KazBERT-пен салыстырып, морфологиялық белгілердің үлесін бағалау
KazBERT_MFA_онтология_граф	Контекстік, морфологиялық және онтологиялық белгілер	Танылған терминдерді 47 онтологиялық түйінмен байланыстыру	KazBERT_MFA-мен салыстырып, онтологиялық-графтық модульдің үлесін бағалау
KazBERT_MFA_онтология_граф_BiLSTM	Контекстік, морфологиялық, онтологиялық белгілер және токендер реті	Токендік көріністерді BiLSTM арқылы қосымша өңдеу	Алдыңғы нұсқамен салыстырып, BiLSTM модулінің үлесін бағалау

Салыстыруда деректер бөлінісі, алдын ала өңдеу тәртібі және бағалау өлшемдері бірізді болуы керек. Әр компоненттің үлесі қалған негізгі шарттары сәйкес келетін модель нұсқалары арқылы бағаланады.

Құрылымдық салыстыру компоненттердің қызметін ажыратады, ал олардың классификация сапасына әсері экспериментте бағаланады. Келесі бөлімдерде осы модельдерге қосылған белгілердің құрамы, оларды қалыптастыру және контекстік көріністермен біріктіру тәсілі сипатталады.

3.2 KazBERT негізіндегі морфологиялық және онтологиялық модульдерді әзірлеу

Бұл бөлімде 2-тарауда негізделген белгілерді модельге беру жолы қарастырылады. Морфологиялық модуль мәтіннің грамматикалық құрамын сандық түрде ұсынады. Онтологиялық-графтық модуль лексикон мен ұғымдар арасындағы байланыстарды пайдаланып, мәтінге қатысты семантикалық көрініс қалыптастырады. Екі модуль де KazBERT-тен алынатын контекстік ақпаратты толықтырады.

Морфологиялық белгілерді алу үшін бастапқы әдістемеде Stanza кітапханасының қазақ тіліне арналған талдауы көрсетілген [91]. Сөз табы мен морфологиялық сипаттамалар Universal Dependencies белгілеу жүйесіне сүйенеді [92]. Мұнда талдағыштың нәтижесі мен модельге берілетін вектор ажыратылады: талдағыш әр сөзге лингвистикалық белгілер тағайындайды, ал

MFA модулі сол белгілерді зерттеуде қабылданған сандық көрініске келтіреді.

Кесте 3.2-де өлшемдердің саны 30 болады. Сөз табына 17 UPOS категориясы бөлінген. Шақ пен жақ топтарындағы қосымша мән тиісті белгі анықталмаған жағдайды белгілейді. Рай тобындағы бесінші мән де дербес грамматикалық рай емес, кодтау кезінде пайдаланылатын анықталмаған мән ретінде беріледі. Бұл түсіндіру 2.2-бөлімдегі 30 өлшемді белгі құрамымен сәйкес келеді.

Кесте 3.2 - Морфологиялық белгі векторының құрамы

Белгі тобы	**Өлшем саны**	**Сипаттамасы**
Сөз табы	17	Зат есім, етістік, сын есім, есімдік, үстеу және басқа UPOS категориялары
Шақ	4	Өткен шақ, осы шақ, келер шақ және анықталмаған жағдай
Жақ	4	1-жақ, 2-жақ, 3-жақ және анықталмаған жағдай
Рай	5	Ашық рай, бұйрық рай, қалау рай, шартты рай және анықталмаған/модальдық мән
Барлығы	30	Мәтінді грамматикалық деңгейде сипаттайтын белгі кеңістігі

Сөз деңгейінде алынған белгілер мәтін деңгейінде жинақталып, m векторын құрайды. Оның координаттары грамматикалық категориялардың сандық сипаттамасын береді. Алайда санақ, салыстырмалы жиілік немесе белгінің бар-жоғын кодтау бірдей операция емес. Осы тәсілдердің қайсысы қолданылғанын алгоритмнің іске асырылуы анықтайды.

Төмендегі математикалық модель бұл операцияны $\mathrm{MFA}_{\mathrm{features}}$ функциясы арқылы белгілейді және бастапқы материалда көрсетілмеген жинақтау ережесін алдын ала бекітпейді.

Морфологиялық белгіні деструктивтіліктің тікелей дәлелі деп қабылдауға болмайды. Мысалы, бұйрық рай не екінші жақтағы тұлға қысым білдіретін сөйлемде де, бейтарап нұсқауда да кездесуі мүмкін. Сондықтан бұл белгілер өздігінен класс тағайындамайды. Олар контекстік көрініспен және онтологияға негізделген семантикалық белгілермен бірге өңделеді.

MFA шығысындағы 30 өлшемді вектор үйретілетін сызықтық түрлендіру арқылы жасырын кеңістікке көшіріледі. Одан кейін активация функциясы мен dropout қолданылады.

Бастапқы морфологиялық вектордың өлшемі мен түрлендірілген вектордың өлшемі бірдей болуға міндетті емес: біріншісі белгі құрамымен, екіншісі модель параметрімен анықталады. Онтологиялық-графтық модуль 2.3-бөлімде сипатталған білім моделіне негізделеді.

Онтология деструктивті веб-контенттің түрін, ашықтығын және нысанасын сипаттайды. Бұл өлшемдер мәтіннің әртүрлі қырын көрсетеді: түрі мазмұнның санатын, ашықтығы оның берілу сипатын, нысанасы кімге немесе қандай топқа бағытталғанын белгілейді. Лексикондағы терминді онтологиялық түйінге сәйкестендіру мен мәтіннің түпкілікті класын болжау бөлек амалдар болып саналады.

Мәтіннен терминдер танылғаннан кейін оларға сәйкес түйіндер белсендіріледі. 2.3-бөлімде берілген иерархиялық сәйкестендіру ережесі төменгі деңгейдегі түйіндермен қатар олардың ата-түйіндерін де ескеруге мүмкіндік береді. Мұндай ереже ұғымдардың онтологиядағы орнын сақтайды. Лексиконда кездеспеген бірліктер онтологиялық арнаға белгі қоспауы мүмкін, бірақ мәтіннің контекстік және морфологиялық өңделуіне қатыса береді.

GNN түйін белгілерін графтың көршілестік байланыстары бойынша жаңартады [93]. Бұл кезеңде терминнің қай түйінге жататыны ғана емес, сол түйіннің өзге ұғымдармен байланысы да ескеріледі.

Мысалы, кемсітуге қатысты лексикалық бірлік онтологияда буллингпен байланыстырылуы мүмкін [68]. Мұндай байланыс онтологияның сәйкестендіру ережесіне сүйенеді; жеке сөздің кездесуі мәтінді автоматты түрде деструктивті деп тануға жеткіліксіз.

Онтологиялық өңдеу үш негізгі амалдан тұрады: мәтіннен терминдерді тану және түйіндерге сәйкестендіру; графтық байланыстар арқылы түйін көріністерін жаңарту; алынған көріністерді мәтінге ортақ онтологиялық векторға жинақтау. Соңғы вектор KazBERT пен MFA белгілерімен біріктіріледі. Граф құрылымы мәтіндерге ортақ болуы мүмкін, ал белсендірілген түйіндер сол мәтінде танылған терминдерге тәуелді болады.

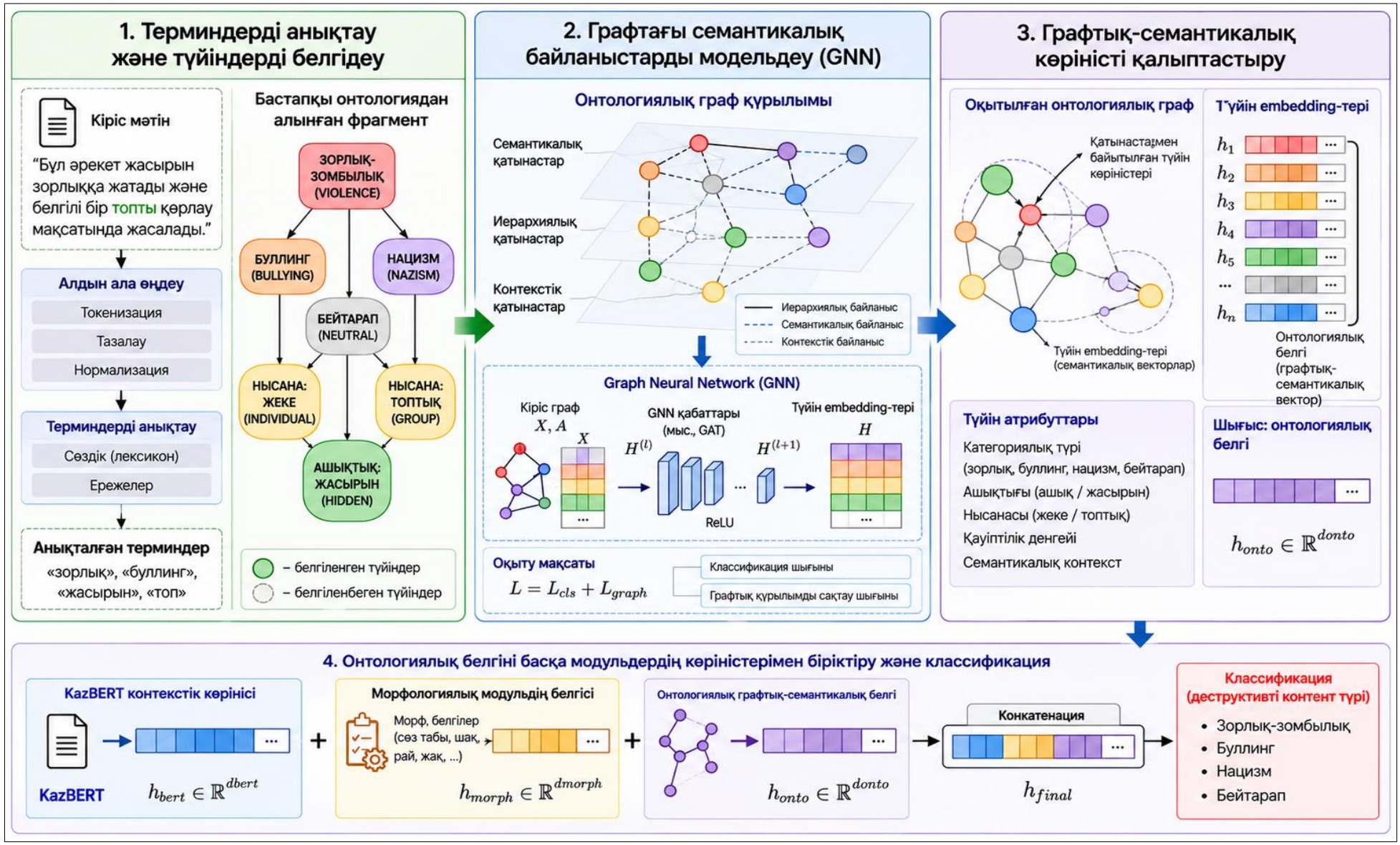
Кесте 3.3-те әр модульдің кірісі, шығысы және атқаратын қызметі көрсетілген. Мұнда 47 түйін саны онтологиялық вектордың жасырын өлшемімен теңестірілмейді. Түйіндер саны граф құрылымына, ал жасырын өлшем GNN параметрлеріне қатысты. Сол сияқты KazBERT алдымен токендік көріністер береді; мәтіндік вектор оларды жинақтағаннан кейін ғана алынады.

Кесте 3.3 - KazBERT негізіндегі қосымша модульдердің қызметі

Компонент	**Кіріс дерек**	**Шығыс белгі**	**Негізгі қызметі**
KazBERT	Қазақ тіліндегі токенделген мәтін	Контекстік көрініс	Мәтіннің жалпы мағынасын және сөздер арасындағы байланысты сипаттау
Морфологиялық модуль	Сөз табы, шақ, жақ, рай	30 өлшемді грамматикалық белгі	Қазақ тілінің морфологиялық ерекшелігін модельге қосу
Онтологиялық-графтық модуль	5000 терминдік лексикон, 47 түйін	онтологияға негізделген семантикалық белгілер	Терминдерді деструктивті категориялармен байланыстыру

Сурет 3.4-те лексикалық сәйкестендіру, онтологиялық түйіндерді белсендіру және графтық өңдеу көрсетілген.

Терминді тікелей мәтіннен табу мен оған мағыналас ұғымды онтология арқылы байланыстыру ажыратылуы керек. Суреттегі түйіндер арасындағы иерархиялық және семантикалық байланыстар да өз мағынасына сай түсіндіріледі.



Сурет 3.4 - Онтологиялық-графтық модульдің жұмыс сызбасы

Осылайша, қосымша модульдер мәтінге қатысты екі түрлі көрініс қалыптастырады: морфологиялық вектор грамматикалық белгілерді, онтологиялық вектор танылған ұғымдар мен олардың байланысын сипаттайды. Оларды контекстік көрініспен біріктіру тәртібі келесі бөлімде беріледі.

3.3 Контекстік, морфологиялық және онтологиялық белгілерді біріктірудің математикалық моделі

Ұсынылған гибридті әдісте қазақ тіліндегі веб-мәтін бірнеше белгі кеңістігі арқылы сипатталады. Бірінші белгі кеңістігі мәтіннің контекстік мағынасын береді, екінші белгі кеңістігі қазақ тілінің морфологиялық құрылымын сипаттайды, ал үшінші белгі кеңістігі деструктивті веб-контентке қатысты онтологияға негізделген семантикалық белгілерге байланысты көрсетеді. Бұл бөлімде осы үш түрлі ақпарат көзін бір классификациялау моделіне біріктірудің математикалық сипаттамасы беріледі.

Кіріс мәтінді $X = \{w_i, w_2, \dots, w_n\}$ деп белгілейміз. Мұндағы w_i – мәтіндегі жеке токен, w_n– токендер саны. Алдымен мәтін KazBERT моделіне беріледі. KazBERT контекстік көріністерді Transformer энкодері арқылы қалыптастырады. Энкодердегі масштабталған скалярлық көбейтіндіге негізделген attention операциясының өрнегі мынадай [34], жұмыста келтірілген (3.1) формула бойынша орындалады:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (3.1)$$

Мұндағы,

Q – сұраныс векторы,

K – кілт векторы,

V – мән векторы,

d_k – кілт векторының өлшемі.

Softmax әр сұраныс үшін кілттер осі бойынша қолданылады. Кірістегі тізбек ұзындығын теңестіруге арналған толтырғыш токендер (padding) бар жағдайда олардың attention есептеуіне қатысуы маскамен шектеледі.

Бір *attention* операциясын толық KazBERT энкодерінің шығысымен теңестіруге болмайды. Энкодер құрамында бірнеше attention басы, алға таратылатын қабаттар, қалдық байланыстар және нормалау амалдары бар. Сондықтан оның соңғы токендік көріністерін $H = KazBERT(X) = (h_1, \dots, h_n)$ деп белгілейміз. Мәтіндік векторды орташа жинақтау арқылы алатын нұсқа үшін:

Мәтін деңгейіндегі жинақталған контекстік белгі орташа мәні (3.2) формула арқылы анықталады:

$$h_{text} = \frac{\sum_{i=1}^{n} a_i h_i}{\sum_{i=1}^{n} a_i} \quad (3.2)$$

Мұндағы,

h_{text} – жинақталған контекстік вектор,

$a_i \in \{0, 1\}$ i– токеннің орташалауға қатысуын белгілейтін маска. Толтырғыш токендер орташаға қосылмайды, ал арнайы токендерді қосу не алып тастау қабылданған жинақтау ережесіне тәуелді. Егер барлық n токен есепке алынса, өрнек олардың қарапайым арифметикалық ортасына келтіріледі. (3.2) – орташа жинақтау қолданылатын нұсқаның сипаттамасы. [CLS] нұсқасында h_{text} ретінде тиісті токеннің көрінісі алынады.

Морфологиялық модуль сөз табы, шақ, жақ және рай белгілерін мәтін деңгейінде жинақтайды:

Бұл белгілер қазақ тіліндегі сөздердің грамматикалық тұлғасын (3.3) формула бойынша сипаттайды:

$$m = MFA_{features}(X),\ m \in \mathbb{R}^{30} \tag{3.3}$$

Мұндағы,

$MFA_{features}$ – мәтіннен морфологиялық сипаттамаларды бөліп алатын функция.

m векторының 30 координаты 3.2-кестедегі белгі топтарына сәйкес келеді.

Бұл функциялық жазу модульдің кірісі мен шығысын анықтайды. Оның ішіндегі категорияларды кодтау мен жинақтау ережелері бағдарламалық іске асырумен нақтыланады.

Морфологиялық белгілер модельдің жасырын кеңістігіне көшіріледі. Бұл үшін сызықтық түрлендіру, активация функциясы және $Dropout$ қабаты (3.4) формулаға сәйкес қолданылады:

$$h_{MFA} = Dropout\big(\sigma(W_M m + b_M)\big) \tag{3.4}$$

Мұндағы,

$W_M \in \mathbb{R}^{d_M \times 30}$ – үйретілетін салмақтар матрицасы,

$b_M \in \mathbb{R}^{d_M}$ – ығысу векторы,

σ – активация функциясы,

$h_{MFA} \in \mathbb{R}^{d_M}$ – түрлендірілген морфологиялық көрініс.

Dropout оқыту кезінде артық үйренуді тежеу үшін қолданылады. Оның үлесі мен активацияның нақты түрі модель конфигурациясында беріледі. Онтологиялық графты $G = (V, E)$ деп белгілейміз.

Мұндағы,

V – түйіндер жиыны,

E – қырлар жиыны,

A – осы графтың көршілестік матрицасы. Мәтінмен сәйкестендірілген түйіндердің бастапқы белгілерін $Z^{(0)}(X)$ арқылы белгілегенде, графтық өңдеудің жалпы түрі былай жазылады:

Үшінші кезеңде онтологиялық-графтық белгілер өңделеді. Онтологиялық граф деструктивті веб-контенттің түрі, ашықтығы және нысанасы бойынша

құрылған түйіндерден тұрады. Мәтіндегі терминдер онтологиялық лексикон арқылы анықталып, сәйкес түйіндермен байланысады. Түйіндер арасындағы семантикалық байланысты өңдеу үшін (3.5) формула қолданылады:

$$Z^{(l+1)}(X) = ReLU\left(GNN_l\left(Z^{(l)}(X), A\right)\right) \quad (3.5)$$

Мұндағы,

$Z^{(l)}(X)$ – l- қабаттағы түйін көріністерінің матрицасы,

A – онтологиялық графтың көршілестік матрицасы,

GNN_l – көрші түйіндерден ақпарат жинақтайтын графтық нейрондық операция,

$ReLU$ – активация функциясы,

$Z(l+1)$ – келесі қабаттағы жаңартылған түйін белгілері.

Контекстік, морфологиялық және онтологиялық белгілер алынғаннан кейін олар бір ортақ векторға (3.6) формула бойынша біріктіріледі:

$$h_{onto} = \frac{1}{|J|}\sum_{j\in J} z_j^{(L)}(X) \quad (3.6)$$

Мұндағы,

L – соңғы графтық қабат,

J – орташаға кіретін түйіндердің бос емес жиыны. Бұл жиынның барлық қолданылатын түйінді ме, әлде белсендірілген түйіндерді ғана ма қамтитыны іске асырудағы жинақтау ережесіне байланысты. Осы айырмашылық вектордың мағынасына әсер етеді, нәтижесі (3.7) формула арқылы көрінеді.

$$F = h_{text} \parallel h_{MFA} \parallel h_{onto} \quad (3.7)$$

Мұндағы,

$\parallel$ – белгісі векторларды бірінен кейін бірін жалғауды білдіреді.

Егер олардың өлшемдері тиісінше d_B, d_M және d_G болса, біріктірілген F векторының өлшемі $d_B + d_M + d_G$ болады. Конкатенация үшін үш вектордың өлшемін өзара теңестіру талап етілмейді. F мәтіннің контекстік көрінісін, грамматикалық белгілерін және онтологияға негізделген семантикалық белгілерін бірге қамтиды.

Соңғы кезеңде біріктірілген вектор классификация қабатына (3.8) формула арқылы беріледі:

$$\hat{y} = argmax_{c\in C}\left[softmax\left(FC\left(Dropout(F)\right)\right)\right]_c \quad (3.8)$$

Мұндағы,

FC – біріктірілген векторды кластар санына сәйкес бағаларға түрлендіретін толық байланысқан қабат,

C – кластар жиыны,

softmax – сол бағаларды ықтималдықтарға айналдыратын функция,

argmax – ықтималдығы ең жоғары кластың индексін таңдау амалы.

Төрт класты нұсқада *FC* шығысының өлшемі төртке тең. Класс атаулары мен олардың реттік сәйкестігі корпусты белгілеуде қабылданған тәртіппен беріледі. (3.1) - (3.8) өрнектері контекстік, морфологиялық және онтологиялық белгілерді мәтін деңгейінде біріктіруді сипаттайды. Рекурренттік нұсқада токендер тізбегі қосымша өңделетіндіктен, оның кірісі мен жинақтау орны бөлек анықталуы керек. Сондықтан бұл формулалар жиынын BiLSTM нұсқасының толық есептеу графы деп қарастыруға болмайды.

3.4 Ұсынылған гибридті классификациялау архитектурасы

Гибридті архитектура кіріс мәтінді үш ақпарат арнасы арқылы сипаттайды. Контекстік арнада KazBERT токендік векторлар тізбегін қалыптастырады. Морфологиялық арнада сөз табы, шақ, жақ және рай белгілері жинақталады. Онтологиялық арнада терминдер ұғымдармен сәйкестендіріліп, графтық байланыстар өңделеді. Бұл арналардың қызметі әртүрлі болғанымен, үшеуі де бір мәтінге қатысты белгілер береді.

Кіріс деректерін дайындау 2.1-бөлімде сипатталған корпус пен алдын ала өңдеу тәртібіне сүйенеді. KazBERT үшін мәтін оның токенизаторы арқылы беріледі. Морфологиялық талдағыш сөз деңгейіндегі сипаттамаларды алады, ал онтологиялық модуль лексиконмен сәйкестендіруді орындайды [29, 94]. Осы амалдарды баяндаудың реті бағдарламалық есептеуде олардың міндетті түрде бірінен кейін бірі орындалатынын білдірмейді. Рекурренттік өңдеу қажет болмаса, одан [CLS] немесе орташа жинақтау арқылы мәтіндік вектор алынады. KazBERT токенизаторы кіріс мәтінді ішкі сөз бірліктеріне бөледі. Тізбектің басына [CLS] арнайы токені, соңына [SEP] (3.9) формулаға сәйкес арнайы токен қосылады:

$$X = (\text{[CLS]},\ x_1,\ x_2,\ \ldots,\ x_n,\ \text{[SEP]}) \tag{3.9}$$

Мұндағы,

x_i – мәтіннің i-ші токені,

n – арнайы токендерсіз алынған токендер саны.

[CLS] токені бүкіл тізбекке ортақ көріністі жинақтау үшін қолданылады, ал [SEP] токені тізбектің (сегменттің) аяқталғанын белгілейді. Бұл токендердің өзінде лексикалық мағына жоқ, олардың векторлары энкодердің қабаттарында басқа токендермен өзара әрекеттесу нәтижесінде қалыптасады. Әр токеннің кіріс векторы токендік, позициялық және сегменттік ендірулердің қосындысы ретінде есептеледі ((3.10) формула):

$$e_i = E_{\text{tok}}(x_i) + E_{\text{pos}}(i) + E_{\text{seg}}(s_i), \qquad i = 0,1,\ldots,n+1 \tag{3.10}$$

Мұндағы,

E_{tok} – токендік ендіру матрицасы,

E_{pos} – позициялық ендіру,

E_{seg} – сегменттік ендіру,

s_i – токеннің сегмент нөмірі. Бір сегментті мәтінде барлық токендер үшін $s_i = 0$. Мұнда $i = 0$ позициясы [CLS] токеніне, $i = n + 1$ позициясы [SEP] токеніне сәйкес келеді. KazBERT энкодерінің шығысы токендік H матрицасы (3.11) формула арқылы анықталады:

$$H = \text{KazBERT}(e_0, e_1, \dots, e_{n+1}) = [h_{\text{CLS}},\ h_1,\ \dots,\ h_n,\ h_{\text{SEP}}] \in \mathbb{R}^{(n+2)\times d} \quad (3.11)$$

Мұндағы,

d – жасырын көріністің өлшемі,

h_i – i-ші токеннің контекстік векторы.

MFA арнасы 30 өлшемді белгілерді жасырын көріністе түрлендіреді. Онтологиялық арна GNN шығысындағы түйін көріністерінен мәтінге ортақ вектор қалыптастырады. Осы векторлардан KazBERT_MFA_онтология_граф тірек нұсқасының классификациялау кірісі құралады. Қосымша баптау кезеңінде [CLS], LSTM, BiGRU және BiLSTM негізіндегі нұсқалар баламалы эксперименттік конфигурациялар ретінде қарастырылады. Олар бір модельге ретімен жалғанатын төрт қабат емес. [CLS] токен көрінісін таңдау тәсілін білдіреді; LSTM, BiGRU және BiLSTM тізбекті өңдейтін модульдер болып саналады. Нұсқалардың барлығы үшін түпкі мақсат – мәтіндік белгілерден класс ықтималдықтарын есептеу. [CLS] нұсқасында контекстік арнадан энкодердің арнайы [CLS] токеніне сәйкес вектор алынады. Ол морфологиялық және онтологиялық көріністермен біріктіріліп, классификациялау қабатына беріледі. Бұл нұсқа қосымша рекурренттік өңдеуді қолданбайды. [CLS] векторын пайдалануды орташа жинақтаумен бір мезгілде орындалатын міндетті амал ретінде көрсетуге болмайды: екеуі мәтіндік көріністі алудың баламалары.

[CLS] нұсқасында мәтіндік вектор ретінде H матрицасының бірінші жолы (3.12) формула ретінде алынады:

$$h_{\text{CLS}} = H_0 \quad (3.12)$$

Баламалы тәсіл ретінде [CLS] және [SEP] токендерін есепке алмай, токендік векторлардың орташа мәні қолданылуы (3.13) формулада көрсетілген:

$$\bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i \quad (3.13)$$

(3.12) және (3.13) формулалары бір мезгілде қолданылатын амалдар емес, мәтіндік көріністі алудың екі баламасы болып табылады.

Контекстік вектор морфологиялық m және онтологиялық o көріністерімен (3.14) формулада біріктіріледі:

$$z_{\text{CLS}} = [h_{\text{CLS}};\ m;\ o] \quad (3.14)$$

Мұндағы,

$[\cdot;\cdot]$ – векторларды тізбектей біріктіру (конкатенация) амалы.

MFA арнасының 30 өлшемді белгілер векторы f жасырын көрініске (3.15) формулада түрлендіріледі:

$$m = \mathrm{ReLU}(W_m f + b_m), \qquad f \in \mathbb{R}^{30} \tag{3.15}$$

Рекуррентік нұсқаларда токендердің реті мен жеке көріністері сақталуы қажет. Сондықтан модульге мәтін деңгейінде жинақталған вектор емес, токендік векторлар тізбегі беріледі.

Әр t-ші токен үшін кіріс векторы KazBERT векторы мен сол токен тиесілі сөздің морфологиялық белгілерінен (3.16) формулада құралады:

$$u_t = \left[h_t;\ W_m f_{w(t)} + b_m\right], \qquad t = 1, \ldots, n \tag{3.16}$$

Мұндағы,

$w(t)$ – t-ші токен кіретін сөздің нөмірі,

$f_{w(t)}$ – сол сөздің морфологиялық белгілер векторы. Бір сөз бірнеше ішкі токенге бөлінсе, олардың барлығына сол сөздің белгілері беріледі.

LSTM нұсқасы токендер тізбегін бір бағытта өңдейді. Оның жасырын күйлері алдыңғы позициялардан келген ақпаратты ескереді. Осылайша, классификациялауға берілетін көрініс токендердің ретіне тәуелді түрде қалыптасады. Қай жасырын күйдің немесе қандай жинақтау нәтижесінің алынатыны осы нұсқаның нақты құрылымында көрсетілуге тиіс. LSTM тізбекті бір бағытта ($t = 1$-ден $t = n$-ге дейін) өңдейді. t-ші қадамдағы есептеулер төмендегі (3.17-3.22) формулаларда көрсетілген:

$$f_t = \sigma\left(W_f u_t + U_f h_{t-1} + b_f\right) \tag{3.17}$$

$$i_t = \sigma(W_i u_t + U_i h_{t-1} + b_i) \tag{3.18}$$

$$o_t = \sigma(W_o u_t + U_o h_{t-1} + b_o) \tag{3.19}$$

$$\tilde{c}_t = \tanh(W_c u_t + U_c h_{t-1} + b_c) \tag{3.20}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{3.21}$$

$$h_t = o_t \odot \tanh(c_t) \tag{3.22}$$

Мұндағы,

f_t – ұмыту қақпасы,

i_t – кіріс қақпасы,

o_t – шығыс қақпасы,

$\tilde{c}_t$ – жадының үміткер күйі,

c_t – жады ұяшығының күйі,

h_t – жасырын күй,
W, U – салмақ матрицалары,
b – ығысу векторлары,
σ – сигмоид функциясы,
$\odot$ – элементтер бойынша көбейту. Бастапқы күйлер $h_0 = 0,\ c_0 = 0$ деп алынады. Классификациялау үшін соңғы жасырын күй (3.23) формула арқылы алынады:

$$s_{\text{LSTM}} = h_n \tag{3.23}$$

BiLSTM нұсқасында тізбек алға және кері бағытта өңделеді. Екі бағыттың көріністері біріктіріліп, токеннің алдындағы және кейінгі позициялар туралы ақпарат қамтылады. KazBERT те токеннің алдындағы және кейінгі контексті ескереді, сондықтан BiLSTM қосу қажеттілігін тек «екі бағытты ескереді» деген уәжбен түпкілікті негіздеу жеткіліксіз. Оның қосымша үлесі осы модулі бар және жоқ нұсқалардың нәтижесін салыстыру арқылы бағаланады.

BiLSTM-де тізбек алға және кері бағытта екі тәуелсіз LSTM арқылы (3.24) формула түрінде өңделеді:

$$\begin{aligned} \overrightarrow{h_t} &= \text{LSTM}_{\rightarrow}\left(u_t,\ \vec{h}_{t-1}\right), \qquad t = 1, \dots, n \\ \overleftarrow{h_t} &= \text{LSTM}_{\rightarrow}\left(u_t,\ \vec{h}_{t+1}\right), \qquad t = 1, \dots, n \end{aligned} \tag{3.24}$$

Әр позициядағы көрініс екі бағыттың күйлерін біріктіру үшін (3.25) формуланы пайдаланады:

$$h_t^{\text{Bi}} = \left[\vec{h}_t;\ \overleftarrow{h}_t\right] \tag{3.25}$$

Мәтіндік вектор ретінде тура бағыттың соңғы күйі мен кері бағыттың соңғы (яғни бірінші позициядағы) күйі (3.26) формуладан алынады:

$$s_{\text{BiLSTM}} = \left[\vec{h}_n;\ \overleftarrow{h}_1\right] \tag{3.26}$$

BiGRU нұсқасы да тізбекті екі бағытта өңдейді. Оның күйін жаңарту GRU қақпаларына негізделеді. Бұл нұсқа BiLSTM-мен салыстырылатын рекурренттік балама ретінде қарастырылады. Есептеу шығыны немесе орындалу жылдамдығы жөніндегі тұжырым нақты өлшемдерге, қабаттар санына және эксперимент нәтижесіне сүйенуі керек. GRU-да күйді жаңарту екі қақпа арқылы жүзеге асады, t-ші қадамдағы есептеулер төмендегі (3.27-3.30) формулалар көмегімен өңделеді:

$$z_t = \sigma(W_z u_t + U_z h_{t-1} + b_z) \tag{3.27}$$

$$r_t = \sigma(W_r u_t + U_r h_{t-1} + b_r) \tag{3.28}$$

$$\tilde{h}_t = \tanh(W_h u_t + U_h(r_t \odot h_{t-1}) + b_h) \quad (3.29)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (3.30)$$

Мұндағы,

z_t – жаңарту қақпасы,

r_t – қалпына келтіру (reset) қақпасы,

$\tilde{h}_t$ – үміткер жасырын күй.

LSTM-нен айырмашылығы – GRU-да жеке жады ұяшығы c_t жоқ, ақпарат тек h_t күйінде сақталады. BiGRU-да да тізбек екі бағытта (3.31-3.32) формулалар арқылы өңделеді:

$$\vec{h}_t = \mathrm{GRU}_{\rightarrow}(u_t, \vec{h}_{t-1}), \qquad \overleftarrow{h}_t = \mathrm{GRU}_{\leftarrow}(u_t, \overleftarrow{h}_{t+1}) \quad (3.31):$$

$$s_{\mathrm{BiGRU}} = [\vec{h}_n; \overleftarrow{h}_1] \quad (3.32)$$

Рекурренттік нұсқаларда онтологиялық вектор токендік деңгейде емес, тізбектен алынған мәтіндік вектормен бірге (3.33) формула арқылы қосылады:

$$z = [s; o], \qquad s \in \{s_{\mathrm{LSTM}}, s_{\mathrm{BiLSTM}}, s_{\mathrm{BiGRU}}\} \quad (3.33)$$

Барлық нұсқа үшін класс ықтималдықтары softmax функциясымен (3.34) формула көмегімен есептеледі:

$$p(y = k \mid X) = \frac{\exp(w_k^{\top} z + b_k)}{\sum_{j=1}^{K} \exp(w_j^{\top} z + b_j)}, \qquad k = 1, \ldots, K \quad (3.34)$$

Мұндағы,

K – кластар саны. Болжамды класс ең үлкен ықтималдық бойынша (3.35) формула арқылы таңдалады:

$$\hat{y} = \underset{k}{\mathrm{argmax}}\, p(y = k \mid X) \quad (3.35)$$

Оқыту кезінде модель параметрлері кросс-энтропия шығын функциясын азайту (3.36) формуласы көмегімен бапталады:

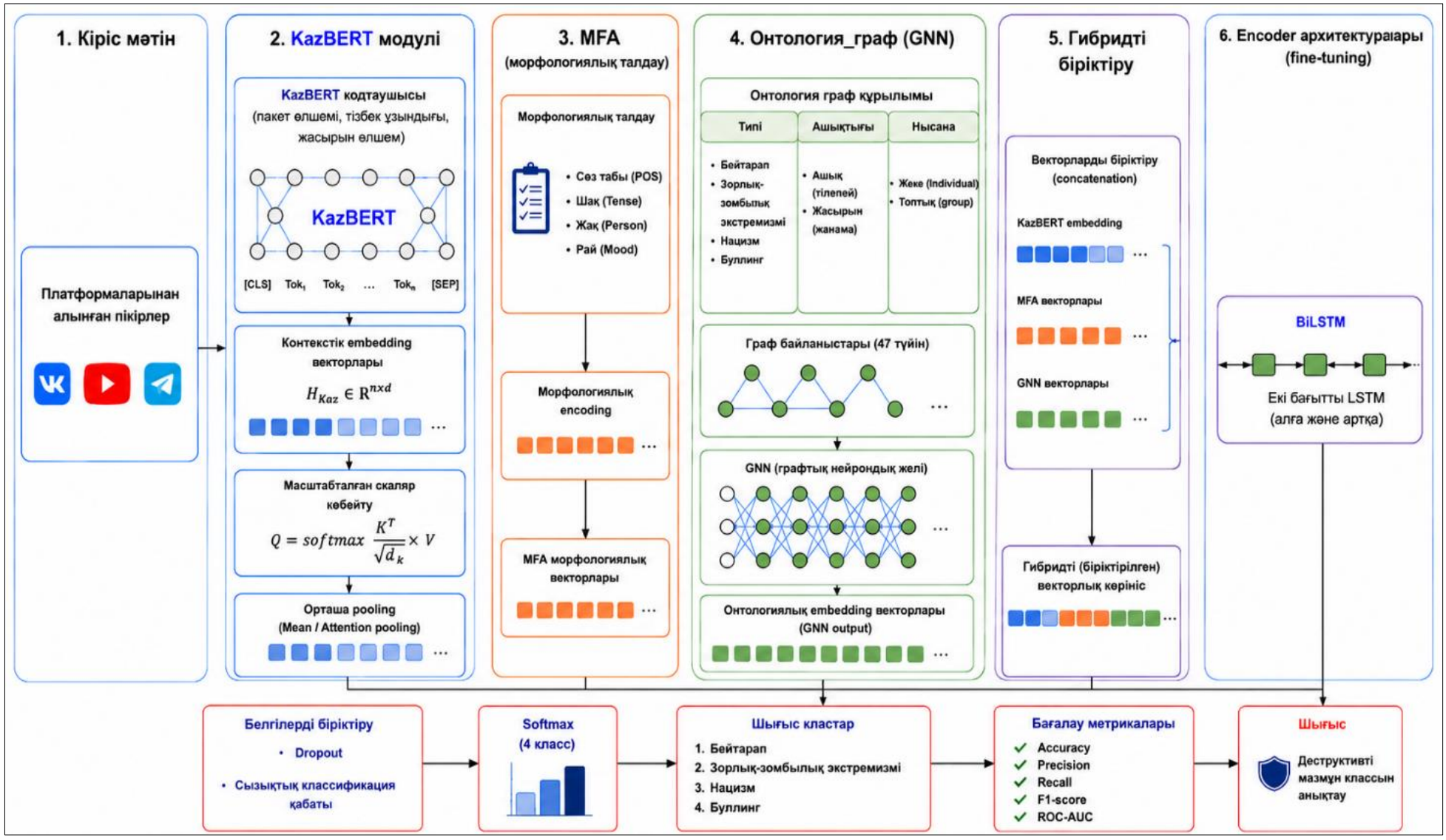
$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log p(y = k \mid X_i) \quad (3.36)$$

Мұндағы,

N – оқыту жиынындағы мәтіндер саны,

$y_{i,k}$ – i-ші мәтіннің шын класы k болса 1-ге, әйтпесе 0-ге тең белгі.

Сурет 3.5-те кіріс мәтіннен бастап контекстік, морфологиялық және онтологиялық белгілерді алу, оларды біріктіру, BiLSTM арқылы өңдеу және класс ықтималдықтарын есептеу кезеңдері көрсетілген [54].



Сурет 3.5 - Ұсынылған гибридті классификациялау архитектурасының жалпы схемасы

Attention операциясы (3.1)-формулаға сәйкес есептеледі. Графикалық сызбадағы жинақтау мен BiLSTM арасындағы байланыс жоғарыда көрсетілген тізбек шартымен бірге оқылуы керек.

Архитектураның шығысы – кластар бойынша ықтималдықтар және солардың ішінен таңдалған болжам. Суреттегі Accuracy, Precision, Recall және F1 көрсеткіштері осы болжамдарды дұрыс класс белгілерімен салыстырып бағалау кезінде есептеледі.

Қосымша баптау параметрлерді классификациялау міндетіне бейімдеуді білдіреді. Оның нақты режимі – энкодердің қандай қабаттары жаңартылатыны, қандай компоненттер бекітілетіні және қандай шығын функциясы қолданылатыны – оқыту конфигурациясымен анықталады. Осы мәліметтер мен эксперименттік нәтижелер модель нұсқаларының әділ салыстырылуына негіз болады.

3.5 3-тарау бойынша қорытынды

Бұл тарауда қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған гибридті әдістің құрамы мен белгілерді біріктіру тәртібі сипатталды. KazBERT базалық контекстік модель ретінде алынды. Морфологиялық модуль сөз табы, шақ, жақ және рай белгілерін 30 өлшемді векторға жинақтайды. Онтологиялық-графтық модуль 5000 терминнен тұратын лексиконға және 47 қолданылатын түйінге сүйеніп, онтологияға негізделген семантикалық белгілер қалыптастырады.

Базалық және кеңейтілген модельдердің айырмашылығы құрылымдық тұрғыдан көрсетілді. Мәтіндік, морфологиялық және онтологиялық көріністерді конкатенациялау, оларды классификациялау қабатына беру және класс болжамын таңдау формалданды. [CLS], LSTM, BiGRU және BiLSTM негізіндегі нұсқалардың баламалы екені, ал рекуррентік өңдеуде токендер тізбегінің сақталуы керектігі ажыратылды.

Ұсынылған компоненттердің классификация сапасына әсері эксперименттік бағалау арқылы анықталады. Эксперименттік бағалау үшін Accuracy, Macro-F1, Precision, Recall, ROC-AUC, PR-AUC және MCC көрсеткіштері қарастырылған [54]. Көпкластық жағдайда көрсеткіштерді біріктіру тәсілі мен әр класс бойынша нәтижелер нақты берілуі тиіс, қосымша модульдердің пайдалы болғаны және қай нұсқаның жоғары нәтиже көрсеткені туралы қорытынды аталған бағалаулар негізінде ғана алынады.

4 Ұсынылған әдістердің эксперименттік зерттелуі және олардың тиімділігін бағалау

Бұл тарауда алдыңғы бөлімде ұсынылған гибридті әдістің эксперименттік тиімділігі бағаланды. KazBERT моделіне морфологиялық және онтологияға негізделген семантикалық белгілерді қосу әсері, сондай-ақ [CLS], LSTM, BiLSTM және BiGRU негізіндегі баламалы классификациялау нұсқалары салыстырылды [50, 54]. Модельдердің нәтижелері Accuracy, Macro-F1, Precision, Recall, ROC-AUC, PR-AUC және MCC метрикалары бойынша бағаланды [95].

4.1 Эксперименттік орта және бағалау әдістемесі

Модельдерді өзара салыстыру үшін барлық эксперименттер бірдей деректер жиыны мен деректерді бөлу тәртібі негізінде жүргізіліп, бағалау кезінде ортақ метрикалар қолданылды [96]. Эксперименттік жағдайлардың бірдей болуы әр модель бойынша алынған нәтижелерді тікелей салыстыруға негіз болды [97].

Деректер жиыны 70/15/15 қатынасында оқыту, валидация және тест бөліктеріне бөлінді. Оқыту бөлігі модель параметрлерін оңтайландыруға пайдаланылды [98], ал валидациялық деректер оқыту барысын бақылау және оңтайлы конфигурацияны таңдау кезінде қолданылды. Тест жиыны бұл кезеңдерге қатыстырылмай, модельдің қорытынды сапасын бағалау үшін ғана сақталды. Мұндай бөлу модельдің оқыту кезінде кездеспеген деректердегі нәтижесін жеке бағалауға жағдай жасайды. Эксперименттік деректердің бөліну тәртібі кесте 4.1-де берілген.

Кесте 4.1 - Эксперименттік деректер жиынын бөлу тәртібі

Деректер бөлігі	**Үлесі**	**Қолданылу мақсаты**
Оқыту жиыны	70%	Модель параметрлерін үйрету
Валидация жиыны	15%	Оқыту барысын бақылау және модель таңдау
Тест жиыны	15%	Соңғы нәтижені бағалау

Нәтижелердің кездейсоқ инициализацияға сезімталдығын бағалау үшін модельдер кездейсоқтық генераторының бастапқы мәндерімен (random seed) қайта іске қосылды [99]. Бірдей деректер бөлінісі сақталған жағдайда бұл тәсіл модель параметрлерінің инициализациясы мен оқытудағы кездейсоқтықтың әсерін бағалауға мүмкіндік береді [100]. Қорытынды көрсеткіштер бес іске қосудың орташа мәні және стандарттық ауытқуы түрінде берілді [95].

Эксперименттік салыстыруда KazBERT контекстік көрінісі, 30 өлшемді MFA векторы және онтологиялық-графтық семантикалық көрініс бірдей дайындық хаттамасы бойынша пайдаланылды.

Эксперименттік орта Python негізіндегі табиғи тілді өңдеу және терең оқыту кітапханалары арқылы ұйымдастырылды. KazBERT және оның кеңейтілген нұсқалары PyTorch және Transformers кітапханалары көмегімен іске асырылды [101]. Бағалау метрикаларын есептеу үшін көпкластық

классификацияда кең қолданылатын Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC және MCC көрсеткіштері алынды, кесте 4.2-де көрсетілген.

Кесте 4.2 - Эксперименттік конфигурацияның негізгі параметрлері

Параметр	**Сипаттамасы**
Базалық модель	KazBERT
Класстар саны	4
Деректерді бөлу	70% / 15% / 15%
Мәтінді ұсыну	KazBERT эмбеддинг
Морфологиялық белгілер	сөз табы, шақ, жақ, рай
Морфологиялық вектор өлшемі	30
Онтологиялық лексикон	5000 термин
Онтологиялық түйіндер	47 қолданылатын түйін
Бағалау метрикалары	Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, MCC

Бағалау әдістемесінде тек жалпы дәлдікке сүйену жеткіліксіз. Себебі деструктивті веб-контентті анықтау міндетінде әр кластың дұрыс танылуы маңызды. Мысалы, модель бейтарап мәтіндерді жақсы танығанымен, буллинг немесе нацизм класын нашар ажыратса, мұндай модельді толық тиімді деп бағалау қиын. Сондықтан Precision, Recall және F1-score көрсеткіштері macro-average түрінде есептелді.

Macro-average әр кластың нәтижесін жеке есептеп [97], кейін олардың орташа мәнін алады. Бұл тәсіл көпкластық классификацияда әр классқа тең мән беруге мүмкіндік береді [95].

Accuracy жалпы дұрыс болжанған мәтіндердің үлесін (4.1) формула арқылы көрсетеді:

$$Accuracy = \frac{\sum_{k=1}^{K} CM_{kk}}{\sum_{i=1}^{K}\sum_{j=1}^{K} CM_{ij}} \quad (4.1)$$

Мұндағы,

$K = 4$ – кластар саны,

CM_{kk} – *confusion matrix* диагоналіндегі дұрыс болжамдар,

бөлім – тест жиынындағы барлық мәтін саны.

Precision модель белгілі бір классқа жатқызған мәтіндердің қаншасы шынымен сол классқа тиесілі екенін (4.2) формула түрінде анықтайды:

$$Precision = \frac{TP}{TP+FP} \quad (4.2)$$

Recall нақты сол классқа жататын мәтіндердің қаншасын модель дұрыс анықтағанын (4.3) формула көмегімен есептейді:

$$Recall = \frac{TP}{TP+FN} \quad (4.3)$$

F1-score Precision мен Recall арасындағы теңгерімі болып табылады ((4.4) формула):

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4.4)$$

ROC-AUC модельдің кластарды ықтималдық деңгейінде ажырату қабілетін бағалау үшін қолданылды. Көпкластық жағдайда бұл көрсеткіш әр класс үшін one-vs-rest тәсілімен есептеліп, кейін орташа мәні алынды. PR-AUC Precision-Recall қисығы бойынша есептеледі және деструктивті кластарды сенімді ажырату сапасын бағалауға мүмкіндік береді [34, 95].

Қосымша түрде Matthews Correlation Coefficient өлшемі қолданылды. Мэттьюс корреляция коэффициенті (MCC) нақты және модель болжаған класс белгілерінің арасындағы сәйкестікті бағалау үшін қолданылады. Екі класты классификациялау жағдайында бұл коэффициент TP, TN, FP және FN шамалары арқылы (4.5) формуламен анықталады.

Зерттеуде төрт класс қарастырылғандықтан, эксперименттік нәтижелерді бағалау кезінде MCC коэффициентінің көпкласты жалпыланған нұсқасы қолданылады. Ол толық 4×4 қателесу матрицасы негізінде есептеледі. Осылайша, (4.5) формула коэффициенттің бинарлық жағдайдағы анықтамасын көрсетеді, ал төрт класты эксперименттің нәтижелері көпкласты MCC арқылы бағаланады. Жеке кластар үшін есептелген MCC мәндерін орташалау қолданылмайды.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4.5)$$

Осылайша, эксперименттік бағалау бір ғана көрсеткішке сүйенбей, модель сапасын бірнеше қырынан қарастыруға бағытталды. Accuracy жалпы дәлдікті көрсетсе, Precision мен Recall әр кластың танылу сапасын нақтылайды. F1-score осы екі көрсеткіштің теңгерімін береді. ROC-AUC және PR-AUC модельдің кластарды ықтималдық мәндері бойынша ажырату қабілетін бағалау үшін қолданылды. MCC классификация нәтижесін кластар бойынша теңгерімді түрде бағалайтын қосымша көрсеткіш ретінде пайдаланылды. Осы бағалау әдістемесі бойынша алынған модель нәтижелері келесі бөлімде салыстырылады.

4.2 Ұсынылған модельді базалық әдістермен салыстыру және бағалау

Қазақ тіліндегі деструктивті веб-контентті классификациялау үшін KazBERT негізінде құрылған бірнеше модель нұсқасының нәтижелері салыстырылды. Салыстырудың негізгі мақсаты – базалық контекстік модельден бастап, оған морфологиялық, онтологиялық-графтық және тізбектік белгілерді қосқан кезде классификация сапасының қалай өзгеретінін анықтау [81]. Осы арқылы ұсынылған гибридті архитектураның базалық KazBERT моделінен

артықшылығы нақты метрикалар негізінде бағаланды [85]. Эксперименттік салыстыруға KazBERT негізіндегі бірнеше модель нұсқасы енгізілді. Олар контекстік белгілерге ғана сүйенетін базалық модельден бастап, морфологиялық, онтологиялық-графтық және тізбектік компоненттермен толықтырылған гибридті модельдерге дейін кезең-кезеңімен ұйымдастырылды [79]. Барлық модельдер бірдей деректер жиыны, бірдей train/validation/test бөлінісі және бірдей бағалау метрикалары бойынша тексерілді [96]. Эксперименттің мұндай жүйелі тәртібі модельдерді объективті түрде салыстыруға негіз болды [98]. Зерттеу логикасы қарапайымнан күрделіге қарай сатылап жүргізілді, бастапқыда тек KazBERT контекстік белгілері талданса, кейіннен оларға морфологиялық және онтологиялық-графтық семантикалық сипаттамалар қосылды. Қорытынды кезеңде қалыптасқан біріктірілген белгілер кеңістігі CLS, LSTM, BiLSTM және BiGRU энкодерлері арқылы бапталып (fine-tuning), олардың классификациялау тиімділігіне әсері бағаланды.

Кесте 4.3-те бойынша базалық KazBERT моделінің Accuracy көрсеткіші 88,015%, Macro-F1 көрсеткіші 88,086% болды. Бұл модель зерттеуде кейінгі архитектуралардың нәтижелерін салыстыруға арналған базалық (baseline) нұсқа ретінде пайдаланылды.

Transformer архитектурасына негізделген BERT модельдерінің мәтіндегі контекстік ақпаратты өңдеу ерекшеліктері [14] еңбегінде қарастырылған, ал KazBERT-тің қазақ тіліндегі веб-контентті талдауда қолданылуы [34] зерттеуінде сипатталған. Базалық KazBERT мәтінді негізінен контекстік көріністер арқылы сипаттайды; морфологиялық белгілер мен деструктивті веб-контентке қатысты онтологияға негізделген семантикалық белгілер бұл модельде жеке ақпарат көзі ретінде қолданылмайды.

Кесте 4.3 - KazBERT негізіндегі модельдердің салыстырмалы нәтижесі

Модель	**Accuracy, %**	**Macro-F1, %**	**Precision, %**	**Recall, %**	**ROC-AUC, %**	**PR-AUC, %**	**MCC, %**
KazBERT	88.015	88.086	88.181	88.033	97.489	94.700	84.005
KazBERT_MFA	88.513	88.601	88.743	88.590	97.867	95.065	84.702
KazBERT_MFA_ онтология_граф	90.492	90.567	90.660	90.527	98.676	96.808	87.319
KazBERT_MFA_ онтология_граф_CLS	90.603	90.686	90.804	90.627	98.681	96.823	87.467
KazBERT_MFA_ онтология_граф_ LSTM	90.547	90.632	90.765	90.576	98.705	96.848	87.400
KazBERT_MFA_ онтология_граф_ BiLSTM	91.045	91.103	91.153	91.096	98.750	96.972	88.054
KazBERT_MFA_ онтология_граф_ BiGRU	90.713	90.802	90.939	90.738	98.722	96.904	87.620

KazBERT_MFA нұсқасында KazBERT қалыптастыратын контекстік көріністерге морфологиялық белгілер қосылды. Осы конфигурацияда Accuracy 88,015%-дан 88,513%-ға, Macro-F1 88,086%-дан 88,601%-ға дейін өсті. Precision 88,743%, Recall 88,590% деңгейінде тіркелді. Алынған көрсеткіштер морфологиялық белгілер қосылған конфигурацияда базалық KazBERT-пен салыстырғанда классификациялау нәтижелерінің жоғарылағанын көрсетеді. Қазақ тілінің морфологиялық ерекшеліктерін автоматты мәтіндік талдауда ескеру мәселесі бұған дейінгі зерттеулерде де қарастырылған [33]. Сөз табы мен грамматикалық категориялар осы жағдайда мәтіннің грамматикалық құрылымы туралы қосымша ақпарат беретін белгілер ретінде қолданылады [67].

KazBERT_MFA моделінде базалық KazBERT қалыптастыратын контекстік көріністер морфологиялық белгілермен толықтырылды. Осы өзгерістен кейін Accuracy және Macro-F1 мәндерінің өсуі байқалды. Бұл нәтиже морфологиялық белгілердің контекстік көріністерге қосымша грамматикалық ақпарат бергенін көрсетеді. Қазақ тілін автоматты өңдеуге арналған алдыңғы зерттеулерде де морфологиялық талдау маңызды бағыттардың бірі ретінде қарастырылған [33].

Келесі кезеңде модельге онтологияға негізделген семантикалық белгілер енгізілді. KazBERT_MFA моделімен салыстырғанда Accuracy 1,979 пайыздық пунктке, ал Macro-F1 1,966 пайыздық пунктке өсті. Морфологиялық белгілерді қосқан кезеңмен салыстырғанда бұл өзгеріс айқынырақ болды. Мұны онтологиялық компоненттің мәтіндегі сөздер мен сөз тіркестерін деструктивті веб-контентке қатысты ұғымдармен байланыстыруы арқылы түсіндіруге болады [54]. Яғни модель контекстік және грамматикалық ақпаратпен қатар пәндік семантикалық байланыстарды да ескере бастайды.

Онтологияға негізделген семантикалық белгілер мәтіндегі лексикалық бірліктерді жеке сөз деңгейінде ғана емес, олардың онтологиялық ұғымдармен байланысы тұрғысынан сипаттайды. Мұндай тәсіл қауіпті мазмұнды онтологиялық білім арқылы сипаттауға бағытталған зерттеулермен үйлеседі [68]. Сондықтан онтологиялық компонент контекстік және морфологиялық белгілерді мәтіннің мағыналық бағыты жөніндегі қосымша ақпаратпен толықтырады [69].

Келесі салыстыруда KazBERT_MFA_онтология_граф конфигурациясы негізінде [CLS], LSTM, BiGRU және BiLSTM нұсқалары қарастырылды. [CLS] және LSTM нәтижелері бір-біріне жақын болды, ал BiGRU екі метрика бойынша да олардан жоғары нәтиже көрсетті. Ең жоғары көрсеткіш BiLSTM қолданылған нұсқада алынды. Бұл айырмашылық токендер тізбегін екі бағытта өңдеудің қарастырылған конфигурацияда тиімді болғанын көрсетеді. LSTM мәтіндегі тізбектік тәуелділіктерді өңдеуге арналған рекуррентті архитектуралардың бірі болып табылады [11], ал BiLSTM тізбекті екі бағытта өңдеу арқылы алдыңғы және кейінгі контексті қатар ескеруге мүмкіндік береді. Салыстырылған конфигурациялар ішінде KazBERT_MFA_онтология_граф_BiLSTM моделі ең жоғары нәтижеге жетті. Базалық KazBERT-пен салыстырғанда оның Accuracy көрсеткіші 3,030 пайыздық пунктке, Macro-F1 көрсеткіші 3,017 пайыздық пунктке, ал MCC көрсеткіші 4,049 пайыздық пунктке жоғары болды. Бұл

айырмашылықты бір ғана компоненттің әсерімен түсіндіру жеткіліксіз. Нәтиже контекстік көріністерді, морфологиялық белгілерді, онтологияға негізделген семантикалық белгілерді және екі бағытты тізбектік өңдеуді бір архитектурада қолданумен байланысты [50].

Жекелеген компоненттердің әсерін нақты ажырату үшін олардың үлесі абляциялық эксперименттер арқылы қосымша бағаланды.

4.3 Ұсынылған гибридті архитектура компоненттерінің тиімділігін талдау

Компоненттік талдау модельдегі төрт ақпарат көзінің үлесін жеке қарастыруға мүмкіндік береді. KazBERT контекстік көріністерді қалыптастырады, MFA морфологиялық белгілер арқылы грамматикалық ақпарат қосады, онтологияға негізделген семантикалық белгілер пәндік мағыналық байланыстарды сипаттайды, ал рекурренттік нұсқалар токендер тізбегін қосымша өңдейді [14, 33, 68]. Әр компоненттің ықпалы негізгі эксперименттік шарттары бірдей, бірақ бір құрылымдық элементі бойынша ерекшеленетін модель нұсқаларын өзара салыстыру арқылы бағаланды.

Кесте 4.4-те Δ мәндері пайыздық пунктпен берілген және әр компонентті қосқаннан кейінгі өзгерісті көрсетеді. Алғашқы екі жол алдыңғы модельге қатысты, ал [CLS], LSTM, BiGRU және BiLSTM нұсқалары KazBERT_MFA_онтология_граф конфигурациясымен салыстырылды.

Кесте 4.4 - Компоненттерді қосу нәтижесіндегі метрикалық өсім

Салыстыру	**Қосылған компонент**	**ΔAccuracy, п.п.**	**ΔMacro-F1, п.п.**	**ΔMCC, п.п.**
KazBERT → KazBERT_MFA	Морфологиялық белгілер	+0.498	+0.515	+0.697
KazBERT_MFA → KazBERT_MFA_онтология_граф	онтологияға негізделген семантикалық белгілер	+1.979	+1.966	+2.617
KazBERT_MFA_онтология_граф → CLS	[CLS] мәтіндік көрінісі	+0.111	+0.119	+0.148
KazBERT_MFA_онтология_граф → LSTM	LSTM қабаты	+0.055	+0.065	+0.081
KazBERT_MFA_онтология_граф → BiGRU	BiGRU қабаты	+0.221	+0.235	+0.301
KazBERT_MFA_онтология_граф → BiLSTM	BiLSTM қабаты	+0.553	+0.536	+0.735

MFA қосылғаннан кейін Accuracy +0,498, Macro-F1 +0,515 және MCC +0,697 пайыздық пунктке өзгерді. Бұл нәтиже морфологиялық ақпараттың

берілген деректер жиынында қосымша пайдалы белгі болғанын көрсетеді [33, 84]. Онтологияға негізделген семантикалық белгілерді енгізгенде ең ірі кезеңдік өсім байқалды: Accuracy +1,979, Macro-F1 +1,966, MCC +2,617 пайыздық пункт [54]. Бұл өзгеріс пәндік семантиканы графтық түрде берудің модель шешіміне қосымша ақпарат енгізгенін көрсетеді [67].

MCC-дің +2,617 пайыздық пунктке өсуі жалпы дәлдікпен қатар көпкластық шешімдердің теңгерімділігі де жақсарғанын көрсетеді. Келесі экспериментте [CLS], LSTM, BiGRU және BiLSTM баламалары бірдей негізгі белгі кеңістігінде салыстырылды; мақсат – қосымша тізбектік өңдеудің үлесін бөлек бағалау.

[CLS] нұсқасында Accuracy +0,111 және Macro-F1 +0,119 пайыздық пунктке өсті. Өсімнің шағын болуы бұл конфигурацияның тірек модельге жақын нәтиже бергенін көрсетеді. Осы айырмадан [CLS] векторының «толыққанды» артықшылығы туралы жеке қорытынды жасауға негіз жеткіліксіз.

LSTM нұсқасының өсімі Accuracy бойынша +0.055, Macro-F1 бойынша +0,065, MCC бойынша +0,081 пайыздық пункт болды. Бұл қосымша тізбектік өңдеудің шағын оң өзгеріспен байланысты болғанын көрсетеді.

BiGRU нұсқасының өсімі Accuracy +0,221, Macro-F1 +0,235 және MCC +0,301 пайыздық пункт болды. Ол LSTM нұсқасынан жоғары, бірақ BiLSTM нұсқасынан төмен нәтиже көрсетті. Рекуррентік баламалар ішіндегі ең жоғары өсім BiLSTM нұсқасында байқалды: Accuracy +0,553, Macro-F1 +0,536 және MCC +0,735 пайыздық пункт. KazBERT өзі де екіжақты контексті модельдейтінін ескергенде, бұл айырмашылықты тек «екі бағыттылықпен» түсіндіру жеткіліксіз; нәтиже осы деректер мен конфигурация жағдайындағы қосымша тізбектік өңдеудің эмпирикалық үлесі ретінде түсіндіріледі.

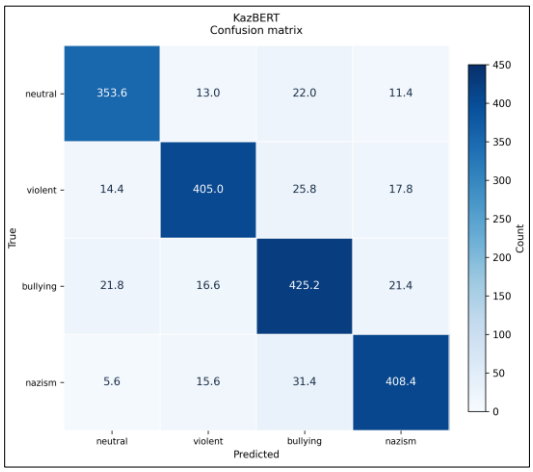
BiLSTM нұсқасының артықшылығы 4.4-кестедегі өсіммен шектеліп сипатталады. Оның нақты қандай тілдік құбылыстарда пайдалы болғанын қателерді талдау арқылы бағалау 4.4-бөлімде жалғасады. Компоненттік салыстыруда ең үлкен кезеңдік өсім онтологияға негізделген семантикалық арнадан, ал рекуррентік баламалар ішіндегі ең жоғары қосымша өсім BiLSTM нұсқасынан алынды. Осы нәтижелер негізінде кейінгі талдау үшін KazBERT_MFA_онтология_граф_BiLSTM финалдық конфигурация ретінде таңдалды [54].

4.4 Қателерді талдау (Error Analysis)

Жинақталған метрикалар модельдің жалпы классификациялау сапасын бағалағанымен, қателердің нақты қай кластар арасында пайда болатынын толық ашпайды. Осы айырмашылықтарды анықтау үшін модель нәтижелері қателесу матрицалары (confusion matrix) арқылы қосымша талданды. Талдау базалық KazBERT-тен бастап, морфологиялық және онтологияға негізделген семантикалық белгілер қосылған модель нұсқаларын қамтыды. Соңғы конфигурация ретінде KazBERT_MFA_онтология_граф_BiLSTM моделінің

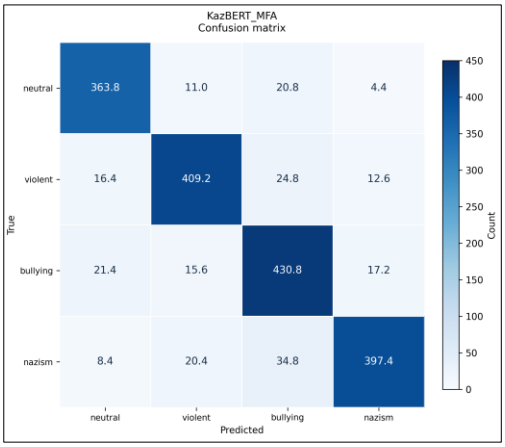
қателері қарастырылды. Қателесу матрицалары әр модельде қандай кластардың жиірек өзара шатасатынын салыстыруға қолданылды.

Сурет 4.1-де базалық KazBERT моделінің қателесу матрицасы берілген. Дұрыс классификацияланған мәтіндердің орташа саны бейтарап класс бойынша 353.6, зорлық-зомбылық экстремизмі класы бойынша 405.0, буллинг класы бойынша 425.2 және нацизм класы бойынша 408.4 болды. Қателердің елеулі бөлігі деструктивті кластар арасында байқалды. Мысалы, нацизм класына жататын мәтіндердің буллинг ретінде қате классификациялануының орташа саны 31.4-ті, ал буллинг мәтіндерінің нацизм ретінде қате классификациялануының орташа саны 21.4-ті құрады. Бұл нәтижелер базалық KazBERT моделі контекстік ақпаратты пайдаланғанымен, мағыналық жағынан жақын деструктивті кластарды ажыратуда қателердің сақталатынын көрсетеді.



Сурет 4.1 - KazBERT моделінің қателесу матрицасы

Сурет 4.2-де морфологиялық белгілер қосылған KazBERT_MFA моделінің қателесу матрицасы берілген. Дұрыс классификацияланған мәтіндердің орташа саны бейтарап класс бойынша 363.8, зорлық-зомбылық экстремизмі класы бойынша 409.2, буллинг класы бойынша 430.8 және нацизм класы бойынша 397.4 болды. Базалық KazBERT-пен салыстырғанда алғашқы үш класс бойынша дұрыс анықталған мәтіндердің орташа саны артты, ал нацизм класы бойынша бұл көрсеткіш 408.4-тен 397.4-ке төмендеді. Бейтарап мәтіндерді деструктивті кластарға қате жатқызудың орташа саны 46.4-тен 36.2-ге азайды. Соған қарамастан нацизм мен буллинг кластары арасындағы шатасу сақталды: нацизм мәтіндерінің буллинг ретінде қате классификациялануының орташа саны 34.8-ді, ал буллинг мәтіндерінің нацизм ретінде қате классификациялануының орташа саны 17.2-ні құрады. Бұл нәтижелер морфологиялық белгілер қосылған конфигурацияда кластарды анықтау сапасының біркелкі өзгермегенін көрсетеді. Нацизм мен буллинг арасындағы шатасудың ықтимал себептерінің бірі – екі класта да қорлау, кемсіту және өшпенділік мәніндегі тілдік бірліктердің кездесуі.

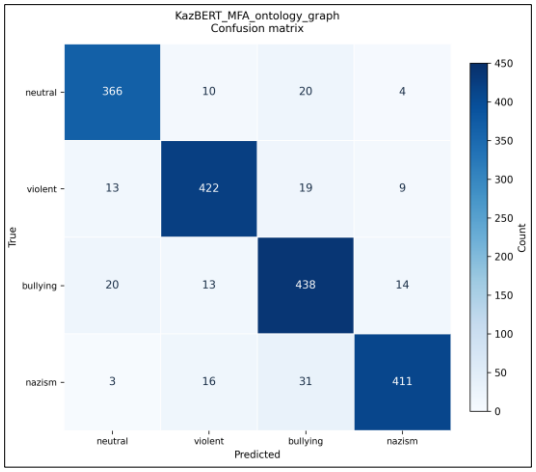


Сурет 4.2 - KazBERT_MFA моделінің қателесу матрицасы

Сурет 4.3-те KazBERT_MFA_онтология_граф моделінің қателесу матрицасы көрсетілген. Дұрыс классификацияланған мәтіндердің орташа саны бейтарап класс бойынша 366, зорлық-зомбылық экстремизмі класы бойынша 422, буллинг класы бойынша 438 және нацизм класы бойынша 411 болды. KazBERT_MFA моделімен салыстырғанда барлық төрт класс бойынша дұрыс анықталған мәтіндердің орташа саны артты. Бұл нәтижелер онтологияға негізделген семантикалық белгілер қосылған конфигурацияда классификация-лау сапасының жақсарғанын көрсетеді.

Мағыналық жағынан жақын деструктивті кластар арасында қателер сақталды. Нацизм мәтіндерінің буллинг ретінде қате классификациялануының орташа саны 31, ал зорлық-зомбылық экстремизмі мәтіндерінің буллинг ретінде қате классификациялануының орташа саны 19 болды. Мұндай шатасулардың ықтимал себептерінің бірі – аталған кластарға жататын мәтіндерде қорлау, кемсіту, өшпенділік және агрессия мәніндегі тілдік бірліктердің кездесуі.

KazBERT_MFA_онтология_граф моделінің Accuracy көрсеткіші 90,492% болып, базалық KazBERT моделінің 88,015% және KazBERT_MFA моделінің 88,513% нәтижелерінен жоғары болды. Дегенмен қателесу матрицасы мағыналық жағынан жақын кластарды ажыратуда қиындықтардың сақталғанын көрсетті. Келесі модельдік конфигурацияда токендер тізбегіндегі тәуелділіктерді екі бағытта қосымша өңдеу үшін BiLSTM қолданылды.

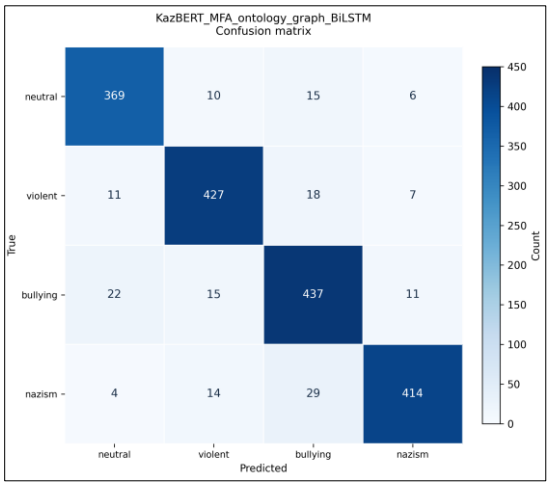


Сурет 4.3 - KazBERT_MFA_онтология_граф моделінің қателесу матрицасы

Сурет 4.4-те толық гибридті модельдің тест жиынындағы қателесу матрицасы берілген. Матрица алынған нәтижелердің орташа мәндері негізінде қалыптастырылған. Қатарлар нақты кластарды, ал бағандар модель болжаған кластарды көрсетеді.

Қателесу матрицасының диагональ элементтері бойынша бейтарап класста 369 мәтін, зорлық-зомбылық экстремизмі классында 427 мәтін, буллинг классында 437 мәтін және нацизм классында 414 мәтін дұрыс классификацияланған. Бұл көрсеткіштер толық гибридті модельдің төрт класс бойынша дұрыс анықтаған мәтіндерінің санын көрсетеді.

Қате классификациялау жағдайларының ішінде нацизм мен буллинг кластары арасындағы шатасу сақталған. Нацизм класына жататын 29 мәтін буллинг ретінде, ал буллинг класына жататын 22 мәтін бейтарап ретінде қате классификацияланған. Мұндай шатасулар мәтіндерде қорлау, кемсіту, өшпенділік және агрессия мәнін білдіретін тілдік бірліктердің әртүрлі кластарда кездесуіне байланысты болуы мүмкін. Соған қарамастан, алдыңғы модельдік конфигурациялармен салыстырғанда қателер құрылымындағы өзгеріс BiLSTM арқылы екі бағытты тізбектік ақпаратты өңдеудің классификациялау нәтижесіне қосымша үлес қосқанын көрсетеді.



Сурет 4.4 - KazBERT_MFA_онтология_граф_BiLSTM қателесу матрицасы

BiLSTM қолданылған нұсқада кейбір қате бағыттарының азаюы жалпы метрикалардағы өсіммен сәйкес келеді. Дегенмен бұл өзгерісті тек архитектураның екі бағытты болуымен байланыстыру жеткіліксіз. Қателесу матрицаларын салыстыру қателердің негізінен семантикалық жағынан жақын деструктивті кластар арасында сақталғанын көрсетті. Модельге қосымша компоненттер енгізілген сайын жекелеген шатасулар азайғанымен, олар толық жойылған жоқ. Сақталған қателердің мазмұнын талдау бірнеше ықтимал факторды бөліп көрсетуге мүмкіндік береді. Біріншісі – кластардың мағыналық жақындығы. Бұл жағдай әсіресе буллинг пен нацизм арасында байқалады. Екі класта да қорлау, кемсіту, өшпенділік, жеке адамды немесе белгілі бір топты төмендетуге бағытталған тілдік бірліктер кездесуі мүмкін. Егер мәтінде ұлттық,

этникалық немесе идеологиялық нысана айқын берілмесе, оның нацизм мен буллинг арасындағы шекарасын ажырату қиындайды.

Екінші фактор деструктивті мағынаның жасырын берілуімен байланысты. Әлеуметтік желілердегі агрессиялық немесе кемсітушілік мазмұн әрдайым тікелей тілдік формада көрінбейді. Ол астарлы мағына, ирония, қысқа тіркес немесе белгілі бір жағдаятқа тәуелді ишара арқылы берілуі мүмкін. Мұндай жағдайда жеке мәтін үзіндісі қажетті контексті толық қамтымаса, деструктивті мағынаны дұрыс анықтау күрделене түседі.

Үшінші фактор – бейтарап мәтіндерде эмоциялық реңктің болуы. Кейбір пікірлерде өткір сын, теріс бағалау немесе әлеуметтік тақырыптарға қатысты эмоционалды тұжырымдар кездеседі, бірақ олардың барлығы деструктивті веб-контентке жатпайды. Сондықтан айқын деструктивті белгісі жоқ, алайда эмоциялық жүктемесі жоғары мәтіндер кей жағдайда деструктивті кластардың біріне қате жатқызылуы мүмкін.

Төртінші фактор интернеттегі қазақ тіліндегі мәтіндердің бейресми сипатымен байланысты. Мұндай мәтіндерде орфографиялық қателер, қысқартулар, аралас тіл, сленг, транслитерация және тыныс белгілерінің толық сақталмауы кездеседі. Бұл ерекшеліктер токенизацияны, морфологиялық белгілерді анықтауды және мәтіндегі терминдерді онтологиялық лексиконмен сәйкестендіруді қиындатуы мүмкін. Соның салдарынан стандартты жазба тілден ауытқыған веб-мәтіндер модель үшін күрделірек жағдай туғызады.

Қателерді талдау модель үшін күрделі болып қалған бірнеше жағдайды айқындады: мағыналық жағынан жақын кластар, астарлы және ирониялық мазмұн, контекстің жеткіліксіздігі, эмоциялық реңкі басым бейтарап мәтіндер және бейресми веб-тіл. Морфологиялық белгілерді, онтологияға негізделген семантикалық белгілерді және тізбектік ақпаратты пайдалану кейбір қате бағыттарының азаюымен қатар жүргенімен, аталған мәселелер толық шешілген жоқ.

4.5 Ұсынылған модель нәтижелерінің статистикалық маңыздылығын бағалау

Модельдерді тек орташа көрсеткіштерімен салыстыру жеткіліксіз болуы мүмкін, себебі терең оқыту нәтижесіне бастапқы салмақтардың инициализациясы, деректерді араластыру тәртібі және оқыту барысындағы кездейсоқ факторлар әсер етеді [102]. Осы себепті ұсынылған модельдің тұрақтылығы да бағаланды [103]. Әр модель бес random seed мәні бойынша іске қосылып, негізгі метрикалардың орташа мәні мен стандарттық ауытқуы есептелді [104].

Модельдің белгілі бір метрика бойынша орташа мәні (4.6) формуласы арқылы анықталады:

$$\bar{M} = \frac{1}{n}\sum_{i=1}^{n} M_i \qquad (4.6)$$

Мұндағы,

$\bar{M}$ – метриканың орташа мәні,

M_i – i -seed бойынша алынған нәтиже,

n – іске қосу саны. Бұл зерттеуде $n = 5$.

Стандарттық ауытқу модель нәтижесінің seed-тер бойынша қаншалықты тұрақты екенін көрсету үшін қолданылды [105], есептелуі (4.7) формулада көрсетілген:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M})^2} \quad (4.7)$$

Мұндағы,

s – стандарттық ауытқу.

Стандарттық ауытқу неғұрлым кіші болса, модель нәтижесі соғұрлым тұрақты деп бағаланады [102].

Кесте 4.5-те 5 түрлі seed бойынша алынған модельдердің орташа нәтижелері мен стандарттық ауытқулары пайызбен берілген. Ең жоғары көрсеткіштерге KazBERT_MFA_онтология_граф_BiLSTM моделі қол жеткізді. Бұл модель Accuracy бойынша 91,045 ± 0,656%, Macro F1 бойынша 91,103 ± 0,654% және MCC бойынша 88,054 ± 0,875% нәтиже көрсетті. Стандарттық ауытқу мәндерінің салыстырмалы түрде төмен болуы, модель нәтижелерінің әртүрлі seed қолданғанда тұрақты екенін көрсетеді [95].

Кесте 4.5 - Бес seed бойынша модель нәтижелерінің орташа мәні және стандарттық ауытқуы

Модель	**Accuracy, %**	**Macro F1, %**	**MCC, %**
KazBERT	88.015 ± 0.486	88.086 ± 0.501	84.005 ± 0.644
KazBERT_MFA	88.513 ± 0.286	88.601 ± 0.283	84.702 ± 0.366
KazBERT_MFA_онтология_граф	90.492 ± 0.702	90.567 ± 0.699	87.319 ± 0.935
KazBERT_MFA_онтология_граф_BiLSTM	91.045 ± 0.656	91.103 ± 0.654	88.054 ± 0.875

Базалық және нәтижелік модельдердің бірдей тест модельдеріндегі жұпталған дұрыс және қате шешімдерін салыстыру үшін McNemar тесті қолданылды [106-107]. Бұл тест 5 seed бойынша алынған орташа метрикаларды емес, бір сынақ жиынындағы бірдей мәтіндер үшін екі классификатордың өзара сәйкес келмейтін шешімдерін бағалауға арналған.

Осы мақсатта әрбір мәтін бойынша базалық KazBERT және ұсынылған KazBERT_MFA_онтология_граф_BiLSTM модельдерінің болжамдары жұптық түрде салыстырылды.

McNemar тестінің қолданылуы үшін екі модельдің болжамдары дәл бір тест жиынында жұпталған түрде алынуы шарт.

Нөлдік гипотеза (4.8) формула арқылы қойылды:

$$H_0: P_{01} = P_{10} \quad (4.8)$$

Мұндағы,

P_{01} – базалық модель дұрыс, ұсынылған модель қате болжаған жағдайлардың ықтималдығы,

P_{10} – базалық модель қате, ұсынылған модель дұрыс болжаған жағдайлардың ықтималдығы. Егер бұл екі ықтималдық арасында статистикалық тұрғыдан маңызды айырмашылық болмаса, онда модельдер арасындағы айырмашылық кездейсоқ болуы мүмкін деп есептеледі.

McNemar статистикасы (4.9) формула арқылы есептеледі:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}} \quad (4.9)$$

Мұндағы,

n_{01} – базалық KazBERT моделі дұрыс, ал ұсынылған модель қате анықтаған мәтіндер саны,

n_{10} – базалық KazBERT моделі қате, ал ұсынылған модель дұрыс анықтаған мәтіндер саны.

KazBERT және ұсынылған KazBERT_MFA_онтология_граф_BiLSTM модельдерінің жұпталған болжау нәтижелері бойынша:

$$n_{01} = 22$$
$$n_{10} = 68$$

Осы мәндерді формулаға қойсақ:

$$\chi^2 = \frac{(|22 - 68| - 1)^2}{22 + 68} = \frac{45^2}{90} = 22.50$$

Алынған McNemar статистикасы бойынша:

$$p = 2.10 \times 10^{-6}$$

Маңыздылық деңгейі ретінде $a = 0.05$ алынды. Себебі $p < 0.05$, нөлдік гипотеза қабылданбайды [95].

Алынған McNemar статистикасы қабылданған маңыздылық деңгейінде нөлдік гипотезаны теріске шығаруға негіз болды. Демек, осы жұпталған тест жиынында базалық KazBERT пен KazBERT_MFA_онтология_граф_BiLSTM модельдерінің қате/дұрыс шешімдерінің айырмашылығы статистикалық маңызды болып бағаланды.

Кесте 4.6-да KazBERT пен ұсынылған модель нәтижелері mcNemar тесті бойынша салыстырылған.

Кесте 4.6 - KazBERT пен ұсынылған модель нәтижелері mcNemar тесті бойынша салыстыру

Салыстырылатын модельдер	n_{01}	n_{10}	χ^2	$p - value$	a	**Қорытынды**
KazBERT және KazBERT_MFA_онтология_граф_BiLSTM	22	68	22.50	2.10×10^{-6}	0.05	Статистикалық маңызды

Бес іске қосу бойынша орташа мәндер мен стандарттық ауытқулар модель нәтижелерінің random seed-ке сезімталдығын сипаттайды. McNemar тесті бірдей тест жиынындағы жұпталған шешімдер айырмашылығының статистикалық маңызды екенін көрсетті. Бұл қорытынды модельдің абсолюттік тиімділігіне емес, осы эксперименттік салыстырудағы айырмашылыққа қатысты.

4.6 Ұсынылған модельдің тиімділігін талқылау және оның практикалық қолданысы

Эксперименттік салыстыруда KazBERT_MFA_онтология_граф_BiLSTM нұсқасы базалық KazBERT-тен жоғары көрсеткіштер берді. Базалық модельде Accuracy 88,015% және Macro-F1 88,086% болса, финалдық конфигурацияда бұл мәндер тиісінше 91,045% және 91,103% болды. Абсолюттік өсім Accuracy бойынша 3,030, Macro-F1 бойынша 3,017 пайыздық пунктті құрады. Бұл өзгеріс 4.3-бөлімдегі компоненттік талдаумен және 4.5-бөлімдегі статистикалық тексерумен бірге қарастырылады.

Accuracy бойынша қате үлесінің салыстырмалы азаюын қосымша сипаттама ретінде есептеуге болады, ол (4.10) формула көмегімен анықталады:

$$RER = \frac{E_{base} - E_{final}}{E_{base}} \times 100\% \quad (4.10)$$

Мұндағы,
E_{base} – базалық модельдің қате үлесі,
E_{final} – ұсынылған модельдің қате үлесі,
RER – қате үлесінің салыстырмалы азаюы.

$$E_{base} = 1 - 0.88015 = 0.11985$$

$$E_{final} = 1 - 0.91045 = 0.08955$$

$$RER = \frac{0.11985 - 0.08955}{0.11985} \times 100\% \approx 25{,}28\%$$

Accuracy метрикасы бойынша жаңа ұсынылған модель базалық KazBERT-пен салыстырғанда қате классификациялау мөлшерін 25,28%-ға төмендетті.

Негізгі қорытынды абсолюттік метрикалар мен статистикалық тест нәтижелеріне сүйенеді.

BiLSTM компонентінің қосымша үлесі 4.3-бөлімде сандық түрде бағаланды; сондықтан бұл бөлімде оның архитектуралық сипаттамасы қайталанбайды. Нәтижелік модель контекстік, морфологиялық, онтологияға негізделген семантикалық және тізбектік ақпаратты бір архитектурада қатар өңдейді. Мұндай біріктіру модельге мәтіннің әртүрлі деңгейдегі сипаттарын бір классификациялау үдерісінде ескеруге жағдай жасайды. Ұсынылған модельді қолданудың негізгі бағыттары кесте 4.7-де келтірілген. Дегенмен кестеде көрсетілген бағыттар модельдің нақты қолданбалы ортадағы жұмыс тәртібін толық сипаттамайды.

Осыған байланысты зерттеу барысында базалық KazBERT пен ұсынылған KazBERT_MFA_онтология_граф_BiLSTM модельдерінің бірдей мәтіндер бойынша нәтижелерін салыстыруға арналған веб-интерфейс прототипі әзірленді [54].

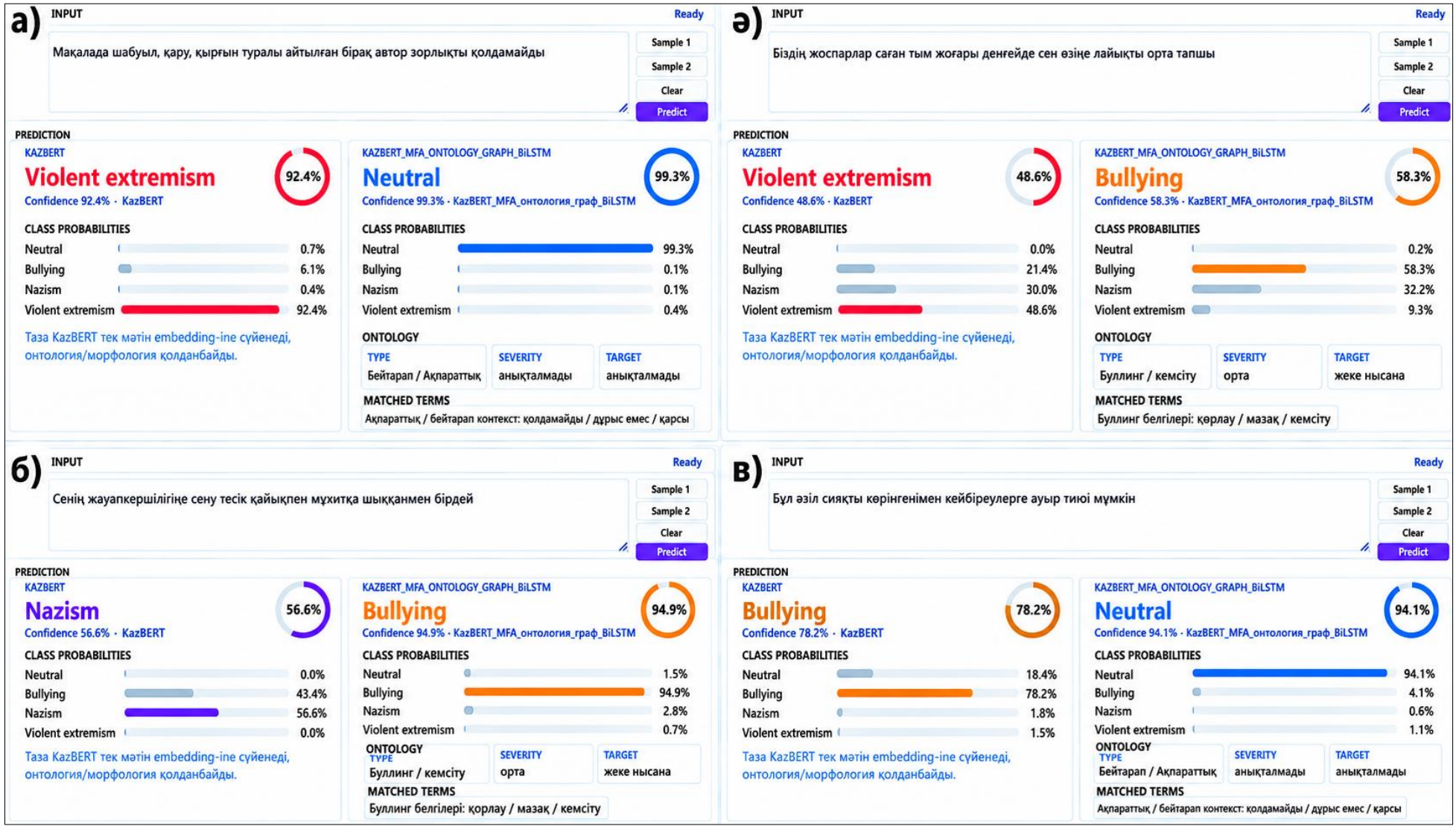
Кесте 4.7 - Ұсынылған модельді практикалық қолдану бағыттары

Қолдану бағыты	**Практикалық мәні**
Әлеуметтік желі мониторингі	Қазақ тіліндегі пікірлерден деструктивті веб-контентті автоматты түрде анықтау
Онлайн-модерация	Күмәнді мәтіндерді алдын ала белгілеп, модераторға ұсыну
Кибербуллингті анықтау	Қорлау, кемсіту және қысым көрсету белгілері бар мәтіндерді табу
Ақпараттық қауіпсіздік	Зорлық-зомбылық экстремизмі және нацистік мазмұнға қатысты мәтіндерді бақылау
Аналитикалық жүйелер	Деструктивті контенттің түрін, таралу бағытын және жиілігін талдау

Веб-интерфейс прототипі бірдей кіріс мәтін үшін базалық KazBERT пен финалдық гибридті модельдің болжамдарын, сенімділік мәндерін және класс ықтималдықтарын қатар көрсетуге арналған. Прототип сандық тестті алмастырмайды; ол модель шешімдерін нақты мысалдарда сапалық тұрғыдан салыстыруға мүмкіндік береді. Прототиптегі мысалдар модельдердің шешім айырмашылықтарын түсіндіретін сапалық иллюстрация ретінде қарастырылады. Олар тест жиынындағы жалпы метрикаларға қосымша түсіндірме береді [54].

Практикалық енгізуде модельді толық автоматты шешімнен гөрі сарапшыға көмекші шешімді қолдау құралы ретінде қолдану орынды: жүйе күмәнді мәтіндерді алдын ала белгілеп, модераторға басымдықпен ұсына алады [37]. Төмендегі мысалдарда екі модель бірдей қазақ тіліндегі мәтіндерге қолданылып, болжамдары эталондық класс белгісімен салыстырылды.

KazBERT_MFA_онтология_граф_BiLSTM моделінің веб-интерфейс арқылы салыстырмалы тестілеу нәтижелері сурет 4.5-те көрсетілген. Тексеру кезінде екі модель де қазақ тіліндегі бірдей мысалдар бойынша бағаланды.



Сурет 4.5 - Базалық KazBERT және ұсынылған модель нәтижелерін веб-интерфейс арқылы салыстырмалы тексеру

Бірінші мысалда мәтінде “шабуыл”, “қару”, “қырғын” сияқты теріс эмоционалдық реңі бар сөздер кездескенімен, жалпы мағынасы зорлықты қолдауға емес, оған қарсы болуға бағытталған. Базалық KazBERT бұл мәтінді “Violent extremism” класына жатқызған болса, ұсынылған модель оны дұрыс түрде “Neutral” категориясына бөлді.

Екінші мысалда жеке тұлғаны кемсітуді және әлеуметтік тұрғыдан төмендетуді сипаттайтын сөйлем берілген. Базалық KazBERT бұл мәтінді “Violent extremism” категориясына қате жіктеді, ал жаңартылған модель сөйлемнің қорлаушылық және кемсіту сипатына назар аудара отырып, оны “Bullying” класына жатқызды.

Үшінші мысалда адамның жауапкершілігін келемеждеу арқылы мазақтау элементі бар сөйлем ұсынылды. Негізгі KazBERT бұл мәтінді “Nazism” класына жатқызған болса, ал ұсынылған модель ұлттық немесе идеологиялық үстемдік белгілерінің жоқтығын ескеріп, оны “Bullying” класы ретінде дұрыс жіктеді.

Төртінші мысалда әзілдің кейбір адамдарға ауыр тиюі мүмкін екенін білдіретін сөйлем қарастырылды. Базалық KazBERT бұл мәтінді «Bullying» класына қате жатқызса, ұсынылған модель сөйлемнің бейтарап мазмұнын ескеріп, оны «Neutral» класына дұрыс жіктеді.

Келтірілген мысалдар екі модельдің нақты мәтіндердегі болжамдарының айырмашылығын көрсетуге мүмкіндік береді. Кейбір жағдайларда базалық KazBERT мәтіндегі жекелеген деструктивті тілдік бірліктерге сүйеніп, оның жалпы мағыналық бағытын дұрыс ажырата алмауы мүмкін. Ұсынылған модельде контекстік көріністер морфологиялық және онтологияға негізделген семантикалық белгілермен толықтырылғандықтан, терістеу, контекст және мәтіннің мағыналық бағыты туралы қосымша ақпарат ескеріледі. Әзірленген веб-прототип осы модельдердің нәтижелерін бірдей мәтіндер бойынша салыстыруға және әлеуметтік желілердегі пікірлерді, форум жазбаларын немесе басқа ашық веб-контентті алдын ала талдау сценарийін көрсетуге арналған.

Ұсынылған модельді толық автоматты шешім ретінде емес, сарапшының немесе модератордың жұмысын қолдайтын құрал ретінде қолданған жөн. Жүйе күмәнді мәтіндерді алдын ала белгілеп, оларды кейінгі тексеруге маманға жібере алады. Осылайша сарапшы барлық жарияланымды бірдей көлемде қарамай, ықтимал деструктивті материалдарға назар аударады.

Қателерді талдау мағыналық жағынан жақын кластар, жасырын агрессия, қысқа контекст және бейресми веб-тіл модель үшін күрделі жағдайлар болып қалатынын көрсетті. Әлеуметтік желілердегі мәтіндерде орфографиялық қателер, қысқартулар, аралас тілдік бірліктер, сленг, ирония және астарлы мағына кездеседі. Мұндай ерекшеліктер токенизацияны, морфологиялық белгілерді анықтауды және мәтіндегі терминдерді онтологиялық лексиконмен сәйкестендіруді қиындатуы мүмкін.

Эксперименттік нәтижелер ұсынылған модельдің базалық KazBERT-пен салыстырғанда қазақ тіліндегі деструктивті веб-контентті классификациялау көрсеткіштерін жақсартқанын көрсетті. Контекстік көріністерді, морфологиялық белгілерді және онтологияға негізделген семантикалық белгілерді бір архитектурада қолдану қазақ тіліндегі веб-мәтіндердің грамматикалық және

мағыналық ерекшеліктерін қатар ескеруге мүмкіндік береді. Алынған нәтижелер ұсынылған тәсілді қазақ тіліндегі веб-контентті мониторингтеу және модерациялау жүйелерінің құрамына енгізіп, деструктивті мазмұнды автоматты талдауда қолдануға болатынын көрсетеді.

4.7 4-тарау бойынша қорытынды

Төртінші тарауда ұсынылған әдістер эксперименттік жағдайда тексеріліп, олардың классификация нәтижелеріне ықпалы салыстырылды. Барлық эксперимент бір деректер жиынында, бірыңғай бағалау хаттамасымен және 5 түрлі seed мәні бойынша жүргізілді. Бұл модельдерді бірдей жағдайда салыстыруға және әр іске қосудағы көрсеткіштердің өзгерісін бөлек бағалауға негіз болды.

Базалық KazBERT моделіне морфологиялық және онтологияға негізделген семантикалық белгілер қосылған конфигурацияларда классификация көрсеткіштері біртіндеп жоғарылады. Ең жоғары нәтижелер KazBERT_MFA_онтология_граф_BiLSTM моделінде тіркелді. Бұл конфигурация Accuracy, Macro F1, Precision, Recall, ROC-AUC, PR-AUC және MCC метрикалары бойынша қарастырылған модельдер арасында ең жоғары мәндерді көрсетті.

Қателесу матрицаларын салыстыру модель құрамына қосымша компоненттер енгізілген сайын қате классификация санының азайғанын көрсетті. Соған қарамастан мағыналық жағынан жақын кластарды ажырату, жасырын деструктивті мағына, қысқа контекст және бейресми веб-тіл модель үшін күрделі жағдайлар болып қалды. Статистикалық маңыздылықты бағалау нәтиелік модель мен базалық KazBERT арасындағы айырмашылықтың статистикалық мәнін тексеруге қолданылды.

Практикалық тексеруде ұсынылған модель веб-интерфейс арқылы базалық KazBERT-пен салыстырылды. Қарастырылған мысалдардың бір бөлігінде базалық модель деструктивті терминдердің болуына көбірек сүйеніп қате классификация жасаған кезде, ұсынылған модель мәтіннің жалпы контексін ескеріп дұрыс класс таңдады. Мұндай нәтижелер модельді қазақ тіліндегі әлеуметтік желі пікірлерін, форум жазбаларын және ашық веб-контентті алдын ала сүзгілеу міндеттерінде қолдану мүмкіндігін көрсетеді.

Жалпы алғанда, контекстік, морфологиялық, онтологияға негізделген семантикалық және тізбектік ақпаратты біріктіретін финалдық конфигурация берілген тест жағдайында базалық KazBERT моделіне қарағанда жоғары классификация көрсеткіштеріне ие болды.

ҚОРЫТЫНДЫ

Диссертациялық жұмыста қазақ тіліндегі деструктивті веб-контентті автоматты анықтау және көпкластық классификациялау мәселесі зерттелді. Негізгі мақсат – қазақ тіліндегі деструктивті веб-контентті классификациялау үшін тереңдетіп оқыту модельдерін кешенді зерттеу және құру.

Теориялық талдау қазақ тілінің агглютинативті құрылымы, күрделі морфологиясы, әлеуметтік желілердегі бейресми жазу, қысқартулар, аралас тіл және астарлы мағына мәтінді тек лексикалық белгілер арқылы сипаттаудың жеткіліксіз екенін көрсетті. Осы ерекшеліктер контекстік, морфологиялық және семантикалық ақпаратты қатар ескеретін тәсілді қолдануға негіз болды.

Зерттеу үшін әлеуметтік желілер мен ашық интернет-платформалардан жиналған қазақ тіліндегі мәтіндерден корпус қалыптастырылып, олар бейтарап контент, зорлық-зомбылық экстремизмі, буллинг және нацизм кластары бойынша аннотацияланды. Корпус модельдерді оқыту, валидациялау, тестілеу және салыстыру кезінде пайдаланылды.

Әртүрлі мәтіндік ұсыну тәсілдері мен классификациялау модельдерін салыстыру нәтижесінде KazBERT базалық модель ретінде таңдалды. Оның контекстік көріністерін қазақ тілінің грамматикалық ерекшеліктерімен толықтыру үшін сөз табы, шақ, жақ және рай белгілерін қамтитын MFA модулі әзірленді.

Деструктивті веб-контенттің семантикалық сипаттарын беру үшін түрі, ашықтығы және нысанасы өлшемдеріне негізделген онтологиялық модель құрылды. Оның құрамына 5000 терминнен тұратын лексикон және 47 түйінді онтологиялық граф енгізілді. Бұл құрылым мәтіндегі терминдерді тиісті ұғымдармен сәйкестендіріп, онтологияға негізделген семантикалық белгілерді қалыптастыруға қолданылды.

Нәтижелік гибридті архитектура KazBERT, MFA, онтологияға негізделген семантикалық белгілер және BiLSTM компоненттерін біріктіреді. Осы құрылым мәтіннің контекстік, грамматикалық, семантикалық және тізбектік ерекшеліктерін бір классификациялау моделінде қатар ескеруге бағытталған.

Онтологияға негізделген семантикалық белгілер мәтіндегі деструктивті мағынаны категориялық деңгейде нақтылады. BiLSTM қабаты сөздердің сөйлем ішіндегі тізбектік байланысын екі бағытта өңдеп, бірнеше сөзден тұратын тіркестер арқылы берілетін мағыналарды жақсырақ тануға мүмкіндік берді. Эксперименттік зерттеу нәтижелері ұсынылған модельдің тиімді екенін көрсетті. Базалық KazBERT моделінде Accuracy көрсеткіші 88,015%, Macro F1 көрсеткіші 88,086% болды. Морфологиялық белгілер қосылған KazBERT_MFA моделінде бұл мәндер сәйкесінше 88,513% және 88,601% деңгейіне дейін артты. Онтологияға негізделген семантикалық белгілер енгізілген KazBERT_MFA_онтология_граф моделінде Accuracy 90,492%, Macro F1 90,567% болды. Ең жоғары нәтиже KazBERT_MFA_онтология_граф_BiLSTM моделінде алынды: Accuracy – 91,045%, Macro F1 – 91,103%, Precision – 91,153%, Recall – 91,096%, ROC-AUC – 98,750%, PR-AUC – 96,972% және MCC – 88,054%.

Базалық KazBERT моделімен салыстырғанда ұсынылған толық модель Accuracy көрсеткішін 88,015%-дан 91,045%-ға дейін арттырды. Бұл абсолюттік өсім 3,030 пайыздық пунктті, ал салыстырмалы өсім шамамен 3,442% құрайды. Macro F1 көрсеткіші 88,086%-дан 91,103%-ға дейін өсіп, 3,017 пайыздық пунктке жақсарды. ROC-AUC мәні 97,489%-дан 98,750%-ға дейін көтеріліп, 1,261 пайыздық пунктке артты. Сонымен қатар Accuracy бойынша қате үлесі 11,985%-дан 8,955%-ға дейін төмендеді, яғни қате үлесі шамамен 25,28% азайды. 1000 мәтінге шаққанда базалық KazBERT орта есеппен 880 мәтінді, ал ұсынылған модель 910 мәтінді дұрыс анықтайды, яғни дұрыс классификацияланған мәтін саны шамамен 30 мәтінге артты.

Қателерді талдау модель компоненттері кезең-кезеңімен қосылған сайын қате классификация санының азайғанын көрсетті. Дегенмен мағыналық жағынан жақын кластар, жасырын агрессия, қысқа контекст және бейресми веб-тіл модель үшін күрделі жағдайлар қатарында қалды. Статистикалық маңыздылықты бағалау базалық KazBERT пен ұсынылған модель нәтижелерінің арасындағы айырмашылықтың статистикалық мәнді екенін көрсетті. Статистикалық бағалау модельдер арасында байқалған айырмашылықты кездейсоқ ауытқумен ғана түсіндіру жеткіліксіз екенін көрсетті.

Практикалық тексеру кезінде ұсынылған модель мен базалық KazBERT бірдей қазақ тіліндегі мәтіндер бойынша веб-интерфейс арқылы салыстырылды. Қарастырылған мысалдардың бір бөлігінде базалық KazBERT жекелеген деструктивті тілдік бірліктерге сүйеніп қате классификация жасаған кезде, ұсынылған модель контекстік, морфологиялық және онтологияға негізделген семантикалық белгілерді бірге ескеріп, мәтінді дұрыс класқа жатқызды. Веб-прототип модельді әлеуметтік желі пікірлерін, форум жазбаларын және ашық веб-контентті алдын ала талдауда көмекші құрал ретінде тексеру үшін пайдаланылды.

Диссертациялық жұмыста қазақ тіліндегі деструктивті веб-контентті сипаттайтын морфологиялық және онтологиялық білім моделі әзірленіп, оның белгілерін KazBERT контекстік көріністерімен біріктіретін гибридті классификациялау архитектурасы ұсынылды. Эксперименттік нәтижелер контекстік көріністерді, морфологиялық белгілерді және онтологияға негізделген семантикалық белгілерді бірлесіп қолданған модельдік конфигурациялардың базалық KazBERT-пен салыстырғанда жоғары классификациялау нәтижелеріне қол жеткізгенін көрсетті.

Жұмыстың негізгі ғылыми және практикалық нәтижелері қазақ тіліндегі деструктивті веб-контентті анықтаудың теориялық және әдістемелік негіздерін талдаудан, әлеуметтік желі платформаларынан жиналған деструктивті және бейтарап мәтіндер негізінде төрт класс бойынша корпус қалыптастырудан тұрады. Қазақ тілінің ерекшеліктерін ескеретін морфологиялық белгілер модулі әзірленіп, деструктивті веб-контентті семантикалық деңгейде сипаттайтын онтологиялық лексикон мен онтологиялық графқа негізделген белгілер қалыптастырылды. KazBERT, MFA, онтологияға негізделген семантикалық белгілер және BiLSTM компоненттерін біріктіретін гибридті классификациялау архитектурасы ұсынылды. Эксперименттік тексеру нәтижелері ұсынылған

модельдің базалық KazBERT және басқа модельдермен салыстырғанда жоғары көрсеткіштерге қол жеткізгенін көрсетті. Сонымен қатар, модельдің практикалық қолдану мүмкіндігі веб-интерфейс түріндегі прототип арқылы көрсетілді.

Зерттеу нәтижелерінің қолданылу саласы – қазақ тіліндегі әлеуметтік желі пікірлерін, мессенджер арналарының ашық жазбаларын, форумдар мен басқа веб-ресурстардағы мәтіндерді алдын ала талдау және модерациялауды қолдау. Ұсынылған модель ақпараттық қауіпсіздік мониторингі жүйелерінде күмәнді хабарламаларды іріктеу, білім беру ұйымдарының цифрлық ортасында буллинг белгілерін анықтау және сарапшының қарауына жіберілетін материалдарды басымдық бойынша реттеу үшін пайдаланылуы мүмкін. Әзірленген корпус, морфологиялық модуль және онтологиялық белгілер қазақ тіліндегі мәтіндерді өңдеу жөніндегі кейінгі ғылыми зерттеулерде де қолдануға жарамды.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. Anjum, Katarya R. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities //International Journal of Information Security. – 2024. – Т. 23. – №. 1. – С. 577-608.

2. Mansur Z., Omar N., Tiun S. Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities //IEEE Access. – 2023. – Т. 11. – С. 16226-16249.

3. Mathew B. et al. Hatexplain: A benchmark dataset for explainable hate speech detection //Proceedings of the AAAI conference on artificial intelligence. – 2021. – Т. 35. – №. 17. – С. 14867-14875.

4. Founta A. et al. Large scale crowdsourcing and characterization of twitter abusive behavior //Proceedings of the international AAAI conference on web and social media. – 2018. – Т. 12. – №. 1.

5. Raut R., Spezzano F. Enhancing hate speech detection with user characteristics //International Journal of Data Science and Analytics. – 2024. – Т. 18. – №. 4. – С. 445-455.

6. Zhu J. et al. Deep prompt multi-task network for abuse language detection //International Conference on Pattern Recognition. – Cham : Springer Nature Switzerland, 2024. – С. 249-263.

7. Yang F. et al. Exploring deep multimodal fusion of text and photo for hate speech classification //Proceedings of the third workshop on abusive language online. – 2019. – С. 11-18.

8. Қазақстан Республикасының Қылмыстық кодексі. – 2014 жылғы 3 шілде. – № 226-V. – 174-бап, 274-бап. – URL: https://adilet.zan.kz/kaz/docs/K1400000226 (қаралған күні: 29.09.2026).

9. Қазақстан Республикасының «Байланыс туралы» Заңы. – 2004 жылғы 5 шілде. – № 567-II. – 41-1-бап. – URL: https://adilet.zan.kz/kaz/docs/Z040000567_ (қаралған күні: 29.09.2026).

10. Қазақстан Республикасының «Киберқауіпсіздік туралы» Заңы (бұрынғы атауы: «Ақпараттандыру туралы»). – 2015 жылғы 24 қараша. – № 418-V. – URL: https://adilet.zan.kz/kaz/docs/Z1500000418 (қаралған күні: 29.09.2026).

11. Beck M. et al. xlstm: Extended long short-term memory //Advances in Neural Information Processing Systems. – 2024. – Т. 37. – С. 107547-107603.

12. Zhang Z., Robinson D., Tepper J. Detecting hate speech on twitter using a convolution-gru based deep neural network //European semantic web conference. – Cham : Springer International Publishing, 2018. – С. 745-760.

13. Kumar A. et al. A hybrid deep BiLSTM-CNN for hate speech detection in multi-social media //ACM Transactions on Asian and Low-Resource Language Information Processing. – 2024. – Т. 23. – №. 8. – С. 1-22.

14. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). – 2019. – P. 4171–4186. – DOI: 10.18653/v1/N19-1423.

15. Mussiraliyeva Sh., Bolatbek M., Zhumakhanova A., Sagynay M., Bagitova K. Development of a Software Module for Collecting and Analyzing Web Content to Determine Extremist Content // Proceedings of the International Conference on Smart Information Systems and Technologies. – 2022. – P. 1–6.

16. Mussiraliyeva Sh., Omarov B., Yoo P., Bolatbek M. Applying Machine Learning Techniques for Religious Extremism Detection on Online User Contents // Computers, Materials & Continua. – 2022. – Vol. 70, No. 1. – P. 915–934. – DOI: 10.32604/cmc.2022.019189.

17. Bolatbek M.A., Mussiraliyeva Sh.Zh., Tukeyev U.A. Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language // Journal of Mathematics, Mechanics and Computer Science. – 2018. – No. 1 (97). – P. 134–142.

18. Шалабаев К., Әліпбай К., Болатбек М., Мусиралиева Ш. Вконтакте әлеуметтік желісіндегі экстремистік мәтіндерді анықтау және жіктеу // Вестник КазНИТУ. – 2019. – №5 (135). – Б. 80–86.

19. Ecker U. K. H. et al. The psychological drivers of misinformation belief and its resistance to correction //Nature reviews psychology. – 2022. – Т. 1. – №. 1. – С. 13-29.

20. Müller K., Schwarz C. Fanning the flames of hate: Social media and hate crime //Journal of the European Economic Association. – 2021. – Т. 19. – №. 4. – С. 2131-2167.

21. Cinelli M. et al. The echo chamber effect on social media //Proceedings of the national academy of sciences. – 2021. – Т. 118. – №. 9. – С. e2023301118.

22. Brady W. J. et al. Estimating the effect size of moral contagion in online networks: A pre-registered replication and meta-analysis //PNAS nexus. – 2025. – Т. 4. – №. 11. – С. pgaf327.

23. Urman A., Katz S. What they do in the shadows: examining the far-right networks on Telegram //Information, communication & society. – 2022. – Т. 25. – №. 7. – С. 904-923.

24. Alkathiri N., Slhoub K. Challenges in machine learning-based social bot detection: a systematic review //Discover Artificial Intelligence. – 2025. – Т. 5. – №. 1. – С. 214.

25. Subramanian M. et al. A survey on hate speech detection and sentiment analysis using machine learning and deep learning models //Alexandria Engineering Journal. – 2023. – Т. 80. – С. 110-121.

26. Bolatbek M., Mussiraliyeva S., Sagynay M. Machine Learning-Based Hate Speech Detection in the Kazakh Language // Research in Language. – 2025. – Vol. 23. – P. 311–319. – DOI: https://doi.org/10.18778/1731-7533.23.19.

27. Bolatbek M., Sagynay M., Mussiraliyeva Sh., Zhastay Y., Machine learning methods for identifying destructive texts in the Kazakh language //Proceedings of the 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology. – 2024. – С. 99-103.

28. Malik J. S. et al. Deep learning for hate speech detection: a comparative study //International Journal of Data Science and Analytics. – 2025. – Т. 20. – №. 4. – С. 3053-3068.

29. Ramos G. et al. A comprehensive review on automatic hate speech detection in the age of the transformer // Social Network Analysis and Mining. – 2024. – Vol. 14. – Article 204. – DOI: 10.1007/s13278-024-01361-3.

30. Hussain F. G. et al. Fake news detection landscape: datasets, data modalities, AI approaches, their challenges, and future perspectives //IEEE Access. – 2025. – Т. 13. – С. 54757-54778.

31. Muneer A. et al. Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT //Information. – 2023. – Т. 14. – №. 8. – С. 467.

32. Aitim, A., & Satybaldiyeva, R. A comparison of Kazakh language processing models for improving semantic search results. Eastern-Eur. J. Enterprise Technol. 1 (2 (133)), 66–75 (2025).

33. Tolegen G., Toleu A., Mussabayev R. A finite state transducer based morphological analyzer for the kazakh language //2022 7th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2022. – С. 01-06.

34. Bolatbek, M., Sagynay, M., Mussiraliyeva, S., & Yeltay, Z. (2025). Detection of offensive content in the Kazakh language using machine learning and deep learning approaches. PeerJ Computer Science, 11, e3027.

35. Baitenova L. et al. Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language //Frontiers in Artificial Intelligence. – 2025. – Т. 8. – С. 1708566.

36. Das S. et al. A survey on automatic online hate speech detection in low-resource languages //arXiv preprint arXiv:2411.19017. – 2024.

37. Arora A. et al. Detecting harmful content on online platforms: what platforms need vs. where research efforts go //ACM Computing Surveys. – 2023. – Т. 56. – №. 3. – С. 1-17.

38. Zampieri M., Malmasi S., Nakov P., Rosenthal S., Farra N., Kumar R. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) // Proceedings of the 13th International Workshop on Semantic Evaluation. – 2019. – P. 75–86. https://aclanthology.org/S19-2010/

39. Kucukkaya I. E., Toraman C. Constructing ensembles for hate speech detection //Natural Language Processing. – 2025. – Т. 31. – №. 3. – С. 745-770.

40. Fetahi E. et al. Enhancing social media hate speech detection in low-resource languages using transformers and explainable AI //Social Network Analysis and Mining. – 2025. – Т. 15. – №. 1. – С. 82.

41. ALBayari R., Abdallah S. Instagram-based benchmark dataset for cyberbullying detection in Arabic text //Data. – 2022. – Т. 7. – №. 7. – С. 83.

42. Mussiraliyeva S., Bolatbek M., Baisylbayeva K. Extremist Ideology Classification in Kazakh: A Multi-Class Approach Using Machine Learning and Psycholinguistic Analysis // IEEE Access. – 2025. – Vol. 13. – P. 140500–140518. – DOI: 10.1109/ACCESS.2025.3596601.

43. Bolatbek M., Mussiraliyeva S., Baisylbayeva K. Detection and Classification of Ideological Texts in the Kazakh Language Using Machine Learning and Transformers // Research in Language. – 2025. – Vol. 23. – P. 341–353. – DOI: 10.18778/1731-7533.23.21.

44. Sakypbekova M., Aubakirova R. Corpus and Baseline Models for Kazakh Toxic Comment Detection // Bulletin of Abai KazNPU. Series of Physical and Mathematical Sciences. — 2026. — № 3 (95). — P. 252–264. DOI: 10.51889/2959-5894.2026.95.3.022.

45. Болатбек М.А., Байсылбаева К.Д., Сағынай М., Мусиралиева Ш.Ж., Жумаханова А.Н. Интернет кеңістігіндегі жастарға бағытталған деструктивті мәтіндерді жинақтауға қажетті парсер бағдарламасын әзірлеу // Известия Национальной Академии Наук Республики Казахстан. Серия физико-математическая. – 2023. – №4. – DOI: 10.32014/2023.2518-1726.223.

46. Тоқтарова А.Б., Ажибекова Ж.Ж., Сұлтан Д.Р., Керимбеков М.А. Онлайн контенттегі қазақ тілді бейәдеп пікірлерді машиналық оқытуда жинақтау // Абай атындағы ҚазҰПУ Хабаршысы. «Физика-математика ғылымдары» сериясы. — 2023. — № 1 (81). — Б. 265–272.

47. Bolatbek, M., Sagynay, M., Mussiraliyeva, Sh., Baisylbayeva, K., & Zhastay, Y. (2024). Kazakh language dataset for hate speech detection on social media text. In 2024 9th International Conference on Computational Intelligence and Applications (ICCIA). https://doi.org/10.1109/ICCIA62557.2024.10719327

48. Mussiraliyeva, Sh., Bolatbek, M., Sagynay, M., Zhumakhanova, A., Yeltay, Z., & Medetbek, Z. (2023). Identifying cyber-threatening texts in the Kazakh segment of web resources. In Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence (pp. 68–72). https://doi.org/10.1145/3633598.3633610

49. Neelakandan S. et al. Deep learning approaches for cyberbullying detection and classification on social media //Computational intelligence and neuroscience. – 2022. – Т. 2022. – С. 2163458.

50. Sagynay, M., Bolatbek, M., Mussiraliyeva, Sh., & Baisylbayeva, K. (2025). Natural language processing and machine learning for detecting harmful speech in Kazakh social media. In 2025 11th International Conference on Computing and Artificial Intelligence (pp. 347–351).https://doi.org/10.1109/ICCAI66501.2025.00060

51. Мусиралиева, Ш.Ж., Байспай, Г.Б., Болатбек, М.А., Сағынай, М., & Терейковский, И.А. (2023). Веб-ресурстардағы экстремисттік мәліметтерді анықтауға арналған машиналық әдістерді оқыту және сынау үшін қазақ тіліндегі мәтіндер корпусын құру. Университет еңбектері, 3(92), 453–459.

52. Mussiraliyeva, Sh., Bolatbek, M., Zhumakhanova, A., Medetbek, Zh., & Sagynay, M. (2023). Comparative analysis of machine learning algorithms to identify extremist texts in the Kazakh language. Scientific Journal of Astana IT University, 14. https://doi.org/10.37943/14DKRN4681

53. Болатбек М.А., Сағынай М., Мусиралиева Ш.Ж. Қазақ тіліндегі деструктивті веб-контентті анықтау үшін машиналық оқыту әдістерін қолдану //Известия НАН РК. Серия физико-математическая. – 2024. – №. 4. – С. 99–111. https://doi.org/10.32014/2024.2518-1726.310

54. Сағынай М., Болатбек М. Қазақ тіліндегі деструктивті веб-мазмұнды жіктеудің морфологиялық және семантикалық тұрғыдан жетілдірілген моделі //Вестник КазАТК. – 2026. – Т. 144. – №. 3. – С. 151-166.

55. Sambetbayeva, M., Nekessova, A., Yerimbetova, A., Bayangali, A., Kaldarova, M., Telman, D., & Smailov, N. (2025). A multi-level annotation model for fake news detection: Implementing Kazakh-Russian corpus via Label Studio. Big Data and Cognitive Computing, 9(8), 215.

56. Lee N., Jung C., Myung J., Jin J., Camacho-Collados J., Kim J., Oh A. Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis. Proceedings of NAACL 2024, Volume 1: Long Papers. 2024. P. 4205–4224.

57. Sairanbekova, A., Nazyrova, A., Bekmanova, G., Zhetkenbay, L., Yergesh, B., & Lamasheva, Z. (2026). Automatic Analysis of Political Discourse: A Comparative Study of Multilingual and Large Language Models. Journal of Applied Data Sciences, 7(2), 1136-1148.

58. Davani A. M., Díaz M., Prabhakaran V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations //Transactions of the Association for Computational Linguistics. – 2022. – Т. 10. – С. 92-110.

59. Wan R., Kim J., Kang D. Everyone's voice matters: Quantifying annotation disagreement using demographic information //Proceedings of the AAAI Conference on Artificial Intelligence. – 2023. – Т. 37. – №. 12. – С. 14523-14530.

60. Fleiss J. L. Measuring nominal scale agreement among many raters // Psychological Bulletin. – 1971. – Vol. 76, No. 5. – P. 378–382. – DOI: 10.1037/h0031619.

61. Moss J. Measures of agreement with multiple raters: Fréchet variances and inference //psychometrika. – 2024. – Т. 89. – №. 2. – С. 517-541.

62. Moons F., Vandervieren E. Measuring agreement among several raters classifying subjects into one or more (hierarchical) categories: A generalization of Fleiss' kappa //Behavior research methods. – 2025. – Т. 57. – №. 10. – С. 287.

63. Toleu A., Tolegen G., Mussabayev R. Language-Independent Approach for Morphological Disambiguation // Proceedings of the 29th International Conference on Computational Linguistics. – 2022. – P. 5288–5297. – URL: https://aclanthology.org/2022.coling-1.470/

64. Abudouwaili G., Abiderexiti K., Yi N., Wumaier A. Joint Learning Model for Low-Resource Agglutinative Language Morphological Tagging // Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. – 2023. – P. 27–37. – URL: https://aclanthology.org/2023.sigmorphon-1.4/

65. Abudouwaili G., Abliz W., Abiderexiti K., Wumaier A., Yi N. Strategies to Improve Low-Resource Agglutinative Languages Morphological Inflection // Proceedings of the 27th Conference on Computational Natural Language Learning. – 2023. – P. 508–520. – URL: https://aclanthology.org/2023.conll-1.34/

66. Abudouwaili G., Ruzmamat S., Abiderexiti K. et al. Research on Morphological Knowledge-Guided Low-Resource Agglutinative Languages-Chinese Translation // Complex & Intelligent Systems. – 2025. – Vol. 11. – Article 168. – DOI: 10.1007/s40747-025-01780-5.

67. Klemen, M., Krsnik, L., & Robnik-Šikonja, M. (2023). Enhancing deep neural networks with morphological information. Natural Language Engineering, 29(2), 360-385.

68. Stranisci M.A., Frenda S., Lai M., Araque O., Cignarella A.T., Basile V., Bosco C., Patti V. O-Dang! The Ontology of Dangerous Speech Messages // Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data. – 2022. – P. 2–8. – URL: https://aclanthology.org/2022.salld-1.2/.

69. Chen, M., Xu, Y., Wu, N., & Pan, Y. (2025). Ontology matching method based on gated graph attention model. Computers, Materials and Continua, 82(3), 5307-5324.

70. Reyero Lobo P., Daga E., Alani H., Fernandez M. Knowledge-Grounded Target Group Language Recognition in Hate Speech // Studies on the Semantic Web. – 2023. – DOI: 10.3233/SSW230002.

71. Cercas Curry A., Abercrombie G., Rieser V. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI // Proceedings of EMNLP. – 2021. – P. 7388–7403. – DOI: 10.18653/v1/2021.emnlp-main.587.

72. Ponzi V., Napoli C. Graph Neural Networks: Architectures, Applications, and Future Directions // IEEE Access. – 2025. – Vol. 13. – P. 62870–62891. – DOI: 10.1109/ACCESS.2025.3558752.

73. Wasi A.T. Explainable Identification of Hate Speech towards Islam using Graph Neural Networks // Proceedings of the Third Workshop on NLP for Positive Impact. – 2024. – P. 250–257. – DOI: 10.18653/v1/2024.nlp4pi-1.23.

74. Mu Y. et al. HA-GCEN: Hyperedge-abundant graph convolutional enhanced network for hate speech detection //Knowledge-Based Systems. – 2024. – Т. 300. – С. 112166.

75. Zhao Y. et al. Enhancing LLM-based hatred and toxicity detection with meta-toxic knowledge graph //Findings of the Association for Computational Linguistics: ACL 2025. – 2025. – С. 24747-24760.

76. Wiemerslage A., Silfverberg M., Yang C., McCarthy A., Nicolai G., Colunga E., Kann K. Morphological Processing of Low-Resource Languages: Where We Are and What's Next // Findings of the Association for Computational Linguistics: ACL 2022. – 2022. – P. 988–1007. – DOI: 10.18653/v1/2022.findings-acl.80. https ://aclanthology.org/2022.findings-acl.80/

77. Inoue G., Khalifa S., Habash N. Morphosyntactic Tagging with Pre-trained Language Models for Arabic and its Dialects // Findings of the Association for Computational Linguistics: ACL 2022. – 2022. – P. 1708–1719. – DOI: 10.18653/v1/2022.findings-acl.135. https://aclanthology.org/2022.findings-acl.135/

78. Nzeyimana A. Low-resource neural machine translation with morphological modeling // Findings of the Association for Computational Linguistics: NAACL 2024. – 2024. – P. 182–195. – DOI: 10.18653/v1/2024.findings-naacl.13.

79. Wu Y., Wan J. A survey of text classification based on pre-trained language model //Neurocomputing. – 2025. – Т. 616. – С. 128921.

80. García-Díaz J.A. Jiménez-Zafra S.M., García-Cumbreras M.A., Valencia-García R. Evaluating feature combination strategies for hate-speech detection in Spanish using linguistic features and transformers // Complex & Intelligent Systems. – 2023. – Vol. 9. – P. 2893–2914. – DOI: 10.1007/s40747-022-00693-x.

81. Ahmed, K., Mathew, A., & Anand, S. (2025). Ontology and AI Integration for Real-Time Detection of Cyberbullying Among University Students. In Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025). Vol. 1. SCITEPRESS. P. 709–716. DOI: 10.5220/0013213500003929.

82. Hur A., Janjua N., Ahmed M. Unifying context with labeled property graph: A pipeline-based system for comprehensive text representation in NLP // Expert Systems with Applications. – 2024. – Vol. 239. – Article 122269. – DOI: 10.1016/j.eswa.2023.122269.

83. Hu L., Liu Z., Zhao Z., Hou L., Nie L., Li J. A Survey of Knowledge Enhanced Pre-Trained Language Models // IEEE Transactions on Knowledge and Data Engineering. – 2024. – Vol. 36, №4. – P. 1413–1430. – DOI: 10.1109/TKDE.2023.3310002.

84. Kapoor S., Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science // Patterns. – 2023. – Vol. 4, №9. – Article 100804. – DOI: 10.1016/j.patter.2023.100804.

85. Sharif W., Abdullah S., Iftikhar S., Al-Madani D., Mumtaz S. Enhancing Hate Speech Detection in the Digital Age: A Novel Model Fusion Approach Leveraging a Comprehensive Dataset // IEEE Access. – 2024. – Vol. 12. – P. 27225–27236. – DOI: 10.1109/ACCESS.2024.3367281.

86. Zhenisbekovna M. S., Aslanbekkyzy B. M., Bolatkyzy B. G. Investigating long short-term memory approach for extremist messages detection in Kazakh language //Expert Systems. – 2025. – Т. 42. – №. 1. – С. e13595.

87. Oralbekova D., Mamyrbayev O., Akhmediyarova A., Kassymova D., Alibiyeva Z. Using Kazakh NER Datasets for Multiclass Classification in the Legal Domain: A Comparative Study of BERT, GPT, and LSTM Models // Proceedings of ICSCA 2025. – 2025. – DOI: 10.1145/3731806.3731855.

88. Nozza D., Bianchi F., Hovy D. What the [MASK]? Making Sense of Language-Specific BERT Models // arXiv preprint. – 2020. – arXiv:2003.02912. – DOI: 10.48550/arXiv.2003.02912.

89. Nugumanova A., Baiburin Y., Alimzhanov Y. Sentiment Analysis of Reviews in Kazakh With Transfer Learning Techniques // 2022 International Conference on Smart Information Systems and Technologies (SIST). – 2022. – P. 1–6. – DOI: 10.1109/SIST54437.2022.9945811.

90. Yeshpanov R., Khassanov Y., Varol H.A. KazNERD: Kazakh Named Entity Recognition Dataset // Proceedings of the Thirteenth Language Resources and Evaluation Conference. – Marseille, France: ELRA, 2022. – P. 417–426. – URL: https://aclanthology.org/2022.lrec-1.44/

91. Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. – 2020. – P. 101–108. – DOI: 10.18653/v1/2020.acl-demos.14.

92. Nivre J., de Marneffe M.-C., Ginter F., Hajič J., Manning C.D., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection // Proceedings of the Twelfth Language Resources and Evaluation Conference. – 2020. – P. 4034–4043. – URL: https://aclanthology.org/2020.lrec-1.497/.

93. Gkarmpounis G. et al. Survey on graph neural networks //IEEE Access. – 2024. – T. 12. – C. 128816-128832.

94. Mukanova A. et al. Development of a geographical question-answering system in the Kazakh language //IEEE Access. – 2024. – T. 12. – C. 105460-105469.

95. Rainio O., Teuho J., Klén R. Evaluation metrics and statistical tests for machine learning // Scientific Reports. – 2024. – Vol. 14. – Article 6086. – DOI: https://doi.org/10.1038/s41598-024-56706-x.

96. Bichri H., Chergui A., Hain M. Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets // International Journal of Advanced Computer Science and Applications. – 2024. – Vol. 15, №2. – P. 331–339. – DOI: 10.14569/IJACSA.2024.0150235.

97. Sivakumar M., Parthasarathy S., Padmapriya T. Trade-off between training and testing ratio in machine learning for medical image processing // PeerJ Computer Science. – 2024. – Vol. 10. – Article e2245. – DOI: 10.7717/peerj-cs.2245.

98. Walsh S., Greaney P. Multiclass hate speech detection with an aggregated dataset // Natural Language Processing. – 2025. – Vol. 31, №6. – P. 1350–1366. – DOI: 10.1017/nlp.2024.62.

99. Bui N.T., Savova G.K., Wang L. Assessing the Macro and Micro Effects of Random Seeds on Fine-Tuning Large Language Models // Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. – 2025. – P. 41–46. – DOI: https://doi.org/10.18653/v1/2025.ijcnlp-short.3.

100. Lones M.A. Avoiding common machine learning pitfalls // Patterns. – 2024. – Vol. 5, №10. – Article 101046. – DOI: https://doi.org/10.1016/j.patter.2024.101046.

101. Ansel J., Yang E., He H. et al. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation // Proceedings of ASPLOS. – 2024. – DOI: https://doi.org/10.1145/3620665.3640366.

102. Pecher B., Srba I., Bielikova M. A Survey on Stability of Learning with Limited Labelled Data and its Sensitivity to the Effects of Randomness // ACM Computing Surveys. – 2024. – Vol. 57, No. 1. – Article 19. – DOI: 10.1145/3691339.

103. Pecher B., Srba I., Bielikova M. On Sensitivity of Learning with Limited Labelled Data to the Effects of Randomness: Impact of Interactions and Systematic Choices // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – 2024. – P. 522–556. – DOI: 10.18653/v1/2024.emnlp-main.32.

104. Pecher B., Cegin J., Belanec R., Simko J., Srba I., Bielikova M. Fighting Randomness with Randomness: Mitigating Optimisation Instability of Fine-Tuning using Delayed Ensemble and Noisy Interpolation // Findings of the Association for Computational Linguistics: EMNLP 2024. – 2024. – P. 11005–11044. – DOI: 10.18653/v1/2024.findings-emnlp.644

105. Lucic A., Bleeker M., Bhargav S., Forde J.Z., Sinha K., Dodge J., Luccioni S., Stojnic R. ACL Tutorial Proposal: Towards Reproducible Machine Learning Research in Natural Language Processing // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. – 2022. – P. 7–11. – DOI: 10.18653/v1/2022.acl-tutorials.2.

106. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages // Psychometrika. – 1947. – Vol. 12. – P. 153–157. – DOI: https://doi.org/10.1007/BF02295996.

107. Dietterich T.G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms // Neural Computation. – 1998. – Vol. 10, №7. – P. 1895–1923. – DOI: https://doi.org/10.1162/089976698300017197.

ҚОСЫМША А



ҚАЗАҚСТАН РЕСПУБЛИКАСЫ

РЕСПУБЛИКА КАЗАХСТАН

АВТОРЛЫҚ ҚҰҚЫҚПЕН ҚОРҒАЛАТЫН ОБЪЕКТІЛЕРГЕ ҚҰҚЫҚТАРДЫҢ МЕМЛЕКЕТТІК ТІЗІЛІМГЕ МӘЛІМЕТТЕРДІ ЕНГІЗУ ТУРАЛЫ

КУӘЛІК

2026 жылғы «27» наурыз № 69171

Автордың (лардың) жөні, аты, әкесінің аты (егер ол жеке басын куәландыратын құжатта көрсетілсе):
САҒЫНАЙ МӘЛДІР, Болатбек Милана Асланбекқызы

Авторлық құқық объектісі: компьютерлік бағдарламалар (бағдарламалық қамтылым)

Объектінің атауы: Қазақ тіліндегі деструктивті веб-контентті классификациялауға арналған BERT негізіндегі терең оқыту модельдерінің тиімділігін Supervised Contrastive Learning (SupCon) әдісі арқылы арттыру

Объектінің жасалған күні: 19.03.2026

Құжат тиісінше "Әділет" ақпараттық-құқықтық жүйесінің сайтында "Авторлық құқық" бөлімінде тексерілуі мүмкін: https://copyright.kazpatent.kz

Подлинность документа возможно проверить на сайте kazpatent.kz в разделе «Авторское право» https://copyright.kazpatent.kz

ЭЦҚ қол қойылды С. Ахметов