КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

ӘОЖ 004.85
ҒТАХА 28.23.25
https://doi.org/10.53364/24138614_2025_39_4_8

А.Б.Омар[1*], Ш.Ж.Мусиралиева[1]
[1]Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы, Қазақстан

E-mail: aiym.omar98@gmail.com[*]

АГРЕССИВТІ МАЗМҰНДЫ ЖІКТЕУ МІНДЕТТЕРІ ҮШІН ФЕДЕРАТИВТІ ОҚЫТУ: ТРАНСФОРМЕР МОДЕЛЬДЕРІ НЕГІЗІНДЕГІ ӘДІС

Аңдатпа. *Цифрлық коммуникацияның қарқынды дамуы интернетте агрессивті мазмұндағы жазбалардың көбеюіне себеп болды. Мұндай мазмұнды жазбаларды автоматты түрде анықтау қазіргі заманның өзекті мәселелерінің бірі болып отыр. Алайда деректерді орталық серверге жинауға негізделген дәстүрлі тәсілдер жеке ақпараттың құпиялығын бұзуы мүмкін. Осы мәселені шешудің бір жолы – федеративті оқыту әдісін қолдану. Бұл әдісте деректерді орталық серверге жібермей-ақ, әрбір пайдаланушы құрылғысында жеке формада модельді оқыту қарастырылады. Зерттеу жұмысын жүргізу барысында ғылыми еңбектерге әдеби шолу жасалып, федеративті оқыту әдісін қолдану тәжірибесі талданды. Деректер жиынтығы ретінде 73 572 жазбадан тұратын агрессивті және агрессивті емес мәтіндерден құралған арнайы корпус пайдаланылды. Модельді оқыту үшін DistilBERT моделі пайдаланылды және деректер жиынтығы үш клиент арасында бөлініп, әрқайсысы тек өз жазбаларын жеке оқытты. Ал сервер әр раундтың соңында FedAvg алгоритмі арқылы барлық клиенттер ұсынған модель параметрлері серверде біріктіріліп, ортақ глобалды модель құрылды. Алынған нәтижелер негізінде федеративті оқыту әдісі екі маңызды артықшылығын анықтады: ол ең алдымен, жоғары дәлдікпен жұмыс жасайды, екіншіден, ақпаратқа қатысты сенімділік пен құпиялылықты қамтамасыз етеді.*

Түйін сөздер: *федеративті оқыту, NLP, DistilBERT, FedAvg, құпиялылықты сақтау, агрессивті мазмұн, классификациялау.*

Кіріспе.

Кибербуллинг, өшпенділік тілі, кемсіту және басқа да вербалды агрессия мазмұнды мәтін әлеуметтік желілердегі пікірлерде, жарияланымдарда және жеке хабарламаларда кең таралған. Мұндай мазмұнды мәтіндер жеке тұлғаның эмоционалдық күйіне кері әсер етіп, интернеттегі қарым-қатынас мәдениетінің бұзылуына алып келуі мүмкін.

Бұл мәселені шешу үшін, яғни мәтіндердегі агрессивті мазмұнды автоматты түрде анықтау мақсатында табиғи тілді өңдеу (NLP) және машиналық оқыту әдістері қолданылады. Алайда, бұл шешімдердің көпшілігі пайдаланушы деректерін жинаудың және оларды орталық серверде сақтаудың дәстүрлі әдістерін қолданады. Бұл құпиялылықтың бұзылуына, ақпараттың ағып кетуіне және деректердің қауіпсіздігіне байланысты басқа да қауіптерге әкелуі мүмкін. Бұл әсіресе жеке ақпаратты пайдалану мен тасымалдауды қатаң шектейтін Деректерді қорғаудың жалпы ережесі (GDPR) сияқты халықаралық заңнама талаптарына сай өте маңызды.

Бұл мәселені шешу үшін федеративті оқытуды (Federated Learning, FL) пайдалану ұсынылады. Федеративті оқыту сервердегі барлық деректерді орталықтандырылған түрде

жинаудың орнына, оқыту процесін тікелей смартфондар немесе компьютерлер сияқты пайдаланушы құрылғыларында жүзеге асырылады. Содан кейін ол серверге деректердің өзін емес, тек оқыту нәтижелерін жібереді. Сервер бұл деректерді барлық құрылғылардан жинайды, оларды біріктіреді және жаһандық модельді жаңартады.

Бұл тәсіл құпия ақпаратты қорғаудың жоғары деңгейін қамтамасыз етеді, өйткені бастапқы деректер клиенттің өзінде қалады және желі арқылы берілмейді, бұл дербес деректерге рұқсатсыз қол жеткізу және ағып кету тәуекелдерін айтарлықтай төмендетеді.

Бұл зерттеудің мақсаты федеративті оқыту әдісін қолдана отырып, DistilBERT моделіне негізделген агрессивті мазмұнды анықтау жүйесін құру болып табылады. Ұсынылған тәсілде деректер бірнеше клиенттер арасында бөлінеді, олардың әрқайсысы өзінің жергілікті моделін үйретеді, содан кейін параметрлер FedAvg алгоритмі арқылы біріктіріледі. Бұл тәсіл пайдаланушы деректерінің құпиялылығын сақтай отырып, жоғары анықтау дәлдігіне мүмкіндік береді.

Материалдар мен зерттеу әдістері.

Модельді оқыту және бағалау үшін агрессивті мазмұнды құрайтын деректер жинағы қолданылды. Деректер жинағы 73 572 мәтіндік жолды құрайды, олардың әрқайсысына бинарлық белгі қойылған: 0 – агрессивті емес пікір, 1 – агрессивті пікір. Хабарламалар ретінде пайдаланушылардың ағылшын тілінде әр түрлі формада жазған қысқа мәтіндері ұсынылған: сабырлы пікірлерден бастап агрессивті және арандатушы сөздерге дейін. Деректер жинағында берілген мәтіндерінің сапасын жақсарту мақсатында, мәтіндерден арнайы таңбалар, эмодзилер мен тыныс белгілері алынып, барлық әріптер кіші әріптерге ауыстырылды. Сонымен қатар, бос жолдар және екі таңбадан қысқа мәтіндер жойылды. Орындалған алдын ала өңдеу жұмыстары бізге жеткілікті нұсқада деректер корпусын қалыптастыруға мүмкіндік берді. Бұл модельді тиімді оқыту және тестілеу үшін қажетті жағдайларды қамтамасыз етті.

Мәтіндерді нейрондық желілік архитектураға жіберуге жарамды сандық форматқа түрлендіру үшін Hugging Face Transformers кітапханасының бөлігі ретінде жүзеге асырылған алдын ала дайындалған DistilBERTTokenizer пайдаланылды. Нәтижесінде көрсетілгендей үш кіріс құрылымы қалыптасты (сурет-1):

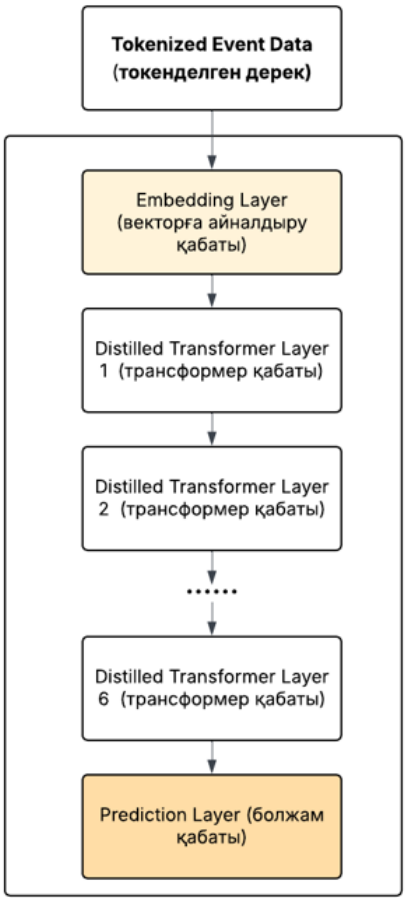
- input_ids (мәтіндердің сандарға айналдырылған нұсқасы);
- attention_mask (маңызды токендерді көрсетуге арналған маска);
- labels (берілген мәтіннің категориясы: 1 – агрессивті, 0 – агрессивті емес).

	input_ids	attention_mask	label
0	[101, 2017, 2123, 2102, 2113, 2073, 2115, 6898...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	1
1	[101, 4388, 24415, 4430, 4140, 14697, 7743, 10...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, ...	1
2	[101, 7632, 22052, 11156, 2400, 2350, 11891, 4...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ...	1
3	[101, 2363, 1998, 5838, 12362, 102, 0, 0, 0, 0...	[1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...	0
4	[101, 1996, 27178, 2869, 10373, 2134, 2102, 25...	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...	0

Сурет 1 – Токенделген үлгідегі деректер жинағы

Классификатор ретінде DistilBERT моделі қолданылды. DistilBERT моделі BERT моделінің жеңілдетілген және оңтайландырылған нұсқасы болып табылады. Оның негізгі артықшылықтарының бірі жоғары өнімділігі және есептеу шығындарының төмендігінде.

DistilBERT модельдің схемалық құрылымы 1-суретте көрсетілген [1]. Ол деректерді өңдеу барысындағы кіріс қабаты мен шығыс қабатындағы болжамға дейінгі қадамдарды бейнелейді.



Сурет 2 – DistilBERT моделінің архитектурасы

Суретте зерттеуде пайдаланылған DistilBERT моделінің архитектурасы көрсетілген. Токенизацияланған кіріс деректері алдымен векторлық көрініске түрленеді, содан кейін 6 трансформер қабаттары арқылы өтеді.

DistilBERT моделі [2] шектеулі есептеу ресурстарына байланысты таңдалды, өйткені ол бастапқы BERT өнімділігінің 97% сақтайды, бірақ параметрлердің жартысын ғана пайдаланады. Бұл оны федеративті оқыту жағдайында агрессивті мазмұнды жіктеу мәселелеріне оңтайлы шешім ретінде ұсынады.

Федеративті оқыту жүйесі.

Федеративті оқыту тұжырымдамасына тоқталсақ [3], бұл – деректерді жергілікті құрылғыларда сақтай отырып, модельдерді оқытуға арналған машиналық оқыту әдісі. Бұл тәсіл деректердің құпиялылығы мен қауіпсіздігін қамтамасыз етуге бағытталған. Деректерді орталық серверге жіберудің орнына, әрбір құрылғы модельдерді өзінде оқытады және тек модель параметрлерінің жаңартуларымен, мысалы, салмақтар мен градиенттермен алмасады. Жиналған жаңартулардың негізінде ортақ жаһандық модель қалыптастырылады.

Федеративті оқытудың архитектурасы, қолданылуы және түрлі салалардағы даму перспективалары туралы көптеген зерттеулер жүргізілген. Бұл жұмыстар федеративті оқытудың келешегі зор екенін растайды. Мәселен 1-кестеде келесі авторлардың еңбектері қарастырылған.

Кесте 1 – Федеративті оқыту әдісі бойынша әдебиеттерге шолу

Әдебиеттер	Автор(лар), жылы	Жұмыстың атауы	Зерттеудің мақсаты	Нәтижесі
[4]	Xin'ao Wang, Huan Li, Ke Chen, Lidan Shou, 2023	FedBFPT: An Efficient Federated Learning Framework for BERT Further Pre-training	Федеративті оқыту арқылы BERT моделін салалық міндеттерге бағыттап, есептеу және байланыс шығындарын азайта отырып жетілдіру.	JNLPBA: 0.71; SciERC: 0.64; RCT-20k: 0.83.
[5]	Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, Marius Portmann, 2023	Cyber Threat Intelligence Sharing Scheme Based on Federated Learning for Network Intrusion Detection	Бірнеше ұйымның қатысуымен киберқауіптерді анықтауға арналған NIDS моделін бірігіп оқыту	91.16–93.08% дәлдік көрсетті, шабуылдарды 93% деңгейде анықтады.
[6]	Dr.Aradhana Sahu, Dr. Yousef және т.б., 2024	Federated LSTM Model for Enhanced Anomaly Detection in Cyber Security	Федеративті оқыту мен LSTM модельдерін біріктіру	NSL-KDD, KDD-99, UNSW-NB15 деректерінде дәлдік – 98.9%
[7]	B.Olanrewaju-George, B. Pranggono, 2025	Federated learning-based intrusion detection system for IoT using unsupervised and supervised DL models	IoT құрылғылары үшін федеративті оқыту арқылы аномалияны анықтау	F1-score – 93%, Recall – 99.9%, Precision – 99.3%;
[8]	Arash Heidari, Nima Jafari Navimipour, Hasan Dag, Samira Talebi, Mehmet Unal, 2024	A Novel Blockchain-Based Deepfake Detection Method Using Federated and Deep Learning Models	Дипфейктерді анонимді әрі қауіпсіз түрде анықтайтын жүйе жасау	ACC ≥ 98.9% AUC ≥ 99.3%

Федеративті оқыту негізінде құрылған жүйелер үш негізгі компоненттен тұрады, олардың әрқайсысы жүйенің үздіксіз жұмыс істеуін қамтамасыз ететін маңызды рөл атқарады [9]:

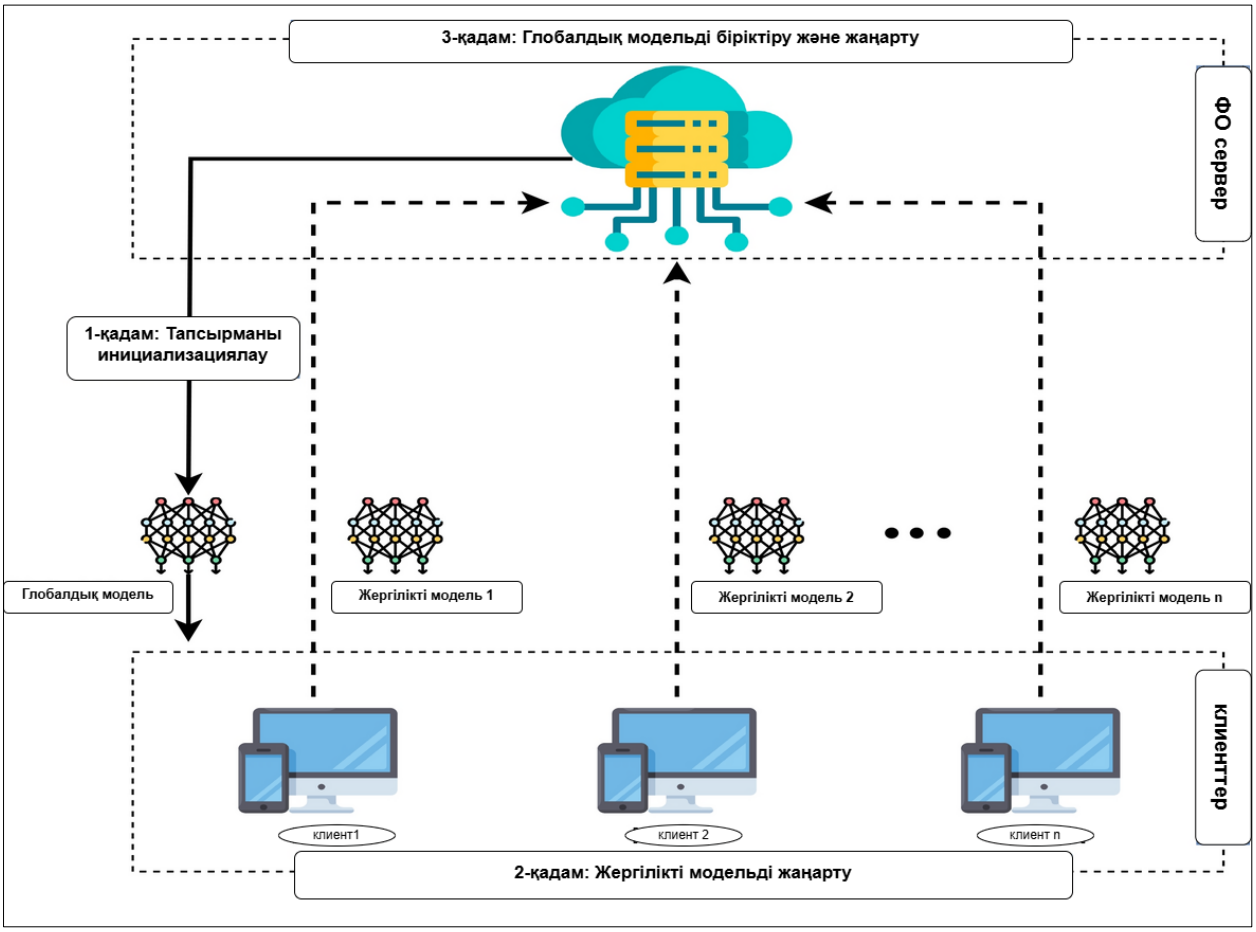
Клиенттер – деректердің иесі болып табылатын құрылғылар немесе ұйымдар. Олар жергілікті модельдерді оқытады, деректерді сыртқа шығармай өңдейді. Мысал: Смартфондар, IoT құрылғылары, ұйым ішіндегі серверлер.

Сервер федеративті оқытудың үйлестірушісі ретінде әрекет етеді. Ол клиенттерден алынған модель жаңартуларын жинақтап, глобалдық модельді қалыптастырады. Мысал: Орталық сервер немесе бұлттық жүйе.

Коммуникациялық-есептеу ортасы – клиенттер мен сервер арасындағы модель параметрлерімен алмасуды қамтамасыз ететін байланыс арнасы. Құпиялылықты сақтау үшін шифрлау технологиялары қолданылады. Мысал: Желі байланысы (Wi-Fi, мобильді байланыс).

Бұл үш компоненттің өзара әрекеттесуі деректердің құпиялылығын сақтай отырып, үлестірілген оқыту процесін тиімді ұйымдастыруға мүмкіндік береді. Клиенттер жергілікті модельдерді оқытса, сервер глобалдық модельді үйлестіреді және агрегаттайды, ал

коммуникациялық-есептеу ортасы параметрлердің алмасуын қамтамасыз етеді. Бұл жүйе орталықтандырылмаған және қауіпсіз түрде модельді оқытуға мүмкіндік береді, әсіресе деректер құпиялылығын сақтау қажет салаларда тиімді. Мәселен 3-суретте автор Shabnam Saki [10] өзінің мақаласында қолданған федеративті оқыту архитектурасы көрсетілген.



Сурет 3 – Федеративті оқыту әдісінің архитектурасы

Бұл архитектура үш негізгі қадамды қамтитын федеративті оқытудың стандартты жұмыс процесін көрсетеді:

1. Тапсырманы инициализациялау: сервер барлық клиенттік түйіндерге бастапқы глобалдық модель параметрлерін жібереді.

2. Жергілікті оқыту: әрбір клиент өзінің жергілікті деректерін пайдалана отырып, өзінің жаңа жергілікті моделін оқытады. Бұл деректердің құпиялылығын сақтауға мүмкіндік береді, себебі деректердің өзі пайдаланушы құрылғысында қалады.

3. Глобалдық агрегация: сервер клиенттерден алынған жергілікті модель жаңартуларын біріктіру арқылы жаһандық модельді жаңартады.

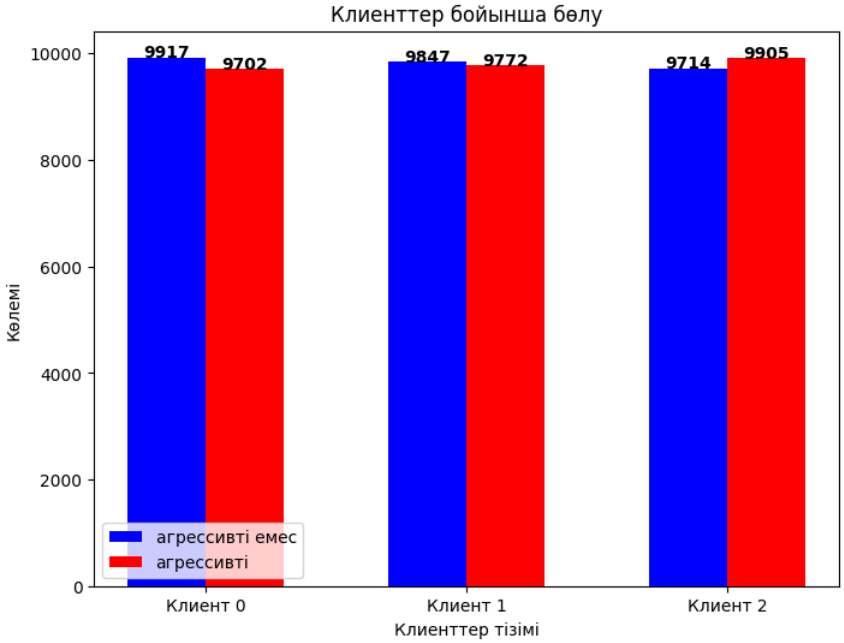
2 және 3-ші қадамдар қажетті дәлдікке жеткенше қайталанады. Бұл федеративті оқыту процесі әртүрлі машиналық оқыту модельдеріне қолданылады.

Клиенттер арасында деректерді бөлу (IID принципі).

Жүргізілген зерттеу жұмысы аясында деректерді үш клиент арасында бөлу IID принципі бойынша жүзеге асады. Әр клиентке шамамен тең пропорцияда агрессивті және агрессивті емес мәтіндер бөлінді, бұл әсіресе 4-суретте айқын көрінеді. Суретте көріп отырғанымыздай, барлық үш клиент IID (Independent and Identically Distributed) әдісін пайдалану арқылы салыстырмалы тең көлемді ақпарат алған. Агрессивті емес

хабарламалар көк түспен, ал агрессивті хабарламалар қызыл түспен көрсетілген. Көлденең осьте клиенттер саны, ал тік осьте тиісті классқа жататын көрсеткіштердің саны бейнеленген. Нәтижесінде:

Клиент 0: 9917 агрессивті емес және 9702 агрессивті мәтін
Клиент 1: 9847 агрессивті емес және 9772 агрессивті мәтін
Клиент 2: 9714 агрессивті емес және 9905 агрессивті мәтін



Сурет 4 – IID принципі бойынша клиенттер арасында деректер жинағын бөлу

Федеративті оқытуда IID әдісі клиенттер арасында деректерді тең бөлу үшін қолданылады. Яғни класстар арасындағы тепе-теңдік болу үшін қолданамыз. Бұл класстарды анықтау барысында дәлдікті қамтамасыз етеді.

IID әдісін математикалық тұрғыдан алғанда [11], толық деректер жиыны –D, көлемі – N болатын жұптардың жиынтығы ретінде (x_i, y_i), мұнда x_i – кіріс мән (мысалы, мәтін), ал $y_i \in \{0; 1\}$ – сәйкес класстық меткасы (агрессивті емес немесе агрессивті):

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad (1)$$

$$D = D_1 \cup D_2 \cup \cdots \cup D_k, \quad D_k \subset D, \ |D_k| \approx \frac{N}{K} \quad (2)$$

Деректерді үлестірілуі кезінде әрбір клиентке өзіне тиесілі деректер жиыны – D_k бастапқы жалпы жиын D-ден кездейсоқ түрде алына отырып класстар арасында тепе-теңдікті сақтайды (агрессивті емес және агрессивті):

$$P(y = c \mid x \in D_k) = P(y = c \mid x \in D), \ \forall k, \forall c \in C, \quad (3)$$

мұндағы C – класстар, біздің жағдайда C = {0,1};

$P(y = c \mid x \in D_k)$ – k-клиенттің деректерінде классстың кездесу ықтималдығы;

$P(y = c \mid x \in D)$ – бастапқы деректер жиынындағы класстардың жиілігі.

Осындай біркелкі бөлудің арқасында әр клиент модельді оқытуда бірдей маңызды рөл атқара алады. Нәтижесінде жергілікті модельдер бір-бірімен үйлесімді және жаһандық модельге біріктірілген кезде ешқандай клиенттік деректер басымдыққа ие болмайды.

Нәтижелер және оларды талқылау.

Бұл бөлімде жүргізілген эксперимент нәтижесінде таңдалған гипер параметрлері, қателік функциясы, бағалау көрсеткіштері және қолданылатын бағдарламалық жасақтама

мен аппараттық орта туралы қосымша ақпарат берілген. Distilbert моделі HuggingFace Transformers кітапханасын пайдаланып PyTorch шеңберінде оқытылды. Қателік функциясы ретінде CrossEntropyLoss қолданылды және оңтайландыру Adam оптимизаторының көмегімен жүзеге асырылды. Сервердегі жаһандық модель жаңартулары FedAvg алгоритмін қолдана отырып біріктірілді.

Федеративті оқыту 7 раундты қамтыды, олардың әрқайсысында барлық 3 клиент 3 дәуірден өтті. Токендер тізбегінің максималды ұзындығы 128, батч өлшемі 16 және оқу жылдамдығы 2e-5 болды. Орнатылған параметрлердің толық көрсеткіштері 2-кестеде берілген.

Кесте 2 – Параметрлердің толық көрсеткіштері

Параметрлер тізімі	**Мәні**
Клиенттер саны	3 клиент
Біріктіру раундтарының саны	7 раунд
Оқыту дәуірі	3 дәуір
Оқу жылдамдығы	2e-5
Батч өлшемі	16 батч
Градиентті оңтайландырғыш	Adam
Шығын функциясы	Cross-entropy
Модель	DistilBERT
Фреймворк	PyTorch + HuggingFace Transformers
Біріктіру	FedAvg
Деректер жиынын бөлу	IID
Бағалау көрсеткіші	F1-score, Accuracy, Precision, Recall

3-кестеде көрсетілгендей, модель барлық негізгі көрсеткіштер бойынша екі класс үшін де (0 және 1) жақсы нәтижелерге қол жеткізгенің көрсек болады. Көрсеткіштердің мұндай нәтижесі модельдің бинарлы классификация негізінде дәл тану қабілетінің жоғарылығын көрсетеді. Сонымен қатар, macro avg және weighted avg көрсеткіштерінің нәтижесі класстардың біркелкі бөлінгенін растайды. Оқытылған модельдер мен нәтижелер Google Drive-та сақталды және оларды келесі эксперименттерде қайта пайдалануға немесе талдау жұмыстарын жүргізуге болады.

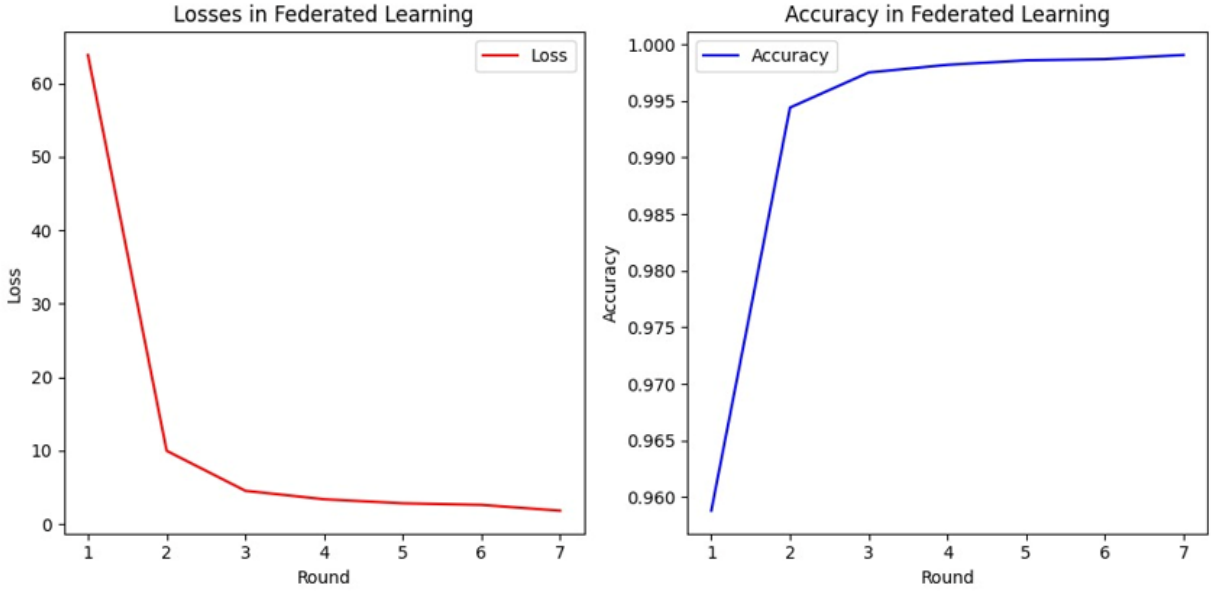
Кесте 3 – Негізгі көрсеткіштер бойынша нәтижелер

	precision	**recall**	**f1-score**	**Support**
0	0.93	0.96	0.94	36786
1	0.96	0.93	0.94	36786
accuracy			0.95	73572
macro avg	0.95	0.93	0.94	73572
weighted avg	0.95	0.94	0.95	73572

Бұл нәтижелер ұсынылған DistilBERT моделі негізіндегі федеративті оқыту әдісін агрессивті мазмұнды анықтауда қолдануға болатынын растайды, бұл тек жоғары дәлдікті ғана емес, сонымен қатар нақты ортада сенімділікті қамтамасыз етеді. 5-суреттен көрініп тұрғандай, модельді федеративті ортада оқыту барысында қателік функциясының төмендеуі және дәлдіктің жоғарылауы графиктерімен анық көрсетілген.

Сол жақтағы график федеративті оқытудың әрбір раундынан кейін loss мәні қалай өзгергенін көрсетеді. Мәселен бірінші раундта 60-тан 10-ға дейін түсті. Келесі раундтарда қателік функциясының мәні біртіндеп төмендей береді және 2-ден аз минималды мәнге жетеді. Оң жақтағы график модельдің дәлдігі қалай өзгергенін көрсетеді:

- бірінші раундта дәлдік шамамен 0,96 құрайды, бұл модельдің дәл болжамдар жасау мүмкіндігін көрсетеді;
- екінші және үшінші раундтарда дәлдік тез артып, 0,995-тен асады;
- 4-тен 7-ге дейін дәлдік мәні 0,999 деңгейіне жетеді, бұл модельдің жоғары сапаға қол жеткізетіндігін және енді маңызды түзетулерді қажет етпейтіндігін көрсетеді.



Сурет 5 – Федеративті оқытудағы қателік мен дәлдік графигі

Бинарлы классификация моделінің сапасын одан әрі бағалау үшін confusion matrix и ROC қолданылды, олардың әрқайсысы болжамдардың қаншалықты жақсы екенін көрсетеді.

ROC қисығы осьтер бойымен құрылады [12]:

– x-осі: False Positive Rate (жалған оң нәтижелердің үлесі),

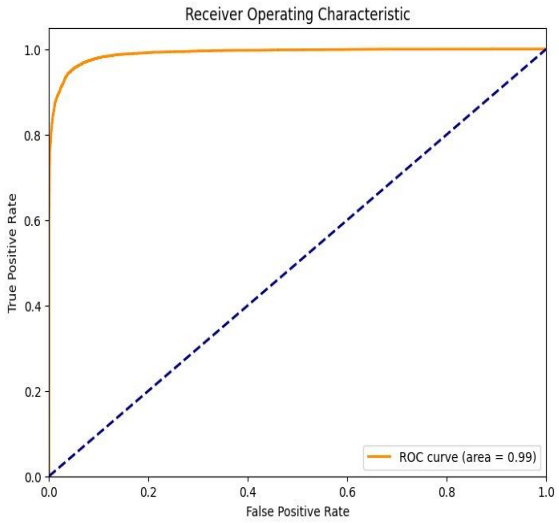
$$FRP = \frac{FP}{FP + TN} \tag{4}$$

– y-осі: True Positive Rate (шынайы оң нәтижелердің үлесі),

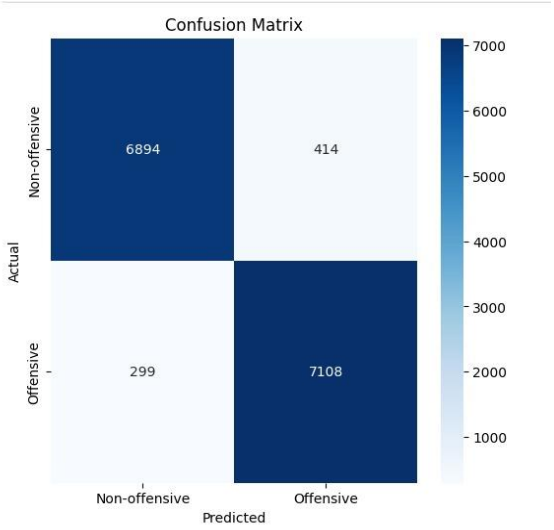
$$TRP = \frac{TP}{TP + FN} \tag{5}$$

6-суретте модельдің әртүрлі шектерде оң және теріс класстарды ажырату мүмкіндігін көрсететін ROC (Receiver Operating Characteristic) қисығы көрсетілген. Сызық графиктің жоғарғы сол жақ бұрышына неғұрлым жақын болса, классификатор соғұрлым жақсы жұмыс істейді. Біздің жағдайда модель идеалды мінез-құлықты көрсетеді, өйткені ROC қисығы сол және жоғарғы жиектерге толығымен бағытталған. Ең маңызды көрсеткіш – AUC (Area Under Curve) – 0.99, бұл модельдің хабарламаларды ажырату қабілетінің жоғары екендігін көрсетеді.

Confusion Matrix модельді екі класс бойынша жіктеу нәтижелері туралы егжей-тегжейлі ақпарат береді. 7-суретте confusion matrix визуализациясы берілген, онда нақты сыныптар тігінен және модель болжаған сыныптар көлденеңінен көрсетілген. Нәтижесінде: 6894 агрессивті мазмұн дұрыс жіктелді; 7108 агрессивті емес мазмұн дұрыс анықталған; 414 агрессивті емес мысалдар қате түрде агрессивті деп жіктеледі; 299 агрессивті хабарлама қате түрде агрессивті емес деп жіктеледі.



Сурет 6 – Агрессивті сөйлемдерді классификациялау үшін ROC-қисық



Сурет 7 – Классификациялауға арналған қателік матрицасы

Эксперимент нәтижелері FedAvg алгоритмін және деректерді біркелкі (IID) бөлу әдісін қолдана отырып, федеративті оқыту жағдайында distilbert моделы негізіндегі ұсынылған архитектура жоғары дәлдік пен тұрақтылықты көрсетеді. Барлық негізгі көрсеткіштердің мәндері – accuracy, precision, recall және F1 – 0.94-тен асады, ал ROC қисығының көрсеткіші (AUC = 0.99) модельдің класстарды ажырата білу қабілетін растайды.

Қорытынды.

Бұл жұмыста агрессивті мазмұнды анықтау тапсырмасы үшін федеративті оқытуды пайдалану мүмкіндігі зерттелді. Бұл міндет әсіресе интернетте қауіпсіз және этикалық тұрақты байланысты қамтамасыз ету қажет цифрландыру дәуірінде өзекті болып табылады. Негізгі мәселелердің бірі – құпиялылықты бұзбай және деректерді орталықтандырылған серверлерге тасымалдамай, пайдаланушы мазмұнын модерациялау қажеттілігі. Бұл мәселені шешу үшін федеравтивті оқыту әдісі ұсыналады.

Негізгі архитектура ретінде DistilBERT трансформаторлық моделі таңдалды, ол толық BERT модельдерімен салыстырғанда есептеу шығындары төмен және жоғары дәлдік пен өнімділікті біріктіреді. Модель FedAvg алгоритмін пайдаланып, үш клиент арасында деректерді біркелкі (IID) үйлестіру арқылы федеративті оқыту орындалды.

Эксперимент нәтижелері ұсынылған тәсілдің жоғары тиімділігін растады. Модель 94% жалпы дәлдікке қол жеткізді, әр класс үшін F1 ұпайы 0,94 көрсеткішіне тең болды. Шатасу матрицасы жалған оң және жалған теріс мәндердің ең аз санымен болжаулардың теңгерімді таралуын көрсетті. Сонымен қатар, оқыту динамикасының графиктері агрегацияның бірінші раундтарынан бастап қателік функциясының тұрақты төмендеуін және дәлдіктің өсуін көрсетті, бұл модельдің жылдам конвергенциясы мен тұрақтылығын көрсетеді.

Алынған нәтижелер көрсеткендей, аздаған оқу раундтары мен клиенттер арасында параметрлердің ең аз алмасуымен де жаһандық модельді өте тиімді оқытуға болады. Ол жоғары классификация сапасын сақтайды және әртүрлі қатысушылардан алынған білімді жақсы қорытындылайды. Бұл жағдайда деректер пайдаланушыларда қалады және серверге берілмейді, бұл ақпараттың ағып кету қаупін айтарлықтай төмендетеді.

Дегенмен, жұмыста шектеулер бар: зерттеу тек деректердің біркелкі (IID) үлестіру жағдайында жүргізілді. Болашақта біркелкі емес (none IID) үлестірімі бар неғұрлым нақты сценарийлерді зерттеу, сондай-ақ басқа агрегаттау алгоритмдерін сынау (FedAdam, FedYogi) және нақты жағдайларға жақындау үшін клиентте санын көбейту жоспарлануда.

Әдебиеттер тізімі

1. Khan, Y., Sánchez, D., & Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: A systematic literature review. *Artificial Intelligence Review, 57*, 320. https://doi.org/10.1007/s10462-024-10970-5

2. Zhang, H., Bosch, J., & Holmström Olsson, H. (2025). Enabling efficient and low-effort decentralized federated learning with the EdgeFL framework. *Information and Software Technology, 178*, Article 107600. https://doi.org/10.1016/j.infsof.2024.107600

3. Kumarappan, J., Rajasekar, E., Vairavasundaram, S., & Others. (2024). Federated learning enhanced MLP–LSTM modeling in an integrated deep learning pipeline for stock market prediction. *International Journal of Computational Intelligence Systems, 17*, 267. https://doi.org/10.1007/s44196-024-00680-9

4. Wang, X., Li, H., Chen, K., & Shou, L. (2023). FedBFPT: An efficient federated learning framework for BERT further pre-training. *Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)* (pp. 4344–4352).

5. Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2023). Cyber threat intelligence sharing scheme based on federated learning for network intrusion detection. *Journal of Network and Systems Management, 31*(3). https://doi.org/10.1007/s10922-022-09691-3

6. Sahu, A., El-Ebiary, Y. A. B., Saravanan, K. A., Thilagam, K., Devi, G. R., Gopi, A., & Taloba, A. I. (2024). Federated LSTM model for enhanced anomaly detection in cyber security: A novel approach for distributed threat. *International Journal of Advanced Computer Science and Applications, 15*(6), 1237–1249.

7. Olanrewaju-George, B., & Pranggono, B. (2025). Federated learning-based intrusion detection system for the Internet of Things using unsupervised and supervised deep learning models. *Cyber Security and Applications, 3*, 100068. https://doi.org/10.1016/j.csa.2024.100068

8. Heidari, A., Navimipour, N. J., Dag, H., Talebi, S., & Unal, M. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. *Cognitive Computation, 16*, 1073–1091. https://doi.org/10.1007/s12559-024-10255-7

9. Schumann, G., Awick, J.-P., & Marx Gómez, J. C. (2023, September). *Natural language processing using federated learning: A structured literature review*. In *2023 International Conference on AI-Based Things*, (pp. 100–112). IEEE. https://doi.org/10.1109/AIBThings58340.2023.10292481

10. Khan, Y., Sánchez, D., & Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: A systematic literature review. *Artificial Intelligence Review, 57*, Article 320. https://doi.org/10.1007/s10462-024-10970-5

11. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

12. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

References

1. Khan, Y., Sánchez, D., & Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: A systematic literature review. *Artificial Intelligence Review, 57*, 320. https://doi.org/10.1007/s10462-024-10970-5

2. Zhang, H., Bosch, J., & Holmström Olsson, H. (2025). Enabling efficient and low-effort decentralized federated learning with the EdgeFL framework. *Information and Software Technology, 178*, Article 107600. https://doi.org/10.1016/j.infsof.2024.107600

3. Kumarappan, J., Rajasekar, E., Vairavasundaram, S., & Others. (2024). Federated learning enhanced MLP–LSTM modeling in an integrated deep learning pipeline for stock market prediction.

International Journal of Computational Intelligence Systems, 17, 267. https://doi.org/10.1007/s44196-024-00680-9

4. Wang, X., Li, H., Chen, K., & Shou, L. (2023). FedBFPT: An efficient federated learning framework for BERT further pre-training. *Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)* (pp. 4344–4352).

5. Sarhan, M., Layeghy, S., Moustafa, N., & Portmann, M. (2023). Cyber threat intelligence sharing scheme based on federated learning for network intrusion detection. *Journal of Network and Systems Management, 31*(3). https://doi.org/10.1007/s10922-022-09691-3

6. Sahu, A., El-Ebiary, Y. A. B., Saravanan, K. A., Thilagam, K., Devi, G. R., Gopi, A., & Taloba, A. I. (2024). Federated LSTM model for enhanced anomaly detection in cyber security: A novel approach for distributed threat. *International Journal of Advanced Computer Science and Applications, 15*(6), 1237–1249.

7. Olanrewaju-George, B., & Pranggono, B. (2025). Federated learning-based intrusion detection system for the Internet of Things using unsupervised and supervised deep learning models. *Cyber Security and Applications, 3*, 100068. https://doi.org/10.1016/j.csa.2024.100068

8. Heidari, A., Navimipour, N. J., Dag, H., Talebi, S., & Unal, M. (2024). A novel blockchain-based deepfake detection method using federated and deep learning models. *Cognitive Computation, 16*, 1073–1091. https://doi.org/10.1007/s12559-024-10255-7

9. Schumann, G., Awick, J.-P., & Marx Gómez, J. C. (2023, September). *Natural language processing using federated learning: A structured literature review*. In *2023 International Conference on AI-Based Things*, (pp. 100–112). IEEE. https://doi.org/10.1109/AIBThings58340.2023.10292481

10. Khan, Y., Sánchez, D., & Domingo-Ferrer, J. (2024). Federated learning-based natural language processing: A systematic literature review. *Artificial Intelligence Review, 57*, Article 320. https://doi.org/10.1007/s10462-024-10970-5

11. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning, 14*(1–2), 1–210. https://doi.org/10.1561/2200000083

12. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

ФЕДЕРАТИВНОЕ ОБУЧЕНИЕ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ АГРЕССИВНОГО КОНТЕНТА: ПОДХОД НА ОСНОВЕ МОДЕЛИ ТРАНСФОРМАТОРА

Аннотация. *Стремительное развитие цифровых коммуникаций привело к увеличению количества постов агрессивного содержания в Интернете. Автоматическое обнаружение такого контента является одной из самых актуальных проблем нашего времени. Однако традиционные подходы, основанные на сборе данных на центральном сервере, могут поставить под угрозу конфиденциальность личной информации. Одним из способов решения этой проблемы является использование метода федеративного обучения. Данный метод подразумевает индивидуальное обучение модели на устройстве каждого пользователя, без отправки данных на центральный сервер. В ходе исследования был проведен обзор литературы научных работ и проанализирован опыт использования метода федеративного обучения. В качестве набора данных использовался специальный корпус, состоящий из 73 572 записей агрессивных и неагрессивных текстов. Для обучения модели использовалась модель DistilBERT, а набор данных был разделен между тремя клиентами, каждый из которых обучал только свои собственные записи по отдельности. В конце каждого раунда сервер использует алгоритм FedAvg для объединения параметров модели, предоставленных всеми клиентами на сервере, создавая общую глобальную модель. На основании полученных результатов можно сделать вывод, что метод федеративного*

обучения имеет два важных преимущества: во-первых, он работает с высокой точностью, а во-вторых, обеспечивает надежность и конфиденциальность информации.

Ключевые слова: *федеративное обучение, обработка естественного языка, DistilBERT, FedAvg, сохранение конфиденциальности, агрессивный контент, классификация.*

A TRANSFORMER MODEL APPROACH TO FEDERATED LEARNING FOR AGGRESSIVE CONTENT CLASSIFICATION TASKS

Abstract. *The rapid development of digital communication has led to an increase in the number of offensive postings on the Internet. Automatic detection of such content is one of the most pressing problems of our time. However, traditional approaches based on collecting data on a central server can compromise the privacy of personal information. One way to address this issue is to use federated learning. This method involves individual model training on each user's device without sending data to a central server. In the course of the study, a literature review of scientific papers was conducted and experiences with the federated learning method were analyzed. A special corpus consisting of 73,572 recordings of aggressive and non-aggressive texts was used as a dataset. The DistilBERT model was used to train the model, and the dataset was divided among three clients, each of which trained only their own recordings separately. At the end of each round, the server uses the FedAvg algorithm to combine the model parameters provided by all of the clients on the server to create a common global model. Based on the results, it can be concluded that the federated learning method has two important advantages: first, it works with high accuracy, and second, it ensures the reliability and confidentiality of information.*

Keywords: *federated learning, natural language processing, DistilBERT, FedAvg, privacy, aggressive content, classification.*

Авторлар туралы мәлімет

Омар Айым Бекболатқызы	Техника ғылымдарының магистрі, «Киберқауіпсіздік және криптология» кафедрасының докторанты, Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан, E-mail: aiym.omar98@gmail.com
Мусиралиева Шынар Женисбековна	PhD, профессор, «Киберқауіпсіздік және криптология» кафедрасының меңгерушісі, Ақпараттық технологиялар факультеті, Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан, E-mail: mussiraliyevash@gmail.com

Сведение об авторах

Омар Айым Бекболатқызы	Магистр технических наук, докторант кафедры «Кибербезопасность и криптология», Факультет информационных технологий, Казахский национальный университет имени аль-Фараби, г.Алматы, Казахстан, E-mail: aiym.omar98@gmail.com
Мусиралиева Шынар Женисбековна	PhD, профессор, заведующая кафедрой «Кибербезопасность и криптология», Факультет информационных технологий, Казахский национальный университет имени аль-Фараби, г.Алматы, Казахстан, E-mail: mussiraliyevash@gmail.com

Information about the authors

Omar Aiym	Master of Technical Sciences, PhD student at the Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh National University, Almaty, Kazakhstan, E-mail: aiym.omar98@gmail.com
Mussiraliyeva Shynar	PhD, Professor, Head of the Department of Cybersecurity and Cryptology, Faculty of Information Technologies, Al-Farabi Kazakh National University, Almaty,Kazakhstan, E-mail: mussiraliyevash@gmail.com