ӘОЖ 004.2
ҒТАХР 20.23.01

https://doi.org/10.55452/1998-6688-2026-23-2-326-339

[1,2]Самбетбаева М.,
PhD, қауымдастырылған профессор, жетекші ғылыми қызметкер,
ORCID ID: 0000-0001-9358-1614,
e-mail: madina_jgtu@mail.ru
[1,3]Еримбетова А.,
PhD, қауымдастырылған профессор, жетекші ғылыми қызметкер,
ORCID ID: 0000-0002-2013-1513,
e-mail:aigerian8888@gmail.com
[1,2]Абдығалым Б.,
магистр, кіші ғылыми қызметкер, ORCID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com
[1,4]*Дайырбаева Э.,
магистр, ғылыми қызметкер, аға оқытушы,
ORCID ID: 0000-0002-4255-5456,
*e-mail: nurbekkyzyelmira@gmail.com
[1]Тұрғанбаев А.,
инженер-бағдарламашы, ORCID ID:0000-0001-8989-7827,
e-mail: m1challenge@inbox.ru

[1]ҚР ҒЖБМ ҒК қарасты «Ақпараттық және есептеу технологиялары институты»,
Алматы қ., Қазақстан
[2]Л.Н. Гумилев атындағы Еуразиялық ұлттық университеті, Астана қ., Қазақстан
[3]META University, Алматы қ., Қазақстан
[4]Қ.И. Сәтпаев атындағы Қазақ ұлттық техникалық зерттеу университеті,
Алматы қ., Қазақстан

ОНТОЛОГИЯЛЫҚ ЖӘНЕ КОРПУС ДЕРЕКТЕРІН БІРІКТІРГЕН ТІЛДІК МОДЕЛЬ

Аңдатпа

Бұл зерттеу әскери және геосаяси оқиғаларға қатысты әлеуметтік желілер мен жаңалықтар платформаларынан жиналған 1000 мәтіннен тұратын аннотацияланған корпус құру арқылы күнделікті дискурстағы цифрлық қауіптерді талдауға бағытталған. Ұсынып отырған зерттеудің мақсаты ақпараттық операциялардың элементтері бар мәтіндердің аннотацияланған корпусын құру арқылы күнделікті дискурстағы цифрлық қауіптерді талдаудың шұғыл қажеттілігін шешу болып табылады. Әсер ету түрін, эмоционалды бояуды, жалған ақпарат белгілерін және авторлық ниеттерді (арандату, қорқыту және т.б.) қамтитын көп деңгейлі аннотация схемасы жасалды. Аннотация сапаны бақылау және контекстті ескере отырып, Label Studio-да орындалады; сенімділік Каппа=0,82 коэффициентімен расталады. Onto-IO-Bert моделі пилоттық экспериментте F1=0,81 көрсетті, бұл негізгі модельдерден асып түсті. Корпустың практикалық қолданылуы нақты Telegram хабарламаларын талдау арқылы расталады. Осы жұмыста Қазақстандық ІІМ кеңістігіндегі әскери ақпараттық операцияларды талдауға бейімделген аннотацияның көп деңгейлі моделі ұсынылған және құрылған корпус қолданыстағы деректер жиынтығындағы маңызды олқылықты толтырып, әскери ақпараттық операцияларды талдауға арналған жаңа ресурс болып табылады. Ұсынылған корпус қазақ, орыс және ағылшын тілдеріндегі мәтіндерді қамтиды. Аннотация құрылымы манипуляциялық мазмұнның лингвистикалық және прагматикалық қырларын модельдеуді жеңілдетеді. Корпус келесі сілтеме бойынша жалпыға қолжетімді: https://github.com/baiangali/multi_mil.

Түйін сөздер: әскери терминдер, аннотацияланған корпус, ақпараттық операциялар, NLP, әлеуметтік желілерді талдау.

Кіріспе

Заман талабына сай жасанды интеллект пен табиғи тілді өңдеу саласында онтологиялық білім базалары мен корпус деректерін біріктіру маңызды ғылыми бағытқа айналып отыр. Онтологиялар ұғымдар арасындағы семантикалық байланыстарды жүйелеуге мүмкіндік берсе, корпус деректері тілдің нақты қолданысын, грамматикалық құрылымдарын және прагматикалық ерекшеліктерін бейнелейді. Бұл екі дерек көзін біріктіру арқылы тілдік модельдер тек статистикалық заңдылықтарды ғана емес, сонымен қатар мағыналық деңгейдегі байланыстарды да қамти алады. Мұндай тәсіл ақпараттық жүйелердің түсіну қабілетін арттырып, көптілді ортада, соның ішінде қазақ тілі секілді ресурстары шектеулі тілдерде де тиімді шешімдер ұсынуға жол ашады [1, 2].

Қазіргі кезеңдегі қарулы қақтығыстар барған сайын гибридті сипат иеленіп, дәстүрлі әскери операциялар ақпараттық ықпал ету әрекеттерімен кешенді түрде толықтырылуда [3]. Бұл үдерісте цифрлық медиа, әсіресе әлеуметтік желілер, бір мезгілде ресми ақпаратты жеткізудің тиімді арнасы әрі ашық және жасырын сипаттағы ақпараттық-психологиялық ықпал ету үшін бәсекелестік алаңы ретінде стратегиялық маңызға ие. Мұндай ықпалдың негізгі көріністеріне мақсатты жалған ақпарат тарату, бейбіт тұрғындар мен әскери қызметкерлердің моральдық-психологиялық жағдайын әлсірету, мемлекеттік және халықаралық институттарға деген сенімді іріткі салу, әлеуметтік дүрбелең туғызу, жекелеген топтарды стигматизациялау, сондай-ақ зорлық-зомбылықты заңдастыруға ұмтылу жатады [4, 5]. Таратылатын ақпараттардың мазмұны көбінесе бөлшектелген, эмоционалды тұрғыдан «зарядталған» және манипуляциялық сипатта болады, үгіт-насихат әдістеріне және әскери терминологияға сүйенеді [6, 7].

Цифрлық қауіптерді бағалауда автоматтандырылған мәтінді талдау маңызды құрал болғанымен, оның тиімділігі оқыту деректерінің сапасы мен өзектілігіне байланысты [8]. NLP модельдері үшін қолданылатын корпустар лингвистикалық тұрғыдан дұрыс болуымен қатар, конфликт жағдайындағы ақпараттық операциялардың ерекшеліктеріне бейімделуі керек. Қолданыстағы NER ресурстары көбіне тұрақты салаларға және ағылшын тіліне [9, 10] бағытталған. Оларда ақпараттық операцияларды талдау үшін қажетті әсер ету түрі, мақсатты аудитория, автор ниеті сияқты маңызды аннотация санаттары мен жалған ақпарат көрсеткіштері жоқ [11]. Зерттеулердің тағы бір шектеуі – құжат деңгейіндегі аннотацияға сүйену [12], бұл модель дәлдігін төмендетеді, себебі, тіпті беделді дереккөздерде де насихат болуы мүмкін [13]. Бұл мәтіннің өзіне негізделген, егжей-тегжейлі және контекстке сезімтал аннотациялаудың қажеттілігін көрсетеді.

Әскери ақпараттық операцияларды талдау үшін сапалы, арнайы аннотацияланған корпустар жетіспейді, бұл қолданбалы шешімдер мен іргелі зерттеулердің әлеуетін шектейді. Бұл мәселені шешу үшін, бұл жұмыс әскери ақпараттық операциялар белгілері бар мәтіндер жинағын ұсынады. Оның басты ерекшелігі – дайындалған топтың қолмен аннотациялауы, ол субъектілер мен прагматикалық ерекшеліктерді (әсер ету түрі, автор ниеті, жалғандық көрсеткіштері, т.б.) қамтитын мамандандырылған схеманы қолданады [14]. Негізгі мақсат – әскери дискурстағы ақпараттық ықпал ету тәжірибесін көрсететін әлеуметтік желілер мен жаңалықтар платформаларынан әдістемелік негізделген, аннотацияланған мәтіндер жинағын жасау.

Ұқсас жұмыстарға салыстырмалы талдау. Әскери дискурстағы ақпараттық операцияларды зерттеу корпус лингвистикасы, сыни дискурстық талдау, когнитивті соғыс және стратегиялық коммуникацияны қамтитын көпсалалы бағыт болып табылады [15]. Дегенмен, аннотацияланған корпустар тіл білімінде кең қолданылғанымен [16], әскери ақпараттық және психологиялық операцияларға арналған арнайы ресурстар әлі де жеткіліксіз.

Әскери контекстегі корпус әдістерінің ерте қолданыстары Әл-Равидің [17] әскери дискурсты кілт сөз жиілігі бойынша талдауынан [18] және әскери тілдің стратегиялық сипатына назар аударғаны [19] жұмысынан көрінеді. Алайда, бұл зерттеулерде ұсақ түйіршікті құрылым мен

прагматикалық аннотация жетіспейді, бұл оларды автоматты өңдеу үшін шектейді. Реттгер және әріптестері [20] LLM-дердің ақпараттық операцияларды талдаудағы әлеуетін көрсетіп, автоматтандырылған нәтижелерді қолмен тексерудің маңыздылығын атап өтті. Когнитивті соғыс ұғымы аудиторияның қабылдауы мен мінез-құлқын манипуляциялауда басты бағытқа айналуда.

Сонымен қатар, оқиғаларды шығару мәтінді тереңірек семантикалық талдау әдісі ретінде танымал болды. Ace2005 [21], MAVEN [22] және DuEE [23] сияқты сөйлем деңгейіндегі деректер жиынтығы жалпы немесе қаржылық домендерге назар аударады, АЛ RAMS [24], WikiEvents [21] және DocEE [26] сияқты құжат деңгейіндегі деректер жиынтығы дискурстың кең контекстін қарастырады. Алайда, бұл дереккөздерде әскери оқиғаларға қатысты оқиғалар туралы ақпарат жоқ. Әскери мақсаттағы ресурстарды игеруде айтарлықтай жетістіктерге қол жеткізілді. MNEE корпусы [27] сөйлем деңгейінде аннотацияланған Қытай мәтіндерін ұсынады, ал жақында CMNEE жобасы [28] сегіз түрлі (мысалы, қақтығыс, орналастыру, эксперимент) 17000 құжаттан және 29000-нан астам іс-шаралардан тұратын толық құрамды ұсынады.және он бір аргумент рөлі. CMNEE әскери корпорациялардың ерекше қиындықтарын, соның ішінде оқиғалардың тығыз құрылымын, терминологиялық әртүрлілігін және күрделі синтаксисін атап көрсетеді.

MilTAC корпусы [29] әскери дискурсты зерттеуге арналған алғашқы корпус болғанымен, оның аннотациясы лексика-семантикалық деңгеймен шектеліп, коммуникативті ниет, эмоционалды тон және жалған ақпарат сияқты прагматикалық аспектілерді қамтымайды. Дипломат корпусы [30] әскери және саяси коммуникациядағы сөйлеу әрекеттері мен аргументтерге баса назар аударады. Ол риторикалық құрылымдар мен аргументтік стратегияларды түсіндіргенімен, эмоционалды манипуляцияларға немесе жалған ақпаратқа қатысты емес, бұл оның ақпараттық операцияларды зерттеудегі қолданылуын шектейді. Осыған байланысты, бұл зерттеуде ұсынылған Multi_mil корпусы оқиға деңгейіндегі, нысан деңгейіндегі және прагматикалық аннотацияларды біріктіру арқылы кемшіліктерді жоюға бағытталған. Бұл әскери дискурсты көп деңгейлі талдауға және көптілді контексте ақпараттық операциялардың негізгі заңдылықтарын анықтауға мүмкіндік береді.

Бұрынғы деректер жиынтықтарында прагматикалық аннотациялар жетіспеді және олар көбіне ағылшын/қытай тілдерімен шектелді, әрі психологиялық операцияларға аз көңіл бөлінді. Біздің корпус бұл олқылықтарды толтырып, лингвистикалық және прагматикалық деңгейлерді біріктіріп, Қазақстандық геосаяси контекстке бейімделген көптілді деректерді ұсынады, бұл аймақтық және тіларалық NLP зерттеулері үшін маңызды. Бұл зерттеу әскери дискурстағы ақпараттық операцияларды талдауға арналған аннотацияланған корпустың қажеттілігін көрсетеді. Ол доктриналық құрылымдарды корпусқа негізделген заманауи әдістермен біріктіріп, зерттеулердегі олқылықты толтырады және академиялық әрі практикалық тұрғыдан құнды ресурс ұсынады.

Материалдар мен әдістер

Әскери ақпараттық операциялардың корпусын құру үшін әлеуметтік желілердегі мәтіндер Label Studio ортасында [31] қолмен аннотацияланып, милитаристік тезаурус кілт сөздері мен ақпараттық операция типологиясына негізделген жалған ақпарат, деморализация, делегитизация, арандатушылық және қорқыту сияқты ықпал ету стратегияларын қамтитын жазбалар таңдалды (Telegram, Instagram және public news aggregator көздерінен). Пилоттық үлгі келесі критерийлер бойынша таңдалған 200 мәтіннен тұрды:

- Әскери және геосаяси тақырыптарға тақырыптық өзектілігі;
- Қайталанбайтын немесе кросс-орналастырылған мазмұнсыз бірегейлік;
- Ақпараттық әсер ету ерекшеліктерімен лингвистикалық қанықтыру.

Аннотациялардың жан-жақты қамтылуын және сәйкестігін қамтамасыз ету үшін корпус бірнеше кезеңдерде жасалды. Бастапқы кезеңде аннотация бойынша нұсқаулықтарды

нақтылау және аннотация схемасын түзету үшін таңдалған және пайдаланылған 200 мәтіннен тұратын эксперименттік ішкі жиын болды. Содан кейін негізгі мәліметтер жиынтығы аяқталған схемаға сәйкес сарапшылардың қолмен аннотациясы үшін платформаға жүктелді.

id	date	sender_id	text_clean
155100	2025-04-28 20:25:14+00:00	1001137332367	вице-президент сша джей ди вэнс считает, что украина не сможет выиграть войну, но существует серьёзный риск ядерной эскалации. россия сша украина
155099	2025-04-28 20:23:24+00:00	1001137332367	украина и европа опасаются, что президент сша дональд трамп прекратит свои посреднические усилия между киевом и москвой из-за отсутствия прогресса. россия сша украина
155098	2025-04-28 20:07:50+00:00	1001137332367	30 тысяч испанских полицейских дополнительно будут развернуты в городах страны для обеспечения безопасности и помощи местному населению на фоне масштабного отключения электроэнергии. в столице страны городе мадриде начинается развертывание армейских подразделений. все атомные электростанции испании приостановили работу и переведены в безопасный режим. европа
155097	2025-04-28 19:57:02+00:00	1001137332367	я подключу тебя к чёртову полиграфу, - сорвался на и.о главы окнд адмирала кристофера грейди министр обороны сша пит хегсет, подозревая офицера в сливе прессе информации о планах привлечь илона маска к секретному брифингу пентагона о ситуации в китае. сша
155096	2025-04-28 19:52:11+00:00	1001137332367	половину своего рабочего времени министр обороны сша пит хегсет тратит на общение с прессой, борьбу с утечками и тренировки с американскими военными для улучшение собственного имиджа. об этом рассказал один из трех высокопоставленных чиновников пентагон колин кэрролхttpst, уволенных за утечку секретной информации.

Сурет 1 – Аннотацияға дейінгі мәтіндік деректердің құрылымы

Сурет 1-де түпнұсқа жаңалықтардың үзіндісін ұсынады.тазартылған мәтіндік деректерді қамтитын csv файлы. Осы деректер жиынындағы әрбір жол келесі өрістерді қамтиды:

- id – әр жазба үшін бірегей идентификатор;
- күні – ISO форматындағы жарияланымның уақыт белгісі;
- sender_id – жарияланым көзінің идентификаторы (мысалы, Telegram арнасы);
- text_clean – шудан, еренсілтемелерден, эмодзилерден және басқа да маңызды емес белгілерден тазартылған алдын ала өңделген мәтін.

Бұл деректер жиынтығы әрі қарай қолмен аннотациялау үшін негіз болды. Талданған корпустың толық көлемі мәтіндерден алынған 15000 таңбалауыштан асады. Хабарламаның орташа ұзындығы 15-тен 20 сөзге дейін, әр хабарламада шамамен 5-тен 10-ға дейін аннотацияланған нысандар бар, бұл семантикалық тұрғыдан маңызды мазмұнның жоғары тығыздығын көрсетеді.

Аннотация нұсқаулығы қолмен лингвистикалық таңбалауға арналып, екі деңгейлі құрылымды қамтиды:

a) субъект деңгейінде – орналасқан жер, әскери терминология, дереккөздер сияқты аталған мәндер;

b) құжат деңгейінде – ақпараттық операция түрі, эмоционалдық реңк, жалған ақпарат белгілері, автордың ниеті және мақсатты аудитория.

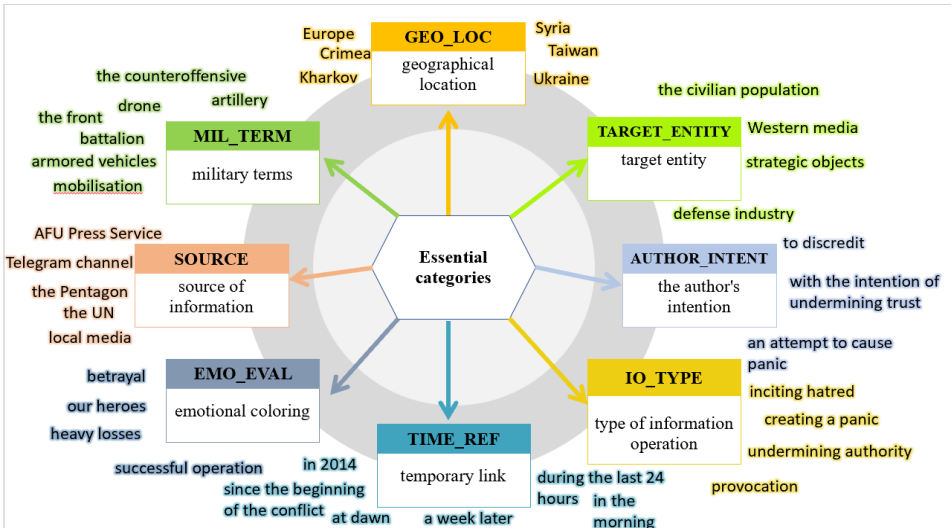
Әр санат ресми критерийлермен, мысалдармен және ескертпелермен толықтырылған. Тегтеу жүйесі халықаралық стандарттарға, ақпараттық операциялар доктриналары мен насихат талдау тәжірибесіне сүйене отырып жасалды.

Сурет 2-де әскери ақпараттық операцияларға қатысты мәтіндер корпусына аннотация жасау кезінде қолданылатын субъектілер санаттарының таксономиясы көрсетілген. Диаграмманың ортасында «um – brella» терминінің субъектілік категориялары орналасқан, олардан сегіз негізгі класс бөлінеді, олардың әрқайсысы корпустан алынған нақты өрнектердің мысалдарымен толықтырылған).

MIL_TERM (ӘСКЕРИ ТЕРМИН) стандартталған әскери терминологияға негізделіп, ISO TC 37 және ACE аннотация хаттамаларында милитаристік әрекеттер мен субъектілерді тануға қолданылады [30]; GEO_LOC және TIME_REF корпус лингвистикасында және ACE бағдарламаларында кеңінен қолданылатын стандартты NER категориялары [32]; ДЕРЕККӨЗ ақпараттың шығу тегін бағалауға мүмкіндік береді [33, 34]; ал TARGET_ENTITY ақпараттық операцияның нақты мақсатын көрсетеді [35, 36] AUTHOR_INTENT психолингвистикалық және үгіт-насихаттық талдау шеңберіне негізделген mes-sage негізіндегі коммуникативті стра-

тегияны ұсынады (мысалы, «жүктелген жергілікті желі», «таңбалау»). Оны қосу манипулятивті әрекетті түсіндіруге мүмкіндік береді.

IO_TYPE (ақпараттық операция түрі) белгісі хабарламаларды әсер ету сипаты бойынша жіктейді – мысалы, жалған ақпарат, көңіл-күйді түсіру, қорқыту-және әскери және халықаралық доктриналық дереккөздерге негізделген. IO_TYPE 1-кесте егжей-тегжейлі көрсетілген сегіз санатты қамтиды.



Сурет 2 – Жіктеу және мысалдарды атап өту

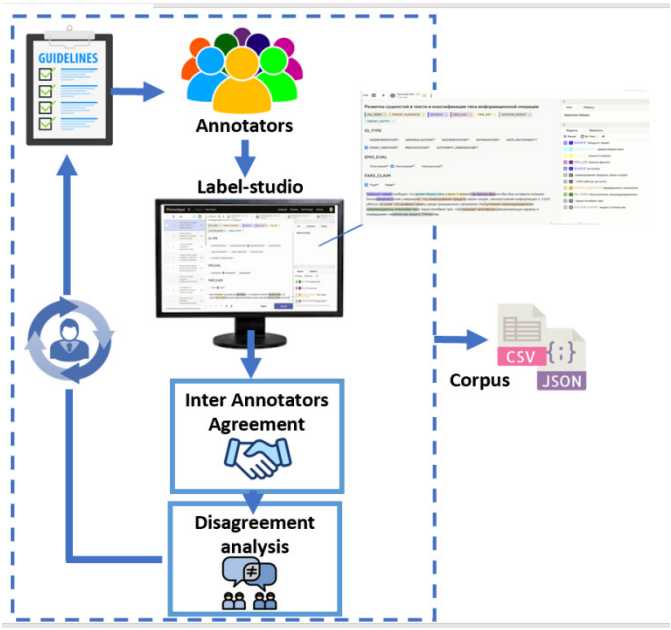
Кесте 1 – IO_TYPE (Ақпараттық операция түрі)

Категория	Сипаттамасы	Мысалдар
DISINFORMATION	Жалған ақпаратты қасақана тарату	KZ: армияның барлық бөлімшелері ұрыссыз берілді EN: All army units surrendered without a fight.
DEMORALIZATION	Моральды бұзу	KZ: майданда хаос орын алды, офицерлер бірінші болып қашуда EN: Chaos reigns on the front line, officers are fleeing first.
DISCREDITATION	Билікке немесе қоғамдық сенімге нұқсан келтіру	KZ: Армия басшылығы жағдайды бақыламайды EN: The army leadership has lost control of the situation.
INTIMIDATION	Қорқыныш немесе қауіп атмосферасын құру	KZ: Шабуыл жаппай шығындармен бірге жүреді EN: The offensive will result in massive casualties.
HATE_INCITEMENT	Өшпенділікті немесе дұшпандықты қоздыру	KZ: Олар -ұлттың жауы EN: They are the enemies of our nation.
PANIC_CREATION	Дүрбелең мен дабылды тарату	KZ: Барлық оқ-дәрі қоймалары жарылды EN: All ammunition depots have been blown up.
PROVOCATION	Жанжал немесе кек алу агрессиясын тудыру	KZ: Олар бітімгершілікті бірінші болып бұзды EN: They were the first to break the ceasefire.
AUTHORITY_ UNDERSCORE	Билікті манипуляциялық түрде баса көрсету немесе бұзу	KZ: Жоғары қолбасшылық осындай маңызды сәтте үнсіз қалады EN: High command remains silent in such a critical moment.

Аннотация шеңберінде мәлімдеменің үнін және оның мақсатты аудиторияға ықтимал психологиялық әсерін түсіру үшін эмоционалды бағалау мүмкіндігі (EMO_EVAL) енгізілді. Бұл жіктеу риторикалық стратегияларды және мәтіндердегі эмоционалды қарқындылық дәрежесін анықтауға мүмкіндік береді (Кесте 2).

Кесте 2 – EMO_EVAL (Emotional Evaluation)

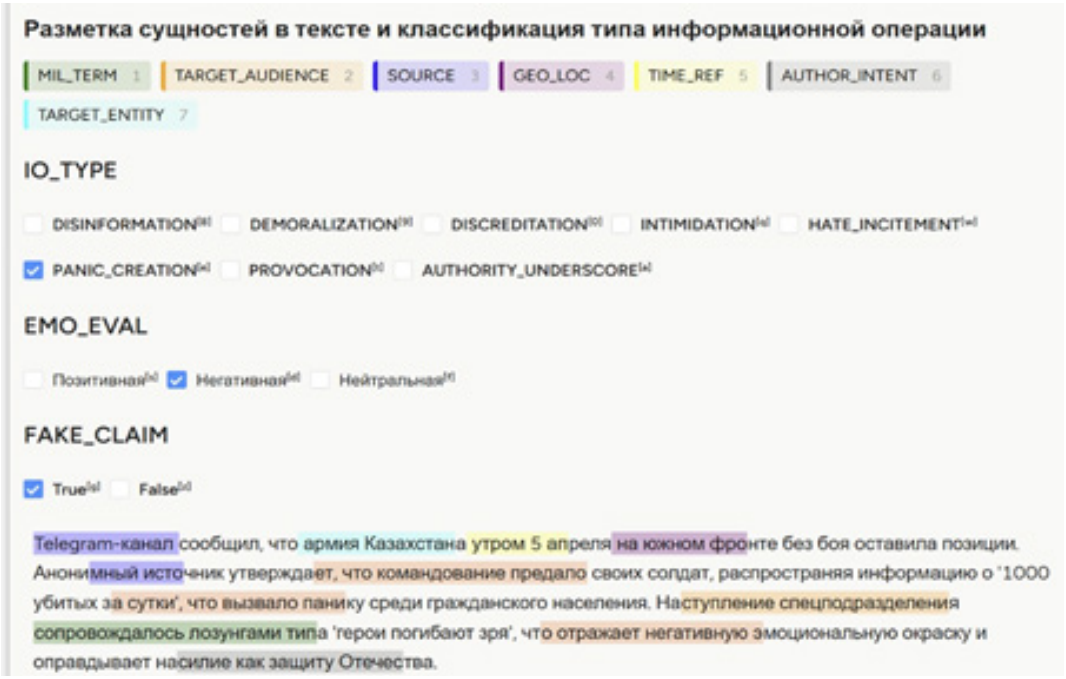
Категория	Сипаттамасы	Мысалдар
Оң пікір	Қолдау, қолдау немесе үміт тудыратын мәлімдемелер	RU: батырлық қорғаныс EN: Heroic defense
Теріс пікір	Қорқыныш, үрей, агрессия немесе менсінбеу көріністері	RU: Ұятпен позициядан айырылу EN: Shameful surrender of positions
Бейтарап	Эмоционалды бояусыз ақпараттық мәлімдемелер	RU: операция барысында екі батальонның күші жұмылдырылды EN: Two battalions were deployed during the operation



Сурет 3 – Корпустың аннотация процесі

Процесті оңтайландыру және топ мүшелері арасында бірізділікті қамтамасыз ету Үшін Label Studio интерфейсі көп пайдаланушы ынтымақтастығын орнату арқылы конфигурацияланды (Сурет 3). Бұл тәсіл бір корпусты бірнеше аннотаторлармен параллель аннотациялауға мүмкіндік берді, бұл аннотацияның жұмыс процесін жеделдетіп қана қоймай, сонымен қатар тұрақты тексеру және топ ішіндегі ішкі кері байланыс механизмдері арқылы жалпы сапаны жақсартты. Аннотация шешімдерін дәйекті диалог пен салыстыруға ықпал ете отырып, бұл қондырғы аннотаторлар арасындағы жоғары келісімді сақтауға және нақты уақыт режимінде түсініксіздіктерді шешуге көмектесті.

Ұсынылған суретте ақпараттық операцияларды автоматты түрде анықтау және жіктеу бойынша зерттеулер шеңберінде мәтіндік аннотацияның орыс тіліндегі мысалы келтірілген (Сурет 4). Қолданылатын белгілеу схемасы деструктивті ақпараттық әсердің әртүрлі аспектілеріне сәйкес келетін негізгі субъектілер мен санаттарды қамтиды (Кесте 3).



Сурет 4 – Нысанға аннотация жасау процесі

Кесте 3 – Деструктивті ақпараттық әсердің әртүрлі аспектілеріне сәйкес келетін негізгі субъектілер мен санаттар (орыс және ағылшын тілдерінде)

Мәні	Мысалдар	Аннотация
MIL_TERM	RU: массированный артиллерийский удар EN: massive artillery strike	[[массированный артиллерийский удар]] MIL_TERM
	RU: казахстанские военные подразделения EN: Kazakhstani military units	[[казахстанские]] GEO_LOC [[военные подразделения]] MIL_TERM
AUTHOR_ INTENT	RU: С целью дискредитации командования EN: With the aim of discrediting the command	С целью [[дискредитации]] AUTHOR_ INTENT: DISCREDIT [[командования]] TARGET_ENTITY
	RU: Направлено на дезориентацию населения EN: Aimed at disorienting the population	Направлено [[на дезориентацию]] AUTHOR_ INTENT: DISINFORMATION [[населения]] TARGET_AUDIENCE
TARGET_ AUDIENCE	RU: Обращение к русскоязычной аудитории EN: Appeal to the Russian-speaking audience	Обращение к [[русскоязычной аудитории]] TARGET_AUDIENCE
	RU: Предупреждение для семей военнослужащих EN: Warning for families of military personnel	Предупреждение для [[семей военнослужащих]] TARGET_AUDIENCE

3-кестенің жалғасы

TARGET_ ENTITY	RU: Обвинения в адрес правительства EN: Accusations against the government	Обвинения в адрес [[правительства]] TARGET_ENTITY
	RU: Недовольство военным руководством EN: Discontent with military leadership	Недовольство [[военным руководством]] TARGET_ENTITY
GEO_LOC	RU: Жанаозенский регион EN: Zhanaozen region	[[Жанаозенский регион]] GEO_LOC
	RU: У границ Казахстана EN: Near Kazakhstan’s borders	У границ [[Казахстана]] GEO_LOC
SOURCE	RU: По данным телеграм-канала «WarNews» EN: According to the Telegram channel “WarNews”	По данным [[телеграм-канала «WarNews»]] SOURCE
	RU: Заявление Министерства обороны EN: Statement by the Ministry of Defense	Заявление [[Министерства обороны]] SOURCE
TIME_REF	RU: утром 24 февраля EN: In the morning of February 24	[[утром]] TIME_REF [[24 февраля]] TIME_ REF
	RU: во время оккупации Крыма EN: During the occupation of Crimea	Во время [[оккупации]]MIL_TERM [[Крыма]] GEO_LOC

Аннотация әр айтылымның семантикасын және оның ақпараттық манипуляциядағы рөлін ескере отырып жүргізілді. Ол оқу корпусын құруға бағытталды, бірақ қолмен тегтеу субъективтілікке тәуелді. Әсіресе AUTHOR_INTENT жапсырмасы терең контекстік талдауды талап ететін екіұштылығымен ерекшеленді. Дегенмен, қолмен аннотация ақпаратқа әсер етудің негізгі заңдылықтарын дәл анықтауға мүмкіндік берді және болашақ автоматтандырудың негізін қалады. Аннотацияның сапасын бағалау үшін аннотаторлар арасындағы Келісімнің (IAA) метрикасы есептелді [37,38] аннотация жасаушылардың пайымдауларының жүйелілік дәрежесін көрсетеді.

Бұл құрылым мәтіндерді мазмұндық деңгейде ғана емес, сондай-ақ прагматика, ниеттілік және мақсатты әсер ету тұрғысынан талдауға мүмкіндік береді. Белгілеуді домен сарапшылары орындап, деректер Label Studio-дан JSON форматында экспортталды, ол spaCy, HuggingFace Трансформаторлары, Flair сияқты NLP құрылымдарымен үйлесімді. Сондай-ақ, CoNLL және CSV форматтарына экспорттау опциялары енгізіліп, дәйекті таңбалау және жіктеу тапсырмаларын қолдауға мүмкіндік жасалды.

Нәтижелер мен талқылау

Аннотация процесінен кейін 1000 мәтіннен (75000-нан астам таңбалауыштан) тұратын мамандандырылған корпус құрылды. Бұл мәтіндер ашық ақпарат көздерінен, негізінен Telegram жаңалықтар арналарынан және әскери тақырыптарға қатысты медиа ресурстардан жинақталған.

Қарастырылған корпус 8000 аннотацияланған нысанды қамтиды және 10 негізгі белгіге негізделген: алты нысанға қатысты (MIL_TERM, GEO_LOC, TIME_REF, SOURCE, TARGET_ ENTITY, TARGET_AUDIENCE) және төрт прагматикалық (IO_TYPE, AUTHOR_INTENT, EMO_EVAL, FAKE_CLAIM).

Бұл мәтін ақпараттық шабуылдарға тән заңдылықтарды көрсетеді, олар баяндауды автоматты тануға негіз бола алады. Белгілер арасындағы тұрақты корреляциялар анықталды: мысалы, жалған мәлімдеме 48 рет жалған ақпаратпен, ал 39 рет дүрбелең тудырумен бірге кездеседі. Автор ниеті көбіне бірнеше ақпараттық операция түрі (IO_TYPE) бар хабарла-

маларда кездеседі, бұл ақпараттық шабуылдардың гибридті сипатын көрсетеді. Сондай-ақ, нысаналы тұлғаның беделін түсіру мен нысаналы аудиторияны деморализациялау арасында байланыс бар. Бұл заңдылықтар агрессивті хабарламалардың ішкі семантикалық құрылымын айқын көрсетеді (4-кесте).

Кесте 4 – Ақпараттық ықпал категорияларының лингвистикалық маркерлері

Категория	Талдау нәтижелері	Мысалдар
DEMORALIZATION	Лексемалардың жоғары жиілігі	KZ: мағынасыз, ешкім көмектеспейді, бітті EN: meaningless, no one will help, it's all ove
DISCREDITATION	Қайталанатын айыптау конструкциялары	KZ: олар сатқындық жасады, шындықты жасырды, адамдарды шығармады EN: betrayed, hid the truth, did not evacuate people
DISINFORMATION	Асыра сілтелген сандар	KZ: мыңдаған техниканың жойылуы туралы хабарламалар EN: Reports of the alleged destruction of thousands of pieces of equipment
Ескерту: авторлар құрастырған		

Аннотацияланған корпустың сенімділігі мен репродуктивтілігін қамтамасыз ету үшін аннотация аралық келісім бағаланды. Үш аннотация жасаушы 1000 мәтінді өз бетінше белгіледі. Бағалау нысандық тегтерге (мысалы, MIL_TERM, GEO_LOC, TIME_REF) және категориялық тегтерге (IO_TYPE, AUTHOR_INTENT, FAKE_CLAIM және т.б.) жүргізілді.

Категориялық белгілерді бағалау мақсатында он екі аннотация жасаушы Cappa Co-hen (к) келісім коэффициенті қолданылды, бұл әдіс кездейсоқ сәйкестікке мүмкіндік береді (5-кесте). Талдау екі режимде жүргізілді:

(a) Ішінара Қабаттасу: Рұқсат етілген таңдалған фрагменттердің ішінара қабаттасуын болжауда қолданылған.

(b) Ішінара Қабаттасу: алынып тасталды аннотациялардың толық сәйкес келуі үшін қатаң талаппен пайдаланылды.

Кесте 5 – Талдау режимдерінің нәтижесі

Категория	Ішінара Қабаттасу: Рұқсат	Ішінара қабаттасу: Алынып тасталды
MIL_TERM (F_1)	0.91	0.84
GEO_LOC (F_1)	0.88	0.79
TIME_REF (F_1)	0.86	0.80
AUTHOR_INTENT (к)	0.72	0.68
TARGET_AUDIENCE (к)	0.76	0.73
TARGET_ENTITY (к)	0.79	0.75
IO_TYPE (к)	0.82	0.80
EMO_EVAL (к)	0.85	0.83
FAKE_CLAIM (к)	0.88	0.87
Ескерту: авторлар құрастырған		

Аннотация нәтижелері жоғары келісімді көрсетті: MIL_TERM, GEO_LOC және TIME_REF белгілері 0,85-тен жоғары F1 ұпайына жетті, ал FAKE_CLAIM және IO_TYPE бойынша к > 0,80 болды. Ең төменгі келісім AUTHOR_INTENT белгісінде тіркеліп, прагматикалық бағытты анықтаудың субъективтілігін көрсетті.

Telegram жаңалықтар арнасында әскери мазмұнға пилоттық талдау жасалды. Хабарламалар жалған ақпарат пен эмоционалдық қысымды анықтау үшін IO_TYPE және AUTHOR_INTENT бойынша белгіленді. 100 хабарлама OntoIO-BERT моделімен жартылай автоматты аннотацияланып, сарапшылармен тексерілді. Мысалы, «АВТОРДЫҢ НИЕТІ: ҚОРҚЫТУ» және «ЖАЛҒАН МӘЛІМДЕМЕ: ШЫНДЫҚ» ретінде жіктелген хабарламалар бейбіт тұрғындар арасында дүрбелең тудыру қаупін көрсетті (5-сурет).

Сурет 5 – Түсіндірілген Telegram әлеуметтік желісіндегі хабарламаның мысалы

Талдау нәтижесінде хабарламалардың 47%-ы жала жабуға, ал 12%-ы үйлестірілген жалған ақпаратқа жататыны анықталды. Бұл корпус әлеуметтік желілердегі дұшпандық ақпаратты бақылауға негіз бола алады. Сарапшылар аннотацияланған деректерді пайдаланып, эмоционалды және қайталанатын мазмұнды жала жабу науқандары ретінде бағалады.

Осы зерттеуде құрылған корпус қолданыстағы деректер жиынтығындағы маңызды олқылықты толтырып, әскери ақпараттық операцияларды талдауға арналған жаңа ресурс болып табылады. Лексикалық-семантикалық немесе риторикалық сипаттамалармен шектелетін Mil-tac, DiPLo-MAT немесе қытай тіліндегі оқиғаларды шығаруға бағытталған CMNEE-ден айырмашылығы, біздің әдіс нысан деңгейіндегі аннотацияны IO_TYPE, AUTHOR_INTENT, EMO_EVAL және FAKE_CLAIM сияқты интерпретациялық санаттармен бірегей түрде ұштастырады. Сонымен қатар, қазақ тілі тәрізді ресурстармен жеткіліксіз қамтылған тілдерді Telegram арқылы таратылатын нақты әскери дискурстармен біріктіру корпустың көптілді, геосаяси және стратегиялық коммуникацияларды зерттеудегі практикалық әрі академиялық құндылығын арттырады.

Дәстүрлі корпустардан айырмашылығы, біздің аннотация жүйеміз соғысқа тән қақтығыстарға, манипуляцияларға және жоғары эмоциялық мазмұнға бейімделген, полисемия, риторикалық стратегиялар мен прагматикалық маркерлерді қамтиды.

Корпустың негізгі артықшылықтары мынадай:

1. Қолмен аннотациялау жоғары сапалы бақылау, екіұшты өрнектерді түсіндіру икемділігі және жасырын риторикалық заңдылықтарды анықтауды жеңілдетті.

2. Аннотация схемасы семантикалық, синтаксистік және прагматикалық көпқабатты құрылымды қамтып, IO_TYPE және EMO_EVAL сияқты маңызды категорияларды қосады.

3. Іске асыру Label Studio интерфейсі арқылы модульдік сипатта, визуализацияны айқын және деректерді бірнеше пішімге экспорттауды қамтамасыз етеді.

4. Пилоттық валидация аннотация нұсқаулығының дәйектілігі мен сенімділігін растап, аннотациялаушылар арасындағы келісімнің қанағаттанарлық екенін көрсетті.

Қорытынды

Осы мақалада көптілді әскери дискурстағы ақпараттық операцияларды анықтау және жіктеу мақсатында әзірленген Multi_mil деп белгіленген көп деңгейлі аннотацияланған корпус көрсетілген. Wikibias, Mil-TAC немесе CMNEE сияқты қолда бар ресурстардан айырмашылығы, әзірленген корпус аннотацияларды бірқатар негізгі бағыттар бойынша біріктіреді: әсер ету түрі (IO_TYPE), аталған нысандар, эмоционалды бояу (EMO_EVAL) және сенімсіздік белгілері (FAKE_CLAIM). Барлық белгілер бір семантикалық схема бойынша рәсімделетіні, осылайша тұтастық пен интерпретацияны қамтамасыз ететіні анық. Ұсынылған корпус қазақ, орыс және ағылшын тілдеріндегі мәтіндерді қамтиды. Аннотация құрылымы манипуляциялық мазмұнның лингвистикалық және прагматикалық қырларын модельдеуді жеңілдетеді.

Қаржыландыру туралы ақпарат: Бұл жұмыс Қазақстан Республикасы Ғылым және жоғары білім министрлігіне қарасты Ғылым комитетінің қолдауымен жүргізілді (AP 23484329).

ӘДЕБИЕТТЕР

1 Yerimbetova, A., Sakenov, B., Sambetbayeva, M., Daiyrbayeva, E., Berzhanova, U., & Othman, M. Creating a Parallel Corpus for the Kazakh Sign Language and Learning. Applied Sciences, 15(5), 2808 (2025). https://doi.org/10.3390/app15052808

2 Sadirmekova, Zh., Daiyrbayeva, E., Karbozova, I., & Bekbolatov, S. Development of a method for automatic extraction of ontology entity names from natural language texts. In UBMK 2024 - Proceedings: 9th International Conference on Computer Science and Engineering, IEEE, 74–79 (2024). https://doi.org/10.1109/UBMK63289.2024.10773503

3 Mussiraliyeva, S., Bolatbek, M., Zhumakhanova, A., Medetbek, Z., & Sagynay, M. Comparative analysis of machine learning algorithms to identify extremist texts in the Kazakh language. Scientific Journal of Astana IT University, 14(14), 71–90 (2023). https://doi.org/10.37943/14DKRN4681

4 Mussiraliyeva, S., Bolatbek, M., Zhumakhanova, A., Baispay, G., & Medetbek, Z. Creating a model of semantic analysis of extremist texts in the Kazakh language. Journal of Mathematics, Mechanics and Computer Science, 121(1), 110–121 (2024). https://doi.org/10.26577/JMMCS2024121111

5 Rzayeva, L., Pogolovkin, D., & Myrzatay, A. Development of a modular messaging analysis service based on NLP for digital forensics. News of the NAS RK. Series of Physics and Mathematics, 2, 212–233 (2025). https://doi.org/10.32014/2025.2518-1726.354 (in Russian).

6 Modzelewski, A., Da San Martino, G., Savov, P., Wilczyńska, M. A., & Wierzbicki, A. MIPD: Exploring manipulation and intention in a novel corpus of disinformation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 1234–1245 (2024). ACL.

7 Klein, I. Information operations and influence campaigns. IEEE Phoenix Computer Society, February 2024.

8 Sambetbayeva, M., Nekessova, A., Yerimbetova, A., Bayangali, A., Kaldarova, M., Telman, D., & Smailov, N. A Multi-Level Annotation Model for Fake News Detection: Implementing Kazakh-Russian Corpus via Label Studio. Big Data and Cognitive Computing, 9(8), 215 (2025). https://doi.org/10.3390/bdcc9080215

9 Lin, Y., Wang, H., & Celikyilmaz, A. A survey on recent advances in named entity recognition from deep learning models. ACM Computing Surveys, 55(2), 1–40 (2023). https://doi.org/10.1145/3514221

10 Derczynski, R., Bontcheva, L., & Roberts, K. Broad Twitter corpus: A diverse named entity recognition dataset. In Proceedings of COLING 2016, 1169–1179 (2016).

11 Barrón-Cedeño, A., Rosso, P., & Yu, S. Fine-grained analysis of propaganda in news articles. In Proceedings of EMNLP 2019 , 564–573 (2019).

12 Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of EMNLP 2017, 2921–2927 (2017).

13 Horn, C., Wiegand, M., & Klakow, D. Towards a multidimensional model of media bias in news articles. In Proceedings of COLING 2018, 498–509 (2018).

14 Reisigl, M., & Wodak, R. The discourse-historical approach. In Methods of Critical Discourse Studies, 3rd ed., 23–61 (2015).

15 Van Dijk, T. Critical discourse studies: A sociocognitive approach. In Methods of Critical Discourse Analysis , 62–86 (2011).

16 Zeldes, A. The GUM corpus: Creating multilayer resources in a university setting. Language Resources and Evaluation, 51, 581–612 (2017). https://doi.org/10.1007/s10579-016-9343-4

17 Al-Rawi, A. Framing the Syrian conflict on Twitter. Global Media and Communication, 10 (2), 153–170 (2014). https://doi.org/10.1177/1742766514536022

18 Vego, E. Effects-based operations: A critique. Joint Forces Quarterly, 41, 51–57 (2006).

19 Dandeker, C. Military language and strategic doctrine. Sociological Review, 54(3), 581–598 (2006). https://doi.org/10.1111/j.1467-954X.2006.00628.x

20 Röttger, P., Schröder, M., Grotov, A., & Augenstein, I. Harmful but legal: Cognitive framing and large language models in online information warfare. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024). ACL.

21 Walker, C., Strassel, S., Medero, J., & Maeda, K. ACE 2005 multilingual training corpus. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2006T06

22 Wang, X., Wang, Z., Han, X., Jiang, W., Han, R., Liu, Z., Li, J., Li, P., Lin, Y., & Zhou, J. MAVEN: A massive general domain event detection dataset. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),1652–1671 (2020).

23 Li, X., Li, F., Pan, L., Chen, Y., Peng, W., Wang, Q., Lyu, Y., & Zhu, Y. DuEE: A large-scale dataset for Chinese event extraction in real-world scenarios. In X. Zhu, M. Zhang, Y. Hong, & R. He (Eds.), Natural Language Processing and Chinese Computing, 534–545 (2020).

24 Ebner, S., Xia, P., Culkin, R., Rawlins, K., & Van Durme, B. Multi-sentence argument linking. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8057–8077 (2020).

25 Li, S., Ji, H., & Han, J. Document-level event argument extraction by conditional generation. In Proceedings of NAACL-HLT 2021, 894–908 (2021).

26 Tong, M., Xu, B., Wang, S., Han, M., Cao, Y., Zhu, J., Chen, S., Hou, L., & Li, J. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In Proceedings of NAACL-HLT 2022, 3970–3982 (2022).

27 Huang, H., Sun, J., Wei, H., Xiao, K., Wang, M., & Li, X. A dataset of domain events based on open-source military news. China Scientific Data, 8(1), 30 (2023). https://doi.org/10.11922/sciencedata.2023.0005.zh

28 Zhu, M., Xu, Z., Zeng, K., Xiao, K., Wang, M., Ke, W., & Huang, H. CMNEE: A large-scale document-level event extraction dataset based on open-source Chinese military news. In Proceedings of LREC-COLING 2024, 3367–3379 (2024).

29 Chen, L.-C., Chang, K.-H., & Yang, S.-C. Integrating corpus-based and NLP approach to extract terminology and domain-oriented information: An example of US military corpus. Acta Scientiarum. Technology, 44, 2022. https://doi.org/10.4025/actascitechnol.v44i1.60486

30 Li, H., Zhu, S.-C., & Zheng, Z. (2023). DiPlomat: A dialogue dataset for situated pragmatic reasoning. arXiv preprint. https://arxiv.org/abs/2306.09030

31 Label Studio Documentation. Label Studio annotator guide, 2023. https://labelstud.io/guide

32 International Organization for Standardization. ISO 24617-1:2012 – Language resource management – Semantic annotation framework (SemAF) – Part 1: Time and events (ISO-TimeML). https://www.iso.org/standard/37331.html

33 International Organization for Standardization. ISO 24617-6:2016 – Language resource management – Semantic annotation framework – Part 6: Principles of semantic annotation (SemAF Principles). https://www.iso.org/standard/60581.html

34 Wardle, C., & Derakhshan, H. Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe, 2017. https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/

35 Joint Chiefs of Staff. (2012). Joint publication 3-13: Information operations. U.S. Department of Defense. https://irp.fas.org/doddir/dod/jp3_13.pdf

36 President of the Republic of Kazakhstan. (2023, March 20). On the approval of the information doctrine of the Republic of Kazakhstan (Decree No. 145). https://adilet.zan.kz/rus/docs/U2300000145

37 Schmitt, M. N. (Ed.). (2017). Tallinn Manual 2.0 on the international law applicable to cyber operations. Cambridge University Press.

38 Artstein, R., & Poesio, M. Survey article: Inter-coder agreement for computational linguistics. Computational Linguistics, 34(4), 555–596 (2008). https://doi.org/10.1162/coli.07-034-R2

[1,2]Самбетбаева М.,
PhD, ассоциированный профессор, ведущий научный сотрудник,
ORCID ID: 0000-0001-9358-1614,
e-mail: madina_jgtu@mail.ru
[1,3]Еримбетова А.,
PhD, ассоциированный профессор, ведущий научный сотрудник,
ORCID ID: 0000-0002-2013-1513,
e-mail:aigerian8888@gmail.com
[1,2]Абдығалым Б.,
магистр, младший научный сотрудник, ORCID ID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com
[1,4]Дайырбаева Э.,
магистр, научный сотрудник, старший преподаватель,
ORCID ID: 0000-0002-4255-5456,
e-mail: nurbekkyzyelmira@gmail.com
[1]Тұрғанбаев А.,
инженер-программист, ORCID ID: 0000-0001-8989-7827,
e-mail: m1challenge@inbox.ru

[1]Институт информационных и вычислительных технологий КН МНВО РК,
г. Алматы, Казахстан
[2]Евразийский университет им. Л.Н. Гумилева, г. Астана, Казахстан
[3]META University, г. Алматы, Казахстан
[4]Казахский национальный исследовательский технический университет им. К.И. Сатпаева,
г. Алматы, Казахстан

ЯЗЫКОВАЯ МОДЕЛЬ, ОБЪЕДИНЯЮЩАЯ ОНТОЛОГИЧЕСКИЕ И КОРПУСНЫЕ ДАННЫЕ

Аннотация

Настоящее исследование направлено на анализ цифровых угроз в повседневном дискурсе через создание аннотированного корпуса из 1000 текстов, собранных из соцсетей и новостных платформ, связанных с военными и геополитическими событиями. Целью предлагаемого исследования является удовлетворение насущной потребности в анализе цифровых угроз в повседневном дискурсе путем создания аннотированного корпуса текстов с элементами информационных операций. Разработана многоуровневая схема аннотации, охватывающая тип воздействия, эмоциональную окраску, признаки дезинформации и авторские намерения (провокация, запугивание и др.). Аннотирование выполнено в Label Studio с контролем качества и учетом контекста; надежность подтверждена коэффициентом Каппа = 0,82. Модель Onto-IO-BERT в пилотном эксперименте показала F1 = 0,81, превзойдя базовые модели. Практическая применимость корпуса подтверждена анализом реальных Telegram-сообщений. В настоящей работе представлена многоуровневая

модель аннотации, адаптированная для анализа военных информационных операций в пространстве Казахстанского МВД, и созданный корпус является новым ресурсом для анализа военных информационных операций, заполнив важный пробел в существующих наборах данных. Представленный корпус содержит тексты на казахском, русском и английском языках. Корпус доступен по ссылке: https://github.com/baiangali/multi_mil .

Ключевые слова: военные термины, аннотированный корпус, информационные операции, NLP, анализ социальных сетей.

[1,2]Sambetbayeva M.,
PhD, Association Professor, leading researcher, ORCID ID: 0000-0001-9358-1614,
e-mail: madina_jgtu@mail.ru,
[1,3]Yerimbetova A.,
PhD, Association Professor, leading researcher, ORCID ID: 0000-0002-2013-1513,
e-mail:aigerian8888@gmail.com
[1,2]Abdygalym B.,
Doctoral student, junior researcher, ORCID ID: 0009-0001-8872-7428,
e-mail: bayangali.abd@gmail.com
[1,4]Daiyrbayeva E.,
Master, researcher, senior lecturer, ORCID ID: 0000-0002-4255-5456,
e-mail: nurbekkyzyelmira@gmail.com
[1]Turganbayev A.,
Software engineer, ORCID ID: 0000-0001-8989-7827,
e-mail: m1challenge@inbox.ru

[1]Institute of Information and Computational Technologies of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Almaty, Kazakhstan
[2]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[3]META University, Almaty, Kazakhstan
[4]Satbayev University, Almaty, Kazakhstan

A LANGUAGE MODEL INTEGRATING ONTOLOGICAL AND CORPUS-DATA

Abstract

This study addresses the urgent need to analyze digital threats in everyday discourse by constructing a 1,000-text annotated corpus from social media and news platforms covering military and geopolitical events. The purpose of the proposed study is to address the urgent need to analyze digital threats in everyday discourse by creating an annotated corpus of texts with elements of information operations. A multi-layered annotation scheme captures semantic actors and pragmatic features – including impact type, emotional tone, disinformation markers, and intent (e.g., provocation, intimidation). Annotation via Label Studio ensured flexibility, quality control, and context sensitivity, with inter-annotator reliability (Cohen's Kappa = 0.82) confirming consistency. In pilot experiments, the Onto-IO-BERT model achieved an F1-score of 0.81, outperforming baseline classifiers. Practical utility was validated through analysis of real Telegram messages. The framework is tailored for studying military information operations within Kazakhstan's Ministry of Internal Affairs and the created corps is a new resource for analyzing military information operations, filling a significant gap in the existing data set. The presented corpus contains texts in Kazakh, Russian and English. The corpus is openly accessible at: https://github.com/baiangali/multi_mil

Keywords: military terms, annotated corpus, information operations, NLP, social media analysis.

Received January 13, 2026; accepted April 20, 2026.