

Г.З. Зиятбекова¹, Ш.Н. Шекербаетова^{1*}, А.И. Акжолова¹,
М.М. Сыдыкова¹, Ж.Б. Полатова¹

¹Алматы технологиялық университеті, Алматы қ., Қазақстан
*e-mail: Sh-shrai@mail.ru

СУИЦИДТІК МІНЕЗ-ҚҰЛЫҚ БЕЛГІЛЕРІ БАР ҚАЗАҚ ТІЛІНДЕГІ МӘТІНДЕР МОДЕЛІН ЖӘНЕ ҚОСЫМШАСЫН ҚҰРУ

Аңдатпа

Мақала суицидтік мінез-құлық белгілері бар қазақ тіліндегі мәтіндер моделін құруға арналған. Есептеу лингвистикасы табиғи тілді өңдеуді, яғни Natural Language Processing (NLP) қамтиды. Осы зерттеу қорытындысы NLP міндеттерін шешуде қолданылары сөзсіз, айталық, мәтін корпустарын анықтауда, білім салаларында, деректерді іздеу жүйелерінде, машиналық аудармада және басқа да салаларда. Мақалада қарастырылған басты мәселе машиналық және терең оқыту әдістерін қолданып, ашық акпарат көздеріндегі мәліметтерден өз-өзіне қол жұмсау әдеттерін бірден байқау болып табылады. Қазақ тіліндегі мәтіндерді талдау арқылы тиімді жіктеу модельдерін әзірлеу және оларды салыстыру көзделеді. Мемлекеттік тілдегі мәтіндерді қайта баптаудың айырмашылықтары, соның арасындағы морфологиялық қиындықтар мен агглютинация құбылысы назарға алынатын болады. Ізденіс жұмысы ұлттық тіліміздегі өзін өлімге қию бағытындағы деректерді айқындауға бейімделген машиналық оқыту қалыптарын дайындау, сонымен қатар өзара әсерлесуін қарастыра отырып көзге түседі. Жасанды интеллектті пайдалана отырып Интернет көздерін мұқият қадағалауды жетілдіру, дағдарыстық жағдайларға ерте жауап беруге мүмкіндік береді.

Түйін сөздер: LSTM, KazBERT модельдері, NLP, RNN, суицидтік мінез-құлық, машиналық оқыту.

Г.З. Зиятбекова¹, Ш.Н. Шекербаетова¹, А.И. Акжолова¹, М.М. Сыдыкова¹, Ж.Б. Полатова¹

¹Алматинский технологический университет, г. Алматы, Казахстан

СОЗДАНИЕ МОДЕЛИ И ПРИЛОЖЕНИЯ К ТЕКСТАМ НА КАЗАХСКОМ ЯЗЫКЕ С ПРИЗНАКАМИ СУИЦИДАЛЬНОГО ПОВЕДЕНИЯ

Аннотация

Статья предназначена для создания модели текстов на казахском языке с признаками суицидального поведения. Вычислительная лингвистика включает в себя обработку естественного языка, то есть обработку естественного языка (NLP). Результаты данного исследования, несомненно, будут использованы для решения задач обработки естественного языка, например, в анализе текстовых корпусов, в предметных областях, в системах поиска данных, в машинном переводе и в других областях. Основная проблема, рассматриваемая в статье, — это оперативное выявление суицидальных наклонностей на основе данных из открытых источников с использованием методов машинного и глубокого обучения. Цель состоит в разработке и сравнении эффективных моделей классификации на основе анализа казахских текстов. Будут учтены различия в адаптации текстов на государственном языке, включая морфологические трудности и феномен агглютинации. Научная работа сосредоточена на разработке моделей машинного обучения, адаптированных для выявления закономерностей суицидального поведения на нашем национальном языке, а также на изучении их взаимодействия. Улучшение мониторинга интернет-источников с помощью искусственного интеллекта позволит оперативно реагировать на кризисные ситуации.

Ключевые слова: модели LSTM, KazBERT, NLP, RNN, суицидальное поведение, машинное обучение.

G. Ziyatbekova¹, Sh. Shekerbayeva¹, A.I. Akzholova¹, M. Sydykova¹, Zh. Polatova¹

¹Almaty Technological University, Almaty, Kazakhstan

DEVELOPMENT OF A MODEL AND APPLICATION FOR KAZAKH-LANGUAGE TEXTS EXHIBITING SIGNS OF SUICIDAL BEHAVIOR

Abstract

This article aims to develop a model for Kazakh-language texts exhibiting signs of suicidal behavior. Computational linguistics encompasses natural language processing—that is, natural language processing (NLP). The results of this study will undoubtedly be used to address natural language processing tasks—for example, in the analysis of text corpora, in specific subject domains, in data retrieval systems, in machine translation, and in other areas. The main problem addressed in the article is the prompt detection of suicidal tendencies based on data from open sources using machine learning and deep learning methods. The goal is to develop and compare effective classification models based on the analysis of Kazakh texts. Differences in the adaptation of texts in the state language will be taken into account, including morphological difficulties and the phenomenon of agglutination. The research focuses on developing machine learning models tailored to identify patterns of suicidal behavior in our national language, as well as on examining the interplay between them. Improving the monitoring of online sources using artificial intelligence will enable a rapid response to crisis situations.

Keywords: LSTM, KazBERT models, NLP, RNN, suicidal behavior, machine learning.

Кіріспе

Бүгінгі таңда ашық интернет көздері мен көптеген қолжетімді сайттарда табиғи тіл жайында ақпараттың көлемі айтарлықтай артуына байланысты, есептеу лингвистикасындағы зерттеулер мен даму жұмыстарының маңызы барған сайын артып келеді. Бізге мәлім болғандай, есептеу лингвистикасы – салыстырмалы түрде жаңа ғылыми бағыт әрі ол ақпараттық телекоммуникациялар саласының түпқазығы іспеттес. Суицид бүкіл әлемде өзекті әлеуметтік қиындық, адамзат өмірінде маңызды рөл атқарады. Бүкіләлемдік денсаулықты қорғау мекемесінің хабарламасына сәйкес жылда жеті жүз мыңдай азамат жеке басына зиян келтіреді. Қазақстанда бұл мәселе алаңдаушылық тудырып отыр: жыл сайын шамамен 4000 адам суицидтен қайтыс болады, бұл күн сайын 10 оқиғаға тең [1].

Соңғы жылдары аяқталған суицид санының 7%-ға азайғанына қарамастан, мәселе өзекті болып қала береді. 2022 жылы 3,7 мың жағдай тіркелді, ал өз-өзіне қол жұмсау әрекеттері 17,5%-ға төмендеді. Алайда, суицид деңгейі ер адамдар арасында жоғары – барлық жағдайлардың 78,9%-ы ер азаматтарға тиесілі. Ең осал топ – 30-54 жас аралығындағы адамдар, олар жалпы оқиғалардың 54,9%-ын құрайды [2]. Әсіресе балалар мен жасөспірімдер арасында суицидтік мінез-құлықтың қалыптасуына әлеуметтік желілердің әсері алаңдатады. Зерттеулер көрсеткендей, желідегі ақпарат көмелетке толмағандардың психологиялық жағдайына кері әсер етіп, суицидтік ойларға түрткі болуы мүмкін [3]. Ол үшін болашақ ұрпаққа зиян тигізбейтіндей мүмкіндік жасай отырып, тәртіп бұзушылықты болдырмауға арналған оң шешімдер қабылдау, сонымен қатар адам психикасына оң әсер ететін іс-шараларды күшейту керек. Жұмыстың мақсаты – машиналық оқытуды пайдалана отырып, суицидтік мінез-құлық белгілеріне әлеуметтік желі мәтіндерін автоматты түрде жіктеу әдістерін әзірлеу және салыстырмалы талдау. Бұл мақсатқа жету үшін табиғи тілді өңдеу әдістері, деректерді жинау, алдын ала өңдеу және векторизациялау, сондай-ақ, логистикалық регрессия, LSTM сияқты машиналық және терең оқыту алгоритмдері қолданылды. Зерттеу бірнеше модельдерді, соның ішінде KazBERT және басқа нейрондық желі архитектурасын оқыту мен бағалауды, қазақ тіліндегі мәтіндер үшін ең тиімді шешімді таңдауды қамтиды. Бұл зерттеудің ғылыми жаңалығы суицидтік мінез-құлық белгілері бар қазақ тіліндегі мәтіндерді жіктеуге арналған машиналық оқыту үлгілерін жасау және салыстырмалы талдау болып табылады. Қолданыстағы шешімдерден айырмашылығы, жұмыста әртүрлі архитектуралардың, соның ішінде қазақ тіліне бейімделген трансформерлық үлгілердің тиімділігі зерттеледі. Мұндай талдауды жүргізу берілген контексте мәтіндік деректерді өңдеу мен жіктеудің ең тиімді

әдістерін анықтауға мүмкіндік береді. Сондай-ақ, зерттеуде қазақ тілінің морфологиялық күрделілігі мен агглютинациясы ескерілген, бұл оның автоматты түрде өңделуін аса қиындатады. Алдын ала өңделген деректерге оқытылған тәсіл болашақта қазақша мәтіндерді талдаудың басқа да жүйелерін құру үшін қолданылуы мүмкін.

Уақыт өте келе, машиналық оқытудың классикалық үлгілері кеңінен қолданыла бастады. Бұл модельдер алдын ала белгіленген мәтіндер жиынтығында оқытылып, мәтіндерді белгілі бір класқа жатқызуды үйренеді. Олардың ішінде Аңғал Байес (Naive Bayes) классификаторы қарапайымдылығымен және мәтіндік деректерде жиі жақсы нәтиже көрсетуімен ерекшеленді. Тірек Векторлар Машинасы (Support Vector Machines, SVM) жоғары өлшемді белгілер кеңістігінде тиімді жұмыс істеп, күрделірек шекараларды анықтауға қабілетті болды. Логистикалық Регрессия (Logistic Regression) да өзінің интерпретациялануының оңайлығымен және тұрақты нәтижелерімен танымал болды. Бұл модельдер әдетте мәтінді алдын ала өңделген белгілер жиынтығы, мысалы, TF-IDF (Term Frequency-Inverse Document Frequency) немесе сөздер қапшығы (Bag-of-Words) арқылы ұсынылған түрде қабылдайды. Олар күрделі емес заңдылықтарды анықтауда және белгілі бір деңгейде жіктеу сапасын қамтамасыз етуде жақсы нәтижелер көрсеткенімен, тілдің терең семантикалық мағынасын және контекстік байланыстарын толық түсінуде шектеулі болды.

Зерттеу әдіснамасы

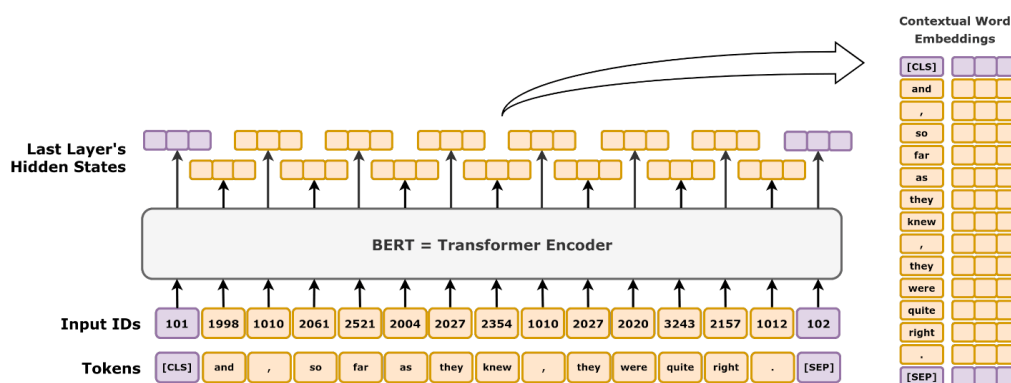
Модельдер мен тәсілдер. Терең оқыту технологияларының қарқынды дамуы бұл саланың ілгерілеуіне жол ашты. Алдымен, қайталанатын нейрондық желілер (RNN), әсіресе олардың жетілдірілген түрлері – ұзақ қысқа мерзімді жады (LSTM) және басқарылатын рекуррентті бірлік (GRU) архитектуралары мәтіннің реттік табиғатын ескеруге және сөздер арасындағы ұзақ мерзімді тәуелділіктерді анықтауға қабілеттілігін көрсетті. Бұл модельдер сөздердің векторлық көріністерін (word embeddings) пайдаланып, мәтіннің мағынасын біртіндеп қалыптастырады, және бұл оларға контексті жақсырақ түсінуге мүмкіндік берді. Алайда, RNN архитектуралары ұзын мәтіндерді өңдеуде және параллель есептеулерде белгілі бір қиындықтарға тап болды. Нағыз серпілістің орнауына Трансформер архитектурасының (мысалы, BERT) пайда болуы себепші болды. Бұл модельдер "назар аудару" (attention) механизмдеріне, әсіресе "өзіне-өзі назар аудару" (self-attention) механизміне негізделген. Бұл механизмдер модельге сөйлемдегі әрбір сөзді өңдеу кезінде сол сөйлемдегі басқа сөздердің маңыздылығын бағалауға мүмкіндік береді, осылайша сөздер арасындағы күрделі контекстік байланыстарды тиімдірек түсінеді. BERT алдын ала үлкен мәтіндік корпустарда екі негізгі тапсырма бойынша оқытылады: маскаланған тіл моделі (Masked Language Model) және келесі сөйлемді болжау (Next Sentence Prediction). Осындай алдын ала оқытудың арқасында BERT тілдің терең заңдылықтарын үйренеді және оны әртүрлі NLP тапсырмаларына, соның ішінде психикалық денсаулыққа қатысты мәселелерді, суицидтік ойларды анықтауға оңай бейімдеуге (fine-tuning) болады. Трансформерлердің параллель өңдеу қабілеті де олардың RNN-ге қарағанда артықшылығы болып табылады. Көптеген осындай зерттеулер мен оларға қажетті деректер жинақтары, әсіресе ағылшын тілі үшін, Kaggle, GitHub сияқты ашық зерттеу платформаларында қолжетімді. Бұл платформалар зерттеушілерге дайын деректермен жұмыс істеуге, модельдерді салыстыруға және саланың дамуына ықпал ететін ғылыми зерттеулердің қарқынды жүргізілуіне мүмкіндік береді. Заманауи тәсілдер көбінесе автоматтандырылған жүйелер мен адамның сараптамалық тәжірибесі арасындағы синергияны қалыптастыруға бағытталған. Автоматтандырылған құралдар алдын ала скрининг жүргізіп, ықтимал қауіпті немесе назар аударуды қажет ететін мазмұнды анықтай алады. Содан кейін бұл ақпарат талдау және нақты шешім қабылдау үшін сарапшы мамандарға (мысалы, психологтар, тәжірибелі модераторлар, дағдарыс орталықтарының қызметкерлері) беріледі. Мұндай гибриді тәсіл ("human-in-the-loop") деструктивті мазмұнды бақылаудың тиімділігін арттырып, жауапкершілік деңгейін жоғарылатуға мүмкіндік береді және қазіргі таңда ең перспективалы бағыттардың бірі болып саналады [4].

Әдебиеттерге шолу. Тақырыпты модельдеудің ең көп тараған әдістері Blei және Дирихле ұсынған жасырын бөлу (LDA) әдісі болып табылады [5]. Гриффитс пен Стейверс [6] зерттеулері көрсеткендей, бұл әдіс тіпті қысқа әлеуметтік медиа мәтіндерінде жасырын тақырыптық құрылымдарды ашуда тиімді.

Машиналық оқыту әдістері қазіргі заманғы әлеуметтік медиа мәтіндік мазмұнды талдау тәжірибесінде орталық орынға ие болды, зерттеу жүргізушілерге қолданушы деректерін дереу сараптау әрі интерпретация жасауға мықты құрылғыларды қолдануын талап етеді. Юрафский және Мартин [7] көрсеткендей, осы тәсілдердің өзі көптеген басты талаптармен сыналды, яғни әрбір сынақ нақты әрі бейімділікті дамытуға ықпал етті. Мәтінді талқылаудың артуының алғашқы белестерінде статистикаға қатысты тәсілдер, қалыпты түрдегі машиналық оқытудың сан алуан түрі көп болды. Де Чоудхури [8] зерттеулері Naive Bayes классификаторының екілік мәтінді жіктеу тапсырмаларында жақсы орындайтынын көрсетті. Қолдау векторлық машинасы (SVM), Ким [9] бойынша, депрессиялық мазмұнды анықтау тапсырмаларында 82%-ға дейінгі дәлдікті көрсетті. Кездейсоқ ормандар шамадан тыс орнатуға төзімді екенін дәлелдеді. Чжан мен Уоллес [10] анықтаған конволюционды нейрондық желілер (CNN) мәтіндердің эмоционалды бояуын талдау тапсырмаларында шамамен 86% дәлдікті көрсетеді. Hochreiter және Schmidhuber [11] зерттеулері LSTM желілерінің тиімділігін растады, суицидтік ниетті анықтауда 91%-ға дейінгі дәлдікке қол жеткізді. Трансформатор архитектурасы, алғаш рет Васвани және т.б. табиғи тілді өңдеуде төңкеріс жасады. Devlin және т.б. [12] BERT үлгісі дағдарыс туралы хабарламаларды жіктеуде 94% дәлдікке қол жеткізетінін көрсетті. Бұл нәтижелер басқа зерттеулермен расталды. Қазіргі зерттеулерде мәтіндік мәліметтерді алдын ала өңдеу мәселелеріне ерекше көңіл бөлінеді. Агглютинативті тілдермен жұмыс істегенде дұрыс токенизация әсіресе маңызды. Маннинг және т.б. [13] талдау сапасын жақсарту үшін деректерді қалыпқа келтірудің маңыздылығын атап көрсетеді.

Зерттеу нәтижелері

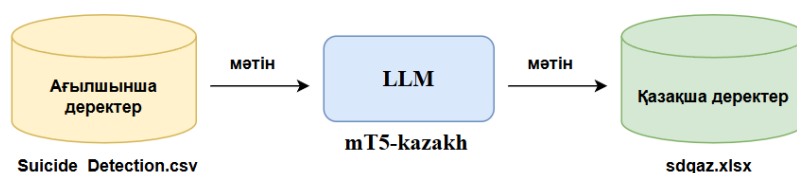
Қазақ тіліне бейімделген ең көп таралған үлгілерге KazBERT, KazT5 және mBERT және XLM-R сияқты көптілді модельдер жатады. Мысалы, KazBERT BERT архитектурасына (1-сурет) негізделген және қазақ корпусы бойынша арнайы дайындалған, бұл оны қазақ тіліндегі мәтінді өңдеу тапсырмалары үшін тиімдірек етеді [14].



Сурет 1. BERT архитектурасы

Сонымен қатар, зерттеулер [15] XLM-R сияқты көптілді модельдер қосымша оқытусыз қазақ тілінде жақсы нәтиже беретінін растайды. XLM-R тек біртілді деректерді пайдалана отырып, көптілді MLM мақсатымен үйретілген Трансформатор үлгісін пайдаланады. Ол әр тілдегі мәтін ағындарын таңдайды және шикі мәтінде қосалқы сөздерді таңбалауды қолданатын Unigram тіл үлгісі бар сөйлем бөлігін және ($\alpha = 0,3$) бірдей іріктеу үлестірімін пайдалана отырып, әртүрлі тілдерден алынған үлгі топтамаларын пайдалана отырып, енгізудегі бүркенішті таңбалауыштарды болжауға үлгіні үйретеді.

Осылайша, қазақ тіліне арналған тілдік үлгілерді әзірлеу жалғасуда және аталған процесте KazNERD-ке келетін басқа да корпустардың орны ерекше. Орындалған ізденіс жұмыстары үлгілерді сапалы етіп, мемлекеттік тілге икемдеуге бейімделген. Ізденісте Kaggle сайтындағы қолжетімді ағылшынша әлеуметтік тораптағы жазба деректерден қамтылған деректер жиыны пайдаланылды. Қазақ тіліне аударылып, барлығы 52995 жазбаны қамтитын деректер жинақталды. Суицидтік контексттегі 26998, суицидтік емес контексттегі 25997 сәйкесінше «1» және «0» цифрларымен белгіленді. Suicide_Detection.csv файлындағы ағылшынша мәтіндер kz-transformers/mT5-kazakh моделі арқылы (2-сурет) қазақ тіліне аударылды. kz-transformers/mT5-kazakh – бұл "text2text" тапсырмаларына бағытталған мамандандырылған LLM. Ағылшын тілінен қазақ тіліне офлайн режимде (API қолданусыз) жоғары сапалы аударма алуға мүмкіндік беретіндіктен деректерді аударуда аталған LLM қолданылды.



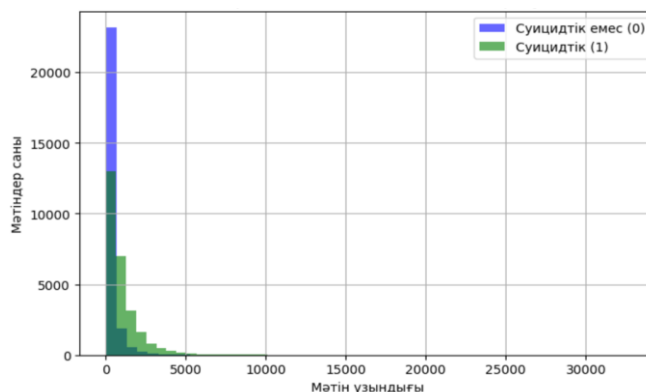
Сурет 2. Деректер аудармасының құрылымы

Жалпы mT5-kazakh моделі арқылы алынған аудармалар бастапқыда датафреймде ‘translated’ бағанына сақталып, кейін ол ‘text’ бағанына өзгертіліп, ал ағылшынша мәтіндер толық алынып тасталды.

Деректер жиынының сипаттамалары. Аударғаннан кейін осы зерттеуде пайдаланылған соңғы деректер жинағы 52 995 мәтіндік жазбадан тұрды. Олардың ішінде:

- 26 998 жазба суицидтік мазмұндағы (шамамен 51%);
- 25 997 жазба суицидтік емес мазмұндағы (шамамен 49%);

- сыныптардың жақсы тепе-теңдігін (51% және 49%) көрсетті, бұл оқытудың классификация үлгілері үшін маңызды фактор болып табылады, өйткені ол үстем сыныпқа қатысты модельдік бейімділікті болдырмайды. Мәтіндік деректердің сипаттамаларын талдау үшін әрбір санаттағы жазбалардың ұзындығы (сөздер саны) зерттелді. 3-суретте суицидтік (жасыл) және суицидтік емес (күлгін) жазбалар үшін мәтін ұзындығының таралуы көрсетілген. X осі мәтіннің ұзындығын (сөздер санын), ал Y осі сәйкес ұзындықтағы мәтіндердің санын көрсетеді.



Сурет 3. Суицидтік және суицидтік емес мәтін ұзындығының класс бойынша таралуы

Тазалаудан өткен мәліметтер токенизация процесінен өтеді. Ол мәтінді шағын топтарға, демек токендерге бөліп ажырату деген ұғымды береді. Мұндағы сөз тіпкестері, абзацтар, қала берді таңбалар да кездесіп қалады. Мақалада әрбір сөздер бөліп қарастырылған. 5-суретте қолданылған токенизация кодының мысалы ұсынылған.

```
# 1. HTML тегтерін жою
def remove_html_tags(text):
    return BeautifulSoup(text, "html.parser").get_text()

# 2. Эмодзилер мен арнайы таңбаларды жою
def remove_emojis_and_symbols(text):
    emoji_pattern = re.compile("[
        u'\U0001F600-\U0001F64F" # смайликтер
        u'\U0001F300-\U0001F5FF" # символдар мен пиктограммалар
        u'\U0001F680-\U0001F6FF" # көлік және белгішелер
        u'\U0001F700-\U0001F7FF" # басқа символдар
        u'\U0001F800-\U0001F8FF"
        u'\U0001F900-\U0001F9FF"
        u'\U0001FA00-\U0001FA6F"
        u'\U0001FA70-\U0001FAFF"
        u'\U00002702-\U000027B0" # қосымша белгілер
        u'\U000024C2-\U0001F251"
    ]+', flags=re.UNICODE)
    return emoji_pattern.sub(r'', text)

# 3. Тыныс белгілері, арнайы таңбалар және сандарды жою
def remove_punctuation_numbers(text):
    return re.sub(r'[^а-яА-ЯёЁқҚңҢөҮұҮһиИа-яА-Яа-зА-З\s]', '', text)

# Барлық тазарту қадамдарын қолдану
def clean_text(text):
    text = remove_html_tags(text)
    text = remove_emojis_and_symbols(text)
    text = remove_punctuation_numbers(text)
    return text
```

Сурет 4. Экран бейнесі. Мәліметтерді өңдеу программасы

```
# Токенизация
max_words = 10000
max_len = 100

tokenizer = Tokenizer(num_words=max_words, oov_token="<OOV>")
tokenizer.fit_on_texts(X_train)

X_train_seq = pad_sequences(tokenizer.texts_to_sequences(X_train), maxlen=max_len, padding="post")
X_test_seq = pad_sequences(tokenizer.texts_to_sequences(X_test), maxlen=max_len, padding="post")
```

Сурет 5. Токенизация коды

Келесі процесс – лемматизация және стемминг. Мұнда әрбір сөз өздерінің бастапқы қалпына орайластыруға бағытталған мәтінді өңдеудің қос тәсілін қамтиды. Лемматизация сөздің түпкі мақсатын айқындауға контекст және сөздің типін басты назарға алып, тіркестерді, морфологиялық талдауларды пайдаланады, ал стемминг сөздің соңынан қосымшаларды қысқарту арқылы жүргізіледі, нәтижесінде сөздің алғашқы түп нұсқасына жақын формасы шығарылады, бірақ бұл әдіс нәтижелердің дәл болмауына әкелуі ғажап емес. Жалпы мақала мағынасы жағынан тартымды, ұтымды болғаны үшін түбір сөздерді орнымен қолдану үшін лемматизация тәсілі пайдаланылды. Ол жайында 6-суретте көрсетілген.

```
# Stanza үшін қазақша модельді жүктейміз
stanza.download("kk")
nlp = stanza.Pipeline("kk", processors="tokenize,mwt,pos,lemma")

# Лемматизация функциясы
def lemmatize_text(text):
    doc = nlp(text)
    lemmas = [word.lemma for sent in doc.sentences for word in sent.words]
    return " ".join(lemmas)

# Лемматизацияны қолдану
df["text_lemmatized"] = df["text"].apply(lemmatize_text)
```

Сурет 6. Лемматизация коды

```
# TF-IDF векторизациясы
vectorizer = TfidfVectorizer(max_features=5000, ngram_range=(1,2))
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)
```

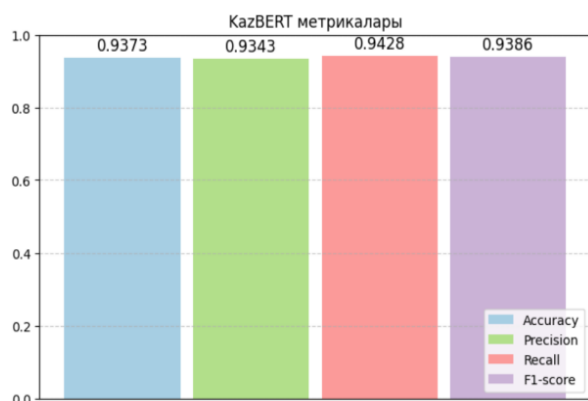
Сурет 7. Векторизация коды

Компьютермен цифрлық мәліметтер ғана өңделетіндіктен, ендігі орындалатын негізгі белес – векторлау тәсілі. Ол мәтінді цифрлық үлгіге, әсіресе, айқын вектор түрінде (цифрлық массивтерге) келтіру үрдісі болып табылады (7-кескіндегі үлгіде көрсетілген).

Дискуссия

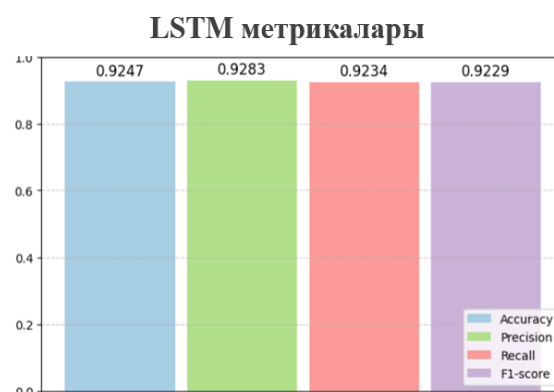
Барлық модельдер оқыту деректерінің 80%-ында жаттықтырылып, 10%-ы валидация үшін және қалған 10%-ы тест жиыны ретінде қолданылды. Модельдер accuracy, precision, recall, F1-score метрикалары бойынша бағаланды. Зерттеуде 3 түрлі машиналық оқыту үлгісі қолданылды. KazBERT – HuggingFace платформасында жарияланған қазақ тіліне бейімделген трансформер үлгісі. Бұл модель қазақ тіліндегі мәтіндерді өңдеуге және мәтін мағынасын тереңірек түсінуге қабілетті. Аталған модель машиналық оқыту арқылы танылған бірнеше тілді BERT-пен дамытылып, қазақша деректермен қосалқы түрде бейнеленді. Тану барысында cross-entropy loss бернесі, сондай-ақ AdamW қосымшасы пайдаланылды.

Аталмыш алгоритмді тану кезінде Trainer API-дің көмегімен, көптеген басты функциялар қойылды. Ол кезде негізгі мәндер сегіз рет қойылды. Бұл параметр құрылғыдағы VRAM ресурстарын тиімді пайдалану үшін таңдалды. Batch size мәнін үлкейту оқыту жылдамдығын арттыруы мүмкін, бірақ GPU жад шектеулері ескерілуі тиіс. Ұсынылған алгоритм барлық мәліметтер тобын 3 мәрте тексереді, сондықтан ол параметр үшке бара бар. Epoch жиілігінің артуы алгоритмнің неғұрлым дұрыс жұмыс істеуіне оң әсерін тигізеді, ал ол шамадан тыс үйретуге әкеп соғады, сол себепті тиімді мәні үшін 3 этап қарастырылды. Аталған алгоритмнің үйрету кезіндегі өнімділік сапасын қадағалауға ықпал етеді. Оған қоса, айтулы параметр әрбір этаптан соң алгоритмді қайта сақтау мүмкіндігі болатынын растайды. Сақтап болған соң олар екімен тоқтайды, ал ол өз кезегінде дискідегі бос кеңістік орнын қалыптастырады. Журналды жүргізіп отыру әрбір бес жүз қадам арқылы орындалды. Оның үйрету ретін қадағалауға жәрдемдесіп әрі алгоритмнің графигін айқындайды. Алгоритм үйрету қорытындысы бойынша жоғары нәтижеге қол жеткізді (8, 9-суреттерде көрсетілген).



Сурет 8. KazBERT метрикалары

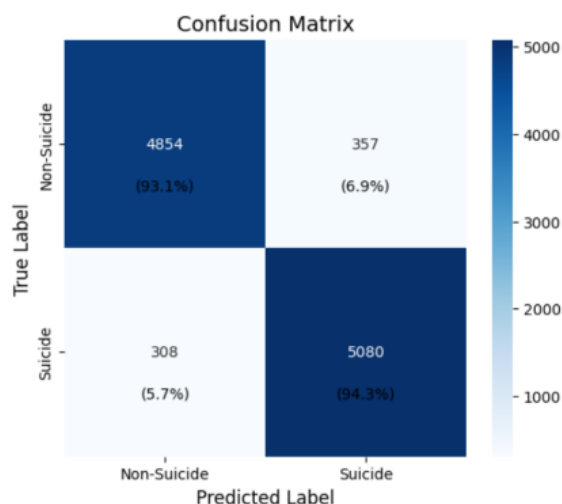
Accuracy: 93,73 %
Precision: 93,43 %
Recall: 94,28 %
F1-score: 93,86 %



Сурет 9. LSTM метрикалары

Accuracy: 92,47 %
Precision: 92,83 %
Recall: 92,34 %
F1-score: 92,29 %

KazBERT моделіне арналған қателіктер матрицасы (10-сурет) оның сыныптарды ажырату қабілетін айқын көрсетеді. Модель 5080 суицидтік мәтінді (шын оң) және 4948 суицидтік емес мәтінді (шын теріс) дұрыс жіктеді. Бұл жағдайда 308 суицидтік мәтін суицидтік емес (жалған теріс), суицидтік емес 253 мәтін суицидтік (жалған оң) деп қате жіктелген. Қателердің екі түрі де аз мөлшерде кездеседі, дегенмен бұл мәндер модельдің жоғары тиімділігін көрсетеді.



Сурет 11. KazBERT моделінің қателіктер матрицасы

Модельдердің тиімділігін салыстырмалы талдау және ең жақсы үлгіні таңдау. Алынған көрсеткіштерді салыстыру KazBERT моделінің негізгі логистикалық регрессия үлгісінен де, LSTM моделінен де барлық негізгі сапа көрсеткіштері бойынша: дәлдік, сенімділік, еске түсіру және F1-өлшемі бойынша асып түсетінін көрсетеді. Бұл үлкен көлемдегі тіл деректерінде алдын ала дайындалған және белгілі бір тілге бейімделген, мәтінді терең семантикалық талдау тапсырмаларында трансформер архитектураларының жоғары тиімділігін көрсетеді (1-кесте).

Кесте 1. Модель метрикаларының салыстырмалы кестесі

Модель метрикасы	KazBERT	Логистикалық регрессия	LSTM
Дәлдік, %	93,73	92,86	92,47
Дәлдік, %	93.43	92,78	92,83
Еске түсіру, %	94.28	93,59	92.34
F1-балл, %	93,86	93,62	92.29

LSTM моделі мәтіндегі дәйекті тәуелділіктерді түсіруде де жақсы жұмыс істегенімен, оның өнімділігі KazBERT-тен сәл төмен болды. Логистикалық регрессия моделі, үйренуге оңайырақ және жылдамырақ болғанымен, нейрондық желі тәсілдерінің дәлдігі, әсіресе күрделі контексттерді түсінуде болжамды түрде төмен орындалды, бірақ ол күрделірек үлгілерді бағалау үшін маңызды эталон ретінде қызмет етті. LSTM моделі логистикалық регрессиямен салыстырғанда ұзақ оқу үдерісіне қарамастан мәтін құрылымын терең түсіне алды.

KazBERT моделінің артықшылығы оның трансформер архитектурасына негізделгендігінде және қазақ тілінің ерекшеліктеріне бейімделіп, алдын ала үлкен көлемдегі қазақша мәтіндерде оқытылуында жатыр. Бұл модель, "назар аудару" (attention) механизмдерін, әсіресе "өзіне-өзі назар аудару" (self-attention) механизмін қолданады. Бұл механизмдер модельге сөйлемдегі әрбір сөзді өңдеу кезінде сол сөйлемдегі басқа сөздердің маңыздылығын бағалауға, яғни контекстті тереңірек түсінуге мүмкіндік береді. Екі бағытты кодтау (bidirectional encoding) арқасында KazBERT әрбір сөз үшін оның алдындағы және кейінгі сөздерді бір мезгілде ескереді, бұл сөздердің байытылған, контекстке негізделген векторлық ұсыныстарын (embeddings) қалыптастыруға ықпал етеді. Осының нәтижесінде KazBERT жасырын мағыналарды, сөз орамдарының нюанстарын және мәтіннің жалпы тонын дәлірек анықтай алады, бұл суицидтік ойларды білдіретін, бірақ тікелей кілт сөздерді қамтымайтын мәтіндерді

жіктеуде өте маңызды. Алдын ала оқытылуы және қазақ тілінің спецификасына бейімделуі оның жалпылама көптілді модельдерге қарағанда артықшылығын қамтамасыз етеді, өйткені ол тілдің морфологиялық және семантикалық ерекшеліктерін жақсырақ "түсінеді". Осы факторлардың жиынтығы KazBERT моделінің LSTM және логистикалық регрессияға қарағанда жоғарырақ нәтиже көрсетуіне негіз болды.

Барлық көрсеткіштер бойынша ең жоғары ұпайларды ескере отырып, KazBERT моделі суицидтік мінез-құлық белгілері бар мәтіндерді талдау және жіктеу қосымшасында одан әрі пайдалану үшін ең жақсысы болып таңдалды. Жасалған үлгілердің практикалық қолданылуын көрсету және қазақ тіліндегі мәтіндерді талдауға арналған құралмен қамтамасыз ету мақсатында интерактивті веб-қосымша әзірленді. Оның негізі ретінде ең жақсы эксперименттік нәтижелер көрсеткен KazBERT моделі таңдалды. Пайдаланушы интерфейсі Gradio кітапханасының көмегімен жүзеге асырылды.

Қосымша архитектурасы және технологиялық стек. Қосымша Python-да құрастырылған және келесі негізгі компоненттерді пайдаланады:

Жіктеу үлгісі: алдын ала оқытылған KazBERT (Eraly-ml/KazBERT) негізіндегі қайта оқытылған BertForSequenceClassification моделі және оған сәйкес BertTokenizer токенизаторы.

Деректерді өңдеу кітапханасы: Excel (.xlsx) файлдарын оқуға және өңдеуге арналған pandas кітапханасы.

Веб-интерфейсті жасауға арналған кітапхана: интерактивті интерфейс ті жылдам орналастыруға арналған gradio кітапханасы.

Тензорлық есептеулерге арналған кітапхана: KazBERT моделінің бәкенді ретінде torch (PyTorch) кітапханасы.

Жұмысты бастамас бұрын модель мен токенизатор жүктеліп, модель бағалау режиміне қойылады (model.eval()) болжау кезінде дұрыс жұмыс істеуін қамтамасыз ету.

Функционалдылық және пайдаланушы интерфейсі. Қосымша Gradio интерфейсінде бөлек қойындылар ретінде жүзеге асырылатын мәтіндермен жұмыс істеудің екі негізгі режимін ұсынады. Жекеленген мәтінді жіктеу («Мәтін» қойындысы). «Мәтін» қойындысында пайдаланушы талданатын мәтінді енгізуі немесе көшіріп қоюы үшін мәтін өрісі беріледі (11-сурет).



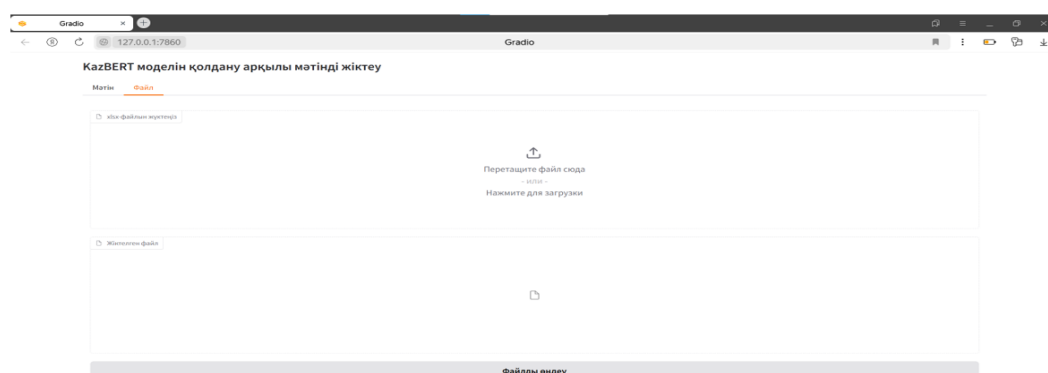
Сурет 11. Қосымшаның «Мәтін» қойындысы

Мәтінді енгізіп, «Анықтау» батырмасын басқаннан кейін енгізілген мәтін predict_single функциясына беріледі. Бұл функцияның ішінде мәтін жүктелген BertTokenizer көмегімен токенизацияланады, тензорларға түрлендіріледі және KazBERT үлгісінің кірісіне беріледі. Модель болжамды қайтарып, оның негізінде екі белгінің бірі "Нәтиже" өрісіне шығарылады: "Суицидті жазба" немесе «Суицидті емес жазба» (12-сурет).



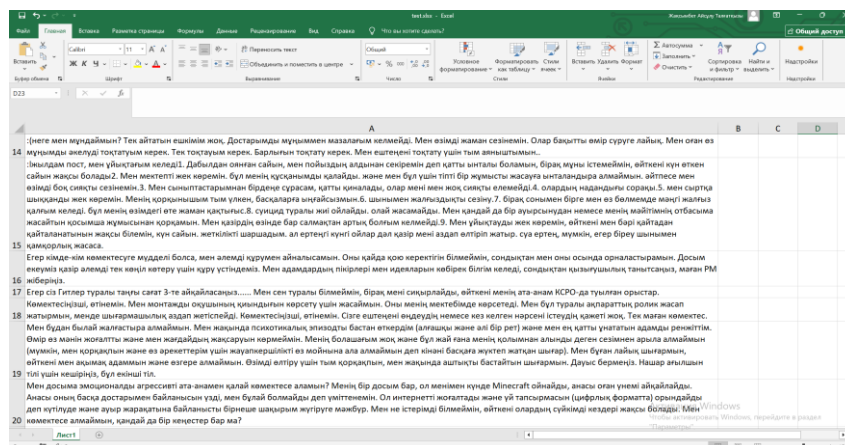
Сурет 12. Қосымианың «Мәтін» қойындысында мәтінді жіктеу

Файлды өңдеу ("Файл" қойындысы). Пайдаланушыға файлды Excel форматында (.xlsx) «xlsx-файлын жүктеу» ұяшығына жүктеп салу мүмкіндігі беріледі (13-сурет).



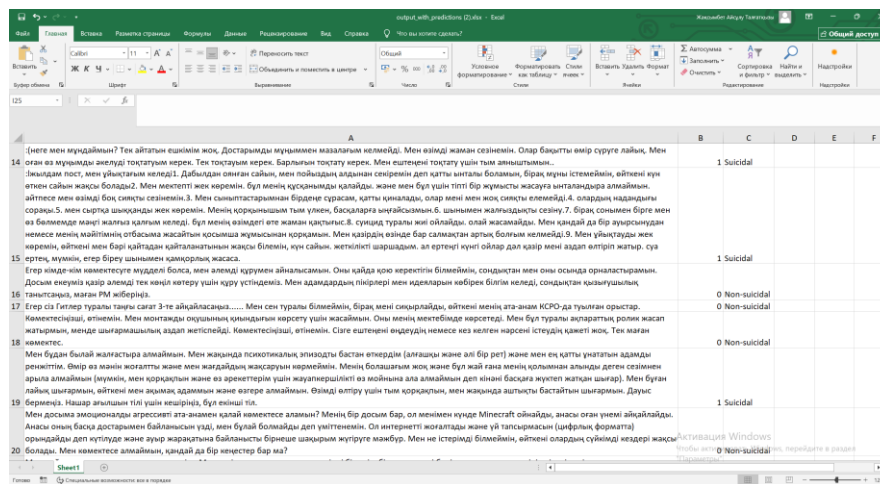
Сурет 13. Қосымианың «Файл» қойындысы

Файлды жүктеп, «Файлды өңдеу» батырмасын басқаннан кейін predict_file функциясы іске қосылады. Функция Excel файлынан деректерді pandas кітапханасының көмегімен оқиды. Қажетті 'text' бағанының болуы тексеріледі (14-суретте жүктелінетін файлдың 'text' бағаны көрсетілген). Ол жоқ болса, «Файлда 'text' бағаны болу қажет» деген хабарлама шығады.



Сурет 14. «Файл» қойындысынан жүктеуге салынатын файл

'text' бағанындағы мәтіндерді KazBERT моделі ретімен өңдейді (токенизация және 0 немесе 1 кластарын болжау). Жүктелген DataFrame-ге екі жаңа баған қосылады: 'prediction' (0 немесе 1 сандық белгілері бар) және 'class' (сәйкесінше "Non-suicidal" немесе "Suicidal" мәтіндік белгілер көрсетіледі) (15-сурет).



Сурет 15. «Файл» қойындысынан жүктелген өңделінген файл

Жаңартылған DataFrame `output_with_predictions.xlsx` атты жаңа Excel файлына сақталады, содан кейін ол пайдаланушыға "Жіктелген файл" компонентінде жүктеп алуға ұсынылады (16-сурет).



Сурет 16. Қосымшаның «Файл» қойындысында файл өңдеу

Қолданбадағы үлгі бойынша мәтінді өңдеу процесі. Енгізу әдісіне қарамастан (бір мәтін немесе файл), әрбір мәтін фрагменті жіктелмес бұрын қолданба ішінде келесі өңдеу қадамдарынан өтеді:

- Токенизация: Мәтін KazBERT үшін бейімделген BertTokenizer токенизаторы арқылы токендер тізбегіне түрлендіріледі. Padding тізбектердің ұзындығын теңестіру үшін пайдаланылады және truncation тым ұзын мәтіндерді үлгі қолдайтын максималды ұзындыққа дейін қысқарту үшін қолданылады (`max_length=512`).
- Модельдің келтірулерін алу: Токенделген деректер BertForSequenceClassification моделінің кірісіне беріледі.
- Жіктеу: Модель екі сыныптың әрқайсысына жататын мәтіннің ықтималдығын есептейді (суицидтік/суицидтік емес).
- Нәтиженің шығуы: Ықтималдығы жоғары класс таңдалады және нәтиже пайдаланушыға сәйкес форматта ұсынылады.

Әзірленген қосымша оқытылған KazBERT үлгісін практикалық қолданудың көрнекі демонстрациясы ретінде қызмет етеді және суицидтік белгілерді анықтауда мәтіндік материалдарды жылдам бағалаудың ыңғайлы құралын ұсынады.

Қорытынды

Бұл зерттеуде әлеуметтік желілердегі посттардың суицидтік контекстін айқындар алдында көптеген машиналық оқытумен қамтып, кейіннен терең талқыланды. Мәліметтерді даярлау, мәтінді векторлау әрі нейрондық желіні пайдалана отырып қолайлы алгоритм таңдалып, өзара сәйкестендірілді. Нәтижелерді салыстырғанда BERT моделі ең жоғары дәлдік көрсетті, себебі

ол контексті терең түсінеді. LSTM алгоритмі және регрессиялық дәлдік аздап кейіндеу қалғанымен, оң көрсеткіш танытты, алғашқысы ауқымды мәтіндік массив арқылы қолайлы жұмыс істесе, келесісі шапшаң және ширақтығымен көзге түсті. Зерттеудің негізгі ресурсы Kaggle платформасындағы деректер жиынынан алынған жазбалардың ағылшын тілінен deep-translator кітапханасының көмегімен қазақ тіліне аудармасынан құралды. Зерттеудің практикалық маңыздылығы Gradio кітапханасын пайдалана отырып, KazBERT моделі негізінде веб-қосымшаны жасау арқылы айтарлықтай арта түсті. Қорыта келе, ашық ақпарат көздеріндегі өзіне қол салу фактілерін білуге тереңдетілген нейрон тораптары және үйретіліп дайын тұрған тілге қатысты алгоритмдер қолайлы болады. Әйтсе де, таңдап алынған тәсілдің қолданыс аясына қатысты ерекшеліктері ескерілген. Болашақта модельдердің тиімділігін арттыру үшін үлкен деректер жиындарын пайдалану, аугментация әдістерін енгізу және ансамбльдік тәсілдерді қарастыру жоспарлануда. Ол зерттеулердің барлығы да жасанды интеллект негізінде орындалады.

References

- [1] Самоубийства в Казахстане. <https://ru.wikipedia.org/?curid=4534555&oldid=146530946>
- [2] UNICEF Kazakhstan. (2014). *Assessment of suicide preventative activities in Kazakhstan: Kyzylorda and East Kazakhstan regions*. https://www.unicef.org/kazakhstan/sites/unicef.org.kazakhstan/files/2018-06/suicide_en.pdf
- [3] Government of Kazakhstan. (2023). *Comprehensive plan for protecting children and preventing suicide in Kazakhstan (2023-2025)*. <https://cabar.asia/en/violence-free-schools-a-comprehensive-plan-to-protect-children-from-bullying-in-kazakhstan-white-paper>
- [4] Zhang, X., & Wallace, B.C. (2015). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2015)*, pp. 1–9.
- [5] Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). *Latent Dirichlet Allocation*. *Journal of Machine Learning Research*, No 3, pp. 993-1022.
- [6] Griffiths, T.L., & Steyvers, M. (2004). *Finding scientific topics*. *Proceedings of the National Academy of Sciences*, No 101(Suppl 1), pp. 5228-5235.
- [7] «Study on prevalence, underlying causes, risk and protective factors in respect to suicides and attempted suicides in Kazakhstan», Astana: The UN Children's Fund (UNICEF) in the Republic of Kazakhstan, 2014. - 104 p. https://www.unicef.org/kazakhstan/sites/unicef.org.kazakhstan/files/2018-06/00001358_0.pdf
- [8] De Choudhuri, S., & Jagannath, P. (2018). *Comparison of Naive Bayes Classifier with Other Machine Learning Algorithms in Text Classification*. *International Journal of Computer Applications*, No 975, pp. 1-5.
- [9] Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1746–1751.
- [10] Zhang, X., & Wallace, B.C. (2015). *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2015)*, pp. 1–9.
- [11] Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. *Neural Computation*, Vol. 9, No 8, pp. 1735–1780.
- [12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *Proceedings of NAACL-HLT 2019*, pp. 4171–4186.
- [13] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, H., Bethard, S., & McClosky, D. (2014). *The Stanford CoreNLP Natural Language Processing Toolkit*. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pp. 55–60.
- [14] Sarker, I. H., & Riazi, H. (2020). *KazBERT: A Pre-Trained Transformer Model for Kazakh Text Classification*. *Proceedings of the International Conference on Artificial Intelligence and Computer Vision*, pp. 102–111.
- [15] Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Joulin, A. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pp. 8440–8451.