ТЕХНИЧЕСКИЕ НАУКИ
И ТЕХНОЛОГИИ

ИНФОРМАЦИОННЫЕ СИСТЕМЫ
INFORMATION SYSTEMS
АҚПАРАТТЫҚ ЖҮЙЕЛЕР

DOI 10.51885/1561-4212_2023_3_78
МРНТИ 20.23.17

А.Т. Бекишев[1], К.Е. Нурсакитов[2], С.Ж. Рахметуллина[3], С.К. Кумаргажанова[4], А.М. Уркумбаева[5]
НАО «Восточно-Казахстанский технический университет имени Д. Серикбаева»,
г. Усть-Каменогорск, Казахстан
[1]*E-mail: a.nomad.b@mail.ru*
[2]*E-mail: Nursakitov@bk.ru**
[3]*E-mail: SRakhmetullina@edu.ektu.kz*
[4]*E-mail: skumargazhanova@gmail.com*
[5]*E-mail: urkumbaeva@mail.ru*

СОЗДАНИЕ СЛОВАРЯ КЛЮЧЕВЫХ СЛОВ ДЛЯ КЛАССИФИКАТОРА ТЕКСТОВ, СОДЕРЖАЩИХ ОПАСНЫЙ КОНТЕНТ В КИБЕРПРОСТРАНСТВЕ КАЗАХСТАНА

ҚАЗАҚСТАН КИБЕРЕКІСТІГІНДЕГІ ҚАУІПТІ МАЗМҰНЫ БАР МӘТІНДЕР ЖІКТІТІРУШІСІНЕ АРНАЛҒАН ТҮЙІН СӨЗДЕР СӨЗДІГІН ҚҰРУ

CREATION OF A DICTIONARY OF KEYWORDS FOR A CLASSIFIER OF TEXTS CONTAINING DANGEROUS CONTENT IN THE CYBERSPACE OF KAZAKHSTAN

Аннотация. *Данная работа является частью исследования создания информационной системы для поиска опасного контента в киберпространстве Казахстана. Целью исследования является создание словаря ключевых слов для работы классификатора текстов, содержащих опасный контент, на примере задачи выявления наличия суицидального риска в текстах предсмертных записок и групп смертников. Для казахского языка не существует такой базы данных. В результате исследования был создан экспериментальный корпус и список ключевых слов на казахском языке. Ключевые слова были добавлены в базу данных с различными морфологическими формами.*

Ключевые слова: *обработка естественного языка, сентимент-анализ, машинное обучение, частота терминов, классификация текста.*

Аңдатпа. *Бұл жұмыс Қазақстанның киберкеңістігінде қауіпті контентті іздеудің ақпараттық жүйесін құру жөніндегі зерттеудің бір бөлігі болып табылады. Зерттеудің мақсаты суицидтік жазбалар мен суицидтік топтар мәтіндерінде суицидтік тәуекелдің болуын анықтау мәселесін мысалға ала отырып, қауіпті мазмұны бар мәтіндер классификаторының жұмысы үшін түйінді сөздер сөздігін жасау болып табылады. Қазақ тіліне арналған мұндай деректер базасы жоқ. Осы зерттеулердің нәтижесінде эксперименттік корпус пен қазақ тіліндегі түйінді сөздер тізімі жасалды. Түйін сөздер әртүрлі морфологиялық формалармен дерекқорға қосылды.*

Түйін сөздер: *табиғи тілді өңдеу, сентимент талдау, машиналық оқыту, термин жиілігі, мәтінді жіктеу.*

Abstract. *This work is part of a study on the creation of an information system for searching dangerous content in the cyberspace of Kazakhstan. The aim of the study is to create a dictionary of keywords for the work of a classifier of texts containing dangerous content, using the example of the problem of identifying the presence of a suicidal risk in the texts of suicide notes and groups of suicidal. There is no such database for the Kazakh language. As a result of this research, an experimental corpus and a list of keywords in the*

Kazakh language were created. Keywords have been added to the database with various morphological forms.

Keywords: *natural language processing, sentiment analysis, machine learning, term frequency, text classification.*

Введение. Проблемы психического здоровья, такие как тревожность и депрессия, вызывают все большую обеспокоенность в современном обществе [1]. Некоторые посты в социальных сетях содержат много негативной информации, порождающей такое проблемное явление, как суицидальный подтекст. Порядка 80 тысяч обращений по данным фактам зафиксировано в Казахстане за 2021 год [2]. Для борьбы с данным явлением правительством РК был разработан законопроект «О внесении изменений и дополнений в некоторые законодательные акты Республики Казахстан по вопросам защиты прав ребенка, образования, информации и информатизации», позволяющий уполномоченному органу блокировать контент с вышеуказанными признаками, в том числе и кибербуллинга [3]. При этом стоит отметить, что последнее является одной из основных причин суицидального поведения и суицидов в целом.

Однако из-за больших объемов данных, публикуемых в сети интернет ежедневно, своевременно находить и блокировать подобные материалы вручную физически невозможно.

В связи с этим становится актуальным автоматический мониторинг интернет-ресурсов с целью выявления текстов суицидальной (в. т.ч. кибербуллинговой) направленности. Данную задачу можно представить в виде задачи бинарной классификации, в которой тексты сообщений в социальных сетях, блогах и других ресурсах будут выступать в роли анализируемых объектов, и решаться с помощью методов машинного обучения [4]. Такой подход требует наличия размеченного корпуса текстов [5] и предопределенного набора анализируемых характеристик, таких как результаты полного лингвистического анализа, список ключевых слов и т. д.

Базы ключевых слов суицидального характера, как для казахского, так и для русского языков не существует. Данная статья является частью исследования по созданию информационной системы семантического анализа в веб-ресурсах для определения суицидального характера в веб-контенте (в т.ч. кибербуллинг направленности в тексте).

Целью исследования является выявление ключевых слов, часто употребляемых, которые будут использоваться для классификации текстов на категории «суицидальные» и «нейтральные» с применением методов машинного обучения, а в дальнейшем – для определения суицидального поведения в сети Интернет.

В мировой статистике число самоубийц превышает количество жертв убийств, террористических актов и войн, вместе взятых. Каждые 20 секунд один человек заканчивает жизнь самоубийством, а каждые 2 секунды кто-то безуспешно пытается свести счеты с жизнью.

Современное состояние исследований по обнаружению опасного контента. В настоящее время существует множество исследований по решению проблемы классификации текстов, содержащих опасный контент, однако большая часть из них – это зарубежные исследования.

В исследовании [6] описано создание прототипа программного обеспечения, способного автоматически идентифицировать буллинг-суицидальные комментарии на платформе социальных сетей ASKfm с использованием методов обработки естественного языка и машинного обучения.

В исследовании [7] используется испанская система предотвращения суицидальных постов, в т.ч. киберзапугивания (SPC), которая опирается на методы обработки

естественного языка (NLP) и различные методы машинного обучения (наивный байесовский метод, метод опорных векторов и логистическая регрессия), используя Twitter в качестве основы для извлечения баз знаний. Для обучения используются несколько показателей точности и переменные размеры корпуса. Результаты обучения достигают максимальной точности 93 %, подтвержденной применением трех учебных примеров.

В статье [8] авторы представляют систему для мониторинга явлений киберзапугивания, сочетая классификацию сообщений и анализ социальных сетей. Оценивают модуль классификации на наборе данных, построенном на сообщениях Instagram, и описывают пользовательский интерфейс мониторинга киберзапугивания.

В работе [9] проведен углубленный анализ 22 исследований по автоматическому обнаружению суицидальных и киберзапугиваний, дополненный экспериментом по проверке существующих практик посредством анализа двух наборов данных. Результаты исследований показали, что сама тема часто неверно представлена в литературе, что приводит к неточным системам, которые мало применимы в реальном мире.

В исследовании [10] приведены глобальный набор данных из 37 373 уникальных твитов из Twitter и семь классификаторов машинного обучения, а именно: логистическая регрессия (LR), машина повышения градиента света (LGBM), стохастический градиентный спуск (SGD), случайный лес (RF), AdaBoost (ADB), наивный байесовский (NB) и машина опорных векторов (SVM). Каждый из этих алгоритмов оценивался с использованием точности, прецизионности, отзыва и оценки F1 в качестве показателей производительности для определения показателей распознавания классификаторов, применяемых к глобальному набору данных.

В статье [11] предложен автоматический метод для обнаружения агрессивного поведения среди пользователей социальных сетей с использованием консолидированной модели глубокого машинного обучения. В данном методе используется многоканальное глубокое обучение, основанное на трех моделях: двунаправленном закрытом рекуррентном блоке (BiGRU), блоке преобразователя и сверточной нейронной сети (CNN), для классификации комментариев Twitter на две категории: агрессивные и неагрессивные. Для оценки эффективности предложенного метода были объединены три известных набора данных о разжигании ненависти. Точность предложенного метода составила около 88 %.

Русскоязычных исследований не так уж и много, при этом большая их часть направлена на выявления тематики на английском языке.

В статье [12] рассматривается суть понятий, в т.ч. кибербуллинга, как виртуальной агрессивной коммуникативной стратегии. Подчеркивается, что он осуществляется на разных интернет-платформах, не ограничен во времени и пространстве, может быть представлен несколькими видами. Утверждается, что поскольку электронная травля запрещена в Сети, автоматическая блокировка сообщений с суицидальными элементами может осуществляться на основе работы системы автоматического распознавания кибербуллинга в виртуальном общении. В рамках инженерного подхода представлены основные принципы организации базы данных.

Основная часть. При использовании метода классификации текста на категории, основанного на тональных словарях, для автоматической классификации текстов необходимо опираться на словарь, в котором содержатся слова с разметкой принадлежности их к определенной категории. В данной работе для примера работы метода классификации была выбрана задача выявления наличия суицидального риска в текстах предсмертных записок и групп смертников.

Для решения задачи выявления суицидального характера в веб-контенте (в т.ч.

кибербуллинга в текстах) была разработана методика, состоящая из пяти этапов:

1) выявление сайтов, размещающих текстовый контент;
2) подготовка к извлечению данных;
3) извлечение данных;
4) анализ данных;
5) классификация.

Согласно методике, на первом этапе определяются веб-ресурсы, которые люди чаще всего используют для обмена сообщениями. В данном случае были рассмотрены наиболее часто используемые социальные сети, такие как Youtube, vk, блоги, форумы и новостные сайты. Второй этап – подготовка к извлечению данных (регистрация на форумах, выбор подходящих прокси-серверов).

Как отмечалось выше, база данных ключевых слов кибербуллинга на казахском и русском языках отсутствуют. По этой причине для данного исследования в первую очередь необходим корпус текстов, написанных на казахском и русском языках. Для создания корпуса был использован открытый датасет «Students anxiety and depression dataset» [17], состоящий из 6500 текстовых сообщений из социальных сетей, комментариев Facebook и т.п. Каждое сообщение аннотировано на две категории: 0 – нейтральное сообщение и 1 – сообщение, содержащее признаки депрессии, суицида или кибербуллинга. Для работы с текстами на казахском и русском языках сообщения были переведены с английского. На данный момент в файле содержится 250 текстов, из них 130 текстов с суицидальным характером и кибербуллингом, остальные тексты относятся к категории «нейтральные», которые содержат комментарий, осуждающий кибербуллинг и новостные тексты.

На рис. 1 приведены примеры комментариев и текстов.

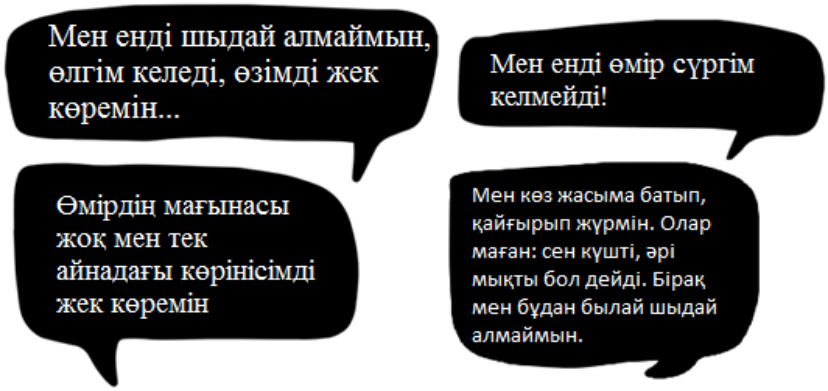

Рисунок 1. Пример комментариев, содержащих суицидальный подтекст

Для определения наиболее часто употребляемых слов исследователи используют разные методы. Например, в [13] используется метод «Part of Speech Tagged». Метод выделяет часто встречающиеся слова и делит текст на группы по частям речи (существительные, глаголы, наречия и т. д.). Данный метод удобно использовать, когда исследователи акцентируют внимание на определенных частях речи. Например, в приведенной выше работе исследователи рассматривают только существительные, так как по результатам исследования ключевыми словами буллинга в английском языке в большинстве случаев являются существительные. В нашем случае этот способ не подходит, так как визуальный осмотр показал, что большинство ключевых слов в казахском языке, помимо существительных, являются глаголами. По этой причине в

данном исследовании был использован метод TF-IDF, который используется для оценки важности слова. Для нахождения слов, характерных для данного типа документов, которые соответствует сообщению форума, находят частоту термина (TF), обратную частоту документа (IDF)[14].

Перед началом обработки текстов следует провести так называемый этап предобработки, который состоит из трех подэтапов [15]:

1. Токенизация – процесс разбиения текста на текстовые единицы (чаще всего слова);
2. Нормализация – серия операций, в результате которых текст приводится к «рафинированному» виду: все слова приводятся к одному регистру, удаляются знаки пунктуации, расшифровываются сокращения, числа приводятся к их текстовому написанию и т.д. Нормализация необходима для унификации методов обработки текста;
3. Удаление стоп-слова – удаление слов, которые не несут никакой смысловой нагрузки. Их еще называют шумовыми словами. Стоп-слова уже давно применяются в алгоритмах поисковых машин. Например, в английском языке – это артикли, в русском – междометия, союзы, и т.д.

После прохождения текстом всех трех этапов предобработки, можно переходить непосредственно к самим вычислениям.

Частота термина *tf(t, d)*, самый простой выбор – использовать необработанный подсчет термина в документе, т. е. количество раз, когда термин *t* встречается в документе *d*.

Обратная частота документа – это мера того, сколько информации предоставляет слово, то есть является ли термин общим или редким во всех документах. Это логарифмически масштабированная обратная дробь документов, содержащих слово, полученная путем деления общего количества документов на количество документов, содержащих термин, и последующего логарифмирования этого частного. IDF определяет с помощью [16].

$$Idf(t, D) = \log \frac{|D|}{|\{d_i \in D\}| t \in d_i|}, \quad (1)$$

где |D| – общее количество документов в корпусе; |{di ∈ D|t ∈ di}| – количество документов, в которых встречается термин t.

Тогда TF-IDF рассчитывается как:

$$tf - id(t,d,D)f = tf(t,d) * idf(t,D). \quad (2)$$

С помощью (2) были рассчитаны значения TF-IDF всех слов. Результаты приведены в порядке убывания в табл. 1.

Таблица 1. TF-IDF значения используемых слов

Ключевое слово	*TF-IDF значение*
Мағынасыз	21.62
Шыдау	18.66
Өмір	10.92
Өлім	17.1
……..	**……..**

В работе представляет интерес только основа слова, без окончаний, т.к. окончания снижают эффективность системы из-за увеличения времени поиска различных форм слова. Поэтому эффективнее рассматривать разные варианты слова с разными окончаниями как одно слово. Например,

слова «өмір» (жизнь), «өмірдің», «өмірге», «өмірден», «өмірде» будут считаться одним и тем же словом. Как было сказано выше, одно слово может иметь несколько вариантов написания, поэтому в базу данных были занесены слова со всеми возможными вариантами. Возможные морфологические варианты слов определялись путем изучения контента веб-форумов, блогов.

Результаты. В результате последовательности процессов предобработки (токенизация, нормализация, удаление стоп-слов) из 250 текстов было выявлено 1045 уникальных слов. На основе этих слов был составлен словарь, фрагмент которого представлен в табл. 2.

Таблица 2. Примеры наиболее часто используемых слов с количеством повторений во всем корпусе текстов

Вариант 1	*Кол-во*	*Вариант 2*	*Частота*	*Вариант 3*	*Кол-во*
Мағынасыз	9	Магынасыз	34	Magynasyz	12
Шыдау	15	Shydau	28	-	-
Өмір	14	Омир	19	Omir	2
Өлім	11	Олим	12	Olim	3
……..	……..	……..	……..	……..	……..

Таблица ключевых слов в базе данных представлена на рис. 2. Были рассчитаны значения TF-IDF каждого слова, по которым слова в базе были упорядочены от самого часто встречаемого к редкому. Результаты приведены в порядке убывания в табл. 1. Эти слова могут быть использованы для повышения достоверности определения суицидального характера текста.

id	var0	var1	var2
1	Өмір	Омир	Omir
2	Мағынасыз	Магынасыз	Magynasyz
3	Шаршадым	-	Sharshadym
4	Жоғалып	Жогалып	Zhogalyp
5	Өлтір	Олтир	Oltir
6	Жалғыз	Жалгыз	Zhalgyz
7	Мәңгі	Манги	Mangi
8	Аяныштy	-	Ayanyshty

Рисунок 2. Список ключевых слов в базе данных

В дальнейшем планируется присвоить выявленным словам эмоциональные оттенки, которые в будущем будут использованы для создания алгоритмов и программного обеспечения для анализа тональности текста (сентимент-анализ).

На этапе анализа наиболее частых слов было выявленно, что казахские буквы часто заменяются кириллическими, например: «каракат» вместо «қарақат» (перевод с рус. – смородина), «муз» вместо «мұз» (перевод с рус. – лед) и т.д. Встречаются случаи, когда

одно слово написано в нескольких вариантах (например, ақымақ (перевод с рус. – дурак), акымак, aqimaq, akymak) и когда комментарии написаны не чисто на казахском языке, а смешаны с русскими словами (например, «привет братка, қалайсын?»).

Вывод. В ходе исследования были найдены и выделены ключевые слова, с разметкой принадлежности их к сентименту потенциально опасного контента *(суицидальный)*, по которым был построен словарь. Данный словарь в дальнейшем будет использован в работе автоматического классификатора текстов. Однако размер словаря на момент написания статьи невелик, и работа над его расширением продолжается. Стоит отметить тот факт, что буллинг также имеет очень тесную взаимосвязь с суицидальным поведением, т.к. его наличие прямо пропорционально суицидам среди подростков. В будущем планируется добавить полярность настроений между [-1;1], которая будет использоваться при анализе настроений. Следующим этапом исследования является классификация входящих текстов с использованием методов машинного обучения, таких как метод Байеса, метод опорных векторов (SVM), случайный лес и логистическая регрессия. Данные исследования будут крайне актуальны и полезны для определения настроений текстов с суицидальными наклонностями в сети интернет, что является более глобальной задачей нашего исследования.

Список литературы

1. Карауш И. С., Куприянова И. Е., Кузнецова А. А. Кибербуллинг и суицидальное поведение подростков // Суицидология. – 2020. – Т. 11. – №. 1 (38). – С. 117-129.
2. https://mk-kz.kz/social/2022/03/25/kiberbulling-realnaya-problema-ili-povod-dlya-uzhestocheniya-cenzury.html
3. https://adilet.zan.kz/rus/docs/Z2200000118
4. Ананьева М. И. и др. Лингвостатистический анализ текстов экстремистской направленности // Ситуационные центры и информационно-аналитические системы класса 4i для задач мониторинга и безопасности (SCVRT1516). – 2016. – С. 210-213.
5. Drewnowski A. et al. Environments perceived as obesogenic have lower residential property values // American journal of preventive medicine. – 2014. – Т. 47. – №. 3. – С. 260-274.
6. Engman L. Automatic detection of cyberbullying on social media. – 2016.
7. León-Paredes G. A. et al. Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language // 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON). – IEEE, 2019. – С. 1-7.
8. Menini S. et al. A system to monitor cyberbullying based on message classification and social network analysis // Proceedings of the third workshop on abusive language online. – 2019. – С. 105-110.
9. Rosa H. et al. Automatic cyberbullying detection: A systematic review // Computers in Human Behavior. – 2019. – Т. 93. – С. 333-345.
10. Muneer A., Fati S. M. A comparative analysis of machine learning techniques for cyberbullying detection on Twitter // Future Internet. – 2020. – Т. 12. – №. 11. – С. 187.
11. Alotaibi M., Alotaibi B., Razaque A. A multichannel deep learning framework for cyberbullying detection on social media // Electronics. – 2021. – Т. 10. – №. 21. – С. 2664.
12. Илюкович Т.С. Принципы организации системы автоматического определения средств выражения кибербуллинга в англоязычных твитах (инженерный подход) // Фундаментальные и прикладные аспекты развития современной науки. – 2020. – С. 93-103.
13. Scrivens R., Frank R. Sentiment-based classification of radical text on the web //2016 European Intelligence and Security Informatics Conference (EISIC). – IEEE, 2016. – С. 104-107.
14. Attestog T., Kukulage S. P. Mapping extremist forums using text mining : дис. – Universitetet i Agder/University of Agder, 2013.
15. Акжолов Р. К., Верига А. В. Предобработка текста для решения задач NLP // Вестник науки. – 2020. – Т.1. – №. 3 (24). – С. 66-68.
16. Чернышова Г.Ю., Овчинников А.Н. Применение методов интеллектуального анализа данных для кластеризации текстовых документов // Информационная безопасность регионов. – 2015.

– № 4 (21). – С. 5-12.
17. Sourav, Saha. Students Anxiety and Depression Dataset [Электронный ресурс]. – URL: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset.

References

1. Karaush I. S., Kupriyanova I. E., Kuznecova A. A. Kiberbulling i suicidal'noe povedenie podrostkov // Suicidologiya. – 2020. – T. 11. – № 1 (38). – S. 117-129.
2. https://mk-kz.kz/social/2022/03/25/kiberbulling-realnaya-problema-ili-povod-dlya-uzhestocheniya-cenzury.html
3. https://adilet.zan.kz/rus/docs/Z2200000118
4. Anan'eva M.I. i dr. Lingvostatisticheskii analiz tekstov ekstremistskoi napravlennosti // Situacionnye centry i informacionno-analiticheskie sistemy klassa 4i dlya zadach monitoringa i bezopasnosti (SCVRT1516). – 2016. – S. 210-213.
5. Drewnowski A. et al. Environments perceived as obesogenic have lower residential property values // American journal of preventive medicine. – 2014. – T. 47. – № 3. – S. 260-274.
6. Engman L. Automatic detection of cyberbullying on social media. – 2016.
7. León-Paredes G.A. et al. Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the Spanish language // 2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON). – IEEE, 2019. – S. 1-7.
8. Menini S. et al. A system to monitor cyberbullying based on message classification and social network analysis // Proceedings of the third workshop on abusive language online. – 2019. – S. 105-110.
9. Rosa H. et al. Automatic cyberbullying detection: A systematic review // Computers in Human Behavior. – 2019. – T. 93. – S. 333-345.
10. Muneer A., Fati S.M. A comparative analysis of machine learning techniques for cyberbullying detection on Twitter // Future Internet. – 2020. – T. 12. – № 11. – S. 187.
11. Alotaibi M., Alotaibi B., Razaque A. A multichannel deep learning framework for cyberbullying detection on social media // Electronics. – 2021. – T. 10. – № 21. – S. 2664.
12. Ilyukovich T.S. Principy organizacii sistemy avtomaticheskogo opredeleniya sredstv vyrazheniya kiberbullinga v angloyazychnyh tvitah (Injenernyi podhod) // Fundametal and priklad aspects razvitiyz sovremennoi nauki – 2020. – S. 93-103.
13. Scrivens R., Frank R. Sentiment-based classification of radical text on the web //2016 European Intelligence and Security Informatics Conference (EISIC). – IEEE, 2016. – S. 104-107.
14. Attestog T., Kukulage S. P. Mapping extremist forums using text mining: dis. – Universitetet i Agder/University of Agder, 2013.
15. Akjolov R.K., Veriga A.V. Predobrabotka teksta dlya resheniya zadach NLP // Vestnik nauki – 2020. – T. 1. – № 3 (24). – S. 66-68.
16. Chernyshova G. YU., Ovchinnikov A. N. Primenenie metodov intellektual'nogo analiza dannyh dlya klasterizacii tekstovyh dokumentov // Informacionnaya bezopasnost' regionov. – 2015. – № 4 (21). – S. 5-12.
17. Sourav, Saha. Students Anxiety and Depression Dataset [Electronic resource]. – URL: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset.