ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ
МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ
INTERNATIONAL UNIVERSITY

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2022 (3) 2
Сәуір-маусым

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншіктенуші: «Халықаралық ақпараттық технологиялар университеті» АҚ (Алматы қ.).
Қазақстан Республикасы Ақпарат және әлеуметтік даму министрлігінің Ақпарат комитетінде – **20.02.2020** жылы берілген.
№ KZ82VPY00020475 мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: ақпараттық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологияларға арналған.
Мерзімділігі: жылына 4 рет.
Тиражы: 100 дана
Редакцияның мекенжайы: 050040, Алматы қ-сы, Манас к-сі, 34/1, 709-кабинет, тел: +7 (727) 244-51-09).
E-mail: ijict@iitu.edu.kz Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

---

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09). E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 3. Is. 2. Number 10 (2022). Pp. 59–66
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/*10.54309/IJICT.2022.10.2.006*

УДК 004.42
УДК 004.056.57

UTILIZING SENTIMENT ANALYSIS FOR CYBER SECURITY LEARNING

A.E. Namiyaly*, B.B. Valiyev, A.O. Sagymbekova, A.Zh. Adil

Azamat E. Namiyaly — Master's student, Department of Information Systems, International Information Technology University;
Bakhytzhan B. Valiyev — Master's student, Department of Computer Engineering, International Information Technology University;
Azhar O. Sagymbekova — Senior Lecturer, Department of Cybersecurity, International Information Technology University;
Altynai Zh. Adil — Master's student, Department of Computer Engineering, International Information Technology University.

Abstract. The main goal of this research article is to apply modern theoretical and practical knowledge in cybersecurity and develop a website by conducting sentiment analysis for cybersecurity learning on the developed platform. Electronic educational resources, information protection technologies, and data integrity preservation have been investigated to create the final product. The work has resulted in a secure platform for increasing user exposure in IT using the key machine learning models and hashing algorithms.

Keywords: Sentiment Analysis, Cyber Security, Learning, Website, Bag-of-words, Words, Server Log Files, SQL Injection, Cross-site-scripting, Genetic Algorithm, Database

For citation: A.E. Namiyaly, B.B. Valiyev, A.O. Sagymbekova, A.Zh. Adil. Utilizing sentiment analysis for cyber security learning // INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES. 2022. Vol. 3. Is. 2. Number 10. Pp. 59–66 (In Russ.). DOI: *10.54309/IJICT.2022.10.2.006.*

КИБЕРҚАУІПСІЗДІКТІ ЗЕРТТЕУ ҮШІН СЕНТИМЕНТАЛДЫ ТАЛДАУДЫ ҚОЛДАНУ

А.Е. Намиялы*, Б.Б. Валиев, А.О. Сагымбекова, А.Ж. Әділ

Намиялы Азамат Ержанович — «Ақпараттық жүйелер» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті;

Валиев Бахытжан Бауржанович — «Компьютерлік инженерия» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті;

Сагымбекова Ажар Орынгалиевна — «Киберқауіпсіздік» кафедрасының сениор лекторы;

Әділ Алтынай Жанарбекқызы — «Компьютерлік инженерия» кафедрасының магистранты, Халықаралық ақпараттық технологиялар университеті.

Аннотация. Ғылыми мақаланың негізгі мақсаты - киберқауіпсіздікте заманауи теориялық және практикалық білімдерді қолдану және платформада киберқауіпсіздікті үйрену үшін сентименталды талдау арқылы вебсайтты құру. Ақырғы өнімді жасау үшін электронды білім беру ресурстары, ақпаратты қорғау технологиялары мен деректердің тұтастығын сақтау әдісі зерттелді. Барлық жұмыстардың нәтижесінде ақпараттық технологиялар саласындағы пайдаланушылардың машиналық оқытудың ең маңызды модельдері мен хэштеу алгоритмдерін қолдана отырып, экспозициясын жоғарылату үшін қауіпсіз платформа жасалды.

Түйін сөздер: сентименнті талдау, киберқауіпсіздік, оқу, вебсайт, сөз қабы, лог-файлы, SQL-инъекция, сайтаралық скриптинг, генетикалық алгоритм, дерекқор

Дәйексөз үшін: А.Е. Намиялы, Б.Б. Валиев, А.О. Сагымбекова, А.Ж. Әділ. Киберқауіпсіздікті зерттеу үшін сентименталды талдауды қолдану// ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ. 2022. Том. 3. Is. 2. Нөмірі 10. 59–66 бет (орыс тілінде). DOI: 10.54309/ IJICT.2022.10.2.006.

ИСПОЛЬЗОВАНИЕ АНАЛИЗА ТОНАЛЬНОСТИ ДЛЯ ИЗУЧЕНИЯ КИБЕРБЕЗОПАСНОСТИ

А.Е. Намиялы*, Б.Б. Валиев, А.О. Сагымбекова, А.Ж. Әділ

Намиялы Азамат Ержанович — магистрант кафедры «Информационные системы», Международный университет информационных технологий;

Валиев Бахытжан Бауржанович — магистрант кафедры «Компьютерная инженерия», Международный университет информационных технологий;

Сагымбекова Ажар Орынгалиевна — сениор лектор кафедры «Кибербезопасность», Международный университет информационных технологий;

Әділ Алтынай Жанарбекқызы — магистрант кафедры «Компьютерная инженерия», Международный университет информационных технологий.

Аннотация. Основная цель научной статьи - применить современные теоретические и практические знания в области кибербезопасности и разработать вебсайт с использованием анализа настроений для изучения кибербезопасности на платформе. Электронные образовательные ресурсы, технологии защиты информации и сохранения целостности данных были исследованы для создания конечного продукта. В результате всей работы была разработана безопасная платформа для увеличения воздействия на пользователей в области информационных технологий с использованием самых ключевых моделей машинного обучения и алгоритмов хеширования.

Ключевые слова: анализ тональности, кибербезопасность, изучение, вебсайт, мешок слов, лог-файлы, SQL-инъекции, межсайтовый скриптинг, генетический алгоритм, базы данных

Для цитирования: А.Е. Намиялы, Б.Б. Валиев, А.О. Сагымбекова, А.Ж. Әділ. Использование анализа тональности для изучения кибербезопасности // МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ. 2022. Том. 3. Is. 2. Номер 10. Стр. 59–66 (на русском языке). DOI: 10.54309/IJICT.2022.10.2.006.

Introduction

Nowadays, the ability and knowledge of securing information technologies from intruders are highly priced. The ongoing epidemics forced many individuals to move away from classic real-world services and switch to online analogues (https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html). The increase in online activity has led to frequent encounters with criminals, such as scammers (https://www.cmu.edu/iso/aware/coronavirus-alerts/coronavirus-scam-awareness.html). Computer-aided learning development, which was conducted by our team, appears to be an effective measure in enhancing the awareness of citizens striving for security.

The replacement of all human activity by information technology requires readiness against intruder attacks. The awareness-raising safety training aims to help avoid any malicious attack on an information system by providing the information needed. The main goal is to develop a web portal to raise citizen awareness and help prepare for information security. Citizens looking for information technology security will be able to use awareness-raising safety training.

Materials and methods

Temporal sentiment analysis. Sentiment time series have been widely used for event prediction, such as political election prediction (Bermingham, 2011), and disaster event detection (Sakaki, 2010). Similarly, in cyberattack scenarios, the goal is to predict the sentiment score of individual messages as well as temporal sentiment changes over time. The goal is to provide insight into how social media users reverse the polarity of sentiment towards public events. Second, understand how sentiment can indicate upcoming attack events. Third, assess the consequences of an attack after its occurrence. To answer these questions, data D about relevant social media posts are collected by

querying relevant keywords. It occurs on social media during a specific time range τ, which covers the time interval before, during, and after the attack event. Our team plots the sentiment time series S = (^y1, y^2, ..., y^m) in chronological order using a previously learned sentiment predictor. We can also group and average sentiments at varying granular levels over time, such as on a per-day basis, to study sentiment variation.

Sentiment analysis is a type of textual order that manages emotional explanations. It is also known as assessment mining because it measures feelings to learn about open recognition. Opinion exploration and assessment mining are similar and are used in reverse throughout the process. It uses Natural Language Processing (NLP) to collect and validate assessment or conclusion words. SA is refined as recognizing people's views on a subject and its main points. The justification for the fame of assessment mining is that people like to consult with others to make intelligent contributions. Revealing emotional points of view in a large amount of useful information is an area of interest in information mining and NLP.

Enterprise systems generate many communication logs. They are used to track Internet traffic directed to a specific source. They create many logs, which are then collected and stored. System-generated logs show running system states and contain important information generated by the system to support diagnostics. Therefore, log analysis is a vital technique for detecting vulnerability scans, XSS, and SQLI attacks.

However, manually analysing a huge number of logs is unrealistic and impractical. Therefore, it is important to use machine learning to automate log analysis.

The objectives of this study can be summarized as follows:

- Analyse log data and detect scanning vulnerabilities.
- Detect XSS and SQLI attacks and check access log files for detection.

Sentiment trends predict cyberattacks before they happen. To study the behaviour of cyberattacks, historical data from the Apache log files is considered to develop a logistic regression classifier for each type of event. The data is divided into a training set and a test set. The training set includes the first 80 % of the time covered by the data, and the test set includes the remaining 20 %. In each test, logistic regression models predict whether each type of attack will occur. The results are shown in Figure 1. The model is much better at predicting XSS attacks or SQL injection attacks. It contrasts with the regular traffic type, with very high accuracy and recall scores for both.

```
******************accuracy_score******************
0.8984
*****************confusion_matrix*****************
[[556  10   4]
 [ 15 250  53]
 [ 18  27 317]]
**************classification_report***************
              precision    recall  f1-score   support

      Normal       0.94      0.98      0.96       570
SQL Injection      0.87      0.79      0.83       318
   XSS attack      0.85      0.88      0.86       362

    accuracy                           0.90      1250
   macro avg       0.89      0.88      0.88      1250
weighted avg       0.90      0.90      0.90      1250
```

Figure 1 - Results for predicting attacks using sentiment sensor

The logical design for using sentiment analysis for cybersecurity education is illustrated in Figure 2. The presented flowchart describes the entire process of the system, starting from the starting point and ending with the completion process.

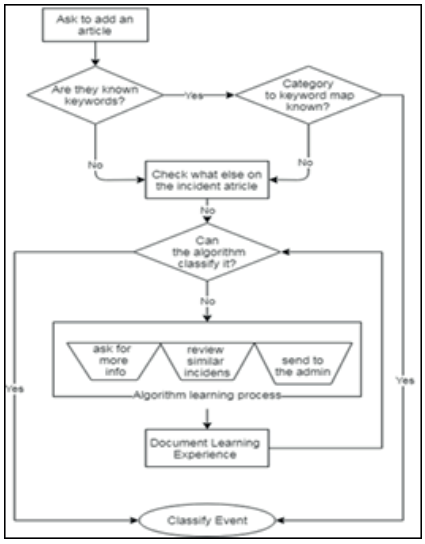

Figure 2 - Block-scheme of article adding process

The user starts the application process by rendering it. The system uploads articles and courses to the system. If the user is authorized, several functions are available: adding materials to favorites; participating in a discussion on a forum; and adding the user's article. Adding user materials is filtered and, provided that the material is suitable, the article is sent to the administrator for re-checking. If these processes are correct, the article is accepted, and the process ends. If the user is not authorized, the system displays only course materials and articles.

Figure 2 shows the article checking process by the system. First, a response to a user adding an article is sent to the system. If the article has known keywords and if the category of the keyword is known, the event is classified successfully. This process takes place during the filtering process. If there are no known keywords, the process starts to check for the presence of other components in the incident article and tries to categorize them. In case of failure, the learning process of the algorithm is started, which includes requesting additional information, viewing similar incidents, and sending them to the administrator. The learning process is then documented and stored.

The three main components that make up the system are the user interface or front-end, the server, and the database, as shown in the UML of the system in Figure 3. The user sends a request to the server through the interface to add a

new article. The request is processed and then saved in the database. After this stage, the server sends the result of the request to the user. The further extraction process is carried out by optimizing the user query using a genetic algorithm, which is extremely useful for rapidly growing databases.

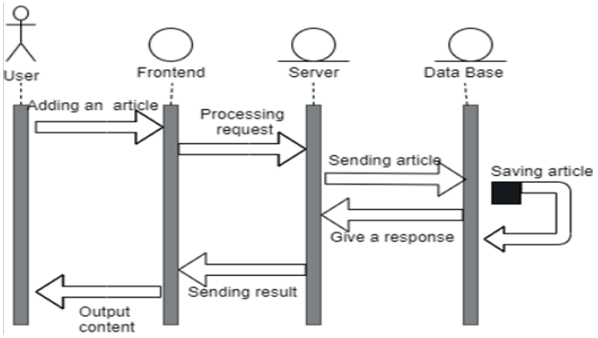

Figure 3 – UML diagram of the system

Structuring data. Analysing data requires it be in certain formats. In our case, the web logs were converted to data frames for easy analysis as shown in Figure 4. Dataframe columns refer to a host IP address, the date including day, month, year and time information, date offset, location in file system, status of success, and size.

	Host	Client_ID	User_Name	Date	Offset	Method	Loc	HTTPver	Status_Code	Size
0	207.46.13.214	-	-	18/Aug/2020:02:33:55	-700	GET	/robots.txt	HTTP/1.1	200	377
1	117.107.147.153	-	-	18/Aug/2020:02:35:33	-700	GET	/	HTTP/1.1	200	12883
2	117.107.147.153	-	-	18/Aug/2020:02:35:34	-700	GET	/css/style.css	HTTP/1.1	404	328
3	172.245.228.28	-	-	18/Aug/2020:02:35:59	-700	GET	/	HTTP/1.1	200	12839
4	172.245.228.28	-	-	18/Aug/2020:02:36:00	-700	GET	/datasets%20description/pe_malware/virusshare....	HTTP/1.1	200	1607

Figure 4 – *Data converted to a data frame*

Results and discussion

Attacks launched on servers. XSS attacks work by embedding script tags in URLs/ HTTP requests and tricking unsuspecting users into clicking on them. It ensures that malicious JavaScript is executed on the victim's machine (https://www.sans.org/reading-room/whitepapers/logging/detecting-attacks-web-applications-log-files-2074).

These attacks exploit the trust between the user and the server and the fact that there is no I/O validation on the server. Since it may reject JavaScript or other active code characters. Here we focus on the most common injection, SQL injection. For SQL injections to work, the attacker must go beyond the original SQL statement. This is usually done with a single quote (') or double dash (//). The single quote acts as a delimiter for the SQL query; double dash is a comment character in Oracle and MS SQL.

Feature building. Once the attacks were identified using regular expression patterns, they were compared against the corresponding data frames, resulting in a new column "State". It identifies the attack as "Xss_attack" OR "Sqli_attack" and normal as "Normal".

After creating new functions, i.e., URL_Length and number of parameters, the DataFrame was coded using Pandas layouts for method and state code columns. The logs were analysed in different ways and the results are reported according to the logs. We have applied a set of common algorithms, such as linear regression, support vector machine, etc and tested their performance.

It was observed that the Gradient Boosting and Random Forest Classifier performed the best algorithm for log attack classification.

Multiple accuracy metrics can be generated from the confusion matrix:

TP – True Positive; TN – True Negative; FP – False Positive; FN – False Negative.

• Accuracy = (all correct / all) = TP + TN / TP + TN + FP + FN. This is the general accuracy of the model. It is not useful, particularly when the classes are not equally represented i.e., the data is imbalanced in terms of the target classes.

• Precision = (true positives / predicted positives) = TP / TP + FP. Precision can be interpreted as the proportion of objects called positive by the classifier. At the same time, real positive, and recall shows what proportion of objects from all objects of a positive class the algorithm found.

• Recall = (true positives / all actual positives) = TP / TP + FN. Sensitivity states, out of all the actual malignant cases, how many cases were correctly identified as malignant. This is a class-level metric. Precision and Recall are critical class-level metrics. However, maximizing precision can pull down the value of recall, and vice versa (Shabbir, 2020).

Results according to the partially shown list in Figure 5 are: Normal – 90.6 %; SQL injection – 4.3 %; XSS attack – 5.1 %. It indicates that majority of the traffic belongs to the normal class, i.e. has no harm. However, remaining 9.4 % contain SQL injection and cross scripting patterns.

дата	HTTPver	Хост	логи	метод	Размер	Статус код	URL_Length	Результат
30/Jan/2020:02:42:36	HTTP/1.1	207.46.13.30	/se|f.logs/error.log.2018-04-14.gz	GET	4113	200	34	Normal
30/Jan/2020:02:42:40	HTTP/1.1	40.77.167.122	/se|f.logs/error.log.2018-10-10.gz	GET	5617	200	34	Normal
30/Jan/2020:02:42:57	HTTP/1.1	207.46.13.30	/se|f.logs/access.log.2019-03-03.gz	GET	24255	200	35	Normal
30/Jan/2020:02:46:08	HTTP/1.1	103.225.13.245	/security-data-analysis/lab_1/conn.log.zip	GET	14491960	200	42	Normal
30/Jan/2020:02:55:37	HTTP/1.0	46.149.86.51	/	GET	12623	200	1	Normal
30/Jan/2020:02:55:37	HTTP/1.0	46.149.86.51	/honeypot/honeypot%20-%20howto%20-%20bsides%20austin.pdf	GET	1895040	200	56	SQL Injection
30/Jan/2020:02:56:21	HTTP/1.0	119.40.82.210	/ftp://download.twlab.fot.se/dataset/	GET	257	404	37	XSS attack
30/Jan/2020:02:57:32	HTTP/1.1	47.111.229.152	/static/.gitignore	GET	294	404	18	Normal
30/Jan/2020:02:57:47	HTTP/1.1	47.111.229.152	/statics/js/admin_common.js	GET	294	404	27	SQL Injection
30/Jan/2020:02:58:53	HTTP/1.1	157.55.39.6	/robots.txt	GET	288	200	11	Normal

Figure 5 - Plot of the score of attacks in log files

Conclusion

The main aim of the project was to utilize sentiment analysis for cyber security learning. It is done to enhance the awareness of the citizens who attain to be secured. The educational platform was created, and it contains articles and video demonstrations about how to be protected from the various type of network attacks. The authorized users are able to share their own articles, which are controlled by the sentiment analysis model and automating this process, thus it became reachable to find out suitable information.

The application can be used to improve efficiency in analysing cybersecurity in everyday situations. For example, to find the latest cybersecurity news and analyse log files for containing any attacks, such as SQL injection or XSS attack.

The web platforms become important avenues for social communication. Thus, the study can serve as an indicator for measuring the extent to which a consumer is truly concerned about the cybersecurity issues and an identifier of the online behaviour effect. These results can help cybersecurity researchers plan and use web platforms to increase user awareness of cybersecurity issues in the future.

REFERENCES

Bermingham A., Smeaton A. (2020). On using twitter to monitor political sentiment and predict election results. In: SAAIP'11 1. Ella Koeze and Nathaniel Popper. The New York Times. [Electronic resource]. *The Virus Changed the way we Internet.* — April

2020. — URL: https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html

Carnegie Mellon University. [Electronic resource]. *Coronavirus (COVID-19) Scams.* December, 2020. URL: https://www.cmu.edu/iso/aware/coronavirus-alerts/coronavirus-scam-awareness.html

Meyer, "Detecting Attacks on Web Applications from Log Files," — 2008. [Online]. Available: https://www.sans.org/reading-room/whitepapers/logging/detecting-attacks-web-applications-log-files-2074.

Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: WWW'10

Shabbir Tayabali (December 11, 2020). [Electronic resource]: A simple guide to building a confusion matrix. URL:

https://blogs.oracle.com/ai-and-datascience/post/a-simple-guide-to-building-a-confusion-matrix.

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

ОТВЕТСТВЕННЫЙ РЕДАКТОР
Ералы Диана Русланқызы

КОМПЬЮТЕРНАЯ ВЕРСТКА
Жадыранова Гульнур Даутбековна

Подписано в печать 15.06.2022.
Формат 60x881/8. Бумага офсетная. Печать - ризограф.7,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59