

DOI 10.20535/2411-1031.2024.12.2.315743

УДК 004.8

ДМИТРО ЛАНДЕ,
ВІРА ГИРДА

ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ІДЕНТИФІКАЦІЇ ФЕЙКОВОЇ ІНФОРМАЦІЇ

Останні роки стали справжньою революцією у галузі штучного інтелекту завдяки появі великих мовних моделей (LLM), таких як GPT-4, Llama-3, Gemini та інших, які успішно застосовуються у широкому спектрі задач – від генерації текстів до аналізу даних. У цій статті ми розглянемо, як ці моделі можна ефективно використовувати для виявлення фейкової інформації.

У цій статті досліджено використання чат-бота ChatGPT для ідентифікації фейкової інформації в контексті кібербезпеки. За допомогою великої мовної моделі було створено рій віртуальних експертів, які генерували інформаційні повідомлення на тему кібербезпеки (фейкові та правдиві) та оцінювали їх як «фейк» чи «правда». Для аналізу було побудовано семантичну мережу, яку згодом візуалізувати за допомогою Gephi.

У дослідженні проаналізовано два масиви повідомлень: створені експертами-людьми та штучними експертами. Кожне повідомлення отримало оцінки, які були переведені у числовий формат для подальшого аналізу. Використовуючи відстань Хеммінга, було проведено перевірку результатів та визначено точність збігів між оцінками.

У результаті побудови семантичної мережі визначено ключові поняття в темі кібербезпеки та встановлено взаємозв'язки між ними. Роєм штучних експертів згенеровано масив повідомлень із фейковим і правдивим змістом, які було оцінено як ними, так і експертом-людиною. Аналіз Хеммінгової відстані між цими оцінками показав, що штучний інтелект має потенціал у виявленні фейкової інформації, проте на даному етапі його робота потребує корегування з боку людини.

Ключові слова: кібербезпека, великі мовні моделі, ChatGPT, класифікація інформації, штучний інтелект

Вступ. В останні роки значна частина інформації в кіберпросторі стала складнішою для обробки через зростаючу кількість спаму, ботів і фейкових новин. Це створює серйозні виклики для суспільства, адже відрізнити достовірну інформацію від недостовірної стає дедалі важче. У цьому контексті штучний інтелект, зокрема великі мовні моделі (LLM), відкривають нові можливості для автоматизованого аналізу інформації, що є критично важливим у сфері кібербезпеки.

Відповідно до схваленого розпорядженням Кабінету Міністрів України від 2 грудня 2020 року № 1556-р Концепції розвитку штучного інтелекту в Україні [1], пріоритетним завданням є використання технологій штучного інтелекту (ШІ) для забезпечення національної безпеки та конкурентоспроможності країни. Крім того, глобальні ініціативи, такі як “Цифровий компас 2030” [2], підкреслюють роль ШІ у цифровій трансформації різних галузей, включаючи сектор безпеки.

Технології ШІ, завдяки здатності швидко аналізувати великі обсяги даних і моделювати можливі сценарії, уже використовуються для моніторингу кіберзагроз, виявлення вразливостей і прогнозування кібератак. У цій статті розглядаються можливості застосування великих мовних моделей для вирішення ще одного ключового завдання – автоматизованого виявлення фейкової інформації, яка становить серйозну загрозу для кіберпростору.

Наявність у штучного інтелекту даних про попередні кібератаки дозволяє ідентифікувати аномальну поведінку системи перед потенційною атакою, сприяючи її прогнозуванню та запобіганню [3].

Швидкий розвиток технологій штучного інтелекту, особливо великих мовних моделей, таких як GPT-4, Llama-3 і Gemini, відкриває нові можливості для автоматизації процесів в освіті, науці та дослідницькій діяльності [4]. Зокрема, Generative Pre-trained Transformer (GPT) – це велика мовна модель, розроблена лабораторією OpenAI, що базується на алгоритмах обробки природної мови та методах навчання штучного інтелекту. Вона здатна генерувати тексти, класифікувати їх, створювати анотації, виконувати машинний переклад і відповідати на запитання [5].

Модель GPT-4 працює на основі глибинного навчання, використовуючи великі масиви даних. Завдяки цьому моделі можуть точно аналізувати текст, проводити семантичний аналіз та визначати ключові слова чи фрази, що є основою для ефективного використання їх у завданнях кібербезпеки та аналізу інформації.

Аналіз останніх досліджень та публікацій показав що з моменту випуску ChatGPT проведено безліч досліджень щодо його використання в освітньому процесі, медицині, науці, програмуванні та судовій діяльності [6]–[9]. У роботі [10], зокрема, досліджувались існуючі методики які використовують для розпізнавання фейкового контенту, які засновані методах штучного інтелекту.

У роботі [11] наведено результати дослідження щодо продуктивності системи ChatGPT в вирішенні 25 завдань, пов'язаних з проблемою обробки природної мови та аналітичних міркувань. До цих завдань входили:

- Двійкова класифікація даних як спам, сарказм чи агресія;
- Класифікація текстів за більшою кількістю класів та міток;
- Розпізнавання емоційного забарвлення текстів;
- Семантична анотація текстів;
- Перевірка на орфографію текстів тощо.

Слід зазначити, що результати які демонструвала велика мовна модель ChatGPT на початку запуску з часом стали більш чіткими, що вказує на те, що система швидко навчається.

Метою статті є дослідження можливостей систем штучного інтелекту, зокрема, ChatGPT, в задачах, пов'язаних з семантичним аналізом, класифікацією інформації як “фейк” чи “правда та її візуалізації”.

Виклад основного матеріалу дослідження

Постановка задачі для визначення фейкової інформації за допомогою великих мовних моделей, зводиться до того, що потрібно обрати один із напрямів дослідження, і як результат, побудувати семантичну мережу та візуалізувати її.

В ході дослідження “попросимо” систему ChatGPT згенерувати ряд правдивих та фейкових повідомлень, що стосуються кібербезпеки, для чого застосовуємо запит (Prompt 1).

Prompt 1 – назвіть 20 пар повідомлень що є правдивою інформацією в кібербезпеці 20 пар повідомлень що є фейковою інформацією. Кожна пара нехай видасться у форматі “повідомлення1”. Кожна така пара понять виводиться з нового рядка.

Результат:

Складний і унікальний пароль для кожного облікового запису значно знижує ризик злому.

Мультифакторна аутентифікація забезпечує додатковий рівень захисту облікового запису.

Регулярне оновлення програмного забезпечення знижує вразливість до кіберзагроз.

Фішингові атаки часто використовують підроблені електронні листи для викрадення облікових даних.

Антивірусна програма допомагає виявляти і видаляти шкідливе ПЗ з пристрою.

Шифрування даних гарантує їхню конфіденційність навіть у разі перехоплення.

Антивірусна програма гарантує абсолютний захист від усіх кібератак.

SSL-сертифікат робить будь-який сайт повністю безпечним для користувачів.

*Фішингові електронні листи завжди легко розпізнати за поганою граматикою.
Використання одного пароля для всіх облікових записів є зручним і безпечним.
Мультифакторна аутентифікація є зайвою, якщо пароль досить складний.
Регулярне оновлення програмного забезпечення не має впливу на безпеку системи.
тощо.*

Виконання багатьох таких запитів, незалежно від часу, буде генерувати різні відповіді (з погляду людини), ці відповіді можна сприймати як відповідь окремих віртуальних експертів.

Для того щоб отримати розширені варіанти, для формування повноти та точності відповідей, попередній запит варто запускати на виконання декілька разів, отримані результати додати до файлу разом із створеними та проаналізованими експертом людиною, та об'єднати відповіді в агрегований csv-файл.

Варто відмітити, що під час генерування відповіді на один і той ж самий запит, траплялись випадки коли штучний експерт відмовлявся надавати фейкову інформацію, спираючись на те, що використання фейкової інформації може завдати шкоди (Рисунок 1), так як існують випадки використання чат-боту для створення фішингових листів та шкідливого програмного забезпечення, але в більшості випадків, таки виконував запит та надавав потрібні відповіді.

ChatGPT

На жаль, надання фейкової чи хибної інформації щодо кібербезпеки може бути небезпечним і призвести до неправильного розуміння або навіть загроз для безпеки. Я можу допомогти з достовірною інформацією та порадами щодо кібербезпеки, але надання неправдивої інформації - не варіант. Якщо у вас є конкретні запитання або потреба у конкретній інформації щодо кібербезпеки, будь ласка, звертайтеся, і я допоможу якнайкраще!

Рисунок 1 – Відповідь-відмова на запит (промпт)

З вивчення отриманих результатів досліджень, видно що система з достатньою точністю класифікує різні види інформації, тому варто спробувати здійснити класифікацію інформації як “фейк” чи “правда”. Для дослідження проаналізуємо масив повідомлень, що стосуються питань кібербезпеки від експерта людини, та масив повідомлень, створений шляхом виконання запропонованого нами запиту в якості віртуального експерта. Ці масиви міститимуть як достовірну інформацію, так і фейкову (без міток), та збережемо все у один файл формату .csv.

Людина експерт здійснює аналіз всієї інформації, що міститься у файлі, та присвоює їй мітку “фейк” чи “правда”, теж саме робить – віртуальний експерт. Необхідно порівняти результати роботи віртуального та реального експертів та оцінити точність запропонованої класифікації.

Для перевірки отриманих результатів використаємо міру Хемінга.

Під відстанню Хемінга розуміють кількість позицій, у якій символи двох однакових за довжиною рядків, відрізняються.

Для цього нам потрібно перевести отримані оцінки експертів “фейк” та “правда” перевести у двійкову систему. 0 відповідатиме за оцінку “фейк”, 1 – “правда”.

$$d_H(x, y) = \sum_{i=1}^n \delta(x_i, y_i), \quad (1)$$

де $d_H(x, y)$ – відстань Хемінга між рядками x та y ;

n – довжина рядків x та y ;

$\delta(x_i, y_i)$ – індикаторна функція, яка дорівнює 1, якщо $x_i \neq y_i$ та 0, якщо $x_i = y_i$.

Мовою програмування Python було написано модуль, що застосовував наведену міру Хемінга, за допомогою якого було перевірено повідомлення на схожість оцінки експерта-людини та віртуального експерта. Як результат, виявилось, що текст, згенерований віртуальним експертом, та оцінений експертом людиною і віртуальним експертом співпадали з точністю 98,31%, а от різниця оцінок інформаційних повідомлень написаних людиною-експертом значно відрізняється, точність складає лише 81,82%. Разом з цим, це вказує на те, що системи штучного інтелекту показують ефективну роботу в різних галузях й можуть застосовуватись для визначення фейкової інформації, але при тісній взаємодії з експертами-людьми.

Наступним кроком, за допомогою великої мовної моделі GPT-4 з масиву повідомлень формується мережа слів (Рисунок 2), для чого застосовується Prompt 2.

Prompt 2 – на основі отриманого масиву повідомлень побудуй мережу слів: IoT пристрої завжди починаються з найсучаснішими захисними заходами. Ризиковий аналіз не має сенсу, оскільки загрози в кібербезпеці завжди однакові. Важливо завжди оновлювати програмне забезпечення для усунення вразливостей. Користування HTTPS забезпечує захист передачі даних через Інтернет. Перевіряйте посилання та відправників електронних листів перед натисканням на них. Використання коду з SMS і пароля гарно захищає ваш обліковий запис. Система управління інцидентами допомагає вчасно виявляти загрози для безпеки. Використання CDN допомагає розподілити трафік та захистити від атак. Система виявлення вторгнень допомагає виявити несанкціонований доступ до системи. Використання VPN забезпечує шифрування вашого інтернет-з'єднання. Встановлення правильних прав доступу обмежує можливість витоку інформації. Використання обмежень доступу та аутентифікації захищає підключені пристрої. Оновлення програмного забезпечення не впливає на безпеку системи. HTTPS не забезпечує безпеку, його можна легко обійти.

Результат:



Рисунок 2 – Мережа слів згенерована із масиву повідомлень

Prompt 3 – згенеруй масив повідомлень фейкової та правдивої інформації за допомогою представленої вище мережі слів.

Результат:

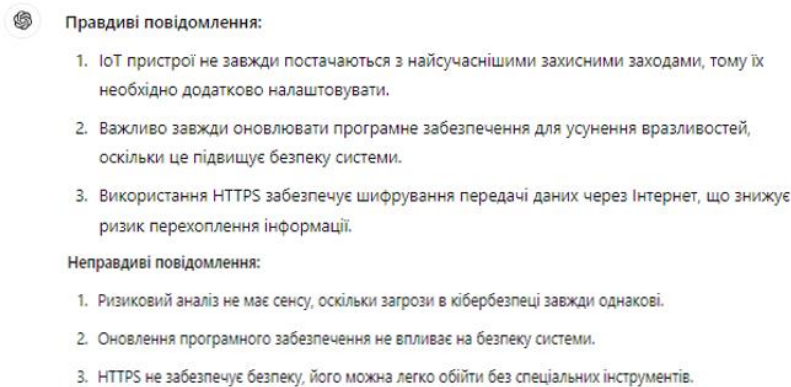
- 
- Правдиві повідомлення:**
1. IoT пристрої не завжди постачаються з найсучаснішими захисними заходами, тому їх необхідно додатково налаштувати.
 2. Важливо завжди оновлювати програмне забезпечення для усунення вразливостей, оскільки це підвищує безпеку системи.
 3. Використання HTTPS забезпечує шифрування передачі даних через Інтернет, що знижує ризик перехоплення інформації.
- Неправдиві повідомлення:**
1. Ризиковий аналіз не має сенсу, оскільки загрози в кібербезпеці завжди однакові.
 2. Оновлення програмного забезпечення не впливає на безпеку системи.
 3. HTTPS не забезпечує безпеку, його можна легко обійти без спеціальних інструментів.

Рисунок 3 – Відповідь CharGPT в результаті опрацювання *Prompt*

Як можна побачити, великі мовні моделі працюють у дві сторони, вони можуть як будувати мережі слів і генерувати з них текстові повідомлення, так і навпаки. Маючи мережу слів чи мережу термінів, за допомогою мовних моделей можна продукувати текст як фейковий, так і достовірний. Не дивно, що в період повномасштабного вторгнення збільшилась кількість фейкового контенту в кіберпросторі, оскільки маючи мережу термінів, будь-яка людина може згенерувати той текст, який був корисний в її цілях, та може нанести інформаційно-психологічний вплив на інших людей, які не вміють відрізнити правдива ця інформація чи ні, або ж не вміють перевіряти її достовірність.

Висновки. У цій роботі проведено дослідження можливостей використання великих мовних моделей, таких як GPT-4, для виявлення фейкової інформації в контексті кібербезпеки. Основними результатами є наступне:

1. За допомогою великої мовної моделі GPT-4 створено семантичну мережу на основі масиву повідомлень, які стосуються кібербезпеки. Визначено ключові поняття та встановлено взаємозв'язки між ними, що дало змогу краще зрозуміти структуру інформаційного простору.

2. Використовуючи метод “рою віртуальних експертів”, створено масив фейкових і правдивих повідомлень. Це дозволило оцінити можливості штучного інтелекту у створенні контенту та його класифікації.

3. Проведено порівняння оцінок повідомлень, здійснених експертами-людьми та штучними експертами, із застосуванням відстані Хеммінга. Було виявлено високу точність співпадіння (98,31%) для повідомлень, згенерованих штучними експертами, оцінених як експертами-людьми, так і самими моделями. Однак для повідомлень, створених людиною, точність збігу становила лише 81,82%, що вказує на необхідність корегування результатів.

4. Було показано, що великі мовні моделі здатні не лише генерувати текст на основі мережі слів, а й будувати ці мережі на основі текстового масиву, що відкриває широкі можливості для роботи з інформацією.

До перспективних напрямів подальших досліджень можна віднести:

– розробку класифікації інформації за трьома категоріями: “фейкова”, “правдива” та “сумнівна”. Остання категорія дозволить визначати випадки, де необхідне втручання експерта – людини.

– подальше вдосконалення методів взаємодії між великими мовними моделями та експертами для оптимізації роботи системи у кібербезпеці.

Отримані результати і оцінки демонструють, що великі мовні моделі є перспективним інструментом у сфері аналізу інформації, але їх застосування потребує відповідального підходу та співпраці з експертами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

- [1] Кабінет Міністрів України (2020, Груд. 12). *Розпорядження № 1556, Про схвалення Концепції розвитку штучного інтелекту в Україні*. [Електронний ресурс]. Доступно:

URL: <https://zakon.rada.gov.ua/laws/show/1556-2020-p#Text>. Дата звернення: Серп. 14, 2024.

- [2] 2030 Digital Compass: the European way for the Digital Decade. Brussels, Belgium: European Commission, 2021. [Online]. Available: <https://eufordigital.eu/library/2030-digital-compass-the-european-way-for-the-digital-decade/>. Accessed on: Oct. 19, 2024.
- [3] S. Hurzhii, “The special features of using the artificial intelligence in the matters of cybersecurity”, *Information and Law*, 2023, no. 4(47), pp. 207-216. doi: [https://doi.org/10.37750/2616-6798.2023.4\(47\).291669](https://doi.org/10.37750/2616-6798.2023.4(47).291669).
- [4] М. Мар’єнко, та В. Коваленко, “Штучний інтелект та відкрита наука в освіті”, *Фізико-математична освіта*, том 38, № 1, с. 48-53, 2023, doi: <https://doi.org/10.311105/2413-1571-2023-038-1-007>.
- [5] D. Lande, and L. Strashnoy, *GPT Semantic Networking: A Dream of the Semantic Web – The Time is Now*. Kyiv, Ukraine: Engineering, 2023.
- [6] D. Johnson et al., “Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model”, *Preprint. Res Sq. 2023; rs.3.rs-2566942*, 2023. doi: <https://doi.org/10.21203/rs.3.rs-2566942/v1>.
- [7] C.K. Lo, “What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature”, *Education Sciences*, vol. 13, no. 4, 410, 2023, doi: <https://doi.org/10.3390/educsci13040410>.
- [8] N. Surameery, and M. Shakor, “Use Chat GPT to Solve Programming Bugs”, *International Journal of Information technology and Computer Engineering*, no. 31. pp. 17-22, 2023, doi: <https://doi.org/10.55529/ijitc.31.17.22>.
- [9] М.В. Белова, та Д.М. Белов, “Імплементація штучного інтелекту в досудове розслідування кримінальних справ: міжнародний досвід”, *Аналітичне-порівняльне правознавство*, № 2, с. 448-453, 2023. doi: <https://doi.org/10.24144/2788-6018.2023.02.78>.
- [10] В. Праздніков, та І. Сугоняк, “Моделі та методи машинного навчання для розпізнавання фейкового контенту”, *Технічна інженерія*, том 92, № 2, с. 132-136, 2023. doi: [https://doi.org/10.26642/ten-2023-2\(92\)-131-136](https://doi.org/10.26642/ten-2023-2(92)-131-136).
- [11] J. Kocoń et al. “ChatGPT: Jack of all trades, master of none”, *Information fusion*, vol. 99, 101861, 2023, doi: <https://doi.org/10.1016/j.inffus.2023.101861>.

Стаття надійшла до редакції 03.11.2024.

REFERENCES

- [1] Cabinet of Ministers of Ukraine (2020, Dec. 12). *Order No. 1556, On Approval of the Concept of Artificial Intelligence Development in Ukraine*. [Online]. Available: <https://zakon.rada.gov.ua/laws/show/1556-2020-p#Text>. Accessed on: Aug. 14, 2024.
- [2] 2030 Digital Compass: the European way for the Digital Decade. Brussels, Belgium: European Commission, 2021. [Online]. Available: <https://eufordigital.eu/library/2030-digital-compass-the-european-way-for-the-digital-decade/>. Accessed on: Oct. 19, 2024.
- [3] S. Hurzhii, “The special features of using the artificial intelligence in the matters of cybersecurity”, *Information and Law*, 2023, no. 4(47), pp. 207-216. doi: [https://doi.org/10.37750/2616-6798.2023.4\(47\).291669](https://doi.org/10.37750/2616-6798.2023.4(47).291669).
- [4] M. Marienko, and V. Kovalenko, “Artificial intelligence and open science in education”, *Physical and mathematical education*, vol. 38, № 1, pp. 48-53, 2023, doi: <https://doi.org/10.311105/2413-1571-2023-038-1-007>.
- [5] D. Lande, and L. Strashnoy, *GPT Semantic Networking: A Dream of the Semantic Web – The Time is Now*. Kyiv, Ukraine: Engineering, 2023.
- [6] D. Johnson et al., “Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model”, *Preprint. Res Sq. 2023; rs.3.rs-2566942*, 2023. doi: <https://doi.org/10.21203/rs.3.rs-2566942/v1>.
- [7] C.K. Lo, “What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature”, *Education Sciences*, vol. 13, no. 4, 410, 2023, doi: <https://doi.org/10.3390/educsci13040410>.

- [8] N. Surameery, and M. Shakor, “Use Chat GPT to Solve Programming Bugs”, *International Journal of Information technology and Computer Engineering*, no. 31. pp. 17-22, 2023, doi: <https://doi.org/10.55529/ijitc.31.17.22>.
- [9] M. Belova, and D. Belov, “Implementation of Artificial Intelligence in Pre-Trial Investigation of Criminal Cases: International Experience”, *Analytical and Comparative Jurisprudence*, no. 2, pp. 448-453, 2023. doi: <https://doi.org/10.24144/2788-6018.2023.02.78>.
- [10] V. Prazdnikov, and I. Sugonyak, “Machine learning models and methods for recognizing fake content”, *Technical Engineering*, vol. 92, no. 2, pp. 132-136, 2023, doi: [https://doi.org/10.26642/ten-2023-2\(92\)-131-136](https://doi.org/10.26642/ten-2023-2(92)-131-136).
- [11] J. Kocoń et al. “ChatGPT: Jack of all trades, master of none”, *Information fusion*, vol. 99, 101861, 2023, doi: <https://doi.org/10.1016/j.inffus.2023.101861>.

DMYTRO LANDE,
VIRA HYRDA

USE OF LARGE LANGUAGE MODELS TO IDENTIFY FAKE INFORMATION

In recent years, the field of artificial intelligence has undergone a true revolution with the emergence of large language models (LLMs) such as GPT-4, Llama-3, Gemini, and others, which have been successfully applied across a wide range of tasks – from text generation to data analysis. This article examines how these models can be effectively used for detecting fake information.

This study explores the use of the ChatGPT chatbot for identifying fake information in the context of cybersecurity. Using a large language model, a swarm of virtual experts was created, which generated informational messages on the topic of cybersecurity (both fake and truthful) and assessed them as either “fake” or “true.” For analysis, a semantic network was constructed and subsequently visualized using Gephi.

The research analyzed two datasets of messages: one created by human experts and the other by artificial experts. Each message was rated and converted into a numerical format for further analysis. Using the Hamming distance, the results were validated, and the accuracy of matches between assessments was determined.

As a result of building the semantic network, key concepts in the field of cybersecurity were identified, along with the relationships between them. A swarm of artificial experts generated a dataset of messages with fake and truthful content, which was assessed both by the artificial experts themselves and by a human expert. Analysis of the Hamming distance between these assessments demonstrated that artificial intelligence has potential in detecting fake information; however, at this stage, its performance requires human oversight and adjustments.

Key words: cybersecurity, generative language models, ChatGPT, information classification, artificial intelligence.

Ланде Дмитро Володимирович, доктор технічних наук, професор, професор кафедри кібербезпеки і застосування інформаційних систем і технологій, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна, ORCID 0000-0003-3945-1178, dwlande@gmail.com.

Гирда Віра Анатоліївна, аспірант, Інститут спеціального зв’язку та захисту інформації Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”, Київ, Україна, ORCID 0000-0002-3858-4086, v.hyrda@kpi.ua

Lande Dmytro, doctor of technical science, professor, professor at the cybersecurity and application of information systems and technologies academic department, Institute of special communication and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

Hyrda Vira, postgraduate student, Institute of special communications and information protection of National technical university of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.