УДК 004.912 DOI: 10.31673/2412-9070.2022.035159

Ю. М. ТЕРТИЧНА, студентка;
О. В. НЕГОДЕНКО, канд. техн. наук, доцент;
О. С. ДЗЯДОВИЧ, аспірант,
Державний університет телекомунікацій, Київ

ВЕРИФІКАЦІЯ ДОКУМЕНТІВ НА ВЕБРЕСУРСІ МЕТОДОМ РОЗПІЗНАВАННЯ ТЕКСТУ

Висвітлено актуальні питання, пов'язані з верифікацією документів на вебресурсі за допомогою методів розпізнавання тексту. Встановлено, що існують різні методи та програмні рішення для цифрової перевірки особистості. Важливим є впровадження технологій штучного інтелекту і можливість доопрацювання під індивідуальні завдання, що підвищує ефективність бізнесу в рази. Вагомим недоліком є розгортання платформ на серверах, у такий спосіб підвищуючи фінансову затратність на інтеграцію та беручи в оренду потрібне програмне забезпечення. Розроблено метод, в якому використовується спосіб інтеграції у вигляді API. Застосовано модуль «Tesseract» для JavaScript, створено модуль для кращого та швидшого методу верифікації, розбитого на кілька блоків визначення коефіцієнтів для підтвердження ліквідності документа. Побудовано алгоритм для знаходження коефіцієнтів для підтвердження ліквідності документа та математичну модель автентифікації документів.

Ключові слова: верифікація; автентифікація; штучний інтелект; API; JavaScript; Erusted Identity Network

Вступ

Більшість фінансових компаній світу нині переходять на роботу з клієнтами в режимі онлайн, використовуючи для цього спеціально створені вебресурси [1]. Процес реєстрації клієнтів має деякі правила, які змінювати не можна. Клієнт сам уводить інформацію про себе в потрібні поля, але є люди, які можуть спробувати створити акаунт на іншу особу чи ввести дані невірно. Щоб такого не сталося, використовують документ підтвердження особистості, наприклад паспорт чи водійські права. Зазвичай вони подаються в електронному вигляді як сканкопії, але іноді постає проблема, коли під час подання інформації та документів їх потік може бути або дуже великим, або менеджери занадто довго перепровіряють усі дані, створюючи черги в роботі онлайн. Відтак з'являються методи автоматичної верифікації документів [2].

Проблеми правового регулювання електронно-цифрового підпису, як засобу захисту електронних документів, розглядали такі вчені, як М. І. Анохин, Ю. В. Бородакий, Н. П. Варновський, В. М. Глушков, М. В. Денісова, М. М. Дутов, А. В. Кобец, Г. І. Купріянова, А. Матвієнко, В. А. Онегов, І. А. Семаєв, В. А. Шахвердов, М. Н. Цивін, В. В. Ященко та ін. Системи онлайн-верифікації громадян досліджувались у працях І. П. Ситника, А. І. Мельниченка, R. Lienhart, A. Kuranov, V. Pisarevsky та ін.

У запропонованій статті розглядається метод верифікації перевірки посвідчувальних документів особи, що відіграє важливу роль у процесах відкриття нових банківських рахунків, оформлення та укладання фінансових угод із використанням технологій штучного інтелекту та методів автентифікації [3].

Основна частина

За умов зростання витоків даних, атак із захопленням облікових записів та крадіжок особистих даних, а також збільшення попиту на віддалені процеси через пандемію COVID-19 підприємствам потрібно виявляти шахрайство з особистими даними та визначати, чи є людина саме тією, за кого вона себе видає в інтернеті. Цифрові методи перевірки особистості, зокрема біометрична перевірка, розпізнавання осіб та перевірка цифрових ідентифікаційних документів, можуть допомогти компаніям, урядам та фінансовим установам перевірити особу людини в мережі [4]. Цифрова перевірка особи може бути застосовна в разі, якщо людина та її посвідчувальний документ не присутні фізично, а також нею можна послуговуватися для прискорення перевірки посвідчень особи, наприклад використання електронних воріт для сканування паспортів в аеропортах.

Цифрова перевірка особи є ключовим етапом у процесі відкриття рахунку та залучення клієнта. Переконавшись в особі заявника, фінансові установи можуть провести перевірку, щоб пересвідчитися, що заявник не є шахраєм, злочинцем, поганим актором або намагається здійснити аферу. Нині вже існують такі програми та ресурси, як, скажімо, «Дія». Ця програма розпізнає людину за фото, яке робиться в момент створення «Дія-підпису», використовуючи теорії розпізнавання образів та нейронну мережу. Проте в такому методі є одна проблема: якщо людина раніше не мала оцифрованої фотографії в державному реєстрі та не має телефона з фронтальною камерою, то вона авторизуватися не зможе.

Існує безліч різних типів цифрової перевірки особистості та рішень щодо її перевірки. Цифрові методи підтвердження особи працюють через порівняння того, що є в людини (наприклад, біометричних даних особи або документа, що посвідчує особу), з перевіреним набором даних (наприклад, даними, які зберігаються в державних органах, зокрема паспортними або біометричними даними, розміщеними на зареєстрованому мобільному телефоні користувача). Перевірка цифрової ідентифікації є зіставленням поданих даних із підтвердженим набором даних, аби впевнитися, що людина є тією самою особою, за яку себе видає. Існує багато різних методів перевірки цифрової ідентифікації, які працюють по-різному [5–7]. Ці методи охоплюють:

- *перевірку документа, що посвідчує особу*: перевіряється легітимність посвідчення особи (наприклад, прав водія, паспорта, державного посвідчення особи);
- *біометричну верифікацію*: використовується селфі для підтвердження того, що людина, яка показує посвідчення особи, є тією самою людиною, фото якої зображено на посвідчувальному документі;
- *виявлення жвавості*: визначається, чи селфі є справжнім, виявляючи підроблені атаки, такі як маски для обличчя або фотографії фотографій;
- *автентифікацію на основі знань* (KBA): генеруються питання з гаманця на основі інформації в особистому кредитному досьє заявника;
- *перевірку за одноразовим кодом* (OTP): передається одноразовий код через SMS або електронною поштою заявнику в процесі перевірки;
- *Erusted Identity Network*: використовуються наявні облікові дані заявника в іншого провайдера для перевірки його особи та уникнення будь-яких непорозумінь у процесі відкриття рахунку та реєстрації;
- *методи баз даних*: використовуються дані із соціальних мереж, автономних баз даних та інших джерел для перевірки інформації, наданої заявником.

Проаналізуємо метод верифікації посвідчувальних документів, оскільки перевірення документів — це метод цифрової перевірки особи, який застосовується для підтвердження легітимності документа, що посвідчує особу заявника (наприклад, паспорт, посвідчення особи, права водія тощо). Метою є збирання, вилучення та аналіз ідентифікаційних даних, аби з'ясувати справжність виданих державними органами документів, що посвідчують особу. Це допомагає відрізнити справжнє від шахрайського. За допомогою автоматизованої перевірки посвідчувальних документів особи такі документи можуть бути перевірені також і в режимі реального часу та протягом кількох секунд. Завдяки вбудованій камері на мобільному або портативному пристрої технологія захоплює зображення документа, що посвідчує особу заявника. Далі для аналізу зображення застосовуються штучний інтелект і передові алгоритми автентифікації, щоб оцінити автентичність і визначити, чи є ідентифікаційний документ справжнім або підробленим [8–10].

Перевірка ідентифікаційних документів дає змогу посвідчувати особу клієнта в цифровому вигляді та в режимі реального часу, незалежно від того, чи перебуває користувач у відділенні, чи віддалено. Для постачальників фінансових послуг технологія прискорює відкриття банківського рахунку, процес реєстрації, кредитування та фінансування, одночасно захищаючи від шахрайства та знижуючи відсоток відмов від послуг у цифрових банківських каналах та каналах онлайн-банкінгу.

Сьогодні на ринку послуг існує така компанія, як «Evergreen», котра надає великий обсяг програмного забезпечення для різних типів бізнесів — від малого до великого. Тут пропонуються готові, надійні, перевірені рішення, відповідні міжнародним практикам і стандартам галузі. Продукти «Evergreen» допомагають підвищити ефективність бізнесу в рази завдяки інтеграції з бізнес-системами компанії, упровадженому штучному інтелекту і можливості доопрацювання під індивідуальні завдання. Але в цій компанії платформу розгорнуто на своїх серверах, а отже, інтеграція та взяття в оренду потрібного програмного забезпечення коштує чималих грошей, тож не кожен може собі її дозволити [11–13].

Нині застосовують: TensorFlow — фреймворк машинного навчання, на якому працівники компанії створили нейронну мережу, Faster-RCNN-Inception-V2 — модель, що оптимально підходить для подальшого навчання. Для розпізнавання написів використовують API Google Cloud Vision, а для пошуку найкращої відповідності серед можливих результатів — повнотекстовий пошук, обчислення відстані Левінштейн і Soundex (алгоритм порівняння двох рядків за їх звучанням, що встановлює однаковий індекс для рядків, котрі мають схоже звучання англійською мовою) [14; 15].

Але такі технології потребують більш високих затрат, а для бізнесу, який тільки-но виходить на ринок, це занадто дорого, тому розроблено та запропоновано метод, який має легший спосіб інтеграції у вигляді API.

Скориставшись модулем «Tesseract» для JavaScript, було розроблено модуль для кращого та швидшого методу верифікації, розбитого на кілька блоків знаходження коефіцієнтів для підтвердження ліквідності документа.

Розглянемо далі алгоритм його роботи.

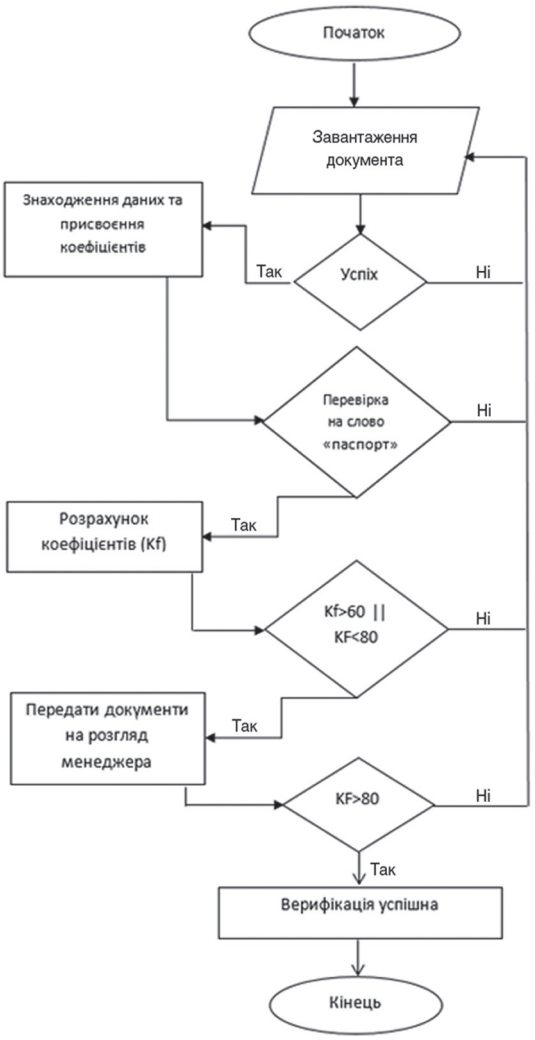
Крок 1. Заповнення полів з особистою інформацією:

1. ПІБ.
2. Номер документа.
3. Серія документа.
4. Дата народження.
5. Дата видачі.
6. Орган видачі.
7. ІПН.

Крок 2. Після введення даних потрібно завантажити фото документа. Зображення має бути або відскановане, або сфотографоване на камеру 8 Мп із розширенням 800×600. Зазначені характеристики є мінімальними, далі скрипт виділяє на зображенні текст:

8. Паспорт (коеф. «обов'язковий»).
9. Дата народження (коеф. «0,01»).
10. ПІБ (коеф. «0,25»).
11. Серія документа (коеф. «0,2»).
12. Номер документа (коеф. «0,04»).
13. Дата видачі (коеф. «0,1»).
14. Орган видачі (коеф. «0,2»).
15. ІПН (коеф. «0,2»).

Крок 3. Після отримання всіх показників потрібно обчислити коефіцієнт та дістати результат (приклад роботи наведено на рисунку).



Алгоритм визначення коефіцієнтів для підтвердження ліквідності документа

Розглянемо математичну модель алгоритму, реалізовану в такий спосіб. Маємо кілька показників, тобто коефіцієнтів:

- $Kf(p)$ — обов'язковий, без якого алгоритм завершується;
- $Kf(pib)$ — ПІБ, дані перевіряються на збіг з уведеними (0,01);
- $Kf(nd)$ — номер документа, дані перевіряються на збіг з уведеними (0,01);
- $Kf(sd)$ — серія документа, дані перевіряються на збіг з уведеними (0,01);
- $Kf(db)$ — дата народження, дані перевіряються на збіг з уведеними (0,01);
- $Kf(dv)$ — дата видачі, дані перевіряються на збіг з уведеними (0,01);
- $Kf(ov)$ — орган видачі, дані перевіряються на збіг з уведеними (0,01);
- $Kf(in)$ — ІПН, дані перевіряються на збіг з уведеними (0,01).

У разі, якщо $Kf(p)$ існує, то виконується така формула:

$$Res = \frac{(Kf(pib) + Kf(nd) + Kf(sd) + Kf(db) + Kf(dv) + Kf(ov) + Kf(in)) \cdot 100}{10},$$

де Res — результат відносності об'єктів на документі та підтвердження його.

Висновки

Наукова новизна здобутих результатів дослідження методу верифікації документів забезпечує доцільність його використання для автоматизованої перевірки документів, що посвідчують особу та можуть бути перевірені в режимі реального часу та протягом кількох секунд. Використовуючи вбудовану камеру на мобільному або портативному пристрої, а також технологію штучного інтелекту, запропонований алгоритм автентифікації дає змогу аналізувати зображення, щоб оцінити автентичність для визначення того, чи є ідентифікаційний документ справжнім або підробленим.

Розглянутий метод переважає меншими фінансовими затратами в обслуговуванні та налаштуванні, технологія не потребує додаткових серверів чи встановлення програмного забезпечення. Також запропонована технологія пришвидшує процес верифікації та її точність, працюючи лише з документами.

Список використаної літератури

*1. **Wang T., Wu D. J., Coates A.** End-to-End Text Recognition with Convolution Neural Networks // IEEE Conf. Pattern Recognition. 2012. P. 3304–3308.*

*2. **Zhu Y., Sun J., Naoi S.** Recognizing Naural Scene Characters by Convolutional Neural Network and Bimodal Image Enhancement // Workshop on Camera-Based Document Analysis and Recognition. 2012. P. 69–82.*

*3. **Chen X., Yang J., Zhang J.** Automatic Detection and Recognition of Signs from Natural Scenes // IEEE Image Processing. 2004. Vol. 13, no. 1. P. 87–99.*

*4. **Васильєв В. Н., Гуров І. П., Потапов А. С.** Математичні методи і алгоритмічне забезпечення аналізу та розпізнавання зображень в інформаційно-телекомунікаційних системах. 2008. С. 22–28.*

*5. **Koo H., Kim D. H.** Scene Text Detection via Connected Component Clustering and Non-text Filtering // IEEE Processing. 2013. Vol. 22, no. 6. P. 2295–2304.*

*6. **Hanif S. M., Prevost L.** Text Detection and Localization in Complex Scenes using Constrained Adaboost Algorithm // IEEE Conf. Document Analysis and Recognition. 2009. P. 1–5.*

*7. **Mosleh A., Bouguila N., Hamza A.** Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform // British Machine Vision Conference. 2012. P. 1–2.*

*8. **Rosset, Zhu, Hastie.** Boosting as a Regularized Path to a Maximum Margin Classifier // Journal of Machine Learning Research. 2004. №5. P. 941–973.*

*9. **Garcia C., Apostolidis X.** Text Detection and Segmentation in Complex Color Images // IEEE Conf. Acoustics, Speech and Signal Processing. 2000. P. 2326–2330.*

*10. **Karatzas D., Antonacopoulos A.** Text Extraction from Web Images Based on a Split-and-Merge Segmentation Method Using Colour Perception // IEEE Conf. Pattern Recognition. 2004. P. 634–637.*

*11. **Lienhart R., Kuranov A., Pisarevsky V.** Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection // MRL Technical Report. 2002. P. 12–15, 17.*

*12. **Hunspell** spell chacker dictionary [Електронний ресурс]. URL: http://hunspell.sourceforge.net/.*

*13. **Jaderberg M., Simonyan K.** Synthetic data and artificial neural networks for natural scene text recognition. NIPS Deep Learning Workshop, 2014.*

*14. **Lucas S. M., Panaretos A.** ICDAR 2003 robust reading competitions: entries, results, and future directions. IJDAR, 2005. P. 105–122.*

*15. **Krizhevsky A., Sutskever I.** Imagenet classification with deep convolutional neural networks. NIPS, 2012.*

Yu. Tertychna, O. Nehodenko, O. Dziadovych

VERIFICATION OF DOCUMENTS ON THE WEB RESOURCE USING THE METHOD OF TEXT RECOGNITION

The article is devoted to the coverage of current issues related to the verification of documents on the web resource using text recognition methods. It has been established that there are various methods and software solutions for digital identity verification. It is important to introduce artificial intelligence technologies and the possibility of refinement for individual tasks, which ii increases business efficiency many times. An important drawback is the deployment of platforms on servers, due to which the financial cost of integration and renting the necessary software increases. This article discusses the verification method of checking identity documents, which plays an important role in the processes of opening new bank accounts, signing and concluding financial agreements, using artificial intelligence technologies and authentication methods A method has been developed that uses the method of integration in the form of an API. The «Tesseract» module for JavaScript was used, a module was developed for a better and faster verification method, divided into several blocks of finding coefficients to confirm the liquidity of the document. An algorithm for finding coefficients to confirm the liquidity of a document and a mathematical model of document authentication have been developed. When using the built-in camera on a mobile or portable device, as well as artificial intelligence technology and the proposed authentication algorithm, it is possible to analyze the image in order to obtain an authenticity assessment to determine whether the identification document is fake or genuine. It has been established that this method prevails due to lower financial costs in maintenance and installation, the technology does not require additional servers or software installation. Also, the proposed technology speeds up the verification process and its accuracy, working only with documents.

Keywords: verification; authentication; artificial intelligence; API; JavaScript; Rusted Identity Network.