УДК 004.4:004.8
DOI https://doi.org/10.32782/EIS/2025-108-9

АНАЛІЗ ВИКОРИСТАННЯ МОДЕРНОВИХ EMBEDDING-МОДЕЛЕЙ ДЛЯ АВТОМАТИЧНОГО ПОШУКУ ПІДСАНКЦІЙНИХ ОСІБ НА ПРИКЛАДІ САНКЦІЙНОГО СПИСКУ OFAC SDN

Павленко Єгор Вікторович,
аспірант кафедри інформаційних технологій та комп'ютерної інженерії
Національного технічного університету «Дніпровська політехніка»
ORCID ID: 0009-0004-0600-3090

Гнатушенко Володимир Володимирович,
доктор технічних наук, професор,
завідувач кафедри інформаційних технологій та комп'ютерної інженерії
Національного технічного університету «Дніпровська політехніка»
ORCID ID: 0000-0003-3140-3788
SCOPUS ID: 6505609275

У статті досліджується ефективність використання сучасних текстових ембедингів і варіацій їх навчання для «наївного» автоматичного пошуку підсанкційних осіб у фінансових транзакціях на прикладі санкційного списку OFAC SDN. Зростання вимог до комплаєнс-процедур та недоліки традиційних методів скринінгу (низька точність, обмежена масштабованість, фрагментарність даних) підкреслюють актуальність дослідження. Авторами запропоновано архітектуру системи, яка інтегрує векторні бази даних з API для Google Embeddings та Gemini API, використовуючи «наївний» підхід до обробки даних без складних процедур попередньої підготовки даних. Проведено експериментальну валідацію із застосуванням чотирьох стратегій векторизації (Stringified JSON, Stringified Non-Empty, Flattened Key-Value, Flattened Non-Empty) та різних типів завдань для ембединг-моделей. Було порівняно результати з існуючими системами скринінгу, включаючи власну реалізацію OFAC. Отримані дані свідчать, що хоча «наївний» підхід забезпечує впевнені результати для подальшої обробки людиною або LLM (у рамках RAG-систем), але для повністю автоматизованих транзакційних систем, що працюють за пороговим значенням, потрібна більш складна попередня підготовка даних. Показано, що традиційні fuzzy-matching-алгоритми (Soundex, Jaro-Winkler), які застосовані у пошуку на сайті OFAC, забезпечують високу точність для імен, що точно збігаються із записами у санкційному списку. Проте їх ефективність знижується за транслітерації та варіацій у транслітерації, при цьому діапазони показників для істинно позитивних і хибнопозитивних результатів перекриваються, що ускладнює визначення єдиного граничного значення. Дослідження підкреслює потенціал модернових ембедингів для підвищення точності та масштабованості санкційного скринінгу, але вказує на необхідність подальшої оптимізації.

Ключові слова: *санкційний скринінг, текстові ембединґи, штучний інтелект, семантичний пошук, фонетичний пошук, обробка природної мови, комплаєнс.*

Pavlenko Iegor, Hnatushenko Volodymyr. An Analysis of Modern Embedding Models for the Automated Identification of Sanctioned Individuals: Evidence from the OFAC SDN List

This article investigates the effectiveness of modern text embeddings and variations in their training for «aive» automatic detection of sanctioned individuals in financial transactions, using the OFAC SDN sanctions list as a case study. The increasing demands on compliance procedures, along with the limitations of traditional screening methods (low accuracy, limited scalability, fragmented data), highlight the relevance of this research. The authors propose a system architecture that integrates vector databases with the Google Embeddings API and the Gemini API, employing a «naive» approach to data processing that avoids complex preprocessing steps. An experimental validation was conducted using four vectorization strategies («Stringified JSON,» «Stringified Non-Empty,» «Flattened Key-Value,» «Flattened Non-Empty») and different task types for embedding models. The results were compared with existing screening systems, including OFAC's own implementation. The findings indicate that, while the «naive» approach provides reliable results for further human or LLM-assisted processing (within RAG systems), fully automated transaction systems operating based on a threshold value require more sophisticated data preprocessing. It is shown that traditional fuzzy-matching algorithms (Soundex, Jaro-Winkler), as applied in the OFAC website search, achieve high accuracy for names that exactly match entries in the sanctions list. However, their effectiveness decreases with transliteration and variations thereof, and the score ranges for true positives and false positives overlap, complicating the selection of a single threshold value. The study highlights the potential of modern embeddings to improve the accuracy and scalability of sanctions screening, but also emphasizes the need for further optimization.

Key words: *sanctions screening, text embeddings, artificial intelligence, semantic search, phonetic search, natural language processing, compliance.*

Актуальність проблеми. В умовах посилення міжнародних санкцій і вимог до комплаєнс-процедур банківські установи змушені забезпечувати ефективний і точний скринінг клієнтів на предмет їх присутності в санкційних списках. Традиційним методам пошуку підсанкційних осіб притаманні висока частка хибнопозитивних результатів, недостатня швидкість обробки великих обсягів даних та обмежена здатність адаптуватися до різних форматів списків.

Системи автоматизованого скринінгу санкційних списків у банківських установах мають низку суттєвих недоліків. Традиційні алгоритми точного та нечіткого пошуку не завжди ефективно ідентифікують підсанкційних осіб через варіативність написання імен, використання різних алфавітів та транслітерації. Санкційні списки різних юрисдикцій (ЄС, США, ООН, національні списки) мають різні структури, формати та рівні деталізації інформації, що ускладнює їх інтеграцію та пошук. Поточні системи не враховують контекстну інформацію про особу, що призводить до помилкових ідентифікацій. Обмежена масштабованість, зростання кількості санкційних списків та їх оновлення потребують швидкої адаптації та обробки суттєвих обсягів даних.

Дослідження застосовує текстові ембединґи SOTA-класу, використовує векторні бази даних, що дає змогу: забезпечити адаптивність до різноманітних форматів санкційних даних; підвищити точність ідентифікації через урахування більш глибокого семантичного контексту за рахунок великої розмірності ембединґ векторів; створити масштабовану архітектуру для обробки великих обсягів транзакційних даних.

Результати можуть бути використані для підвищення ефективності комплаєнс-процедур; зменшення ризиків порушення санкційних режимів; скорочення часу обробки фінансових транзакцій; зниження операційних витрат на ручну верифікацію спрацювань системи; покращення якості обслуговування зменшенням помилкових блокувань.

Аналіз останніх досліджень і публікацій. Публікацій про автоматизовані системи фінансового моніторингу і запобігання відмиванню грошей доволі багато. Існують публікації і в Україні. Але вони більше присвячені або розробленню концепцій, або таких частин системи, як інтерфейс користувача. В іноземному сегменті більшу частину робіт присвячено системам боротьби із шахрайством та відмиванням грошей. Багато робіт розглядають загальні або теоретичні підходи [1–3]. Завдання модернового семантичного пошуку розглядається в [4; 5].

Мета дослідження – оцінка ефективності застосування сучасних моделей текстових ембединґів для побудови системи автоматизованого виявлення підсанкційних осіб із метою забезпечення високої точності та швидкодії скринінгу фінансових транзакцій. Дослідження передбачає використання даних санкційних списків у первинному вигляді (as is), без застосування складних процедур попереднього очищення чи нормалізації. Розробити архітектуру системи, що поєднує векторні бази даних для семантичного пошуку з реляційними/документними сховищами, які містять базові дані санкційних списків, а також API для ідентифікації потенційних збігів із підсанкційними особами. Забезпечити інтеграцію гетерогенних санкційних списків із різних джерел та юрисдикцій. Розробити методи уніфікації та гармонізації даних санкційних списків, використовуючи спрощені підходи, що не передбачають складних процедур попереднього очищення та нормалізації.

Виклад основного матеріалу дослідження. Архітектура системи, що запропонували автори, така (рис. 1):

1. Рівень збору та обробки даних. Складається з модулів парсингу санкційних списків на мові програмування Python 3 та бази даних MongoDB для зберігання структурованих санкційних даних.

2. Рівень векторизації та індексування. Складається з модулів на мові програмування Python для генерації векторних представлень інформації санкційних списків із використанням Google Embeddings API та векторної бази даних для зберігання та пошуку векторизованої інформації.

3. Рівень обробки запитів. Складається з Web-серверу на фреймворку Express для обробки HTTP-запитів та модулів на мові програмування JavaScript для отримання векторного представлення даних, що аналізуються. Для подальшого аналізу даних за використання на реальних системах зроблено логування процесу обробки запитів до бази даних MongoDB.

За розробленим авторами алгоритмом (рис. 2) виконується таке: прийом HTTP POST-запиту в JSON-форматі; векторизація запиту через Google Embeddings API; векторний пошук у векторній БД за косинусною подібністю; формування відповіді з топ-K результатів та повної інформації. До результату пошуку додаються початковий запит та дані для моніторингу, відповідь форматується та повертається як результат HTTP-запиту.

Інтеграція гетерогенних санкційних списків. Санкційні списки різних юрисдикцій

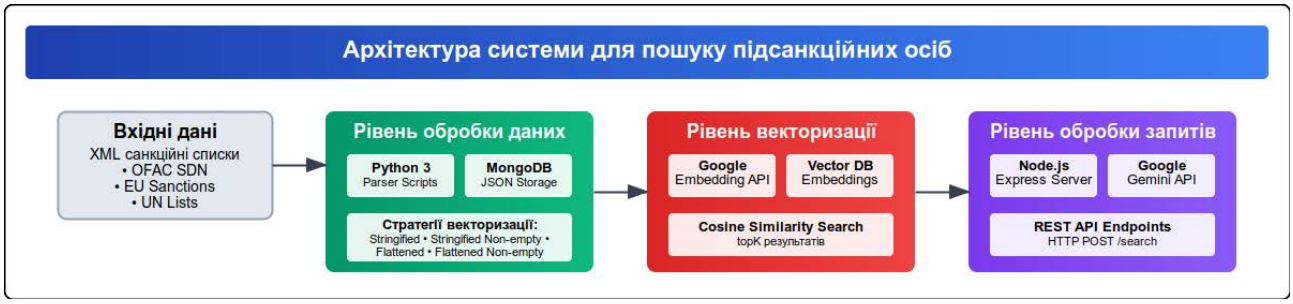

Рис. 2. Алгоритм роботи системи

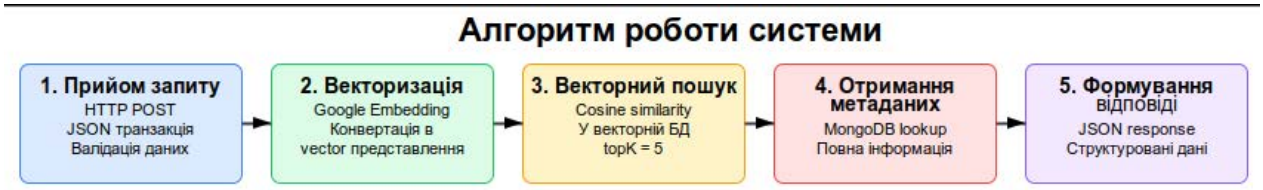

Рис. 2. Алгоритм роботи системи

характеризуються значною гетерогенністю форматів та структур даних. Виклики інтеграції включають: форматні відмінності; переважне використання XML-форматів із різними схемами; відмінності в структурі вкладених елементів; різні способи кодування символів та транслітерації; семантичні відмінності; варіативність полів імен та ідентифікаторів; різні рівні деталізації адресної інформації.

Для подолання гетерогенності санкційних списків запропоновано перевірку «наївного» підходу. Для реалізації пошуку за варіаціями даних про особу з уникненням специфічної для кожного санкційного списку попередньої обробки даних записи списків використовуються повністю для побудови векторного виду запису. Пошук осіб відбувається виключно за рахунок семантичного пошуку за допомогою ембединґ-моделі спеціально претренованої для пошуку в документі більшого розміру за запитом меншого розміру. Для ембединґ-моделей від Google [6] таку можливість імплементовано передаванням із запитом до API [7] параметру task_type зі значеннями: RETRIEVAL_DOCUMENT – для отримання вектору документа, в якому буде відбуватися пошук, та RETRIEVAL_QUERY – для отримання вектору тексту для пошуку подібності [8].

Етап 1. Парсинг санкційного списку. Створено Python-скрипт для парсингу XML-файлів із конвертацією у JSON-формат. Дані записуються зі збереженням початкової структури XML, що забезпечує підтримку вкладених структур (адреси, ідентифікатори), збереження всіх атрибутів та метаданих, можливість відтворення оригінального формату.

Етап 2. Підготовка до векторизації за чотирма стратегіями обробки даних (рис. 3):

Stringified JSON – перетворення JSON-об'єкта в рядок за допомогою функції Python json.dumps().

Stringified Non-Empty – попереднє видалення пустих полів. І потім перетворення JSON-об'єкта в рядок за допомогою функції Python json.dumps().

Flattened Key-Value – перетворення ієрархічної структури на плаский список «ключ – значення».

Flattened Non-Empty – попереднє видалення полів із пустими значеннями і потім перетворення на плаский список.

Приклад stringified вигляду: «{«name»: «John Doe», «dob»: «1980-01-01», «program»:[«SDN» «EU»], «address»:{«country»: «Russia», «city»: «Moscow»}}».

Приклад «плаского» вигляду: «name: John Doe/ndob: 1980-01-01/nprogram: SDN, EU/naddress.country: Somecountryname/naddress.city: Somecityname».

Додатково для порівняння з «наївним» підходом відокремлено дані: прізвище + ім’я + у разі наявності по-батькові (або middle name). Пошук виконується лише відносно повного імені підсанкційної особи без інших даних про особу. Векторизацію виконано зі значенням параметра task_type = SEMANTIC_SIMILARITY, ембединґи для пошуку семантично подібних текстів.

Рис. 3. Стратегії підготовки даних для векторизації

Стратегії векторизації санкційних даних.

Були отримані такі вектори для кожного запису санкційного списку:

stringified-версії з параметром API task_type = RETRIEVAL_DOCUMENT

stringified-версії з параметром API task_type = SEMANTIC_SIMILARITY

flattened-версії з параметром API task_type = RETRIEVAL_DOCUMENT

flattened-версії з параметром API task_type = SEMANTIC_SIMILARITY

для повного імені з параметром API task_type = RETRIEVAL_DOCUMENT

для повного імені з параметром API task_type = SEMANTIC_SIMILARITY

Розмір векторів 768. Отримані дані векторів записано як точки до векторної бази даних. Пошук подібності векторів відбувався засобами векторної бази даних із використанням коефіцієнта «косинусної подібності» (*cosine similarity).*

Експериментальна валідація стратегій. Порівняльний аналіз ефективності стратегій векторизації проводився на реальних даних санкційного списку OFAC SDN. Як дані пошуку (потенційні санкційні особи) було використано синтетично згенеровані дані з прізвищ реальних підсанкційних осіб та прізвищ, які в санкційному списку не зустрічаються, – для реалізації різних випадків.

Використана модель ембединґів Google text-embedding-004 розрахована на англійську мову [1], тому використано англомовний санкційний список OFAC.

Запропоновані варіанти обробки даних і запитів та їх взаємодію подано в табл. 1. Для аналізу використано набір імен який, складається з двох груп імен (загалом 110): 1) імена, що є схожими на наявні у санкційному списку (загалом 77); 2) імена, що є схожими на відсутні у санкційному списку (загалом 33). Ці дві групи поділяються на підгрупи. Для осіб у санкційному списку три підгрупи:

1) підгрупа, де імена написано англомовним написом латинськими буквами, як є у санкційних списках, лише деякі імена написано повністю як у запису санкційного списку, а деякі мають тільки ім'я та прізвище, наприклад не мають частини імені «по батькові», або мають неповне ім'я, наприклад дві частини повного імені замість чотирьох частин для імен іспанського походження (13 імен);

2) підгрупа, де імена написано англомовним написом латинськими буквами, але імена транслітеровано. Із різними варіаціями транслітерації, не тільки такими, що відповідають постанові Кабінету Міністрів України з транслітерації, а й навмисно хибними, для перевірки різних можливих варіацій даних (39 імен);

3) підгрупа, де використано імена, написані англомовним написом латинськими буквами, з написанням, як вони фігурують у санкційних списках, але не з основних полів для запису імен, а з полів для псевдонімів (25 aliases).

Список осіб, які відсутні у санкційному списку, поділяється на дві підгрупи аналогічно двом першим підгрупам для санкційних осіб:

1) підгрупа, де імена написано англомовним написом латинськими буквами, з «правильним» написанням імені (9 імен);

2) підгрупа, де імена написано англомовним написом латинськими буквами, але імена транслітеровані. Із різними варіаціями транслітерації, не тільки такими, що відповідають постанові Кабінету Міністрів України з транслітерації [4], а й навмисно хибними, для перевірки різних можливих варіацій даних. Також до цієї групи входять імена, які є схожими до імен підсанкційних осіб (24 імені).

Пошук за псевдонімами (aliases), згідно з рекомендаціями OFAC, не вважається обов'язковим, але включений до аналізу, передусім для перевірки, наскільки так званий «наївний підхід» із майже відсутньою попередньою обробкою даних може виявляти записи санкційних осіб не за основними іменами, а за псевдонімами. Отже, пошук за цією групою виконується і для інших підходів для виявлення, чи ці підходи враховують цю можливість, та для порівняння результатів.

Для пошуку за транслітерованими та зміненими іменами було підготовано 39 варіацій імен підсанкційних осіб. Розподілення варіацій не є рівномірним. Не три варіації для кожного

Таблиця 1

Варіанти обробки даних списку і запитів та їх взаємодія

	Іноземні імена та імена з прийнятою транслітерацією які входять до санкційного списку (перевірка на позитивні та хибно негативні)	**Варіації в транслітерації імен для осіб, що входять до санкційного списку (перевірка на позитивні та хибно негативні)**	**Іноземні імена та імена з прийнятою транслітерацією, які не входять до санкційного списку (перевірка на хибно позитивні та негативні)**	**Варіації в транслітерації імен для осіб, що не входять до санкційного списку (перевірка на хибно позитивні та негативні)**
«stringified» task_type = RETRIEVAL_ DOCUMENT	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			
«stringified» task_type = SEMANTIC_ SIMILARITY	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			
«flattened» task_type = RETRIEVAL_ DOCUMENT	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			
«flattened» task_type = SEMANTIC_ SIMILARITY	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			
Full Name task_type = RETRIEVAL_ DOCUMENT	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			
FullName task_type = SEMANTIC_ SIMILARITY	запити векторизовано з task_type = RETRIEVAL_ QUERY			
	запити векторизовано з task_type = SEMANTIC_ SIMILARITY			

імені, як можна подумати з відношення кількостей (13 до 39). Деякі імена мали більше варіацій, аніж інші. Кількість варіацій на ім'я залежала від варіацій імен іншомовного походження в українській мові та їх транслітерацій.

На даному етапі аналізу був використаний випадок пошуку за прізвищем та ім'ям особи. Для такого пошуку і враховуючи варіанти, результати показали таке. Ембединґ-вектори, отримані за різними task_types-параметрами, дійсно відрізняються за значеннями. Зі порівняння stringified-запису з flattened-записом, векторизованих за типом RETRIEVAL_DOCUMENT і запитом RETRIEVAL_QUERY, загалом варіант flattened показує гірші результати. Спостерігаються лише окремі кращі результати для flattened-записів у разі пошуку осіб за псевдонімами.

Цікаві результати отримано для випадку, коли записи векторизовано за типом SEMANTIC_ SIMILARITY, а пошуковий рядок – за типом RETRIEVAL_QUERY. Такий варіант дає більш точні результати в top-5 (більш точні для flattened-версії, ніж для stringified) для пошуку за псевдонімом, а також для важких випадків пошуку, наприклад пошуку за двома частинами імені, коли ім'я складається з чотирьох і більше частин. Хоча коефіцієнти подібності в такому наборі даних не є високими.

Під час аналізу пошуку псевдонімів випадок, де запис складається тільки з імені, порівняння має сенс лише тоді, коли псевдонім є подібним до основного імені, оскільки у записах для імен використовувалися тільки основні імена (псевдоніми не оброблялись як записи для пошуку). Із результатів можна побачити, що для таких

випадків коефіцієнти подібності є відповідно високими.

Усі пари пошуків показали погані результати для пошуку імені, яке має чотири і більше частин за двома частинами, як може бути урізане таке ім'я в деяких країнах. Наприклад, пошук за ім'ям Maria Hernandez, коли повне ім'я HERNANDEZ PULIDO Maria Elda або SALAZAR HERNANDEZ Maria Alejandrina. Також для таких типів імен, коли використовується ім'я матері для чоловічого імені (як у прикладі в попередньому абзаці), результати показали багато хибних спрацювань на санкційних осіб жінок із Росії та Білорусі з відповідними іменами. Хоча можна було очікувати, що модернові ембединґ-моделі, які побудовані не за лінгвістичною подібністю, а дистильовані з великих мовних моделей і можуть мати в собі простори загальних, а не лише лінгвістичних сенсів, мають відрізняти ці випадки.

Загальний висновок порівняння різних типів векторного пошуку. Хоча підходи з мінімальною підготовкою даних типу stringified або flattened демонструють достатньо стабільні результати у векторному пошуку, у контексті побудови системи автоматичного скринінгу постає інше завдання. Основним викликом тут є необхідність отримання єдиної числової метрики ймовірності, яку можна порівняти з установленим пороговим значенням: якщо значення нижче порогу, сповіщення не генерується, якщо вище – система повідомляє про можливу наявність підсанкційної особи у транзакції. При цьому необхідно враховувати не лише випадки пошуку осіб, що дійсно перебувають у санкційних списках, а й перевірку осіб, які не є підсанкційними, але мають імена, подібні до імен осіб зі списків. Оскільки векторний пошук завжди повертає найближчі значення, неминучими є хибнопозитивні спрацювання. Засобом їх відокремлення є порівняння обчислених показників подібності: для справді позитивних результатів вони будуть вищими, тоді як для хибнопозитивних – нижчими. У разі коли потрібно отримати одне інтегральне значення для порівняння, найкращі результати в межах можливого діапазону подібності забезпечує пошук типу SEMANTIC_SIMILARITY, застосований до даних, підготовлених лише на основі повного імені.

Таким чином, можна стверджувати, що для систем, де інформація подається на подальшу обробку людиною, підходи з «наївним пошуком» без складної попередньої обробки можуть розглядатися для використання. Також потребує подальшого дослідження можливість використання таких спрощених систем як retrieval-частини системи Retrieval-Augmented Generation (RAG), де подальший аналіз виконується не людиною, а за допомогою великої мовної моделі (LLM). Такі системи можуть використовуватися не для аналізу транзакцій, а для аналізу клієнтів під час реєстрації або перед наданням їм послуг для виконання принципу «знай свого клієнта» (KYC), або в інших випадках, коли ми маємо інформацію не лише про ім'я особи, а й про офіційні документи, місце проживання, дату народження, іншу інформацію, яка також є в записах про санкційних осіб. Водночас для автоматизованих систем обробки транзакцій, які працюють за граничним значенням, виглядає необхідною більш складна підготовка санкійних записів, передусім із виокремленням як основних, так і додаткових імен.

Аналіз реалізації пошуку підсанкційних осіб від OFAC. Office of Foreign Assets Control (OFAC) має свою публічно доступну реалізацію пошуку підсанкційних осіб [9]. З інформації сайту OFAC на сторінці відповідей на поширені запитання [10] алгоритм пошуку в санкційному списку OFAC використовує нечіткий (fuzzy) пошук за полем імені. Цей інструмент поєднує фонетичний алгоритм (Soundex) та алгоритм порівняння рядків (Jaro–Winkler) для обчислення оцінки збігу імен. Спочатку пошук фільтрує кандидатів (мають починатися з тієї ж літери та мати ≥50% схожість за редагуванням), потім обчислює дві оцінки: (а) Jaro–Winkler для повного імені, та (б) Jaro–Winkler плюс Soundex для кожної частини імені окремо. Вища із цих оцінок використовується як підсумкова оцінка збігу. У 2021 р. OFAC додала третій (не розкритий) алгоритм fuzzy matching для покращення результатів, однак Soundex і Jaro–Winkler залишилися в основі роботи пошуку.

Алгоритми: Soundex (фонетичне кодування) і Jaro–Winkler (відстань між рядками). Стратегія пошуку: порівняння повного імені та частин імені окремо. Поріг: результати повинні пройти початкову фільтрацію (та сама перша літера та ≥50% схожості за відстанню редагування). Нечітка логіка: тільки поле імені обробляється з використанням нечіткого порівняння, інші поля перевіряються на точну відповідність [11]. Оновлення 2021 р.: додано новий (неопублікований) алгоритм fuzzy matching, при цьому Soundex та Jaro–Winkler залишилися у використанні. Дослідження в галузі перевірки санкцій підтверджують загальний підхід. Наприклад, Kim & Yang [12] розглядають методи перевірки санкцій, відзначаючи, що міри відстані редагування (наприклад, Левенштейн)

є широко використовуваними в пошуку імен. Ці дослідження підкреслюють, що алгоритми схожості рядків (Levenshtein/Jaro–Winkler) у поєднанні з фонетичним кодуванням (Soundex або Metaphone) є стандартом у перевірках санкційних списків. Ні Kim & Yang, ні Nino та ін. не описують безпосередньо власну реалізацію OFAC, але підтверджують використання відповідних методів fuzzy matching у цій сфері.

Для аналізу пошуку у реалізації від OFAC було використано той самий набір імен, що й для аналізу семантичного пошуку за допомогою ембедингів та для аналізу реалізації у фінансовій установі.

Пошук за використаними псевдонімами (aliases) на сайті OFAC показав мінімальне значення оцінки збігу (score) 98 і найчастіше значення дорівнює 100, із чого можна зробити висновок, що пошук за псевдонімами відбувається. Унаслідок того, що використані значення псевдонімів збігаються з тими, що наявні у записах санкційного списку OFAC SDN, або дуже подібними (у випадку значень 98), було отримано високі значення оцінок збігу імен, переважно 100.

Пошук від OFAC за іменами підсанкційних осіб показав упевнені результати. Для кожного імені (13 імен), яке було використано в пошуку, першим результатом з коефіцієнтом подібності 100 було знайдено відповідний запис, навіть у разі використання для пошуку тільки імені та прізвища, за наявності в запису також імені по батькові. У випадку імен іспанського походження пошук тільки за двома частинами імені з чотирьох наявних частин у санкційному запису також показав перші два результати з коефіцієнтом подібності 100 для двох імен у санкційному списку, де присутні дві частини імені, за якими відбувався пошук. Також упевнений результат пошуку з коефіцієнтом подібності 100 було отримано для імен кириличного походження.

Пошук за транслітерованими та зміненими іменами підсанкційних осіб показав результати, які можна поділити на п’ять груп за значенням подібності (табл. 2).

У першій групі з 16-ти варіацій першим записом у результаті була особа, з імені якої складалися варіації подібності. Коефіцієнт подібності був у діапазоні 92–98, як для випадків пошуку за трьома частинами з трьох наявних у санкційному запису, так і для випадків пошуку за двома частинами імені з трьох у списку.

Друга група з дев’яти варіацій також показала в першому результаті ім’я, з якого складалися відповідні варіації, але з меншими коефіцієнтами подібності – 78–90. До цих варіацій увійшли варіації на імена SHAHRIARI, Behnam, KIRIYENKO, Vladimir Sergeevich, KARIMIAN, Mohammad Sadegh; також можна зазначити, що різні варіації на ім’я ROLDUGIN, Sergei Pavlovich увійшли як до першої, так і до другої групи.

Третя група – це дві варіації пошуку Volodymyr Kyryenko та Volodymyr Kyryienko для імені KIRIYENKO, Vladimir Sergeevich, де відповідна особа була не перша у результатах пошуку, з коефіцієнтом 82 та 84, а першою була особа зі санкційного списку з іменем VOLODYMYR BILYY з коефіцієнтом подібності 91. Отже, якщо враховувати тільки першу відповідь, маємо невірний результат пошуку.

Четверта група – дев’ять варіацій. Із них шість варіацій імені Hernandez, Maria та три варіації імені GHANI, Esmail, де відповідні імена не були знайдені в перших п’яти записах результату, а коефіцієнт подібності перших записів у відповідях (не правильні відповіді) був у діапазоні 83–88.

П’ята група з трьох варіацій, для яких не було знайдено відповідних імен у перших п’яти записах відповіді, коефіцієнт подібності був 71–79. Дві з цих варіацій мали першою буквою символ не латинського, але кириличного алфавіту, третя варіація цієї групи Yvan Abramov для імені Ivan Abramov, незважаючи на відмінність тільки в одну букву, показала погані результати пошуку. Це може бути пов’язано з тією особливістю алгоритму пошуку, що, як вище написано в його опису, під час пошуку кандидатів

Таблиця 2

Значення коефіцієнтів подібності пошуку OFAC за транслітерованими та зміненими іменами підсанкційних осіб

	Кількість	Мін.	Макс.	Середнє	Середнє без мін. та макс.	Медіана
1 група (вірні значення)	16	92	98	95	95	96
2 група (вірні значення)	9	78	90	85	85	85
3 група (вірні значення)	2	82	84	83	-	83
3 група (невірні значення)	2	91	91	91	-	91
4 група (невірні значення)	9	83	88	85	85	84
5 група (невірні значення)	3	71	79	75	74	74

алгоритм шукає таких, що починаються з однакової букви. Усі три варіації цієї групи мають особливості в першій букві імені.

Виходячи з того, що для непідсанкційних осіб для всіх реалізацій, що досліджуються, пошук виконується і здебільшого повертає деякі результати, фактично результати пошуку для групи несанкційних осіб є хибно позитивними.

Для осіб із групи непідсанкційних першої підгрупи (імена з вірним написанням) у пошуку від OFAC маємо значення подібності від 80 до 99. Із дев'яти пошукових запитів шість знаходяться в діапазоні 84–87. Один має значення 80 (табл. 3). Для пошуку Thomas Anderson знайдене значення THOMAS, Мае має високе значення – 91. Для пошуку Elena Petrova знайдене значення TIMCHENKO, Elena Petrovna, має дуже високе значення – 99. Схоже прізвище Petrova було уподоблено до по батькові Petrovna, імовірно, саме це зумовлює високе значення коефіцієнта подібності відносно інших випадків.

До другої підгрупи непідсанкційних осіб (варіації написання імен) додано імена, які схожі на імена підсанкційних осіб. Підгрупа складається з 24 імен (табл. 4). Здебільшого результати цієї підгрупи відповідають результатам попередньої першої підгрупи.

Збереглися аномально високі оцінки для пошуку з різним написанням імені Elena Petrova. Можна відзначити результати BENGUIAT JIMENEZ, Alberto David, CABELLO RONDON, Jose David, CAMPBELL LICONA, David Elias, CHAVARIN PRECIADO, David Alonso – усі з оцінкою подібності 100 для пошуку David Li. Для імен, які схожі з іменами підсанкційних осіб у різних варіаціях, отримані оцінки подібності від 75 під час пошуку, який співпадає тільки за прізвищем, наприклад Oleg Fridman, Oleh Fridman, до 89 за більшої схожості імен, наприклад отримано значення FRIDMAN, Mikhail Maratovich з оцінкою 89 для пошуку Fridman Mykhail Davydovych.

Висновки розподілу коефіцієнтів із наявними даними в реалізації OFAC.

1. Для випадків пошуку за іменами та псевдонімами з написанням, яке відповідає написанню в записах санкційного списку, можна засвідчити впевнений пошук із коефіцієнтами подібності в діапазоні значень 98–100.

2. Для випадків пошуку за транслітерованими та зміненими іменами підсанкційних осіб маємо випадки як вірного, так і невірного пошуку.

При цьому діапазони коефіцієнтів подібності для вірних та помилкових груп перетинаються, ба більше, деякі значення невірних коефіцієнтів сягають значення 91, при тому, що для досить великої групи вірних значень середнє значення знаходиться біля, а медіанне дорівнює 85.

3. При цьому медіана хибно позитивних значень для правильного написання імен дорівнює значенню 86. Медіана для хибнопозитивних значень групи імен, що є транслітеровані з помилками або схожі на імена підсанкційних осіб, є 83,5.

4. Якщо потрібно вибрати одне число як граничне значення, число з діапазону 83–85 виглядає таким, яке буде розділяти найбільші групи вірних та хибних значень. При цьому залежно від цілей (менше хибнопозитивних значень або менше хибнонегативних значень) можна вибирати менше або більше значення граничного коефіцієнта, але якщо написання імен будуть мати помилки або імена будуть схожі на імена підсанкційних осіб, хибні випадки пошуку будуть траплятися. Виходячи з того, що пошук вірогіднісний, фактично він завжди повертає значення, і хибнопозитивні значення можуть мати високий коефіцієнт подібності, це може призводити до багатьох випадків хибнопозитивних

Таблиця 3

Значення коефіцієнтів подібності пошуку OFAC за іменами не підсанкційних осіб (хибнопозитивні випадки)

	Кількість	Мінімум	Максимум	Середнє	Середнє без мін. та макс.	Медіана
єдина група	9	80	99	87	86	86

Таблиця 4

Значення коефіцієнтів подібності пошуку OFAC за іменами непідсанкційних осіб, які транслітеровані з помилками або схожі на імена підсанкційних осіб (хибнопозитивні випадки)

	Кільк.	Мін.	Макс.	Середнє	Серед. без мін. та макс.	Медіана
єдина група	24	70	100	84	84	84

спрацювань, що може порушувати автоматичну обробку, тому що позитивні спрацювання зазвичай відправляються на підтвердження людиною. А також може призводити до помилкових зупинок операцій для осіб, не причетних до санкцій.

5. Слід також відзначити нестабільність роботи пошукового сайту OFAC: трапляються випадки, а подекуди й тривалі періоди некоректного функціонування публічного сервісу, коли на запити не повертаються результати.

Аналіз наявної реалізації. Наявна реалізація одного з банків України для пошуку підсанкційних осіб працює з іменами осіб як кириличнім, так і транслітерованим написом. Виходячи з того, що модель, використана для семантичного пошуку, працює тільки з англійською мовою, для порівняння будемо використовувати тільки ту частину наявної реалізації, яка працює з транслітерованими іменами. Використовуючи наявну реалізацію за принципом «чорна коробка», орієнтуючись на дані, які доступні в результатах пошуку, можна стверджувати, що в наявній реалізації в частині обробки транслітерованих імен використовуються методи, які порівнюють імена як повні, так і розбиваючи їх на частини. Якщо повне ім'я має прізвище, ім'я, по батькові, буде відбуватися пошук входження в повному імені, яке записане в санкційному списку, окремо для повного імені і потім, окремо для прізвища, окремо для ім'я, окремо для по батькові. Пошук реалізовано у двох режимах: за повним збігом та за текстовою подібністю, за якої голосні літери трактуються як довільні символи. Результати цієї реалізації мають дискретний характер із характерними значеннями 100; 66,66; 50; 33,33; 25 (табл. 5–7). Коефіцієнт співпадіння залежить від співвідношення кількості знайдених співпадаючих частин і кількості частин імені в тексті для пошуку. Наприклад, як можна бачити в табл. 5, для тексту пошуку Ivan Abramov і підсанкційної особи ABRAMOV, Ivan Nikolayevich удалося отримати максимальний коефіцієнт подібності 66,66, тому що в повному імені присутнє ім'я по батькові, а в пошуковому тексті воно відсутнє.

Виходячи з дискретного характеру результатів, можливі значення граничних коефіцієнтів, які будуть мати практичне значення, також є дискретним набором. Наприклад, граничні

Таблиця 5

Пошук Behnam Shahriyari для підсанкційної особи SHAHRIYARI, Behnam

Знайдений запис	Коефіцієнт	Тип	Текст за яким знайдено подібність
SHAHRIYARI, Behnam	100,00	Співпадіння	Behnam Shahriyari
SHAHRIYARI, Behnam	100,00	Фонетична подібність	B*hn*m Sh*hr***r*
BEHNAM SHAHRIYARI TRADING COMPANY	50,00	Співпадіння	Behnam Shahriyari
BEHNAM SHAHRIYARI TRADING COMPANY	50,00	Фонетична подібність	B*hn*m Sh*hr***r*
BAHNAM, Hikmat Jarjes	33,33	Фонетична подібність	B*hn*m

Таблиця 6

Пошук Ivan Abramov для підсанкційної особи ABRAMOV, Ivan Nikolayevich

Знайдений запис	Коефіцієнт	Тип	Текст за яким знайдено подібність
ABRAMOV, Ivan Nikolayevich	66,66	Співпадіння	Abramov Ivan
ABRAMOV, Ivan Nikolayevich	66,66	Фонетична подібність	*br*m*v *v*n
ABRAMOV, Valeri Vyacheslavovich	33,33	Співпадіння	Abramov
ABRAMOV, Valeri Vyacheslavovich	33,33	Фонетична подібність	*br*m*v

Таблиця 7

Приклад пошуку для непідсанкційної особи Thomas Anderson

Знайдений запис	Коеф.	Тип	Текст за яким знайдено подібність
DIXON, Ian Thomas	33,33	Співпадіння	Thomas
DIXON, Ian Thomas	33,33	Фонетична подібність	Th*m*s
THOMAS, Mae Toussaint	33,33	Співпадіння	Thomas
THOMAS, Mae Toussaint	33,33	Фонетична подібність	Th*m*s
SUAREZ ANDERSON, Lourdes Benicia	25,00	Співпадіння	Anderson
SUAREZ ANDERSON, Lourdes Benicia	25,00	Фонетична подібність	Anderson

значення 70 або 80 не будуть створювати фактичної різниці.

Висновки. Дослідження показує, що модернові текстові ембединґи мають значний потенціал для підвищення точності та масштабованості систем санкційного скринінгу. Вони дають змогу враховувати більш глибокий семантичний контекст та адаптуватися до різних форматів даних без надмірної попередньої обробки. Проте для повністю автоматизованих систем обробки транзакцій, де критично важливим є отримання єдиної метрики вірогідності для порівняння з пороговим значенням, необхідна складніша підготовка даних, зокрема з чітким виокремленням основних та додаткових імен. Це допоможе зменшити кількість хибнопозитивних спрацювань та підвищити надійність прийняття рішень.

Подальші дослідження можуть бути зосереджені на інтеграції з Retrieval-Augmented Generation (RAG), що дасть змогу враховувати ширший контекст для підвищення точності ідентифікації. Варто дослідити можливість використання моделей мультимовних або донавчених на специфічних даних для покращення обробки транслітерованих та варіативних імен.

ЛІТЕРАТУРА:

1. Vu N.T.H., Wisniewski T.P., Skovoroda R. Textual Analysis in Financial Research and Practice: A Literature Review. 2025. DOI: https://dx.doi.org/10.2139/ssrn.5393334

2. Oztas B. et al. Transaction monitoring in anti-money laundering: A qualitative analysis and points of view from industry.Future Generation Computer Systems. 2024. T. 159. P. 161–171. DOI: https://doi.org/10.1016/j.future.2024.05.027

3. Mallela I.R. et al. Deep Learning Techniques for OFAC Sanction Screening Models.International Journal of Computer Science and Engineering. 2023. T. 12(2). P. 89–114.

4. Masoudi A. AggroDoc-Semantic Search using Vector Databases and Large Language Models. Luleå: Luleå University of Technology. 2024.

5. Kabasta S. Multi-language semantic search model for free-text insurance clauses. Brno: Masaryk University. 2025.

6. Embeddings. Gemini API docs. Google AI for Developers. URL: https://ai.google.dev/gemini-api/docs/embeddings (дата звернення: 21.07.2025).

7. Embeddings. API Reference. Google AI for Developers. URL: https://ai.google.dev/api/embeddings (дата звернення: 21.07.2025).

8. Enhancing your gen AI use case with Vertex AI embeddings and task types. Google Cloud. 2024. URL: https://cloud.google.com/blog/products/ai-machine-learning/improve-gen-ai-search-with-vertex-ai-embeddings-and-task-types (дата звернення: 21.07.2025).

9. OFAC Sanctions List Service. Office of Foreign Assets Control URL: https://sanctionslist.ofac.treas.gov/Home/index.html (дата звернення: 21.07.2025).

10. How is the Score calculated? Frequently Asked Questions. Office of Foreign Assets Control. 2021. URL: https://ofac.treasury.gov/faqs/249 (дата звернення: 21.07.2025).

11. How does Sanctions List Search work? Frequently Asked Questions. Office of Foreign Assets Control. 2021. URL: https://ofac.treasury.gov/faqs/246.

12. Kim S., Yang S. Accuracy improvement in financial sanction screening: is natural language processing the solution? Frontiers in Artificial Intelligence. 2024. T. 7: 1374323. URL: https://doi.org/10.3389/frai.2024.1374323

REFERENCES:

1. Vu, N.T.H., Wisniewski, T.P., & Skovoroda, R. (2025). Textual Analysis in Financial Research and Practice: A Literature Review. Available at SSRN 5393334.

2. Oztas, B., Cetinkaya, D., Adedoyin, F., Budka, M., Aksu, G., & Dogan, H. (2024). Transaction monitoring in anti-money laundering: A qualitative analysis and points of view from industry. Future Generation Computer Systems, 159, 161–171.

3. Mallela, I.R., Vadlamani, S., Kumar, A., Goel, O., Gopalakrishna, P.K., & Agarwal, R. (2023). Deep Learning Techniques for OFAC Sanction Screening Models. International Journal of Computer Science and Engineering (IJCSE), 12(2), 89–114.

4. Masoudi, A. (2024). AggroDoc-Semantic Search using Vector Databases and Large Language Models. Luleå University of Technology.

5. Kabasta, S. (2025). Multi-language semantic search model for free-text insurance clauses. Masaryk University.

6. Google AI for Developers. (2025). Embeddings. Gemini API docs. Retrieved July 21, 2025, from https://ai.google.dev/gemini-api/docs/embeddings

7. Google AI for Developers. (2025). Embeddings. API Reference. Retrieved July 21, 2025, from https://ai.google.dev/api/embeddings

8. Google Cloud. (2024). Enhancing your gen AI use case with Vertex AI embeddings and task types. Retrieved July 21, 2025, from https://cloud.google.com/blog/products/ai-machine-learning/improve-gen-ai-search-with-vertex-ai-embeddings-and-task-types

9. Office of Foreign Assets Control. (2025). OFAC Sanctions List Service. Retrieved July 21, 2025, from https://sanctionslist.ofac.treas.gov/Home/index.html

10. Office of Foreign Assets Control. (2021). How is the Score calculated? Frequently Asked Questions. Retrieved July 21, 2025, from https://ofac.treasury.gov/faqs/249

11. Office of Foreign Assets Control. (2021). How does Sanctions List Search work? Frequently Asked Questions. Retrieved July 21, 2025, from https://ofac.treasury.gov/faqs/246

12. Kim, S., & Yang, S. (2024). Accuracy improvement in financial sanction screening: is natural language processing the solution?. Frontiers in Artificial Intelligence, 7, 1374323. https://doi.org/10.3389/frai.2024.1374323

Стаття надійшла до редакції 20.09.2025
Стаття прийнята 10.10.2025
Статтю опубліковано 09.12.2025