

## **OPIS WYZWANIA – #FakeJobHunter!**

Waszym zadaniem będzie stworzenie mechanizmu webscrapingu w celu wyszukania oszukańczych ofert pracy i handlowych ofert mających realnie zatuszować sprzedaż danych osobowych poprzez zakup np. kont bankowych, czy zarejestrowanych kart SIM. Następnym krokiem będzie stworzenie algorytmów w oparciu o modele statystyczne lub modele AI, pozwalających na analizę dostarczonych wyników, które pozwolą określić czy ogłoszenie ma charakter oszukańczy.

### **OCZEKIWANE REZULTATY**

Oczekiwany rezultat jest stworzenie narzędzia wykorzystującego mechanizmy, które wyszukują oferty na dostępnych platformach, jak również pomocy algorytmów dostarczy analitykowi ustrukturyzowany raport z określeniem prawdopodobieństwa charakteru ogłoszenia wraz z zabezpieczeniem materiału źródłowego. Wymagane jest, aby narzędzie miało zaimplementowaną listę, która daje analitykowi możliwość szczegółowego sprawdzenia ogłoszenia. Aplikacja powinna również posiadać moduł umożliwiający implementację szablonu pisma, jak i podpięcia serwera wysyłkowego.

### **JAKIE SĄ KORZYŚCI Z TEGO ROZWIĄZANIA?**

Projekt ten umożliwi zwiększenie efektywności wyłapywania oszukańczych ogłoszeń. Dostarczone algorytmy wspomogą Urząd Komisji Nadzoru Finansowego w proaktywnej ochronie konsumentów, jak również we wsparciu profesjonalnych uczestników rynku w ich procesach związanych z przeciwdziałaniem praniu pieniędzy i finansowaniu terroryzmu oraz „poznaj swojego klienta” (AML / KYC). Dzięki uda się nam wspólnie zwiększyć bezpieczeństwo uczestników rynku finansowego i ich środków.

### **Co dostarczymy:**

1. Przykładowy Dataset do modelu
2. Przygotowany Bag of Words

### **Obszary developerskie:**

W naszym wyzwaniu znajdziecie takie obszary deweloperskie jak:

- frontend,
- backend,
- machine learning,
- wizualizacja danych.

## **Specyfikacja techniczna zadania:**

### **I. APLIKACJA WEBOWA:**

#### **1. Środowisko SDK**

- JavaScript / TypeScript / Dart,
- React / Next / Sevlte / Vue / Flutter,
- Docker (opcjonalne)

#### **2. Wizualizacja danych - DASHBOARD**

- Analiza ogłoszenia (podstawowe informacje: treść, data, link, login, słowa kluczowe itp.),
- Szczegółowy podgląd ogłoszenia (Szczegółowe informacje, które uda się zdobyć np. poprzez Who.is itp. W przypadku portali typu Olx czy Sprzedajemy24 - analiza użytkownika (jeżeli jest to możliwe)),
- Utworzenie białej / czarnej listy, do której będą wpadały ogłoszenia, które mają według modelu odpowiedni procent prawdopodobieństwa. Analityk korzystający z aplikacji powinien móc nadawać wagi ryzyka ogłoszenia (np. od 1 do 5, gdzie 1 jest to niskie ryzyko, a 5 wysokie),
- Dostarczanie zbiorowej analizy statystycznej (np. na jakich stronach jest najwięcej ogłoszeń, jaka jest najbardziej popularna treść, charakterystyka właściciela ogłoszenia (nowe konto, podobieństwo nazw itp.),
- Wykorzystanie filtrów do grupowania ogłoszeń,
- Generowanie raportów CSV / TXT,
- Wbudowany system alertów w aplikacji.

### **II. BACKEND:**

#### **1. Środowisko SDK**

- Python / Node / Golang,

- Metody statystyczne / uczenie nadzorowane / uczenie nienadzorowane / uczenie głębokie,
- Docker (opcjonalnie).

## **2. Funkcjonalność:**

- Scraping stron WWW (olx, sprzedajemy24, make-cash, poszukiwanie ofert w indeksach google itp.) i archiwizacja dowodowa dla ogłoszeń (np zrzut strony),
- Zarządzanie modelem. Stworzenie szeregu funkcji oraz mechanizmów pozwalających na przeprowadzenie analizy ogłoszenia w oparciu o model uczenia maszynowego, sieci neuronowych lub metod statystycznych,
- Stworzenie modułowej architektury, pozwalającej na modyfikacje aplikacji, dodawanie nowych funkcjonalności.
- Obsługa aplikacji frontendowej. Web serwer powinien wysyłać oraz odbierać zdarzenia,
- Obsługa bazy danych. Web serwer powinien wysyłać oraz odbierać zapytania do bazy danych,
- Dostarczanie analizy słów kluczowych - Procentowe występowanie fraz - zwrócenie semantyki,
- Poszukiwanie danych kontaktowych (o ile jest możliwe),
- Konteneryzacja. Wykorzystanie platform konteneryzacji systemów np. Docker / Kubernetes.

## **III. MODEL**

*(\*\*\* Wybrać przynajmniej jedno \*\*\*)*

- Metody statystyczne,
- Uczenie Nadzorowane,
- Uczenie Nienadzorowane,
- Uczenie głębokie.