Kolekcjoner fejków

FOLLOW THE FAKE

Opis wyzwania HackNation

Krótki opis wyzwania

Zbuduj narzędzie, które z chaosu pojedynczych podejrzanych reklam wydobędzie wspólne ślady i połączy je w większe kampanie oszustw. Rozwiązanie ma automatycznie analizować treść, warstwę wizualną i publicznie dostępne metadane infrastruktury, a następnie pokazywać podobieństwa i zależności między incydentami w czytelnej formie.

1. Wprowadzenie - opis organizacji, sytuacji i stanu aktualnego

Fałszywe reklamy wykorzystujące wizerunki znanych osób są publikowane masowo, często prowadzą do spreparowanych stron internetowych i funkcjonują na krótko żyjących domenach. Pojedyncze zgłoszenie takiej reklamy ma ograniczoną wartość analityczną: istotne informacje ujawniają się dopiero po zestawieniu wielu przypadków i porównaniu ich cech, narracji, materiałów wizualnych oraz śladów infrastrukturalnych.

Obecnie brakuje ustrukturyzowanego sposobu, który pozwoliłby automatycznie wyciągać z takich materiałów cechy, porządkować je w jednolitym formacie, usuwać duplikaty i łączyć pozornie niezależne reklamy w spójne kampanie. Reklamy mogą wykorzystywać podobną narrację, te same osoby, zbliżony szablon produkcyjny, wspólne domeny, mechanizmy przekierowań lub inne drobne sygnały, które dopiero w połączeniu wskazują na wspólne źródło albo sposób działania.

Wyzwaniem jest więc nie tylko wykrycie, że pojedynczy materiał jest podejrzany, lecz przede wszystkim zbudowanie metody analitycznej pozwalającej odkrywać relacje między wieloma materiałami. Taka wiedza może pomóc w szybszym rozpoznawaniu kolejnych odsłon tych samych kampanii, charakteryzowaniu ich wzorców oraz budowaniu lepszych mechanizmów ich wykrywania.

Wyzwanie zostało zgłoszone przez Ośrodek Badań nad Bezpieczeństwem Sztucznej Inteligencji, działający w strukturach NASK PIB. NASK (Naukowa i Akademicka Sieć Komputerowa) to wiodąca polska instytucja technologiczna, odpowiedzialna za cyberbezpieczeństwo kraju, rozwój nowoczesnych rozwiązań teleinformatycznych oraz prowadzenie rejestru domeny narodowej .pl.

2. Wyzwanie

Zaproponuj narzędzie, które na podstawie materiałów pochodzących z podejrzanych reklam internetowych i powiązanych z nimi stron automatycznie pozyskuje informacje, identyfikuje wspólne cechy oraz wykrywa zależności między incydentami.

Głównym celem wyzwania jest opracowanie metody korelacji i grupowania materiałów na podstawie różnorodnych śladów, m.in. tekstowych, wizualnych, semantycznych i infrastrukturalnych. Rozwiązanie powinno umożliwiać identyfikowanie duplikatów, rozpoznawanie wariantów tej samej kampanii oraz odkrywanie podobieństw między odrębnymi kampaniami. Jednym z kluczowych elementów zadania jest również analiza nowych materiałów i określanie, czy można je powiązać z wcześniej zidentyfikowaną kampanią, czy też reprezentują nowy wzorzec działania.

Nie oczekujemy stworzenia idealnego detektora pojedynczej fałszywej reklamy. Najważniejsze jest opracowanie skutecznej, możliwej do wyjaśnienia metody łączenia niezależnych materiałów w większe struktury, z uwzględnieniem dostępnych danych i wiarygodności wykrytych powiązań.

3. Oczekiwany rezultat

Oczekiwanym rezultatem jest działający demonstrator wraz z opisaną metodyką analizy i grupowania materiałów. Rozwiązanie powinno w szczególności:

- przyjmować materiały do analizy, np. adres URL reklamy lub strony, zrzut ekranu, log komunikacji, plik multimedialny, dane pochodzące z udostępnionego zbioru albo inną wskazówkę pozwalającą rozpocząć analizę;
- automatycznie zbierać/pozyskiwać i porządkować dostępne informacje o materiale, np. tekst, słowa kluczowe, tematykę, wykorzystane wizerunki osób, logotypy, cechy wizualne, domeny, przekierowania, linki, strukturę stron oraz publicznie dostępne metadane domen i infrastruktury;
- zamieniać zebrane dane na zestaw cech umożliwiających porównywanie materiałów;
- identyfikować duplikaty i materiały będące wariantami tego samego przekazu;
- zaproponować metodę określania podobieństwa i grupowania materiałów w kampanie lub grupy podobnych kampanii;
- umożliwiać analizę nowego materiału względem wcześniej przetworzonego zbioru oraz wskazać potencjalne nowe kampanie;
- prezentować znalezione zależności i powiązania w czytelnej formie, np. jako klastry, graf zależności lub listę podobnych incydentów;
- wskazywać cechy i przesłanki, które wpłynęły na powiązanie konkretnych materiałów
- umożliwić przeprowadzenie procesu analizy za pomocą prostego interfejsu graficznego;
- udokumentować zastosowaną metodykę, w szczególności sposób pozyskiwania cech, obliczania podobieństwa i grupowania materiałów.

Rozwiązanie powinno uwzględniać możliwość występowania niekompletnych danych, w tym niedostępnych stron internetowych lub brakujących materiałów źródłowych.

Dodatkowo punktowane mogą być funkcjonalności wykraczające poza podstawowy zakres, np. atutem będzie mechanizm automatycznego wyszukiwania podejrzanych reklam w publicznie dostępnych źródłach, takich jak biblioteki reklam, (według słów kluczowych, daty wystąpienia albo innych reguł), okresowe aktualizowanie informacji lub identyfikowanie nowych materiałów powiązanych z wcześniej wykrytymi kampaniami. Wypracowane rozwiązania i metodyki mogą stanowić podstawę dalszego rozwoju systemu służącego do detekcji i analizy oszukańczych treści.

4. Wymagania formalne

Projekt powinien zawierać:

- działający demonstrator z prostym interfejsem graficznym umożliwiającym przeprowadzenie i zaprezentowanie procesu analizy;
- opis zastosowanej metodyki, w tym sposobu pozyskiwania cech, określania podobieństwa i grupowania materiałów;
- instrukcję instalacji, konfiguracji i uruchomienia rozwiązania wraz z informacją o wymaganych zależnościach;
- kod źródłowy udostępniony w repozytorium Git (np. GitHub, GitLab) lub w innej formie umożliwiającej organizatorowi pobranie i weryfikację rozwiązania, np. Google Drive
- pomocnym będzie skonteneryzowane środowisko ułatwiające przenoszenie i uruchomienie projektu.

Nie narzuca się konkretnych technologii, języków programowania ani bibliotek. W przypadku wykorzystania zewnętrznych usług, modeli lub API wymagających autoryzacji należy opisać sposób uzyskania dostępu oraz zapewnić możliwość weryfikacji działania rozwiązania podczas oceny. Nie wymaga się przekazywania prywatnych kluczy API ani certyfikatów. Uczestnicy powinni wskazać wykorzystane biblioteki, modele i zewnętrzne usługi wraz z informacją o ich licencjach oraz ewentualnych ograniczeniach wykorzystania

5. Wymagania techniczne

Technologia implementacji jest dowolna. Demonstrator może być przygotowany np. w Streamlit, Gradio lub innym prostym środowisku webowym. Estetyka interfejsu nie jest przedmiotem oceny - interfejs ma przede wszystkim umożliwiać uruchomienie i zaprezentowanie pełnego procesu analizy.

Zakres analizy oraz sposób określania podobieństwa i grupowania pozostają otwarte. Zespół może wykorzystać metody regułowe, podobieństwo ważone, modele uczenia maszynowego, embeddingi, silniki klastrujące, grafy zależności lub inne podejścia. Sposób reprezentacji i wizualizacji relacji również jest dowolny, jednym z możliwych kierunków jest grafowa baza danych, np. Neo4j.

Rozwiązanie powinno umożliwiać porównywanie nowych materiałów z wcześniej przeanalizowanymi przypadkami, uwzględniając możliwość występowania niekompletnych danych, np. brakujących metadanych, niedostępnych stron internetowych lub części materiałów źródłowych.

W przypadku identyfikacji osób widocznych w materiałach dopuszczalne jest przygotowanie niewielkiej bazy referencyjnej obejmującej najczęściej występujące osoby. Jej rozmiar nie jest kluczowy - ma jedynie pomóc wykazać użyteczność proponowanego rozwiązania. Dobór metod, modeli i analizowanych cech pozostaje w gestii zespołu. Istotna jest ich przydatność do identyfikowania powiązań między materiałami oraz możliwość uzasadnienia uzyskanych wyników. Rozwiązanie powinno być możliwie łatwe do przenoszenia i uruchomienia - preferowane jest przygotowanie wersji skonteneryzowanej wraz z instrukcją.

6. Sposób testowania i/lub walidacji

Rozwiązania będą oceniane na podstawie demonstracji działania oraz testów przeprowadzonych z wykorzystaniem materiałów przygotowanych przez organizatora.

Uczestnikom zostanie udostępniony przykładowy zbiór danych umożliwiający zapoznanie się z charakterystyką incydentów, przygotowanie i rozwijanie rozwiązania oraz przeprowadzenie wstępnych testów. Podczas oceny organizator wykorzysta również dodatkowy, wcześniej nieudostępniony zbiór testowy zawierający nowe warianty znanych kampanii, jak i materiały reprezentujące inne kampanie.

Walidacja będzie obejmować w szczególności:

- skuteczność identyfikowania podobnych materiałów i wariantów tych samych kampanii
- poprawność przypisywania nowych materiałów do istniejących grup
- zdolność identyfikowania materiałów, które nie powinny zostać połączone z wcześniej rozpoznanymi kampaniami
- przydatność i różnorodność wykorzystywanych cech
- odporność na duplikaty, niekompletne dane i różnice w formacie materiałów
- interpretowalność wyników, w tym możliwość wskazania przesłanek wykrytych powiązań
- kompletność działania rozwiązania, od wczytania danych do prezentacji wyników

Podczas demonstracji zespół powinien zaprezentować pełny proces analizy, w tym przetwarzanie materiału wejściowego, wyodrębnienie cech, porównanie z innymi przypadkami, grupowanie oraz prezentację uzyskanych wyników.

Ocena jakości korelacji i grupowania będzie uwzględniała wyniki testów oraz uzasadnienie przyjętej metodyki. Organizator nie narzuca konkretnego algorytmu ani sposobu reprezentacji danych.

Istotna będzie nie tylko końcowa jakość grupowania, lecz również sensowność zastosowanej metodyki: dobór źródeł danych i cech, sposób obliczania podobieństwa, odporność na duplikaty oraz możliwość wyjaśnienia, dlaczego konkretne materiały zostały ze sobą połączone.

7. Dostępne zasoby

- Przykładowe reklamy typu scam
- inne materiały udostępnione na kanale Discord

8. Zasady korzystania z danych

Uczestnicy otrzymają zbiór przykładowych materiałów służących do opracowania i weryfikacji rozwiązania. Organizator zastrzega sobie możliwość przeprowadzenia dodatkowych testów na materiałach nieuwzględnionych w udostępnionym zbiorze.

Dopuszczalne jest wykorzystanie dodatkowych, pozyskanych danych i publicznie dostępnych źródeł informacji, pod warunkiem wskazania ich pochodzenia.

Materiały udostępnione przez organizatora mogą być wykorzystywane wyłącznie na potrzeby realizacji wyzwania. Niedozwolone jest ich publiczne rozpowszechnianie.

9. Kryteria oceny

- Jakość metody korelacji i grupowania materiałów - 30%
- Użyteczność i trafność pozyskiwanych cech oraz źródeł danych - 30%
- Kompletność demonstratora i działanie procesu end-to-end - 15%
- Czytelność i interpretowalność prezentowanych powiązań - 15%
- Potencjał wdrożeniowy, skalowalność i przenośność rozwiązania - 10%

10. Dodatkowe uwagi / kontekst wdrożeniowy

Z perspektywy organizatora ważnym rezultatem wyzwania, poza działającym demonstratorem, jest wiedza uzyskana podczas realizacji zadania, w szczególności: wzbogacony i uporządkowany zbiór danych, opis zidentyfikowanych wzorców kampanii oszukańczych, metodyka łączenia cech oraz wskazanie użytecznych źródeł metadanych. Wypracowane rozwiązania i metodyki mogą bezpośrednio wspierać dalszy rozwój narzędzi Instytutu, w kierunku identyfikowania i analizy całych kampanii oszustw internetowych. Szczególnie istotna jest możliwość wykorzystania proponowanych metod do rozpoznawania nowych wariantów znanych kampanii oraz odkrywania powiązań między pozornie niezależnymi incydentami. Wiedza zdobyta dzięki temu wyzwaniu zostanie wykorzystana w dalszym rozwoju narzędzi do skutecznej ochrony w sieci.

10. Kontakt

Podczas wydarzenia dostępni będą mentorzy zarówno techniczni jak i merytoryczni w specjalnie oznaczonym punkcie konsultacyjnym. Pytania dotyczące wyzwania można również kierować za pośrednictwem kanału na Discord.

THE KOLEKCJONER FEJKÓW

FOLLOW THE FAKE

“Matrix jest wszędzie, otacza nas ze wszystkich stron. Nawet tu i teraz. Widzisz go wyglądając przez okno i włączając telewizor. Czujesz, gdy idziesz do pracy, czy do kościoła, gdy płacisz podatki. To świat, który postawiono ci przed oczami, by przesłonić prawdę.”

- Morfeusz, Matrix

Społeczeństwo przyzwyczajone jest do tego, że to, co zobaczą na ekranie jest prawdziwe. To jest nasz Matrix - nie wiemy, gdzie jest granica między prawdą, a kłamstwem. Często bywamy rozbrajani z ostrożności przez ładne, spreparowane reklamy, kolorowe filmiki, albo obietnice potwierdzone przez dziesiątki osób. Gdy jednak zanurzymy się w świat kłamstwa, możemy znaleźć pewne zależności, a może nawet dotrzeć do ich źródła.

Twoim przeznaczeniem jest wzięcie udziału w wojnie z kłamstwem, którym cały czas jesteśmy bombardowani. Wspólnie musimy wejść w posiadanie narzędzi, które w jasny sposób oddzielą prawdę od oszustwa w świecie po drugiej stronie ekranu. Każdy powinien czuć się bezpiecznie we wspólnej przestrzeni - nie tylko tej fizycznej, ale i wirtualnej. A przecież żyjemy w świecie, w którym świat wirtualny wpływa na prawdziwy. Fałszywe wypowiedzi polityków, nieprawdziwe wypowiedzi lekarzy, narracje budujące nienawiść do różnych grup społecznych przekładają się na prawdziwe zachowania. Zadaniem jest znaleźć metodę, która z chaosu pojedynczych reklam wyciągnie strukturę. Rozpoznajcie cechy wspólne, określcie podobieństwo i połączcie materiały należące do tych samych lub podobnych kampanii.

Na fałszywe materiały najczęściej możemy natknąć się w mediach społecznościowych oraz portalach o charakterze rozrywkowym. Musimy posiąść wiedzę, jak dostać się do bibliotek takich reklam, aby przeprowadzić ich pełną analizę. Pomoże nam to w pozyskaniu informacji o źródle fałszu i jego wyeliminowaniu.
Reklamy zwykle tworzone są na domenach o krótkim czasie życia, ale mogą współdzielić między sobą inne metadane. Po pozyskaniu materiału możemy sprawdzić informacje o domenie, ewentualne przekierowania, odnośniki, motyw przewodni, słowa kluczowe, wykorzystane wizerunki osób.

Dysponując m.in takimi informacjami możemy stworzyć system grupowania kampani reklamowych. Kluczowa jest wiedza o tym, jak skutecznie łączyć reklamy o podobnych cechach: wizualnych, semantycznych, konceptualnych i wielu innych.Klaster może powstać z różnych powodów. Reklamy mogą wykorzystywać podobną narrację (tekst), te same osoby, podobny szablon produkcji, wspólną domenę albo zestaw innych pozornie drobnych sygnałów, które pojedynczo wydają się nieistotne, lecz razem wskazują na jedno źródło. Taka wiedza pozwala identyfikować wzorce wspólne dla różnych kampanii i wykorzystywać je do szybszego wykrywania nowych odsłon tych samych fałszywych narracji. Gdy kolejne warianty można powiązać ze wspólnym źródłem i mechanizmem działania, iluzja informacyjnego Matrixa zaczyna się rozpadać.

“Umiesz tym latać?”
- Neo, Matrix (fragment dialogu)

Przejdźmy do kwestii technicznych. Bo po to tu jesteśmy.

>_ Rozwiązanie powinno zawierać prosty demonstrator (prosta stronka, np. z wykorzystaniem Streamlit, Gradio, lub innych; jej estetyka nie podlega ocenie. Jest oceniana 0/1 - jest, albo jej nie ma. Chodzi tylko o to, aby mieć graficzne środowisko uruchomieniowe). Strona może przyjąć bezpośredni adres URL, zrzut ekranu zawierający link, lub inną wskazówkę jak nazwa produktu, albo plik zawierający podejrzaną reklamę, aby uruchomić proces analizy. Dodatkowym, pozytywnym akcentem może być zaszyty scheduler, który samodzielnie wykonuje skan zidentyfikowanych bibliotek reklam, znajduje te podejrzane i wykonuje ich analizy.

>_ Zakres analizy jest dowolny. Ocenie podlegają mechanizmy i ich przydatność w kontekście zadania. Przykład: słowa kluczowe z tekstu, najważniejsze elementy wizualne, wyizolowane strony internetowe, metadane tych stron.

>_ Sposób identyfikacji podobieństwa jest dowolny. Może to być podobieństwo ważone, może drobny model, a może jakiś silnik klastrujący. Zdaj się na swoją wiedzę. Skup się. Jest to jeden z kluczowych elementów zadania.

>_ Jak wyciągnąć tożsamość osoby widocznej na materiale? Dowolnie. Sugeruję zrobić wstępną bazę referencyjną kilku, najczęściej występujących osób. Obszerność tej bazy jest nieistotna. Ma tylko przyczynić się do udowodnienia użyteczności narzędzia.

>_ Reprezentacja powiązań? Zdaj się na swoją wiedzę. Może Neo4j?

Nie chodzi o stworzenie idealnego detektora pojedynczej fałszywej reklamy. Zbudujcie sposób na okrywanie, jak skutecznie łączyć wiele pozornie niezależnych materiałów w większe kampanie na podstawie pozostawianych przez nie wspólnych śladów.

PS. Widział_ś stronę z zajawką? >_ Zeskanuj mnie
https://science.nask.pl/hacknation