#AdsDetect – mechanizm wykrywania fałszywych reklam

KONTEKST ROZWIĄZANIA

Reklamy są coraz częstszym nośnikiem wszelkiego rodzaju oszustw i cyberprzestępstw. Zorganizowane grupy wykorzystują je do kierowania użytkowników na fałszywe strony celem wyłudzenia danych. Często wchodząc na tego typu reklamę nie zdajemy sobie sprawy, że możemy trafić na fałszywą stroną bankowości elektronicznej, która przechwyci nasze hasło. Po wprowadzeniu tam danych są one przesyłane bezpośrednio do cyberprzestępców. Odwiedzenie takiej reklamy w Internecie potrafi skończyć się dla użytkownika utratą oszczędności całego życia. Bardzo popularną praktyką jest także tworzenie reklam fałszywych okazji inwestycyjnych. Chcąc przyciągnąć uwagę użytkowników, oszuści oferują zazwyczaj bardzo wysokie zyski bez ryzyka utraty pieniędzy. Reklamy często wykorzystują wizerunki znanych organizacji lub też rozpoznawalnych osób z pierwszych stron gazet. Z tego typu reklamami użytkownicy mogą spotkać się w mediach społecznościowych, różnych portalach informacyjnych czy też wyszukiwarkach internetowych. Zespół CSIRT KNF stara się zwalczać te zagrożenia, samodzielnie wyszukując reklamy i zgłaszając je właścicielowi danego medium.

W ramach działań zespołu CSIRT KNF w 2022 roku wykryliśmy i zgłosiliśmy 17 899 fałszywych reklam w Internecie. To jednak nie rozwiązuje problemu, cyberprzestępcy bowiem cały czas wymyślają nowe sposoby na wykorzystywanie reklam do dystrybucji linków prowadzących do stron phishingowych.

Chcielibyśmy zautomatyzować i ulepszyć proces przeszukiwania, analizowania i wykrywania potencjalnie niebezpiecznych reklam w Internecie. Pomóżcie nam przeszukać cyberprzestrzeń w celu wykrycia fałszywych i niebezpiecznych reklam.

OPIS I CHARAKTERYSTYKA ZADANIA

Waszym zadaniem jest stworzenie sytemu, który umożliwi wyszukiwanie oraz wstępną analizę reklam w serwisach internetowych na podstawie wskazanych słów kluczowych. Jeżeli uda się Wam uwzględnić w rozwiązaniu kontekst wyświetlania reklam dla użytkownika – będzie to dodatkowo punktowane.

OCZEKIWANE REZULTATY

System powinien przeszukiwać wskazane serwisy pod kątem obecności reklam oraz listować jest w ustrukturyzowanej formie, uwzględniając dane odczytane za pomocą OCR. System powinien uwzględniać popularne serwisy informacyjne, wyszukiwarki internetowe oraz opcjonalnie media społecznościowe, które zostały określone poniżej. W kolejnych etapach system powinien umożliwić przeszukiwanie wyników pod kątem określonych słów kluczowych. Duża część z reklam obecnych w Internecie to reklamy kontekstowe. Dodatkowo punktowane będą także rozwiązania, które uwzględnią określony kontekst użytkownika w trakcie wyszukiwania reklam.

JAKIE SĄ KORZYŚCI Z TEGO ROZWIĄZANIA?

Dostarczone rozwiązanie pozwoli na usprawnienie naszych mechanizmów wykrywania i analizowanie reklam o charakterze oszukańczym. Dzięki temu możliwe będzie szybsze ich zgłoszenie oraz zablokowanie, tak aby w maksymalny sposób zmniejszyć liczbę potencjalnie oszukanych użytkowników.

CO OD NAS DOSTANIECIE?

- Listę serwisów, które powinno uwzględniać rozwiązanie.
- Przykłady reklam oszukańczych wykorzystywanych przez cyberprzestępców.
- Słowa kluczowe, które mogą być uwzględnione w trakcie budowania kontekstu.

CO MUSICIE DOSTARCZYĆ?

- Kod działającego rozwiązania opublikowany na otwartym githubie zawierający pliki readme.txt, install.txt, requirments.txt.
- Prezentację przedstawiającą Wasze rozwiązanie – maksimum 10 slajdów w formacie PDF.
- Dokumentację rozwiązania, która powinna zawierać w szczególności:
    - listę serwisów, które uwzględnia Wasze rozwiązanie (nie musicie zrobić wszystkich ale im więcej zrobicie, tym większa będzie wasza szansa na zwycięstwo),
    - dokładny opis w jaki sposób symulujecie kontekst użytkownika, jeżeli Wasze rozwiązanie zapewnia taką funkcjonalność,`
    - dokładny opis monitorowania reklam w mediach społecznościowych.
- Wyniki działania Waszego rozwiązania w postaci listy zidentyfikowanych reklam na wskazanych serwisach internetowych.
- Formularz www z działającym rozwianiem do testów, który udostępnicie w sieci lokalnej. Rozwiązanie powinno przyjmować zapytania i zwracać odpowiedzi zgodnie ze schematem opisanym w specyfikacji zadania.

Całość wyników wrzućcie na platformę Challenge Rocket, ułatwi nam to ocenę Waszych rozwiązań.

SPECYFIKACJA TECHNICZNA ZADANIA

Celem zadania jest stworzenie API, które będzie przyjmowało na wejściu adres URL serwisu. Mechanizm powinien wykryć wszystkie reklamy dostępne w badanym serwisie. Następnie zwracaną jest ustrukturyzowana lista reklam, z uwzględnieniem OCR tekstu obrazu z reklamy. Powinna istnieć możliwość zawężenia wyników na podstawie słów kluczowych wskazanych w parametrze query.

Przykład danych wejściowych:

```
Parametry wejściowe (form data lub json)

{
        "url": "https://gazeta.pl",
        "query": "Inwestycje w złoto",
        "user-agent": "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:47.0) Gecko/20100101 Firefox/47.0"
        "context": "Ewentualny kontekst przeglądarki"
}
```

Rozwiązanie powinno umożliwiać także wyszukiwanie reklam wyświetlanych w wyszukiwarkach internetowych takich jak Google i Bing. W tym przypadku API powinno przyjmować dodatkowy parametr „search”, którego wartość zostanie wykorzystana jako fraza w wyszukiwarce.

Przykład danych wejściowych dla wyszukiwarki:

```
{
        "url": "https://www.google.com/search?q=baltic+pipe",
        "search": "baltic pipe",
        "query": "Dochód pasywny",
```

```
        "user-agent": "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:47.0) Gecko/20100101 Firefox/47.0"

        "context": "Ewentualny kontekst przeglądarki"

}
```

Opisy pól wejściowych:

url - adres serwisu, w którym szukamy reklam.

search – parametr przy wyszukiwaniu reklam w wyszukiwarkach internetowych. Przekazywany jako fraza do wyszukania (np. dla googla `"search": "baltic pipe"` oznacza w praktyce `https://www.google.com/search?q=baltic+pipe)/`

query - pole opcjonalne, na podstawie którego filtrowane są otrzymywane wyniki. Zwracane wyniki powinny być ograniczanie do tych zawierających w treści wartości przekazane w parametrze.

user-agent – pole, w którym możemy zdefiniować z jakim parametrem user-agent przedstawi się nasz skrypt przy weryfikacji linku destination_url. Strony wykorzystywane przez cyberoszustów, często weryfikują wartość tego parametru.

context - pole opcjonalne, w którym przekazujemy context (np. ciasteczka) profilu.

W ramach odpowiedzi rozwiązania powinno zwracać tablice z listą reklam spełniających kryteria wyszukiwania. Dodatkowo dobrze gdyby rozwiązanie wykonywało zrzut ekranu ze zidentyfikowaną reklamą.

Odpowiedź powinna mieć następującą strukturę

```
{
        "url": "https://gazeta.pl",

        "user-agent": "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:47.0) Gecko/20100101 Firefox/47.0"

        "context": "Kontekst przeglądarki"


        "ads": [

                "name": "Unique name",

                "destination_url":            ["https://tracking_url_1.pl",            "https://tracking_url_2.pl",
"https://firmaxyz.pl/produkt"],

                "words": ["Firma Inwestcyjna", "Produkt"],

                "screenshot_ads": http://localhost/img/screenshots/ads/utuy8767687y8778.jpg,

        ]

}
```

Opisy pól wyjściowych:

url - adres serwisu, w którym szukamy reklam.

context - pole opcjonalne, w którym przekazujemy context (ciasteczka) profilu.

ads – tablica w której zwracane są wszystkie wyniki zawierające reklamy.

name - nazwa reklamy – unikalna nazwa np. na podstawie tytułu reklamy lub jego treści.

destination_url - docelowy url reklamy, przy czym jeśli link prowadzi przez kilka serwisów nim dotrze do celu, to zbieramy wszystkie przejścia (0 - docelowy adres).

words - słowa występujące w reklamie – zarówno w tekście jak i odczytane z wykorzystaniem OCR

Dodatkowe informacje.

screenshot_ads – zrzut ekranu wykrytej reklamy, która spełnia kryteria wyszukiwania.

- Zapytanie w celu weryfikacji destination_url powinno być wysyłane z ustawionym polem „Referer", na którym znajduje się reklama (np. google.pl) - jest to istotne ponieważ wiele stron phishingowych weryfikuje w ten sposób czy komunikacja pochodzi z właściwego serwisu.
- W przypadku braku wskazania parametru query rozwiązanie powinno zwrócić w wynikach wszystkie reklamy, które udało się zidentyfikować w badanym serwisie.
- W projekcie nie powinniśmy korzystać z zewnętrznych API, które mogą być odpłatne.

W przypadku wątpliwości – pytajcie na Discordzie lub bezpośrednio mentorów w trakcie wydarzenia.

MEDIA SPOŁECZNOŚCIOWE – DODATKOWE PUNKTY

Fałszywe reklamy w mediach społecznościowych to obecnie istna plaga. Jeżeli Wasze rozwiązanie będzie umożliwiało wyszukiwanie reklam także w najpopularniejszych serwisach społecznościowych, otrzymacie za to dodatkowe punkty. Pamiętajcie aby dobrze opisać to w dokumentacji rozwiązania i pomysły (najciekawsze pomysły i koncepcje będą również premiowane dodatkowymi punktami).

Przykłady serwisów społecznościowych, w których występują oszukańcze reklamy:

- facebook.com
- youtube.com
- instagram.com
- linkedin.com

KONTEKST UŻYTKOWNIKA – DODATKOWE PUNKTY

Wiele złośliwych reklam wyświetlanych jest z wykorzystaniem „kontekstu użytkownika". Monitorowanie złośliwych reklam jest z tego powodu bardzo utrudnione. Jeżeli uda Wam się opracować rozwiązanie, które zaadresuje te problemy i pozwoli na uwzględnienie kontekstu będzie to dodatkowo punktowane. Nawet jeżeli nie uda Wam się w pełni opracować gotowego rozwiązania, możecie dostać punkty za opisanie pomysłów i koncepcji na rozwiązanie w dokumentacji. Opiszcie to najdokładniej jak się da, możecie dostać za to dodatkowe punkty.

Słowa kluczowe do kontekstu:

dochód pasywny, zarabianie, wysokie zyski, inwestycje, kredyt, akcje, inwestowanie w akcje, zarabianie w internecie, oszczędzanie, konta inwestycyjne, wzbogacenie się, inwestycje kryptowalutowe, szybki zarobek.

SERWISY, KTÓRE POWINNY BYĆ UWZGLĘDNIONE PRZEZ WASZE ROZWIĄZANIE

Rozwiązanie powinno uwzględniać poniższe serwisy informacyjne:

- msn.com
- wykop.pl
- onet.pl
- wp.pl
- interia.pl

- gazeta.pl
- wyborcza.pl
- fakt.pl
- money.pl
- tvn24.pl
- tvp.pl
- polsatnews.pl
- bankier.pl
- businessinsider.com.pl
- fakt.pl
- forsal.pl
- rp.pl
- dziennik.pl
- tvn24.pl
- wprost.pl
- newsweek.pl
- ceneo.pl
- biznes.pl
- pb.pl
- parkiet.com
- newseria.pl
- strefabiznesu.pl
- gazetaprawna.pl
- tokfm.pl
- radiozet.pl
- rmf24.pl
- se.pl
- superexpress.pl
- naszdziennik.pl
- wpolityce.pl

Wyszukiwarki:

- google.pl
- bing.com

Media społecznościowe (opcjonalnie):

- facebook.com
- youtube.com
- instagram.com
- linkedin.com

Jeżeli rozwiązanie obsługuje więcej serwisów, lub jest na tyle uniwersalne, że potrafi wyszukiwać reklam w dowolnym serwisie – opiszcie to w dokumentacji – otrzymacie za to dodatkowe punkty.

DODATKOWE POMYSŁY = DODATKOWE PUNKTY

Dostaniecie od nas przykłady, jak w praktyce wyglądają fałszywe reklamy w serwisach internetowych. Jeżeli wpadniecie na dodatkowe pomysły w jaki sposób można je wykrywać i analizować, opiszcie to koniecznie i spróbujcie zaimplementować. Jeżeli okażą się ciekawe otrzymacie od nas dodatkowe punkty (również za opisy, jeżeli zabraknie czasu na implementację).