netidee

PROJEKTE

Konzept | Crawler Entwicklung

Fraud Seeker | Projekt ID 6326

Inhalt

1 Crawler-Entwicklung 3

1.1 Inhaltliche Grundlagen 3
1.2 Suchmaschinen-Crawler: Sucharchitektur & Suchphrasen 3
1.3 meta-Werbebibliothek: Sucharchitektur und Suchphrasen 7

2 Technische Grundlagen 9

2.1 Anforderungsprofil 9
2.2 Dashboard 10
2.3 Anbindung an die Datenbank des Fake-Shop Detector 10
2.4 Qualitätssicherung 11

1 Crawler-Entwicklung

1.1 Inhaltliche Grundlagen

Das Team der Watchlist Internet beobachtet in den vergangenen Jahren eine starke Zunahme von betrügerischen Tradingplattformen, die einen enormen finanziellen Schaden verursachen. Auch Fake-Shops für Medikamente und Nahrungsergänzungsmittel sind im Internet stark verbreitet. Hier kommen neben den finanziellen Schäden mögliche Gesundheitsschäden hinzu.

Beiden Betrugsformen ist gemein, dass sowohl Fake-Shops als auch Tradingplattformen mit den immer gleichen Textbausteinen arbeiten. Textbausteine, die auf den Websites selbst oder in Anzeigen für die Websites (bspw. auf Facebook und Instagram) zu finden sind. Für die Betrugsdetektion bedeutet dies, dass durch die Suche nach solchen „Copy+Paste-Textbausteinen“ eine Vielzahl an betrügerischen Websites ausfindig gemacht werden können. Das gleiche gilt auch für die Suche in der Meta-Werbebibliothek. Eine Automatisierung dieser Suchen mittels Crawler ermöglicht eine schnelle Erkennung von sowie Warnung vor neuen betrügerischen Websites.

1.2 Suchmaschinen-Crawler: Sucharchitektur & Suchphrasen

„Tesler is an automated trading software created for everyone.”

Das ist ein Beispiel einer Textphrase, die Cyberkriminelle für eine Vielzahl von betrügerischen Tradingplattformen verwenden. Um auf die jeweiligen betrügerischen Websites zu kommen, reicht die Suche nach dieser Textphrase (in Anführungszeichen stehend, damit die Phrase 1:1 gesucht wird) in einer Suchmaschine wie Google:

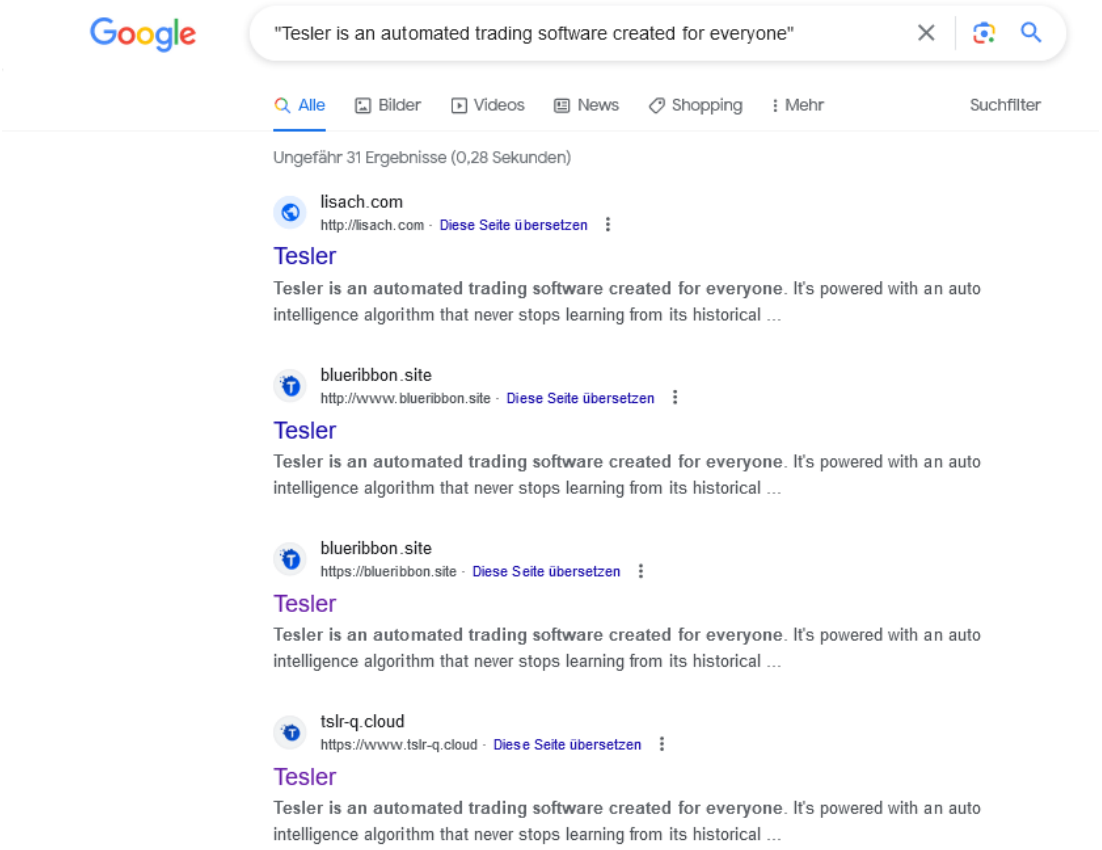

Abb. 1. Am 17.05.2021 wurden mit dem Textbaustein „Tesler is an automated trading software created for everyone." 31 Ergebnisse gefunden. Die Erfahrung zeigt, dass auch zukünftig immer wieder neue Websites mit diesem Textbaustein von Betrüger:innen registriert werden.

Automatisiert werden kann diese Suche mit einem Crawler, der die Suchmaschine regelmäßig auf neu indexierte Domains zu dieser Textphrase überprüft. Zusätzlich muss eine Whitelist mit Domains aufgebaut werden. Damit wird sichergestellt, dass seriöse Suchergebnisse (bspw. Seiten, die vor Betrug warnen und in der Warnung den Textbaustein erwähnen) vom Crawler automatisiert herausgefiltert werden.
Der Erfolg des Crawlers hängt von den ausgewählten Suchphrasen ab. Der Zugang dazu kann durch Meldungen betrügerischer Websites (sowohl Tradingplattformen als auch Fake-Shops für Nahrungsergänzungsmittel bzw. für Medikamente) an das Team der Watchlist Internet sichergestellt werden. Die dort immer wieder verwendeten Textbausteine werden gesammelt und anschließend in den Crawler integriert.

Zusätzlich können je nach Thema weitere Ressourcen (Blacklists) genutzt werden, um aktuelle betrügerische Seiten, inkl. den jeweiligen Textbausteinen zu finden:

Investmentbetrug

- Investorenwarnungen der FMA
- CryptoscamDB
- Swiss Financial Market Supervisory Authority FINMA Warning List
- RED List CFTC (Registration Deficient List)
- ABE Blacklist
- Cyprus Securities and Exchange Commission ('CySEC') List of non-approved domains

Gefälschte Medikamente

- AGES/BASG-Warnungen
- Gute Pillen – Schlechte Pillen
- Not Recommended List (safe.pharmacy)
- Internet Pharmacy Warning Letters (U.S. Food & Drug)

Nahrungsergänzungsmittel

- Verbraucherwarnungen von Klartext Nahrungsergänzung
- AGES/BASG-Warnungen
- Gute Pillen – Schlechte Pillen
- RASFF Window (Europäische Kommission)
- RAPEX-Meldungen

Je nach Thema muss dabei der Fokus anders gelegt werden: So finden sich „Copy+Paste-Textbausteine" bei Tradingplattformen oftmals bereits auf der Startseite. Bei Fake-Apotheken können Textbausteine aus den „Über Uns" Texten generiert werden. Ebenfalls erfolgsversprechender ist die Suche nach Dateinamen der verwendeten Medikamentenbilder. In puncto Nahrungsergänzungsmittel wird auf die Suche nach Krankheitsbeschreibungen/Heilungsversprechungen sowie nach Produktnamen gesetzt.

Tradingplattformen	*“Here are some of the added benefits of trading with our cutting-edge trading technology."*
	"Neben Bitcoin ist der Handel mit exotischen Devisenpaaren wie Euro/Türkische Lira, US Dollar/Schwedische Krone, US Dollar/Norwegische Krone“
	"Tesler is an automated trading software created for everyone."
	"bietet normalen Menschen die Möglichkeit, in kurzer Zeit reich zu werden."
GefŶlschte Medikamente	*"Alle Medikamente, die man in unserem Produkt-Listen sehen kann, sind generisch."*
	"Potenzmittel rezeptfrei: Viagra, Cialis, Levitra, Kamagra, Priligy und mehr"
	"Wir sind das professionelle Team, das sich um Sie und um Ihre Familie kümmert, und wollen, dass jeder die günstigsten Gesundheitsprodukte der Welt hat."
	"bietet seinen Kundinnen und Kunden ein einfaches und bequemes Shoppingerlebnis."
	"cialis-super-active"
NahrungsergŶnzungsmittel	*"Ich hatte Schmerzen in meinen Fingern. Ich konnte sie nicht beugen."*
	"LEISTUNGSSTARKE NEUE FORMEL LÖST DIE FETTVERBRENNENDE KETOSE AUS!"
	"stimuliert die Synthese von Hyaluronsäure, wobei es die Bindegewebsstrukturen stärkt"
	Ketoxplode
	Hondrolife

1.3 meta-Werbebibliothek: Sucharchitektur und Suchphrasen

„ist dies das Ende Ihrer Karriere“

Hinter diesem Satz versteckt sich Werbung für eine betrügerische Tradingplattform, die aktuell auf Facebook und Instagram geschaltet wird. Sucht man auf https://www.facebook.com/ads/library/ nach dieser Phrase und schränkt die Suchregion dabei auf „Österreich“ ein, werden aktuell 78 Ergebnisse mit betrügerischen Werbeanzeigen gefunden.

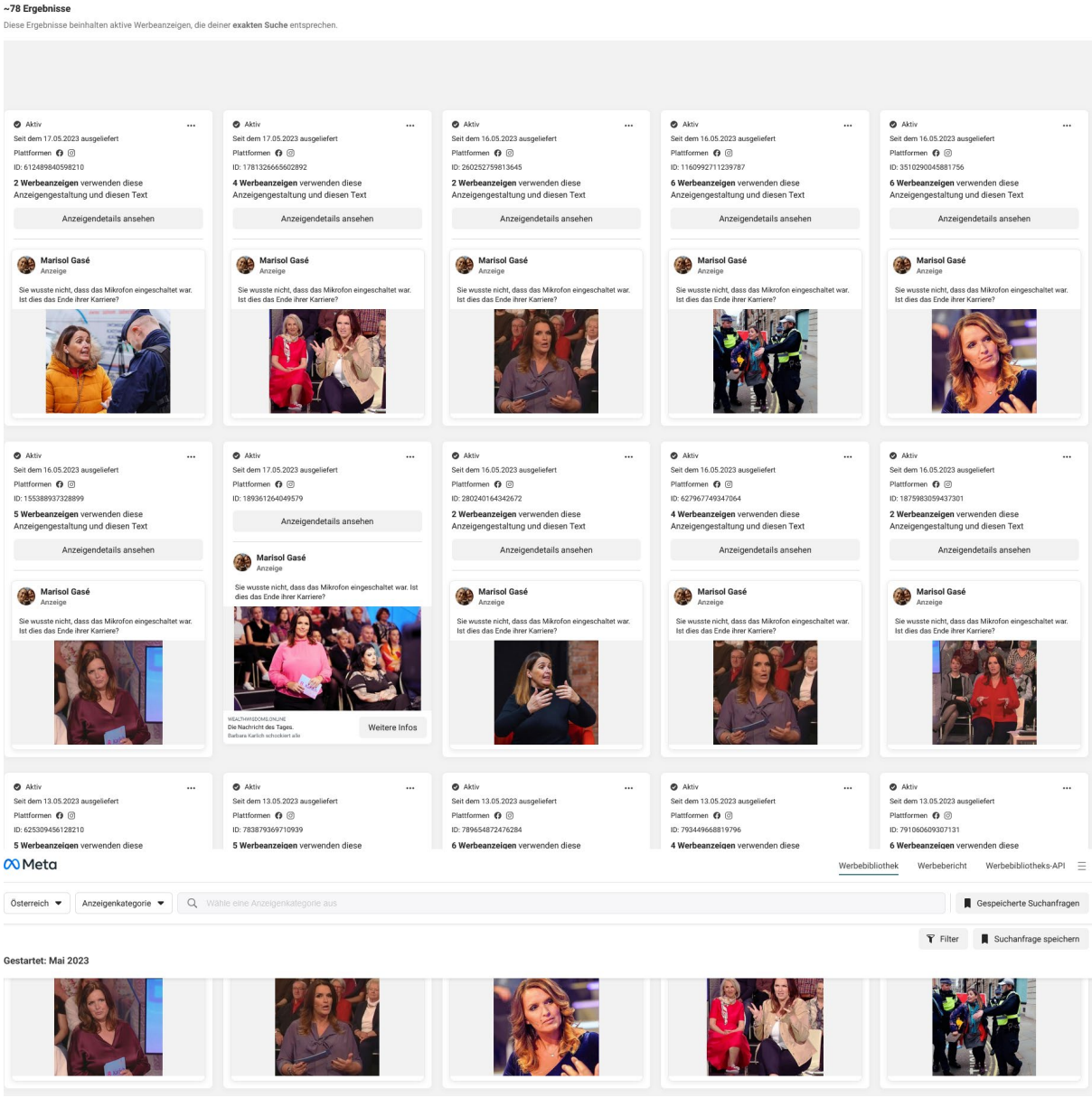

Abb. 2. Am 17.05.2021 wurden mit dem Textbaustein „ist dies das Ende Ihrer Karriere“ 78 Ergebnisse gefunden. ALLE Ergebnisse bilden die österreichische Moderatorin Barbara Karlich ab. Die Links führen zu Fake-News, in denen behauptet wird, dass Barbara Karlich aus Versehen ein Trading-Geheimnis bei eingeschaltetem Mikrofon ausgeplaudert hätte. Tatsächlich verbirgt sich dahinter eine betrügerische Trading-Plattform.

Automatisiert werden kann die Suche durch einen Crawler, der mit Hilfe der Werbebibliothek API regelmäßig nach den Textphrasen sucht. Neben einen Screenshot der Werbeanzeige selbst muss die Domain, die in diesen Anzeigen zu finden ist, als Ergebnis ausgespielt werden.

Die Suche nach Werbeanzeigen auf Facebook und Instagram gestaltet sich schwieriger als die Suchmaschinen-Suche, da die Anzeigen nur selten von Konsument:innen gemeldet werden bzw. erfährt das Team der Watchlist Internet oftmals nur, dass die Betroffenen über Werbung auf Social Media zur Falle gefunden haben. Wie diese Werbung aussieht und welche Texte verwendet werden, ist nur selten nachvollziehbar.

Um auf entsprechende Textbausteine zu kommen, werden daher in einem ersten Schritt bestimmte Keywords genutzt. Die damit angezeigten Ergebnisse werden auf betrügerische Inhalte durchsucht und die Keywords entsprechend verfeinert, um eine bessere Suche zu erhalten.

Beispiel:

- Suche in der meta-Werbebibliothek nach „tesler“ (angebliches „Trading-System“ mit dem viele Tradingplattformen werben)
- Zahlreiche Ergebnisse, auch seriöse, werden angezeigt und vom Team gescannt
- Textbaustein „a secure future in 5 minutes“ wird in verschiedenen betrügerischen Anzeigen gefunden
- Suche in der meta-Werbebibliothek nach „a secure future in 5 minutes“ zeigt rund 30 betrügerische Anzeigen

Aufgrund des höherschwelligen Zugangs zu geeigneten Suchphrasen liegt der Fokus bei der Suche in der Werbebibliothek in einem ersten Schritt auf Investmentbetrug.
Aktuell werden folgende Textphrasen gesucht:

- "a secure future in 5 minutes"
- "OpenAI verändern das Leben einfacher Menschen"
- "Ist dies das Ende ihrer Karriere?"

In einem zweiten Schritt wird die Suche auch auf die Themen Medikamentenbetrug und unseriöse Nahrungsergänzungsmittel ausgeweitet.

Probleme

Bei der Suche sind aber auch Probleme aufgetreten. Das betrifft vor allem die folgenden Punkte:

- 1. Suchbegriffe ändern sich öfter als bei Websites (Bspw. ""OpenAI verändern das Leben einfacher Menschen" Hier gab es Anfang März viele Anzeigen, zum Zeitpunkt Juli 2023 gar keine mehr).
- 2. Es passiert mittlerweile öfter, dass wir zahlreiche Ergebnisse über die adLibrary erhalten, aber keines der Ergebnisse auf eine brauchbare Domain führt. Die Domain, die in der adLibrary angezeigt wir, ist eine andere als jene, die im Facebook- oder Instafeed angezeigt wird. Hier müssen wir testen, welche Domains wir über die Suche mit der Werbebibliothek-API erhalten können.

Diesen Fragestellungen wird in der zweiten Projekthälfte nachgegangen.

2 Technische Grundlagen

2.1 Anforderungsprofil

Folgende technische Anforderungen stellen sich an die Crawler:

- Anlegen und Bearbeiten von Kategorien (Trading-Plattformen, Fake-Shops für Medikamente, Fake-Shops für Nahrungsergänzungsmittel)
- Anlegen und Bearbeiten von Suchbegriffen
- Automatisierte Suche nach angelegten Suchbegriffen
- Wöchentliches Ausspielen der neu indexierten Ergebnisse in Tabellenform mit folgenden Metadaten:
    a. anklickbare Domain
    b. URL auf der die genaue Suchphrase zu finden ist
    c. Kategorie
    d. Suchphrase
    e. Textsnippets
- Ergebnisse nicht (oder hervorgehoben) anzeigen, wenn:
    a. Domain sich auf einer Liste der Watchlist Internet befindet

b. es sich um eine Whitelist-Domain handelt

- Filtermöglichkeit der Ergebnisse nach Kategorie
- Schnittstelle für die automatisierte Anbindung an die Datenbank des Fake-Shop Detector (siehe 2.3.)

2.2 Dashboard

Die Crawler werden vom gesamten Team der Watchlist Internet genutzt und müssen dementsprechend eine möglichst intuitive Bedienoberfläche bereitstellen, um die Crawler zu konfigurieren und die Ergebnisse einzusehen. Gleichzeitig sollte der Crawler automatisiert arbeiten. Manuelle Eingriffe sind nur zu Beginn (Eingabe von Suchphrasen) und am Ende des Prozesses (Qualitätssicherung) geplant. Das Dashboard gilt es also dementsprechend einfach zu halten. Folgende drei Menüpunkte sollen im Dashboard angezeigt werden:

- **Nutzer:innen-Verwaltung:** Möglichkeit, Passwort zu ändern
- **Eingabemaske – Kategorien / Texte bearbeiten:** Hier können Redakteur:innen Kategorien anlegen. Für die jeweiligen Kategorien können wiederum Suchphrasen erstellt und hinzugefügt werden. Sowohl die Kategorien als auch die Suchphrasen müssen bearbeitbar sein.
- **Ausgabemaske – Ergebnisse anzeigen:** Die Ergebnisse müssen wöchentlich in einer übersichtlichen und verständlichen Tabelle dargestellt werden. Damit das gewährleistet wird, müssen die angezeigten Metadaten (siehe 2.1. Anforderungsprofil) möglichst gering gehalten werden. Hier sollen die Redakteur:innen schnell einen Überblick über spezifische Betrugsmuster und aktuelle Trends zu erhalten.

2.3 Anbindung an die Datenbank des Fake-Shop Detector

Die eigentliche Bearbeitung der Ergebnisse findet in der Datenbank des Fake-Shop Detector statt. Dafür müssen die Crawler eine Schnittstelle zur Verfügung stellen, mit Hilfe derer die Suchergebnisse (insbesondere die Domains) automatisiert an die Datenbank des Fake-Shop Detector übergeben werden. Weitere Metadaten werden vorerst in einem Kommentarfeld angezeigt. Geplant ist außerdem eine Graph-DB basierte Lösung, damit zukünftig Metadaten der Crawler (aber auch von anderen Quellen) in der Fake-Shop Detector Datenbank beliebig dargestellt werden können.

Außerdem muss in der Datenbank ersichtlich werden, dass die Domains von den Crawlern stammen. Damit wird gewährleistet, dass wir die Effektivität dieses Tools messen können und dass keine Crawler-Ergebnisse verloren gehen.

Landen die gefundenen Domains in der Datenbank, werden diese (1) von der Künstlichen Intelligenz des Fake-Shop Detector mit einem Risikoscore bewertet, (2) entsprechend der Qualitätssicherung vom Team der Watchlist Internet bearbeitet und (3) betrügerische Websites automatisiert auf den entsprechenden Listen der Watchlist Internet veröffentlicht.

2.4 Qualitätssicherung

Die Websites landen dabei sowohl auf den Listen der Watchlist Internet, wenn der Fake Shop Detector diese mit eine Risikoscore von „sehr hoch“ (> 90%) bewertet oder wenn die Domain vom Team der Watchlist Internet als Finanzbetrug oder als Fake-Shop eingestuft wird. Eine solche manuelle Überprüfung der Suchergebnisse garantiert die Qualitätssicherung der durch die Crawler erhaltenen Daten.
Da die Crawler-Ergebnisse letztendlich auf den Listen der Watchlist Internet veröffentlicht werden, aber auch im Fake-Shop Detector vor diesen Seiten gewarnt wird, handelt es sich dabei um ein sicherheitsrelevantes Service für Konsument:innen. Dieses Service beruht auf Vertrauen, falsche Einschätzungen müssen daher möglichst vermieden werden. Eine Qualitätssicherung der Expert:innen der Watchlist Internet ist daher unerlässlich.