Konzept: Workflow-Management zur Teilautomatisierung der Arbeitsprozesse der Watchlist Internet

Die Watchlist Internet (WL) ist die größte Präventionsplattform gegen Internetbetrug im deutschsprachigen Raum. Als niederschwellige, redaktionell betreute Plattform ermöglicht sie es Konsument:innen, betrügerische Onlineangebote zu melden und sich gleichzeitig über aktuelle Betrugsmaschen zu informieren. Täglich bearbeitet die Redaktion der WL hunderte User-Meldungen, prüft maschinell detektierte Risiken und Wertungen des Fake-Shop Detector (FSD) sowie Suchergebnisse von Crawlern.

Dadurch kann die WL theoretisch vor mehr Betrugsfällen warnen als je zuvor. In der Praxis zeigt sich jedoch, dass die Redaktion mit den in den vergangenen Jahren entwickelten Automatisierungs-Tools an ihre Grenzen stößt. Aktuell bleiben zahlreiche verdächtige Domains unbearbeitet. Notwendig ist eine Optimierung der Arbeitsprozesse.

Ziel des hier vorgestellten Konzepts ist die Entwicklung eines Workflows-Management-Systems, das durch regel- und risikobasierte Klassifikation eine teilautomatisierte Verarbeitung ermöglicht, um die Domaineinnträge auf den Warnlisten der WL zu erhöhen und gleichzeitig den hohen Anspruch an Qualitätssicherung der WL zu wahren.

1. Workflow: Status Quo

Der aktuelle Arbeitsprozess der Watchlist Internet ist stark manuell geprägt: Aktuell sichtet die Redaktion der Watchlist Internet tagtäglich hunderte Meldungen, in denen Konsument:innen Hinweise auf Betrugsfallen geben. Sie analysieren Impressums- und Whois-Daten, gleichen neu gemeldete Domains mit bestehenden Warnlisten ab und recherchieren Erfahrungsberichte von Konsument:innen. Neu erkannte Betrugsseiten werden manuell in die Datenbank des Fake-Shop Detector (FSD) übertragen, von wo sie einmal täglich in die öffentlichen Warnlisten ausgespielt werden.

Parallel liefern Crawler (siehe netidee-Projekt Fraud Seeker), Scraper und das FSD-Plugin kontinuierlich Domains mit potenziellem Betrugsrisiko. Wurden diese Daten lange Zeit nach gewissen Regeln (bspw. nur .at-Domains, nur Risikoscore hoch etc.) manuell von der Redaktion überprüft und so auf die Warnlisten übertragen, bleiben sie aktuell aufgrund der Menge der Daten unbearbeitet. Damit besteht das Risiko, dass Konsument:innen nicht rechtzeitig vor

betrügerischen Websites gewarnt werden. Notwendig sind weitere Analysen, um einen Teil der betrügerischen Domains automatisiert auf die Warnlisten zu übertragen bzw. einen weiteren – stark reduzierten – Teil an die manuelle Qualitätssicherung zu übergeben.

2. Workflow-Management-System

2.1. Zielsetzung

Durch die Entwicklung eines Workflow-Management-Systems sollen Domains aus unterschiedlichen Datenquellen teilautomatisiert und risikobasiert verarbeitet und bewertet werden. Damit werden folgende Ziele verfolgt:

- Entlastung der Redaktion der Watchlist Internet in ihrem derzeitigen Arbeitsprozess
- Veröffentlichung automatisiert detektierter Domains auf den Warnlisten der Watchlist Internet
- Beibehaltung des derzeitigen Anspruchs hinsichtlich Transparenz und Qualitätssicherung

2.2. Workflow-Architektur

Um diese Ziele umzusetzen, werden Domains basierend auf Risikobewertung, Quelle und weiteren Indikatoren automatisiert eingestuft. Hochriskante Fälle werden direkt veröffentlicht, während bei mittleren Risikofällen Human-in-the-Loop Ansätze integriert werden.

Die Workflow-Architektur gliedert sich wie folgt:

Datenquellen:

- Meldungen von Konsument:innen
- Crawler- und Scraper-Ausgaben (z. B. Google-Crawler, Warnlisten)
- KI-basierte Risikobewertung aus dem FSD-Plugin

Outcomes:

- Automatisiertes Ausspielen
  - auf die Warnlisten der WL
  - in das Plugin des Fake-Shop Detector
- Human-in the-Loop Qualitätssicherung
  - Qualitätssicherung durch Expert:innen der WL
  - Qualitätssicherung durch Clickworker

Workflows

- **Meldungen von Konsument:innen ➔ Qualitätssicherung durch Expert:innen:** Expert:innen prüfen die gemeldeten Domains und tragen bestätigte Fälle in die FSD-Datenbank ein. Von dort werden sie auf die Warnlisten der Watchlist Internet ausgespielt und im FSD-Plugin als betrügerisch angezeigt.
- **Crawler- und Scraper-Ausgaben ➔ regel- und risikobasierte Klassifikation mittels Impressums-Check und Trustpilot-Analyse (Automatisiertes Ausspielen oder Qualitätssicherung durch Expert:innen):** Crawler- und Scraper-Ausgaben werden mittels Impressum-Check und Trustpilot-Analysen überprüft und nach einer regelbasierten Klassifikation (siehe 2.3.) entweder direkt auf die Watchlist und das FSD-Plugin übertragen oder der Redaktion zur Überprüfung übergeben.
- **KI-basierte Risikobewertungen aus FSD-Plugin ➔ regel- und risikobasierte Klassifikation mittels KI-Bewertung (Automatisiertes Ausspielen oder Qualitätssicherung durch Clickworker):** Über das Plugin aufgerufene Domains erhalten einen KI-basierten Risikoscore. Domains, die mit einem sehr hohen Risikoscore, werden auf die WL ausgespielt; Domains mit einem hohen Risikoscore werden durch Clickworker überprüft.

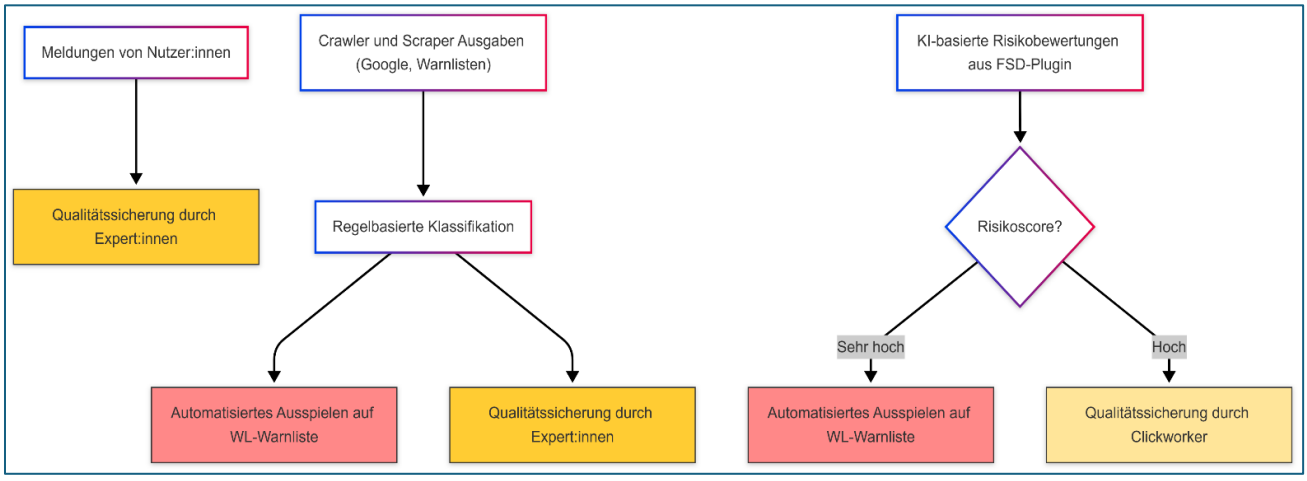

Abbildung 1: Workflow-Architektur

2.3. Regel- und risikobasierte Klassifikation mittels Impressums-Check & Trustpilot-Reviews

Für die teilautomatisierte Verarbeitung von Domains aus unterschiedlichen Quellen wird eine regel- und risikobasierte Klassifikation angewandt. Dabei werden die Daten mittels Impressums-

Check und Trustpilot-Analyse in Risikostufen eingeteilt. Abhängig von Quelle und Score ergeben sich unterschiedliche Workflows, die entweder zu einer direkten automatischen Veröffentlichung auf der Watchlist Internet führen oder eine manuelle Überprüfung durch Expert:innen erfordern.

Impressums-Check

Ein Großteil, der auf der WL gelisteten betrügerischen Websites zeichnet sich dadurch aus, dass kein Impressum zu finden ist. Bei wenigen gut gemachten Fake-Shops ist zwar ein Impressum vorhanden, eine kurze Recherche zeigt jedoch, dass die Daten nicht stimmen oder zu einem anderen – seriösen – Unternehmen gehören.

Entsprechend ist die Überprüfung des Impressums für die Expert:innen der WL ein zentrales Kriterium zur Bewertung von Domains. Der Impressums-Check automatisiert diesen Prozess entlang folgendem Flow:

- Scraping Impressum & UID (Umsatzsteuer-Identifikationsnummer)
- Validierung UID (über das MwSt-Informationsaustauschsystems, MIAS, der Europäischen Kommission)
- Google-Suche UID & Abgleich der ersten drei Google-Ergebnisse mit der Ursprungsdomain

Darauf aufbauend wird der Risikoscore wie folgt berechnet:

- **Risikoscore == 100**
  - UID wurde nicht gefunden
  - Validierung der UID fehlgeschlagen
  - Innerhalb der ersten drei Google-Suchergebnisse wurde eine Anti-Fraud Website gefunden
- **Risikoscore == 80**
  - Ursprungs-URL wurde nicht in den ersten drei Google-Ergebnissen gefunden
- **Risiko == 0**
  - keine der oben genannten Bedingungen trifft zu

Trustpilot-Reviews

Die automatisierte Analyse von Trustpilot-Reviews ist ein weiteres Tool, dass bei der Einschätzung von betrügerischen und unseriösen Domains unterstützen kann. Für diese Analyse werden sofern vorhanden folgende Daten gescrapt:

- **Angaben zum Unternehmen:**
  - Profil beansprucht (ja/nein, wenn ja: Datum der Beanspruchung)
  - Unternehmen verifiziert: Identitätsnachweis (ja/nein), Kontaktdaten (ja/nein), Eigentümerschaft (ja/nein), Bankkonto (ja/nein)
- **Bewertungen:**
  - Anzahl der Bewertungen
  - Trustscore[1]
  - Anteil der unterschiedlichen Bewertungen (5-Sterne, 4-Sterne, 3-Sterne, 2-Sterne, 1-Sterne) in %
- **Interaktion vonseiten des Unternehmens**
  - Anzahl der Antworten auf negative Bewertung
  - Datum letzte Antwort auf negative Bewertung
  - Durchschnittliche Tage für Antworten
- **Kennzahlen zu Bewerter:innen**
  - Durchschnittliche Anzahl der Bewertungen der User
  - Prozentsatz der User mit Profilbild
  - Prozentsatz der User aus Österreich
  - Prozentsatz der User aus Deutschland
- **Trustpilot-Profil geschlossen (ja/nein)**

Auf Basis dieser Daten wird ein Risikoscore[2] wie folgt berechnet:

- **Risikoscore == 100**
  - Trustpilot-Profil wurde geschlossen
  - ≥ 60% 1-Sterne-Bewertungen
  - insgesamt ≥ 50% 1- und 2-Sterne-Bewertungen
- **Risikoscore == 0**
  - ≥ 70% 5-und 4-Sterne-Bewertungen
  - keine der oben genannten Bedingungen trifft zu UND Profil wurde beansprucht UND min. eine der Verifizierungsmethoden wurde durchgeführt
- **Risiko == 50**
  - keine der oben genannten Bedingungen trifft zu UND Profil wurde beansprucht UND keine der Verifizierungsmethoden wurde durchgeführt

---

[1] Laut Trustpilot wird der Trustscore anhand der durchschnittlichen Bewertungen berechnet, gleichzeitig fließen aber auch Faktoren wie Anzahl und Alter der Bewertungen ein oder ob ein Unternehmen seine Kund:innen aktiv einlädt zu bewerten.

[2] Der Risikoscore wird nur berechnet, wenn zu einer Domain mehr als zehn Bewertungen abgegeben wurden.

Workflow

In einem iterativen Prozess wurde das Zusammenspiel zwischen Trustpilot-Reviews, Impressums-Check und den bestehenden Crawlern und Scrapern erprobt und folgender Automatisierungs-Flow je Quelle erarbeitet:

Quelle: Google-Crawler / Datenbank

- Automatisiertes Ausspielen auf die WL und in das FSD-Plugin
  - Impressums-Check Risikoscore == 100
  - Trustpilot geschlossen == ja
- Qualitätssicherung durch Expert:innen
  - Impressums-Check Risikoscore == 80
  - Trustpilot-Risikoscore == 50

Quelle: Scraper I

- Automatisiertes Ausspielen auf die WL und in das FSD-Plugin
  - Trustpilot-Check Risikoscore == 100
  - Trustpilot geschlossen == ja
- Qualitätssicherung durch Expert:innen
  - Impressums-Check Risikoscore == 100 oder 80
  - Trustpilot-Risiko == 50

Quelle: Scraper II

- Automatisiertes Ausspielen auf die WL und in das FSD-Plugin
  - Impressums-Check Risikoscore == 100
  - Trustpilot-Risikoscore == 100
  - Trustpilot geschlossen == ja
- Qualitätssicherung durch Expert:innen
  - Impressums-Check Risikoscore == 80
  - Trustpilot-Risikoscore == 50

Entspricht die Quelle eine der genannten, sind jedoch keine Bedingungen erfüllt, verbleibt die Domain ohne Klassifikation in der Datenbank. Für diese Fälle wird ein allgemeiner Risiko-Score von überdurchschnittlich vergeben.

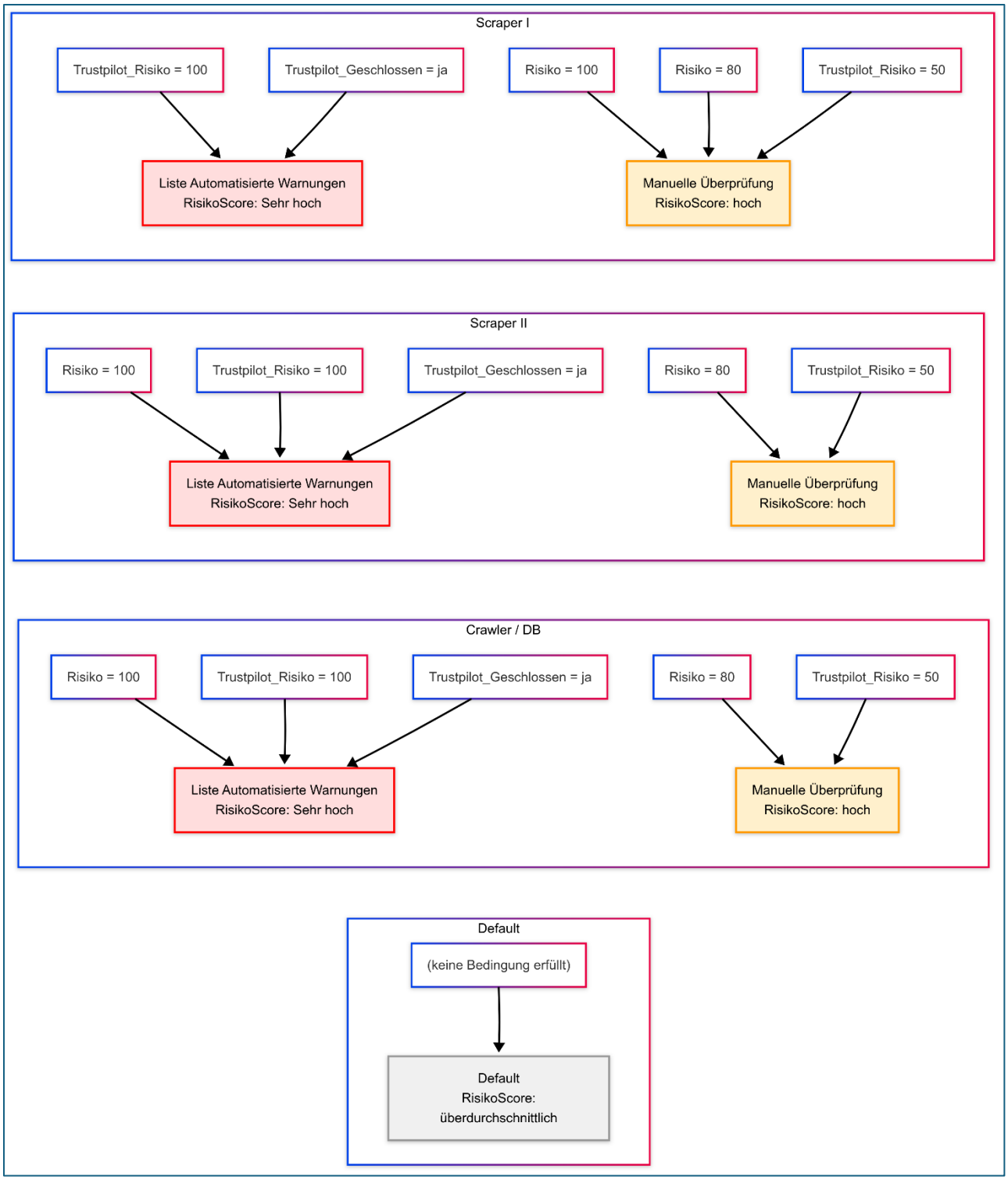

Abbildung 2: Workflow der regel- und risikobasierten Klassifikation

2.4. Human-in-the-Loop Integration

Die erstellte Workflow-Architektur, inkl. regel- und risikobasierter Klassifikation ermöglicht es einen großen Teil der Domains automatisiert zu verarbeiten und so das Team der Watchlist Internet zu entlasten.

Die Einbindung menschlicher Expertise bleibt jedoch unverzichtbar. Insbesondere in Grenzfällen, bei neu auftauchen Betrugsmustern oder unvollständigen Daten kann eine ausschließlich maschinelle Bewertung der Domains zu Fehlklassifikationen führen. Eine Human-in-the-Loop (HITL) Integration stellt sicher, dass die Automatisierung mit manueller Qualitätssicherung kombiniert wird und baut auf folgenden Arten der menschlichen Überprüfung auf:

- **Qualitätsüberprüfung durch Expert:innen:** Expert:innen der WL überprüfen Domains, die mit hoher Wahrscheinlichkeit betrügerisch sind. Dabei erfolgt eine manuelle Überprüfung des Impressums und weiterer Website-Inhalte, der Whois-Daten sowie von Konsument:innen-Erfahrungen.
- **Qualitätsüberprüfung durch Clickworker:** Für standardisierte Prüfungen können externe Clickworker eingesetzt werden. Erprobt werden muss erst, welche Bewertungen von Clickworkern zufriedenstellend erarbeitet werden können.
- **Kontinuierliches Feedback:** Die Ergebnisse der menschlichen Qualitätssicherung fließen zurück in die Automatisierungs-Logik, die dadurch sowohl kontinuierlich verbessert wird als auch an neue Betrugsformen angepasst werden kann.

3. *Fazit*

Das entwickelte Workflow-Management-System entlastet die Redaktion, in dem die Arbeitsprozesse deutlich effizienter gestaltet werden. Die regelbasierte Klassifikation ermöglicht es hochriskante Fälle automatisiert auf die Warnlisten zu übernehmen, während Verdachtsfälle mit hohem Risiko gezielt mittels HITL-Integration an die menschliche Überprüfung weitergegeben werden. Die Redaktion wird so von Routineaufgaben entlastet, gleichzeitig kann die Watchlist Internet bestehende Tools optimal nutzen, um die Quantität und die Aktualität der Warnlisten zu erhöhen - ohne den hohen Qualitätsanspruch der Plattform zu verlieren.

So kann durch die Kombination aus Automatisierung und menschlicher Expertise auch ein skalierbares System geschaffen werden, das auch mit wachsender Zahl an Meldungen sowie an mittels Automatisierung detektieren Daten Schritt halten kann. Langfristig trägt das Workflow-Management-System dazu bei, dass Konsument:innen noch schneller und umfassender vor betrügerischen Angeboten geschützt werden und die Watchlist Internet ihre Rolle als führende Präventionsplattform im deutschsprachigen Raum weiter ausbauen kann.