GBS

Content Recognition Engine (CORE)

- Intelligente Inhaltsanalyse -

Umfassende Klassifikation und Spam-Abwehr

Expertise matters

Inhalt

1 Spamabwehr – Ein Überblick 2

1.1 Effiziente Spamabwehr – Vorgehensweise 3

2 Einsatzmöglichkeiten von CORE 4

3 Wie funktioniert die CORE-Technologie? 5

3.1 Überblick über die Funktionsweise der SVM 6

3.2 Warum SVM? 7

4 Praktischer Einsatz von CORE 8

4.1 Textanalyse mit CORE 8

4.2 Praxistipps 9

4.2.1 Kategorisieren 9

4.2.2 Trainieren und Validieren 9

4.2.3 Nachkategorisieren 10

5 CORE - Highlights und Features 11

1 Spamabwehr – Ein Überblick

SPAM – jeder kennt ihn, keiner mag ihn und wirklich niemand will ihn haben.

Die Abwehr von SPAM ist nach wie vor eine technologische Herausforderung, die durch den ständigen Wandel und den Einfallsreichtum der Spammer immer neue Abwehrmaßnahmen zu erfordern scheint. Was gibt es für Möglichkeiten, diesen elektronischen Müll bereits zu erkennen, bevor er im Postfach des Adressaten landet?

Das Spektrum der Spamabwehrmaßnahmen erstreckt sich von einfachsten String-Operationen bis hin zu komplexen Skriptprogrammen. Es kann in folgende Methoden eingeteilt werden:

- **Wortlisten** stellen eines der ältesten Verfahren zur Spamabwehr dar. Das Risiko von sogenannten *False Positives* (= erwünschte Mails, die als SPAM klassifiziert werden) ist recht hoch und das Verfahren als primärer Abwehrmechanismus ungeeignet.
- **Realtime Blacklists** (RBLs oder DNSBLs) beschreiben eine Methode, bei der im Internet verfügbare Blacklist-Server angefragt werden, um einen E-Mail-Versender als Spammer zu erkennen. Dieses Verfahren ist sehr unsicher, da häufig Mailserver seriöser Versender nach einer Attacke von Spammern auf solche Listen geraten. Außerdem sind die dort gespeicherten Adresslisten sehr schnell veraltet. Die Fehlerrate von RBLs liegt bereits heute bei 60% (d.h. False Positives und nicht erkannte SPAM-Mails) und mehr.
- **Checksummen-Verfahren** bilden für jede eingehende E-Mail eine eindeutige Checksumme und legen diese nach einer Klassifikation in Internet-Datenbanken ab. Andere E-Mail-Server können eingehende E-Mails mit dieser Datenbank vergleichen und als SPAM eingeordnete Mail erkennen. Einige der existierenden Lösungen bieten einen Dienst an, der, vergleichbar mit der Vorgehensweise von Virenschutzprogrammen, aktuelle Updates/“Pattern“ für die Spamerkennung zur Verfügung stellt. Das Checksummen-Verfahren basiert auf der Annahme, dass SPAM-Mails potenziell vervielfältigte Kopien der selben E-Mail sind und somit eindeutig auf allen Empfängerservern zugeordnet werden können. Damit es funktioniert, ist es auf ein möglichst großes Netzwerk an Teilnehmern angewiesen. Mittlerweile wird das Verfahren jedoch von Spammern häufig dadurch umgangen, dass die generierten SPAM-Mails sich nur in der Checksumme aber nicht im lesbaren Text unterscheiden oder die Massenmails als personalisierte Einzelmails verschickt werden.
- **HTML Decoding** begegnet der Tatsache, dass Spammer in steigendem Maße HTML-basierte E-Mails versenden, um Standardmethoden der SPAM-Bekämpfung zu umgehen. Spammer setzen dazu die Tagging-Möglichkeiten von HTML in einer Form ein, die es zwar dem Client erlaubt, eine lesbare E-Mail anzuzeigen, jedoch im HTML-Quellcode keinen zusammenhängenden Text enthält. Die HTML Decoding Methoden dekodieren die HTML-Mail und prüfen von typischen Formatierungen wie z.B. Großbuchstaben und Farbe bis hin zu enthaltenen HTML-Links.
- **Skriptfilter** stellen eine effektive, aber hochkomplexe und aufwändige Methode dar, mit Hilfe von Perl oder Sieve Skripten eine maßgeschneiderte SPAM-Abwehr zu erreichen. Hoher Entwicklungs- sowie fortlaufender Wartungsaufwand machen Methoden dieser Kategorie allerdings ineffizient und unüberschaubar.

- **Heuristische Ansätze** versuchen in E-Mails bestimmte Textmuster zu erkennen, die eine Klassifikation in SPAM oder Nicht-SPAM ermöglichen. Neben diversen Ansätzen neuronaler Netze und Bayes-Filtern gehört auch das innovative CORE (= Content Recognition Engine) Analyseverfahren zu dieser Kategorie.
- Die naiven Bayes-Filter sind ein häufig eingesetztes Verfahren. Diese statistischen Verfahren gehen auf den Pfarrer Thomas Bayes zurück, der vor 250 Jahren die zugrunde liegende Wahrscheinlichkeitsformel entwickelte. Anhand kalkulierter Wahrscheinlichkeiten werden neue E-Mails als SPAM oder Nicht-SPAM klassifiziert. Naive Bayes-Filter stellen keinen Zusammenhang zwischen einzelnen Dokumentenmerkmalen her und sind damit z.B. für eine Multikategorisierung ungeeigneter bzw. es müssen dafür Zusatzmodule implementiert werden.
- CORE dagegen eliminiert einerseits die Probleme herkömmlicher Verfahren wie beispielsweise die langsame und aufwändige Anpassung an neue SPAM-Methoden und andererseits den im Allgemeinen hohen Trainingsaufwand anderer heuristischer Verfahren.

1.1 Effiziente Spamabwehr – Vorgehensweise

Für eine effiziente und leistungsstarke Spamabwehr werden die verschiedenen Methoden kombiniert. Dabei hat sich in der Praxis folgende Reihenfolge bei der E-Mail Analyse bewährt:

1. Adressprüfung mit Hilfe von Blacklists (verbotene E-Mail-Adressen und Domänen) und firmenspezifischen Whitelists (erlaubte Adressen und Domänen). Die Whitelists enthalten geschäftsrelevante Absenderadressen von z.B. Kunden, Lieferanten, Newsletter, Diskussionsforen.
2. Prüfung der Betreffzeile mittels Wortlistenprüfung auf einfache Schlüsselworte (100% Stoppworte).
3. Prüfung des Nachrichtentextes mittels Wortlistenprüfung und HTML-Analyse. Die Wortlisten in diesem Prüfschritt sollten wie die Wortliste für den Betreff 100% Stoppworte enthalten, um entsprechende Mails sofort als SPAM aussortieren zu können. Wortlisten mit 100% Stoppworten sind in der Regel kürzer und erfordern weniger Pflegeaufwand.
4. Prüfung des E-Mail-Inhalts mit CORE.

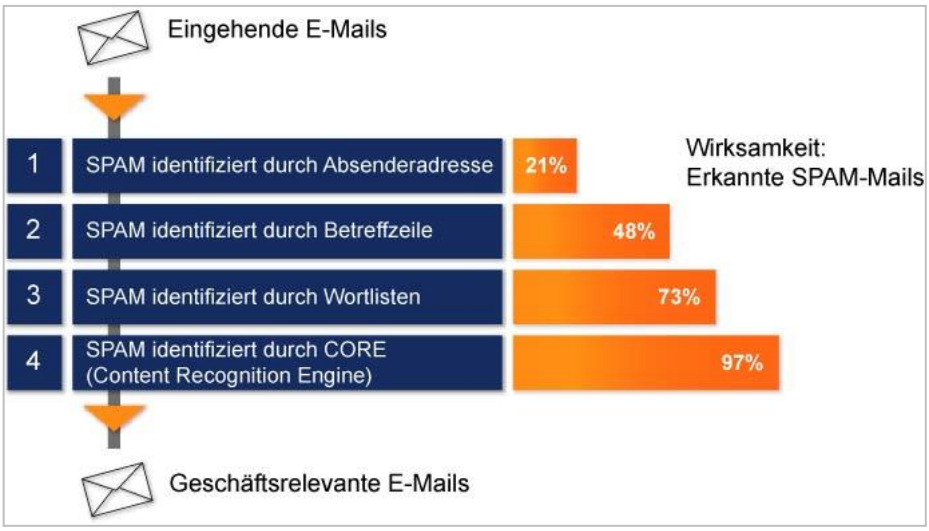

Abbildung 1: SPAM-Abwehr mit iQ.Suite Wall

Ohne den Einsatz von CORE können höchstens 73% der SPAMs (s. Abb.) beseitigt werden – mit abnehmender Tendenz, da neue Spammer-Tricks diese statischen Methoden umgehen können. Nur mit Hilfe von CORE können 97% der eingehenden SPAM-Mails erkannt werden – auch zukünftig, da CORE lernfähig ist.

Merkmale von CORE für Anti-SPAM:

- Erkennungsrate von SPAM ist höher als 95%
- Reduzierung der False Positives1 innerhalb der als SPAM erkannten E-Mails auf unter 0,1%.
- Einfache Kategorisierung in nur zwei Klassen: SPAM und NOSPAM.
- Multikategorisierung zur Unterscheidung verschiedener Klassen von Mail: SPAM, Newsletter, Angebote, Aufträge, private Post … Oft haben z.B. Newsletter und SPAM-Mails einen ähnlichen Aufbau und liegen auch inhaltlich ziemlich nahe beieinander. Mit CORE ist eine genaue Zuordnung durch die Einführung von mehreren Kategorien möglich.

2 Einsatzmöglichkeiten von CORE

CORE analysiert und klassifiziert E-Mails. Außerdem ist es lernfähig. Damit ergeben sich folgende praktische Einsatzmöglichkeiten:

- **Anti-SPAM:** Spammer verwenden immer ausgeklügeltere Methoden, um ihre E-Mails an den statischen Filtern zur Schlüsselwort-, Betreff- oder Text-Analyse vorbei zu schleusen. So werden Wörter oder Sätze derart manipuliert, dass sie zwar in keiner Wortliste stehen, vom Browser bzw. E-Mail-Client des Benutzers jedoch korrekt dargestellt werden oder vom menschlichen Leser verstanden werden. Häufig werden deswegen Verschleierungen im Betreff wie Burn F_a_t, H:A:R:D:C:O:R:E E:X:T:R:E:M:E eingesetzt.

[1] **False Positives sind erwünschte Mails, die als SPAM klassifiziert werden**

- Mit herkömmlichen Methoden unmöglich zu erkennen sind die immer häufiger auftretenden SPAMs mit sehr geschäftsmäßig oder privat aufgemachten Nachrichtentexten, die mittels lexikalischer Analyse von normaler Korrespondenz nicht zu unterscheiden sind.
- Beim Einsatz von CORE nützen alle derartigen Verschleierungstaktiken dem Spammer überhaupt nichts, seine Mails werden trotzdem als das erkannt, was sie sind: SPAM. CORE beschränkt sich nicht auf einzelne Wörter oder Satzteile, sondern analysiert den gesamten Inhalt der E-Mail. Durch seine Lernfähigkeit ist CORE damit auch zukünftigen, neuen Spammer-Tricks gewachsen.
- **Dokumentenschutz:** CORE ermöglicht einen besseren Schutz von Dokumenten, die Firmeninterna enthalten. Wenn eine entsprechende Kategorie angelegt ist, können z.B. alle ausgehenden Mails auf firmenrelevante Inhalte geprüft werden. Eine Inhaltsprüfung von E-Mail-Anhängen ist ebenfalls möglich.
- **E-Mail Response Management:** Eine weitere mögliche Einsatzmöglichkeit von CORE wäre die Optimierung der internen Betriebsabläufe durch E-Mail Response Management. Durch geeignete Klassifizierungen könnten z.B. E-Mails an den Kundensupport automatisch erkannt, klassifiziert und an den zuständigen Bearbeiter gesandt werden oder mit einer automatisch erzeugten Antwort direkt beantwortet werden.

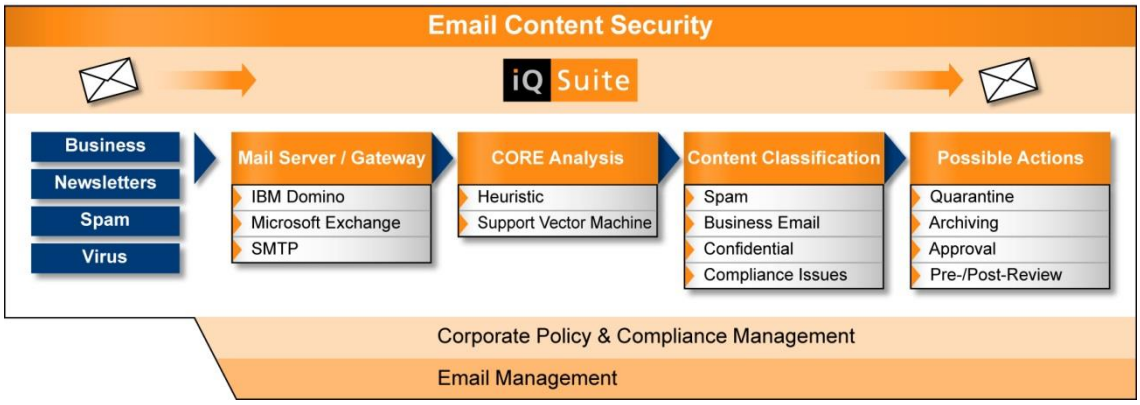

Abbildung 2: Sicherheit durch intelligente Klassifizierung

3 Wie funktioniert die CORE-Technologie?

CORE basiert auf der Methode der Support Vector Machines (SVM). SVM ist eine neue Generation von Lernsystemen, die auf Fortschritten in der statistischen Lerntheorie basieren. SVM ist eine leistungsfähige Methode für praxisnahe Anwendungen, wie z.B. Kategorisieren von Text oder Klassifizieren von Bildern.

Das Ziel von SVM beim Einsatz in CORE ist die optimale Einordnung neuer Dokumente in festgelegte Kategorien. Um dieses Ziel zu erreichen, wird mit Trainingsdokumenten ein Klassifikator trainiert. Die verwendeten Dokumente stellen eine repräsentative Menge der in einer Firma ein- und ausgehenden E-Mails (inklusive SPAM) dar. Je repräsentativer die Auswahl ist, desto besser arbeitet das Verfahren im Echtbetrieb.

Geeignete Trainingsdokumente stammen entweder direkt aus einem Mailjob, der die E-Mails in die Trainingsdatenbank sendet oder sie werden per Copy & Paste aus Benutzer-Mailboxen, Safe-Archiven oder Quarantäne-Datenbanken in die Trainingsdatenbank kopiert.

Ungeeignet sind weitergeleitete E-Mails, da die bei der Weiterleitung erzeugten Informationen mit in das Training eingehen und das Ergebnis verfälschen.

3.1 Überblick über die Funktionsweise der SVM

Die Methode der Support Vector Machines (SVM) arbeitet mit Vektoren. Jedes Dokument wird durch einen Vektor abgebildet. Dazu wird aus der Menge der in den einzelnen Texten (Anzahl „*m*“) enthaltenen Terme ein Vektor der Länge „*n*“ erzeugt. Terme sind alle in dem Dokument enthaltenen Teile, wie z.B. Worte oder HTML-Tags. Danach wird jedes einzelne Dokument auf diesen Vektor abgebildet, so dass ein *n x m* dimensionaler Vektorraum aufgespannt wird.

Die einzelnen Terme werden mit Hilfe des TF-IDF Verfahrens normiert und gewichtet. „TF“ steht dabei für Termfrequenz und repräsentiert die Häufigkeit eines Wortes in einem Dokument. „IDF“ ist die inverse Dokumentenfrequenz und sagt aus, in wie vielen Dokumenten ein Term vorkommt. Je häufiger ein Term in einem Text auftaucht, desto relevanter ist er für die Klassifizierung. Taucht derselbe Term dagegen in sehr vielen Dokumenten auf, sinkt seine Bedeutung für die Klassifizierung.

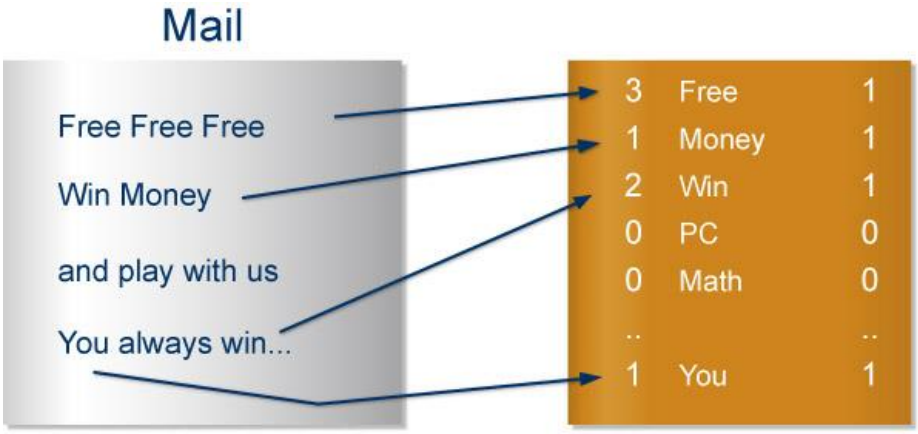

Abbildung 3: Beispiel für einen Dokumentenvektor

SVM ermittelt im Vektorraum eine Hyperebene, die positive und negative Trainingsdokumente einer Kategorie optimal voneinander trennt. Dazu ist ein lineares Optimierungsproblem zu lösen. Das Ergebnis ist die Klasse der Trainingsvektoren, die der Hyperebene am nächsten liegen. Diese Vektoren heißen Supportvektoren.

Im Unterschied zu anderen Vektor-Klassifizierungstechniken geht bei SVM die Komplexität des Klassifizierers mit in den Algorithmus ein. Dies verhindert, dass der gelernte Klassifizierer „übertrainiert“ wird und nur noch die Trainingsdokumente richtig kategorisiert. Zu diesem Übertrainieren oder „Overfitting“ kann es auch leicht bei den naiven Bayes-Filtern kommen, da diese aufgrund ihrer Selbstlernfähigkeit und fehlender Normierung besonders anfällig sind.

Folgende Abbildung veranschaulicht den Mechanismus von SVM im zweidimensionalen Raum für die Kategorie „SPAM“.

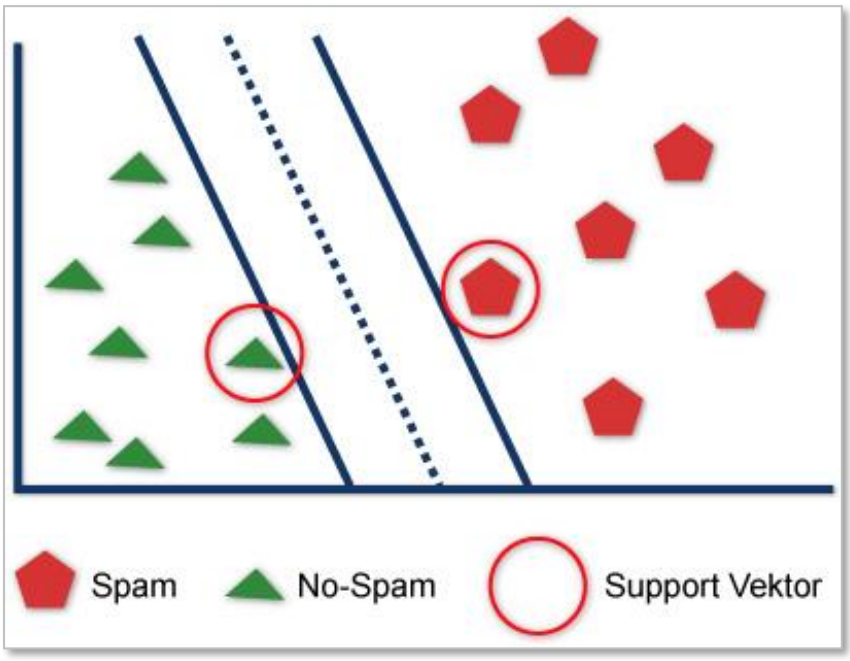

Abbildung 4: Konstruktion der Hyperebene, welche die Grenze zwischen zwei Klassen maximiert

3.2 Warum SVM?

Das in CORE eingesetzte SVM Verfahren bietet folgende Vorteile:

- SVM minimieren die Anzahl der Klassifizierungsfehler
- SVM sind robust gegen Overfitting
- SVM besitzen sehr gute Performanz, da effiziente Algorithmen zur Lösung des Optimierungsproblems existieren
- SVM liefern sehr gute Klassifizierungsergebnisse
- Text/Inhalt-Erkennung durch Einsatz eines modernen statistischen Verfahrens
- Einfache Wartbarkeit durch Lernvorgang
- Einfache Adaptierbarkeit an den firmenspezifischen Mailverkehr durch Lernvorgang

SVM sind heutzutage das beste frei verfügbare statistische Verfahren zur Textklassifizierung. So erreichte der Dortmunder Informatiker Thomas Joachims bereits 1997 mit SVM beim Klassifizieren der Reuters Textsammlung eine Genauigkeit von 86% über alle Kategorien und eine Genauigkeit von 91,4% über die zehn größten Kategorien. Die Reuters-Textsammlung gilt als die Standard Benchmark-Sammlung für Textklassifizierer. Sie besteht aus 9603 Dokumenten mit 118 Kategorien.

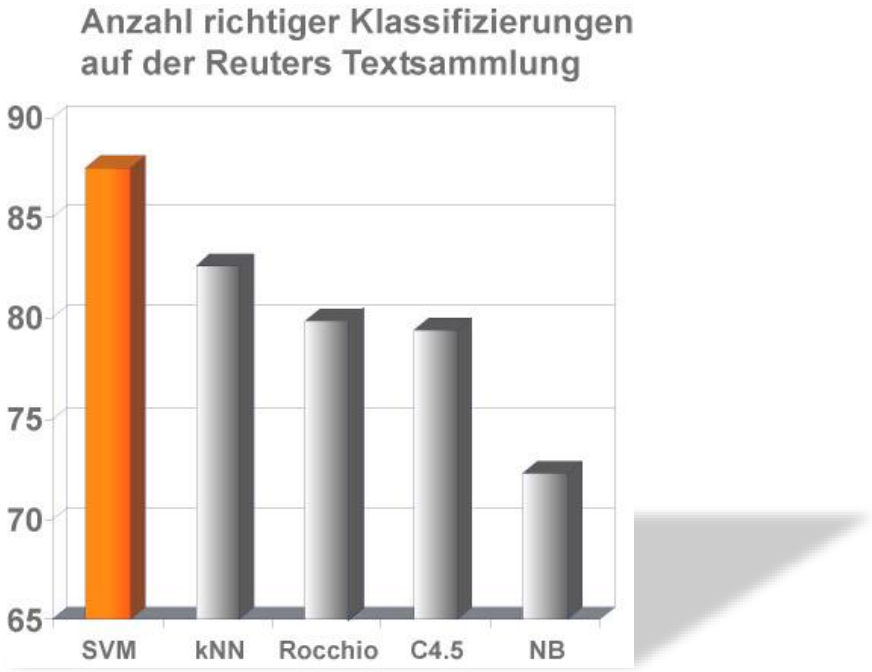

Abbildung 5: SVM als Klassenbester im Vergleich mit anderen Algorithmen: kNN = k (Anzahl) Nearest Neighbours, Rocchio = Roccio Algorithmen , C4.5 = Ross Quinlan Algorithmus, NB = Naïve Bayes

4 Praktischer Einsatz von CORE

Folgende Komponenten werden für den effektiven Einsatz von CORE verwendet.

- iQ.Suite Wall als Basis für die Prüfung der eingehenden E-Mails und den Lernvorgang
- Eine Referenzmenge an E-Mails, die vom Administrator oder anderen autorisierten Personen kategorisiert werden
- Eine Trainingsdatenbank als E-Mail-Container für den Lernvorgang

4.1 Textanalyse mit CORE

Der prinzipielle Ablauf zum erstmaligen Einrichten der Textanalyse mit CORE ist wie folgt:

1. Erstellen einer Trainingsdatenbank mit einer Referenzmenge an E-Mails
2. Kategorisieren der E-Mails durch den Administrator bzw. eine entsprechende Person
3. Konfigurieren des Datenbank-Trainingsjobs
4. Ausführen des Datenbank-Trainingsjobs
5. Konfigurieren des Datenbank-Validierungsjobs
6. Ausführen des Datenbank-Validierungsjobs
7. Konfigurieren des Mail-Prüfjobs – Testbetrieb
8. Ausführen des Mail-Prüfjobs – Testbetrieb

9. Nachtrainieren oder Nachkategorisieren der Trainingsdatenbank
10. Scharfschalten des Mail-Prüfjobs – Echtbetrieb

4.2 Praxistipps

4.2.1 Kategorisieren

- Zunächst wird vom Administrator oder anderen autorisierten Personen festgelegt, welche E-Mail-Kategorien es überhaupt gibt. Als Grundlage kann eine E-Mail-Referenzmenge dienen, die z.B. aus allen E-Mails eines Geschäftstages besteht.
- Bei einer Anti-Spam-Prüfung werden anschließend nur zwei Kategorien definiert: SPAM und NOSPAM. Es ist auch möglich, mehr als zwei Kategorien zu definieren und genauestens nach Art der Mails zu unterscheiden, z.B. Newsletter, SPAM, Business und diese dann evtl. auch noch nach Sprachen aufzugliedern.
- Pro Kategorie müssen mindestens zehn E-Mails vorhanden sein. Wenn Sie mehrere Kategorien definiert haben, ist es besser, eine kleinere Lernmenge (ca. 25 – 50 E-Mails) mit typischen Vertretern der gewünschten Kategorie zu nehmen, als eine zu große. Bei nur zwei Kategorien nehmen Sie mindestens 200, besser etwa 500 Mails pro Kategorie.
- Für manche Newsletter und Diskussionsforen sollten Adressausnahmen konfiguriert werden, die diese Mails schon vorher aussortieren, da CORE sie in der Regel als SPAM klassifizieren wird. Geht dies nicht, wie z.B. bei Yahoo-Groups, die immer von einem anderen Absender kommen, sollten Sie mehrere Kategorien (mehr als SPAM und NOSPAM) aus Ihrer Referenzmenge bilden und für diese Gruppen eine neue Kategorie anlegen.

4.2.2 Trainieren und Validieren

- Die E-Mails für das Training müssen direkt aus einer E-Mail-Referenzmenge in die Trainingsdatenbank gesendet oder mit Copy und Paste hineingebracht werden. Im Nachrichtentext dürfen keinesfalls E-Mail-Header oder Informationen einer Weiterleitung auftauchen, da das den Text und damit die Textanalyse verfälscht.
- Eine Referenzmenge an E-Mails wird am einfachsten erzeugt, indem ein iQ.Suite Wall-Job angelegt wird, der alle eingehenden Mails z.B. über einen Tag in eine zusätzlich angelegte Quarantänedatenbank kopiert.
- HTML-Mails müssen in der Urform in die Trainingsdatenbank gestellt werden, da der HTML-Code bei der Vektorisierung berücksichtigt wird.
- Alle Mails für die Trainingsdatenbank sollten längere Texte sein. Bei der Erstellung von mehreren Kategorien sollte insbesondere die Kategorie Business möglichst aus Mails bestehen, die nicht nur zwei Sätze beinhalten.
- Trainings- und Validierungsjobs sollten manuell angestoßen und sofort auf Erfolg geprüft werden. Sie sollten keinesfalls automatisch im Intervall laufen.

4.2.3 Nachkategorisieren

- Die *False Positives* aus der Quarantäne werden zur Nachkategorisierung per „Resend“ in die Trainingsdatenbank gesendet.
- Sind bei der Nachkategorisierung zu viele Mails in einer Kategorie, sollte die Kategorie entweder geteilt oder die kurzen Mails aus der Trainingsdatenbank entfernt werden.

Folgende Abbildung zeigt ein Beispiel für das Kategorisieren einer E-Mail-Referenzmenge bei mehr als zwei Kategorien

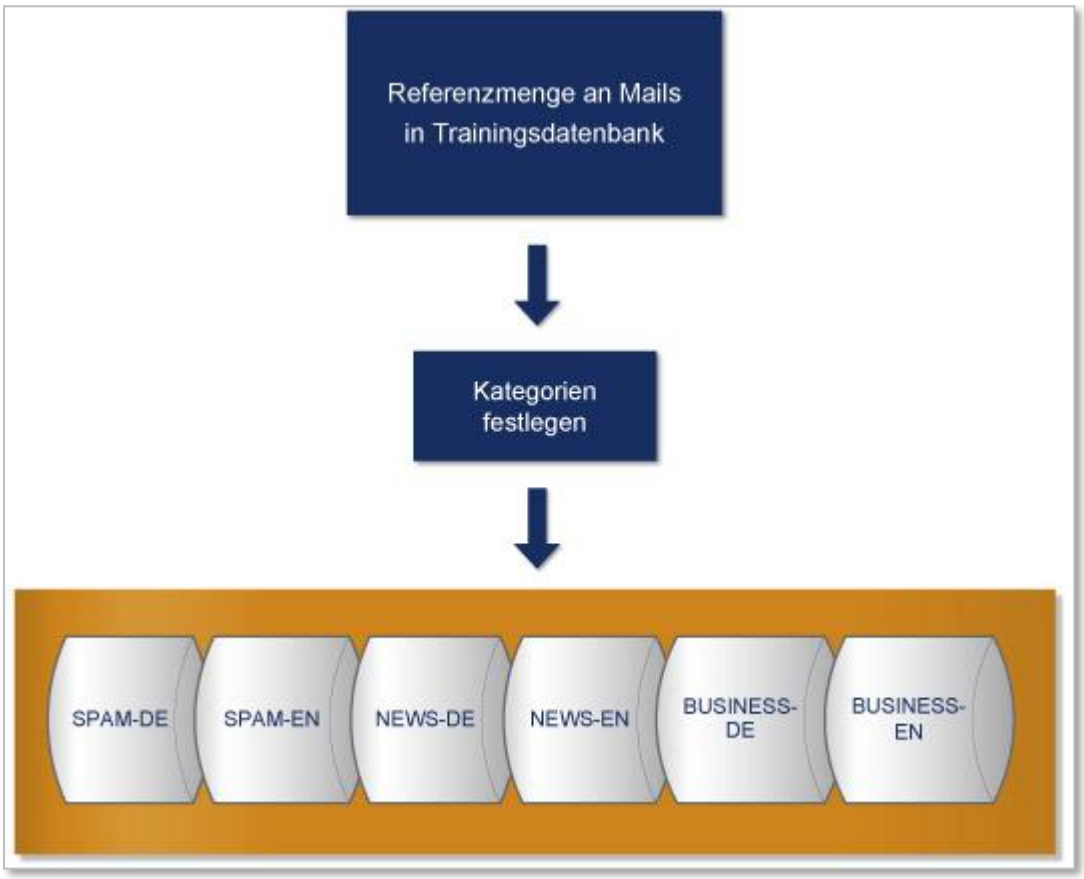

Abbildung 6: Kategorisieren von E-Mails (Beispiel)

5 CORE - Highlights und Features

Highlights

- Spam-Erkennung at its best

  Lernfähige Verfahren sind hervorragend geeignet für die Aufgabe der Spamabwehr. SPAM-Mails ändern sich so schnell; dass Wortlisten-basierte Techniken und alle sonstigen lexikalischen Verfahren nicht mit den Anforderungen mithalten können.

- Integration in die iQ.Suite

  CORE setzt nahtlos auf dem bestehenden Prüfablauf von E-Mails auf, ohne dass es zu Performance-Verlusten kommt.

- Modernste Technik zur Textklassifizierung

  Das in CORE verwendete Verfahren der Support Vector Machines (SVM) ist das modernste statistische Verfahren zur Textklassifizierung. Es kann ohne weitere Module für mehr als zwei Kategorien verwendet werden

- Firmenindividueller Zuschnitt

  Individuelle E-Mail-Kategorien mit firmenspezifischen Trainingsdokumenten ermöglichen optimale Anpassung und Unabhängigkeit. Im Gegensatz zu webgestützten Spam-Services ist zur Prüfung keine Verbindung zu externen Servern notwendig.

- Inhaltsanalyse und Dokumentenschutz

  Transparentes Management und Monitoring aller eingehenden und ausgehenden E-Mails. Firmeninterne Unterlagen können klassifiziert und entsprechende E-Mails erkannt werden.

Features

- Bei der Klassifikation von SPAM erreicht CORE Genauigkeiten von über 95%
- Weniger als 0,1% False Positives
- Robust gegen Overfitting
- Bestes statistisches Verfahren zur Textklassifizierung der Reuters-Textsammlung mit 86% Genauigkeit über alle 118 Kategorien und 92% über die größten 10 Kategorien
- Minimierte Anzahl der Klassifizierungsfehler (Structural Risk Minimization)
- Sehr gute Performanz, da effiziente Algorithmen zur Lösung des Optimierungsproblems existieren
- Multikategorisierung möglich
- Zugrunde liegender SVM-Algorithmus ist frei verfügbar (Open Source) und wird permanent weiterentwickelt und optimiert.
- Bestehende iQ.Suite Wall Jobs als Basis für die Prüfung
- Neue iQ.Suite Wall Jobs als Basis für den Lernvorgang
- Neuer Analyzer für die Prüfung der Dokumente
- In den Analyzer integrierter Trainer für den Lernvorgang
- Dokumentencontainer mit Copy & Paste-Möglichkeiten für den Lernvorgang

Über GBS

GROUP Business Software ist führender Anbieter von Lösungen und Services in den Bereichen Messaging Security und Workflow für die IBM und Microsoft Collaboration Plattformen. Weltweit vertrauen mehr als 5.000 Kunden und 4 Millionen Anwender auf die Expertise von GBS. Der Konzern ist in Europa, Nordamerika sowie Asien tätig.

Weitere Informationen unter www.gbs.com