Arbeitspapiere der FOM

Klumpp, Matthias / Marner, Torsten / Sandhaus, Gregor (Hrsg.)

ild Schriftenreihe Logistikforschung
Band 35

Maschinelles Lernen zur Erkennung von SMS-Spam

Keinhörster, Mark / Sandhaus, Gregor

FOM Hochschule
ild
Institut für Logistik- & Dienstleistungsmanagement
der FOM University of Applied Sciences

Keinhörster, Mark / Sandhaus, Gregor

Maschinelles Lernen zur Erkennung von SMS-Spam

FOM Hochschule
ild Institut für Logistik- & Dienstleistungsmanagement
Schriftenreihe Logistikforschung

Band 35, August 2013

ISSN 1866-0304

Essen

Die Autoren danken *Christian Witte* für Korrekturhinweise zu dieser Publikation.

Inhalt

Abkürzungsverzeichnis

BSI	Bundesamt für Sicherheit in der Informationstechnik
GSM	Global System for Mobile Communications
ISP	Internet Service Provider
KNN	Künstliches Neuronales Netz
MAPS	Mail-Abuse Prevention System
MLP	Multi-Layer Perceptron
OECD	Organisation for Economic Co-operation and Development
RBL	Realtime Blackhole List
ROC	Receiver Operating Characteristics
SLP	Single-Layer Perceptron
SMS	Short-Message-Service
SMSC	Short Message Service Center
UDH	User-Data-Header
URL	Uniform Resource Locator

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

For more than 2 decades the popularity of text messages has been continuously increasing. At the same time the abuse of short message services (SMS) in terms of SMS-Spam is also increasing and leads to high costs and data security issues for mobile users as well as for service providers.

In this working paper the concepts of machine learning are elaborated to identify automatically possible SMS-Spam and therefore improve the service quality of telecommunication companies.

The recognition of SPAM is hereby based upon the classification of text messages with the help of the

a) Bayes' theorem by fragmenting the text message into n-grams and to determine the likelihood for SPAM. Deploying vertical n-grams is hereby a new approach.
b) Multilayer perceptron (neural networks) and the Backpropagation learning algorithm.

The ROC curve (receiver operation characteristic) and corresponding measures are used to compare and evaluate both methods.

1 Einführung

Ursprünglich als reines Signal für den Verbindungsaufbau ins GSM-Netz gedacht, wurde am 3. Dezember 1992 die erste Kurznachricht mit dem Inhalt ‚Merry Christmas' versendet. Schnell wurde das Potenzial für persönliche Textnachrichten erkannt. Der dadurch entstandene Kurznachrichtendienst (Short-Message-Service) erfährt bis heute enormen Zuwachs.[1]

Doch dieser explosionsartige Gewinn an Popularität[2] bringt auch negative Effekte mit sich. Ein Beispiel ist Korea, wo bereits im Jahr 2003 die Menge der ungewollten Kurznachrichten die der Spam-Mails überstieg.[3] Aber auch Indien ist von diesen Effekten betroffen. Das dortige Mobilfunknetz wird täglich mit über 100 Millionen ungewollter SMS belastet.[4]

In der westlichen Welt ist dieses Problem bisher weniger weit verbreitet, da das Versenden von Kurznachrichten, im Gegensatz zu einem Medium wie der E-Mail, noch zu teuer ist.[5] Doch auch dort sinken die Kosten für das Versenden von Kurznachrichten mit der Folge, dass die Profite mit Spamnachrichten steigen.[6][7] Damit wird auch in Europa der Short-Message-Service zu einem profitablen Medium für Nachrichten dieser Art.[8] Bestätigt wird dies unter anderem durch die Sicherheitsfirma Sophos, die bereits im Jahr 2008 einen hohen Zuwachs an Spam per SMS in Europa verzeichnete.[9]

Da Mobiltelefone und Smartphones in der heutigen Zeit kaum mehr wegzudenken sind, resultieren aus dieser Entwicklung nicht nur Folgen seitens der Netzbetreiber, wie Überlastungen und Ausfälle,[10] sondern auch Sicherheitsrisiken und Kostenfallen auf der Nutzerseite.[11] Trojaner wie ‚Android.Pikspam' und ‚SpamSoldier' tarnen sich beispielsweise als Android-Applikation, um anschließend Nachrichten direkt vom Mobiltelefon des Opfers zu versenden.[12] Zusätzlich zu den auftretenden Kosten für den Nutzer, wird auch die Erkennung dieser Art des Spamversands erschwert, da er nicht mehr auf Grundlage der Telefonnummer geblockt werden kann. Sollte sich der Netzbe-

[1] Vgl. computerwelt.at (2012), 25. Jun. 2013.
[2] Vgl. Yadav, K. et al. (2011).
[3] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[4] Vgl. Yadav, K. et al. (2011).
[5] Vgl. Gómez Hidalgo, J.M. et al.. (2006).
[6] Vgl. cloudmark.com (2012), 03. Dez. 2013.
[7] Vgl. Gsma.com (2011), 03. Dez. 2013.
[8] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[9] Vgl. sophos.com (2007), 12. Jun. 2013.
[10] Vgl. Yadav, K. et al. (2011).
[11] Vgl. Almeida, T., Hidalgo, J.M.G., Silva, T.P. (2013).
[12] Vgl. blog.lookout.com (2012), 24. Jul. 2013; symantec.com (2012), 24. Jul. 2013.

treiber aufgrund des hohen Spamaufkommens trotzdem dazu entscheiden, wäre der Nutzer nicht mehr in der Lage SMS zu senden oder zu empfangen.[13]

1.1 Problemstellung und Zielsetzung

Mit der zuvor erläuterten Entwicklung geht die Problemstellung dieser Arbeit hervor. Denn trotz der steigenden Tendenz gibt es bisher nur wenige erfolgreiche Unternehmungen, den Versand von Spam über Kurznachrichten einzudämmen.[14]

Daher ist das Ziel dieses Arbeitspapieres die Entwicklung und Bewertung neuer Verfahren zur Erkennung von SMS-Spam. Als Grundlage dieser Verfahren dienen der bereits bei Emails etablierte Bayes-Klassifikator sowie das in diesem Bereich weniger weit verbreitete neuronale Netz.

1.2 Vorgehensweise

Um die zuvor definierten Ziele zu erreichen, werden zunächst die grundlegenden Charakteristika von SMS-Spam erläutert, indem der Begriff definiert und abgegrenzt wird. In diesem Zuge erfolgt auch eine Erläuterung des technischen Aufbaus sowie der textuellen Merkmale von Kurznachrichten. Im Anschluss daran wird die ökonomische Relevanz für die Erkennung von SMS-Spam seitens der Netzbetreiber sowie auch der Nutzer erarbeitet. Darauf folgt ein rudimentärer Überblick über bisherige Verfahren im Bereich der Spamerkennung.

Nach der Erläuterung der grundlegenden Eigenschaften von SMS-Spam wird der Fokus auf die verwendeten Verfahren zur Erreichung der Zielsetzung gelegt. Dazu wird zu Anfang das Konzept des maschinellen Lernens erläutert und das Problem der Spamerkennung als Klassifikationsproblem in dieses Konzept eingeordnet. Anschließend werden die theoretischen Grundlagen der verwendeten Ansätze zur Erkennung von Spam thematisiert, angefangen bei dem statistischen Ansatz nach Bayes. Im Anschluss daran wird das Konzept des mehrlagigen Perzeptrons erläutert.

Auf der Grundlage der theoretischen Basis werden im nächsten Schritt die beiden Verfahren auf die Erkennung von SMS-Spam angewandt. Dazu werden zunächst die Phasen des inhaltsbasierten Lernens erläutert. Aufbauend auf diesen Phasen wird dann der bayessche Klassifikationsansatz konkretisiert. In diesem Schritt werden die einzel-

[13] Vgl. Computerworld.com (2012), 28. Aug. 2013.
[14] Vgl. Sohn, D.-N., Lee, J.-T., Rim, H.-C. (2009); Yadav, K. et al. (2011).

nen Phasen durchlaufen und es wird aufgezeigt, auf welche Art und Weise inhaltliche Merkmale extrahiert, zahlenmäßig dargestellt und gelernt werden, um darauffolgend deren statistische Verrechnung zu erläutern. Entsprechend wird im Anschluss daran mit dem Konzept des mehrlagigen Perzeptrons verfahren. Dabei wird zu Beginn die Zerlegung von Nachrichten in Messgrößen thematisiert, die durch das Perzeptron verarbeitet werden können. Anschließend wird auf Grundlage dieser Größen eine geeignete Netzarchitektur festgelegt, die für die Klassifikation verwendet wird.

Im Anschluss an die Konkretisierung der Verfahren erfolgt der Evaluierungsteil. Innerhalb dieses Teils werden zunächst geeignete Kennzahlen festgelegt, anhand derer sich die Qualität der einzelnen Verfahren messen lässt. Daraufhin wird das Konzept der ROC Kurve erläutert und aufgezeigt, wie es zur Feinabstimmung der Klassifikatoren eingesetzt werden kann. Anschließend werden die einzelnen Verfahren evaluiert und deren Ergebnisse in Form von Kennzahlen festgehalten.

Abschließend werden die wichtigsten Ergebnisse dieser Ausarbeitung zusammengefasst und ein Ausblick auf zukünftige Trends gegeben.

1.3 Zielgruppe und Rahmenbedingungen

Im Folgenden werden die Zielgruppe sowie die Rahmenbedingungen dieser Arbeit festgelegt. Während der Abschnitt zur Zielgruppe einen potenziellen Leserkreis aufzeigt, werden im Abschnitt der Rahmenbedingungen Restriktionen erläutert, denen die Entwicklung sowie Bewertung der Erkennungsverfahren unterliegen.

Zielgruppe

Das Arbeitspapier richtet sich als erste Zielgruppe insbesondere an Entwickler von Verfahren zur Spamerkennung. Da die Verfahren in dieser Arbeit jedoch auf jede Art von Texten angewendet werden können, lässt sich die Zielgruppe allgemein auf die Entwickler von Werkzeugen zur Textklassifikation ausweiten.

Als zweite Zielgruppe richtet sich die Arbeit an Personen mit einem mathematisch-technischen Hintergrund und Interesse an den Konzepten des maschinellen Lernens. Aufgrund der mathematischen Ansätze sind Vorkenntnisse im Bereich der Statistik notwendig.

Rahmenbedingungen

Die Rahmenbedingungen dieses Arbeitspapieres ergeben sich aus den Nachrichten, anhand derer die Verfahren trainiert und getestet werden. Diese Sammlung der Trainings- und Testnachrichten wird auch ‚Korpus' genannt. Innerhalb dieses Arbeitspapie-

res wurden die Datensätze der „SMS Spam Collection v0.1“[15] verwendet, die aus 5.574[16] Kurznachrichten besteht.

Damit die maschinellen Lernverfahren ausreichend trainiert werden können, werden zu Beginn 300 zufällig gewählte Nachrichten einer jeden Klasse als Trainingskorpus festgelegt. Die restlichen Nachrichten dienen der Evaluation.

Des Weiteren bestehen die SMS lediglich aus englischsprachigen Inhalten ohne Kopfdaten oder sonstigen technischen Informationen. Aus diesem Grund ist die Sprache, anhand derer die Erkennungsverfahren entwickelt, trainiert und getestet werden, Englisch.

[15] Vgl. dt.fee.unicamp.br (o.J.), 03. Dez. 2013.
[16] 4827 Ham-Nachrichten und 747 Spam-Nachrichten.

2 Grundlagen zur Erkennung von SMS-Spam

Im Folgenden Kapitel werden die grundlegenden Eigenschaften von SMS-Spam erläutert. Dabei wird zunächst Spam als Begriff definiert und abgegrenzt. Anschließend wird auf dessen historischen Hintergrund eingegangen und bisherige Verfahren zu dessen Erkennung vorgestellt, um abschließend die ökonomische Relevanz der Spamerkennung.

2.1 Begriffsdefinition

Für den Begriff Spam gibt es in der Literatur keine einheitliche Definition.[17] Während CROSS ET AL. sowie auch das Bundesamt für Sicherheit in der Informationstechnik (BSI) den Begriff als unaufgeforderte Massen-E-Mails definieren,[18][19] bietet die Organisation for Economic Co-operation and Development (OECD) durch Aufteilung der Spezifika in primäre und sekundäre Eigenschaften eine genauere Definition. Die primären Eigenschaften gelten übergreifend für alle Spam-Nachrichten, die sekundären hingegen können, müssen jedoch nicht zutreffen. Weiterhin löst sich die Definition der OECD von der E-Mail als Trägermedium und macht sie auch für den Short-Message-Service (SMS) gültig.[20] Dem Begriff Spam steht die Bezeichnung Ham als vom Nutzer erwünschte Nachrichten gegenüber.[21]

Primäre Merkmale

Ein primäres Merkmal von Spam ist der elektronische Versand. Dabei unterscheidet die OECD nicht zwischen unterschiedlichen Technologien. Die Definition gilt für E-Mails gleichermaßen wie auch für Nachrichten per SMS oder Instant Messenger. Des Weiteren werden Spam-Nachrichten ohne Aufforderung und Einwilligung an eine große Empfängerzahl versendet. Zusätzlich ist der Versand von Spam kommerzieller Natur, mit der Absicht Gewinne zu erzielen. Dabei kann es sich um Werbung handeln oder auch um politische Themen oder schadhaften Code.[22]

Sekundäre Merkmale

Zu den sekundären Merkmalen der OECD zählt unter anderem die Verwendung von Zieladressen, die ohne Zustimmung des Empfängers ermittelt werden sowie die Unmöglichkeit, den Versand der Nachricht nach Bekanntgabe der Empfängeradresse zu

[17] Vgl. Zdzdiarski, J.A. (2005).
[18] Vgl. Cross, F.B., Miller, R.L. (2011).
[19] Vgl. Topf, J. et al. (2005).
[20] Vgl. OECD (2004).
[21] Vgl. Topf, J. et al. (2005).
[22] Vgl. OECD (2004).

unterbinden. Aber auch sich wiederholende oder nur leicht abgewandelte Inhalte zählen dazu. Dabei ist der Inhalt nicht an bestimmte Zielpersonen gerichtet, sondern spricht eine breite Masse an Adressaten an. Zusätzlich sind die Sender anonym oder täuschen eine falsche Identität vor. Letztes sekundäres Merkmal sind illegale oder anstößige Inhalte wie beispielsweise betrügerische Geschäftsabsichten oder Pornografie.[23]

2.2 Aufbau und Merkmale von Kurznachrichten

Das folgende Kapitel gibt einen kurzen Überblick über die technischen sowie textuellen Merkmale von Kurznachrichten. Dabei wird zu Beginn der Short-Message-Service erläutert, über den der Versand der Nachrichten erfolgt. Anschließend wird der Fokus auf den Aufbau und den daraus folgenden textuellen Merkmalen gelegt.

Arbeitsweise des Short-Message-Services

Der Short-Message-Service arbeitet nach dem Store-and-Forward-Verfahren. Das bedeutet die Nachrichten werden zunächst vom Absender an einen Intermediär im Netzwerk des Mobilfunkoperators gesendet, dem Short Message Service Center (SMSC). Nach dem Erhalt leitet das SMSC sie an ein weiteres Zentrum oder den Empfänger weiter.[24]

Aufbau von Kurznachrichten

Kurznachrichten existieren in zwei Formaten, dem ‚SMS-SUBMIT' sowie dem ‚SMSDELIVER'.[25] Ersteres wird für Nachrichten verwendet, die vom Mobiltelefon zum SMSC gesendet werden. SMS-DELIVER wiederum dient dem Versand vom Intermediär zum eigentlichen Empfänger.[26] Da die in diesem Arbeitspapier entwickelten Verfahren lediglich mit bereits empfangenen Nachrichten arbeiten, wird im weiteren Verlauf der Fokus auf den Aufbau des SMS-DELIVER Formats und die für dieses Arbeitspapier relevanten Felder gelegt.

[23] Vgl. OECD (2004).
[24] Vgl. Mulliner, C., Miller, C. (2009).
[25] Vgl. European Telecommunications Standards Institute (1998).
[26] Vgl. Rafique, M.Z., Farooq, M. (2010).

Aufbau des DELIVER-Formats

Das Deliver-Format ist in zwei Blocks untergliedert, dem Header, der Standardinformationen beinhaltet,[27] sowie dem Payload der den textuellen Inhalt der Nachricht enthält.[28] Die beiden Bestandteile werden nachfolgend genauer betrachtet.

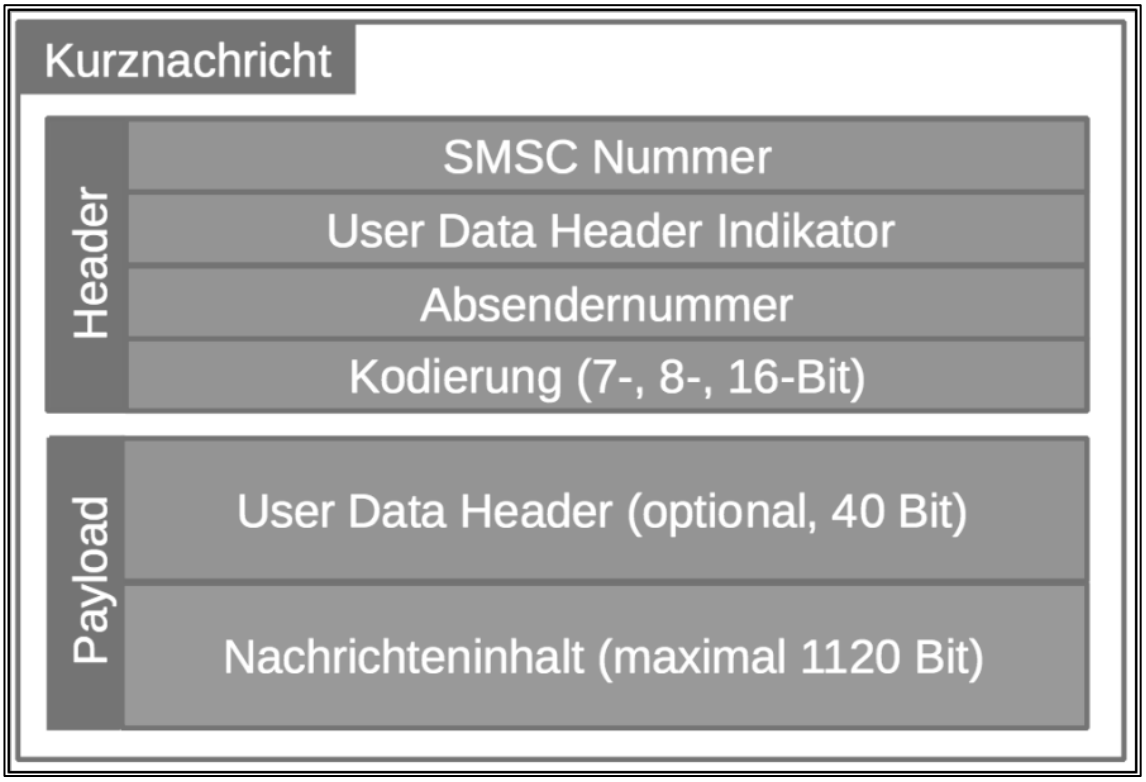
Zusammensetzung des Headers

Der Header besteht unter anderem aus der Nummer des SMSC, einem Flag, das als Indikator für den User-Data-Header (UDH) dient, der Absendernummer sowie der Codierung.[29] Die Codierung gibt die Anzahl der Bits an, die je Zeichen im Payload verwendet werden. Standardmäßig sind die Nachrichtentexte im 7-Bit-Format codiert. Bei der maximalen Größe der Nutzdaten von 1120 Bit ergibt sich somit eine mögliche Textlänge von 160 Zeichen. Das 16-Bit-Format hingegen verwendet den Unicode Zeichensatz und reduziert die Länge auf 70 Zeichen.[30] Datennachrichten wie Bilder oder Musik codieren die Informationen mit 8-Bit.[31] Ist der Indikator für den UDH gesetzt, gehen für den Header noch einmal 40-Bit vom möglichen Textinhalt ab. Er wird verwendet um zusätzliche Verarbeitungsinformationen für den Payload bereitzustellen, wie beispielsweise der Verknüpfung aufgeteilter Nachrichten die mehr als 160 Zeichen enthalten.[32]

Abbildung 1 fasst die für dieses Arbeitspapier relevanten Komponenten des SMS-DELIVER Formats zusammen.

[27] Vgl. Kolcz, A., Chowdhury, A., Alspector, J. (2004).
[28] Vgl. Rafique, M.Z., Farooq, M. (2010).
[29] Vgl. Rafique, M.Z., Farooq, M. (2010).
[30] Vgl. Mahmoud, T.M., Mahfouz, A.M. (2012).
[31] Vgl. Rafique, M.Z., Farooq, M. (2010).
[32] Vgl. Mulliner, C., Miller, C. (2009).

Abbildung 1: Aufbau von Kurznachrichten im SMS-Deliver Format



Quelle: In Anlehnung an Rafique, M.Z., Farooq, M. (2010) und Kolcz, A., Chowdhury, A., Alspector, J. (2004).

Da der verwendete SMS-Korpus nur den Inhalt der Nachrichten enthält, werden in dieses Arbeitspapier keine Headerinformationen extrahiert und verwertet.

Textuelle Merkmale des Payload

Es existiert bereits eine große Vielfalt an Verfahren, die auf die inhaltliche Erkennung von E-Mail Spam spezialisiert sind. Diese Verfahren können prinzipiell auch für die Erkennung von SMS-Spam eingesetzt werden. Trotzdem es sich in beiden Fällen um die Verarbeitung textueller Inhalte handelt, kann jedoch nicht sichergestellt werden, dass die einfache Anwendung bereits bestehender und erprobter Verfahren für E-Mails auch zu den gleichen, positiven Resultaten bei SMS-Nachrichten führt.[33]

Diese Problematik resultiert aus der Längenbegrenzung von Kurznachrichten auf 160 Zeichen. Sie impliziert auf der einen Seite weniger extrahierbare Merkmale für die Verfahren,[34] auf der anderen Seite jedoch auch sprachliche Eigenheiten, die E-Mails von

[33] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[34] Vgl. Yadav, K. et al. (2011).

SMS unterscheiden.[35] Die Unterschiede äußern sich dabei oftmals in einer hohen Anzahl von mehrdeutigen Sprachkonstrukten, wie beispielsweise Emoticons, Abkürzungen oder Akronymen, die in kurzen SMS-Texten deutlich häufiger auftreten als es in E-Mails der Fall ist.[36]

2.3 Ökonomische Relevanz

Kurznachrichten sind mittlerweile allgegenwärtig und zählen aufgrund ihrer Einfachheit in der Benutzung und den hohen Erträgen für die Service Provider zu den populärsten Kommunikationsmitteln mobiler Endgeräte. So ist der Kurznachrichtendienst bereits in Bereichen wie dem mobile Banking oder der Benutzerauthentifizierung weit verbreitet.[37] Mit dem ansteigenden Nutzungsniveau ist jedoch auch ein enormer Anstieg von SMS-Spam zu verzeichnen.[3839] Dessen Folgen für die Netzbetreiber und Nutzer werden im weiteren Verlauf dieses Kapitels thematisiert, um die Notwendigkeit geeigneter Filterungsverfahren aufzuzeigen.

Relevanz für die Netzbetreiber

Aufgrund der großen Mengen von Spam-Nachrichten, die versendet werden, kann seitens des Mobilfunkbetreibers ein sehr hohes Trafficvolumen aufkommen, welches das Mobilfunknetz überlastet. Die Folgen einer solchen Überlastung reichen von Verzögerungen in der Nachrichtenzustellung bis hin zu Netzausfällen.[40] Aufgrund dessen entstehen höhere Netzinfrastruktur- und Betriebskosten, denen Provider entgegenwirken müssen.[41] Mit einem Spam-Anteil von 20 bis 30% am gesamten SMS-Datenverkehr dienen China und Indien als Paradebeispiele dieses immer stärker wachsenden Problems.[42] So wurden im Jahr 2008 innerhalb einer Woche in China 200 Milliarden Spam-Nachrichten zugestellt.[43]

Neben den finanziellen Aspekten hat SMS-Spam jedoch auch negative Auswirkungen auf den Ruf des Betreibers. Die unerwünschten Nachrichten senken das Vertrauen in den Dienstleister und schaden somit dem Markenimage.[4445]

---

[35] Vgl. Delany, S.J., Buckley, M., Greene, D. (2012).
[36] Vgl. Delany, S.J., Buckley, M., Greene, D. (2012).
[37] Vgl. Rafique, M.Z., Farooq, M. (2010).
[38] Vgl. Mahmoud, T.M., Mahfouz, A.M. (2012).
[39] Vgl. Rafique, M.Z., Farooq, M. (2010).
[40] Vgl. Yadav, K. et al. (2011).
[41] Vgl. Delany, S.J., Buckley, M., Greene, D. (2012).
[42] Vgl. gsma.com (2011), 03. Dez. 2013.
[43] Vgl. Delany, S.J., Buckley, M., Greene, D. (2012).
[44] Vgl. Xu, Q. et al. (2012).

Relevanz für die Nutzer

Doch SMS-Spam hat nicht nur Folgen für Mobilfunkbetreiber, sondern auch für die Endnutzer. Dazu zählen zum einen finanzielle und sicherheitsrelevante Folgen, zum anderen jedoch auch Behinderungen bei der alltäglichen Nutzung des Mobiltelefons.

Sicherheitsrelevante Aspekte

Das hohe Vertrauen der Nutzer in ihre Mobiltelefone führt zu einem leichtfertigeren Umgang mit Spam-SMS[46] im Hinblick auf die Nutzung kritischer Dienste wie Authentifizierung oder mobile Banking.[47] Dadurch entstehen nicht nur finanzielle Risiken, wie ungewollte Anrufe bei kostenpflichtigen Hotlines und Abschlüsse teurer Abonnements, sondern auch sicherheitsrelevante Bedrohungen wie Attacken durch Phishing und Malware-Programme.[48] In bestimmten Ländern wird außerdem auch der Empfang von Kurznachrichten abgerechnet, dies stellt einen weiteren Kostenpunkt für den Nutzer dar, der durch SMS-Spam hervorgerufen wird.[49]

Nutzungsrelevante Aspekte

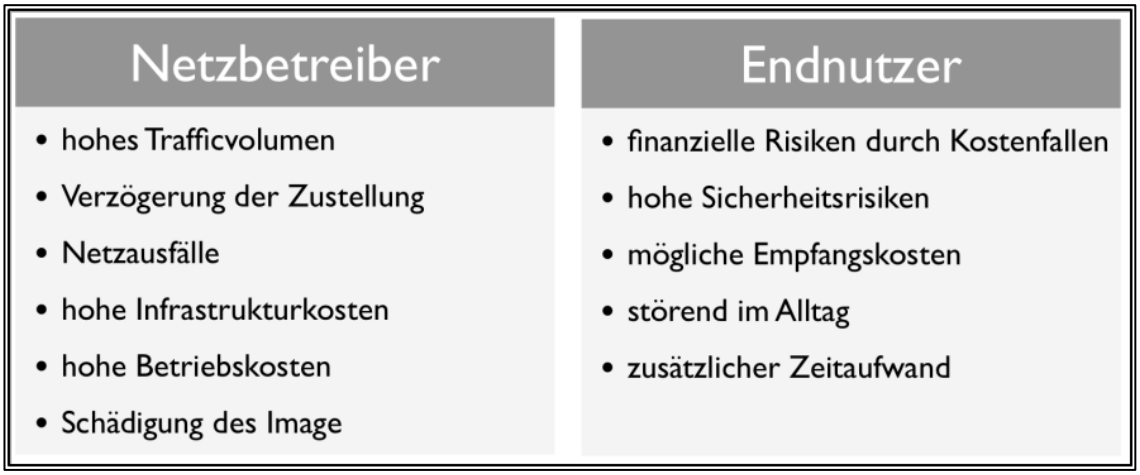
Neben finanziellen und sicherheitsrelevanten Risiken entstehen jedoch auch Behinderungen in der alltäglichen Nutzung der Mobiltelefone. So werden eingehende Kurznachrichten im Gegensatz zur E-Mail häufig durch Empfangstöne und Hinweise bestätigt, die die Aufmerksamkeit auf sich ziehen, auch wenn sie durch ihren Spam-Charakter für den Leser nicht relevant sind. Außerdem mangelt es häufig an geeigneten Benutzerschnittstellen in SMS-Applikationen, um die unerwünschten Kurznachrichten ohne viel Aufwand zu löschen. Dies führt zu einem beträchtlichen Aufwand, gerade bei günstigen Mobiltelefonen, denn oft müssen die sie einzeln gelesen und gelöscht werden.[50]

Abbildung 2 fasst noch einmal die in diesem Kapitel erarbeiteten Folgen von SMS-Spam für Netzbetreiber und Endkunden zusammen.

---

[45] Vgl. Delany, S.J., Buckley, M., Greene, D. (2012).
[46] Vgl. Almeida, T., Hidalgo, J.M.G., Silva, T.P. (2013).
[47] Vgl. gsma.com (2011), 03. Dez. 2013.
[48] Vgl. Xu, Q. et al. (2012).
[49] Vgl. Almeida, T., Hidalgo, J.M.G., Silva, T.P. (2013).
[50] Vgl. Juanid, M.B., Farooq, M. (2011).

Abbildung 2: Folgen von Spam für Netzbetreiber und Endnutzer



Quelle: Eigene Darstellung.

2.4 Klassische Erkennungskonzepte

Das folgende Kapitel stellt die wichtigsten klassischen Verfahren zur Erkennung von Spam vor, die nicht auf maschinellen Lernverfahren basieren. Zu diesen Verfahren zählen:

- Primitive Sprachanalyse
- Signaturbasierte Filter
- White- und Blacklisting
- Heuristische Filter

Dabei wird nicht zwischen den Übertragungsarten wie SMS oder E-Mail unterschieden, sondern vielmehr ein Überblick über grundsätzliche Erkennungsmethoden gegeben.

2.4.1 Primitive Sprachanalyse

Die primitive Sprachanalyse beschreibt den Einzelvergleich spezifischer Textbausteine mit den Inhalten der zu prüfenden Nachricht. Enthält die Nachricht einen der definierten Indikatoren für Spam, wird sie verworfen. Dabei wurden nicht nur inhaltlich typische Textteile abgeglichen, sondern auch definierte Absenderadressen.[51]

[51] Vgl. Zdzdiarski, J.A. (2005).

Mitte der 90er Jahre, als Spam noch nicht die Ausmaße annahm, wie es in der heutigen Zeit der Fall ist, war es möglich auf diese Weise rund 80% der erhaltenen Spam-Nachrichten herauszufiltern.[52]

Der Vorteil dieses Verfahrens liegt in der Einfachheit und leichten Anpassbarkeit. Jedoch birgt dies gleichzeitig auch den größten Nachteil, da das Verfahren eine konsequente Wartung der Schlüsselwörter voraussetzt, um aktuell zu bleiben. Hinzukommt die hohe Rate falscher Positive.[53] Denn auch wenn die genutzten Schlüsselphrasen häufig für Spam sprechen, kann es passieren, dass sie trotzdem in legitimen Nachrichten vorkommen. Somit kann ein Schlüsselwort, das ausnahmsweise außerhalb des Spamkontexts verwendet wurde, zum vorschnellen Verwerfen einer eigentlich legitimen Nachricht führen.[54]

2.4.2 Signaturbasierte Filter

Signaturbasierte Filter zur Erkennung von Spam sind eine Alternative zu maschinellen Lernverfahren.[55] Das Verfahren generiert einen einzigartigen Hash-Wert (Signatur) für jede bekannte Spam-Nachricht. Während der Nutzung werden diese zuvor ermittelten Hashes mit denen der eingehenden Nachrichten verglichen und bei Übereinstimmung jeweils als Spam klassifiziert.

Diese Technik macht es statistisch unmöglich, eine legitime Nachricht als Spam zu klassifizieren, wodurch sie eine sehr niedrige Falsch-Positiv-Rate zum Vorteil hat.[56] Jedoch haben einfache signaturbasierte Filter den Nachteil, dass kleinste Änderungen wie ein eingefügtes Leerzeichen den Hash-Wert der Spam-Nachricht ändern. Dies hat zur Folge, dass vorhandene Nachrichten nicht mehr erkannt werden. Aus diesem Grund arbeiten signaturbasierte Filter am effektivsten bei identischen Nachrichten, die in großen Mengen an eine hohe Zahl von Adressaten verschickt werden.[57]

[52] Vgl. Zelkowitz, M. (2011).
[53] Eine Nachricht wird als Spam klassifiziert obwohl sie kein Spam ist.
[54] Vgl. Zdzdiarski, J.A. (2005).
[55] Vgl. Koppel, M., Schler, J., Argamon, S. (2009).
[56] Vgl. Carpinter, J., Hunt, R. (2006).
[57] Vgl. Koppel, M., Schler, J., Argamon, S. (2009).

3 Maschinelles Lernen als Grundlage der Spamerkennung

Das folgende Kapitel erläutert das Konzept sowie die Eigenschaften des maschinellen Lernens. Dabei wird zunächst der Begriff definiert und dessen unterschiedliche Artenvorgestellt. Im nächsten Schritt wird die Klassifikation als ein Teilbereich thematisiert und in diesem Zuge der Grundaufbau von Klassifikatoren erläutert. Abschließend wird der Fokus auf Methoden zur Merkmalsselektion sowie -extraktion gelegt, um eine ganzheitliche Basis für die praxisorientierten Kapitel zu schaffen.

3.1 Definition und Konzepte

Durch die stetige Weiterentwicklung der Computertechnologie entstehen immer mehr Möglichkeiten zur Speicherung und Verarbeitung großer Datenmengen. Supermärkte mit landesweiten Filialen sind ein gutes Beispiel dafür. Jeder getätigte Verkauf wird mit seinen gesamten Details von den Kassen registriert und persistiert. Diese Daten nützen jedoch nur, wenn sie auch analysiert und in brauchbare Informationen umgewandelt werden. Durch diese Analyse soll ein Prozess hergeleitet werden, der bestimmte Beobachtungen in den erhobenen Daten erklärt. Dabei geht es nicht um die Klärung aller Details, sondern um die Annäherung an bestimmte Muster, wie zum Beispiel der Kauf von Eis im Sommer oder der Zusammenhang von Kartoffelchips und Bier.[58]

Genau in diesem Bereich setzt das maschinelle Lernen an. Es dient der Auffindung von Regelmäßigkeiten in großen Datenmengen, um zukunftsnahe Prognosen zu erstellen.[59] Eine genaue Definition des maschinellen Lernens liefert ALPAYDIN:

„Maschinelles Lernen heißt, Computer so zu programmieren, dass ein bestimmtes Leistungskriterium anhand von Beispieldaten oder Erfahrungswerten aus der Vergangenheit optimiert wird.“[60]

Um ein System nach den Anforderungen dieser Definition zu entwickeln, wird ein Modell mit spezifischen Parametern definiert. Der Lerneffekt wird durch stetige Anpassung der Parameter mithilfe eines Trainingsprogramms hervorgerufen, das mit eigens dafür zusammengestellten Trainingsdaten arbeitet.

Grundsätzlich untergliedert sich maschinelles Lernen in folgende Konzepte:[61]

- überwachtes Lernen
- unüberwachtes Lernen

[58] Vgl. Alpaydin, E. (2008).
[59] Vgl. Segaran, T. (2008).
[60] Alpaydin, E. (2008).
[61] Vgl. Alpaydin, E. (2008).

- bestärkendes Lernen

Da die einzelnen Konzepte unterschiedliche Ziele verfolgen, werden sie nachfolgend genauer erläutert.

Überwachtes Lernen

Überwachtes Lernen wird zur Erkennung von Musterpaaren auf Ein- und Ausgabeparametern verwendet. Dabei sind die korrekten Ausgaben, wie beispielsweise Spam oder Nicht-Spam, bekannt. Das Verfahren wird mithilfe von beispielhaften Eingabedaten sowie deren erwartete Ausgaben trainiert. Bei einer fehlerhaften Aussage über den Ausgabezustand werden die Parameter zum richtigen Zustand hin korrigiert. Dieser Prozess wird wiederholt, bis das Verfahren präzise Aussagen treffen kann.[62] Bekannte Vertreter für diese Art des maschinellen Lernens sind die Klassifikation und Regression.[63]

Unüberwachtes Lernen

Im Gegensatz zum überwachten Lernen wird beim unüberwachten Lernen das Ziel verfolgt, Regelmäßigkeiten in den Eingabeparametern zu erkennen, wobei die Ausgabewerte unbekannt sind. Dadurch sollen bestimmte Muster in der Eingabe erkannt werden, die häufiger auftreten als andere. In der Statistik wird ein solches Verfahren auch ‚Dichteschätzung' genannt.[64]

Bestärkendes Lernen

Bestärkendes Lernen bildet die Eingaben nicht auf Zustände, sondern auf Aktionen ab. Relevant sind dabei deren korrekte und zielführende Abfolgen (Sequenz, Taktik), nicht die Aktionen im Einzelnen. Sie sind dabei erst dann als gut zu bewerten, wenn sie auch Teil einer guten Sequenz sind. Ein solches Verfahren sollte damit in der Lage sein, von früheren Sequenzen zu lernen und auf dieser Basis die Effizienz neuer abzuschätzen oder sie gar zu generieren.[65]

---

[62] Vgl. Haykin, S. (2009).
[63] Vgl. Alpaydin, E. (2008).
[64] Vgl. Alpaydin, E. (2008).
[65] Vgl. Alpaydin, E. (2008).

3.2 Klassifikation als Ansatz zur Spamerkennung

Die Klassifikation ist ein Teilbereich des überwachten Lernens. Sie steht für die mathematische Zuordnung einer Eingabemenge anhand ihrer Merkmale in eine definierte Ausgabekategorie. Diese Kategorien werden auch Klassen genannt. Sie sind von Anfang an bekannt, womit auch die Einordnung in das überwachte Lernen begründet wird.[66] Bei der Spamerkennung existieren zwei Klassen, Spam und Nicht-Spam. Die Eingabeparameter sind die spezifischen Merkmale der zu klassifizierenden Kurznachricht.

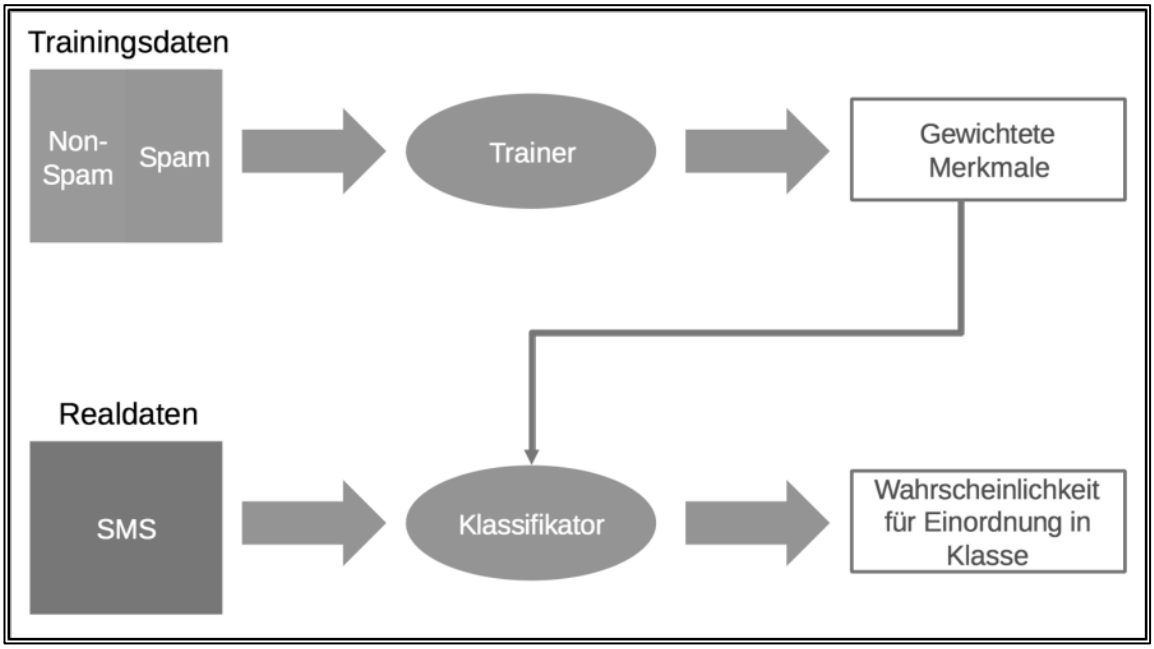
Der Klassifikationsprozess

Um diese Aufgabe zu erfüllen, werden Klassifikatoren entwickelt, die Muster aus den Parametern der Beispieldaten lernen und auf deren Basis neue Eingabeparameter ihren adäquaten Klassen zuordnen. Dieser Prozess besteht aus zwei Phasen, einer Trainings- und einer Klassifikationsphase. Damit der Klassifikator korrekt arbeitet, wird in der Trainingsphase das Wissen des Klassifikators aufgebaut. In der Klassifikationsphase wird dieses Wissen genutzt um neue Daten zu klassifizieren.[67] Die in diesem Arbeitspapier verwendeten Klassifikationsverfahren sind der Klassifikationsansatz nach BAYES sowie die Klassifikation durch mehrlagige Perzeptronen[68].

Der Klassifikationsprozess zur Erkennung von SMS-Spam ist in Abbildung 3 dargestellt. Die im Training erlernten Attribute stellen als gewichtete Merkmale die Klassifikationsgrundlage dar. Da die Erkennung von Spam eine binäre Klassifikation ist, gibt es nur einen Ausgabeparameter: die Wahrscheinlichkeit für die Einordnung der geprüften SMS in die Kategorie Spam.

[66] Vgl. Alpaydin, E. (2008);Heaton, J (2008).
[67] Vgl. Fathi, M., Adly, N., Nagi, M.(2004).
[68] Form eines künstlichen neuronalen Netzes.

Abbildung 3: Klassifikationsprozess



Quelle: Eigene Darstellung.

Klassifikationsfehler durch Rauschen

Ein weiterer wichtiger Punkt im Hinblick auf Klassifikationsverfahren ist das ‚Rauschen'. Es beschreibt Anomalien in den Trainingsdaten, durch die eine Klasse schwerer zu erlernen ist.[69]

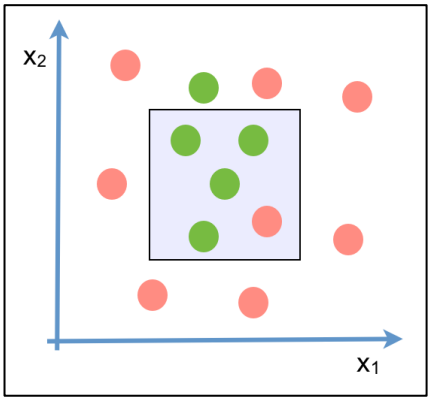
Dieser Effekt kann nach ALPAYDIN drei mögliche Ursachen haben:[70]

- ungenaue Trainingsdaten
- falsche Klassenzuordnung der Merkmale
- prägnante Merkmale die nicht beachtet werden

Abbildung 4 veranschaulicht den Rauscheffekt in einem Koordinatensystem. Es wird deutlich, dass eine eindeutige Abgrenzung der grünen von den roten Punkten durch ein Rechteck aufgrund des Rauschens nicht möglich ist. Somit wird entweder eine komplexere Form der Abgrenzung benötigt, oder es muss eine gewisse Fehlerrate eingeräumt werden.

[69] Vgl. Duda, R.O., Hart, P.E., Stork, D.G. (2012).
[70] Vgl. Alpaydin, E. (2008).

Abbildung 4: Rauschen in Daten



Quelle: In Anlehnung an Alpaydin, E. (2008).

In Bezug auf SMS-Spam äußert sich Rauschen beispielsweise in Wörtern, die in Spam sowie auch legitimen Nachrichten vorkommen können. Dieser Effekt wird von Spambetreibern häufig durch ‚out-of-context –Wörter' ausgenutzt. Das sind zusammenhangslose Wörter die häufig in legitimen Nachrichten vorkommen und in Spam injiziert werden, um Textklassifikatoren zu überlisten.[71]

3.3 Klassifikationsansatz nach Bayes

Um Klassifikatoren zu entwickeln, die in der Lage sind, aus gegebenen Daten Rückschlüsse zu ziehen, werden neben Methoden der Informatik, auch statistische Methoden benötigt.[72] Bayessche Klassifikatoren bedienen sich dabei insbesondere Methoden der Bayes-Statistik.[73] Dieser statistische Ansatz wird im Folgenden genauer erläutert und vom Ansatz der klassischen Statistik abgegrenzt. Im Anschluss daran wird der Satz von Bayes und dessen begriffliche Terminologien als Kernelemente der bayesschen Klassifikation vorgestellt.

3.3.1 Abgrenzung der Bayes-Statistik

Das nachfolgende Kapitel dient der Abgrenzung der Bayes-Statistik nach BAYES[74] von der klassischen nach FISCHER[75] sowie NEYMAN und PEARSON.[76]

[71] Vgl. Zdzdiarski, J.A. (2005).
[72] Vgl. Alpaydin, E. (2008).
[73] Vgl. Zdzdiarski, J.A. (2005).
[74] Vgl. Bayes, T. (1763).
[75] Vgl. Fischer, R.A. (1923); Fischer, R.A. (1925).
[76] Vgl. Neyman, J., Pearson, E.S. (1928).

Klassische Statistik

In der klassischen Statistik werden Hypothesen aufgestellt, die die Vorstellung eines bestimmten Sachverhalts repräsentieren.[77] Dabei ist es wichtig, dass die Aussage, unter der Beobachtung der ihr zugrunde liegenden Informationen, grundsätzlich wahr sein könnte. Um dies empirisch zu bestätigen, wird eine zufällige Stichprobe aus einer für die Hypothese relevanten Population gewählt, anhand derer ihr Wahrheitswert gemessen wird.[78]

Statistischer Ansatz nach Bayes

Im Gegensatz zur klassischen Statistik, werden beim Ansatz nach BAYES Vorinformationen in die Hypothesenbewertung mit einbezogen, die nicht auf die beobachtete Stichprobe zurückzuführen sind. Diese Informationen sind bereits vor der Bewertung verfügbar und werden als a-priori-Wahrscheinlichkeiten[79] in der Berechnung berücksichtigt.[80]

Der Unterschied zwischen dem klassischen sowie dem Statistikansatz nach Bayes ist noch einmal in Abbildung 5 zusammengefasst.

Abbildung 5: Unterschiede zwischen der klassischen und der Bayes-Statistik

Klassische Statistik	Bayes-Statistik
Aussagen allein auf Grundlage der Stichprobe.	Aussagen auf Grundlage der Stichprobe + Vorinformationen

Quelle: In Anlehnung an Strelec, H. (1989).

[77] Vgl. Alpaydin, E. (2008).
[78] Vgl. Wendt, D. (1983), Xu, C., Chen, Y., Chiew, K. (2010).
[79] Wahrscheinlichkeit für Klassenzugehörigkeit vor Auswertung.
[80] Vgl. Wendt, D. (1983), Xu, C., Chen, Y., Chiew, K. (2010); Gigerenzer, G. (2004).

3.3.2 Satz von Bayes zur Klassifikation

Die grundlegende Aufgabe eines Klassifikators ist die Schätzung der Wahrscheinlichkeiten für die Zugehörigkeit einer definierten Eingabe zu zuvor definierten Klassen. Bayes-Klassifikatoren lösen diese Aufgabe mithilfe des Satzes von Bayes.[81]

Um das Grundkonzept zur Lösung dieses Problems aufzuzeigen wird es nachfolgend theoretisch auf die Klassifikation von Spam angewendet. Dabei werden die Grundlagen und Definitionen aus der Literatur nach ALPAYDIN[82], RUNKLER[83], LIU[84], ZDZIARSKI[85] sowie BISHOP[86] zugrunde gelegt und auf den konkreten Fall angepasst. Die zuvor beschriebene Aufgabe lässt sich daraus ableitend in die folgende Problemstellung übersetzen:

Die Aufgabe eines Spam-Klassifikators ist es, die a-posteriori-Wahrscheinlichkeiten[87] *P* einer SMS *S* für die Klassen K_i aus der Menge der Klassen *K* zu schätzen. *K* besteht dabei aus den Klassen *Spam = 1* und *Ham = 0*. Die Schätzung basiert dabei auf dem Merkmalsvektor S_x, der die einzelnen Merkmale von *S* enthält.

Auf dieser Basis ergibt sich die in Gleichung (1) dargestellte mathematische Kurzform. $P(K = Spam \mid S_x)$ ist dabei die Wahrscheinlichkeit für das Auftreten einer SMS *S* mit dem Attributsvektor S_x unter der Bedingung dass die Nachricht der Klasse Spam angehört.

$$wähle \begin{cases} K = Spam\ falls\ P(K = Spam \mid S_x) > P(K = Ham \mid S_x) \\ K = Ham\ falls\ P(K = Spam \mid S_x) \leq P(K = Ham \mid S_x) \end{cases} \quad (1)$$

Die Besonderheit dieser Aufgabe liegt in der Abhängigkeit der Klassen aus *K* vom Attributsvektor S_x. An diesem Punkt wird der Satz von Bayes verwendet. Er ermöglicht die Kombination der a-priori-Wahrscheinlichkeit $P(K_i)$ mit der Wahrscheinlichkeit für das Auftreten der vorhandenen Daten $P(S_x)$. Die Definition des Bayes-Theorems ist in Gleichung (2) dargestellt. Dessen Komponenten $P(S_x)$ und $P(S_x \mid K_i)$ werden nachfolgend erläutert.

$$P(Ki \mid Sx) = \frac{P(Ki) * P(Sx \mid Ki)}{P(Sx)} \quad (2)$$

In diesem Arbeitspapier soll mithilfe des Satzes von Bayes die Wahrscheinlichkeit $P(K = Spam \mid S_x)$ ermittelt werden.

---

[81] Vgl. Liu, B. (2007).
[82] Vgl. Alpaydin, E. (2008).
[83] Vgl. Runkler, T.A. (2009).
[84] Vgl. Liu, B. (2007).
[85] Vgl. Zdzdiarski, J.A. (2005).
[86] Vgl. Bishop, C.M. (1995).
[87] Wahrscheinlichkeit für Klassenzugehörigkeit nach Auswertung.

Evidenz

Die Wahrscheinlichkeit für das klassenunabhängige Auftreten der beobachteten Daten $P(S_x)$ wird auch als Evidenz oder Randwahrscheinlichkeit bezeichnet. Ihre Zusammensetzung ist in (3) dargestellt.

$$P(S_x) = P(S_x \mid K = Spam) * P(K = Spam) + P(S_x \mid K = Ham) * P(K = Ham) \quad (3)$$

Klassen-Likelihood

$P(S_x \mid K_i)$ ist die bedingte Wahrscheinlichkeit für das Auftreten der beobachteten Daten aus dem Vektor S_x in der Klasse K_i. Diese Wahrscheinlichkeit wird als Klassen-Likelihood bezeichnet. Beispielsweise ist $P(S_x \mid K = Spam)$ die Wahrscheinlichkeit für das Auftreten der Attribute S_x in einer Spam-SMS.

Bayes-Klassifikator

Nach der Berechnung der a-posteriori-Wahrscheinlichkeiten $P(K_i \mid S_x)$ wählt der Klassifikator die Klasse mit der höchsten Wahrscheinlichkeit (4). Da es sich in diesem Arbeitspapier um die Erkennung von Spam handelt wird zwischen $K_0 = Ham$ oder $K_1 = Spam$ unterschieden.

$$wähle\ K_i\ wenn\ P(K_i \mid S_x) = \max_{N=2} P(K_n \mid S_x) \quad (4)$$

Diese Entscheidungsformel lässt sich in diesem Fall noch einmal vereinfachen. Aufgrund der Normalisierung durch die Evidenz innerhalb des Bayes-Theorems ergibt die Summe der Ergebnisse der beiden Klassen stets den Wert eins (5).

$$P(K_1 \mid S_x) + P(K_0 \mid S_X) = 1$$

Durch dieses Wissen kann die Entscheidungsformel aus Gleichung (4) dahingehend vereinfacht werden, als dass die Klasse gewählt wird, dessen Wahrscheinlichkeit größer 0,5 ist (6).

$$wähle\ K_i\ wenn\ P(K_i \mid S_x) > 0{,}5 \quad (6)$$

3.4 Klassifikation durch mehrlagige Perzeptronen

Innerhalb dieses Kapitels soll zunächst der grundlegende Aufbau und die Arbeitsweise künstlicher neuronaler Netze erläutert werden. Im Anschluss daran wird der Fokus auf das mehrlagige Perzeptron, als die in diesem Arbeitspapier verwendete Netzarchitektur, gelegt. Nach der Erläuterung der Grundbegriffe mehrlagiger Perzeptronen werden

dessen Komponenten sowie auch die Trainingsmöglichkeit durch Backpropagation betrachtet.

3.4.1 Aufbau und Funktionsweise

Künstliche neuronale Netze (KNN) sind vereinfachte, dem Gehirn nachempfundene Modelle,[88] die das zentrale Nervensystem abbilden.[89] Solche Netze bestehen aus eng miteinander verbundenen und parallel arbeitenden Prozessorelementen, den Neuronen, die Eingaben durch simpelste Entscheidungen an andere Neuronen weiterleiten.[90] Diese Verbindungen werden auch Synapsen genannt.[91] Aufgrund der extrem rudimentären Funktionsweise wird für komplexe Problemstellungen eine hohe Anzahl an Neuronen benötigt, die die einfachen Entscheidungen einzelner Elemente zu höherwertigen Entscheidungsprozessen kombinieren.[92] Obwohl eine Vielzahl verschiedener Arten von KNN für unterschiedlichste Problemstellungen existieren,[93] gibt es für sie bislang keine einheitliche Definition die als Oberbegriff dient.[94]

Rey und Wender präzisieren den Begriff des KNN daher anhand der drei übergeordneten Gemeinsamkeiten:

- Informationsaufnahme,
- Informationsverarbeitung und Netzmodifikation,
- sowie Informationsausgabe.

Diese Gemeinsamkeiten werden im Folgenden näher betrachtet.

Informationsaufnahme

Künstlichen neuronalen Netzen werden als Eingabeparamater wiederholt Informationen zur Verfügung gestellt, die einen zu verarbeitenden Realitätsausschnitt wiederspiegeln. Dargestellt wird dieser Ausschnitt in Zahlen, die diese Netze verarbeiten können.

Informationsverarbeitung und Netzmodifikation

Mithilfe der Eingangsparameter werden KNN modifiziert. Dies geschieht auf der Basis von Lernregeln, die den Eingangsvektor mit dem aktuellen Stand des Netzes, ebenfalls

---

[88] Vgl. Alpaydin, E. (2008); Zaun, D.P. (1999).
[89] Vgl. Patterson, D.W. (1997).
[90] Vgl. Heaton, J. (2008).
[91] Vgl. Alpaydin, E. (2008).
[92] Vgl. Heaton, J. (2008).
[93] Vgl. Segaran, T. (2008); Rey, G.D., Wender, K.F. (2010).
[94] Vgl. Patterson, D.W. (1997).

als Zahlenvektor dargestellt, kombiniert. Dieser Lernprozess ist iterativ und wird mehrfach durchgeführt. Während die zu verarbeitenden Informationen das Netz durchlaufen, werden sie modifiziert und am Ende als Ausgangsparameter ausgegeben. Die Umformung der Eingangsparameter geschieht auch während der späteren Nutzung, nachdem das Netz angelernt wurde.

Informationsausgabe

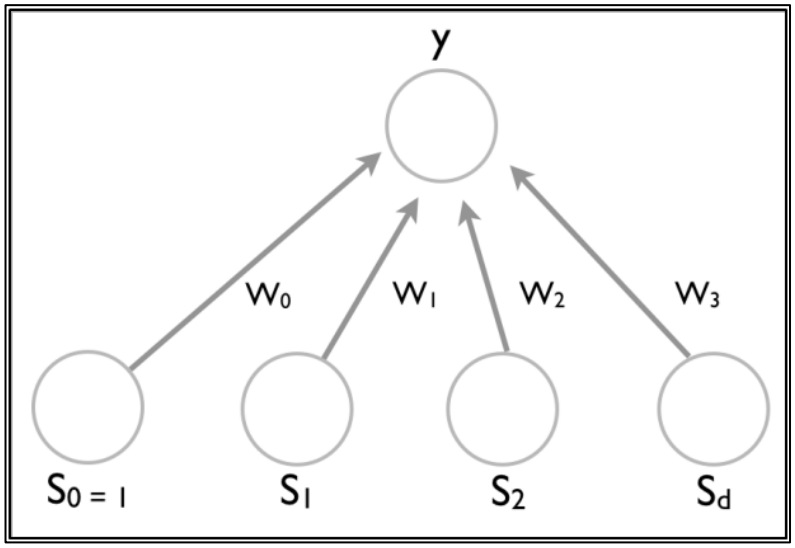
Die Ausgabeparameter sind das Ergebnis des Netzes nach Verarbeitung der Eingaben. In Bezug auf die Spamerkennung wäre die zugeordnete Klasse, derer die Kurznachricht angehört, eine mögliche Ausgabe für einen eingegebenen Attributsvektor einer SMS.

3.4.2 Aufbau und Funktionsweise

Das Perzeptron (SLP), erstmals entwickelt von ROSENBLATT, gilt als grundlegende Verarbeitungseinheit eines KNN. Die Eingaben eines Perzeptrons können entweder von außerhalb stammen oder Ausgabeeinheiten anderer Perzeptronen sein. Mit jeder dieser Eingaben ist ein synaptisches Gewicht verbunden. Auf der Grundlage dieser Informationen, den Eingaben sowie deren korrespondierenden synaptischen Gewichten, setzt sich anschließend die Ausgabe zusammen. Diese kann, beispielsweise im einfachsten Fall, die Summe der gewichteten Eingaben sein.[95]

[95] Vgl. Rosenblatt, F. (1958); Alpaydin, E. (2008).

Abbildung 6: Einlagiges Perzeptron



Quelle: In Anlehnung an Alpaydin, E. (2008).

Ein simples einlagiges Perzeptron ist in Abbildung 6 dargestellt. *Sj* mit *j = 1..d* stellt die Eingaben des Perzeptrons mit *wj* als dazugehörige Verbindungsgewichte dar. *d* entspricht dabei der Gesamtzahl der Eingabeeinheiten. Die Variable *y* dient als Ausgabeeinheit. Eingabeeinheit S_0 nimmt als Einheit für die Verzerrung eine besondere Rolle ein. S_0 dient als Schwellwert oder Bias, ab dem das Perzeptron schaltet und wird in diesem Fall fix auf eins gesetzt.

Die Ausgabe *a* des Perzeptrons aus Abbildung 6 kann beispielsweise die Summe der anhand der Synapsen gewichteten Eingaben sein. Eine entsprechende Funktion ist in Gleichung (7) dargestellt. Der Schwellwert S_0 stellt dabei den Ordinatenabschnitt der Funktion dar. Er liegt typischerweise bei eins.

$$a = \sum_{j=1}^{d} w_j * S_j + w_0 * S_0 = \sum_{j=1}^{d} w_j * S_j + w_0 \qquad (7)$$

Klassifikation

Da die Funktion in Gleichung (7) eine Hyperebene beschreibt, kann sie dafür genutzt werden, einen eingegebenen Vektorraum in einen positiven und einen negativen Halbraum zu trennen. Die Trennung kann anhand des Vorzeichens der Ausgabe vorgenommen werden. Bei der Anwendung dieser Funktion zur Klassifikation ist sie demnach in der Lage zwei Klassen zu unterscheiden. Die passende Schwellwertfunktion *s(a)* dazu ist in (8) dargestellt.

$$s(a) = \begin{cases} 1 \; falls \; a > 0 \\ 0 \; falls \; a \leq 0 \end{cases} \qquad (8)$$

Auf Basis dieser Schwellwertfunktion lässt sich somit eine spezifische Klassifikationsregel ableiten (9), analog der mathematischen Kurzform zur Schätzung bei einem Bayesklassifikator (1).

$$wähle \begin{cases} C1 \; falls \; s(a) = s(\sum_{j=1}^{d} w_j * S_j + w_0) > 0 \\ C2 \; falls \; s(a) = s(\sum_{j=1}^{d} w_j * S_j + w_0) \leq 0 \end{cases} \qquad (9)$$

Um im Anschluss an die Klassifikation die a-posteriori-Wahrscheinlichkeit für das Ergebnis zu ermitteln, kann die Sigmoid-Funktion auf die Ausgabe angewandt werden. Dies ist möglich, da lediglich zwischen zwei Klassen getrennt und somit nur ein Wert in der Schwellwertfunktion verglichen wird. Die dazugehörige Formel zur Ermittlung der Wahrscheinlichkeit ist in Gleichung (10) dargestellt.

$$y = sigmoid(a) = \frac{1}{1+e^{-\sum_{j=1}^{e} w_j * S_j + w_0}} \qquad (10)$$

Training des Perzeptrons

Neuronale Netze werden für gewöhnlich iterativ trainiert, indem das Netz der Reihe nach mit unterschiedlichen Merkmalsvektoren durchlaufen wird. Nach jedem dieser Durchläufe passt es seine synaptischen Gewichte leicht an, um Fehlklassifikationen so gering wie möglich zu halten. Zu Beginn werden alle Gewichte mit Zufallswerten initialisiert. Dieser Ansatz wir allgemeinhin als Online-Lernen bezeichnet und ist Teil des überwachten Lernens. Ein Durchlauf durch alle Merkmalsvektoren innerhalb des Trainingskorpus wird ‚Epoche' genannt.

Um das Netz trainieren zu können wird für die jeweilige Iteration *t* zunächst ein Trainingstupel bestehend aus Merkmalsvektor *S* und Klassenzugehörigkeit *r* definiert (11). Dabei repräsentiert C_1 die Klasse Spam und C_2 die Klasse Ham.

$$(S^t, r^t) \; mit \; r^t = 1 \; falls \; S^t \in C_1 \; und \; r^t = 0 \; falls \; S^t \in C_2 \qquad (11)$$

Mithilfe des Tupels sowie dessen jeweilige Ausgabe y^t lässt sich die Fehlerfunktion, auch Kreuzentropie genannt, für die Sigmoidfunktion herleiten. Sie bezeichnet das Unterschiedlichkeitsmaß zwischen der angenommenen Klasse und der tatsächlichen.[96] Die Formel zur Berechnung des Fehlermaßes ist in Gleichung (12) aufgelistet. Auf die

[96] Vgl. Polasek, W. (1994); Carstensen, K.U. et al. (2010).

Herleitung dieser Formel wird im Rahmen dieses Arbeitspapieres verzichtet und stattdessen auf die entsprechende Literatur verwiesen.[97]

$$E^t(w_j \mid S^t, r^t) = -r^t * log(y^t) + (1 - r^t) * log(1 - y^t)$$

Stochastischer Gradientenabstieg

Der stochastische Gradientenabstieg wird verwendet, um die Kreuzentropie aus Gleichung (12) iterativ zu minimieren. Dazu wird die Fehlerfunktion für das jeweilige Gewicht w_j partiell abgeleitet. Das Ergebnis dieser Ableitung, der Gradient, gibt dann die Richtung an, in der die Fehlerfunktion wächst. Folglich gibt die Negation des Gradienten die Richtung an, in welcher der Fehler minimiert wird.

$$\Delta w_j = -\eta \frac{\delta E}{\delta w_j} \qquad (13)$$

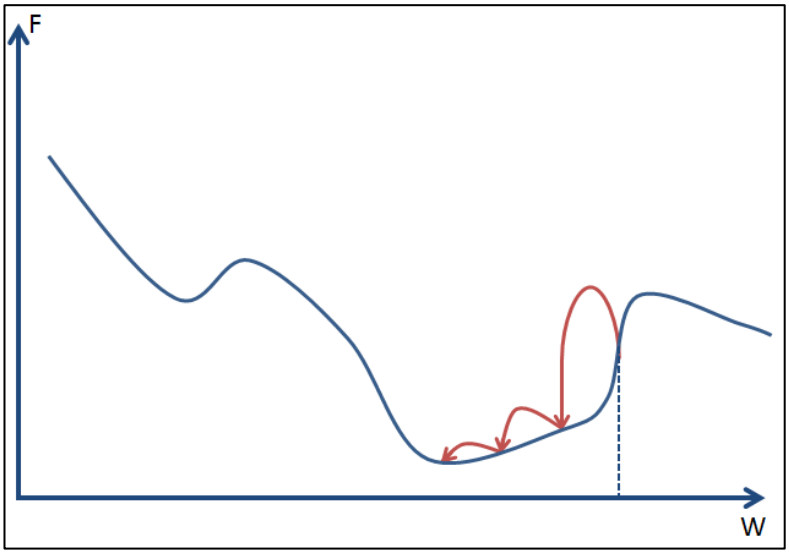
Die Ermittlung des Anpassungswerts ist in Gleichung (13) dargestellt. Dieser Wert stellt die Änderung dar, mit dem das Gewicht w_j aktualisiert wird. Zusätzlich zum Gradientenabstieg geht in die Ermittlung auch der Parameter η ein. Dieser Parameter wird auch als Lernfaktor oder Schrittweite bezeichnet und legt fest, wie weit der Abstieg zum lokalen Minima der Kreuzentropie in die jeweilige Richtung geht. Die Anpassung selbst ist in Gleichung (14) dargestellt.

$$w_j = w_j + \Delta w_j \qquad (14)$$

Abbildung 7 veranschaulicht noch einmal den generellen Ablauf des Abstiegs anhand eines zweidimensionalen Raums. Mit jeder Iteration nähert sich das Ergebnis der Fehlerfunktion einem lokalen Minimum an.

[97] Vgl. Alpaydin, E. (2008), Patterson, D.W. (1997), Polasek, W. (1994) und Carstensen, K.U. et al. (2010).

Abbildung 7: Gradientenabstiegsverfahren



Quelle: In Anlehnung an Bishop, C.M. (1995).

Die spezifische Aktualisierungsregel für das vorgestellte einlagige Perzeptron ergibt sich nun durch Einsetzen der Fehlerfunktion (12) in die Gleichung (13) zur Ermittlung des Anpassungswerts. Da die Herleitung der Aktualisierungsregel auf Basis des Gradientenabstiegs nicht Kern dieses Arbeitspapieres ist, wird lediglich die Formel in Gleichung (15) aufgezeigt. Zur Herleitung dieser Regel sei auf die Literatur verwiesen.[98]

$$\Delta w_j = \eta * (r^t - y^t) * S_j^t \qquad (15)$$

3.4.3 Mehrlagige Perzeptronen

Das folgende Kapitel stellt das Konzept des mehrlagigen Perzeptrons (MLP) vor. Im Gegensatz zu dem zuvor vorgestellten einlagigen Perzeptron, ist diese Netzarchitektur in der Lage, auch Funktionen zu approximieren, die nicht linear separierbar sind.[99]

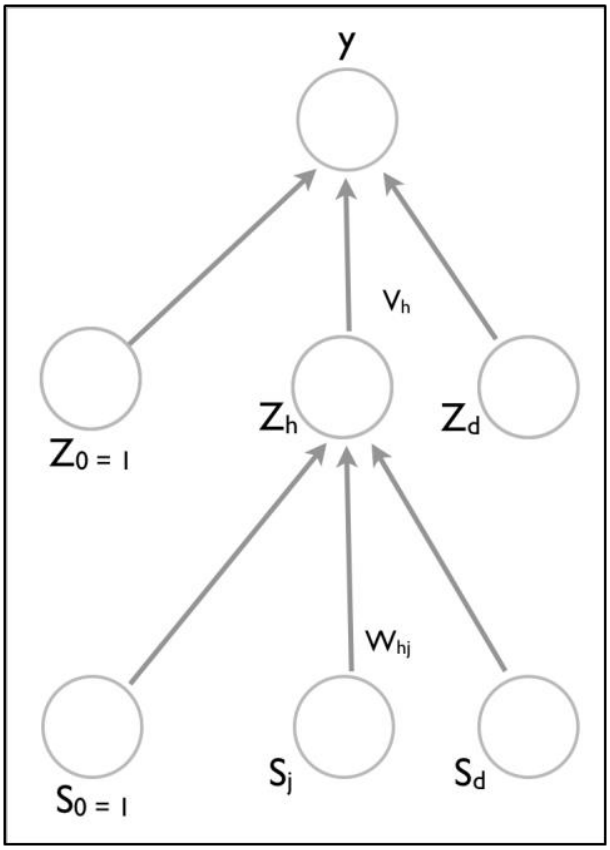
Zur Vorstellung dieses Konzepts wird zunächst auf die grundlegende Architektur von mehrlagigen Perzeptronen und Feedforward-Netzen eingegangen. Im Anschluss daran wird das Trainingskonzept mithilfe des Backpropagation Algorithmus erläutert.

[98] Vgl. Alpaydin, E. (2008), Rey, G.D., Wender, K.F. (2010), Weinberger, G. (2009) und Kruse, R. et al. (2012).
[99] Vgl. Alpaydin, E. (2008); Patterson, D.W. (1997); Minsky, M.L., Papert, S. (1969).

Grundlegende Architektur

Das zuvor erläuterte SLP besteht aus exakt zwei Schichten, einer Eingabe- sowie einer Ausgabeschicht, die mittels synaptischer Gewichte miteinander verbunden sind. Diese Schichten werden im MLP um weitere Zwischenschichten, den sogenannten versteckten Schichten, erweitert. Als versteckte Schichten werden dabei grundsätzliche alle Schichten bezeichnet, die keine Aus- oder Eingabeschicht sind.[100] Ein Beispiel für ein solches mehrlagiges Perzeptron ist in Abbildung 8 dargestellt.

Abbildung 8: Mehrlagiges Perzeptron



Quelle: In Anlehnung an Bishop, C.M. (1995).

Jede Einheit innerhalb dieser verborgenen Schicht *z* ist ein Perzeptron für sich. Wie auch das einlagige Perzeptron zuvor, wendet es die Sigmoidfunktion auf die gewichtete Summe der Eingaben an. w_{hj} entspricht dabei den Gewichten der ersten Schicht, v_{ih} denen der zweiten. Da innerhalb der Eingabeschichten keine Berechnungen stattfinden, werden diese nicht mitgezählt. Daraus ergibt sich für das MLP aus Abbildung 8 die Bezeichnung eines zweilagigen Perzeptrons, das aus einer verborgenen sowie

[100] Vgl. Bishop, C.M. (1995).

einer Ausgabeschicht besteht. Grundsätzlich bestehen keinerlei Restriktionen für die Anzahl der Schichten. Es ist jedoch zu beachten, dass die Analyse eines Netzes mit steigender Anzahl versteckter Lagen immer schwieriger wird.

Feedforward-Netze

Das in Abbildung 8 dargestellte Netz wird auch als Feedforward-Netz bezeichnet. Der Begriff definiert eine azyklische Netzstruktur, in denen die gewichteten synaptischen Verbindungen lediglich nach vorn gerichtet sind. Dies hat zur Folge, dass die Informationen in dem Netz auch nur in eine Richtung fließen können, von der Eingabe hin zur Ausgabe. Eine Rückkopplung in eine vorherige Schicht ist nicht möglich.[101]

Lernen durch Backpropagation

Das bereits erläuterte Gradientenabstiegsverfahren benötigt unter anderem die gewünschte Ausgabe zur Gewichtsanpassung. Aus diesem Grund kann es nicht auf Synapsen verborgener Schichten angewendet werden, die nicht auf das Ausgabeneuron verweisen. Zur Anwendung auf mehrlagige Perzeptronen muss das Verfahren daher erweitert werden, indem die Kreuzentropie auch nach den Gewichten der anderen Synapsen abgeleitet wird.

Die Erweiterung basiert auf der Theorie, dass die Ergebnisse der verborgenen Schichten die Ausgabeneuronen nur indirekt über deren Nachfolger beeinflussen. Zur Gewichtsanpassung würde der Fehler somit rückwärts, von der Ausgabe über die versteckten Schichten bis zur Eingabe hin, zurück propagiert werden. Daher stammt auch der Begriff ‚Backpropagation'. Die Gleichung für die Aktualisierung ist in (16) dargestellt.

$$\Delta w_{hj} = \eta * \frac{\partial E}{\partial w_{hj}} = \eta * \frac{\partial E^t}{\partial y^t} * \frac{\partial y^t}{\partial z_h^t} * \frac{\partial z_h^t}{\partial w_{hj}} \quad (16)$$

Nach Einsetzen von E^t, y^t und z^t_h ergibt die Auflösung von (16) die Aktualisierungsregel für die versteckten Schichten (17).

$$\Delta w_{hj} = \eta * \sum((r^t - y^t) * v_h * z_h^t * (1 - z_h^t) * S_j^t) \quad (17)$$

$(r^t - y^t) * v_h$ stellt den Fehlerterm für die verborgene Einheit *h* dar. Dabei ist *$(r^t - y^t)$* der letztliche Ausgabefehler der durch das Gewicht *v_h* für die verborgene Einheit justiert

[101] Vgl. Heaton, J. (2008).

wird. $z^t_h * (1 - z^t_h)$ ist die Ableitung der Sigmoidfunktion des versteckten Neurons, S^t_j ergibt sich wiederrum durch die Ableitung der gewichteten Summe nach w_{hj}.

Es ist wichtig zu beachten, dass die Aktualisierungsgleichung zur Anpassung der Gewichte der ersten Schicht die Gewichte der zweiten verwendet. Daher sollten die Gewichte der ersten Schicht vor denen der zweiten geändert werden, um eine korrekte Aktualisierung zu gewährleisten.

Wie bereits zuvor, wird auch in diesem Fall auf die detaillierte Herleitung dieser Regel verzichtet und stattdessen auf die Literatur verwiesen.[102]

[102] Vgl. Alpaydin, E. (2008), Rey, G.D., Wender, K.F. (2010), Weinberger, G. (2009) und Kruse, R. et al. (2012).

4 Anwendung der Klassifikationsverfahren auf SMS-Spam

Aufbauend auf den Grundlagen der vorherigen Kapitel, werden nun die beiden Klassifikationsansätze auf die Erkennung von SMS-Spam angewendet. Dabei werden zunächst die grundlegenden Schritte im Lernprozess beleuchtet. Im Anschluss daran wird auf den statistischen Ansatz nach BAYES eingegangen. Dabei werden insbesondere Möglichkeiten zur Selektion und Extraktion beleuchtet. Im Anschluss wird das statistische Verfahren zur Verrechnung der Merkmale und deren Eigenschaften angewandt. Nach der konkreten Zusammenstellung des Bayesklassifikators wird der Fokus auf das neuronale Netz gelegt. In diesem Zuge wird analog zum Ansatz nach BAYES auch auf die Merkmalsverarbeitung eingegangen und im Anschluss daran eine geeignete Netzarchitektur definiert.

4.1 Verarbeitungsschritte inhaltsbasierten Lernens

Für die Anwendung inhaltsbasierter Klassifikationsverfahren soll im Folgenden der grundlegende Verarbeitungsablauf einzelner Nachrichten zur Erlernung der Klassen definiert und erläutert werden. Die einzelnen Schritte werden in Anlehnung an GÓMEZ HIDALGO u. a. und GOWEDER u. a. wie folgt untergliedert:

1. Auswahl relevanter Teilsegmente
2. Zerlegung in Teilfragmente
3. Repräsentation der Fragmente
4. Aussortieren ausdrucksschwacher Tupel
5. Lernen der aussagekräftigen Tupel

Auswahl relevanter Teilsegmente

Im ersten Schritt werden zunächst nicht relevante Elemente aus der Nachricht entfernt und der Fokus auf die für die Klassifikation verwendeten Segmente gelegt. Das in diesem Arbeitspapier verwendete Segment, der textuelle Inhalt des Payload, wurde bereits im Grundlagenteil festgelegt. [103]

Zerlegung in Teilfragmente

Während der Zerlegung in Teilfragmente wird die vorverarbeitete Kurznachricht in semantisch zusammenhängende Bestandteile aufgeteilt. Dies kann beispielsweise die

[103] Vgl. Gómez Hidalgo, J.M. et al. (2006); Goweder, A.M. et al. (2008).

Aufteilung des Inhalts aus dem Payload in Wörter bedeuten. Die daraus entstehenden Komponenten werden Token genannt.[104]

Repräsentation der Fragmente

Zur Repräsentation werden den zuvor entstandenen Tokens verarbeitbare Werte zugewiesen. Die Nachricht wird somit in eine Menge aus Token-Wert-Tupel konvertiert, die im Anschluss daran weiter verarbeitet werden. [105] Diese Werte können beispielsweise binär oder auch als Wahrscheinlichkeiten dargestellt sein.[106]

Aussortieren ausdrucksschwacher Tupel

Im nächsten Schritt, dem Aussortieren, werden die wenig aussagekräftigen Tupel aus der Menge entfernt.[107] Für einen Bayes-Klassifikator können dies beispielsweise Tokens sein, die in Ham und Spam gleichermaßen häufig auftreten oder deren Häufigkeit im Korpus sehr gering ist und somit für die Klassifikation wenig aussagekräftig sind.[108]

Lernen der aussagekräftigen Tupel

In der Lernphase wird aus der zuvor selektierten Tupelmenge ein Klassifikationsmodell (Klassifikator) erstellt. Die Struktur des Klassifikators wird dabei durch das jeweilige Klassifikationsverfahren bestimmt. Die in diesem Arbeitspapier verwendeten Verfahren sind, wie bereits vorgestellt, der Bayes-Klassifikator sowie das neuronale Netz.

Anschließende Verarbeitung zu klassifizierender Nachrichten

Nach dem Anlernen des Klassifikators, wird jede eingehende Nachricht die klassifiziert werden soll, gemäß dieses Prozesses vorverarbeitet, in Teilfragmente aufgeteilt, in Zahlen repräsentiert und anschließend an den jeweiligen Klassifikator übergeben, der auf Basis der Informationen über die Nachrichtenkategorie entscheidet.[109]

---

[104] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[105] Vgl. Salton, G. (1989).
[106] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[107] Vgl. Gómez Hidalgo, J.M. et al. (2006).
[108] Vgl. Zdziarski, J.A. (2005).
[109] Vgl. Gómez Hidalgo, J.M. et al. (2006).

Heutige maschinelle Lernverfahren zur Spamerkennung setzen ihren Fokus auf die merkmalsverarbeitenden Schritte, da die Qualität der zahlenbasierten Repräsentation einen großen Einfluss auf die Genauigkeit der Verfahren haben.[110]

4.2 Klassifikationsansatz nach Bayes

Nachfolgend wird ein bayesscher Klassifikator zur Erkennung von SMS-Spam entwickelt. Dazu werden der Reihe nach die einzelnen Schritte des Verarbeitungsprozesses betrachtet und für den Klassifikator konkretisiert.

4.2.1 Aufteilung in Fragmente

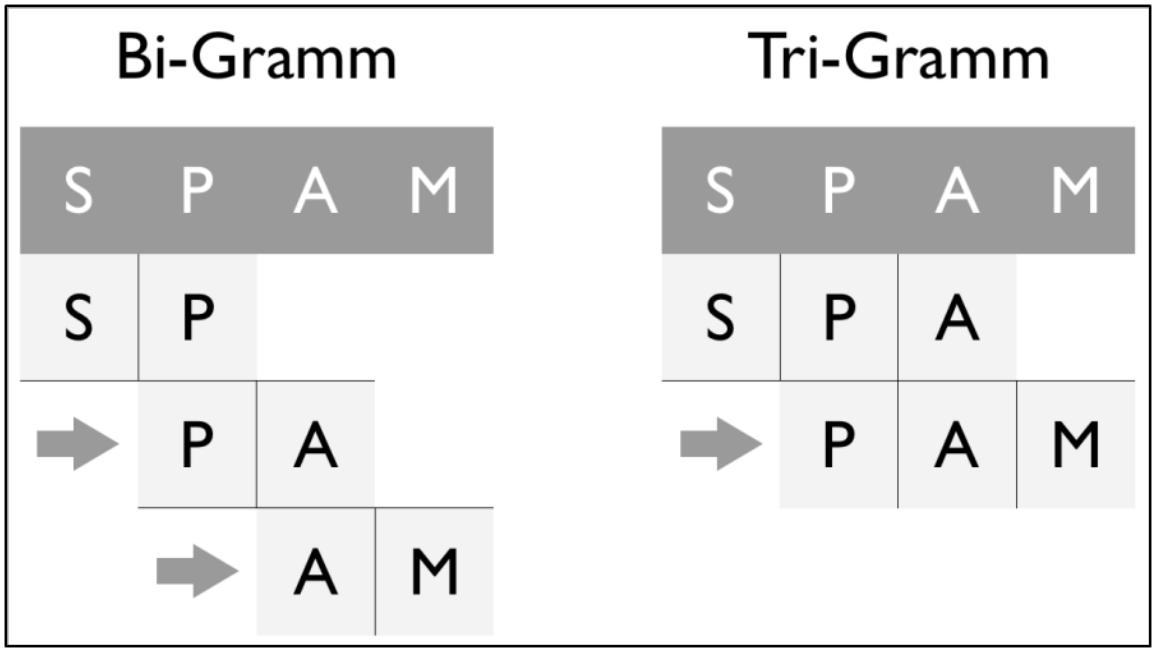
Aufbauend auf den selektierten Segmenten aus Kapitel 4.1 folgt nun die Zerlegung der Segmente in Teilfragmente, um klassifikationsrelevante Informationen zu extrahieren. Für den Bayes-Klassifikator wird dazu das Verfahren der n-Gramm-Fragmentierung verwendet.

n-Gramme

Ein n-Gramm ist eine Abfolge einzelner Textbausteine. Diese Bausteine können unterschiedlicher Art sein, wie beispielsweise Zeichen, Wörter oder auch separierte Token.[111] Zur Verdeutlichung wird in Abbildung 9 das Wort ‚SPAM' in n-Gramme der Länge n=2 (Bi-Gramm) sowie n = 3 (Tri-Gramm) zerlegt.

[110] Vgl. Bratko, A., Filipic, B. (2005).
[111] Vgl. Cavnar, W.B., Trenkle, J.M. (2010).

Abbildung 9: Zerlegung in n-Gramme



Quelle: Eigene Darstellung.

Das Verfahren findet sehr häufig Anwendung in den Sprachwissenschaften und dort speziell in der Textklassifikation.[112] Im Hinblick auf die Klassifikation von Textinhalten bieten n-Gramme nach MIAO u. a., CAVNAR und TRENKLE, XU u. a. und KANARIS u. a. mehrere Vorteile:[113]

- Robustheit gegenüber Rechtschreibfehlern
- vereinfachte Erkennung von Wortstämmen
- explizites Tokenizing von Textbausteinen entfällt
- variable Anzahl extrahierbarer Features

Die Robustheit gegenüber Rechtschreibfehler wird durch die Extraktion von Textsequenzen statt einzelner Tokens begründet. Der Klassifikator lernt somit nicht ein bestimmtes Wort sondern eine Reihe von Teilsequenzen die in diesem Wort vorkommen. Enthält es einen Rechtschreibfehler, sind nur die Sequenzen die diesen Fehler enthalten betroffen, die Teile die gleich geschrieben sind werden jedoch weiterhin beachtet und korrekt verwertet.[114]

Die Extraktion von Sequenzen macht das Verfahren jedoch nicht nur robuster gegenüber Rauscheffekten, sondern ermöglicht auch die Erkennung ganzer Wortstämme.

[112] Vgl. Mason, J.E. (2009).
[113] Vgl. Miao, Y., Kešelj, V., Milios, E. (2005), Cavnar, W.B., Trenkle, J.M. (2010), Xu, C., Chen, Y., Chiew, K. (2010), Kanaris, I. et al. (2006).
[114] Vgl. Cavnar, W.B., Trenkle, J.M. (2010).

Wie auch schon zur Vermeidung von Rechtschreibfehlern wird bei der Extraktion von Wortstämmen auch die Schnittmenge der einzelnen Sequenzen zweier Worte mit dem gleichen Stamm beachtet.[115]

Des Weiteren vereinfachen n-Gramme das Tokenizing. Es müssen keine bestimmten Delimiter gefunden werden, sondern der Text kann nach bestimmten Regeln vorverarbeitet und anschließend konsistent in n-Gramme fragmentiert werden.[116]

Zuletzt wird durch die Länge der n-Gramme auch die Granularität des lernenden Systems beeinflusst, denn je größer die Länge der n-Gramme ist, desto spezifischer sind die zu extrahierenden Sequenzen. Die maximale Anzahl f_{max} der möglichen Fragmente in Abhängigkeit ihrer Länge ergibt sich aus Gleichung (18). Dabei sei n die Länge der Fragmente und k die Anzahl der möglichen Zeichen im Text.[117]

$$f_{\max} = k^n \qquad (18)$$

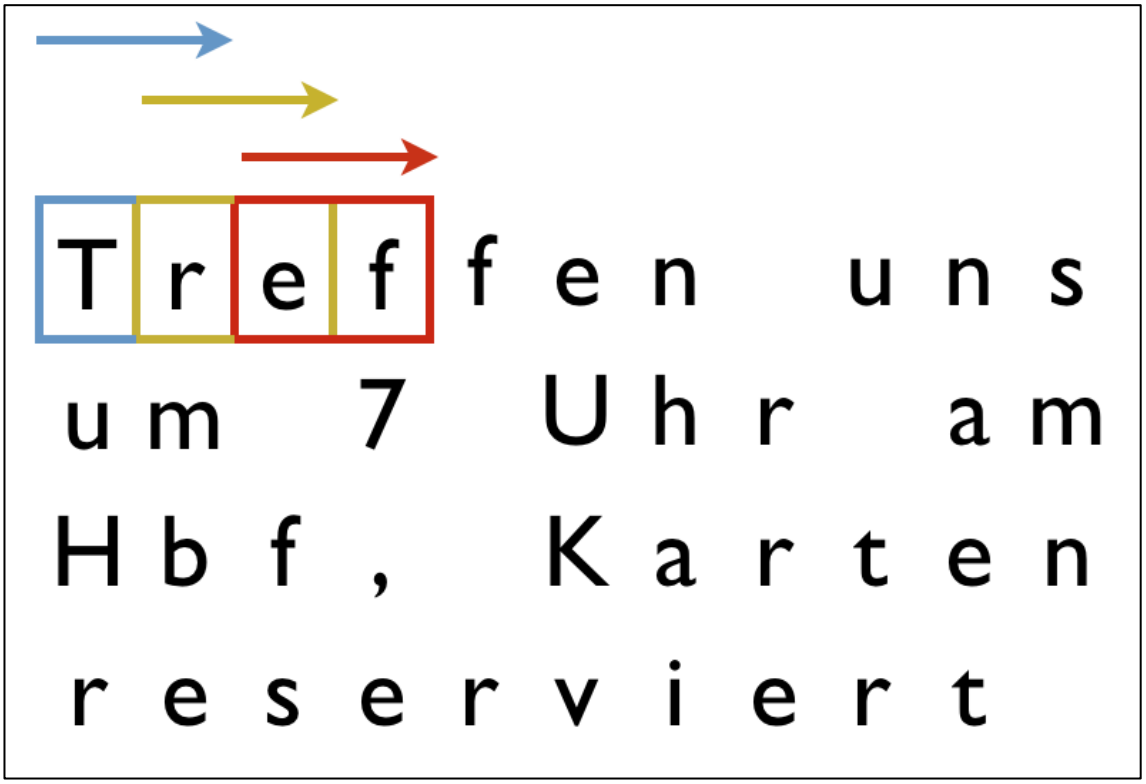
Bei einer standardmäßigen Kodierung einer Kurznachricht im 7-Bit Format und einer Fragmentlänge von drei Zeichen ergeben sich somit 128^3 = 2:097:152 mögliche Merkmale die extrahiert und verarbeitet werden können.

Horizontale Fragmentierung

In der Literatur und Forschung werden n-Gramme anstelle von Schlüsselwörtern genutzt.[118] Dabei wird der Text horizontal durchlaufen und inhaltliche Signaturen anhand des Textverlaufs extrahiert. Dieses Vorgehen ist in Abbildung 10 mit einer n-Gramm-Länge von n = 2 dargestellt.

[115] Vgl. Miao, Y., Kešelj, V., Milios, E. (2005).
[116] Vgl. Kanaris, I. et al. (2006).
[117] Vgl. Xu, C., Chen, Y., Chiew, K. (2010).
[118] Vgl. Miao, Y., Kešelj, V., Milios, E. (2005), Cavnar, W.B., Trenkle, J.M. (2010), Xu, C., Chen, Y., Chiew, K. (2010), Kanaris, I. et al. (2006), Zdziarski, J.A. (2005), Abou-Assaleh, T., Cercone, N., Sweidan, R. (2003).

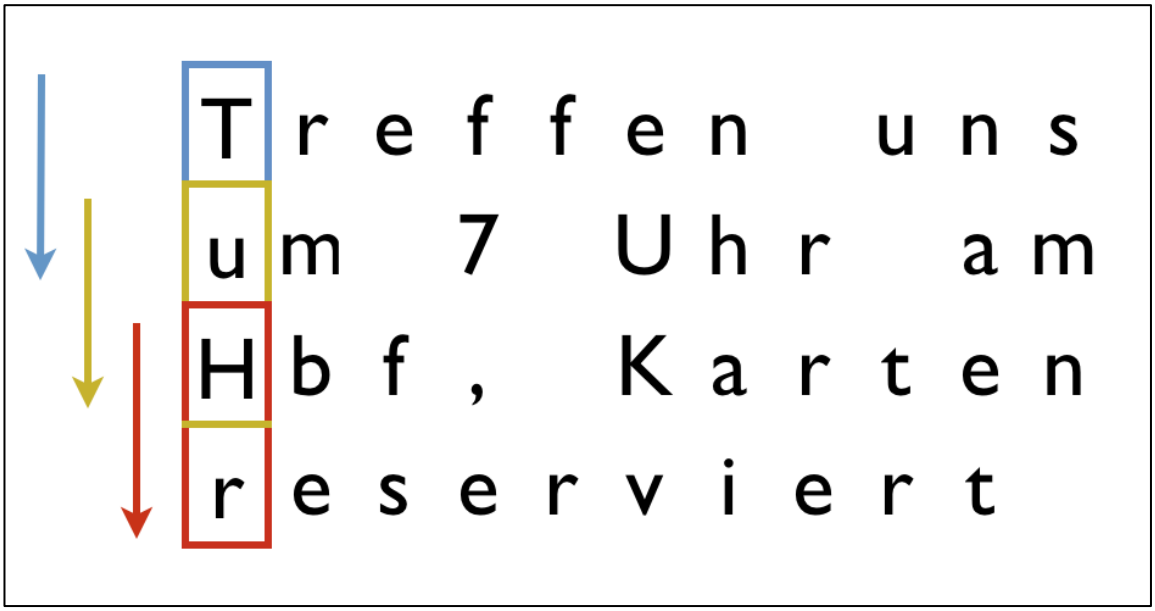
Abbildung 10: Horizontale n-Gramm-Fragmentierung



Quelle: Eigene Darstellung.Vertikale Fragmentierung

Ein neues Vorgehen, das in diesem Arbeitspapier zusätzlich zum Standardverfahren eingesetzt werden soll, ist die vertikale Fragmentierung des SMS-Payloads. Dabei werden die n-Gramme zeilenweise gebildet, statt wie zuvor von links nach rechts. Dargestellt ist dieses Vorgehen in Abbildung 11 mit einer Fragmentlänge von n = 2. Da der Inhalt von Kurznachrichten sehr kurz ist, lassen sich auf diese Weise weitere Features extrahieren, die zur Klassifikation herangezogen werden können. Dabei wird der Fokus weniger auf den Inhalt und mehr auf die Struktur und Formatierung gelegt, wodurch auch Textschablonen besser erkannt werden können.

Abbildung 11: Vertikale n-Gramm-Fragmentierung



Quelle: Eigene Darstellung.

Die optimale n-Gramm-Länge gilt es zu evaluieren, standardmäßig wird zu Beginn jedoch eine Länge von *n = 3* für beide Varianten, horizontal und vertikal, festgelegt.

4.2.2 Repräsentation der Fragmente

In der folgenden Phase werden den n-Grammen während des Trainings Zahlenwerte zugewiesen, die das Klassifikationsmodell für die Ermittlung der korrekten Klasse verwendet. Dazu wird die Klassen-Likelihood der einzelnen Token nach GRAHAM ermittelt.[119] Wie in Kapitel 3.3.2 beschrieben, gibt dieser Wert die Wahrscheinlichkeit dafür an, dass das gegebene n-Gramm der Klasse *K = Spam* angehört.

Die Formel zur Berechnung der Klassen-Likelihood nach GRAHAM ist in Gleichung (19) dargestellt. X_{Spam} stellt dabei die absolute Häufigkeit des n-Gramms im Spamkorpus dar und X_{Ham} die Häufigkeit im Hamkorpus. Dementsprechend stellt T_{Spam} die Anzahl der gesamten, im Training verwendeten, Spamnachrichten und T_{Ham} die Anzahl der verwendeten Hamnachrichten dar.

$$P(S_x \mid K = Spam)Graham = \frac{\frac{X_{Spam}}{T_{Spam}}}{\left(\frac{X_{Spam}}{T_{Spam}}\right)+\left(\frac{X_{Ham}}{T_{Ham}}\right)} \qquad (19)$$

[119] Vgl. paulgraham.com (2002), 12. Jun. 2013.

Das Ergebnis ist ein Wahrscheinlichkeitswert zwischen null und eins. Fragmente mit einer Wahrscheinlichkeit größer 0;5 gelten als Indikatoren für Spam. Fragmente mit einem Wert kleiner als 0,5 gelten als Ham. Somit gelten Fragmente mit einem Wert von 0,5 als neutral.[120][121]

Bias

Um die Falsch-Positiv-Rate während der Klassifikation zu minimieren, wurde die Berechnung der Klassen-Likelihood nach Graham durch Einsatz eines Bias-Werts optimiert. Indem der Wert X_{Ham} mit 2 multipliziert wird, wird implizit die Klassen-Likelihood des Fragments gesenkt. Somit gelten nur Fragmente als Indikatoren für Spam, wenn sie in Spam auch deutlich häufiger auftreten als in Ham. Die optimierte Gleichung für den Ansatz nach Graham ist in Gleichung (20) dargestellt.

$$P(S_x \mid K = Spam)Graham = \frac{\frac{X_{Spam}}{T_{Spam}}}{\left(\frac{X_{Spam}}{T_{Spam}}\right)+\left(\frac{2*X_{Ham}}{T_{Ham}}\right)} \qquad (20)$$

Einseitig vorkommende Tupel

Einseitige Merkmale sind starke Indikatoren für eine Klasse, da sie auch nur in einer vorkommen. Jedoch entstehen dadurch absolute Wahrscheinlichkeitswerte, 1 für Spam und 0 für Ham. Dies kann im späteren Teil die statistische Kombination der einzelnen Merkmale negativ beeinflussen oder, indem durch 0 geteilt wird, das System zum Absturz führen. Aus diesem Grund werden bei einseitig vorkommenden Tupel fixe Wahrscheinlichkeiten verwendet. Im Fall von einer Einseitigkeit in der Klasse Spam wird die Wahrscheinlichkeit auf $P(S_x \mid K = Spam)Graham = 0{,}99$ gesetzt und im Fall von Ham auf $P(S_x \mid K = Spam)Graham = 0{,}01$.[122]

Klassen-Likelihood nach Robinson

ROBINSON[123] erweiterte die Formel zur Berechnung der Klassen-Likelihood um einen Anpassungsmechanismus, der die Gesamtanzahl der Fragmente in den Trainingsdaten miteinbezieht. Die Erweiterung basiert auf der Theorie, dass Fragmente die weni-

[120] Vgl. Zdziarski, J.A. (2005).
[121] Vgl. paulgraham.com (2002), 12. Jun. 2013.
[122] Vgl. Zdziarski, J.A. (2005).
[123] Vgl. Robinson, G. (2003).

ger häufig in den Trainingsdaten vorkommen, auch eine geringere Gewichtung haben sollten.[124]

Um dies zu realisieren führte Robinson drei Variablen ein:

- N, absolute Auftrittshäufigkeit des N-Gramms
- X, Wert der angenommen würde, wenn das N-Gramm nicht in den Trainingsdaten vorkäme
- S, Glättungswert zur Feinanpassung

Diese Variablen werden in Gleichung (21) mit dem zuvor erläuterten Wahrscheinlichkeitswert nach Graham kombiniert.[125]

$$P(S_x \mid K = Spam)Robinson = \frac{(S*X)+(N*P_{Graham})}{S+N} \qquad (22)$$

Durch die Formel werden die Gewichte für Fragmente, deren Auftrittshäufigkeit sehr gering ist, geschwächt. Demnach fallen sie während der Klassifikation auch weniger stark ins Gewicht.[126]

4.2.3 Aussortieren ausdrucksschwacher Tupel

Im Anschluss an die Zusammenstellung werden die nicht aussagekräftigen Tupel aussortiert. Dazu wird zunächst geprüft, ob die einzelnen n-Gramme häufig genug im Trainingskorpus vorkommen. Dies wird anhand eines prozentualen Schwellwerts festgelegt. Der Wert gibt an, wie häufig ein Fragment mindestens in den Trainingsdaten vorkommen muss, um in der Lernphase berücksichtigt zu werden. Je höher der Schwellwert ist, desto schneller ist auch die Klassifikation. Im Gegenzug dazu verliert das Verfahren jedoch an Genauigkeit, da Fragmente bewusst ignoriert werden. Der Schwellwert zur Filterung relevanter Merkmale wird für das Verfahren auf eine Häufigkeit von 0,01% gesetzt.

Ein weiteres Kriterium zur Ausdrucksstärke ist die ermittelte Klassen-Likelihood. Tupel mit einer Likelihood von 0,5 gelten dabei als nicht aussagekräftig, da sie kein klassenspezifisches Merkmal darstellen. Sie werden aus diesem Grund auch nicht in der Lernphase berücksichtigt.

[124] Vgl. Robinson, G. (2003); Zdziarski, J.A. (2005).
[125] Vgl. Robinson, G. (2003).
[126] Vgl. Zdziarski, J.A. (2005).

4.2.4 Lernen aussagekräftiger Tupel

In der abschließenden Phase werden alle relevanten Tupel zusammengefasst und in eine, für das Klassifikationsmodell, nutzbare Form transformiert. Die hier vorgestellte Form entspricht dabei der nach Zdziarski[127] und wird nachfolgend als „Datenbasis" bezeichnet.

Aufbau der Datenbasis

Die Datenbasis nach Zdziarski besteht im Grunde genommen aus den zuvor ermittelten Fragmenten und ihren dazugehörigen Klassen-Likelihoods. Sie ist mit einem Katalog vergleichbar, der die erlernten Features enthält und während der eigentlichen Klassifikation zur Ermittlung der Klasse herangezogen wird.[128]

Unbekannte Fragmente während der Klassifikation

Ein weiterer Punkt der beachtet werden sollte, ist der Umgang mit Fragmenten, die nicht in der Datenbasis existieren, also nicht vom System gelernt wurden. Dies kann entweder daran liegen, dass das betroffene Fragment nicht im Trainingskorpus vorkam oder aber es wurde aufgrund fehlender Prägnanz aussortiert. GRAHAM und ZDZIARSKI empfehlen in diesem Fall, dem betroffenen Fragment einen neutralen Likelihood-Wert von 0,4 oder 0,5 zuzuweisen. Damit wird zum einen verhindert, dass neue Features in legitimen Nachrichten zu einer Einordnung in die Klasse Spam führen, zum anderen führen somit auch zufällige Zeichenketten, die bewusst in eine Spam-Nachricht injiziert werden, nicht zu einer fälschlichen Einordnung in die Kategorie Ham.[129] In diesem Fall liegt der neutrale Likelihood-Wert bei 0,5.

4.2.5 Statistische Verrechnung

Nachdem im Kapitel zuvor das Training des Klassifikators thematisiert wurde, wird der Fokus nun auf die statistische Kombination der gelernten Features während der Klassifikation gelegt. Dazu wird der Satz von Bayes[130] auf den konkreten Fall der Erkennung von Spam angewendet, um aufzuzeigen wie die Verrechnung innerhalb des zu entwickelnden Systems funktioniert.

---

[127] Vgl. Zdziarski, J.A. (2005).
[128] Vgl. Zdziarski, J.A. (2005).
[129] Vgl. Zdziarski, J.A. (2005).
[130] Vgl. Bayes, T. (1763).

Die theoretischen Grundlagen des Bayes-Theorems wurden bereits in Kapitel 3.3.2 erläutert. Nun soll das Verfahren mithilfe der trainierten Datenbasis auf zu prüfende Kurznachrichten angewandt werden. Dazu werden, wie bereits beschrieben, folgende Wahrscheinlichkeiten benötigt:

- die Klassen-Likelihood $P(S_x \mid K = Spam)$ der Fragmente aus der Datenbasis
- die Evidenzen $P(S_x)$ der Nachrichtenfragmente
- die a-Priori-Wahrscheinlichkeit $P(K_{Spam})$ für Spam-Nachrichten
- die a-Priori-Wahrscheinlichkeit $P(K_{Ham})$ für Ham-Nachrichten

Die gesamte Klassen-Likelihood ergibt sich aus dem Produkt der einzelnen Merkmalswahrscheinlichkeiten (22).

$$P(S_x \mid K = Spam) = \prod_{i=1}^{x} p(S_x = i \mid K_{Spam}) \quad (22)$$

Die Klassen-Likelihood für die Klasse K = Ham ergibt sich wiederrum aus der Gegenwahrscheinlichkeit von $P(S_x \mid K = Spam)$, dargestellt in Gleichung (23).

$$P(S_x \mid K = Ham) = 1 - \prod_{i=1}^{x} p(S_x = i \mid K_{Spam}) \quad (23)$$

Wie in Kapitel 3.3.2 beschrieben, lässt sich mithilfe von Gleichung (3) die Evidenz aus den beiden a-priori-Wahrscheinlichkeiten der Klassen sowie der Klassen-Likelihood berechnen. Für ein besseres Verständnis wird die Formel zur Berechnung der Evidenz noch einmal in Gleichung (24) dargestellt. Die Klassenwahrscheinlichkeiten $P(K_{Spam})$ und $P(K_{Ham})$ erhalten den Wert 0,5.

$$P(S_x) = P(S_x \mid K = Spam) * 0{,}5 + P(S_x \mid K = Ham) * 0{,}5 \quad (24)$$

Mit diesen Informationen lassen sich die einzelnen Wahrscheinlichkeiten für die Zugehörigkeit der Fragmente zu der Klasse K = Spam ermitteln, indem der in Kapitel 3.3.2 beschriebene Satz von Bayes verwendet wird. Zum besseren Verständnis ist die Formel noch einmal in Gleichung (25) dargestellt.

$$P(K_{Spam} \mid S_x) = \frac{0{,}5 * p(S_x \mid K_{Spam})}{p(S_x)} \quad (25)$$

Durch das Einsetzen der Evidenz-Formel aus Gleichung (24) lässt sich der Satz von Bayes nochmals vereinfachen. Das Ergebnis der Vereinfachung in Gleichung (26) dargestellt [16].

$$P(K_{Spam} \mid S_x) = \frac{p(S_x \mid K_{Spam})}{p(S_x \mid K_{Spam}) + p(S_x \mid K_{Ham})} \quad (26)$$

Es gilt noch zu erwähnen, dass der Satz von Bayes nicht auf alle Fragmente der zu klassifizierenden Nachricht angewendet wird, sondern nur auf die prägnantesten.

GRAHAM empfiehlt die 15 aussagekräftigsten Fragmente mit der höchsten oder niedrigsten Klassen-Likelihood zu wählen und zu kombinieren.[131]

Das im Laufe dieses Kapitels entwickelte mathematische Verfahren (26), entspricht in dieser Form dem von SAHAMI u. a. entwickelten bayesschen Spamfilter, der durch die Ausarbeitung ‚A plan for spam' von Graham auch heute noch sehr populär ist.

4.3 Klassifikation durch mehrlagige Perzeptronen

Nachdem zuvor ein konkreter Klassifikationsansatz mithilfe des Satzes von Bayes erarbeitet wurde, soll der Fokus im Folgenden auf das mehrlagige Perzeptron zur Erkennung von SMS-Spam gelegt werden.

Verarbeitung struktureller Merkmale

Die Verarbeitung struktureller Merkmale unterscheidet sich in diesem Arbeitspapier in mehreren Punkten von der Verarbeitung inhaltlicher. Aus diesem Grund wird der bereits vorgestellte Klassifikationsprozess nachfolgend leicht modifiziert, um mithilfe des MLP strukturelle statt inhaltliche Textmerkmale zu lernen.

Zunächst einmal werden keine semantisch zusammenhängenden Fragmente extrahiert, sondern strukturbasierte Messgrößen. Eine solche Messgröße kann beispielsweise die Nachrichtenlänge, die durchschnittliche Wortlänge oder auch die Anzahl der Ziffern in einer Nachricht sein. Somit wird die Nachricht nicht in Teilfragmente, sondern in textübergreifende Messgrößen zerlegt. Die Messgrößen müssen keinen Repräsentationsschritt durchlaufen, da sie bereits in Zahlenform dargestellt sind. Zuletzt ist auch die Anzahl der Features entsprechend den Eingabeparametern des MLP fix, wodurch die Phase zur Aussortierung, genauso wie die der Repräsentation, wegfällt.

Verarbeitungsschritte strukturbasierten Lernens

Der modifizierte Lernprozess für das neuronale Netz besteht somit lediglich aus drei Schritten:

1. Auswahl relevanter Teilsegmente
2. Zerlegung in Messgrößen
3. Lernen der Messgrößen

[131] Vgl. Zdziarski, J.A. (2005).

Der letzte Schritt unterscheidet sich kaum von der ursprünglichen Phase. Statt des Lernens relevanter Textfragmente, lernt das Netz nun Muster innerhalb der übergebenen Messgrößen.

4.3.1 Zerlegung in Messgrößen

Nachfolgend werden die einzelnen Messgrößen vorgestellt, die für das Training des Klassifikators sowie für die anschließende Klassifikation verwendet werden. Dazu werden zunächst einige grundlegende Eigenschaften definiert, die bereits bei der Klassifikation von E-Mails sowie auch SMS Spam verwendet werden. Zusätzlich zu diesen werden anschließend noch Messgrößen aus dem Bereich der Linguistik vorgestellt, die auch als Eingabeparameter dienen.

Grundlegende strukturelle Eigenschaften von Texten

Im Folgenden werden die in diesem Arbeitspapier verwendeten strukturellen Textmerkmale aufgezeigt und erläutert, die auch bereits in der Literatur und Forschung auf ähnliche Problemstellungen angewendet wurden:[132]

- Länge der Nachricht in Zeichen
- Anzahl der Worte
- Anzahl von Ziffern normalisiert durch Länge in Zeichen
- enthält URL (binäres Merkmal)
- durchschnittliche Wortlänge

Durch die Begrenzung von Kurznachrichten auf 160 Zeichen gewinnen die Nachrichtenlänge sowie die Anzahl der verwendeten Worte deutlich mehr an Bedeutung, als es für andere Medien, wie beispielsweise E-Mails, der Fall ist.[133] Auch wurde festgestellt, dass SMS Spam deutlich häufiger Zahlen und URLs enthält als Ham-Nachrichten.[134] Die durchschnittliche Wortlänge, als weiteres Merkmal, wurde bereits von CHENG u. a. zur Klassifikation von E-Mails anhand des Geschlechts des Autors eingesetzt.[135] Statt zwischen Geschlechtern soll in diesem Fall zwischen Ham- und Spam-Autor unterschieden werden.

[132] Vgl. Uysal, K.A. et al. (2013), Cheng, N. et al. (2009), Miner, G. et al. (2012).
[133] Vgl. Uysal, K.A. et al. (2013).
[134] Vgl. Uysal, K.A. et al. (2013).
[135] Vgl. Cheng, N. et al. (2009).

Linguistische Messgrößen zum Wortschatzreichtum

Für weitere Eingabeparameter, anhand derer das MLP trainiert werden soll, wird auf Strukturmerkmale aus der Linguistik zurückgegriffen. Die nachfolgenden Messgrößen stellen dabei Indikatoren für die Reichhaltigkeit des verwendeten Vokabulars in Texten dar:[136]

- Hapax Legomenon
- Hapax Dislegomenon
- Sichel's Messgröße S
- Simpson's Messgröße D

Sie werden in diesem Arbeitspapier verwendet, um eventuelle Unterschiede zwischen Spam und Ham auch anhand quantitativer Analysen des Wortschatzes zu erkennen. Für ein besseres Verständnis werden die einzelnen Spezifika nachfolgend kurz vorgestellt.

Hapax Legomenon und Hapax Dislegomenon

Der Begriff Hapax Legomenon steht für Anzahl der Worte die nur einmalig in einem Text vorkommen[137] und gilt nach Muller als stilistisches Textmerkmal.[138] Hapax Dislegomenon wiederum steht für die Anzahl aller Worte die zweimal in einem Text vorkommen.[139] Durch die beiden Indikatoren soll das Netz in der Lage sein, den Schreibstil einer Spam-Nachricht von dem einer Ham-Nachricht zu unterscheiden.

Sichel's Messgröße S

Die Messgröße S nach Sichel basiert auf der Zahl der Hapax Dislegomenon. Die Gleichung zur Berechnung von S ist in (27) dargestellt. V steht für die Anzahl unterschiedlicher Wörter im Text und V2 für die Zahl der Hapax Dislegomenon.[140]

$$S = \frac{V2}{V} \qquad (27)$$

[136] Vgl. Lüdeling, A. et al. (2009), Simpson, E.H. (1949), Koppel, M., Schler, J., Argamon, S. (2009).
[137] Vgl. Köhler, R., Altmann, G., Piotrovski˘i, R.G. (2005).
[138] Vgl. Muller, C. (1972).
[139] Vgl. Lüdeling, A. et al. (2009).
[140] Vgl. Lüdeling, A. et al. (2009).

Simpson's Messgröße D

Die Messgröße D nach Simpson ist die Wahrscheinlichkeit dafür, das zwei zufällig gewählte Wörter eines Textes nicht identisch sind. Berechnet wird der Wert mit Formel (28).[141] N steht für die Anzahl aller Wörter im Text S. ni steht für die Anzahl die Wort i in S vorkommt.

$$D = 1 - \sum_{i=1}^{S} \frac{n_i * (n_i - 1)}{N * (N - 1)} \qquad (28)$$

4.3.2 Zerlegung in Messgrößen

Als Netzarchitektur wird ein zweilagiges feedfoward Netz definiert, welches zunächst neun Eingabeneuronen und ein Ausgabeneuron besitzt. Da die Trainingsdaten, die in diesem Arbeitspapier verwendet werden, nicht vollständig gesichtet sind, kann nicht garantiert werden, dass die Klassen linear separierbar sind. Aus diesem Grund wird eine verborgene Schicht verwendet.

Anzahl der verborgenen Einheiten

Die Bestimmung der Anzahl der verborgenen Einheiten erweist sich als komplex. LINOFF und BERRY beschreiben in ihrer Ausarbeitung, dass die optimale Anzahl der Einheiten nicht nur auf den zu erkennenden Mustern beruht, sondern auch auf der Anzahl der Trainingsdaten. Bei einer zu großen Zahl verborgener Einheiten und einem zu kleinen Trainingsset kann es passieren, dass nicht mehr alle Gewichte ausreichend genau justiert werden können. Im Zuge dieses Arbeitspapieres soll daher die Bestimmung der optimalen Perzeptronenzahl in der verborgenen Schicht systematisiert werden. Dazu wird nachfolgend eine Faustregel in Form einer Gleichung aufgestellt. [142]

Systematische Ermittlung

Mithilfe von Gleichung (29) lässt sich die Anzahl c der Gewichte innerhalb eines MLP berechnen.[143] h steht dabei für die Anzahl der verborgenen Einheiten, i für die Anzahl der Eingabeneuronen. Der Wert 1 repräsentiert die Zahl der Ausgabeneuronen für diesen konkreten Fall.

---

[141] Vgl. Simpson, E.H. (1949).
[142] Vgl. Linoff, G.S., Berry, M.J. (2011).
[143] Vgl. Linoff, G.S., Berry, M.J. (2011).

$$c = h * (i + 1) + h + 1 \qquad (29)$$

LINOFF und BERRY beschreiben, dass etwa 100 Trainingsnachrichten pro Gewicht verwendet werden sollen. Ziel ist es nun, eine Formel zu entwickeln, die es ermöglicht, die Ermittlung der optimalen Anzahl der verborgenen Einheiten zu systematisieren. Dazu wird Gleichung (29) zunächst umgestellt und um die Parameter x, t und e erweitert. Der Parameter t gibt die Anzahl aller Nachrichten im Korpus an, e die Anzahl der Trainingsepochen und x die Anzahl der gewünschten Justierungen je Gewicht. Die erweiterte Gleichung ist in (30) dargestellt.

$$x * (h * (i + 1) + h + 1) = t * e \qquad (30)$$

Die linke Seite des Terms repräsentiert die Anzahl der benötigten Nachrichten für das Training. Die rechte Seite steht für die Zahl der vorhandenen Nachrichten im Trainingskorpus, multipliziert mit der Anzahl der Trainingsepochen. Durch Umstellen der Formel nach h und herstellen eines Gleichgewichts zwischen den beiden Seiten kann nun ein hypothetisches Optimum an verborgenen Einheiten h ermittelt werden. Die entsprechende Formel ist in (31) dargelegt. Die Testversuche haben ergeben, dass das neuronale Netz nach 200 Trainingsepochen die besten Ergebnisse zeigt.

$$h = \frac{t*e}{x*(i+2)} - \frac{1}{i+2} \qquad (31)$$

Nach Einsetzen von *i = 9*, *t = 600*, *x = 100* und *e = 200* ergibt sich ein grober Richtwert von *h = 110* verborgenen Einheiten.

5 Evaluation der Klassifikatoren mittels ROC-Analysen

Im folgenden Kapitel werden die Ergebnisse der beiden Klassifikationsansätze, durch die Anwendung auf den Testkorpus, vorgestellt. Dazu werden zu Beginn die grundlegenden Qualitätsindikatoren erläutert. Im Anschluss daran wird auf das Konzept der ROC-Kurven eingegangen, um eine Feinabstimmung der einzelnen Klassifikatoren zu ermöglichen. Darauf basierend werden dann die Klassifikationsergebnisse ausgewertet. Dabei werden zwei Varianten des Bayesklassifikators evaluiert, die klassische Variante, die lediglich horizontale n-Gramme verwendet, sowie die neue Variante, die horizontale sowie vertikale n-Gramme zur Entscheidungsfindung heranzieht. Im darauffolgenden Teil werden die Ergebnisse des mehrlagigen Perzeptrons evaluiert und abschließend eine Bewertung der drei Ergebnisse durchgeführt.

5.1 Qualitätsindikatoren für Klassifikatoren

Die in diesem Arbeitspapier verwendeten Indikatoren zur Messung der Klassifikatorqualität bauen auf den Kennzahlen der Konfusionsmatrix auf, die in Tabelle 1 dargestellt ist. Sie stellt das Gesamtergebnis eines Klassifikators, über einen Testlauf, bezogen auf einen Testkorpus dar.[144][145]

Tabelle 1: Aufbau der Konfusionsmatrix

	Vorhergesagte Klasse	
Wahre Klasse	Spam	Ham
Spam	Wahres Positiv (WP)	Falsches Negativ (FN)
Ham	Falsches Positiv (FP)	Wahres Negativ (WN)

Quelle: In Anlehnung an Alpaydin, E. (2008).

Die in diesem Papier verwendeten Kennzahlen sind die Fehlerrate, die Rate von Fehlalarmen, die Trefferrate (Recall), die Präzisionsrate (Precision) sowie das F-Maß. Diese Kennzahlen sollen im Folgenden kurz vorgestellt werden.

Fehlerrate

[144] Vgl. Alpaydin, E. (2008).
[145] Vgl. Fawcett, T. (2004).

Die Fehlerrate ergibt sich aus dem Verhältnis der Summe der falschen Positive und falschen Negative zur Gesamtzahl der Klassifikationsergebnisse. Sie gibt den Anteil der im Testkorpus falsch klassifizierten Nachrichten an und ist in Gleichung (32) dargestellt.[146]

$$Fehlerrate = \frac{FP+FN}{WP+FN+FP+WN} \qquad (32)$$

Rate von Fehlalarmen

Die Rate von Fehlalarmen gibt den Anteil der Nachrichten an, die eigentlich der Klasse Ham angehören, aber als Spam klassifiziert wurden. Die Formel zur Ermittlung der Kennzahl ist in Gleichung (33) dargestellt.[147]

$$Fehlalarme = \frac{FP}{FP+WN} \qquad (33)$$

Trefferrate (Recall)

In Gleichung (34) ist die Trefferrate dargestellt, auch Recall genannt. Sie gibt den Anteil der Nachrichten an, der der Klasse Spam angehört und auch als Spam klassifiziert wurde.[148]

$$Trefferrate = \frac{WP}{WP+FN} \qquad (34)$$

Präzisionsrate (Precision)

Die Präzisionsrate gibt das Verhältnis der richtig klassifizierten Spam-Nachrichten zu allen, als Spam klassifizierten, Nachrichten an. Sie ist in Gleichung (35) dargestellt.[149]

$$Präzisionsrate = \frac{WP}{WP+FP} \qquad (35)$$

F-Maß

[146] Vgl. Alpaydin, E. (2008).
[147] Vgl. Alpaydin, E. (2008); Fawcett, T. (2004).
[148] Vgl. Alpaydin, E. (2008); Fawcett, T. (2004).
[149] Vgl. Fawcett, T. (2004).

Das F-Maß, zu sehen in Gleichung (36), wurde entwickelt um Klassifikatoren anhand einer einzigen Kennzahl vergleichbar zu machen. Es wird aus dem harmonischen Mittel von Precision und Recall berechnet und bewegt sich zwischen 0 als schlechtesten und 1 als besten Vergleichswert.[150]

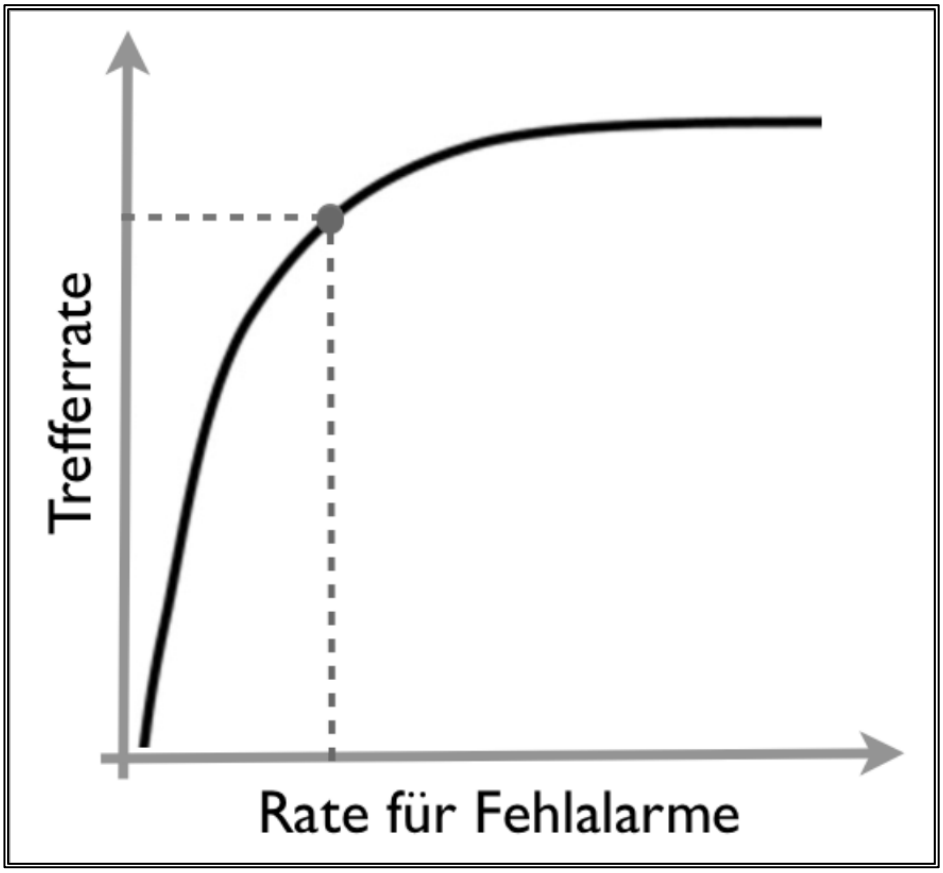
$$F = \frac{2 * Recall * Precision}{Recall + Precision} \qquad (36)$$

5.2 Konzept der ROC-Kurve

Das Konzept der ROC-Kurve stellt die Trefferrate eines Klassifikators der Rate der Fehlalarme gegenüber. ROC-Kurven haben zwei Aufgaben. Zum einen stellen sie die Performance eines Klassifikators über dem Trainingskorpus grafisch dar, zum anderen dienen sie zur Feinabstimmung der Klassifikatoren. [151] Der typische Verlauf einer solchen Kurve ist in Abbildung 12 skizzenhaft dargestellt. Da das übergeordnete Ziel die Erhöhung der Treffer bei Senkung der falschen Positive ist, ist die Performance eines Klassifikators umso besser, je mehr sich die Kurve der oberen linken Ecke annähert.

[150] Vgl. Kowalski, G. (2010).
[151] Vgl. Krzanowski, W.J., Hand, D.J. (2009); Zou, K.H., Liu, A., Bandos, A.I. (2011); Alpaydin, E. (2008).

Abbildung 12: Skizzenhafte ROC-Kurve



Quelle: In Anlehnung an Alpaydin, E. (2008).

Eine Funktion, die diese Kurve für die jeweiligen Klassifikatoren beschreibt, existiert nicht. Ein Punkt auf dieser Kurve zeigt lediglich das mit den Testdaten ermittelte Verhältnis zwischen Treffern und falschen Positiven.

Feinabstimmung von Klassifikatoren durch ROC-Kurven

Jedes Klassifikationsverfahren besitzt einen Parameter, mit dem das Verhältnis zwischen Treffern und Fehlalarmen beeinflusst werden kann. Im Rahmen dieses Arbeitspapieres ist dieser Parameter die Schwelle t, ab der dem Klassifikationsergebnis vertraut wird. Eine Erhöhung dieses Schwellwerts führt demnach zur Verringerung der falschen Positive, gleichzeitig jedoch auch zu einer Verringerung der Recall-Rate. Bei der ROC-Analyse existiert zu jedem Punkt auf der Kurve ein Wahrscheinlichkeitswert, der die Entscheidungsschwelle des Verfahrens darstellt. Je nachdem wie wichtig die Zahl der Treffer im Vergleich zu den Fehlalarmen ist, wird die Schwelle anhand des Punktes auf der Kurve verändert.[152] Um die Komplexität der Grafiken so gering wie

[152] Vgl. Alpaydin, E. (2008).

möglich zu halten sind die konkreten Schwellwerte jedoch lediglich in den Testergebnissen hinterlegt und nicht in den Diagrammen.

Messung der gewichteten Distanz

Das Ziel der Feinabstimmung ist in diesem Arbeitspapier die Ermittlung eines optimalen Verhältnisses zwischen Treffer und Fehlalarmen. Dieser Wert wird festgelegt, indem für jeden Punkt auf der ROC-Kurve die Distanz zum optimalen Ergebnis (100% Treffer und 0% Fehlalarme) gemessen wird. Der Punkt mit der geringsten Distanz stellt anschließend das Ergebnis der Feinabstimmung dar.

In diesem Arbeitspapier gilt jedoch die zusätzliche Bedingung, die Fehlalarme möglichst gering zu halten, auch wenn dies höhere Verluste in der Trefferquote bedeutet. Um den falschen Positiven ein höheres Gewicht zu verleihen wurde die Gleichung zur Messung der Distanz *d* zwischen zwei Punkten um einen Bias *b = 2* erweitert. Die entsprechende Formel zur Anwendung auf ROC-Kurven ist in Gleichung (37) dargestellt. Die Distanz wird zwischen dem Punkt P(0,1), der als Optimum gilt, und P(x,y), der auf der Kurve liegt, gemessen.

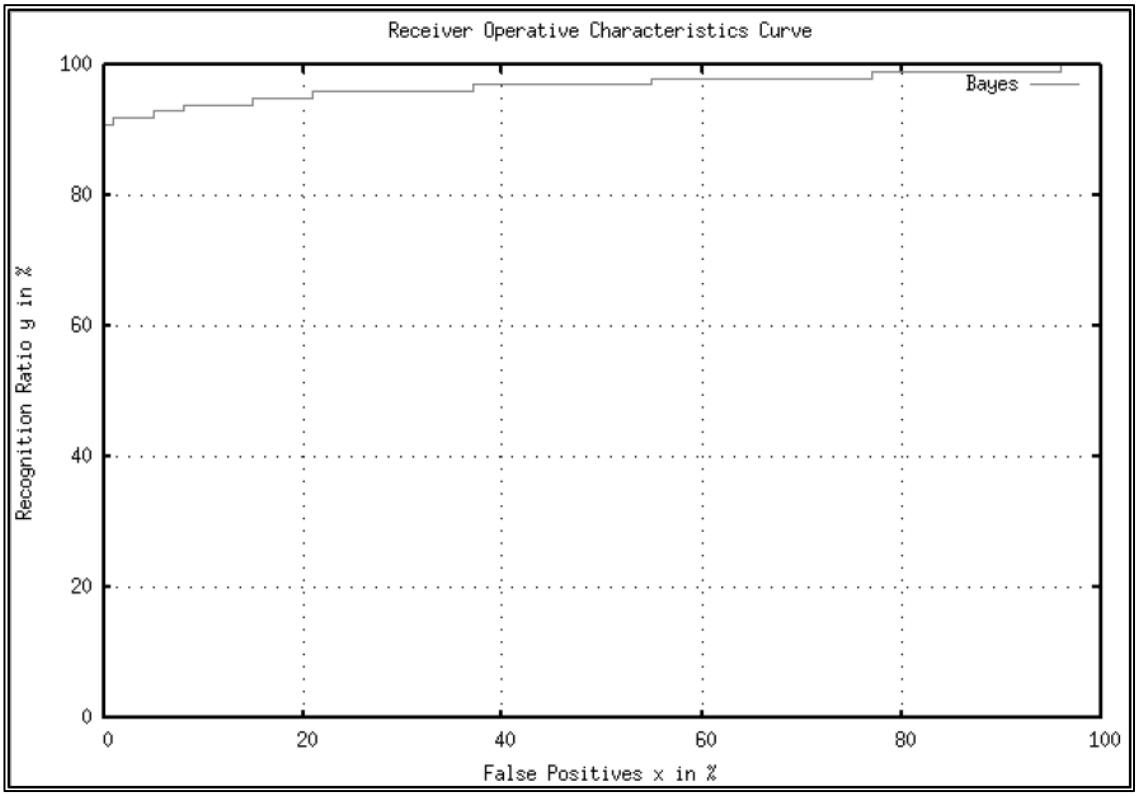
$$d = \sqrt{(0-x)^2 + b * (1-y)^2} \qquad (37)$$

5.3 Bayes-Klassifikation bei horizontalen Fragmenten

Nachfolgend wird das Ergebnis des Bayes-Klassifikators für horizontal fragmentierte Nachrichten vorgestellt. Da sich in den Testläufen eine optimale Fragmentlänge von n=3 herausstellte, wird auf Variationen mit unterschiedlichen n-Gramm-Längen nicht weiter eingegangen. Die dazugehörige ROC Kurve ist in Abbildung 13 dargestellt.

Wie sich erkennen lässt, verläuft die Kurve deutlich im oberen linken Quadranten des Koordinatensystems, was grundsätzlich Rückschlüsse auf eine hohe Klassifikationsqualität ziehen lässt.

Abbildung 13: ROC-Kurve des Bayes-Klassifikators mit horizontalen Fragmenten



Quelle: Grafik mit Gnuplot generiert, gnuplot.info (2013), 30. Aug. 2013.

Der ermittelte Schwellwert, der den optimalen Punkt auf der Kurve repräsentiert, liegt bei $t = 0{,}9*10^{-15}$. Er scheint auf den ersten Blick hoch zu sein, jedoch muss erwähnt werden, dass das Verfahren nach Bayes für extreme Wahrscheinlichkeiten bekannt ist.[153]

Auf der Grundlage dieses Schwellwerts wird die dazugehörige Konfusionsmatrix aufgebaut, die in Tabelle 2 zu sehen ist.

[153] Vgl. Zdziarski, J.A. (2005).

Tabelle 2: Konfusionsmatrix des Bayes-Klassifikators mit horizontalen Fragmenten

	Vorhergesagte Klasse	
Wahre Klasse	Spam	Ham
Spam	89,93%	10,07%
Ham	0,55%	99,45%

Quelle: Eigene Darstellung.

Fehlerrate

Basierend auf der Konfusionsmatrix ergibt sich eine Fehlerrate bei der Klassifikation des Testkorpus von 5,31%.

Rate von Fehlalarmen

Die Rate von Fehlalarmen konnte durch die Feinabstimmung über die ROC Kurve nahezu eliminiert werden und liegt bei 0,55%.

Trefferrate (Recall)

Die Recall-Rate liegt bei 89,93% und ist der niedrigen falsch Positiv Rate geschuldet. Wie in Abbildung 13 zu sehen, ist eine Rate nahe der 100% durch die sehr hohe Anzahl an Fehlalarmen nicht mehr als effektiv zu bewerten.

Präzisionsrate (Precision)

Wenn auch der Recall-Wert niedrig ist, so liegt die Precision-Rate mit 99,4% auf einem hohen Niveau und stellt dadurch die Genauigkeit des Klassifikators im Hinblick auf Fehlalarme unter Beweis.

F-Maß

Zuletzt wird das F-Maß für das Verfahren als übergreifende Qualitätskennzahl berechnet. Das Maß liegt bei 0,94 und somit nahe am optimalen Wert. Aus diesem Grund sind die Ergebnisse des Klassifikators durchaus als positiv zu werten.

Die in diesem Kapitel ermittelten Kennzahlen werden noch einmal für $t = 0{,}9*10^{-15}$ in Tabelle 3 zusammengefasst.

Tabelle 3: Kennzahlen des Bayes-Klassifikators mit horizontalen Fragmenten

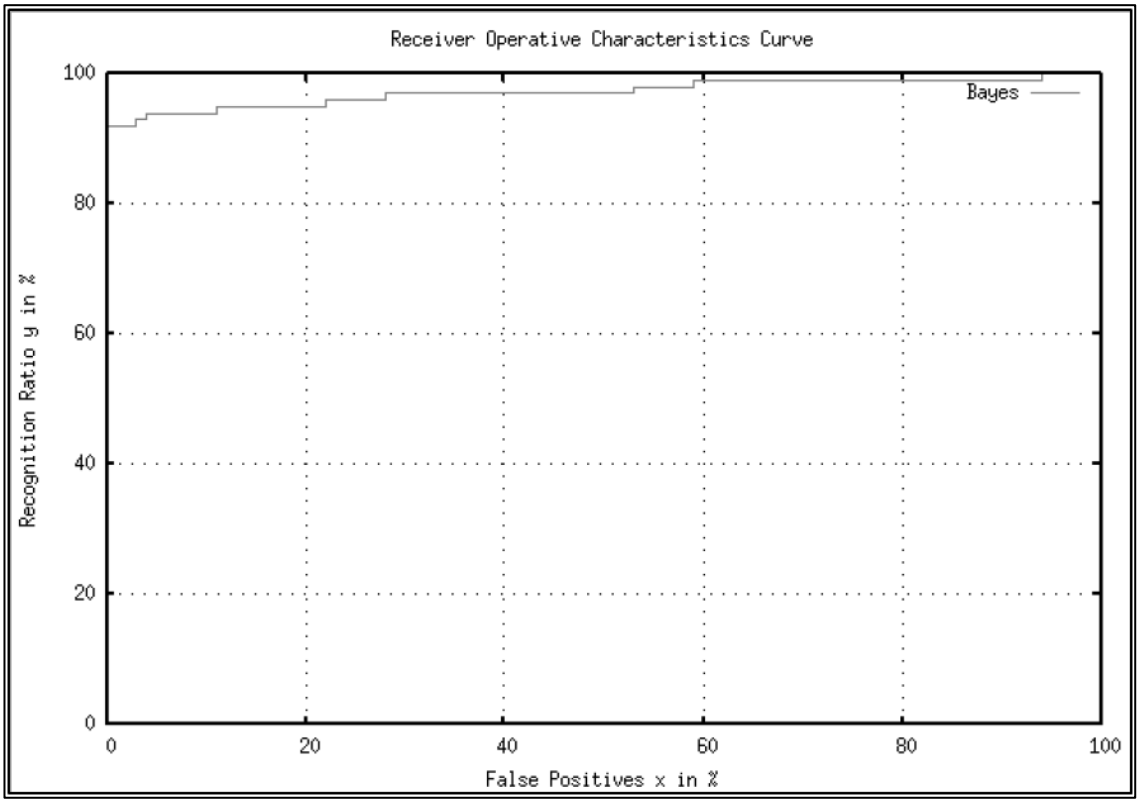
Kennzahl	Wert
Fehlerrate	5,31%
Falsch-Positiv-Rate	0,55%
Recall	89,93%
Precision	99,4%
F-Maß	0,94

Quelle: Eigene Darstellung.

5.4 Bayes-Klassifikation bei horizontalen und vertikalen Fragmenten

In diesem Kapitel wird das Ergebnis des Bayes-Klassifikators für horizontal und vertikal fragmentierte Nachrichten vorgestellt. Auch hier bleibt die Fragmentlänge mit n = 3 unverändert. Die entsprechende ROC Kurve ist in Abbildung 14 dargestellt. Da diese Kurve, wie bereits beim Verfahren mit rein horizontalen Fragmenten, auch im oberen linken Quadranten verläuft, deutet sie auf eine hohe Performance hin.

Abbildung 14: ROC-Kurve des Bayes-Klassifikators mit allen Fragmenten



Quelle: Grafik mit Gnuplot generiert, gnuplot.info (2013), 30. Aug. 2013.

Der für dieses Verfahren ermittelte Schwellwert liegt auch bei $t = 0{,}9*10^{-15}$. Die darauf basierende Konfusionsmatrix ist in Tabelle 4 dargestellt.

Tabelle 4: Konfusionsmatrix des Bayes-Klassifikators mit allen Fragmenten

	Vorhergesagte Klasse	
Wahre Klasse	Spam	Ham
Spam	92,39%	7,61%
Ham	3,16%	96,84%

Quelle: Eigene Darstellung.

Fehlerrate

Die Fehlerrate für diesen Testkorpus liegt nun bei 5,38% und ist damit höher als bei der Klassifikation durch rein horizontale n-Gramme.

Rate von Fehlalarmen

Auch die Rate der Fehlalarme liegt mit 3,16% deutlich höher. Auch wenn dieser Wert nicht als schlecht zu bewerten ist, so war das Verfahren ohne vertikale n-Gramme in dieser Hinsicht nahezu um das Siebenfache effektiver.

Trefferrate (Recall)

Die Recall-Rate liegt mit 92,39% leicht höher als die des klassischen Verfahrens.

Präzisionsrate (Precision)

Die hohe Trefferquote fällt jedoch zu Lasten der Präzision. Sie ist mit 96,69% zwar immer noch hoch, jedoch geringer als beim Ansatz mit rein horizontaler Fragmentierung.

F-Maß

Zuletzt wird auch für dieses Verfahren das F-Maß berechnet. Mit einem Maß 0,95 ist dieses Verfahren sogar positiver zu bewerten als das F-Maß bei der rein horizontalen Fragmentierung.

Grundsätzlich liegen die beiden Verfahren hinsichtlich ihrer Performance sehr nah beieinander. Für welches sich im Realfall entschieden wird ist davon abhängig ob und wie viele Fehlalarme zugunsten der Trefferrate in Kauf genommen werden wollen.

Die Performance des Klassifikators wird in Tabelle 5 noch einmal zusammenfassend für $t = 0{,}9*10^{-15}$ dargestellt.

Tabelle 5: Kennzahlen des Bayes-Klassifikators mit allen Fragmenten

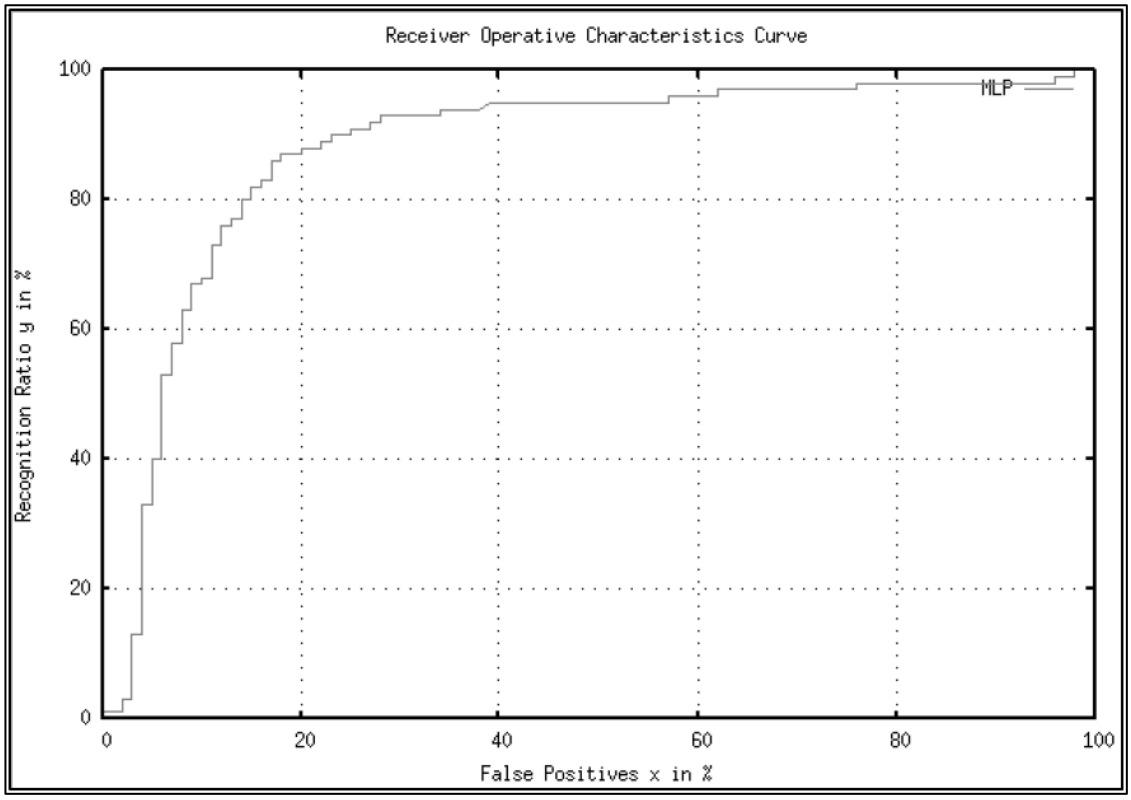
Kennzahl	Wert
Fehlerrate	5,38%
Falsch-Positiv-Rate	3,16%
Recall	92,39%
Precision	96,69%
F-Maß	0,95

Quelle: Eigene Darstellung.

5.5 Klassifikation durch mehrlagige Perzeptronen

Nachdem zuvor die Ansätze nach Bayes auf der Grundlage der Textinhalte evaluiert wurden, werden im Folgenden die Ergebnisse des mehrlagigen Perzeptrons zur Erkennung von SMS-Spam vorgestellt. Die ROC-Kurve in Abbildung 15 zeigt die Ergebnisse des Verfahrens.

Abbildung 15: ROC-Kurve des mehrlagigen Perzeptrons



Quelle: Grafik mit Gnuplot generiert, gnuplot.info (2013), 30. Aug. 2013.

Die Kurve des MLP fällt im Vergleich zu denen des Bayes-Ansatzes deutlich schlechter aus. Der ermittelte Schwellwert, nach der Methode der gewichteten Distanz, beträgt t = 0,991807287657152. Die darauf aufsetzende Konfusionsmatrix wird in Tabelle 6 dargestellt.

Tabelle 6: Konfusionsmatrix des mehrlagigen Perzeptrons

	Vorhergesagte Klasse	
Wahre Klasse	Spam	Ham
Spam	76,96%	23,04%

Ham	13,06%	86,94%

Quelle: Eigene Darstellung.

Fehlerrate

Die Fehlerrate des Verfahrens bei der Anwendung auf den Testkorpus liegt mit 18,05% deutlich höher als bei den bayesschen Ansätzen zuvor.

Rate von Fehlalarmen

Die Rate der Fehlalarme liegt bei 13,06%. Dieser Wert ist sehr hoch und könnte bei der Verwendung des Verfahrens auf einem Mobiltelefon als störend empfunden werden. Dies sollte vor einem produktiven Einsatz des Verfahrens geprüft werden.

Trefferrate (Recall)

Die Recall-Rate beträgt 76,96%. Sie ist im Vergleich zu den bayesschen Ansätzen nicht besonders hoch, senkt jedoch die Zahl der empfangenen Spamnachrichten auf weniger als ein Viertel, wodurch trotzdem eine positive Wirkung eintritt.

Präzisionsrate (Precision)

Mit 85,49% liegt die Präzisionsrate auf einem moderaten Niveau. 14,51% der als Spam erkannten Nachrichten waren eigentlich Ham Nachrichten. Es gilt, wie bereits erwähnt, zu evaluieren, inwiefern dies als störend empfunden wird.

F-Maß

Wie bereits durch die ROC-Kurve und den einzelnen Kennzahlen angedeutet, bestätigt das F-Maß mit einem Wert von 0,81 eine schlechtere Gesamtperformance gegenüber den Bayes-Klassifikatoren. Er liegt dennoch im positiven Bereich und zeigt, dass sich mehrlagige Perzeptronen auch zur Klassifikation von SMS-Spam eignen.

Die einzelnen Kennzahlen des MLP bei der Anwendung auf den Testkorpus werden für den Schwellwert t = 0,991807287657152 noch einmal in Tabelle 7 zusammengefasst.

Tabelle 7: Kennzahlen des Bayes-Klassifikators mit allen Fragmenten

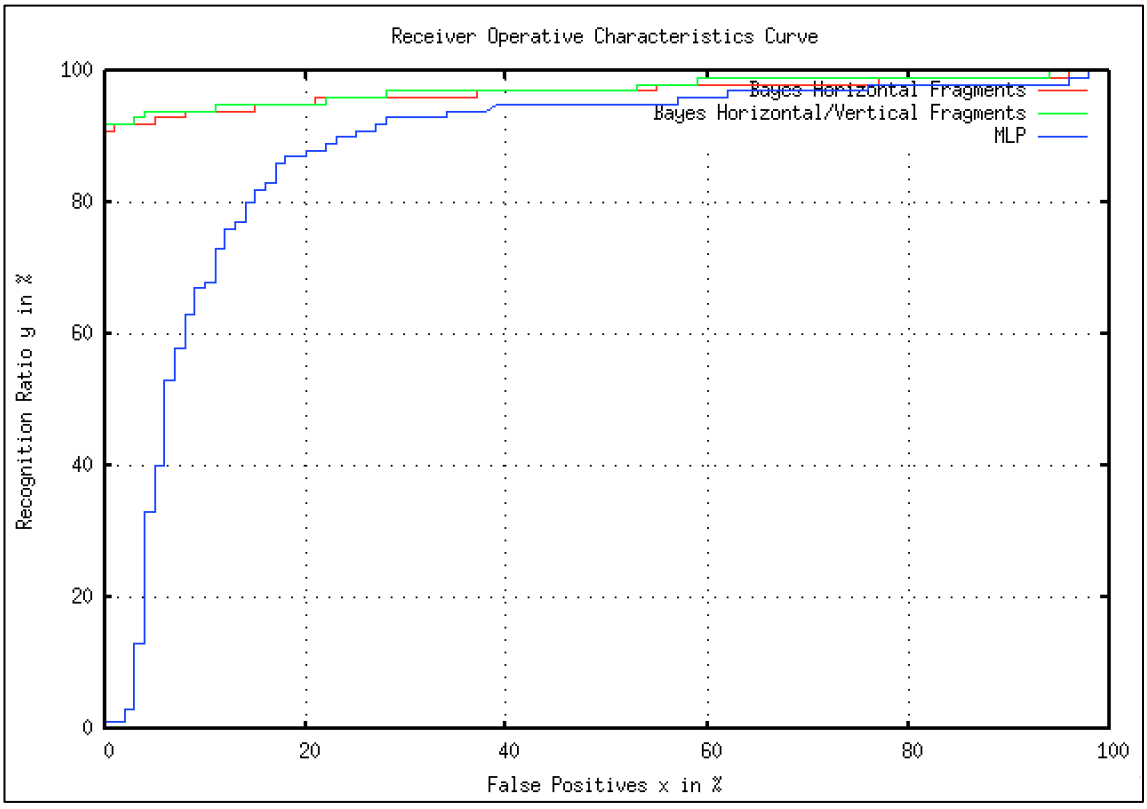
Kennzahl	Wert
Fehlerrate	18,05%
Falsch-Positiv-Rate	13,06%
Recall	76,96%
Precision	85,49%
F-Maß	0,81

Quelle: Eigene Darstellung.

5.6 Bewertung der Ergebnisse

Nachdem zuvor die Ergebnisse der Klassifikatoren im Einzelnen vorgestellt wurden, folgt nun eine ganzheitliche Betrachtung sowie Bewertung der Verfahren. Abbildung 16 zeigt dazu die ROC Kurven aller Klassifikatoren in einem Diagramm.

Abbildung 16: ROC-Kurven aller Klassifikatoren



Quelle: Grafik mit Gnuplot generiert, gnuplot.info (2013), 30. Aug. 2013.

Während sich der Kurvenverlauf für das mehrlagige Perzeptron (blau dargestellt) eher dem der skizzierten Kurve in Abbildung 12 angleicht, beginnen die Kurvenverläufe der beiden Bayes-Klassifikatoren (rot für die horizontale Fragmentierung, grün für die horizontal-vertikal Fragmentierung) erst im oberen Teil des Diagramms.

Im Gesamtvergleich zeigt dies die Überlegenheit der beiden Verfahren nach Bayes bei der Klassifikation von SMS Spam gegenüber dem MLP. Sie liefern bei niedrigeren Falsch-Positiv-Raten deutlich höhere Erkennungsraten und weisen daher bereits ohne weitere Optimierungen sehr gute Ergebnisse auf. Dies beweisen auch die nahezu optimalen F-Maße von 0,94 für das Verfahren mit horizontalen Fragmenten und 0,95 für das Verfahren mit horizontalen und vertikalen Fragmenten.

Trotzdem sind die Ergebnisse des MLP nicht als negativ zu bewerten. Auch wenn sich das Verfahren aufgrund seiner vergleichsweise hohen Falsch-Positiv-Rate von 13,06% noch nicht für den produktiven Einsatz eignet, kann es bereits in seiner jetzigen Form in Kombination mit den anderen Verfahren eingesetzt werden. Auf diese Weise können Synergien aus eventuellen Stärken und Schwächen der einzelnen Verfahren gezogen werden.

Die zuvor bewerteten Ergebnisse der einzelnen Klassifikationsverfahren werden abschließend noch einmal in Tabelle 8 gegenübergestellt.

Tabelle 8: Kennzahlen der drei Klassifikationsverfahren

Kennzahl	Bayes (horizontal)	Bayes (horizontal/vertikal)	MLP
Fehlerrate	5,31%	5,38%	18,05%
Falsch-Positiv-Rate	0,55%	3,16%	13,06%
Recall	89,93%	92,39%	76,96%
Precision	99,4%	96,69%	85,49%
F-Maß	0,94	0,95	0,81

Quelle: Eigene Darstellung.

6 Fazit

Im Folgenden Kapitel werden die erarbeiteten Ergebnisse dieses Arbeitspapieres kurz zusammengefasst. Im Anschluss daran wird ein Ausblick auf Trends und zukünftige Technologien im Bereich der Spamerkennung von Kurznachrichten gegeben.

6.1 Zusammenfassung

In diesem Arbeitspapier wurden, basierend auf dem Konzept des maschinellen Lernens, drei Verfahren zur Erkennung von SMS-Spam entwickelt:

- Bayes-Klassifikation mit horizontaler n-Gramm-Fragmentierung
- Bayes-Klassifikation mit horizontaler und vertikaler n-Gramm-Fragmentierung
- Mehrlagige Perzeptronen mit rein strukturellen Textmerkmalen

Die Bayes-Klassifikation mit horizontaler n-Gramm-Fragmentierung gilt als klassischer Ansatz, der auch häufig zur Erkennung von E-Mail-Spam angewendet wird. Sein als sehr gut zu bewertendes F-Maß von 0,94 zeigt, dass sich dieser Ansatz ohne Modifikation auch auf die Klassifikation von Kurznachrichten übertragen lässt.

Die Bayes-Klassifikation mit horizontalen und vertikalen n-Grammen ist ein gänzlich neuer Ansatz, der im Verlauf dieses Arbeitspapieres entwickelt wurde. Bei diesem Ansatz wird der zu klassifizierende Textinhalt zusätzlich in vertikale n-Gramme zerlegt, mit dem Ziel, mehr klassenspezifische Merkmale aus dem, auf 160 Zeichen begrenzten, Inhalt zu extrahieren. Auch das für den Ansatz ermittelte F-Maß von 0,95 zeigt, dass er sich zur Erkennung von SMS eignet.

Das dritte Verfahren basiert auf einem mehrlagigen Perzeptron zur Klassifikation von SMS. Im Gegensatz zu den anderen beiden Verfahren verwendet es jedoch rein strukturelle Textmerkmale, die sich nicht auf die Inhalte der Nachrichten beziehen. Auch wenn dieses Verfahren mit einem F-Maß von 0,81 als weniger effektiv zu bewerten ist wie die beiden Bayes-Klassifikatoren, weist es mit seinem Wert trotzdem positive Ergebnisse vor. Im produktiven Einsatz könnte dieses Verfahren jedoch aufgrund seiner Fehlerrate von 18,05% als störend empfunden werden. Daher bietet es sich vorerst an, es in Kombination mit anderen Klassifikatoren zu verwenden.

Zusammenfassendist festzuhalten , dass sich alle der in diesem Arbeitspapier entwickelten Verfahren auf die Erkennung von SMS-Spam anwenden lassen. Während die Bayes-Klassifikatoren aufgrund ihrer sehr guten Ergebnisse bereits produktiv eingesetzt werden können, zeigt sich jedoch für das MLP noch Verbesserungsbedarf. Trotz-

dem kann es in seiner aktuellen Form bereits in Kombination mit anderen Klassifikatoren genutzt werden.

6.2 Zusammenfassung

Mobiltelefone gelten mittlerweile als stetige Begleiter und sind aus dem heutigen Alltag nicht mehr wegzudenken.[154] Diese rasante Entwicklung des Mobilfunkmarktes ist ein Beweis dafür, dass das Filtern ungewollter Kurznachrichten auch in Zukunft benötigt wird.[155]

Entwicklungen in diesem Bereich können jedoch in unterschiedlichen Richtungen erfolgen. Eine dieser Richtungen ist beispielsweise die Entwicklung neuer Klassifikationsverfahren. MAHMOUD und MAHFOUZ nutzen in ihrer Ausarbeitung beispielsweise ein künstliches Immunsystem zur Klassifikation, dass dem Abwehrsystem des menschlichen Körpers nachempfunden ist.[156]

Durch die hohe Leistung heutiger Smartphones, sowie deren Offenheit für neue Applikationen, ist es außerdem möglich, die in diesem Arbeitspapier entwickelten Verfahren direkt auf dem Smartphone einzusetzen und somit unabhängig vom Netzbetreiber zu sein. [157] In diesem Bereich wurde mit der Applikation ‚SMSAssassin'bereits eine erste Applikation für Smartphones entwickelt. [158] Damit wird dem Nutzer ermöglicht sich selbst zu schützen.

Zuletzt lassen sich die drei Verfahren auch zur Klassifikation anderer Texte verwenden. Dabei bieten sich insbesondere Nachrichten des Microblogging-Dienstes ‚Twitter' an, da diese Nachrichten auf 140 Zeichen begrenzt und somit der SMS sehr ähnlich sind. [159] Aber auch Blogeinträge und Kommentare lassen sich, den Trainingsdaten entsprechend, in unterschiedliche Klassen separieren.

[154] Vgl. Yadav, K. et al. (2011).
[155] Vgl. Dt.fee.unicamp.br. (o.J.), 03. Dez. 2013.
[156] Vgl. Mahmoud, T.M., Mahfouz, A.M. (2012).
[157] Vgl. Yadav, K. et al. (2011).
[158] Vgl. Yadav, K. et al. (2011).
[159] Vgl. support.twitter.com (2013), 28. Aug. 2013.

Literaturverzeichnis

Abou-Assaleh, T., Cercone, N., Sweidan, R. (2003): N-gram-based detection of new malicious code. In: Proceedings of the 28th Annual International Computer Software and Applications Conference, IEEE CSP, Seiten 10-1109.

Almeida, T., Hidalgo, J.M.G., Silva, T.P. (2013): Towards SMS Spam Filtering: Results under a New Dataset. In: International Journal of Information Security Science, Vol. 2, Nr. 1, Seiten 1–18.

Alpaydin, E. (2008): Maschinelles Lernen, Oldenbuourg Wissenschenschaftsverlag, München.

Bayes, T. (1763): An essay towards solving a problem in the doctrine of chances. In: Phil. Trans. of the Royal Soc. of London, Vol. 53, Seiten 370-418.

Bishop, C.M. (1995): Neural Networks for Pattern Recognition, Oxford University Press, New York.

blog.lookout.com (2012): Security Alert: SpamSoldier, https://blog.lookout.com/blog/2012/12/17/security-alert-spamsoldier, 24. Jul. 2013

Bratko, A., Filipic, B. (2005): Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track. In: Voorhees, E.M., Buckland, L.P. (Hrsg.): TREC Bd. Special Publication Seiten 500-266, National Institute of Standards and Technology (NIST).

Carpinter, J., Hunt, R. (2006): Tightening the net: A review of current and next generation spam filtering tools. In: Computers and Security, Vol. 25, Nr. 8, Seiten 566–578.

Carstensen, K.U., Ebert, C., Endriss, C., Jekat, S., Langer, H., Klabunde, R. (2010): Computerlinguistik und Sprachtechnologie, Spektrum Akademischer Verlag GmbH.

Cavnar, W.B., Trenkle, J.M. (1994): N-Gram-Based Text Categorization. In: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, Seiten 161-175.

Cheng, N., Chen, X., Chandramouli, R., Subbalakshmi, K.P.: Gender identification from E-mails. In: CIDM, IEEE, 2009, Seiten 154–158, URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.158.5183&rep=rep1&type=pdf.

cloudmark.com (2012): GSMA Spam Reporting Service Solutions Guide / Cloudmark, http://www.cloudmark.com/releases/docs/solutionguides/ gsma-srs-solutions-guide-2012-october.pdf, 03. Dez. 2013

computerwelt.at (2012): Happy Birthday: 20 Jahre SMS, http://www.computerwelt.at/news/hardware/smartphone-tablet/detail/artikel/happy-birthday-20-jahre-sms, 25. Jun. 2013

computerworld.com (2012): Android botnet sends SMS spam through Android phones, http://www.computerworld.com/s/article/9234838/Android_botnet_sends_SMS_spam_through_Android_phones?taxonomyId=85&pageNumber=2, 28. Aug. 2013

Cross, F.B., Miller, R.L. (2011): The Legal Environment of Business: Text and Cases: Ethical, Regulatory, Global, and Corporate Issues. South-Western Cengage Learning.

Delany, S.J., Buckley, M., Greene, D. (2012): Review: SMS spam filtering: Methods and data. In: Expert Syst. Appl., Vol. 39, Nr. 10, Seiten 9899–9908.

dt.fee.unicamp.br (o.J.): SMS Spam Collection v.1, http://www.dt.fee.unicamp.br/˜tiago/smsspamcollection, 03. Dez. 2013

Duda, R.O., Hart, P.E., Stork, D.G. (2012): Pattern Classification, Wiley.

European Telecommunications Standards Institute (1996): Digital Cellular Telecommunications System (Phase 2+), Technical realization of the Short Message Service (SMS), Point-to-Point (PP), (ETSI TS 300 901, GSM 03.40 version 5.3.0 Release 1996), December 1998.

Fahrmeir, L., Künstler, R., Pigeot, I., Tutz, G. (2007): Statistik. Springer (limitiert, Springer-Lehrbuch), London.

Fathi, M., Adly, N., Nagi, M. (2004): Web Documents Classification Using Text, Anchor, Title and Metadata Information. In: ISCA International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CSITeA). Cairo. https://cs.uwaterloo.ca/~m2ali/pubs/CSITeA.pdf

Fawcett, T. (2004): ROC Graphs: Notes and Practical Considerations for Researchers / Intelligent Enterprise Technologies Laboratory, Forschungsbericht.

Fisher, R.A. (1922): On the Mathematical Foundations of Theoretical Statistics. In: Philosophical transactions of the Royal Society of London: Mathematical and physical sciences 222 (1922), Seiten 309–368.

Fisher, R.A. (1925): Statistical methods for research workers, Oliver & Boyd, Edinburgh (Biological Monographs and Manuals).

Gigerenzer, G. (2004): Die Evolution des statistischen Denkens. In: Unterrichtswissenschaft, Vol. 32, Nr. 1, Seiten 4-22.

gnuplot.info (2013): Gnuplot, http://www.gnuplot.info, 30. Aug. 2013

Gómez Hidalgo, J.M., Bringas, G.C., Sánz, E.P., García, F.C. (2006): Content based SMS spam filtering. In: Proceedings of the 2006 ACM symposium on Document engineering. New York, NY, USA, ACM, 2006 (DocEng ’06), Seiten 107–114.

Goweder, A.M., Rashed, T., Elbekai, A., Alhammi, H.A. (2008): An Anti-Spam System Using Artificial Neural Networks and Genetic Algorithms. In: Proceedings of the 2008 International Arab Conference on Information Technology, 2008, Seiten 1–8.

gsma.com (2011): http://www.gsma.com/technicalprojects/wp-content/uploads/2012/04/srssmsspamandmobilemessagingattacksthreatsandtrendswp.pdf, 03. Dez. 2013.

Haykin, S. (2009): Neural networks and learning machines, 3. Aufl., Prentice Hall, London.

Heaton, J. (2008): Introduction to Neural Networks for Java, 2. Aufl., Heaton Research.

Junaid, M.B., Farooq, M. (2011): Using evolutionary learning classifiers to do MobileSpam (SMS) filtering. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation. New York, NY, USA, ACM, 2011 (GECCO '11), Seiten 1795–1802.

Kanaris, I., Kanaris, K., Houvardas, I., Stamatatos, E. (2006): Words vs. Character N-grams for Anti-spam Filtering. In: International Journal on Artificial Intelligence Tools.

Kolcz, A., Chowdhury, A., Alspector, J. (2004): The Impact of Feature Selection on Signature-Driven Spam Detection. In: CEAS-The Conference on Email and Anti-Spam.

Koppel, M., Schler, J., Argamon, S. (2009): Computational methods in authorship attribution. In: J. Am. Soc. Inf. Sci. Technol., Vol. 60, Nr. 1, Seiten 9–26.

Kowalski, G. (2010): Information Retrieval Architecture and Algorithms, Springer, USA (Computer science).

Kruse, R., Borgelt, C., Klawonn, F., Möwes, C., Russ, G., Steinbrecher, M. (2012): Computational Intelligence, Springer, Berlin.

Krzanowski, W.J., Hand, D.J. (2009): ROC Curves for Continuous Data. Taylor & Francis (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

Köhler, R., Altmann, G., Piotrovskiˇi, R.G. (2005): Quantitative Linguistik [electronic resource]: Ein internationales Handbuch, Walter de Gruyter GmbH & Company KG (Handbucher Zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science Series).

Linoff, G.S., Berry, M.J. (2011): Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley (IT Pro).

Liu, B. (2007): Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer-Verlag GmbH(Data-Centric Systems and Applications).

Lüdeling, A., Kytö, M. (2009): Corpus Linguistics. De Gruyter.

Mahmoud, T.M., Mahfouz, A.M. (2012): SMS Spam Filtering Technique Based on Artificial Immune System. In: International Journal of Computer Science Issues IJCSI, Vol. 9, Nr. 1, Seiten 589–597.

Mason, J.E. (2009): An N-gram Based Approach to the Automatic Classification of Web Pages by Genre, Halifax, Dalhousie University.

Mcdonald, A. (2005): SpamAssassin, Addison Wesley in Pearson Education Deutschland (Open Source Library).

Miao, Y., Kešelj, V., Milios, E. (2005): Document clustering using character N-grams: A comparative evaluation with term-based and word-based clustering. In: Proceedings of the 14th ACM international conference on Information and knowledge management. New York, NY, USA, ACM, 2005 (CIKM '05), Seiten 357–358.

Miner, G., Elder, J., Hil, T., Delen, D., Fast, A. (2012): Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Academic Press.

Minsky, M.L., Papert, S. (1969): Perceptrons: An Introduction to Computational Geometry. MIT Press.

Muller, C. (1972): Einführung in die Sprachstatistik, Max Hueber Verlag (Hueber Hochschulreiche).

Mulliner, C., Miller, C. (2009): Injecting SMS messages into smart phones for security analysis. In: Proceedings of the 3rd USENIX conference on Offensive technologies. Berkeley, CA, USA, USENIX Association, 2009 (WOOT'09), Seite 5.

Neyman, J., Pearson, E.S. (1928): On the use and interpretation of certain test criteria for purposes of statistical inference. In: Biometrika, Vol. 20, Nr. 1/2, Seiten 175-240 und 263-294.

OECD (2004): Background Paper for the OECD Workshop on Spam / OECD Publishing, Vol. 78.

Patterson, D.W. (1997): Künstliche neuronale Netze: das Lehrbuch, Prentice Hall.

paulgraham.com (2002): A plan for spam, http://paulgraham.com/spam.html, 12. Jun. 2013

Polasek, W. (1994): EDA Explorative Datenanalyse, Springer (Springer- Lehrbuch).

Rafique, M. Z., Alrayes, N., Khan, M.K. (2011): Application of evolutionary algorithms in detecting SMS spam at access layer. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation, New York, NY, USA, ACM, 2011 (GECCO '11), Seiten 1787–1794.

Rafique, M.Z., Farooq, M. (2010): SMS SPAM detection by operating on byte-level distributions using hidden markov models (HMMs). In: Proceedings of the 20th virus bulletin international conference. http://www.nexginrc.org/~zubair.rafique/SMSspam.pdf

Rey, G.D., Wender, K.F. (2010): Neuronale Netze-Eine Einführung in die Grundlagen, Anwendungen und Datenauswertung, 2. vollständig überarbeitete und erweiterte Auflage, Huber Hans, Bern. 5

Robinson, G. (2003): A statistical approach to the spam problem. In: Linux J., Nr. 107, Seite 3.

Rosenblatt, F. (1958): The perceptron: a theory of statistical separability in cognitive systems (Project Para), Cornell Aeronautical Laboratory (Report). –

Runkler, T.A. (2009): Data Mining: Methoden und Algorithmen Intelligenter Datenanalyse, Vieweg Verlag, Friedrich & Sohn Verlagsgesellschaft mbH.

Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. (1998): A Bayesian Approach to Filtering Junk E-Mail. In: Learning for Text Categorization: Papers from the 1998 Workshop. Madison, Wisconsin, AAAI Technical Report WS-98-05.

Salton, G. (1989): Automatic text processing-the transformation, analysis, and retrieval of information by computer, Addison-Wesley, Amsterdam.

Segaran, T. (2008): Kollektive Intelligenz-analysieren, programmieren und nutzen, erste Aufl., O'Reilly Germany, Köln.

Simpson, E.H. (1949): Measurement of diversity. In: Nature, Vol. 163, Nr. 4148, Seite 688.

Sohn, D.-N., Lee, J.-T., Rim, H.-C. (2009): The contribution of stylistic information to content-based mobile spam filtering. In: Proceedings of the ACLIJCNLP 2009

Conference Short Papers. Stroudsburg, PA, USA, Association for Computational Linguistics, 2009 (ACLShort ’09), Seiten 321-324.

sophos.com (2007): Das ’Dreckige Dutzend’: Spam-Versand per SMS nimmt zu, http://www.sophos.com/de-de/press-office/press-releases/2007/04/pr_de_dirtydozapr07.aspx, 12. Jun. 2013

Strelec, H. (1989): Bayes-Statistik und Statistische Qualitätskontrolle. In: Schriftenreihe zur Didaktik der Mathematik der Österreichischen Mathematischen Gesellschaft (ÖMG), Nr. 17, Seiten 129–147.

support.twitter.com (2013): TWITTER: Posten eines Tweets, https://support.twitter.com/articles/495853, 28. Aug. 2013

symatec.com (2012): RESPONSE, Symantec S.: Pikspam: An SMS Spam Botnet, http://www.symantec.com/connect/blogs/pikspam-sms-spam-botnet, 24. Jul. 2013

Topf, J., Etrich, M., Heidrich, J., Romeo, L., Thorbrügge, M., Ungerer, B., BSI (Hrsg.) (2005): Antispam - Strategien. Unerwünschte E-Mails erkennen und abwehren, BSI, Bonn.-BSI-Bundesamt für Sicherheit in der Informationstechnik

Uysal, A.K., Gunal, S., Ergin, S., Gunal, E.S. (2013): The Impact of Feature Extraction and Selection on SMS Spam Filtering. In: Electronics & Electrical Engineering, Vol. 19, Nr. 5, Seiten 67-72.

Weinberger, G. (2009): Identifikation von Spam-Mail mit künstlichen neuronalen Netzen: Entwicklung eines Verfahrens. Igel Verlag (Recht-Wirtschaft-Steuern).

Wendt, D. (2010): Statistische Entscheidungstheorie und Bayes-Statistik. In: Hypothesenprüfung. Enzyklopädie der Psychologie, Themenbereich B, Serie 1, Seiten 471–529.

Xu, C., Chen, Y., Chiew, K. (2010): An Approach to Image Spam Filtering Based on Base64 Encoding and N-Gram Feature Extraction. In: Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, Vol. 1, Washington, DC, USA, IEEE Computer Society, 2010 (ICTAI ’10), Seiten 171-177Xu, Q., Xiang, E.W., YANG, Q., Du, J., Zhong, J. (2012): SMS Spam Detection Using Noncontent Features. In: IEEE Intelligent Systems, Vol. 27, Nr. 6, Seiten 44-51.

Yadav, K., Kumaraguru, P., Goyal, A., Gupta, A., Naik, V. (2011): SMS Assassin: crowdsourcing driven mobile-based system for SMS spam filtering. In: Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, New York, NY, USA, ACM, 2011 (HotMobile ’11), Seiten 1–6.

Zaun, D.P. (1999): Künstliche neuronale Netze und Computerlinguistik, Max Niemeyer, Tübingen.

Zdziarski, J.A. (2005): Ending Spam: Bayesian Content Filtering and the Art of Statistical Language Classification, No Starch Press, San Francisco, CA, USA. –

Zelkowitz, M. (2011): Advances in Computers: Software Development, Elsevier Science.

Zou, K.H., Liu, A., Bandos, A.I. (2011): Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis. CRC Press/Taylor & Francis (A Chapman & Hall book).

Die Publikationsreihe

Schriftenreihe Logistikforschung / Research Paper Logistics

In der Schriftenreihe Logistikforschung des Institutes für Logistik- & Dienstleistungsmanagement (ild) der FOM werden fortlaufend aktuelle Fragestellungen rund um die Entwicklung der Logistikbranche aufgegriffen. Sowohl aus der Perspektive der Logistikdienstleister als auch der verladenden Wirtschaft aus Industrie und Handel werden innovative Konzepte und praxisbezogene Instrumente des Logistikmanagement vorgestellt. Damit kann ein öffentlicher Austausch von Erfahrungswerten und Benchmarks in der Logistik erfolgen, was insbesondere den KMU der Branche zu Gute kommt.

The series research paper logistics within Institute for Logistics and Service Management of FOM University of Applied Sciences addresses management topics within the logistics industry. The research perspectives include logistics service providers as well as industry and commerce concerned with logistics research questions. The research documents support an open discussion about logistics concepts and benchmarks.

Band 1 Klumpp, M. / Bovie, F.: Personalmanagement in der Logistikwirtschaft

Band 2 Jasper, A. / Klumpp, M.: Handelslogistik und E-Commerce [vergriffen]

Band 3 Klumpp, M.: Logistikanforderungen globaler Wertschöpfungsketten [vergriffen]

Band 4 Matheus, D. / Klumpp, M.: Radio Frequency Identification (RFID) in der Logistik

Band 5 Bioly, S. / Klumpp, M.: RFID und Dokumentenlogistik

Band 6 Klumpp, M.: Logistiktrends und Logistikausbildung 2020

Band 7 Klumpp, M. / Koppers, C.: Integrated Business Development

Band 8 Gusik, V. / Westphal, C.: GPS in Beschaffungs- und Handelslogistik

Band 9 Koppers, L. / Klumpp, M.: Kooperationskonzepte in der Logistik

Band 10 Koppers, L.: Preisdifferenzierung im Supply Chain Management

Band 11 Klumpp, M.: Logistiktrends 2010

Band 12 Keuschen, T. / Klumpp, M.: Logistikstudienangebote und Logistiktrends

Band 13 Bioly, S. / Klumpp, M.: Modulare Qualifizierungskonzeption RFID in der Logistik

Band 14 Klumpp, M.: Qualitätsmanagement der Hochschullehre Logistik

Band 15 Klumpp, M. / Krol, B.: Das Untersuchungskonzept Berufswertigkeit in der Logistikbranche

Band 16 Keuschen, T. / Klumpp, M.: Green Logistics Qualifikation in der Logistikpraxis

Band 17 Kandel, C. / Klumpp, M.: E-Learning in der Logistik

Band 18 Abidi, H. / Zinnert, S. / Klumpp, M.: Humanitäre Logistik – Status quo und wissenschaftliche Systematisierung

Band 19 Backhaus, O. / Döther, H. / Heupel, T.: Elektroauto – Milliardengrab oder Erfolgsstory?

Band 20 Hesen, M.-A. / Klumpp, M.: Zukunftstrends in der Chemielogistik

Band 21 Große-Brockhoff, M. / Klumpp, M. / Krome, D.: Logistics capacity management – A theoretical review and applications to outbound logistics

Band 22 Helmold, M. / Klumpp, M.: Schlanke Prinzipien im Lieferantenmanagement

Band 23 Gusik, V. / Klumpp, M. / Westphal, C.: International Comparison of Dangerous Goods Transport and Training Schemes

Band 24 Bioly, S. / Kuchshaus, V. / Klumpp, M.: Elektromobilität und Ladesäulenstandortbestimmung – Eine exemplarische Analyse mit dem Beispiel der Stadt Duisburg

Band 25 Sain, S. / Keuschen, T. / Klumpp, M.: Demographic Change and ist Effect on Urban Transportation Systems: A View from India

Band 26 Abidi, H. / Klumpp, M.: Konzepte der Beschaffungslogistik in Katastrophenhilfe und humanitärer Logistik

Band 27 Froelian, E. / Sandhaus, G.: Conception of Implementing a Service Oriented Architecture (SOA) in a Legacy Environment

Band 28 Albrecht, L. / Klumpp, M. / Keuschen, T.: DEA-Effizienzvergleich Deutscher Verkehrsflughäfen in den Bereichen Passage/Fracht

Band 29 Meyer, A. / Witte, C. / Klumpp, M.: Arbeitgeberwahl und Mitarbeitermotivation in der Logistikbranche

Band 30 Keuschen, T. / Klumpp, M.: Einsatz von Wikis in der Logistikpraxis

Band 31 Abidi, H. / Klumpp, M.: Industrie-Qualifikationsrahmen in der Logistik

Band 32 Kaiser, S. / Abidi, H. / Klumpp, M.: Gemeinnützige Kontraktlogistik in der humanitären Hilfe

Band 33 Abidi, H. / Klumpp, M. / Bölsche, D.: Kompetenzen in der humanitären Logistik

Band 34 Just, J. / Klumpp, M. / Bioly, S.: Mitarbeitermotivation bei Berufskraftfahrern – Eine empirische Erhebung auf der Basis der AHP-Methode

Band 35 Keinhörster, M. / Sandhaus, G.: Maschinelles Lernen zur Erkennung von SMS-Spam

Die 1993 von Verbänden der Wirtschaft gegründete staatlich anerkannte gemeinnützige FOM Hochschule verfügt über 30 Studienorte in Deutschland.

Als praxisorientierte Hochschule fördert die FOM den Wissenstransfer zwischen Hochschule und Unternehmen. Dabei sind alle wirtschaftswissenschaftlichen Studiengänge der FOM auf die Bedürfnisse von Berufstätigen zugeschnitten. Die hohe Akzeptanz der FOM zeigt sich nicht nur in der engen Zusammenarbeit mit staatlichen Hochschulen, sondern auch in zahlreichen Kooperationen mit regionalen mittelständischen Betrieben sowie mit internationalen Großkonzernen. FOM-Absolventen verfügen über solide Fachkompetenzen wie auch über herausragende soziale Kompetenzen und sind deshalb von der Wirtschaft sehr begehrt.

Weitere Informationen finden Sie unter **fom.de**

Institut für Logistik- & Dienstleistungsmanagement
der FOM University of Applied Sciences

Das Ziel des ild Institut für Logistik- & Dienstleistungsmanagement ist der konstruktive Austausch zwischen anwendungsorientierter Forschung und Betriebspraxis. Die Wissenschaftler des Instituts untersuchen nachhaltige und innovative Logistik- und Dienstleistungskonzepte unterschiedlicher Bereiche, initiieren fachbezogene Managementdiskurse und sorgen zudem für einen anwendungs- und wirtschaftsorientierten Transfer ihrer Forschungsergebnisse in die Unternehmen. So werden die wesentlichen Erkenntnisse der verschiedenen Projekte und Forschungen unter anderem in dieser Schriftenreihe Logistikforschung herausgegeben. Darüber hinaus erfolgen weitergehende Veröffentlichungen bei nationalen und internationalen Fachkonferenzen sowie in Fachpublikationen.

Weitere Informationen finden Sie unter **fom-ild.de**

ISSN 1866-0304