Unsere Machbarkeitsstudie *Technik für den digitalen Jugendschutz: Automatische Erkennung von Sexting und Cybergrooming* von 2018 hat untersucht, wie Methoden des maschinellen Lernens dabei helfen können, den Schutz von Minderjährigen in digitalen Räumen zu verbessern. Dabei ging es um zwei Fragestellungen: Können Sexts, also erotische bzw. sexualisierte Selfies, auf Smartphones automatisiert erkannt und somit potenziell auch verhindert werden? Kann Cybergrooming automatisiert erkannt werden, indem das Alter der Chatpartner erkannt oder ein bereits auffällig gewordener Teilnehmer wiedererkannt wird? Beide Themen stehen im Zusammenhang: Sexts von Minderjährigen können das Ergebnis von erfolgreichem Cybergrooming sein, bei dem ein Erwachsener einen Minderjährigen in einem Chat unter Vortäuschung einer falschen Identität zu entsprechenden Aufnahmen überredet. Beide Sachverhalte sind also Aspekte der Gefährdung Minderjähriger, wobei Sexts von Minderjährigen auch ohne Cybergrooming entstehen können und eine Gefährdung dann primär durch eine unkontrollierte Weiterverbreitung erfolgen kann.

Maschinelles Lernen im Jugendschutz

TEXT: MARTIN STEINEBACH

Lokale Sexting-Erkennung durch maschinelles Lernen

Die von uns untersuchten Mechanismen zum Aufdecken und Eindämmen solcher Phänomene stammen aus der auf Deep Learning basierten Bildverarbeitung sowie aus dem Natural Language Processing (NLP). Beides sind Ausprägungen des maschinellen Lernens, welche in den vergangenen Jahren eine signifikante Steigerung ihrer Leistungsfähigkeit erfahren haben. Tatsächlich sind die oben genannten Fragestellungen erst durch maschinelles Lernen ausreichend erfolgreich adressierbar geworden: So gibt es schon lange Ansätze, Nacktheit durch Algorithmen automatisiert zu erkennen. Diese sind immer wieder daran gescheitert, dass Bilder komplex und unstrukturiert sind. Ein Ansatz, der beispielsweise Pixel in einem Bild zählt, welche Hautfarben darstellen, kann auf vielfältige Weise scheitern: Bilder können durch Über- und Unterbelichtung untypische Farben haben, Haut erscheint dann plötzlich eher weißlich grau statt rosa. Unproblematische Strandfotos können einen hohen Anteil an Hautfarben aufweisen. Pornografische Inhalte müssen nicht unbedingt einen großen Anteil nackter Haut aufweisen. Natürlich kann hier versucht werden, durch eine Analyse der Bilder oder das Erkennen von Bildobjekten eine Verbesserung der Erkennungsleistung zu erreichen. In der Praxis war vor maschinellem Lernen das Erkennen von Nacktheit durch Software aber immer unzureichend und wies hohe zweistellige Fehlerraten auf.

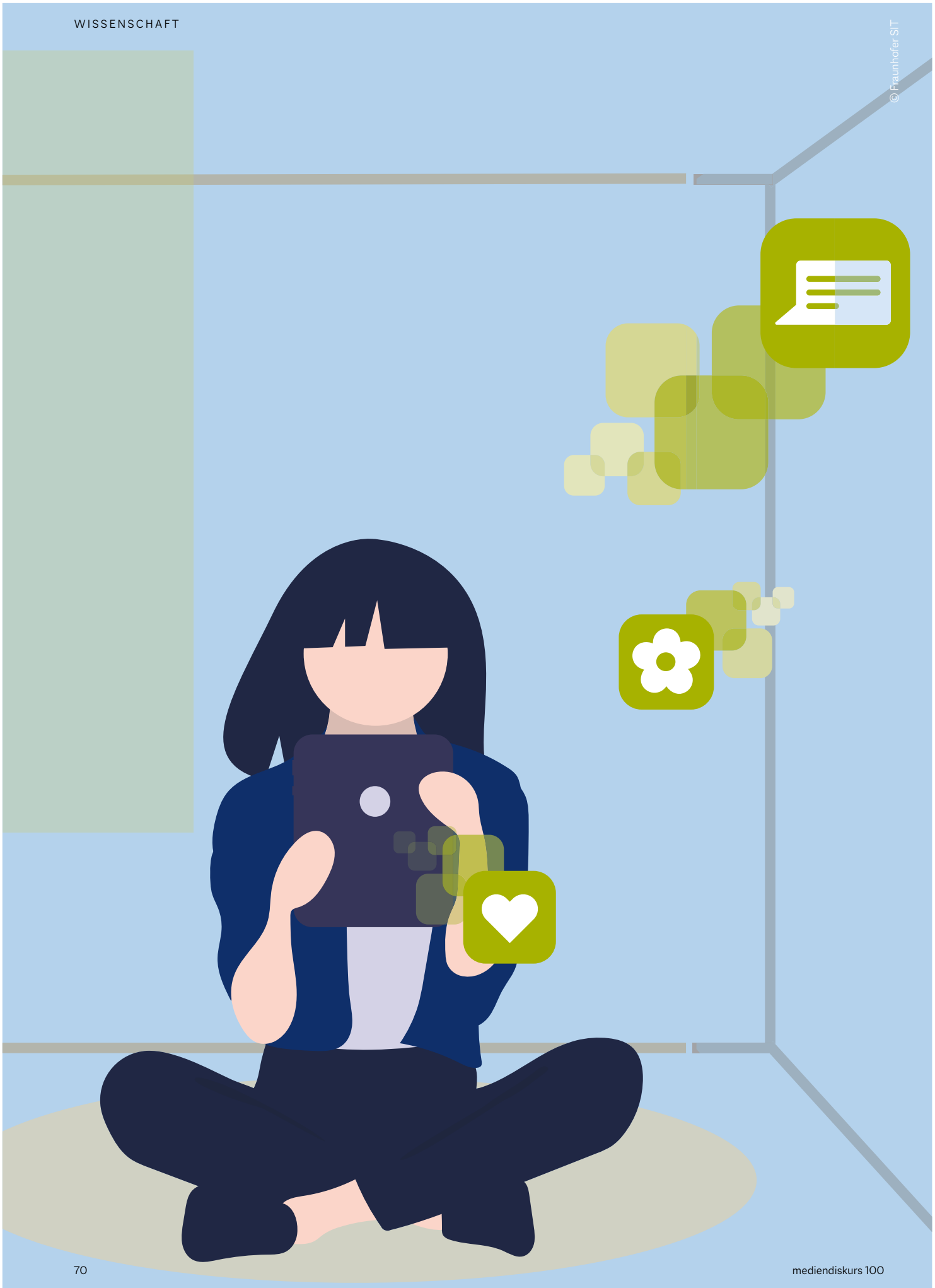
Durch maschinelles Lernen wurden hier sehr schnell Verbesserungen erzielt, bei denen die Fehlerraten im niedrigen einstelligen Bereich sind. Die Modelle, die neuronale Netze von vorgegebenen Beispielbildern ableiten, sind besser in der Lage, auch bei komplexen Bildern mit Störungen zurechtzukommen und korrekte Einschätzungen abzugeben. So können entsprechende Verfahren inzwischen sehr gut einen Menschen vom Bildhintergrund abgrenzen. Sie unterscheiden beispielsweise erfolgreich eine Strandszene von einer Aufnahme eines Schlafzimmers, weil sie für beide Ausprägungen genug Beispiele mit entsprechender Annotation zur Verfügung haben.

In unserer Studie war die Aufgabe ausschließlich, das von der Smartphone-Kamera erzeugte Foto hinsichtlich von Nacktheit zu analysieren. Die Grundidee war dabei, dass Eltern ihren minderjährigen Kindern zwar ein Smartphone zur Verfügung stellen, auf diesem aber vorher Mechanismen aktivieren, welche Nacktaufnahmen verhindern. Die Kinder würden somit daran gehindert, Nacktaufnahmen von sich zu erstellen und zu verbreiten. Technisch ist diese Erkennung hier allerdings nicht auf die Kinder beschränkt: Auch der Versuch, eine andere nackte Person zu fotografieren, würde scheitern – ebenso wie das Abfotografieren von Bildern aus Magazinen oder von Kunstwerken, die Nacktheit zeigen. Allerdings ist es natürlich möglich, das Erkennen um einen biometrischen Faktor zu erweitern und damit dem Smartphone beizubringen, nur Fotos von Personen zu verhindern, die es vorher angelernt bekommen hat.

Grenzen der Automatisierung

Das Erkennen von Nacktheit ist in manchen Fällen nicht ausreichend. Würde man versuchen, mit einem entsprechenden Verfahren allgemein kinder- oder jugendpornografische Inhalte zu erkennen, müsste auch das Alter der auf den Bildern gezeigten Personen korrekt geschätzt werden. Erst aus der Kombination von Nacktheit und dem Unterschreiten einer Altersgrenze würde dann eine Erkennung entsprechender Inhalte entstehen. Und die Herausforderungen sind oft noch komplexer. So muss in manchen Fällen erkannt werden, ob die Bildinhalte eine erotische Aussage haben, was wiederum ein interpretierendes Verständnis von Bildern voraussetzt. Eine Unterscheidung, ob auf einem Foto ein harmloser Badeurlaub oder Missbrauch gezeigt wird, ist sonst für eine automatische Einschätzung nur schwer zu erreichen und somit fehleranfällig. Dabei würden Fehler eben genau in diesen Grenzbereichen auftreten: Bilder, die nicht oder nur wenig bekleidete Kinder zeigen.

An dieser Stelle entsteht dann ein grundlegendes Problem, welches je nach Design der Anwendung, in der die Erkennung eingesetzt wird, substanzielle Auswirkungen haben kann. In der jüngeren Vergangenheit gab es eine Reihe von Aktivitäten, die anstrebten, Kinderpornografie automatisiert auf Endgeräten zu erkennen. Im Gegensatz zu dem Vorgehen in unserer Studie war hier aber der nächste Schritt, verdächtige Bilder von Menschen beurteilen zu lassen. Die Lösung war also kein Filter, der anstrebt, Konsum und Verbreitung illegalen Bildmaterials zu unterbinden, sondern ein moderierter Meldemechanismus, der letztendlich mit der Übergabe der Bilder als Beweismittel bei Polizeibehörden enden sollte. Demzufolge steigt das Risiko, dass fremde Personen private Aufnahmen begutachten, abhängig von den Motiven der Fotos, die der Besitzer eines entsprechend ausgestatteten Smartphones macht. Der Skiurlaub wird eher keinen Alarm auslösen, ein Strandurlaub vielleicht schon eher.



Vielfältige Herausforderungen

Die Erkennung von Kinderpornografie wird bereits lange automatisiert betrieben, insbesondere auch von der Polizei im Rahmen der massenhaften Sichtung von Datenträgern. Dabei muss aber unterschieden werden, ob es sich um ein Wiedererkennen bereits bekannter Bilder anhand von Hashverfahren handelt oder um ein Einordnen von vorher unbekannten Bildern als Kinderpornografie auf Basis von maschinellem Lernen. Die Fehlerraten sind hier sehr unterschiedlich. Bei Hashverfahren kann je nach Methode ein Fehler ausgeschlossen werden oder die Fehlerrate liegt im Promillebereich. Maschinelles Lernen weist hier noch immer Fehler im Bereich von Prozenten auf. Bei einer großen Menge von Bildern führt dies schnell zu einer hohen Wahrscheinlichkeit zahlreicher Fehler.

Erwähnt werden muss auch, dass maschinelles Lernen bzw. die Klassifizierung durch maschinelles Lernen angreifbar ist. Durch spezielle Verfahren lässt sich die Einordnung eines trainierten Netzes verschieben. Ein Bild, welches einen Hund zeigt, wird dann als Bild einer Katze erkannt. Dies geschieht durch das gezielte Hinzufügen von Rauschmustern, welche so aufgebaut sind, dass sie eine Fehlerkennung auslösen. Mit solchen Angriffen wäre es möglich, sowohl harmloses Material als Kinderpornografie als auch Kinderpornografie als harmlos einordnen zu lassen. Hier ist allerdings der Zugang zum trainierten Netz notwendig, welches zur Einordnung verwendet wird. Dieser kann bei einer lokalen Lösung auf einem Smartphone leichter erreicht werden als bei einer Serverlösung.

Cybergrooming

Neben der bisher diskutierten Erkennung von Nacktheit war auch Cybergrooming ein Thema der Studie. Hier können verschiedene Ansätze verfolgt werden, die alle auf NLP basieren, welches in unserem Fall die Vereinigung aus computergestützter Linguistik und maschinellem Lernen darstellt.

Zum einen kann versucht werden, durch Analyse von in einem Chat geschriebenen Texten das Alter des Verfassers einzuschätzen. Dies geschieht durch Autorenprofiling. Der Computer lernt hier, wie Vertreter verschiedener Altersgruppen üblicherweise schreiben, und kann dann vorliegende Texte mit dem Erlernten vergleichen. Einfluss können hier beispielsweise der Wortschatz oder die Komplexität der Sätze haben. Eine grobe Einordnung genügt hier, denn die Aufgabe ist es, Angaben in einem Nutzerprofil mit der Einschätzung durch die Maschine zu vergleichen. Eine Person, die vorgibt, 12 zu sein, aber von der Maschine als Mitte 40 eingeordnet wird, ist auffällig. Ob ein Kind 12 oder 14 Jahre alt ist, sich die Maschine also um wenige Jahre irrt, wird in den meisten Fällen unwichtig sein. Eine relevante Grenze könnte hier höchstens die Volljährigkeit mit 18 Jahren darstellen. Das in unserer Studie umgesetzte Verfahren verschätzt sich um rund fünf Jahre, was in vielen Fällen ausreichend genau sein dürfte.
Zum anderen kann aber auch versucht werden, eine bestimmte Person anhand ihres Schreibstils zu erkennen. Das ist dann die Autorschaftsattribution. Das Szenario wäre hier, eine Person, die bereits aus einer Chatplattform ausgeschlossen wurde und sich unter einem falschen Namen wieder anmeldet, anhand ihres Schreibstils wiederzuerkennen. Hier ist eine höhere Genauigkeit notwendig als bei der Einschätzung des Alters, da es ja um die Erkennung einer bestimmten Person geht. Der von

uns untersuchte Ansatz wies Fehler zwischen 5 und 20 % auf. Bei einer großen Anzahl von Nutzern kann dies zu einer problematischen Fehleinschätzung und demzufolge zu zahlreichen Fehlalarmen führen. Eventuell kann dieses Verfahren aber eingesetzt werden, um Verdachtsfälle in einem Forum zu betrachten: Löst eine Alterserkennung Alarm aus, kann mit der Attribution geprüft werden, ob die stilistischen Merkmale zu einem bereits auffällig gewordenen ehemaligen Teilnehmer passen. Ebenso könnte hiermit der Verdacht weiterverfolgt werden, dass sich ein Nutzer parallel unter mehreren Namen im Forum anmeldet, um seine Aktivitäten zu verschleiern.

Einsatzmöglichkeiten

Ob die oben genannten Methoden dabei helfen können, Minderjährige im digitalen Raum zu schützen, ist von zahlreichen Faktoren abhängig. Insbesondere sollte neben der technischen Machbarkeit auch der Datenschutz betrachtet werden. Ob beispielsweise der Kampf gegen Kindesmissbrauch einen Zugriff auf eigentlich private Aufnahmen auf den Endgeräten von Nutzern rechtfertigt, ist eine ethische und juristische Frage, die Entwickler von Technologie nicht beantworten können und auch nicht sollten. Was aber die Technik beisteuern kann, sind Alternativen und Handreichungen zum besseren Verständnis der technischen Grenzen und Risiken.

Unsere Studie hat beispielsweise gezeigt, dass eine Erkennung von Nacktheit problemlos lokal auf einem Smartphone durchgeführt werden kann. Ein Hochladen auf einen Server zur Analyse ist nicht notwendig. Das ermöglicht einen deutlich höheren Grad an Privatheit. Ebenso kann die Erkennung aber auch in einem Forum (und damit dann wieder auf einem Server) geschehen und dazu verwendet werden, Nacktaufnahmen, die von Nutzern hochgeladen und geteilt werden, zu unterdrücken. Der erste Ansatz verhindert dann das Entstehen von Sexts, der zweite nur deren Verbreitung. Die Erkennung von Cybergrooming hingegen wird eher auf den Plattformen der Betreiber umsetzbar sein. Nur dort können Beiträge eines Nutzers aus mehreren Chats zusammengefügt werden, um eine gute Grundlage für eine Analyse hinsichtlich des Alters zu bieten. Und nur hier könnten stilistische Eigenschaften eines Autors zur Wiedererkennung gespeichert werden.

Zusammenfassend kann festgehalten werden: Maschinelles Lernen kann auf vielfältige Weise helfen, den Schutz von Minderjährigen im digitalen Raum zu erhöhen. Je nach Ansatz erfordert dies die Integration in Endgeräte oder Plattformen. Ein technisches Verständnis der Möglichkeiten und Grenzen der Verfahren ist eine notwendige Grundlage für eine Diskussion, die allerdings keinesfalls rein technisch geführt werden sollte.

Prof. Dr. Martin Steinebach ist Abteilungsleiter für Media Security und IT Forensics am Fraunhofer SIT. Er studierte Informatik und promovierte mit dem Thema digitaler Audiowasserzeichen. Steinebach leitet zahlreiche Projekte zu IT-Forensik und Big-Data-Sicherheit.