Wie die KI Fakes aufspürt – Die künstlich-intelligente Jagd nach Fake News

Beim alltäglichen Scrollen durch Social Media zeigen sich Fotos von Bekannten, Infoposts, Comedy, und…der Papst im „Balenciaga" Mantel? Da stimmt doch etwas nicht. Doch wie lässt sich nun überprüfen, ob der Papst tatsächlich auf Designermode umgestiegen ist, oder es sich schlichtweg um eine überzeugende Fälschung handelt? Und welche Rolle spielt KI bei der Entlarvung von Fakes im Internet?

Das mithilfe von Künstlicher Intelligenz (KI) generierte Bild von Papst Franziskus, gekleidet in einer großen Pufferjacke der Luxusmarke „Balenciaga" machte weltweit Schlagzeilen. Es ist ein Beispiel für die Reichweite und Aufmerksamkeit, die Falschinformationen – Neudeutsch Fake News – weltweit erreichen können. Das Foto wurde millionenfach geteilt und erreichte zahlreiche Menschen, die den neuen Stil des Papsts sofort für echt hielten. Als sich das Bild als KI-Fake entpuppte, war die Sorge um die Gefahr von Künstlicher Intelligenz für das Erschaffen von Fake News groß. Doch dass KI auch Tools bereitstellen kann, die Falschinformationen als solche identifizieren können, zeigt etwa das Forschungsprojekt *defalsif-AI*.

Entstanden ist das Projekt *defalsif-AI* („Detektion von Falschinformation mittels AI") im Rahmen einer Förderung des Sicherheitsforschungs-Förderprogrammes KIRAS durch das österreichische Bundesministerium für Finanzen. Nach einer Laufzeit von 24 Monaten wurde es im September 2022 abgeschlossen. Es stellt einen wesentlichen Baustein eines Forschungsschwerpunktes am Austrian Institute of Technology (AIT) dar. Unter den Projektpartnern finden sich das Bundeskanzleramt, das Außenministerium, das Verteidigungsministerium, der ORF und die Austria Presseagentur. Bei der Projekthälfte wurde das Tool für die Projektpartner erstmals zum Testen freigegeben – für Privatpersonen ist es allerdings noch nicht nutzbar.
Zu beschreiben ist das Ergebnis des Projekts als eine Art Werkzeugkoffer, der verschiedenste Tools zur Ermittlung von Falschinformationen enthält.
Die Funktionsweise lasse sich allerdings nicht mit herkömmlichen Verifizierungsmaßnahmen vergleichen, sagt defalsif-AI Projektleiter Martin Boyer.
„Wenn man es gewöhnt ist, mit Personenregister, Suchmaschinen, Bildrückwärtssuche und Kartendiensten zu arbeiten, kommt man vielleicht nicht auf

die Idee, dass es einen Algorithmus gibt, dem man ein Bild zeigt, und allein aufgrund der visuellen Pixelinformation schätzt das Analysemodul einen Punkt am Globus, wo dieses Bild aufgenommen wurde. Also keine Metadaten, keine Rückwärtssuche, sondern der Algorithmus wurde auf Millionen von Bildern trainiert, und schätzt eine Geoposition aufgrund der Pixel."
Weiters gebe es unter anderem ein Analysemodul, das künstlich generierte Gesichter als solche erkennen soll, sowie eines, das Texte auf ihren Wahrheitsgehalt überprüft. Wurde in einem Text beispielsweise viel mit hasserfüllter Sprache, subjektivem Schreibstil und politischen Claims gearbeitet, erkennt das Tool diese als Indikatoren für Fake News, erzählt Boyer. Das Tool warne dann: *Möglicherweise handelt es sich um einen Text, der Falschinformationen enthält.*

Schutz der Demokratie vor Fake News

Florian Schmidt von der Austria Presseagentur (APA) ist Faktenchecker und nutzt selbst verschiedenste KI-Technologien zur Verifizierung von Inhalten. KI-gestützte Tools wie Google Lens bieten besonders bei der Rückwärtssuche von Bildern Vorteile. Sie können nicht nur identische Bilder aufspüren, sondern auch einzelne Elemente innerhalb eines Bildes erkennen und in anderer Form im Internet finden. Von großer Wichtigkeit sei KI in Schmidt's Arbeitsumfeld allerdings noch nicht:
„Ich würde nicht sagen, dass das derzeit so eine große Rolle spielt, weil diese Tools teilweise leider noch sehr unzuverlässig sind. Also, es gibt schon sehr viele Anbieter, die versuchen, mit KI-gestützten Tools etwas anzubieten, mit dem man Fakes überführen kann. Aber meistens ist das nicht so, dass ich da irgendwas eingebe, und dann zu 100 Prozent sicher sein kann, dass das wirklich korrekt überführt worden ist, oder verifiziert worden ist. Für uns als Faktenchecker ist vor allem Transparenz wichtig, damit wir sehen, wir sehen, wie ein Tool zu einer Entscheidung gekommen ist und welche Prozesse stattfinden."
Einen solchen Versuch hat die APA im Rahmen der Zusammenarbeit mit dem AIT während der Entwicklung des Forschungsprojekts *defalsif-AI* unterstützt. Dabei lag die Hauptaufgabe von Schmidt und seinen Kolleg*innen vor allem im Schildern von Erfahrungen aus der Redaktion, auf deren Basis Anforderungen für das Tool definiert werden konnten.
Das Projektziel von defalsif-AI geht über das Bereitstellen eines Tools zum Faktenchecken hinaus – es hat zum Ziel, die Demokratie zu schützen. „Ich würde

sagen, gesicherte Informationen sind immer ein Beitrag zur Demokratie, weil natürlich Falschinformationen vor allem darauf abzielen, dass sie den Meinungsbildungsprozess von einer Publikumsgruppe beeinflussen“, so Schmidt. Vor allem von Regierungen gesteuerte Falschinformationen, etwa in Diktaturen, zielten oft durch Desinformationskampagnen darauf ab, eine Bevölkerung etwas Falsches glauben zu lassen. „Natürlich kann ich mir nur eine wirklich starke Meinung bilden, wenn ich gesicherte Informationen habe. Weil sonst glaube ich vielleicht irgendetwas, das gar nicht stimmt. Insofern würde ich schon sagen, dass das ein Schutz von Demokratie ist, wenn wir mithelfen, Falschinformationen zu überführen.“, schildert er.

Keine Chance ohne Media Literacy

Neben dem Schutz der Demokratie liegt das Ziel des Forschungsprojekts laut Boyer bei der Unterstützung der Menschen. Als unterstützendes Tool sollen die Anwendungen durch Einschätzungen und Indikatoren auf mögliche Falschinformationen hinweisen.

Nicht aber ließen sich die Tools mit einem Richter gleichsetzen, der ein klares Urteil über die Authentizität eines Fotos, Videos oder Textes fällen kann. „Man kann es sich nie so vorstellen, dass diese Tools sagen: *Schwarz. Weiß. Das ist manipuliert. Das nicht*. Sondern diese Tools selbst geben eine Einschätzung ab: *Möglicherweise manipuliert*. Oder *Mit dieser Wahrscheinlichkeit schätze ich das.“*

Um zukünftig die Fähigkeit der Menschen zu unterstützen, Falschinformationen als solche erkennen zu können, reiche es daher nicht, sich ausschließlich auf Künstliche Intelligenz zu verlassen. Es bedürfe vor allem einer gut ausgeprägten Media Literacy, oder Medienkompetenz, die auch eine kritische Herangehensweise an Nachrichten voraussetzt.

Die Fähigkeit, mit der enormen Informationsfülle, mit der man vor allem in Zeiten von Social Media täglich konfrontiert ist, umgehen zu können, sei selbst mit der Hilfe von KI von absoluter Wesentlichkeit, meint Boyer. Eine endgültige Verifizierung sei trotz KI ohne den Menschen selbst nicht möglich.

Dazu kommt laut Boyer die Aufnahmefähigkeit oder Aufmerksamkeitsschwelle, die vor allem bei der Nutzung von Social Media nebenbei eingeschränkt sein kann:

„Wenn ich jetzt sage *‘Sehen Sie hin, dieses Bild ist generiert‘,* dann werden alle suchen, wie beim Zehn Fehlersuchbild. Das ist ja wunderbar, aber in unserer täglichen

Auffassung scrollen wir da drüber und sagen *‘Ah, okay.‘* Und es hat auch nicht immer diese Relevanz, das muss man schon dazusagen. Es kommt nicht immer über diese Aufmerksamkeitsschwelle. Wenn ich in der U-Bahn sitze, werde ich einen Artikel anders konsumieren, als wenn ich zu Hause am Arbeitstisch sitze und mich konzentriere. Das ist nun mal so. Das ist ein Aspekt.“

Und jetzt?

Bis heute steht *defalsif-AI* ausschließlich den Projektpartnern*innen zur Verfügung. Laut Boyer ist jedoch geplant, das Tool weiterzuentwickeln, und in diesem Zuge auch zur Nutzung zu bringen. „Der Auftrag des AIT ist ja auch, dass wir die österreichischen Unternehmen – die österreichische Industrie beispielsweise – mit Innovation beleben. Da ist natürlich unser Ansporn, dass wir zum Beispiel Teile daraus in die Industrie bringen, und in die Umsetzung bringen. Das ist auf jeden Fall der nächste Schritt.“ Auch Faktencheckerinnen und Faktencheckern soll das Tool zukünftig als Unterstützung angeboten werden. Schmidt kann sich ebenfalls vorstellen, dass das Projektergebnis einmal in die Verwendung kommen können wird. Bis dahin brauche es allerdings noch einiges an Entwicklung. Das liege allerdings nicht daran, dass nicht gut gearbeitet worden sei. „Wir haben das wirklich sehr gut entwickelt, und das meiste von diesen Tools ist „State of the Art“, also das Beste, was derzeit technisch möglich ist. Aber es reicht eben derzeit nicht aus, damit man das einfach einem Redakteur oder gar einer Privatperson in die Hand drückt.“ Einige Tools aus dem „Werkzeugkoffer“ seien jedoch durchaus bereits verwendbar.

Boyer sieht Künstliche Intelligenz in der Gesellschaft gerade am Anfang. Gerade beginnen die Menschen, Anwendungen der KI auszuprobieren und zu sehen, was in dieser Richtung möglich sei. In seiner Prognose für die zukünftige Entwicklung der KI spricht Boyer als Beispiel auch von ChatGPT – einem Chatbot, der aufgrund seiner Funktionen und Nutzung in den letzten Monaten weltweit Schlagzeilen machte. „Denken Sie daran, wie gut ChatGPT *jetzt* funktioniert. Was wird in zehn Jahren sein? Da werden die Dinge ja weiterentwickelt sein. Und das gilt auch für Tools wie die, an denen wir arbeiten. Es ist eben ein Prozess, und unser Projekt – dieses „Werkzeugset“ – war sozusagen der erste Schritt. Natürlich glaube ich, dass das in

Zukunft uns allen zur Verfügung stehen wird. Es ist wichtig, dieses Thema voranzutreiben, und auch die Forschung daran.“

Mia Weisz