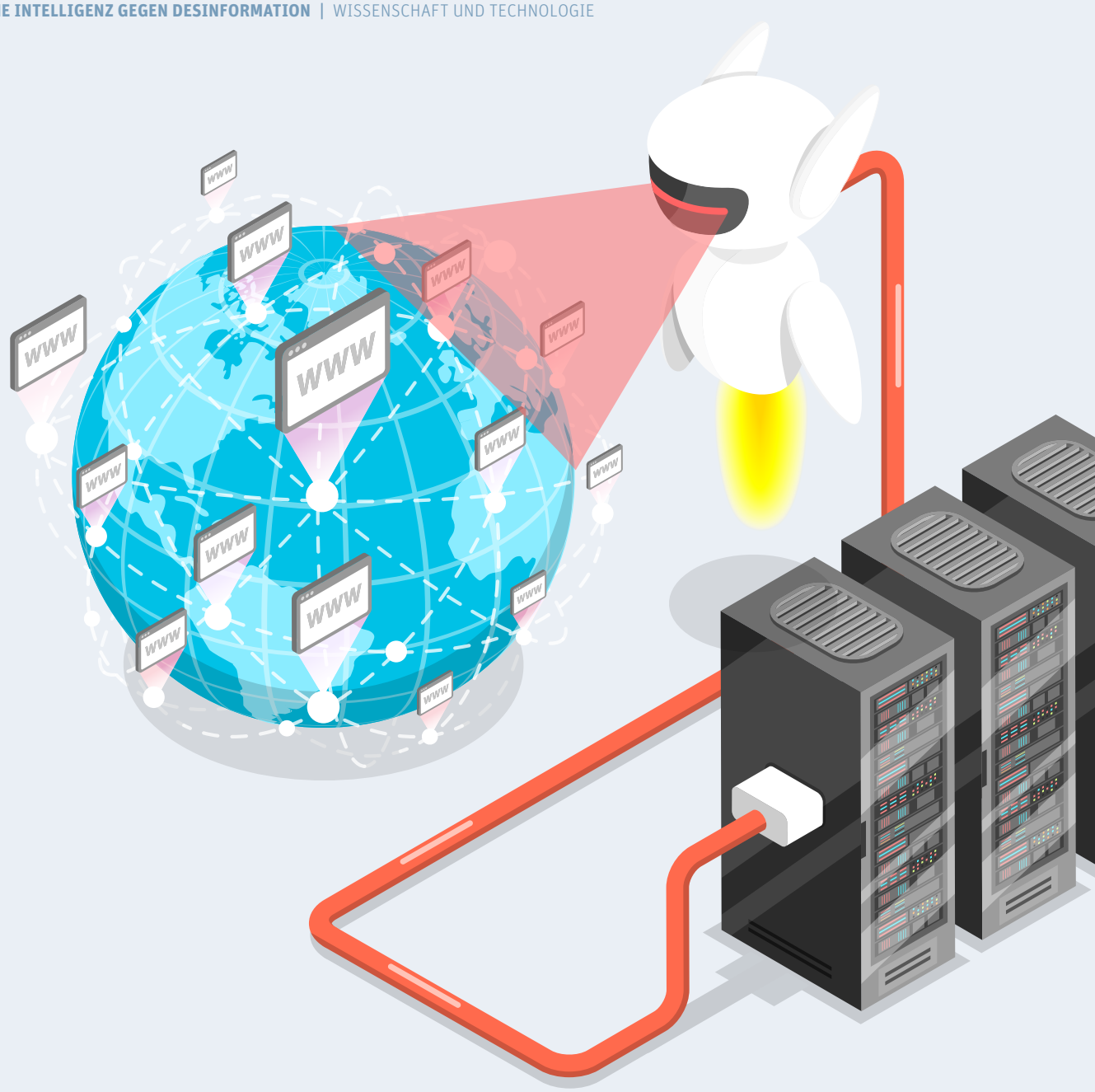


Künstliche Intelligenz im Einsatz gegen Desinformation in sozialen Netzwerken

Soziale Netzwerke wie Twitter und Facebook sind allgegenwärtig und aus unserer digitalen Welt nicht mehr wegzudenken. Sie werden jedoch häufig als Plattform für Desinformationskampagnen genutzt, die ein Land destabilisieren können. Wie können Fehlinformationen angesichts der heutigen Informationsflut erkannt werden? Künstliche Intelligenz (KI) bietet hier innovative Lösungen.

Text: Dr. Gérôme Bovet und Sarah Frei

Tagtäglich werden Millionen, ja sogar Milliarden von Bildern und Texten in den sozialen Medien geteilt. Der Grossteil dieser Informationen ist harmlos – es handelt sich um Fotos aus dem Urlaub, von der Katze oder einfach um den Ausdruck der eigenen Meinung zu einer bestimmten Thematik. Dennoch zielt ein geringer Teil dieser Informationen darauf ab, die Meinung von Userinnen und Usern in den sozialen Medien zu beeinflussen und damit gezielt Falschinformationen zu verbreiten. Staatliche und nichtstaatliche Desinformationskampagnen können sich an die gesamte Gesellschaft oder nur an bestimmte Teile davon richten, um sie bezüglich eines Themas zu täuschen oder zu verwirren. Da jeder Nutzer und jede Nutzerin Informationen selbst generieren und verbreiten kann, entsteht eine grosse Menge an Daten, die zu einer Informationsflut führen kann. Für den Einzelnen ist es unmöglich, alle Informationen zu verarbeiten und gleichzeitig jene Inhalte zu identifizieren, die gezielt Einfluss auf die öffentliche Meinung nehmen wollen. Erschwerend kommt hinzu, dass die Autorenschaft dieser Beiträge oft versucht, ihre Spuren zu verwischen. So setzen staatliche Akteure bei gezielten Desinformationskampagnen sogenannte Social Bots ein, um die gewünschten Desinformationen zu verbreiten. Für die Nutzerin und den Nutzer ist es daher schwierig, die Herkunft der Informationen nachzuvollziehen.

Künstliche Intelligenz bietet erhebliche Vorteile bei der wirksamen Bewältigung der beschriebenen Herausforderungen. Der entscheidende Nutzen der KI liegt in ihrer Fähigkeit, Muster in grossen Datenmengen zu erkennen. KI eignet sich sehr gut für die Analyse von sozialen Medien. Denn einerseits erfordert das Trainieren von Modellen für maschinelles Lernen oder für Deep-Learning-Methoden eine beträchtliche Menge an Daten. Andererseits können soziale Medien genau diese Menge an Daten bereitstellen.

Die Gruppe Data Science von armasuisse Wissenschaft und Technologie (W+T) ist bestrebt, mit Hilfe von künstlicher Intelligenz innovative Lösungen zu finden, damit Destabilisierungs- und Radikalisierungsversuche in sozialen Netzwerken frühzeitig erkannt werden können. Einige dieser Projekte werden im Folgenden erläutert.

Früherkennung von Radikalismus in sozialen Netzwerken

Verdächtige Aktivitäten in sozialen Netzwerken umfassen ein breites Spektrum an Verhaltensweisen. Sie können sich sowohl auf Inhalte, wie etwa das Erstellen und Verbreiten von Falschinformationen, als auch auf Konten, wie beispielsweise die Unterschlagung von Konten oder den Identitätsdiebstahl, beziehen. Alle diese Aktivitäten zielen darauf ab, Schaden anzurichten. Es ist daher wichtig, die Absichten dieser Personen zu erkennen, bevor das entsprechende Ereignis eintritt. Dies kann ein frühzeitiges Eingreifen ermöglichen und die Radikalisierungstendenzen einer Person abschwächen.

armasuisse W+T beschäftigt sich mit der Früherkennung von Radikalisierungstendenzen. Der Moment der Radikalisierung beschreibt den Zeitpunkt, zu dem der oder die

Nutzende ein bestimmtes radikales Verhalten – sei es in politischer, religiöser oder in anderer Hinsicht – in sozialen Netzwerken zeigt. Dieses Ereignis zeichnet sich in der Regel durch einen sogenannten Radikalisierungsprozess aus, dem – wenn auch je nach Thema unterschiedlich – oft ein Schneeballeffekt durch den Konsum von extremistischem Material vorausgeht. Dieser Prozess ist relativ langwierig, da die betreffende Person Schritt für Schritt in die jeweilige Ideologie abgleitet.

Bei der Früherkennung von Radikalismus ist es wichtig, den Fokus auf die Entwicklung des Nutzerverhaltens in den sozialen Medien zu legen und nicht nur auf den Wortlaut einzelner Posts. Ausgehend von den Daten bestätigter radikalisierter Nutzer/-innen, die den chronologischen Verlauf von Ereignissen in sozialen Netzwerken verfolgt und aufgezeichnet haben, und unter Berücksichtigung der verfügbaren Informationen (Text- und/oder Bildveröffentlichungen), ist es möglich, Modelle zu trainieren, die auf sprachlichen und sentimentalen Merkmalen basieren. Diese Modelle können dann die Vorläufersignale einer Radikalisierung erkennen. Sie sind dementsprechend in der Lage, die Emotionalität einer Information einzuordnen. Das ist wichtig, weil die Verbreitung falscher beziehungsweise radikaler Aussagen oft mit einer hohen Emotionalität des Beitrags einhergeht. Solche Modelle könnten zum Beispiel dazu dienen, Warnsignale an Nutzer und Nutzerinnen zu senden, die sich ihrer schleichenden Radikalisierung nicht bewusst sind.

Analyse des Kontoverhaltens auf Twitter

Die Interaktionen zwischen verschiedenen Konten, z. B. das Retweeten oder Folgen, können als gerichtete Graphen dargestellt werden. Anhand eines solchen Graphen kann ein Algorithmus durch Berechnung verschiedener statistischer Parameter bestimmen, wie sich ein Beitrag in sozialen Netzwerken, in diesem Fall Twitter, verbreitet hat. Zu diesen Parametern gehören der Vernetzungsgrad eines Kontos (wie viele Nutzer und Nutzerinnen dem Konto folgen und wie vielen Nutzern und Nutzerinnen das Konto selbst folgt).

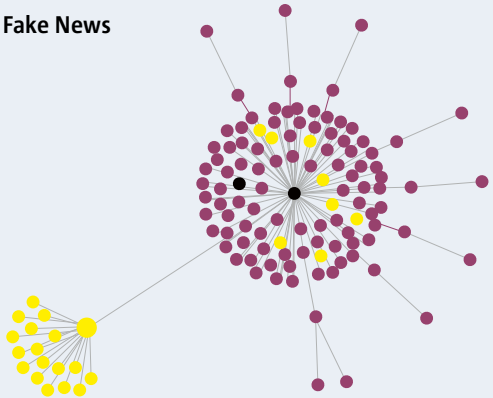
Bei einem viral verbreiteten Beitrag lassen sich der Ursprung und der Pfad dieser Verbreitung ermitteln. Durch die Analyse von Nutzerkonten kann der Tsunami zurückverfolgt werden, um herauszufinden, wie sich diese Informationen so stark verbreitet haben. Die oben beschriebenen Social Bots spielen bei diesem Tsunami oft eine wichtige Rolle, weil sie die Verbreitung der Informationen in Zusammenarbeit mit anderen Bots verstärken. Die Bots zeichnen sich durch ein charakteristisches Verhalten aus. Sie sind in der Lage, in kürzester Zeit eine grosse Anzahl von Followern (die selbst auch Bots sind) aufzubauen, was im Allgemeinen ein Gefühl der Glaubwürdigkeit vermittelt. Denn Tatsache ist, dass Menschen dazu neigen, den Beiträgen eines Nutzers zu vertrauen, dem eine grosse Community folgt. Außerdem neigen diese Bots dazu, Inhalte von anderen Bots oder Propaganda-Accounts zu jeder Tageszeit massenhaft neu zu veröffentlichen. Diese Art von Verhalten kann mathematisch modelliert werden, was die Identifizierung von

SOCIAL BOTS

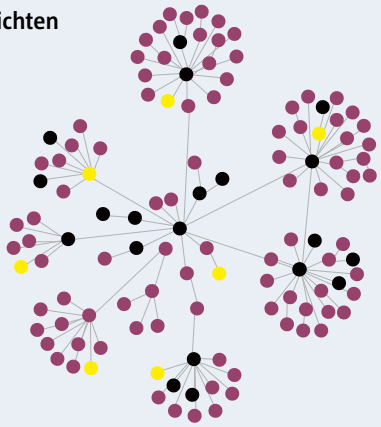
Als Social Bots werden automatisierte Programme bezeichnet, die z. B. auf bestimmte Hashtags mit einer vorab programmierten Antwort reagieren und bestimmte Inhalte in sozialen Medien teilen. Sie wirken wie übliche Benutzerkonten von Personen oder Unternehmen mit einem Profilbild, Beiträgen und einem interaktiven Netzwerk. Cyborgs sind teilautomatisierte Benutzerkonten, die von Menschen betrieben werden und damit über ein authentischeres Auftreten als Bots verfügen.

Tweets Verbreitungsbaum

von Fake News



von echten Nachrichten



– der zentrale Wurzelknoten zeigt einen Nachrichtenbericht an
– die schwarz markierten Knoten zeigen hoch einflussreiche Nutzer
– die pink markierten Knoten zeigen Retweets
– die gelb markierten Knoten zeigen Zitate der originalen Nachricht oder von einem Retweet

Bot-Spuren ermöglicht. Eine Herausforderung bei der Erkennung von Bots ist ihre relativ kurze Lebensdauer. Bots gewinnen schnell an Stärke, indem sie zahlreiche Verbindungen (Follower) aufbauen, ihren Desinformations- oder Propagandaauftrag ausführen und indem das Konto rasch aufgegeben oder gelöscht wird. So bleibt nur wenig Zeit, sie zu identifizieren.

Ziel ist es, die manuelle Analyse durch den Algorithmus zu skalieren und zu automatisieren. Der Algorithmus lernt selbst und entwickelt sich basierend auf den Gegebenheiten weiter.

Erkennung und Klassifizierung von Memes

Eine weitere populäre Methode zur Verbreitung von Falschinformationen ist die Verwendung von so genannten IWT-Memes (englisch: Image with Text-IWT, deutsch: Bild mit Text). Dank der Forschungstätigkeit von armasuisse W+T sollen künftig mittels Methoden der künstlichen Intelligenz diejenigen IWT-Memes identifiziert werden, die potenziell Fehlinformationen verbreiten.

Der erste Schritt bei der Abwehr von Meme-gesteuerten Desinformationen besteht darin, sie von anderen in sozialen Medien zahlreich verfügbaren Bilddaten zu unterscheiden. Zu diesem Zweck wurde ein Algorithmus entwickelt, der eine automatische Klassifizierung von IWT-Bildern im Vergleich zu Nicht-IWT-Meme-Bildern vornimmt (z. B. Ferienfotos, Screenshots etc.). Konkret beruht der Algorithmus auf sogenannten Convolutional Neural Networks, die dem Feld des Deep Learnings zuzuordnen sind. Die Methode wird auf verschiedenen IWT-Meme- und Nicht-IWT-Meme-Datensätzen trainiert und kann dadurch eine binäre Klassifikation von neu empfangenen Bilddaten in die Kategorien IWT-Meme-Bild bzw. Nicht-IWT-Meme-Bild vornehmen. Durch eine anschliessende Charakterisierung des Inhalts sowie der Nutzer und Nutzerinnen der so detektierten IWT-Memes zielt das Forschungsprojekt in Zukunft darauf ab, diejenigen Memes herausfiltern zu können, bei denen es sich möglicherweise um visuelle Desinformation handelt. Dazu soll zuerst der Inhalt des IWT-Memes analysiert werden. Bei der Ermittlung des Inhalts werden durch die Bestimmung des Themas und der Emotionalität des Inhalts Rückschlüsse darauf gezogen, ob es sich um Desinformation handeln könnte oder nicht. Desinformation beinhaltet häufig Themen, welche sozial spaltend wirken und, damit verbunden, negative Gefühle beim Betrachter oder der Betrachterin verstärken können. Ausserdem wird eine Charakterisierung der Verwender und Verwenderinnen der Memes angestrebt. Dabei wird der Frage nachgegangen, von wem das IWT-Meme verbreitet wird. Handelt es sich allenfalls um eine verdächtige Verbreitung der IWT-Memes durch Social-Bot-ähnliche Accounts? Und wie kann die Absicht des Nutzers oder der Nutzerin hinter der Verbreitung beurteilt werden? Lässt sich beispielsweise hinter der Verbreitung des Memes eine wahrscheinliche Absicht zur Diskreditierung einer politischen Person oder zur Spaltung von Bevölkerungsgruppen erkennen, so kann dies ein Indiz für das Vorliegen von Desinformation sein. Diese automatischen Analysemethoden können die Menge der auszuwertenden Bilddaten in Open-Source-Intelligence-Aufträgen drastisch reduzieren und ermöglichen so eine schnellere Bearbeitung von sicherheitsrelevanten Desinformationskampagnen durch Analysten und Analystinnen.

IWT-MEMES
(englisch: Image with Text-IWT, deutsch: Bild mit Text).

Die IWT-Memes dienen als Medium zur Verbreitung von Ideen, gestützt auf einer Kombination von Bild und Text. Sie zeichnen sich durch ihre charakteristische virale Verbreitung aus. Sie sind als Forschungsgegenstand besonders interessant, weil sie ein wirksames Mittel zur Beeinflussung von Online-Narrativen sind und daher häufig in Desinformationskampagnen eingesetzt werden.

Zukunftsthemen und Herausforderungen

Die künstliche Intelligenz bietet, wie in diesem Artikel ausgeführt, innovative Lösungen zur Bekämpfung von Desinformation. Dennoch sollten diese Methoden nicht als Universallösungen betrachtet werden. Das Training von Modellen wird immer ein entscheidender Schritt bleiben, der oft labelisierte Datensätze erfordert. Das Label gibt dabei an, ob es sich beispielsweise bei einem Bild um ein IWT-Meme handelt oder nicht. Das Modell findet dann selbständig heraus, was die Eigenschaften eines IWT-Memes sind. Da die Labelisierung von Datensätzen nach wie vor häufig nur manuell möglich ist, wird es schwierig, neue Trainingssätze zu erstellen, insbesondere, wenn sie viele Fälle abdecken müssen.

Eine weitere grosse Herausforderung ist die Analyse der Interaktionen zwischen den Nutzern und Nutzerinnen, die in Form von Graphen dargestellt werden. Tatsächlich entwickeln sich diese Graphen ständig weiter, zum Beispiel, wenn ein Nutzer oder eine Nutzerin Inhalte veröffentlicht oder wenn ihm oder ihr ein anderer folgt. Bei jeder Änderung muss das Modell daher den gesamten Graphen analysieren. Dies stellt eine enorme Datenmenge dar, die derzeit nicht in Echtzeit verarbeitet werden kann.

Letztlich nutzen die Autoren von Desinformationskampagnen, seien es Staaten, radikale Organisationen, Lobbies oder gar Privatpersonen, selbst die künstliche Intelligenz, um die entsprechenden Inhalte zu erstellen und in den sozialen Netzwerken zu propagieren. Aus diesem Grund ist es notwendig, dass die Modelle laufend neu trainiert werden, um diese neuen Taktiken und Formen der Desinformation zu erkennen. Folglich werden die Akteure den Tools, die zweifelhaftes Verhalten und entsprechende Beiträge erkennen können, stets einen Schritt voraus sein. Die von den Forschern von armasuisse W+T zusammen mit akademischen Partnern durchgeführten Forschungsarbeiten zielen darauf ab, die Verbreitung von Falschinformationen oder radikalen Ideologien frühzeitig zu erkennen, bevor sie sich viral verbreiten können. a