Fake News technisch begegnen – Detektions- und Behandlungsansätze zur Unterstützung von NutzerInnen

7

Katrin Hartwig und Christian Reuter

Zusammenfassung

Die Bedeutung des Umgangs mit Fake News hat sowohl im politischen als auch im sozialen Kontext zugenommen: Während sich bestehende Studien vor allem darauf konzentrieren, wie man gefälschte Nachrichten erkennt und kennzeichnet, fehlen Ansätze zur Unterstützung der NutzerInnen bei der eigenen Einschätzung weitgehend. Dieser Artikel stellt bestehende Black-Box- und White-Box-Ansätze vor und vergleicht Vor- und Nachteile. Dabei zeigen sich White-Box-Ansätze insbesondere als vielversprechend, um gegen Reaktanzen zu wirken, während Black-Box-Ansätze Fake News mit deutlich größerer Genauigkeit detektieren. Vorgestellt wird auch das von uns entwickelte Browser-Plugin TrustyTweet, welches die BenutzerInnen bei der Bewertung von Tweets auf Twitter unterstützt, indem es politisch neutrale und intuitive Warnungen anzeigt, ohne Reaktanz zu erzeugen.

Schlüsselwörter

Fake News · Desinformation · White-Box · Detektion · NutzerInnenunterstützung · Medienkompetenz

K. Hartwig (✉) · C. Reuter
Lehrstuhl Wissenschaft und Technik für Frieden und Sicherheit, Technische Universität Darmstadt, Darmstadt, Deutschland
E-Mail: hartwig@peasec.tu-darmstadt.de; reuter@peasec.tu-darmstadt.de

P. Klimczak, T. Zoglauer (Hrsg.), *Wahrheit und Fake im postfaktisch-digitalen Zeitalter*, ars digitalis, https://doi.org/10.1007/978-3-658-32957-0_7

7.1 Einleitung

Seit geraumer Zeit dienen soziale Netzwerke wie Facebook und Twitter immer mehr als wichtige Nachrichten- und Informationsquellen. Dabei entsteht eine teilweise vom professionellen Journalismus unabhängige Verbreitung von Information. Die großen Mengen an zur Verfügung stehenden Daten und Informationen können überfordernd sein. In diesem Kontext wurde der Begriff „Information Overload" geprägt (Kaufhold et al. 2020). Gleichzeitig wird die Verbreitung zweifelhafter oder gefälschter Inhalte erleichtert. Steinebach et al. (2020) nennen „ein hohes Tempo, Reziprozität, niedrige Kosten, Anonymität, Massenverbreitung, Passgenauigkeit und Unsichtbarkeit" als Merkmale, welche im Internet und insbesondere in sozialen Netzwerken die Verbreitung von Desinformation begünstigen. Darüber hinaus können auch im professionellen Journalismus ähnliche Phänomene wie die Verbreitung falscher Gerüchte oder Clickbaiting auftreten, begünstigt durch den stark auf Aufmerksamkeit basierenden Online-Markt.

Seit der Präsidentschaftswahl in den Vereinigten Staaten 2016 ist der Begriff Fake News weit verbreitet und wird sowohl im wissenschaftlichen Kontext als auch in öffentlichen Debatten aufgegriffen. Fake News werden von der EU-Kommission definiert als „alle Formen falscher, ungenauer oder irreführender Informationen, welche erfunden, präsentiert und verbreitet werden um Gewinne zu erzielen oder bewusst öffentlichen Schaden anzurichten" (European Commission 2018). Allcott und Gentzkow (2017, S. 213, übersetzt aus dem Englischen) definieren Fake News als „Nachrichtenartikel, welche absichtlich und nachprüfbar falsch sind und die LeserInnen irreführen können". Analysen haben gezeigt, dass Fake News häufig durch kleinere Änderungen in den Formulierungen entstehen, sodass sich beispielsweise die Grundstimmung ändert, und weniger vollkommen frei erfunden werden (Rashkin et al. 2017).

Auch in Deutschland wurden die Bundestagswahlen 2017 von Diskussionen über den Einfluss von Fake News begleitet. Die Forschungsergebnisse einer Studie von Sängerlaub (2017) zeigen jedoch, dass es keine bedeutenden Fake News während des Wahlkampfes gab, welche die Wahlergebnisse beeinflusst hätten. Diese Beobachtungen lassen vermuten, dass sich die Wahrnehmung von Fake News durch die Bevölkerung vom tatsächlichen Einfluss unterscheidet. Oft fällt Menschen die Unterscheidung von Fake News und wahren Nachrichten schwer, da kaum eine gefälschte Nachricht völlig falsch ist und wahre Nachrichten ebenfalls Fehler beinhalten können (vgl. Potthast et al. 2018).

Darüber hinaus werden auch jüngere Ereignisse mit einer Flut von Falschinformationen begleitet. So wurden insbesondere auf der Videoplattform TikTok hunderttausendfach problematische Clips zur Ausbreitung des Coronavirus aufgerufen. Um dem entgegenzuwirken, werden TikTok-NutzerInnen „verstärkt daran erinnert, Inhalte zu melden" (Breithut 2020). Videos mit irreführenden Informationen werden von dem Unternehmen entsprechend gelöscht.

Aktuelle Forschungsergebnisse zeigen weiterhin, dass nur eine begrenzte Anzahl von Personen tatsächlich anfällig ist, durch Fake News beeinflusst zu werden (Dutton und

Fernandez 2019). Eine Twitteranalyse in den USA ergab, dass „nur 1 % der Nutzer 80 % der Fake News ausgesetzt ist und 0,1 % der Nutzer für das Teilen von 80 % der Fake News verantwortlich ist“ (Grinberg et al. 2019, übersetzt aus dem Englischen). Obwohl die tatsächlichen Auswirkungen von Fake News noch immer ein kontroverses Thema sind und Forschungsergebnisse nahelegen, dass nur wenige NutzerInnen dafür anfällig sind, scheinen bereits große Teile der Bevölkerung Fake News begegnet zu sein. Eine repräsentative Umfrage in Deutschland aus dem Jahr 2017 zeigt, dass Fake News in der Wahrnehmung der Bevölkerung eine erhebliche Rolle spielen. 48 % gaben dabei an, Fake News schon einmal wahrgenommen zu haben. Weiter waren 84 % der Meinung, dass Fake News eine Gefahr darstellten und die Meinung der Bevölkerung manipulieren könnten. 23 % gaben an, Fake News schon einmal gelöscht oder gemeldet zu haben. Hingegen gaben nur 2 % an, selbst schon einmal Fake News erstellt zu haben (Reuter et al. 2019). Ein Überblick der Ergebnisse ist in Abb. 7.1 dargestellt.

Zusammenfassend lässt sich festhalten, dass Fake News durchaus negative Auswirkungen, beispielsweise auf Demokratie und das öffentliche Vertrauen haben können (Zhou et al. 2019). Tatsächlich gab es bereits Fälle, in denen die Verbreitung von Fake News erheblichen Schaden angerichtet hat. Im Jahr 2013 verursachte beispielsweise ein Fake Tweet des gehackten Accounts der US-Nachrichtenagentur Associated Press einen Börsenschaden von 130 Milliarden Dollar, in welchem fälschlicherweise von Explosionen im Weißen Haus berichtet wurde (Rapoza 2017). Weiter führten Fake News der #PizzaGate Verschwörungstheorie zu einer Schießerei in einer Pizzeria in Washington D.C. (Aisch et al. 2016).

Technische Lösungen zum Umgang mit Fake News, insbesondere in sozialen Netzwerken, haben großes Potenzial mit weniger Nutzeraufwand dem Einfluss von Fake News entgegenzuwirken. In der Entwicklung technischer Unterstützungsansätze zum Umgang mit Fake News sind prinzipiell zwei Schritte notwendig: Fake News detektieren und Behandlungsmaßnahmen zum Schutz und zur Unterstützung von NutzerInnen treffen (Potthast et al. 2018). Diese sind in Tab. 7.1 näher erläutert.

Zu berücksichtigen ist auch bei wem letztlich die Verantwortlichkeit liegt. Die repräsentative Studie von Reuter et al. (2019) untersuchte Meinungen der deutschen Bevölkerung zum Umgang mit Fake News. Dabei wurden TeilnehmerInnen unter anderem gebeten, die folgenden Vorschläge zum Umgang mit Fake News auf einer 5-stufigen Likert-Skala zu bewerten: Schnelle Reaktionen der Behörden, BetreiberInnen müssen böse und erfundene Inhalte löschen, BetreiberInnen sollen Fake News markieren, transparenter und selbstkritischer Journalismus sowie die Einrichtung staatlicher IT-Abwehrzentren. Die meisten TeilnehmerInnen gaben an, mit allen vorgeschlagenen Maßnahmen einverstanden zu sein. Dabei zeigt die Idee zur Einrichtung staatlicher IT-Abwehrzentren für die Bekämpfung von Fake News mit immerhin 72 % im Vergleich zu den anderen Items die geringste Akzeptanz (Reuter et al. 2019).

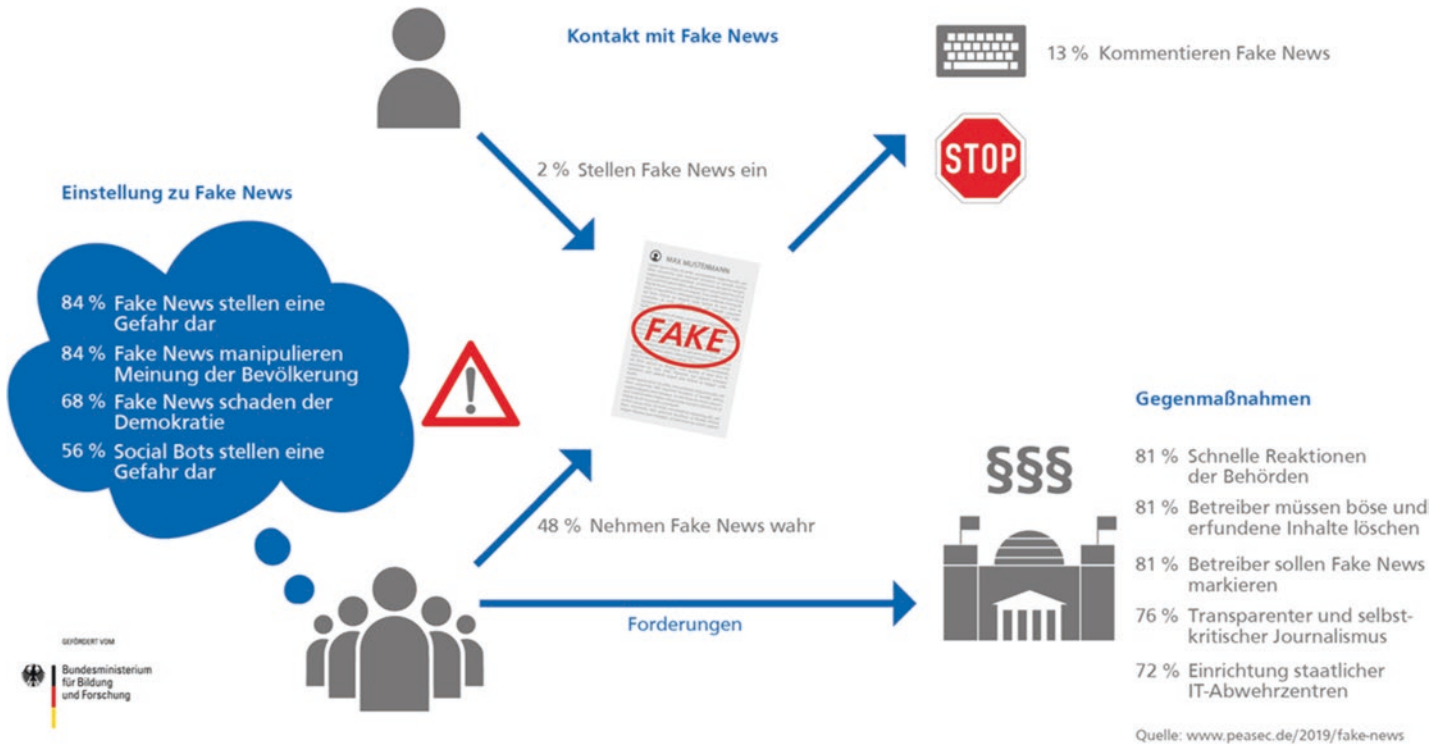

Abb. 7.1 Ergebnisse der repräsentativen Umfrage zur Wahrnehmung von Fake News (Reuter et al. 2019)

Tab. 7.1 Schritte zur technischen Unterstützung im Umgang mit Fake News

Schritte	1. Detektion	2. Behandlungsansätze
Beschreibung	Erkennen von Desinformationen; beispielsweise aus einer Menge von Tweets jene Tweets identifizieren, die Fake News enthalten.	Maßnahmen treffen, um NutzerInnen vor den Auswirkungen von Fake News zu schützen und sie zu befähigen, selbst Inhalte zu bewerten.

7.2 Detektion von Fake News in Sozialen Medien

7.2.1 Ansätze

Da sich Fake News nach Angaben von Vosoughi et al. (2018) schneller verbreiten als wahre Nachrichten, sind interdisziplinäre Ansätze essenziell, um die damit verbundenen komplexen Herausforderungen anzugehen. Zur Detektion von Fake News in sozialen Medien existieren bereits verschiedene Methoden. Plattformen können ihren NutzerInnen beispielsweise erlauben, verdächtige Inhalte zu melden. Weiter können professionelle FaktencheckerInnen die gemeldeten Inhalte manuell verifizieren oder widerlegen. Zudem wächst das Forschungsfeld der automatisierten Detektion von Fake News durch technische Lösungen, zum Beispiel Style-based Fake News Detection, Propagation-based und Context-based Fake News Detection (Potthast et al. 2018; Zhou et al. 2019).

Einen guten Überblick zu gängigen Detektionsansätzen für Fake News bietet die Arbeit von Steinebach et al. (2020). Die AutorInnen unterscheiden dabei die Erkennung von Desinformationen bezüglich Texten, Bildern und Bots. Weiter differenzieren Zhang und Ghorbani (2020) automatische Detektionsverfahren nach drei Kategorien – *component-based, data mining-based* und *implement-based* Ansätze. Dabei untersuchen component-based Detektionsverfahren beispielsweise die AutorInnen von Fake News oder NutzerInnen sozialer Medien anhand von Sentimentanalysen. Die Sentimentanalyse gehört zum Bereich des Text Mining und untersucht beispielsweise anhand von Signalwörtern automatisiert, welche Empfindungen und Stimmungen in Texten bestimmter AutorInnen vorherrschen. Weiter untersuchen component-based Detektionsverfahren Nachrichteninhalte anhand von linguistischen (z. B. besonders viele Ausrufezeichen), semantischen (z. B. besonders aufmerksamkeitserregende Titel, welche inhaltlich im Konflikt zum Textkörper stehen), wissensbasierten (z. B. Webseiten, welche ExpertInnenwissen nutzen) oder stilbasierten (z. B. Schreibstil mit besonders hoher Anzahl an emotionalen Wörtern) Merkmalen sowie den sozialen Kontext anhand von Nutzernetzwerkanalysen oder Verbreitungsmustern. Die Kategorie der data mining-basierten Detektionsverfahren hingegen unterscheidet überwachtes und unüberwachtes Lernen. Weiter unterscheidet die Kategorie der implement-based Ansätze Echtzeit- und Offlinedetektion von Fake News (vgl. Zhang und Ghorbani 2020). Die Kategorisierung der Detektionsverfahren von Fake News ist in Abb. 7.2 zu finden.

Viele Ansätze setzen den Fokus auf Charakteristika von Textinhalten (Granik und Mesyura 2017; Gravanis et al. 2019; Hanselowski et al. 2019b; Potthast et al. 2018; Rashkin et al. 2017; Zhou et al. 2019). Eine Grundlage für Machine-Learning-Ansätze zum

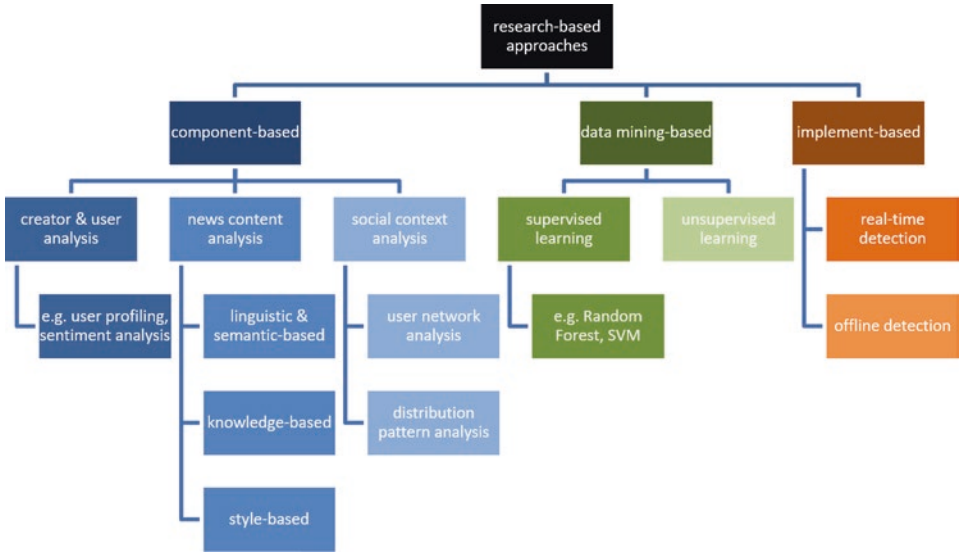

Abb. 7.2 Automatische Detektionsverfahren von Fake News nach Zhang und Ghorbani 2020

automatisierten Faktenprüfen bietet der annotierte Korpus von Hanselowski et al. (2019a). Andere untersuchen die Interaktion von NutzerInnen (Long et al. 2017; Ruchansky et al. 2017; Shu et al. 2019b; Tacchini et al. 2017) oder die Ausbreitung von Inhalten innerhalb sozialer Netzwerke (Monti et al. 2019; Shu et al. 2019a; Wu und Liu 2018). Weitere Arbeiten behandeln das Verhältnis der Überschrift zum Textkörper (Bourgonje et al. 2018), Argumentationen (Sethi 2017) und widersprüchliche Sichtweisen auf ein Thema (Jin et al. 2016).

Anlehnend an existierende Ansätze zur Identifizierung von Spamnachrichten werden oftmals Naive Bayes-Klassifikatoren für die Detektion und Wahrscheinlichkeitsberechnung von Fake News verwendet. Hier werden Objekte anhand des mathematischen *Satzes von Bayes* einer Klasse (z. B. (a) Fake News oder (b) korrekte Information) zugeordnet, der sie mit größter Wahrscheinlichkeit am ähnlichsten sind. Da Artikel, die Fake News beinhalten, oft dieselben Wortgruppen aufweisen, kann mithilfe von Naive Bayes-Klassifikatoren berechnet werden, mit welcher Wahrscheinlichkeit Artikel Fake News enthalten (Granik und Mesyura 2017). Sowohl Pérez-Rosas et al. (2017) als auch Potthast et al. (2018) greifen auf linguistische und semantische Merkmale (z. B. bestimmte N-Gramme, Satz- und Wortproportionen) für die Detektion von Fake News zurück. Potthast et al. (2018) fokussieren dabei insbesondere stilistische Merkmale bei Nachrichten, die links- oder rechtsextreme Inhalte beinhalten. Hierbei fällt auf, dass sich trotz sehr unterschiedlicher politischer Orientierung die verwendeten Schreibstile sehr ähnlich sind. Des Weiteren zeigt sich, dass Superlative und Übertreibungen vermehrt in Fake News verwendet werden (Rashkin et al. 2017).

Für die Erkennung bestimmter Sprachmuster, die für das Detektieren von Fake News essenziell sind, werden algorithmische Ansätze des maschinellen Lernens, z. B. von Per-

spective API, angewendet. Diese prüfen Aussagen und Nachrichten beispielsweise auf kurze Sätze und bestimmte Zeitformen. Faktenprüfende Websites wie PolitiFact ordnen überprüfte Artikel auf einer Skala, die von „wahr“ bis „absolut falsch“ skaliert sind, ein (Rashkin et al. 2017, S. 2931). Ein weiterer Ansatz von Gravanis et al. (2019) beschreibt, dass für die Identifikation von Fake News ein Werkzeug notwendig sei, das die Profile von Menschen, die Fake News erstellen, erkennen kann. Auch Castillo et al. (2011) nutzen verschiedene Merkmale von BenutzerInnenprofilen (z. B. Registrierungsalter), um Fake News zu identifizieren. Neben inhaltlichen und stilistischen Überprüfungsmechanismen gibt es den *Propagation-based Fake News Detection* Ansatz. Dieser untersucht, wie Nachrichten in sozialen Netzwerken propagiert werden (Zhou et al. 2019).

7.2.2 Black-Box vs. White-Box

Die genannten Detektionsalgorithmen haben jedoch einen signifikanten Nachteil: Sie sind Blackbox-basiert und liefern den EndnutzerInnen entsprechend keine Erklärung zur automatisierten Entscheidungsfindung. Die NutzerInnen können den Input (z. B. ein Tweet) und den Output (z. B. die Markierung des Tweets als Fake News) beobachten, erhalten jedoch keine Information darüber, was dazwischen geschieht (z. B. warum ein Tweet als Fake News markiert wurde). Das Gegenstück zu Blackbox-Ansätzen nennt man Whitebox-Ansatz. Bei diesem können die internen Vorgänge zwischen Input und Output beobachtet werden. Im Kontext von Fake News ermöglichen Whitebox-Ansätze die Nachvollziehbarkeit von Indikatoren für falsche Inhalte. Entsprechend haben NutzerInnen hier Zugang zu allen notwendigen Informationen, um zu verstehen, warum der Algorithmus einen spezifischen Output generiert hat. Eine entsprechende Visualisierung ist in Abb. 7.3 dargestellt.

In anderen Kontexten, in denen maschinelles Lernen Anwendung findet, wird der Bedarf an „Interpretierbarkeit, Erklärbarkeit und Vertrauenswürdigkeit“ bereits hervorgehoben und vermehrt diskutiert (Conati et al. 2018, S. 24, übersetzt aus dem Englischen). Erklärbares maschinelles Lernen setzt sich zum Ziel, das Vertrauen der NutzerInnen in die Ergebnisse von Systemen aufzubauen (Ribeiro et al. 2016). Bisher gibt es jedoch wenige Ansätze zur Detektion von Fake News, welche mit erklärbarem maschinellem Lernen arbeiten. Zu erwähnen sind dabei die Ansätze von Reis et al. (2019) und Yang et al. (2019).

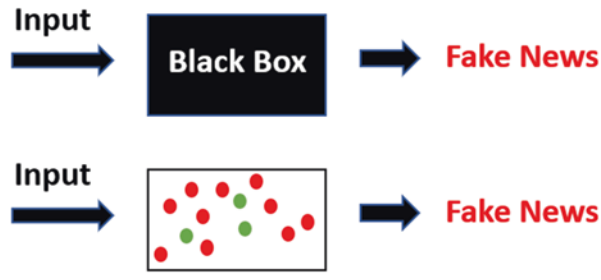

Abb. 7.3 Visualisierung vom Blackbox- (oben) und Whitebox-Ansatz (unten)

Andere Whitebox-Ansätze setzen den Fokus auf die Bildung der NutzerInnen mit dem Ziel der verbesserten Medienkompetenz. Studien haben gezeigt, dass eine verbesserte Medienkompetenz vielversprechende gegensteuernde Auswirkungen im Umgang mit Fake News haben kann (Kahne und Bowyer 2017; Mihailidis und Viotty 2017). Wenn die Fähigkeit zur autonomen Bewertung von Online-Inhalten durch Whitebox-Ansätze verbessert wird, kann dies Reaktanzen vermindern und den Backfire-Effekt verhindern. Als Backfire-Effekt bezeichnen Nyhan und Reifler (2010) die Entstehung von Wut und Trotz, wenn insbesondere politische Inhalte einen Warnhinweis enthalten. NutzerInnen neigen dazu, den Inhalten dann erst recht zu glauben, da sie „die Korrektur als einen illegalen Persuasionsversuch wahrnehmen“ (Müller und Denner 2017, S. 17). Hartwig und Reuter (2019) haben ein Browser-Plugin entworfen, welches politisch neutrale und transparente Hinweise zu Charakteristiken eines Tweets auf Twitter liefert, welche auf nicht vertrauenswürdige Inhalte hindeuten. In einem ähnlichen Ansatz stellen Bhuiyan et al. (2018) ein Browser-Plugin vor, welches NutzerInnen auf Twitter durch Nudging zu besserer Einschätzung der Glaubwürdigkeit von Nachrichtenartikeln verhelfen soll. Gezielte Fragen (z. B. Erzählt der Beitrag die gesamte Story?) dienen dabei als Nudge, um NutzerInnen zu reflektiertem Denken anzuregen. Weiter stellen Fuhr et al. (2018) einen Ansatz vor, in dem sie ähnlich wie bei Nährwertangaben auf Lebensmitteletiketten Onlinetexte zum Beispiel bezüglich Fakten und Emotionen kennzeichnen und somit LeserInnen bei der informierten Beurteilung unterstützen. Anstelle von eindeutigen Blackbox- oder Whitebox-Ansätzen nutzen Plattformen in der Regel Kombinationen verschiedener Strategien, um Fake News zu detektieren. So bietet Facebook NutzerInnen die Möglichkeit, verdächtige Inhalte zu melden und wendet gleichzeitig Algorithmen zur Detektion und Priorisierung von Falschnachrichten an, welche im Folgenden von unabhängigen FaktencheckerInnen untersucht werden (McNally und Bose 2018; Mosseri 2016).

7.3 Behandlungsmaßnahmen zur Unterstützung von NutzerInnen

Eine große Zahl wissenschaftlicher Arbeiten hat bereits Möglichkeiten untersucht, um Fake News automatisch zu erkennen. Weniger Aufmerksamkeit ist jedoch dem nächsten Schritt zuteil geworden, nämlich was zu unternehmen ist, wenn Desinformationen in den sozialen Medien schließlich detektiert wurden. Wurden Desinformationen erfolgreich identifiziert? Gibt es verschiedene Ansätze, um im Folgenden damit umzugehen?

Da sich Falschinformationen vorrangig über soziale Medien ausbreiten, haben Plattformen wie Facebook, Twitter und Instagram begonnen, gegenzusteuern. Viele der Ansätze sind für die NutzerInnen direkt sichtbar und beeinflussen das Erleben in sozialen Netzwerken. Insbesondere Facebook hat seit 2016 eine Reihe von Verfahren als potenzielle Gegenmaßnahmen eingesetzt (Tene et al. 2018). Beispielsweise begann Facebook nach den US-Wahlen 2016 damit, Warnungen unter umstrittenen Beiträgen anzuzeigen (Mosseri 2016). Medienberichten zufolge wurde dieses Feature nach anhaltender Kritik jedoch zu-

rückgezogen. Seitdem nutzt Facebook subtilere Techniken, um die Reichweite umstrittener Beiträge einzuschränken, wie etwa das Verkleinern der Beitragsgröße, die Auflistung von Faktencheck-Artikeln und das Herabsetzen des Beitragsrankings im Newsfeed (McNally und Bose 2018). Diese Gegenmaßnahmen scheinen in etwa den gewünschten Effekt zu erzielen, indem sie die Verbreitung von Fake News im sozialen Netzwerk reduzieren. Seit der Einführung im Jahr 2016 konnte die Interaktion mit Fake News auf Facebook um mehr als 50 % reduziert werden (Allcott et al. 2019). Kirchner und Reuter (2020) stellen in ihrer Arbeit verschiedene genutzte Techniken der sozialen Medien überblicksartig dar. Weiter vergleichen sie die Effektivität und NutzerInnenakzeptanz verschiedener Maßnahmen wie beispielsweise der Anzeige von Warnungen oder verwandten Artikeln und der Bereitstellung zusätzlicher Informationen.

Das Kennzeichnen und Löschen von falschen Inhalten kann unter Umständen jedoch nicht effektiv und manchmal sogar kontraproduktiv sein. Als eine vielversprechende Strategie dagegen sehen viele ForscherInnen stattdessen ein Training der Medienkompetenz (Müller und Denner 2017; Stanoevska-slabeva 2017; Steinebach et al. 2020). Studien haben gezeigt, dass Menschen mit hoher Medienkompetenz in der Lage sind, einen Großteil deutschsprachiger Fake News anhand verschiedener Faktoren wie beispielsweise dem Textaufbau leicht zu erkennen, da Desinformationen im Textkörper oftmals mehr als zwei Rechtschreibfehler, durchgängige Großschreibung oder Fehler bei der Zeichensetzung aufweisen (vgl. Steinebach et al. 2020). Da jedoch die meisten Ansätze, um automatisch Fake News zu erkennen und zu kennzeichnen, Blackbox-Algorithmen nutzen und dies auch bei vielen verbreiteten Machine-Learning-Techniken der Fall ist, können diese meist nicht klarstellen, warum sie bestimmte Inhalte als Fake News labeln. NutzerInnen ein Label zu präsentieren kann sogar zu Reaktanz führen, wenn dieses nicht der eigenen Wahrnehmung entspricht. Dieser Effekt wird durch den sogenannten Bestätigungsfehler (confirmation bias) erzeugt, der auftritt, wenn Nachrichten gerade dann als wahr angesehen werden, wenn sie der eigenen Ideologie entsprechen (Kim und Dennis 2018; Nickerson 1998; Pariser 2011).

Bode und Vraga (2015) untersuchten die Möglichkeit, mit korrigierenden Informationen in der Sektion „Verwandte Artikel“ unter dem jeweiligen Beitrag, Falschinformationen zu bekämpfen. ForscherInnen haben bereits gezeigt, dass Warnungen vor Falschformationen deren wahrgenommene Richtigkeit reduzieren (Ecker et al. 2010; Lewandowsky et al. 2012; Sally Chan et al. 2017), diese jedoch auch fehlschlagen können (Berinsky 2017; Nyhan und Reifler 2010; Nyhan et al. 2013). Beispielsweise haben Garrett und Weeks (2013) sofortige mit verspäteter Richtigstellung bei Falschformationen verglichen. Dabei zeigte sich, dass die sofortige Richtigstellung den signifikantesten Einfluss auf die wahrgenommene Richtigkeit hat. Wenn jedoch Falschinformationen die Meinung der NutzerInnen bestätigt, ist das Potenzial für einen Backfire-Effekt größer (Kelly Garrett und Weeks 2013). Pennycook et al. (2018) zeigen, dass sich ein verwandtes Phänomen – wiederholter Konsum von Falschinformationen erhöht die wahrgenommene „Illusory Truth Effect“-Genauigkeit – auch auf Fake News in den sozialen Medien übertragen lässt. Zudem fanden sie heraus, dass Warnungen die wahrgenommene Richtigkeit der Inhalte ver-

ringern können. Pennycook et al. (2019) bestätigen den positiven Effekt solcher Warnungen. Mit einem bayesschen „Implied Truth"-Modell argumentieren sie, dass das Zeigen von Warnbenachrichtigungen bei Falschnachrichten nicht nur den Glauben an deren Richtigkeit reduziert, sondern auch den Glauben an die Richtigkeit von Nachrichten ohne angehängte Warnung erhöht. Clayton et al. (2019) vergleichen mehrere Arten von Warnungen. Neben spezifischen Warnungen vor falschen Schlagzeilen testen sie auch eine generelle Warnung ohne Bezug zu einem bestimmten Beitrag. Facebook hatte im April 2017 und Mai 2018 solch eine Warnung über den Newsfeed von NutzerInnen angezeigt und darin generell vor Falschinformationen gewarnt. Darüber hinaus untersuchen sie zwei unterschiedliche Formulierungsweisen für bestimmte Warnungen zu Schlagzeilen: „umstritten" und „als falsch eingestuft". Ihre Ergebnisse zeigen, dass generelle Warnungen nur einen minimalen Effekt, spezifische Warnungen jedoch einen signifikanten Effekt haben. Somit bestätigen sie die Ergebnisse von Pennycook et al. (2019). Diese Gruppe von WissenschaftlerInnen gelangte zu dem Ergebnis, dass „als falsch eingestuft"-Warnungen signifikant effektiver sind als solche, die mit dem Label „umstritten" versehen werden.

7.4 TrustyTweet: Ein Whitebox-Ansatz zum Umgang mit Fake News

Wie in Abschn. 7.3 aufgezeigt, gilt die Erhöhung der Medienkompetenz als vielversprechende Strategie im Umgang mit Fake News. Unter Angabe transparenter und identifizierbarer Indikatoren für Fake News können NutzerInnen bei der Meinungsbildung zu Online-Inhalten unterstützt werden. In diesem Kontext ist es wichtig, für das Trainieren von Medienkompetenz zwischen Assistenz-Systemen, die neutrale Hinweise basierend auf transparenten Indikatoren geben, und Systemen die Reaktanz hervorrufen, zu differenzieren, um einem Backfire-Effekt entgegenzuwirken. Die Nutzung eines Whitebox- anstatt eines Blackbox-Ansatzes ist ein wichtiger Schritt, um Reaktanz zu verringern oder zu verhindern.

Im Folgenden wird das Browser-Plugin TrustyTweet vorgestellt, welches Twitter-NutzerInnen im Umgang mit Fake News auf Twitter unterstützen soll, indem politisch neutrale, transparente und intuitive Hinweise gegeben werden (Hartwig und Reuter 2019). Dieser Ansatz zielt insbesondere darauf ab, ein hilfreicher Assistent zu sein, ohne zu Reaktanz zu führen. Die NutzerInnen werden dadurch nicht ihres eigenen Urteilsvermögens beraubt. Ziel ist es, einen Lerneffekt bezüglich Medienkompetenz herbeizuführen, der nach längerer Nutzung das Plugin überflüssig macht. Im Gegensatz zu anderen Ansätzen ist TrustyTweets deshalb auf einer Whitebox-Technologie aufgebaut. Das Plugin wurde in einem NutzerInnen-zentrierten Designprozess unter Zuhilfenahme des „Design Science"-Ansatzes entwickelt. Es wurden potenzielle Indikatoren für Fake News durch das Abwägen von Ansätzen identifiziert, die sich in wissenschaftlichen Arbeiten bereits als erfolgversprechend herausgestellt haben. Der Fokus liegt auf Heuristiken, die von Menschen intuitiv und erfolgreich genutzt werden und einfach zu verstehen sind. Wichtig ist jedoch

zu betonen, dass dieser Ansatz nicht alle relevanten Indikatoren für Fake News umfassen kann.

Als potenzielle Indikatoren werden folgende Charakteristiken verwendet (Tab. 7.2):

TrustyTweet wurde für den Webbrowser Firefox entwickelt. Seine Hauptbestandteile sind eine Textbox, welche alle in einem Tweet erkannten Indikatoren beinhaltet und als Warnbenachrichtigung dient, zwei verschiedene Icons, um anzuzeigen, ob Indikatoren im Tweet erkannt wurden und schließlich ein weiteres Icon, durch welches man zu den Einstellungen gelangt, welche in einem Popup-Fenster geöffnet werden. Neben jedem Indikator befindet sich ein Link, um allgemeine Informationen zu selbigem Indikator in einem Popup-Fenster aufzurufen. Bewegt man die Maus über einen Indikator wird die entsprechende Komponente im Tweet dynamisch hervorgehoben (siehe Abb. 7.4). So können NutzerInnen sofort sehen, warum eine Warnung angezeigt wird. Das Haupticon des Plugins dient als Umschaltknopf für die Textbox. Die NutzerInnen können entscheiden, ob sie alle erkannten Indikatoren neben dem jeweiligen Tweet sehen möchten oder ob sie nur ein Icon sehen und bei Bedarf zur Textbox umschalten möchten, um zu sehen, warum die aktuelle Warnung angezeigt wird. Ein wesentliches Feature von TrustyTweet ist der Konfigurations-Popup. Durch die Nutzung von Kontrollkästchen können NutzerInnen einzelne Indikatoren zur Untersuchung von Tweets an- und abschalten. So wird durch unser Plugin ein stärkeres Gefühl von Autonomie vermittelt und Bevormundung entgegengetreten.

Evaluiert wurden Benutzbarkeit und NutzerInnenerlebnis des Plugins im Rahmen von ersten qualitativen Thinking-Aloud Studien mit insgesamt 27 TeilnehmerInnen. Dabei wurde das Unterstützungstool größtenteils als hilfreich und intuitiv bewertet. Weiter bringen die Erkenntnisse unserer Studie Hinweise auf die folgenden Design-Implikationen für Unterstützungswerkzeuge im Umgang mit Fake News mit sich:

Tab. 7.2 Potenzielle Indikatoren für Fake News

Indikator	Beispiel	Literatur
fortlaufende Großschreibung	fortlaufende GROßSCHREIBUNG	(Steinebach et al. 2020; Wanas et al. 2008; Weerkamp und De Rijke 2008; Weimer et al. 2007)
übermäßige Nutzung von Satzzeichen	übermäßige Nutzung von Satzzeichen **!!!**	(Morris et al. 2012; Wanas et al. 2008)
falsche Zeichensetzung am Satzende	falsche Zeichensetzung am Satzende **!!1**	(Morris et al. 2012; Weimer et al. 2007)
übermäßige Nutzung von Emoticons und insbesondere aufmerksamkeitserregenden Emoticons		(Wanas et al. 2008; Weerkamp und De Rijke 2008)
die Nutzung des Standard-Profilbildes		(Morris et al. 2012)
Fehlen der offiziellen Account-Verifizierung, besonders bei Berühmtheiten		(Morris et al. 2012)

Abb. 7.4 Beispieloutput von TrustyTweet (Hartwig und Reuter 2019)

1. Personalisierung, um die persönliche Autonomie zu erhalten: Das Konfigurations-Feature ist wichtig, um die Autonomie zu erhöhen und Reaktanz zu verhindern.
2. Unterstützung der NutzerInnen durch transparente und objektive Informationen: Die Indikatoren benötigen detaillierte Beschreibungen, aus denen klar wird, warum sie relevant zur Erkennung von Fake News sind. Unseren TesterInnen zufolge ist es von großer Bedeutung, dass die Beschreibungen politisch neutral und in objektiver Art und Weise formuliert sind.
3. Eindeutige Zuordnung von Warnungen: Bestandteile eines Tweets beim Darüberfahren mit der Maus hervorzuheben, wenn eine Warnung ausgelöst wurde, wurde als eines der hilfreichsten Plugin-Features angesehen und ist unabdingbar, um einen Lerneffekt zu erzielen.
4. Personalisierbarkeit der Auffälligkeit: Das Umschalt-Feature der Warnungen wurde ebenfalls positiv aufgenommen. Vielen TeilnehmerInnen gefiel die Funktion, detaillierte Textboxen nur bei Bedarf anzeigen zu lassen und ansonsten hauptsächlich die Farbe des Icons zu beachten.
5. Minimierung von Falschalarmen: Wie auch in vielen anderen Kontexten (z. B. Warnapps) ist es sehr wichtig, Falschalarme zu minimieren, da NutzerInnen sonst die Aufmerksamkeit für das Plugin verlieren oder dieses deinstallieren könnten, bevor ein Lerneffekt eingetreten ist. Um das Plugin diesbezüglich zu verbessern, schlugen einige ProbandInnen die Anzeige von graduellen Warnungen (beispielsweise in Ampelfarben) als mögliche Alternative vor.

7.5 Fazit und Ausblick

Der Umgang mit Fake News ist aktuell eine große Herausforderung für Gesellschaft und Politik (vgl. Granik und Mesyura 2017). Studien haben gezeigt, dass es einen großen Bedarf an Assistenz-Systemen gibt, um NutzerInnen sozialer Medien zu unterstützen. Bisher

hat sich die Forschung insbesondere darauf konzentriert, mithilfe von Machine-Learning-Algorithmen Fake News zu erkennen und zu labeln. Beispielsweise präsentieren Gupta et al. (2014) ein Browser-Plugin, welches automatisch den Wahrheitsgehalt von Inhalten auf Twitter bewertet. Weitere Ansätze (z. B. Fake News KI) nutzen ebenfalls maschinelles Lernen. Wieder andere Ansätze basieren auf Whitelists und Blacklists (z. B. B.S. Detector) zum Erkennen von Fake News. Blackbox-Verfahren bergen jedoch die Gefahr, Reaktanz hervorzurufen, da sie keine Gründe für ihre Fake-News-Warnungen geben können.

In unseren Augen und der Meinung anderer Studien folgend (Müller und Denner 2017; Stanoevska-slabeva 2017), ist das Verbessern individueller Medienkompetenz eine zentrale Strategie im Umgang mit Fake News. Die ersten empirischen Ergebnisse der durchgeführten Studie zeigen, dass unser Indikator-basierter Whitebox-Ansatz zur Unterstützung von Twitter-NutzerInnen im Umgang mit Fake News potenziell vielversprechend ist, wenn folgende fünf Design-Implikationen beachtet werden: Personalisierbarkeit zur Steigerung der Autonomie, transparente und objektive Informationen, Unzweideutigkeit der Warnungen, personalisierte Wahrnehmbarkeit und Minimierung von Fehlwarnungen. Für zukünftige Studien ist eine Kombination aus automatischer Detektion von Fake News und anschließendem Einsatz von TrustyTweet als Unterstützungsmaßnahme geplant. Hier könnten die Vorteile beider Verfahren genutzt werden: die transparenten und einfach zu verstehenden Indikatoren unsererseits und die akkurate Detektion von Blackbox-Verfahren auf der anderen Seite. Ein entsprechendes repräsentatives Onlineexperiment zur Wirksamkeit von TrustyTweet in Kombination mit automatischen Detektionsverfahren als Ergänzung zur durchgeführten qualitativen Studie ist in Planung.

7.6 Danksagung

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – SFB 1119 – 236615297 (CROSSING) sowie vom Bundesministerium für Bildung und Forschung (BMBF) und vom Hessischen Ministerium für Wissenschaft und Kunst (HMWK) im Rahmen ihrer gemeinsamen Förderung für das Nationale Forschungszentrum für angewandte Cybersicherheit ATHENE.

Dieser Artikel basiert in Teilen auf dem Artikel „*Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation*“ (Reuter et al. 2019) sowie „*TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter*“ (Hartwig und Reuter 2019). Überdies basiert er in Teilen auf dem Konferenzbeitrag „*Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness*“ (Kirchner und Reuter 2020). Wir bedanken uns bei Jan Kirchner für seine Unterstützung.

Literatur

Aisch G, Huang J, Kang C (2016) Dissecting the #PizzaGate conspiracy theories. New York Times. https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html. Zugegriffen am 18.04.2020

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236. https://doi.org/10.1257/jep.31.2.211

Allcott H, Gentzkow M, Yu C (2019) Trends in the diffusion of misinformation on social media. Res Politics 6(2). https://doi.org/10.1177/2053168019848554

Berinsky AJ (2017) Rumors and health care reform: experiments in political misinformation. Br J Polit Sci 47(2):241–262. https://doi.org/10.1017/S0007123415000186

Bhuiyan MM, Zhang K, Vick K, Horning MA, Mitra T (2018) Feed reflect: a tool for nudging users to assess news credibility on Twitter. In: Companion of the 2018 ACM conference on computer supported cooperative work and social computing – CSCW '18, S 205–208. https://doi.org/10.1145/3272973.3274056

Bode L, Vraga EK (2015) In related news, that was wrong: the correction of misinformation through related stories functionality in social media. J Commun 65(4):619–638. https://doi.org/10.1111/jcom.12166

Bourgonje P, Moreno Schneider J, Rehm G (2018) From clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In: Proceedings of the 2017 EMNLP workshop: natural language processing meets journalism, S 84–89. https://doi.org/10.18653/v1/w17-4215

Breithut J (2020) Falschinformationen im Netz: So reagieren Facebook, Google und TikTok auf das Coronavirus. Spiegel Online. https://www.spiegel.de/netzwelt/web/coronavirus-wie-facebook-google-und-tiktok-auf-falschinformationen-reagieren-a-6bc449fc-2450-4964-a675-7d6573316ad9. Zugegriffen am 03.02.2020

Castillo C, Mendoza M, Poblete B (2011) Information credibility on twitter. In: Proceedings of the International Conference on World Wide Web, Hyderabad, S 675–684

Clayton K et al (2019) Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. Polit Behav. https://doi.org/10.1007/s11109-019-09533-0

Conati C, Porayska-Pomsta K, Mavrikis M (2018) AI in education needs interpretable machine learning: lessons from open learner modelling. In: Proceedings of 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018). http://arxiv.org/abs/1807.00154. Zugegriffen am 22.04.2021

Dutton WH, Fernandez L (2019) How susceptible are internet users? InterMedia 46(4). https://doi.org/10.2139/ssrn.3316768

Ecker UKH, Lewandowsky S, Tang DTW (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. Mem Cogn 38(8):1087–1100. https://doi.org/10.3758/MC.38.8.1087

European Commission (2018) A multi-dimensional approach to disinformation. Report of the independent High Level Group on fake news and online disinformation (bd 2). https://doi.org/10.2759/0156.

Fuhr N et al (2018) An information nutritional label for online documents. ACM SIGIR Forum 51(3):46–66. https://doi.org/10.1145/3190580.3190588

Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: 2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 – Proceedings, S 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

Gravanis G, Vakali A, Diamantaras K, Karadais P (2019) Behind the cues: a benchmarking study for fake news detection. Expert Syst Appl 128:201–213. https://doi.org/10.1016/j.eswa.2019.03.036

Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Political science: fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425):374–378. https://doi.org/10.1126/science.aau2706

Gupta A, Kumaraguru P, Castillo C, Meier P (2014) TweetCred: real-time credibility assessment of content on Twitter. In: International conference on social informatics, S 228–243. http://arxiv.org/abs/1405.5490. Zugegriffen am 22.04.2021

Hanselowski A, Stab C, Schulz C, Li Z, Gurevych I (2019a) A richly annotated corpus for different tasks in automated fact-checking. In: Proceedings of the 23rd conference on computational natural language processing, S 493–503. https://doi.org/10.18653/v1/k19-1046

Hanselowski A et al (2019b) UKP-Athene: multi-sentence textual entailment for claim verification. In: Proceedings of the First Workshop on Fact Extraction and Verification (FEVER), S 103–108. https://doi.org/10.18653/v1/w18-5516

Hartwig K, Reuter C (2019) TrustyTweet: an indicator-based browser-plugin to assist users in dealing with Fake News on Twitter. In: Proceedings of the international conference on Wirtschaftsinformatik (WI). http://www.peasec.de/paper/2019/2019_HartwigReuter_TrustyTweet_WI.pdf. Zugegriffen am 18.04.2020

Jin Z, Cao J, Zhang Y, Luo J (2016) News verification by exBploiting conflicting social viewpoints in microblogs. In: 30th AAAI conference on Artificial Intelligence, AAAI 2016, Phoenix, S 2972–2978

Kahne J, Bowyer B (2017) Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation. Am Educ Res J 54(1):3–34. https://doi.org/10.3102/0002831216679817

Kaufhold M, Rupp N, Reuter C, Habdank M (2020) Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. Behav Inform Technol 39(3):319–342

Kelly Garrett R, Weeks BE (2013) The promise and peril of real-time corrections to political misperceptions. In: Proceedings of the ACM conference on Computer Supported Cooperative Work, CSCW, S 1047–1057. https://doi.org/10.1145/2441776.2441895

Kim A, Dennis A (2018) Says who?: how news presentation format influences perceived believability and the engagement level of social media users. In: Proceedings of the 51st Hawaii international conference on System Sciences. https://doi.org/10.24251/hicss.2018.497

Kirchner J, Reuter C (2020) Countering fake news: a comparison of possible solutions regarding user acceptance and effectiveness. In: Proceedings of the ACM: human computer interaction (PACM): computer-supported cooperative work and social computing, ACM, Austin, USA

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 13(3):106–131. https://doi.org/10.1177/1529100612451018

Long Y, Lu Q, Xiang R, Li M, Huang C-R (2017) Fake news detection through multi-perspective speaker profiles, Bd 2, 8. Aufl. In: Proceedings of the eighth international joint conference on Natural Language Processing, Taipei, S 252–256

McNally M, Bose L (2018) Combating false news in the Facebook news feed: Fighting abuse @ Scale. https://atscaleconference.com/events/fighting-abuse-scale/. Zugegriffen am 24.01.2020

Mihailidis P, Viotty S (2017) Spreadable spectacle in digital culture: civic expression, Fake News, and the role of media literacies in "post-fact" society. Am Behav Sci 61(4):441–454. https://doi.org/10.1177/0002764217701217

Monti F, Frasca F, Eynard D, Mannion D, Bronstein MM (2019) Fake news detection on social media using geometric deep learning. [Preprint]

Morris MR, Counts S, Roseway A, Hoff A, Schwarz J (2012) Tweeting is believing? Understanding microblog credibility perceptions. In: Proceedings of the ACM conference on Computer Supported Cooperative Work, CSCW, S 441–450. https://doi.org/10.1145/2145204.2145274.

Mosseri A (2016) Addressing hoaxes and fake news. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. Zugegriffen am 24.01.2020

Müller P, Denner N (2017) Was tun gegen „Fake News"? Friedrich Naumann Stiftung Für die Freiheit, Bonn

Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2:175–220

Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Polit Behav 32(2):303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan B, Reifler J, Ubel PA (2013) The hazards of correcting myths about health care reform. Med Care 51(2):127–132. https://doi.org/10.1097/MLR.0b013e318279486b

Pariser E (2011) The filter bubble: how the new personalized Web is changing what we read and how we think. Penguin, London

Pennycook G, Cannon TD, Rand DG (2018) Prior exposure increases perceived accuracy of fake news. J Exp Psychol Gen 147(12):1865–1880. https://doi.org/10.1037/xge0000465

Pennycook G, Bear A, Collins E (2019) The implied truth effect: attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. In: Management Science. http://www.mnsc.2019.3478.pdf. Zugegriffen am 18.04.2020

Pérez-Rosas V, Kleinberg B, Lefevre A, Mihal R (2017) Automatic detection of fake news. In: Proceedings of the 27th international conference on Computational Linguistics. https://www.aclweb.org/anthology/C18-1287. Zugegriffen am 22.04.2021

Potthast M, Kiesel J, Reinartz K, Bevendorff J, Stein B (2018) A stylometric inquiry into hyperpartisan and fake news. In: ACL 2018 – 56th annual meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers) Vol. 1, S 231–240. https://doi.org/10.18653/v1/p18-1022

Rapoza K (2017) Can „fake news" impact the stock market? In: Forbes. https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#33dc99c02fac. Zugegriffen am 24.01.2020

Rashkin H, Choi E, Jang JY, Volkova S, Choi Y (2017) Truth of varying shades: analyzing language in fake news and political fact-checking. In: EMNLP 2017 – conference on Empirical Methods in Natural Language Processing, Proceedings, S 2931–2937. https://doi.org/10.18653/v1/d17-1317

Reis JCS, Correia A, Murai F, Veloso A, Benevenuto F (2019) Explainable machine learning for fake news detection. In: WebSci 2019 – Proceedings of the 11th ACM conference on Web Science, S 17–26. Association for Computing Machinery, Inc. https://doi.org/10.1145/3292522.3326027

Reuter C, Hartwig K, Kirchner J, Schlegel N (2019) Fake news perception in Germany: a representative study of people's attitudes and approaches to counteract disinformation. In: Proceedings of the international conference on Wirtschaftsinformatik. http://www.peasec.de/paper/2019/2019_ReuterHartwigKirchnerSchlegel_FakeNewsPerceptionGermany_WI.pdf. Zugegriffen am 18.04.2020

Ribeiro MT, Singh S, Guestrin C (2016) „Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD international conference on Knowledge Discovery and Data Mining, S 1135–1144. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. In: International conference on Information and Knowledge Management, Proceedings, S 797–806. Association for Computing Machinery. https://doi.org/10.1145/3132847.3132877

Sally Chan M, Jones CR, Hall Jamieson K, Albarraci D (2017) Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 28(11):1531–1546. https://doi.org/10.1177/0956797617714579

Sängerlaub A (2017) Verzerrte Realitäten – Die Wahrnehmung von „Fake News" im Schatten der USA und der Bundestagswahl. Stiftung Neue Verantwortung, Berlin. https://www.stiftung-nv.de/sites/default/files/fake_news_im_schatten_der_usa_und_der_bundestagswahl.pdf. Zugegriffen am 18.04.2020

Sethi RJ (2017) Crowdsourcing the verification of fake news and alternative facts. In: HT 2017 – proceedings of the 28th ACM conference on Hypertext and Social Media, S 315–316. Association for Computing Machinery, Inc. https://doi.org/10.1145/3078714.3078746

Shu K, Bernard HR, Liu H (2019a) Studying fake news via network analysis: detection and mitigation. In: Emerging research challenges and opportunities in computational social network analysis and mining, S 43–65. https://doi.org/10.1007/978-3-319-94105-9_3

Shu K, Wang S, Liu H (2019b) Beyond news contents: the role of social context for fake news detection. In: WSDM 2019 – proceedings of the 12th ACM international conference on Web Search and Data Mining, S 312–320. Association for Computing Machinery, Inc. https://doi.org/10.1145/3289600.3290994

Stanoevska-slabeva K (2017) Teaching social media literacy with storytelling and social media curation. In: Twenty-third Americas conference on Information Systems, S 1. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1524&context=amcis2017. Zugegriffen am 18.04.2020

Steinebach M, Bader K, Rinsdorf L, Krämer N, Roßnagel A (2020) Desinformation aufdecken und bekämpfen: Interdisziplinäre Ansätze gegen Desinformationskampagnen und für Meinungspluralität, Bd 45, 1. Aufl. Nomos Verlagsgesellschaft mbH & Co. KG, Baden-Baden. https://doi.org/10.5771/9783748904816

Tacchini E, Ballarin G, Della Vedova ML, Moret S, de Alfaro L (2017) Some like it Hoax: automated fake news detection in social networks. In: CEUR Workshop Proceedings (Vol. 1960). https://developers.facebook.com/docs/graph-api. Zugegriffen am 22.04.2021

Tene O, Polonetsky J, Sadeghi A-R (2018) Five freedoms for the momo deus. IEEE Secur Priv 16(3): 15–17. https://ieeexplore.ieee.org/abstract/document/8395137/. Zugegriffen am 22.04.2021

Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151

Wanas N, El-Saban M, Ashour H, Ammar W (2008) Automatic scoring of online discussion posts. In: International conference on Information and Knowledge Management, Proceedings, S 19–25. https://doi.org/10.1145/1458527.1458534

Weerkamp W, De Rijke M (2008) Credibility improves topical blog post retrieval. In: ACL-08: HLT – 46th annual meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Columbus, S 923–931

Weimer M, Gurevych I, Mühlhäuser M (2007) Automatically assessing the post quality in online discussions on software. In: Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, S 125–128. Association for Computational Linguistics. https://doi.org/10.3115/1557769.1557806

Wu L, Liu H (2018) Tracing fake-news footprints: characterizing social media messages by how they propagate. In: WSDM 2018 – proceedings of the 11th ACM international conference on Web Search and Data Mining, S 637–645. Association for Computing Machinery, Inc. https://doi.org/10.1145/3159652.3159677

Yang F et al (2019) XFake: explainable fake news detector with visualizations. In: The Web Conference 2019 – proceedings of the World Wide Web Conference, WWW 2019, S 3600–3604. Association for Computing Machinery, Inc. https://doi.org/10.1145/3308558.3314119

Zhang X, Ghorbani AA (2020) An overview of online fake news: characterization, detection, and discussion. Inf Process Manag 57(2):102025. https://doi.org/10.1016/j.ipm.2019.03.004

Zhou X, Jain A, Phoha VV, Zafarani R (2019) Fake news early detection: a theory-driven model. Digit Threats Res Pract. http://arxiv.org/abs/1904.11679. Zugegriffen am 22.04.2021