Im Kampf gegen Hate Speech

Was tun soziale Netzwerke dagegen?

Jenny F. Schneider
Medienradar, 10/2020

Im Internet auf Hate Speech zu stoßen, gehört für viele Nutzer*innen bereits zum Alltag. Besonders in den sozialen Netzwerken tummeln sich viele Hassredner*innen, die gerne ihren Unmut verbreiten. Doch die sozialen Netzwerke reagieren – und zwar immer schneller und effektiver. Der nachfolgende Artikel gibt Einblicke in die Mechanismen der sozialen Netzwerke, um Hate Speech zu bekämpfen und macht deutlich, unter welch enormem zeitlichen Druck sie stehen. Denn je länger eine Hassrede online bleibt, desto höher ihre Reichweite.

Das Internet im (schier unmöglichen) Kampf gegen Hate Speech

Kinder und Jugendliche werden schon frühzeitig im Internet mit Hate Speech konfrontiert – oftmals sogar mehr als Erwachsene. Doch was tut das Internet eigentlich gegen Hate Speech? So simpel die Frage ist, so komplex ist deren Beantwortung. Denn „das Internet" ist unüberschaubar groß, die Anzahl der existierenden Websites beläuft sich auf über eine Milliarde (Stand 2020) und die Richtlinien und Gesetzeslagen hinsichtlich des Umgangs mit Hate Speech variieren nicht nur von Plattform zu Plattform, sondern auch von Land zu Land. So müssen selbst länderübergreifend genutzte Plattformen je nach Land der Nutzung ggf. unterschiedlich intervenieren.

Daher scheint es sinnvoll, die Frage etwas einzugrenzen und sich vornehmlich die Plattformen anzuschauen, auf denen Jugendliche in Deutschland am meisten unterwegs sind und am häufigsten mit Hate Speech konfrontiert werden – ob nun direkt oder indirekt. Der Fokus richtet sich daher auf Instagram, YouTube, Snapchat, TikTok, Twitter und Facebook.

Länderübergreifende Grundsätze der Plattformen

Generell kann erst einmal festgehalten werden, dass die von den Jugendlichen gern genutzten Plattformen in ihren jeweiligen Community-Richtlinien Hate Speech als Regelverstoß verankert haben – und zwar unabhängig davon, von welchem Land aus die Nutzer*innen die Plattform oder das App-Angebot nutzen. So halten Facebook und sein Tochterunternehmen Instagram in ihren gemeinsamen Gemeinschaftsrichtlinien fest, dass „Hassrede" gegen ihre Richtlinien verstößt und somit

nicht zulässig ist – dem schließt sich TikTok in seinen Community-Richtlinien an. Twitter verbietet in seinen Regeln und Richtlinien „Hass schürendes Verhalten“, in den Snapchat-Community-Richtlinien sind es „Hassbotschaften“, die nicht erlaubt sind, und YouTube widmet dem Thema „Richtlinien zu Hassrede“ sogar ein eigenes Video.

Viele der Plattformen bieten zudem eine ausführliche Definition des Begriffes Hate Speech. Aber wie gehen sie nun damit um, wenn dennoch Hate-Speech-Inhalte veröffentlicht werden?

Maßnahmen gegen Hate Speech – wie schnell reagieren die Plattformen?

„Hassrede schafft ein Umfeld der Einschüchterung und Ausgrenzung und kann offline Gewalt fördern. Deshalb lassen wir Hassrede auf Facebook nicht zu.“, so äußert sich Facebook einleitend zum Thema Hassrede in seinen Community-Standards. Um Hate Speech zu stoppen und einzudämmen, bieten die Plattformen den Nutzer*innen die Möglichkeit, Hate-Speech-Inhalte als Verstoß gegen die Community- bzw. Gemeinschaftsrichtlinien zu melden. Diese Meldungen werden geprüft, bewertet und – falls sie gegen die Richtlinien verstoßen – gelöscht oder, bei Twitter etwa, unsichtbar geschaltet.

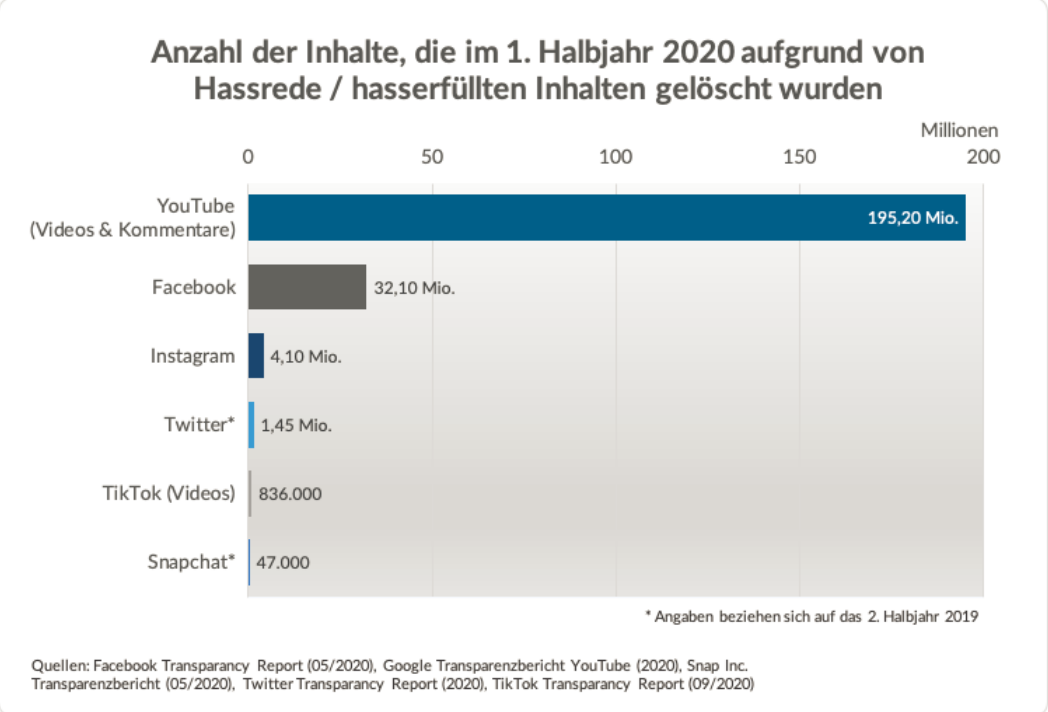
Darüber hinaus haben mittlerweile alle sechs Plattformen einen 2016 aufgesetzten Verhaltenskodex der Europäischen Kommission zur Bekämpfung illegaler Online-Hassreden unterzeichnet („EU Code of conduct on countering illegal speech online“), um gemeinsam gegen Hate Speech vorzugehen und Hate Speech-Inhalte möglichst schnell zu entfernen. Ein unabhängiger Bericht von der Europäischen Kommission evaluiert in regelmäßigen Abständen die Maßnahmen und die Effizienz der Plattformen; der fünfte und jüngste Bericht wurde im Juni 2020 veröffentlicht.

Aus diesem Bericht geht hervor, dass Facebook mit 95,7 % die meisten Hate-Speech-Meldungen, die im sechswöchigen Untersuchungszeitraum (04.11.–13.12.2019) durch insgesamt 39 Organisationen gemeldet wurden, in unter 24 Stunden bewertet hat. Instagram schaffte 91,8 %, YouTube etwa vier Fünftel (81,5 %) und Twitter drei Viertel (76,6 %). Während Facebook und YouTube einen Großteil der gemeldeten Inhalte gelöscht haben (87,6 % bzw. 79,7 %), wurden von Instagram nicht einmal die Hälfte (42 %) und von Twitter sogar nur etwa ein Drittel (35,9 %) gelöscht. Zu Snapchat wurden im Bericht keine Angaben gemacht und TikTok ist der Vereinbarung erst nach Veröffentlichung des fünften Berichts beigetreten. Insgesamt hat der Verhaltenskodex der Europäischen Kommission seit Beginn in 2016 zu erheblichen Fortschritten im Hinblick auf die Bekämpfung von Hate Speech beigetragen.

Doch wie entscheiden die Plattformen darüber, welche gemeldeten Inhalte letztlich gelöscht werden und welche nicht?

Löschen oder nicht löschen?

Um eine fundierte Entscheidung fällen zu können, müssen die Plattformen verschiedene Aspekte bewerten. Dafür haben sie ihre eigenen Kriterienkataloge entwickelt, die ihnen bei der klaren Abgrenzung von Hassrede helfen. Wie schon Facebook in einem Beitrag von 2017 verdeutlicht, ist eine klare Benennung von Hate Speech in Einzelfällen jedoch sehr schwierig, insbesondere, wenn der Kontext fehlt, die Intention nicht klar erkennbar ist, die gewählten Wörter mehrdeutig sind und somit zu unterschiedlichen Interpretationen führen können oder der Inhalt ggf. satirisch gemeint ist. So ist Hate Speech in bestimmten Kontexten sogar zulässig, zum Beispiel wenn es im pädagogischen Rahmen zu Lehrzwecken oder in dokumentarischen Zusammenhängen auftaucht und weder die Hate-Speech-Redner*innen noch deren Anschauungen unterstützt werden. Ein genauer Blick in den Kontext und ggf. eine Kommunikation mit dem*der entsprechenden Nutzer*in sind in solchen Fällen unerlässlich.



[Grafik: Medienradar]
YouTube liegt hinsichtlich der wegen Hate Speech gelöschten Inhalte ganz weit vorne, wobei die Kommentare (195 Mio.) den größten Teil ausmachen, während im gleichen Zeitraum nur 187.000 Videos aufgrund von hasserfüllten Inhalten gelöscht wurden.

Darüber hinaus verwenden viele Plattformen, so auch YouTube, Facebook, Instagram und TikTok, mittlerweile zusätzliche Technologien aus dem Bereich der künstlichen Intelligenz, um Hate-Speech-Inhalte automatisiert zu löschen, noch bevor sie

von Nutzer*innen gemeldet oder überhaupt gesehen werden. So sind fast 95 % der auf Facebook gelöschten Hate-Speech-Inhalte auf künstliche Intelligenz zurückzuführen (ca. 21 Millionen Inhalte), bei Instagram sind es etwa 84 % (ca. 2,8 Millionen Inhalte), wie aus dem aktuellen Facebook-Transparenz-Bericht aus dem zweiten Quartal 2020 hervorgeht. Beide Plattformen verzeichnen hinsichtlich ihrer Erkennungsraten durch künstliche Intelligenz so hohe Quoten wie noch nie zuvor.

Ob mittels menschlicher oder künstlicher Intelligenz entschieden – Fehler und Irrtümer sind nicht ausgeschlossen und bergen ein hohes Konfliktpotenzial, da fälschlicherweise gelöschte Inhalte schnell als Zensur kritisiert werden können und Inhalte, die trotz Meldung nicht gelöscht werden, den Vorwurf nach sich ziehen können, dass die Plattformen ihren eigenen Gemeinschaftsrichtlinien nicht genügend nachkommen. Die Nutzer*innen haben jedoch die Möglichkeit, die Entscheidungen der Plattform anzufechten, sodass sie einer erneuten und intensiveren Prüfung unterzogen werden.

Konsequenzen für die Nutzer*innen

Hinsichtlich der Konsequenzen für die Nutzer*innen setzen die Plattformen je nach Schweregrad des Verstoßes auf unterschiedliche Härtegrade des Bestrafens. Dabei spielt auch eine Rolle, ob es sich um einen erstmaligen oder einen wiederholten Regelverstoß handelt. So hat der erste Regelverstoß für YouTuber*innen noch keine Konsequenzen, beim wiederholten Verstoß erhalten sie eine Warnung, bei drei Warnungen wird ihnen der YouTube-Kanal gekündigt. Twitter macht vom Schweregrad des Regelverstoßes abhängig, ob es den Account vorübergehend in einen schreibgeschützten Modus versetzt (Nutzer*in kann nicht mehr tweeten, retweeten oder „Gefällt mir“ markieren) oder den Account dauerhaft sperrt. Und auch bei Facebook, Instagram, TikTok und Snapchat können mehrmalige Verstöße zu einem Deaktivieren oder Kündigen des Kontos führen. Eine erneute Anmeldung ist den Nutzer*innen dann für gewöhnlich untersagt.

Deutschland: anderes Land, andere Regeln und ein länderspezifisches Gesetz

Wie eingangs geschildert, gibt es einige länderspezifische Unterschiede hinsichtlich der Richtlinien zu Hate Speech, so zum Beispiel im Hinblick auf Anstiftung zum Hass, wie Richard Allen von Facebook erklärt: „In Germany, for example, laws forbid incitement to hatred; you could find yourself the subject of a police raid if you post such content online. In the US, on the other hand, even the most vile kinds of speech are legally protected under the US Constitution.“ (Facebook, 2017), übersetzt etwa: „In Deutschland zum Beispiel verbietet das Gesetz Anstiftung zum Hass. Man könnte plötzlich selbst zum Gegenstand einer Polizeirazzia werden, wenn man

solche Inhalte online veröffentlicht. In den USA hingegen sind selbst die abscheulichsten Reden gesetzlich durch die US-Verfassung geschützt."

Darüber hinaus ist in Deutschland 2017 das Netzwerkdurchsuchungsgesetz (NetzDG) in Kraft getreten, dessen Ziel es ist, Hasskriminalität, strafbare Falschnachrichten und andere strafbare Inhalte in sozialen Netzwerken schneller und wirksamer zu bekämpfen. Grundlage dafür sind gesetzliche Compliance-Regeln, an die sich soziale Netzwerke mit Gewinnerzielungsabsicht und mehr als zwei Millionen in Deutschland registrierten Nutzer*innen halten müssen. Gemäß den Compliance-Regeln müssen die Plattformen beanstandete und offensichtlich strafbare Inhalte innerhalb von 24 Stunden nach Meldungseingang sperren oder entfernen sowie halbjährig einen auf Deutsch verfassten Transparenz-Bericht über ihre Beschwerdeverfahren veröffentlichen. Aus diesem soll hervorgehen, wie viele Beschwerden eingegangen sind, wie schnell diese bearbeitet wurden, wie genau die Übermittlung von Beschwerden funktioniert und in welcher Form der*die Beschwerdeführer*in und der*die Nutzer*in, dessen*deren Inhalt beanstandet wurde, Rückmeldung erhalten.

Nutzer*innen in Deutschland können seither Inhalte in sozialen Netzwerken als Verstoß im Sinne des NetzDG melden, welches auf insgesamt 21 Straftatbestände verweist. Das Melden ist meist über einen eigenen Reiter im Meldebereich möglich oder über ein separates Meldeformular.

Die gemeldeten Inhalte werden dann einer zweistufigen Prüfung unterzogen. Zunächst wird geprüft, ob der gemeldete Inhalt allgemein gegen die Gemeinschaftsrichtlinien des sozialen Netzwerkes verstößt. Wenn dem so ist, greifen die bereits beschriebenen Maßnahmen: der Inhalt wird entfernt – und zwar weltweit. Verstößt der Inhalt nicht gegen die Gemeinschaftsrichtlinien, wird in einem zweiten Schritt geprüft, ob er gegen die im NetzDG aufgeführten Bestimmungen des deutschen Strafgesetzbuches verstößt. Folglich kann es passieren, dass Inhalte, die nicht gegen die Gemeinschaftsrichtlinien einer Plattform verstoßen, hingegen aber gegen die Bestimmungen des deutschen Strafgesetzbuches verstoßen, ausschließlich in Deutschland gesperrt werden, während sie in anderen Ländern online bleiben.

Zudem müssen die sozialen Netzwerke gemäß dem NetzDG gewährleisten, dass sowohl die beschwerdeführenden als auch die betroffenen Nutzer*innen, deren Inhalte gemeldet wurden, im Rahmen des NetzDG eine Rückmeldung erhalten – auch hinsichtlich der getroffenen Maßnahme.

Die Schwachstellen des NetzDG

Die Effizienz des NetzDG ist stark umstritten. Zwar werden die auf Grundlage des NetzDG gemeldeten Inhalte in den meisten Fällen sehr schnell bewertet, aber nur

ein recht geringer Teil davon wird tatsächlich entfernt. Ein vom Counter Extremism Project (CEP) Berlin durchgeführter Stresstest im Untersuchungszeitraum vom 31.01. bis 14.02.2020 hat ergeben, dass die sozialen Netzwerke zum Teil noch deutliche Schwachstellen beim Entfernen offensichtlich rechtswidriger Inhalte, die auf Grundlage des NetzDG gemeldet wurden, aufweisen. Die Sperrung der strafbaren Inhalte nach Beschwerde sei „unzureichend [...] und für die Reduzierung rechtswidriger Inhalte Online nicht ausreichend" (CEP Policy Paper, S. 1). Von den im Rahmen des Stresstests gemeldeten Inhalten, die als offensichtlich rechtswidrig einzustufen sind, hat YouTube nur etwa ein Drittel entfernt, während Facebook und Instagram alle vom CEP gemeldeten Inhalte entfernt haben. Möglicherweise ist dies darauf zurückzuführen, dass YouTube im gleichen Zeitraum vergleichsweise deutlich mehr NetzDG-Meldungen erhält (1. Halbjahr 2020: 388.824 Meldungen) als die anderen beiden Plattformen (Facebook: 4.292, Instagram: 2.025) und demnach hinsichtlich der Bearbeitung unter größerem Zeitdruck steht.

Die Effizienz im Hinblick auf Bekämpfung von Hate Speech bleibt dennoch fraglich, da die Plattformen die eingehenden Meldungen zwar reaktiv bearbeiten, aber nicht proaktiv nach strafbaren Inhalten suchen. So hat das CEP weiterhin feststellen müssen, dass zwar ein gemeldetes, strafbares Bild auf Facebook innerhalb weniger Stunden entfernt wurde, aber mehrere weitere Bilder mit ähnlichem strafbaren Motiv, die sich im gleichen Ordner desselben Facebook-Profils befanden, weiterhin online blieben – weil sie eben nicht gemeldet wurden.

Ein weiterer, großer Schwachpunkt des NetzDG sind die Messengerdienste (WhatsApp, Facebook-Messenger, Telegram u. a.), auf die das Gesetz keine Anwendung findet, da sie der Individualkommunikation zugeordnet werden und keine Gewinnerzielungsabsicht haben. Dies ist besonders problematisch, da Messengerdienste längst nicht mehr nur zur direkten Kommunikation, sondern auch zur Kommunikation in Gruppen verwendet werden. Besonders die App Telegram stellt eine Gefahr dar, da die Nutzer*innen dort neben der Privatkommunikation auch ganzen Kanälen von Personen folgen können, die aufgrund der Follower-Anzahl längst einen halb-öffentlichen Kommunikationsraum darstellen und damit durchaus zur Plattform für Hate Speech werden können – und zwar vollkommen unreguliert.

Quellennachweise

1. Allan, Richard: *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?* vom 27.06.2017, in: Facebook Newsroom, https://about.fb.com/news/2017/06/hard-questions-hate-speech/ (abgerufen am 13.10.2020).

2. Bundesministerium der Justiz und für Verbraucherschutz (BMJV): *Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsuchungsgesetz – NetzDG)*, Bundesgesetzblatt Jahrgang 2017 Teil I Nr. 61, ausgegeben zu Bonn am 7. September 2017,

www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/BGBl_NetzDG.pdf;jsessionid=1E76AE0F0CA264035E83A3AB6953ECA4.1_cid297?__blob=publicationFile&v=2 (abgerufen am 13.10.2020).

3. Bundesministerium der Justiz und für Verbraucherschutz (BMJV): Netzwerkdurchsuchungsgesetz, www.bmjv.de/DE/Themen/FokusThemen/NetzDG/NetzDG_node.html (abgerufen am 13.10.2020).

4. Commission européenne: *TikTok joins EU Code of Conduct against illegal online hate speech* vom 08.09.2020, https://ec.europa.eu/luxembourg/news/tiktok-joins-eu-code-conduct-against-illegal-online-hate-speech_fr (abgerufen am 13.10.2020).

5. Counter Extremism Project Berlin: *CEP Policy Paper – NetzDG 2.0 – Empfehlungen zur Weiterentwicklung des Netzwerkdurchsetzungsgesetzes (NetzDG) und Untersuchung zu den tatsächlichen Sperr- und Löschprozessen von YouTube, Facebook und Instagram*, veröffentlicht am 12.03.2020, https://www.counterextremism.com/sites/default/files/CEP%20NetzDG%202.0%20Policy%20Paper.pdf (abgerufen am 13.10.2020).

6. European Commission: *Countering illegal hate speech online – 5th evaluation of the Code of Conduct*, https://ec.europa.eu/info/sites/info/files/codeofconduct_2020_factsheet_12.pdf (abgerufen am 13.10.2020).

7. European Commission: The EU Code of conduct on countering illegal hate speech online, https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (abgerufen am 13.10.2020).

8. Facebook: Gemeinschaftsstandards – 12. Hassrede, https://www.facebook.com/communitystandards/hate_speech (abgerufen am 13.10.2020).

9. Facebook: *NetzDG Transparenzbericht*, 07/2020, https://about.fb.com/wp-content/uploads/2020/07/facebook_netzdg_July_2020_German.pdf (abgerufen am 13.10.2020).

10. Facebook Transparency Report: Community Standards Enforcement Report, 08/2020, https://transparency.facebook.com/community-standards-enforcement (abgerufen am 13.10.2020).

11. Google Transparenzbericht YouTube: *Entfernungen von Inhalten nach dem Netzwerkdurchsetzungsgesetz*, https://storage.googleapis.com/transparencyreport/legal/netzdg/YT-NetzDG-TR-Bundesanzeiger-latest.pdf (abgerufen am 13.10.2020).

12. Hemmes, Anne: *Messengerdienst Telegram: Wo die Verschwörungstheorien sprießen*, in: Br24 (Bayrischer Rundfunk), veröffentlicht am 28.05.2020, https://www.br.de/nachrichten/netzwelt/messengerdienst-telegram-wo-die-verschwoerungstheorien-spriessen,S0HDUq1 (abgerufen am 13.10.2020).

13. Instagram: Gemeinschaftsrichtlinien, https://help.instagram.com/477434105621119 (abgerufen am 13.10.2020).

14. Instagram: *NetzDG Transparenzbericht*, 07/2020, https://scontent-ber1-1.xx.fbcdn.net/v/t39.8562-6/116715787_2373553682941263_6359719088636124711_n.pdf?_nc_cat=109&_nc_sid=ae5e01&_nc_ohc=-eMUpc6qIDsAX9J9KmP&_nc_ht=scontent-ber1-1.xx&oh=bb852081bac0845a5b64a87511aad7f7&oe=5FAE10C3 (abgerufen am 13.10.2020).

15. Instagram: Nutzungsbedingungen, https://help.instagram.com/581066165581870 (abgerufen am 13.10.2020).

16. Netcraft: *January 2020 Web Server Survey*, https://news.netcraft.com/archives/2020/01/21/january-2020-web-server-survey.html (abgerufen am 13.10.2020).

17. Snap Inc.: Community-Richtlinien, https://www.snap.com/de-DE/community-guidelines (abgerufen am 13.10.2020).

18. Snap Inc.: Transparenzbericht, veröffentlicht am 27.05.2020, https://www.snap.com/de-DE/privacy/transparency (abgerufen am 13.10.2020).

19. TikTok: Community-Richtlinien, https://www.tiktok.com/community-guidelines?lang=de (abgerufen am 13.10.2020).

20. TikTok: Nutzungsbedingungen, https://www.tiktok.com/legal/terms-of-use?lang=de (abgerufen am 13.10.2020).

21. TikTok: TikTok Transparency Report, veröffentlicht am 22.09.2020, https://www.tiktok.com/safety/resources/transparency-report-2020-1?lang=en (abgerufen am 13.10.2020).

22. Twitter: Transparancy Report – Rules Enforcement, https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jul-dec (abgerufen am 13.10.2020).

23. Twitter Hilfe-Center: Richtlinie zu Hass schürendem Verhalten, https://help.twitter.com/de/rules-and-policies/hateful-conduct-policy (abgerufen am 13.10.2020).

24. Twitter Hilfe-Center: Unsere verschiedenen Durchsetzungsmaßnahmen, https://help.twitter.com/de/rules-and-policies/enforcement-options (abgerufen am 13.10.2020).

25. YouTube-Hilfe: Richtlinien zu Hassrede, https://support.google.com/youtube/answer/2801939?hl=de&ref_topic=9282436 (abgerufen am 13.10.2020).

Link zum Artikel

https://medienradar.de/hintergrundwissen/artikel/im-kampf-gegen-hate-speech