

**Developing Community Media to Mitigate
the Impact of Fake News**

**Modul 2–
Modelle der
Inhaltsmoderation &
Anwendbarkeit**



HALLO!

Überblick

Inhaltsmoderation ist ein Prozess, dem kaum Beachtung geschenkt wird und der bei der Internetnutzung oft unbemerkt bleibt. Dabei ist er notwendig, um Usern ein sicheres Surfen und Vergnügen zu gewährleisten.

Im Rahmen dieses Workshops erhältst du einen Überblick über den Begriff sowie die damit verbundenen Themen und Ansätze.

Themen

- Was bedeutet Moderation von Inhalten und wie wichtig ist sie?
- Modelle der Inhaltsmoderation
- Technische Tools für die Moderation von Inhalten
- Blacklists & Filter nach der Veröffentlichung
- Praktiken der Inhaltsmoderation in sozialen Medien

Der Workshop versucht, die folgenden Themen abzudecken:

- Was bedeutet Moderation von Inhalten und wie wichtig ist sie?
- Modelle der Inhaltsmoderation
- Technische Tools für die Moderation von Inhalten
- Blacklists & Filter nach der Veröffentlichung
- Praktiken der Inhaltsmoderation in sozialen Medien

lernziele

- Verstehen, was Inhaltsmoderation ist und ihre Notwendigkeit
- Kennenlernen verschiedener Moderationsarten und deren Vorteile
- Erfahren, wie man eine "Blacklist" erstellt und Filter nach der Veröffentlichung anwendet
- Die Praktiken von Social-Media-Diensten verstehen, um Auffälligkeiten in der Verbreitung zu erkennen

Durch die Nutzung des Trainingsangebots und des Materials zum selbstgesteuerten Lernen des Moduls werden die Lernenden:

Die verschiedenen Arten von Moderationen und ihre Vorteile kennenlernen
Grundkenntnisse über verschiedene technische Tools erwerben
In der Lage sein, Merkmale von absichtlich negativem und provokativem Social Media-Inhalt zu erkennen
Erfahren, wie man eine "Blacklist" erstellt und Filter nach der Veröffentlichung anwendet
Die Praktiken von Social-Media-Diensten verstehen, um Auffälligkeiten in der Verbreitung zu erkennen

Präsentation der Theorie

Moderation von
Inhalten und ihre
Bedeutung



Einführung

- Die zunehmende Verbreitung digitaler Technologien mit der Verbreitung des Internets hat zu einer "Demokratisierung der Information" geführt
- BürgerInnen verlassen sich nicht mehr nur auf traditionelle Medien (TV, Radio & Zeitungen) für Informationen
- Sie haben Zugang zu einer Fülle von Quellen sowie die Möglichkeit, ihre eigenen Gedanken und Erfahrungen einzubringen

Durch die Nutzung des Trainingsangebots und des Materials zum selbstgesteuerten Lernen des Moduls werden die Lernenden:

Die verschiedenen Arten von Moderationen und ihre Vorteile kennenlernen
Grundkenntnisse über verschiedene technische Tools erwerben
In der Lage sein, Merkmale von absichtlich negativem und provokativem Social Media-Inhalt zu erkennen
Erfahren, wie man eine "Blacklist" erstellt und Filter nach der Veröffentlichung anwendet
Die Praktiken von Social-Media-Diensten verstehen, um Auffälligkeiten in der Verbreitung zu erkennen

Einführung

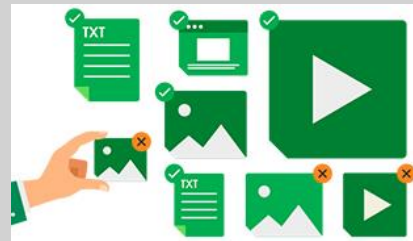
- Trotz des anfänglichen Optimismus wird diese Entwicklung nicht mehr als völlig positiv gesehen
- Das Internet wurde zu einem Raum, in dem bedrohliches und bösartiges Material verbreitet wurde.
- Es müssen Schritte unternommen werden, um sicherzustellen, dass die Inhalte, die von den Usern geteilt und aufgerufen werden, nicht schädlich sind.

Diese Entwicklung wurde mit großer Begeisterung und Hoffnung begrüßt. Viele AnalytikerInnen sahen in dieser Entwicklung den Anbruch eines großen Zeitalters, in dem jede einzelne Person besser informiert sein wird und der öffentliche Diskurs durch die aktive Teilnahme aller BürgerInnen und die Vertretung aller Standpunkte in hohem Maße profitieren wird. Dieser Optimismus wurde jedoch bald erschüttert, da die digitale Sphäre auch einen fruchtbaren Boden für den Austausch und das Gedeihen von bedrohlichem und oft hasserfüllem Material sowie die unkontrollierte und oft absichtliche Verbreitung von Fehlinformationen geboten hat.

Daher ist es für jede Form von Medien von größter Wichtigkeit, Verfahren zur Inhaltsmoderation einzuführen und zu etablieren, um sicherzustellen, dass die Inhalte, die veröffentlicht oder gehostet werden, nicht schädlich sind. Konkrete Schritte sind erforderlich, um sicherzustellen, dass potenziell schädliche Inhalte in Form von Material über Kindesmissbrauch, gewalttätigen und extremen Inhalten, Hassreden, grafischen Inhalten, sexuellen Inhalten, grausamen und unsensiblen Materialien und Spam-Inhalten blockiert werden (Ofcom: 2019). Zu diesem Zweck möchte das vorliegende Modul spezifische Medienmoderationsmodelle mit Ressourcen vorstellen, um Lernende mit Interesse an der Entwicklung von Community-Medien dabei zu unterstützen, solche Prozesse in ihrer Arbeit effektiv durchzuführen.

Präsentation der Theorie

Die Moderation von Inhalten im Überblick



Was ist Inhaltsmoderation?

- Das Cambridge Dictionary definiert Moderation als die Eigenschaft, etwas innerhalb von angemessenen Grenzen zu tun.
- Im Zusammenhang mit Medien bezieht sich dieses Prinzip auf die Festlegung von Regeln und Grenzen hinsichtlich der Informationen und des Materials, die geteilt oder gehostet werden dürfen
- Sie bezieht sich auf die Überwachung der hochgeladenen Inhalte und das Sperren oder Entfernen von Inhalten, die als nicht akzeptabel erachtet werden

Das Cambridge Dictionary definiert Moderation als die Eigenschaft, etwas innerhalb angemessener Grenzen zu tun. Im Kontext von Medieninhalten bezieht sich dieses Prinzip auf das Festlegen von Regeln und Grenzen für Informationen und Materialien, die von einer Plattform geteilt oder gehostet werden, sei es ein Medienunternehmen, eine Social-Media-Seite, eine Webseite oder ein Blog, und das Ergreifen von Maßnahmen, um sicherzustellen, dass diese eingehalten werden. Mit anderen Worten: Es geht darum, die Inhalte zu überwachen, die hochgeladen werden sollen oder bereits hochgeladen wurden, und keine Inhalte zuzulassen oder zu entfernen, die nicht mit dem Regelwerk der betreffenden Plattform vereinbar sind (Grimes-Viort: 2010). "Je nach den spezifischen Anforderungen der jeweiligen Seite kann die Moderation mehr oder weniger stark ausgeprägt sein. Einige Internetforen sind stolz auf ihre Meinungsfreiheit, während andere, wie z. B. soziale Netzwerke, eine schwierige Balance zwischen Benutzerfreundlichkeit und dem Schutz ihrer jüngeren User finden müssen" (Smith:2019).

Was ist Inhaltsmoderation?

- Die überwiegende Mehrheit der Social-Media-Plattformen und Medien (sowohl digitale als auch traditionelle) haben spezifische Moderationsrichtlinien, die bis zu einem gewissen Grad vorschreiben, welche Art von Inhalten für die Verbreitung akzeptabel ist
- Unter Einhaltung dieser Regeln und durch den Einsatz verschiedener Techniken und Ansätze zur Inhaltsmoderation überwachen und bewerten sie das Material, das bei ihnen geteilt wird

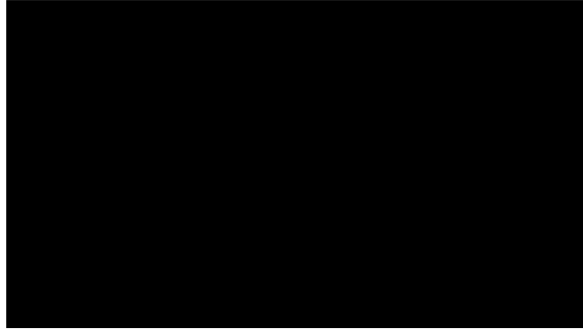
Die meisten beliebten sozialen Medien und Medienkanäle (sowohl digitale als auch traditionelle) haben spezifische Moderationsrichtlinien, die bis zu einem gewissen Grad vorschreiben, welche Art von Inhalten für die Verbreitung über ihre Plattform akzeptabel ist. Unter Einhaltung dieser Regeln und durch den Einsatz verschiedener Techniken und Ansätze zur Inhaltsmoderation überwachen und bewerten sie das geteilte Material, um sicherzustellen, dass alle Arten von beleidigenden oder anstößigen Inhalten wie Pornovideos, explizite Bilder oder Bilder, die nicht für alle Altersgruppen geeignet sind, nicht veröffentlicht oder entfernt werden (Cogito Tech: 2020).

Wozu eine Inhaltsmoderation?

- Es muss unbedingt sichergestellt werden, dass die User eines Mediums oder einer Plattform keinen anstößigen oder schädlichen Inhalten ausgesetzt werden
- Das Hosten oder Teilen von unangemessenem, schädlichem oder illegalem Verhalten verursacht nicht nur der Plattform/dem Medium Probleme, es kann auch ernste Folgen für das Publikum haben, insbesondere junge Menschen
- Daher ist besondere Achtsamkeit bei den betreffenden Medien und Plattformen äußerst wichtig.

Bildmoderation, Videomoderation und Textmoderation sind entscheidend, um sicherzustellen, dass das Publikum/die User eines Mediums oder einer Plattform keinen anstößigen oder schädlichen Inhalten ausgesetzt sind (Schomer: 2019). Das Hosten oder Teilen von unangemessenem, schädlichem oder illegalem Verhalten führt nicht nur zu Problemen für die Plattform/das Medium, sondern kann auch ernsthafte Auswirkungen auf das Publikum haben, insbesondere auf junge Menschen, die das Internet und vor allem soziale Medien nutzen, um Informationen und Nachrichten zu erhalten sowie für soziale Interaktionen. Daher ist es von äußerster Wichtigkeit, dass die betreffenden Medien und Plattformen wachsam sind und in die Überwachung und Moderation der über sie verfügbaren Inhalte investieren (puremoderation: 2020).

Inhaltsmoderation im Überblick





AKTIVITÄT Nr. I

Diese Bedrohungen müssen ModeratorInnen ansprechen

Schicke mit deinem Telefon auf das Mentimeter die Bedrohungen, die dir einfallen

Präsentation der Theorie: Modelle der Inhaltsmoderation



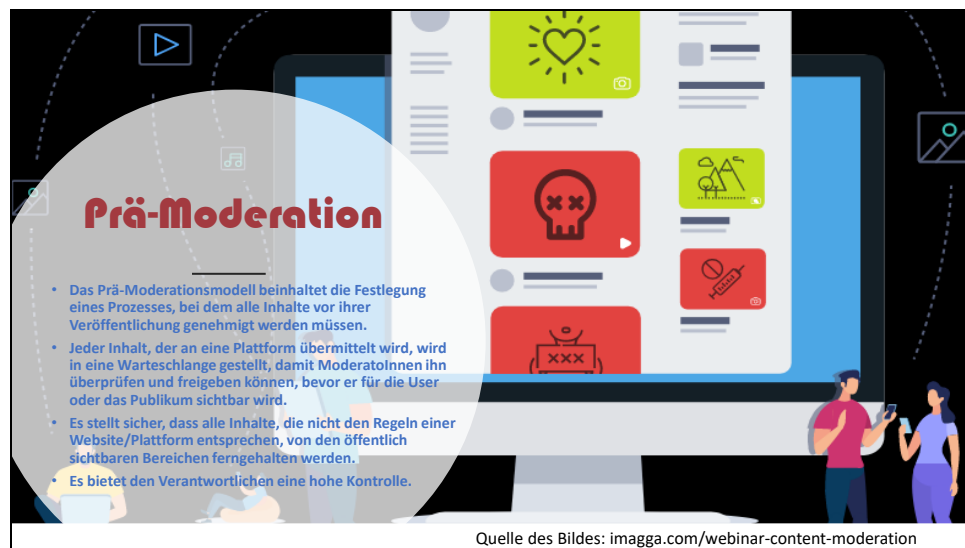
Es gibt verschiedene Ansätze oder Modelle der Inhaltsmoderation, die eine Plattform anwenden kann

Diese Modelle schließen sich bis zu einem gewissen Grad nicht gegenseitig aus, so dass es sich anbietet, mehrere Ansätze zu verfolgen.

Letztendlich muss jedes Unternehmen entscheiden, was für es am effektivsten ist und am besten funktioniert

Wenn es um die Moderation von Inhalten geht, gibt es verschiedene Ansätze oder Modelle, die eine Plattform je nach ihren eigenen Prinzipien, ihrem Modus Operandi, ihren Zielen und Arbeitsprozessen anwenden kann. Diese Modelle schließen sich bis zu einem gewissen Grad nicht gegenseitig aus, sodass man sich für mehrere Ansätze entscheiden kann. Letztendlich sollte man sich für einen Ansatz entscheiden, um das bestmögliche Überwachungs- und Filterergebnis zu erzielen, das das Publikum vor schädlichen und bedenklichen Inhalten schützt und gleichzeitig ein gewisses Maß an Ordnung in den Arbeitsabläufen aufrechterhält (Cogito Tech: 2019).

Die verschiedenen Arten der Inhaltsmoderation zu verstehen, zusammen mit ihren Stärken und Schwächen, kann ModeratorInnen helfen, jenes Modell zu wählen, das für ihren Zweck und ihre Arbeitsweise am besten geeignet ist.



Wie der Name schon sagt, beinhaltet das Prä-Moderationsmodell einen Prozess, bei dem alle Inhalte vor ihrer Veröffentlichung genehmigt werden müssen. In der Praxis bedeutet dies, dass jeder Inhalt, der an eine Plattform gesendet wird, in eine Warteschlange gestellt wird, damit ein/eine ModeratorIn ihn überprüfen und genehmigen kann, bevor er für die User sichtbar wird. Die Vormoderation stellt sicher, dass (in den Händen einer gut moderierenden Person) jeder Inhalt, der schädlich oder unangemessen ist oder nicht den Regeln einer Website/Plattform entspricht, von den öffentlich sichtbaren Bereichen ferngehalten wird. Dadurch bietet es der managenden-moderierenden Person eine hohe Kontrolle.

Nachteile der Prä-Moderation

- Kann zu Beschwerden seitens der Person führen, die den Inhalt einreicht
- Kann zu Verzögerungen führen, die sich negativ auf den Inhalt auswirken können, da er konversations- und zeitgebunden ist
- Großer Aufwand, der unüberschaubar werden kann

Das hat allerdings auch Nachteile, da es zu Beschwerden seitens der Person, die den Inhalt einreicht, oder zu Verzögerungen beim Hochladen führen kann, was sich negativ auf Inhalte auswirken kann, die unterhaltsam und zeitgebunden sind. Außerdem ist dies ein Prozess, der hohe Kosten verursacht, wenn eine Plattform wächst und die Beiträge einen Grenzwert an nutzergenerierten Inhalten überschreiten, der von den Moderierenden nicht mehr bewältigt werden kann (Grimes-Viort: 2010).

Post-Moderation

- Zulassen des Einstellens von Inhalten/Materialien auf einer Plattform ohne vorherige Prüfung und anschließende Bewertung
- Reduziert Verzögerungen auf ein Minimum und ermöglicht den schnelleren Betrieb einer Plattform und der auf/über ihr stattfindenden Aktivitäten
- Ermöglicht ModeratorInnen, Sicherheit zu gewährleisten, während Verhaltens- und rechtliche Probleme rechtzeitig erkannt und angegangen werden können



Quelle des Bildes: www.telusinternational.com/articles/mastering-content-moderation

Bei diesem Ansatz wird das Posten von Inhalten/Materialien auf einer Plattform ohne anfängliche Überprüfung erlaubt, woraufhin der/die Moderierende(n) diese bewerten und bei Bedarf entfernen können. Der Vorteil dieser Art der Moderation besteht darin, dass sie Verzögerungen minimiert und zu einem schnelleren Tempo der Plattform und der auf ihr/über sie stattfindenden Interaktionen beiträgt. "Die Menschen erwarten ein gewisses Maß an Unmittelbarkeit, wenn sie im Web interagieren, und die Post-Moderation ermöglicht dies, während die ModeratorInnen gleichzeitig sicherstellen können, dass Sicherheits-, Verhaltens- und rechtliche Probleme rechtzeitig erkannt werden und darauf reagiert werden kann" (Grimes-Viort: 2010).

Nachteile der Post-Moderation

- Kann dazu führen, dass einige User auf schädliche oder unangemessene Inhalte stoßen.
- Ein Prozess, der mit zunehmender Popularität und Reichweite einer Plattform und der damit verbundenen Zunahme der dort eingereichten Inhalte kostspielig werden kann.

Dennoch kann dies dazu führen, dass einige User auf schädliche oder unangemessene Inhalte stoßen. Darüber hinaus ist es ein Prozess, der kostspielig werden kann, wenn die Sichtbarkeit und Reichweite einer Plattform wächst und damit auch die dort eingereichten Inhalte zunehmen.



Reaktive Moderation ist ein Ansatz, der sich darauf verlässt, dass die User einen Inhalt melden, wenn sie ihn für unangemessen halten. Es wird ein Meldeverfahren eingerichtet, das den Usern die Möglichkeit gibt, den/die AdministratorIn oder Moderierende(n) zu informieren und darum zu bitten, dass der besagte Inhalt überprüft und entfernt wird. Die reaktive Moderation ist in sozialen Medien weit verbreitet, wo die User aktiv an der Moderation der Inhalte mitwirken können. Sie kann als eine von den Nutzern vorgenommene Moderation von Inhalten beschrieben werden und gibt Macht und Eigentum zweifellos in die Hände des User-Publikums einer Plattform. (Cogito Tech: 2019).


Nachteile der reaktiven Moderation

- Kann zu übermäßiger Abhängigkeit von den Usern und einer eingeschränkten Kontrolle führen.
- Die Moderation von Inhalten wird zu einer Frage der Interpretation und Diskussion zwischen den Usern.

Der Hauptvorteil dieser Moderationsmethode ist, dass sie sich an die Sichtbarkeit einer Plattform anpassen kann und es ihr ermöglicht, sich der Verantwortung für unangemessene Inhalte zu entziehen, die von Usern hochgeladen werden, solange sie auf Anfragen zur Überprüfung und Entfernung von solchem Material eingeht. Sie birgt jedoch ein Risiko, da sie zu einer übermäßigen Abhängigkeit von den Usern führt (Grimes-Viort: 2010).

Verteilte Moderation

- Verfahren zur Moderation von durch User generierten Inhalten, das auf einem Bewertungssystem basiert
- User stimmen darüber ab, ob Beiträge entweder mit den Erwartungen der Community oder mit den Nutzungsregeln übereinstimmen
- Sie trägt somit durch die Unterstützung von engagierten und erfahrenen ModeratorInnen zur Kontrolle von Kommentaren, bzw. Forenbeiträgen bei



Quelle des Bildes: www.eunagi.com/user-generated-content-moderation-challenges-beyond/

Eine Methode zur Moderation Inhalten, die User eingestellt haben und die auf einem Bewertungssystem basiert, mit dem die Mitglieder der Community darüber abstimmen, ob Beiträge entweder den Erwartungen der Community oder den Nutzungsregeln entsprechen. Es trägt somit, mit der Unterstützung von engagierten und erfahrenen ModeratorInnen, zur Kontrolle von Kommentaren oder Forenbeiträgen bei.

Nachteile der verteilten Moderation

- Verlässt sich zu sehr auf die Fähigkeit und Bereitschaft der Online-Community zur Selbstmoderation

Wichtig ist dabei zu berücksichtigen, dass dieser Ansatz auf die Fähigkeit und Bereitschaft der Online-Community zur Selbstmoderation angewiesen ist.

Automatisierte Moderation

- Beinhaltet die Einführung und Nutzung von technischen Werkzeugen und automatisierten Filterprozessen, um Inhalte und Einsendungen zu prüfen und zu bewerten
- Das gängigste Tool ist der Wortfilter, der einen Text auf zuvor definierte verbotene Wörter untersucht und diese entweder ersetzt oder den Text ganz sperrt.
- Ähnliche Tools gibt es für die Bild- und Videomoderation, die besagte Multimedia-Inhalte hinsichtlich unangemessener Titel kennzeichnen

Zusätzlich zu allen oben genannten, von Menschen betriebenen Moderationssystemen ist die automatisierte Moderation eine wertvolle Waffe im Arsenal der ModeratorInnen. Sie beinhaltet die Einführung und Nutzung von technischen Tools und automatisierten Filterprozessen, um Inhalte und Beiträge zu prüfen und zu kontrollieren. Das typischste Werkzeug ist der Wortfilter, der einen Text auf zuvor definierte verbotene Wörter untersucht und diese entweder ersetzt oder den Text ganz sperrt. Ähnliche Tools gibt es für die Bild- und Videomoderation, die besagte Multimedia-Inhalte auf unangemessene Bildunterschriften hin kennzeichnen. (Grimes-Viort: 2010).

Nachteile der automatisierten Moderation

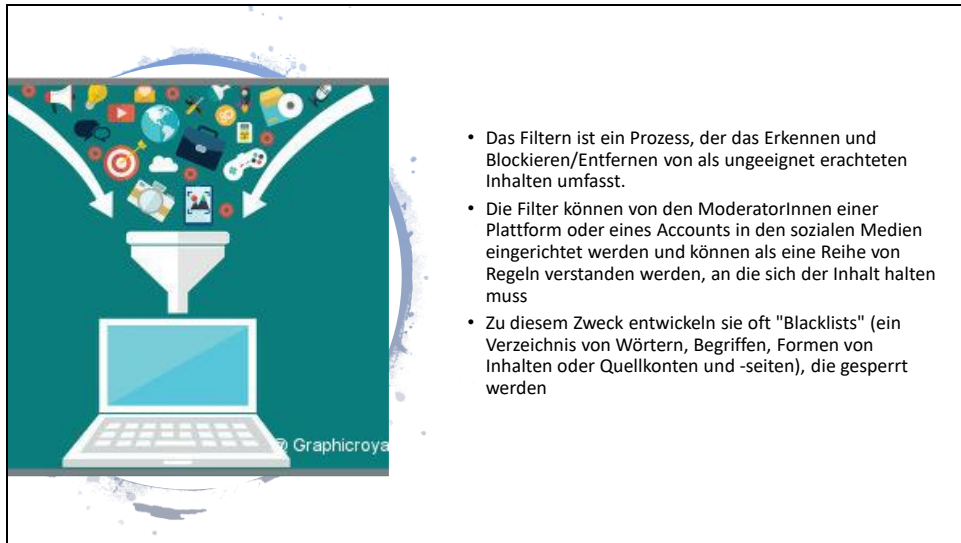
- Automatisierte Systeme können überlistet werden oder bei der automatischen Entfernung von schädlichem oder bedrohlichem Material versagen
- Können den Kontext oder die Herangehensweise des Materials, das sie überprüfen, nicht verstehen



AKTIVITÄT Nr. 2

**Erstellen einer Blacklist & Einstellen von
Publikationsfiltern**

Schicke mit deinem Telefon auf das Mentimeter die Bedrohungen, die dir einfallen



Filtern ist ein Prozess, der das Erkennen und Blockieren/Entfernen von Inhalten beinhaltet, die als unangemessen erachtet werden. Die Filter können von den ModeratorInnen einer Plattform oder eines Social-Media-Kontos eingerichtet werden und können als eine Reihe von Regeln verstanden werden, denen der Inhalt entsprechen muss, wenn er auf einer Plattform veröffentlicht werden oder angezeigt werden soll (Europarat: 2017). Zu diesem Zweck entwickeln sie oft "schwarze Listen" (ein Verzeichnis von Wörtern, Begriffen, Formen von Inhalten oder Quellkonten und Seiten), die für das Posten, Teilen oder Kommentieren auf der Plattform oder dem Konto gesperrt werden.

- Content-ModeratorInnen erstellen ein Archiv von Wörtern oder Begriffen, die für ihre Plattform nicht zulässig sind, sowie eine Liste von Websites, E-Mail-Konten oder Usern, die für das Posten oder Teilen auf ihrer Plattform gesperrt werden sollen
- Dies ist ein Prozess, der durch Tools und Plugins unterstützt werden kann, die diese Arbeit automatisch ausführen
- Diese Tools unterstützen ModeratorInnen bei der fachgerechten Durchführung dieser Aufgabe sowie bei der regelmäßigen Aktualisierung ihrer Sperrliste durch Hinzufügen oder Entfernen von Begriffen, Accounts, Inhaltsquellen

An illustration featuring a red outline of an open umbrella on a white background. The background is filled with various red symbols and text, including the word 'SPAM', '+18', 'XXX', and symbols like a heart with a slash, a skull and crossbones, and a radiation symbol. The umbrella is positioned in the center, with its handle curving upwards. The overall theme is digital content moderation and filtering.

Mit anderen Worten: Content-ModeratorInnen können ein Archiv von Wörtern oder Begriffen erstellen, die für ihre Plattform nicht zulässig sind, sowie eine Liste von Websites, E-Mail-Konten oder Usern, die für das Posten oder Teilen auf ihrer Plattform gesperrt werden sollen. Dies ist ein Prozess, der durch Tools und Plugins unterstützt werden kann, die diese Arbeit automatisch ausführen können und ModeratorInnen dabei unterstützen, diese Aufgabe fachmännisch zu erledigen und ihre Standards in ihren Plattformen oder Konten aufrechtzuerhalten sowie ihre Blacklist/Blockliste regelmäßig zu aktualisieren, indem sie Begriffe, Konten, Inhaltsquellen usw. hinzufügen oder entfernen.



Heutzutage verlassen wir uns auf die **Internetmedien**, um Informationen häufiger zu erhalten als über traditionelle Nachrichtenquellen.

Online-Foren erlauben es uns, Wissen zu teilen, Kommentare abzugeben, Nachrichten zu schreiben oder ein bestimmtes Thema mit anderen Menschen mit den gleichen Interessen zu diskutieren.

Soziale Netzwerke oder Websites, wie Facebook, Instagram, Twitter, YouTube, Tumblr, LinkedIn, Snapchat, Quora, Reddit, Pinterest, werden von vielen Menschen auf der ganzen Welt genutzt und ermöglichen es uns, dort beliebige Informationen zu finden und zu teilen.

Podcasts ermöglichen es allen, ihr Wissen zu teilen und mit der Welt durch eine Reihe von Audios zu kommunizieren, die sich auf ein bestimmtes Thema oder eine Thematik konzentrieren.

Tools der Inhaltsmoderation

Der erhöhte Bedarf an Moderation von Inhalten, die über Online-Plattformen geteilt werden, hat zur Entwicklung einer Vielzahl von Tools und Plattformen geführt, die die ModeratorInnen/AdministratorInnen bei diesem Prozess unterstützen.

Diese Tools reichen von KI-gesteuerten Filterprozessen bis hin zu einfachen Plugins, die ihnen helfen, eine gewisse Kontrolle und Übersicht zu behalten.

Schließlich kann eine Plattform beschließen, keine Form der Moderation oder Überwachung einzurichten. Dies ist nicht ratsam, da das Fehlen von Moderation in der Praxis bedeutet, dass der/die EigentümerIn/AdministratorIn einer Plattform keine Kontrolle über die Inhalte hat, was dazu führen kann, dass sie zu einem Tummelplatz für unangemessenes, bedrohliches oder illegales Material und Ansichten wird.

Tools der Inhaltsmoderation

Kommentar-Moderations-Tools für Blogger

Disqus

Ein Plugin, mit dem AdministratorInnen die Kommentare überprüfen und moderieren sowie User-Sperrlisten, Wortfilter und Spam-Kontrollen erstellen und verwalten können.

Facebook-Kommentare

Ein Plugin, das seinen Usern erlaubt, ihre Kommentare mit ihren Facebook-Profilen zu verknüpfen und den ModeratorInnen erlaubt, Kommentare über die Facebook-App oder über einen Browser zu moderieren und Massenaktionen auf Kommentare anzuwenden.

Disqus

Discus ist ein Tool, das einfach über einen Drop-in-Code oder als Plugin in einem Blog installiert werden kann und es AdministratorInnen ermöglicht, die Kommentare zu den Artikeln über ein einziges Dashboard zu überprüfen und zu moderieren. Das Dashboard bietet ihnen auch die Möglichkeit, unter anderem User-Bannlisten, Wortfilter und Spam-Kontrollen zu erstellen.

Facebook-Kommentare

Blogs können das Facebook-Kommentar-Plugin verwenden, das es ihren Usern ermöglicht, ihre Kommentare mit ihren Facebook-Profilen zu verknüpfen, ohne sich im besagten Blog registrieren/anmelden zu müssen. Mit diesem Tool können die ModeratorInnen die Kommentare nach dem Zeitpunkt des Postings oder nach ihrem Engagement ordnen. Das Wichtigste ist, dass sie die Möglichkeit haben, Kommentare über die Facebook-App oder über einen Browser zu moderieren und Massenaktionen auf Kommentare anzuwenden. Sie können auch entscheiden, ob sie Kommentare, die von Facebook oder anderen Usern markiert wurden (z. B. weil sie beleidigend oder anstößig sind), als "öffentlich" zulassen oder ausblenden. Wenn die Moderation einer Seite/eines Blogs von einem Team übernommen wird, können die ModeratorInnen schließlich auch bestimmte Kommentare verschiedenen Teammitgliedern zuweisen.

Tools der Inhaltsmoderation

Kommentar-Moderations-Tools für Blogger

IntenseDebate

Es ermöglicht ModeratorInnen, Kommentare zu überprüfen und per E-Mail zu beantworten sowie das Moderieren innerhalb eines Teams von AdministratorInnen aufzuteilen. Diese können auch Filter anwenden, nach Schlüsselwörtern, IP-Adressen oder E-Mail-Adressen suchen und/oder Kommentare automatisch löschen und bei Bedarf User sperren.

Livefyre

Livefyre ermöglicht ModeratorInnen die Überprüfung von Kommentaren vor deren Veröffentlichung, die Bearbeitung durch User sowie die Erstellung von Sperrlisten für User und Regeln für die automatische Verwaltung von Kommentaren

IntenseDebate

IntenseDebate ist ein Tool, das mit einer Vielzahl von Website-Plattformen wie Wordpress, Tumplr oder Blogger synchronisiert werden kann. Es ermöglicht ModeratorInnen, Kommentare zu überprüfen und per E-Mail zu beantworten sowie die Aufgabe der Moderation innerhalb eines Teams von AdministratorInnen aufzuteilen. Das Wichtigste ist, dass die Funktionen den Moderatoren die Möglichkeit bieten, Kommentare nach Schlüsselwort, IP-Adresse oder E-Mail-Adresse zu filtern, zu suchen und/oder automatisch zu löschen und bei Bedarf User zu sperren.

Livefyre

Livefyre verfügt über eine Vielzahl von anpassbaren Einstellungen, mit deren Hilfe der Moderationsprozess automatisch durchgeführt werden kann. ModeratorInnen haben die Möglichkeit, die Kommentare vor der Veröffentlichung zu überprüfen, sie von Usern bearbeiten zu lassen sowie Sperrlisten zu erstellen, die bestimmte User am Kommentieren hindern. Zusätzlich können sie Regeln für die Kommentare festlegen, so dass z.B. ein Inhalt, der von mehreren Usern markiert wurde, automatisch gelöscht wird. Das Gleiche gilt für den Profanityfilter (Schimpfwortfilter) des Netzwerks.

Tools der Inhaltsmoderation

Tools zum Umgang mit Missbrauch in sozialen Medien - Twitter

Stummschalten

Beim Empfang von Online-Beschimpfungen über Twitter sollten ModeratorInnen die betreffenden Konten stummschalten, anstatt sie zu sperren. Diese Vorgehensweise mildert die Auswirkungen des Missbrauchs, da das betroffene Konto keine Benachrichtigungen von dem stummgeschalteten Konto erhält.

Blockieren

Das letzte Mittel gegen Konten, die andauernd Spam oder anstößige Inhalte versenden.

Tools der Inhaltsmoderation

Tools zum Umgang mit Missbrauch in sozialen Medien - Twitter

Melden

ModeratorInnen melden in der Regel Tweets oder Accounts an Twitter, die potenziell glaubwürdige und unmittelbare Drohungen verbreiten oder gewalttätiges Bildmaterial enthalten.

Tools der Inhaltsmoderation

Tools zum Vorgehen gegen Missbrauch in sozialen Medien - Facebook

Kommentar löschen

Facebook bietet SeitenadministratorInnen die Möglichkeit, einen Kommentar zu löschen, den sie als beleidigend, bedrohlich oder abwertend empfinden.

Kommentar ausblenden

ModeratorInnen können sich dafür entscheiden, einen beleidigenden Kommentar in einem Beitrag lediglich auszublenden. Dies wird jedoch als weniger effektiv angesehen als das Löschen, da der betreffende Kommentar dadurch für den User und seine Freunde sichtbar bleibt.

Tools der Inhaltsmoderation

Tools zum Vorgehen gegen Missbrauch in sozialen Medien - Facebook

Jemanden von der Seite sperren

Wenn ein User wiederholt Kommentare postet, die gegen die Standards einer Seite verstoßen und die Werte einer gesunden Diskussion untergraben, dann ist es ratsam, ihn zu sperren.

Kommentare deaktivieren/ausschalten

Diese Funktion steht nur bei Videobeiträgen zur Verfügung und ist eine Maßnahme, die von ModeratorInnen gewählt wird, wenn sie das Gefühl haben, dass sie nicht genügend Kapazitäten haben, um den Fluss der Kommentare zu einem Video oder Live-Stream erfolgreich zu überwachen.

Tools der Inhaltsmoderation

Tools zum Vorgehen gegen Missbrauch in sozialen Medien - Facebook

Wörter blockieren

ModeratorInnen können die Wirksamkeit des Schimpfwortfilters ihrer Seite durch das Sperren bestimmter Wörter beeinflussen und so sicherstellen, dass Kommentare, die diese Wörter enthalten, nicht auf ihrer Seite veröffentlicht werden.

Melden

Wenn eine moderierende Person der Meinung ist, dass ein Kommentar/Beitrag gegen die Facebook-Standards verstößt, kann sie den User oder die Seite melden, die ihn abgegeben hat.

Tools der Inhaltsmoderation

Automatisierte Tools/Plattformen zur Moderation von Inhalten

Akismet

Akismet ist ein Plugin, das am besten zu WordPress-basierten Plattformen passt und sich hauptsächlich auf das Blockieren von Spam konzentriert. Es kann auch als Scan-Tool verwendet werden, um Beiträge und Seiten zu überwachen sowie Kommentare zu überprüfen, damit sichergestellt wird, dass keine bösartigen Inhalte auf einer Website erscheinen. Mehr unter:

<https://akismet.com/>

Utopia AI Moderator

Utopia AI Moderator ist ein vollautomatisches Tool zur Überwachung und Unterbindung von missbräuchlichen, betrügerischen und Spam-Inhalten. Es erlaubt AdministratorInnen, ihre Moderationsrichtlinien zu definieren und kann Inhalte in mehreren Sprachen bearbeiten. Mehr unter:

<https://utopiaanalytics.com/utopia-ai-moderator/>

Akismet

Akismet ist ein Plugin, das am besten zu WordPress-basierten Plattformen passt und sich hauptsächlich auf das Blockieren von Spam konzentriert. Es kann auch als Scan-Tool verwendet werden, um Beiträge und Seiten zu überwachen sowie Kommentare zu überprüfen, damit sichergestellt wird, dass keine bösartigen Inhalte auf einer Website erscheinen. Erwähnenswert ist, dass dieses Plugin sowohl benutzerfreundlich als auch kostenlos ist.

Mehr unter: <https://akismet.com/>

Utopia AI Moderator

Utopia AI Moderator ist ein vollautomatisches Moderationstool, das speziell dafür entwickelt wurde, missbräuchliche, betrügerische und Spam-Inhalte zu überwachen und zu beseitigen. Das Tool moderiert die eingehenden Inhalte zu 100 % und es bleibt auf dem neuesten Stand, indem es während der Arbeit aus den Veröffentlichungsentscheidungen lernt, die menschliche ModeratorInnen in der Vergangenheit getroffen haben. Es erlaubt AdministratorInnen, ihre Moderationsrichtlinien zu definieren, die es automatisch befolgt und kann Inhalte in mehreren Sprachen bearbeiten, Links und Adressen verarbeiten. Utopia AI Moderator ist auch in der Lage, den Kontext eines zu bewertenden Inhalts zu verstehen, unabhängig von der Sprache und der Rechtschreibung, während es eine automatische Moderation von Videos und Bildern bietet.

Mehr unter: <https://utopiaanalytics.com/utopia-ai-moderator/>

Tools der Inhaltsmoderation

Automatisierte Tools/Plattformen zur Moderation von Inhalten

Implio

Ein Tool, das eine automatisierte und manuelle Inhalts- sowie Kommentar-Moderation unterstützt. Es bietet eine manuelle Moderation über eine Schnittstelle zur Steuerung und ermöglicht es ModeratorInnen außerdem, eine breite Palette von Automatisierungsregeln festzulegen, um den gesamten Prozess zu rationalisieren.

Mehr unter: <https://besedo.com/implio-features/>

PicPurify

PicPurify ist eine Echtzeit-Bildmoderations-API, die entwickelt wurde, um Bilder mit unerwünschten Inhalten automatisch zu erkennen und zu filtern. Sie ist in der Lage, störende Elemente in Bildern zu identifizieren und ModeratorInnen können die von ihnen gewünschten Parameter einstellen.

Mehr unter: <https://www.picpurify.com/>

Implio

Implio ist ein Tool, das sowohl die automatisierte als auch die manuelle Moderation von Inhalten und Kommentaren unterstützt. Es bietet eine Schnittstelle für die manuelle Moderation, um den Prozess zu steuern, und ermöglicht es den Usern, ihre benutzerdefinierten Filter einzurichten, um unerwünschte Inhalte mit vorhersehbaren Mustern leicht zu finden und zu entfernen. Darüber hinaus können sie eine breite Palette von Automatisierungsregeln einrichten, um den gesamten Prozess zu rationalisieren.

Mehr unter: <https://besedo.com/implio-features/>

PicPurify

PicPurify ist eine Echtzeit-Bildmoderations-API, die entwickelt wurde, um Bilder mit unerwünschten Inhalten automatisch zu erkennen und zu filtern. Sie ist in der Lage, negative Elemente in Bildern wie Nacktheit, Drogen, Hass zu identifizieren und sicherzustellen, dass sie nicht auf einer Plattform erscheinen. Die ModeratorInnen sind in der Lage, einen maßgeschneiderten Ansatz zu erstellen, der ihren Bedürfnissen und ihrer Perspektive entspricht, und das Tool arbeitet auf einer 24-Stunden-Basis, um sicherzustellen, dass ihre Anforderungen erfüllt werden.

Mehr unter: <https://www.picpurify.com/>

Tools der Inhaltsmoderation

Automatisierte Tools/Plattformen zur Moderation von Inhalten

WebPurifys automatisierte intelligente Moderation

Ein Tool, das einen Rund-um-die-Uhr-Schutz vor den mit User-generierten Bildern verbundenen Risiken bietet, indem es Nacktheit und andere unangemessene Inhalte in Echtzeit erkennt und entfernt. Mehr unter:

<https://www.webpurify.com/photo-moderation/automated/>

Azure Content Moderator

Ein erkenntnisorientierter Dienst, der Inhalte in verschiedenen Formen wie Text-, Bild- und Videoinhalte überwacht, prüft und entsprechende Kennzeichnungen (Flaggen) für Material anbringt, das anstößig, riskant oder anderweitig unerwünscht sein könnte. Mehr unter: <https://docs.microsoft.com/en-in/azure/cognitive-services/content-moderator/overview>

WebPurifys automatisierte intelligente Moderation

Der automatisierte intelligente Moderationsdienst von WebPurify (AIM) bietet einen Rund-um-die-Uhr-Schutz vor den Risiken, die mit von Usern generierten Bildern verbunden sind, indem Nacktheit und andere unangemessene Inhalte in Echtzeit erkannt und entfernt werden. Die Tools können Bilder erkennen, die Nacktheit, Alkohol, Drogen, anstößige Gesten und Hass-Symbole und -Texte enthalten, um sicherzustellen, dass der Inhalt einer Plattform vor solchen anstößigen oder betrügerischen Inhalten "sicher" ist. Darüber hinaus bietet es benutzerdefinierte Moderationsmodelle, die sich an die Bedürfnisse der jeweiligen Plattform anpassen und diese am besten bedienen können.

Mehr unter: <https://www.webpurify.com/photo-moderation/automated/>

Azure Content Moderator

Azure Content Moderator ist ein erkenntnisbasierter Dienst, der Inhalte in verschiedenen Formen wie Text-, Bild- und Videoinhalte überwacht, prüft und entsprechende Kennzeichnungen (Flaggen) für Material anbringt, das beleidigend, riskant oder anderweitig unerwünscht sein könnte. Der Content-Moderator-Dienst umfasst auch das webbasierte Review-Tool, das die Inhaltsüberprüfungen hostet, die von menschlichen ModeratorInnen bearbeitet werden. Durch die Kombination der Arbeit des Dienstes mit den menschlichen Überprüfungsteams können die PlattformmoderatorInnen die richtige Balance zwischen Effizienz und Genauigkeit finden. Das Review-Tool bietet außerdem eine benutzerfreundliche Oberfläche für verschiedene Content-Moderator-Ressourcen.

Mehr unter: <https://docs.microsoft.com/en-in/azure/cognitive-services/content-moderator/overview>



Geistiges Eigentum

Videoquelle: <https://www.youtube.com/watch?v=UqZJPuyK9VY>

Die Notwendigkeit der Inhaltsmoderation in sozialen Medien

- Social-Media-Plattformen sind zu den wichtigsten Quellen für Beiträge und Informationen geworden
- Dies hat zwar zu einer Revolution in der Art und Weise geführt, wie Nachrichten und Informationen konsumiert und geteilt werden, aber sie haben auch ein Vakuum für die Verbreitung von Desinformation, Hass, Cybermobbing und bösartigen Inhalten geschaffen.
- Infolgedessen ist die Nachfrage nach Social-Media-Plattformen gestiegen, gegen solche Formen von Inhalten vorzugehen, was sie dazu veranlasst, Schritte zur Überwachung und Moderation zu unternehmen.

Social-Media-Plattformen haben sich zu den wichtigsten Quellen für Inhalte und Informationen entwickelt. Ein großer Teil der User verlässt sich auf Social Media, um auf Informationen, Nachrichten und Standpunkte zuzugreifen und sich mit Gleichgesinnten auszutauschen. Während dies zu einer Revolution in der Art und Weise geführt hat, wie Nachrichten und Informationen konsumiert, geteilt und erörtert werden, haben Social-Media-Plattformen auch ein Vakuum für die Verbreitung von Desinformation, Hass, Cybermobbing und bösartigen Inhalten geschaffen.

Infolgedessen ist die Nachfrage nach Social-Media-Plattformen gestiegen, gegen solche Formen von Inhalten vorzugehen, was sie dazu veranlasst, Schritte zur Überwachung und Moderation zu unternehmen.

Reaktion der Social-Media-Plattformen

- Social Media-Plattformen haben Community-Standards entwickelt, die festlegen, was akzeptabel ist und was nicht
- Sie setzen aufwändige Content-Moderationsstrategien ein, die sowohl menschliche ModeratorInnen als auch automatisierte Systeme umfassen, um sicherzustellen, dass ihre User so weit wie möglich vor anstößigen Inhalten und Missbrauch geschützt werden.
- Sie bitten User, Inhalte zu melden, die sie als anstößig empfinden oder die gegen die Regeln der Plattform verstoßen

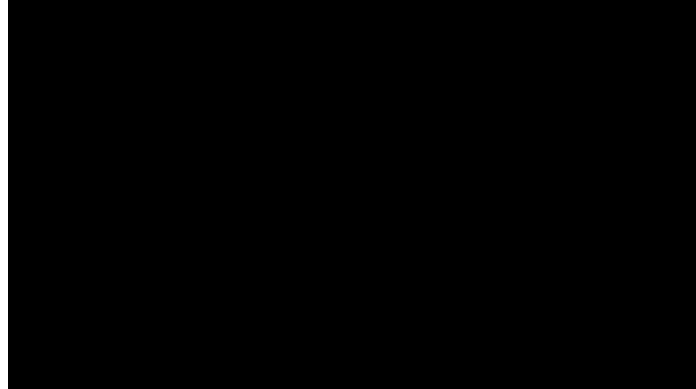
Social-Media-Plattformen haben sich zu den wichtigsten Quellen für Inhalte und Informationen entwickelt. Ein großer Teil der User verlässt sich auf Social Media, um auf Informationen, Nachrichten und Standpunkte zuzugreifen und sich mit Gleichgesinnten auszutauschen. Während dies zu einer Revolution in der Art und Weise geführt hat, wie Nachrichten und Informationen konsumiert, geteilt und erörtert werden, haben Social-Media-Plattformen auch ein Vakuum für die Verbreitung von Desinformation, Hass, Cybermobbing und böartigen Inhalten geschaffen.

Infolgedessen ist die Nachfrage nach Social-Media-Plattformen gestiegen, gegen solche Formen von Inhalten vorzugehen, was sie dazu veranlasst, Schritte zur Überwachung und Moderation zu unternehmen.







Moderation & freie Meinungsäußerung in sozialen Medien



Wenn es um die Moderation von Inhalten in sozialen Medien geht, kommen die Redefreiheit und die Angst vor Zensur oder Kontrolle der geteilten Meinungen zur Sprache. Sieh dir dieses Video als Anregung für eine Diskussion an, die du in der Gruppe führen kannst




Vielen Dank!




If you exchange information internationally, you must strengthen data protection. Those are two sides of the same coin.


— Gijs de Vries —


Vielen Dank!





Wires Crossed


 **dante**
ISTITUTO NAZIONALE PER L'ADULT EDUCATION


 **ALK**
ADULT EDUCATION INSTITUTE


 **Speha Fresia**
SOCIETÀ COOPERATIVA

 **JUGEND- & KULTURPROJEKT EV.**

 **The Rural Hub**

 **CARDET**
CENTRE FOR THE ADVANCEMENT OF RESEARCH & DEVELOPMENT IN EDUCATIONAL TECHNOLOGY

 **ACUMEN TRAINING**

 Co-funded by the
Erasmus+ Programme
of the European Union

*Die Unterstützung der Europäischen Kommission für die Erstellung dieser Veröffentlichung stellt keine Billigung des Inhalts dar, welcher nur die Ansichten der Verfasser wiedergibt, und die Kommission kann nicht für eine etwaige Verwendung der darin enthaltenen Informationen haftbar gemacht werden. 2019-1-DE02-KA204-006115