





Im Rahmen des Landesaktionsplans "Wir in Niedersachsen. Für Vielfalt. Gegen Rassismus."



Automatisierten Moderation auf Social Media

Detaillierte Schritt-für-Schritt-Erklärungen

Social Media lebt von **Austausch**, **Diskussion** und **Vernetzung** – gerade für politische Content Creator*innen und Community Influencer*innen. Gleichzeitig können **Hass**, **Trolle** und **Desinformationen** schnell zur **Belastung** werden. **Moderation** ist deshalb **wichtig**, aber oft **aufwendig** und **nervenraubend**.

Diese Sammlung von Anleitungen bietet **detaillierte Schritt-für-Schritt-Erklärungen** zur Einrichtung und Nutzung der automatisierten Moderations-Tools auf den Social-Media-Plattformen: YouTube, TikTok, Twitch und den Meta-Plattformen (Facebook & Instagram).

YouTube

YouTube bietet eine Reihe von leistungsstarken Werkzeugen, um die Kommentarbereiche deines Kanals zu verwalten und eine positive Community zu fördern. Diese Anleitung führt dich durch die verschiedenen Einstellungen und zeigt dir, wie du die automatisierte Moderation effektiv nutzt.



1. Zugriff auf die Moderationseinstellungen

Alle zentralen Moderationseinstellungen findest du im YouTube Studio.

- 1. Öffne das **YouTube Studio**.
- 2. Klicke im linken Menü auf Einstellungen.
- 3. Wähle den Tab Community aus → hier findest du die "Automatisierten Filter" und die "Standardeinstellungen".

2. Automatisierte Filter - digitale Türsteher*innen

Die **automatisierten Filter** sind das **Herzstück** der **Moderation** auf YouTube. Hier legst du die **Regeln** fest, nach denen **Kommentare automatisch behandelt** werden.

Genehmigte Nutzer*innen (Approved Users):

Füge hier Kanäle von **vertrauenswürdigen Nutzer*innen** hinzu. Kommentare von diesen Personen werden immer **sofort veröffentlicht** und **umgehen** alle **Filter** (einschließlich der Wortfilter). Das ist ideal für andere Creator*innen, mit denen du **kollaborierst**, oder für **langjährige**, **positive Mitglieder** deiner **Community**.

Ausgeblendete Nutzer*innen (Hidden Users):

Wenn du Nutze*innen hier hinzufügst, werden die Kommentare und Live-Chat-Nachrichten nirgendwo auf deinem Kanal mehr angezeigt. Die Person kann weiterhin Kommentare schreiben, aber nur sie selbst wird diese sehen. Dies ist die effektivste Methode, um Trolle oder wiederholte Störer*innen unsichtbar zu machen, ohne sie direkt zu blockieren.

Blockierte Wörter (Blocked Words):

Dies ist eine der wichtigsten Funktionen. Hier kannst du eine **Liste** von **Wörtern** und **Phrasen** erstellen, die du in deinem **Kommentarbereich nicht sehen** möchtest.

- Funktionsweise: Kommentare, die eines dieser Wörter oder eine sehr ähnliche Schreibweise enthalten, werden automatisch zur Überprüfung zurückgehalten. Sie erscheinen nicht öffentlich, bis du sie manuell freigibst.
- **Einrichtung**: Gib die Wörter oder Phrasen durch Kommas getrennt in das Feld ein. Du musst sowohl die Einzahl als auch die Mehrzahl und mögliche Variationen hinzufügen, um die Effektivität zu maximieren.
 - → **Beispiel**: *beleidigung*, *beleidigungen*, *dummes video*, *hass*
- **Live-Chat**: Diese Liste gilt auch für den Live-Chat. Nachrichten mit diesen Begriffen werden dort ebenfalls blockiert.

Links blockieren (Block links):

Setze hier ein **Häkchen**, um zu **verhindern**, dass neue **Kommentare** mit **Hashtags** und **URLs** veröffentlicht werden. Diese werden **ebenfalls** zur **Überprüfung** zurückgehalten. Dies ist eine extrem wirksame Methode, um **Spam** und **Eigenwerbung** zu unterbinden.

Ausnahmen sind:

- Deine eigenen Kommentare
- Kommentare deiner Moderator*innen
- Kommentare genehmigter Nutzer*innen

3. Standardeinstellungen für Kommentare

Unter dem Reiter **Standardeinstellungen** legst du fest, wie Kommentare bei **neuen Videos** und auf deinem **Community-Tab standardmäßig** behandelt werden sollen.

Kommentare zu deinen neuen Videos:

- Alle Kommentare zulassen: Jeder Kommentar wird sofort veröffentlicht (nicht empfohlen, wenn du viel Spam erhältst).
- Potenziell unangemessene Kommentare zur Überprüfung zurückhalten: Dies ist die empfohlene Standardeinstellung. YouTubes KI versucht, Spam, Beleidigungen und andere unangemessene Inhalte zu erkennen und hält sie zur manuellen Überprüfung zurück. Du kannst die Strenge dieses Filters erhöhen, indem du die Option "Strenge erhöhen" aktivierst.
- Alle Kommentare zur Überprüfung zurückhalten: Jeder einzelne Kommentar muss von dir oder Moderator*innen manuell freigeschaltet werden. Dies gibt dir maximale Kontrolle, erfordert aber auch den größten Aufwand.
- Kommentare deaktivieren: Niemand kann Kommentare unter deinen neuen Videos hinterlassen.

Kommentare in deinem Community-Tab:

Die gleichen vier Optionen wie oben gelten auch für deine Beiträge im Community-Tab.

4. Manuelle Überprüfung: Der "Zur Überprüfung zurückgehalten"-Tab Alle Kommentare, die von den automatisierten Filtern oder deinen Standardeinstellungen abgefangen wurden, landen im YouTube Studio unter dem Menüpunkt Kommentare im Tab "Zur Überprüfung zurückgehalten".

Hier hast du für jeden Kommentar folgende Möglichkeiten:

- Genehmigen (Häkchen): Der Kommentar wird öffentlich sichtbar.
- Entfernen (Papierkorb): Der Kommentar wird endgültig gelöscht.
- Melden (Fähnchen): Der Kommentar wird an YouTube zur Überprüfung auf einen Verstoß gegen die Community-Richtlinien gemeldet.
- Nutzer*in auf dem Kanal ausblenden (Durchgestrichenes Nutzersymbol): Der*die Nutzer*in wird zur Liste der "Ausgeblendeten Nutzer" hinzugefügt.

5. Moderator*innen ernennen

Du musst die **Moderation** nicht alleine bewältigen. Du kannst **vertrauenswürdige Nutzer*innen** zu Moderator*innen ernennen, die dir helfen.

- Standard-Moderator*in: Kann Kommentare im "Zur Überprüfung zurückgehalten"-Tab überprüfen und Live-Chat-Nachrichten entfernen.
- Verwaltende*r Moderator*in: Hat mehr Rechte und kann auch die Filtereinstellungen (z.B. blockierte Wörter) anpassen und andere Moderator*innen hinzufügen oder entfernen.

So fügst du eine*n Moderator*innen hinzu:

- 1. Gehe zu Einstellungen > Community.
- 2. Füge die Kanal-URL des Nutzers in das Feld "Moderatoren" ein.

TikTok

TikTok bietet eine Vielzahl von **Werkzeugen**, um die **Kontrolle** über deinen **Kommentarbereich** zu behalten und eine **sichere** und **positive Umgebung** für deine **Community** zu schaffen. Diese **Anleitung** führt dich durch alle **wichtigen Einstellungen** zur Kommentar-Moderation.



1. Grundlegende Kommentareinstellungen: Wer darf kommentieren?

Zuerst legst du fest, welche Nutzer*innengruppe überhaupt

Kommentare unter deinen Videos hinterlassen darf. Diese Einstellung gilt für deinen gesamten Account.

So findest du die Einstellungen:

- 1. Öffne die TikTok-App und gehe zu deinem Profil.
- 2. Tippe auf das Menü-Symbol (≡) oben rechts und wähle Einstellungen und Datenschutz.
- 3. Gehe zu Datenschutz und tippe dann auf Kommentare.

Unter "Kommentare zulassen von" hast du folgende Optionen:

- Alle (nur für öffentliche Konten): Jeder auf TikTok kann deine Videos kommentieren.
- Follower*innen (nur für private Konten): Nur Personen, die dir folgen, können kommentieren.
- Freunde (Follower*innen, denen du ebenfalls folgst): Nur gegenseitige Follower*innen können kommentieren.
- **Niemand**: Kommentare werden für alle deine Videos global deaktiviert.

2. Kommentar-Filter: Dein intelligentes Schutzschild

TikToks Filter-Optionen sind mächtig und erlauben eine **granulare Kontrolle**. Du findest sie ebenfalls unter Einstellungen und **Datenschutz** > **Datenschutz** > **Kommentare**.

Alle Kommentare filtern:

Wenn du diese Option aktivierst, wird **jeder einzelne Kommentar** unsichtbar geschaltet, bis du ihn **manuell überprüfst** und **genehmigst**. Dies bietet **maximale Kontrolle**, ist aber auch sehr **zeitaufwendig**.

Unerwünschte Kommentare filtern (Creator Care Mode):

Dies ist eine **KI-gestützte Funktion**, die automatisch Kommentare herausfiltert, die als **unangemessen**, **beleidigend** oder als **Spam** eingestuft werden. Sie **lernt** mit der Zeit dazu und **berücksichtigt** auch Kommentare, die in der Vergangenheit von **dir** oder anderen **Nutzer*innen gemeldet** oder **gelöscht** wurden.

→ Es wird dringend empfohlen, diese Funktion zu aktivieren.

Kommentare mit Schlüsselwörtern filtern:

Ähnlich wie bei YouTube kannst du hier eine **eigene Liste** mit **Wörtern** und **Phrasen** erstellen, die du **nicht** sehen möchtest.

- 1. Aktiviere "Schlüsselwörter in Kommentaren filtern".
- 2. Tippe auf "Schlüsselwörter hinzufügen" und gib die gewünschten Begriffe ein.

Kommentare, die diese **Schlüsselwörter** enthalten, werden zur **Überprüfung** zurückgehalten.

3. Überprüfung der gefilterten Kommentare

Alle Kommentare, die von deinen **Filtern** abgefangen wurden, findest du hier:

- 1. Gehe zu Einstellungen und Datenschutz > Datenschutz > Kommentare.
- 2. Tippe auf "Gefilterte Kommentare überprüfen".

Hier siehst du eine Liste **aller zurückgehaltenen Kommentare** und kannst für jeden **einzelnen entscheiden**:

- Genehmigen: Der Kommentar wird öffentlich sichtbar.
- Löschen: Der Kommentar wird endgültig entfernt.

4. Kommentare für einzelne Videos verwalten

Du kannst die **Kommentareinstellungen** auch für **jedes Video** individuell anpassen.

Vor dem Posten eines neuen Videos:

- 1. Im **letzten Schritt** vor dem **Veröffentlichen**, auf dem Posten-Bildschirm, tippe auf **Weitere Optionen**.
- 2. Hier kannst du den **Schalter** für **Kommentare zulassen** ein- oder ausschaltem.

Für ein bereits veröffentlichtes Video:

- 1. Gehe zu dem entsprechenden Video.
- 2. Tippe auf die **drei Punkte** (...) für **weitere Optionen**.
- 3. Wische nach links und wähle **Datenschutzeinstellungen**.
- 4. Schalte Kommentare zulassen ein oder aus.

5. Massen-Löschung und Meldung von Kommentaren

Wenn du mit einer **Welle** von **unerwünschten Kommentaren** konfrontiert wirst, bietet **TikTok** eine nützliche Funktion zur **Massenverwaltung**:

- 1. Öffne den Kommentarbereich eines Videos.
- 2. Tippe auf das **Filter-Symbol** oben oder halte einen **Kommentar lange gedrückt** und wähle "**Mehrere Kommentare**" verwalten.
- -Du kannst nun bis zu 100 Kommentare auswählen, um sie gleichzeitig zu löschen oder an TikTok zu melden.

6. Moderation im LIVE-Stream

Während eines **TikTok LIVE-Streams** kannst du **Moderator*innen** ernennen, die dir helfen, den **Chat sauber** zu halten.

- Zugriff: Vor oder während deines LIVE-Streams kannst du über die Einstellungen im Chat-Panel auf die Moderationseinstellungen zugreifen.
- **Funktionen**: Du kannst Moderator*innen hinzufügen oder entfernen. Diese können dann Kommentare filtern, stummschalten oder Nutzer*innen aus dem LIVE-Stream blockieren.

Twitch mit AutoMod

Twitch's AutoMod ist ein entscheidendes Werkzeug für jede*n Streamer*in, um eine sichere und einladende Chat-Umgebung zu schaffen. Es nutzt maschinelles Lernen, um potenziell schädliche Nachrichten proaktiv zurückzuhalten, damit sie von Moderator*innen überprüft werden können. Diese Anleitung erklärt detailliert, wie du AutoMod konfigurierst und optimal nutzt.



1. Was ist AutoMod und wie funktioniert es?

AutoMod ist kein Bot, der Nutzer*innen bannt oder timeoutet. Seine einzige Funktion ist es, **Nachrichten**, die es als **riskant einstuft**, **abzufangen** und für **menschliche Moderator*innen** zur Überprüfung sichtbar zu machen. Die Moderator*innen **entscheiden** dann, ob die Nachricht **zugelassen** (Allow) oder **abgelehnt** (Deny) wird.

- Transparenz für Nutzer*innen: Der*die Absender*in der Nachricht erhält eine Benachrichtigung, dass seine Nachricht zur Überprüfung zurückgehalten wird.
- **Keine Bestrafung**: AutoMod selbst bestraft keine Nutzer*innen. Moderator*innen behalten jedoch die volle Kontrolle und können eine*n Nutzer*in nach einer abgelehnten Nachricht weiterhin timeouten oder bannen.

2. Zugriff auf die AutoMod-Einstellungen

Die Konfiguration von AutoMod erfolgt im **Creator Dashboard**.

- 1. Gehe zu deinem <u>Creator Dashboard</u>.
- 2. Klicke auf das **Hamburger-Menü** (**≡**) oben links und dann auf **Einstellungen > Moderation**.
- 3. Unter dem Abschnitt **AutoMod-Steuerung** findest du alle relevanten Optionen.

3. Die AutoMod-Stufen: Finde deine Balance

Twitch bietet ein einfaches Stufensystem, um die allgemeine Strenge von AutoMod festzulegen. Du kannst eine von fünf Stufen wählen:

- **Stufe 0**: Keine Filterung. (Deine manuell blockierten Begriffe sind weiterhin aktiv).
- Stufe 1 (Standard): Filtert hauptsächlich Diskriminierung.
- **Stufe 2**: Filtert zusätzlich sexuelle Inhalte und verstärkt die Prüfung auf Anfeindungen.
- Stufe 3: Eine noch strengere Filterung der genannten Kategorien.
- **Stufe 4** (**Stärkste Einstellung**): Maximale Filterung in allen Kategorien, einschließlich Obszönitäten/ Vulgärsprache.

Empfehlung: Beginne mit Stufe 1 oder 2 und beobachte, wie sich dein Chat verhält. Du kannst die Stufe jederzeit anpassen.

4. Individuelle Konfiguration: Die Moderationskategorien

Für eine präzisere Steuerung kannst du jede der vier Hauptkategorien individuell anpassen, anstatt eine globale Stufe zu wählen. Jede Kategorie kann auf einer Skala von 0 (keine Filterung) bis 4 (stärkste Filterung) eingestellt werden.

- **Diskriminierung, Beleidigungen und Verunglimpfungen**: Filtert Hassrede basierend auf Herkunft, Religion, Geschlecht etc.
- **Sexuelle Inhalte**: Filtert Nachrichten mit sexuellen Anspielungen oder Begriffen.
- Anfeindungen: Filtert Mobbing, Provokationen und Belästigungen (Inzivilität)
- **Obszönitäten, Vulgärsprache**: Filtert Schimpfwörter und Flüche. Ideal für einen familienfreundlichen Stream.

5. Blockierte und Erlaubte Begriffe: Deine persönliche Filterliste

Dies ist der Bereich, in dem du die Automatisierung von AutoMod mit deinen eigenen Regeln verfeinerst.

Blockierte Begriffe und Phrasen:

Hier fügst du Wörter oder Phrasen hinzu, die du unter keinen Umständen in deinem Chat sehen möchtest, bspw. um persönliche Informationen zu schützen, spezifische Beleidigungen zu blockieren oder bekannte Spam-Phrasen zu filtern.

- Wildcards: Du kannst das Sternchen (*) als Platzhalter verwenden. Wenn du hate* blockierst, werden auch "haters" und "hateful" blockiert. Dies ist extrem nützlich, um Variationen eines Wortes zu erfassen.
- Links filtern: Um alle Links zu blockieren, kannst du einen Eintrag wie *.com* hinzufügen, um die meisten URLs abzufangen.
- **Privat vs. Öffentlich**: Du kannst Begriffe als "Privat" markieren. Nur du kannst diese sehen, nicht deine Moderator*innen. Das ist nützlich für sehr persönliche Informationen.

Erlaubte Begriffe und Phrasen:

Dies ist die Whitelist. Wenn AutoMod ein harmloses Wort fälschlicherweise immer wieder filtert (ein "False Positive"), kannst du es hier eintragen. Nachrichten, die dieses Wort enthalten, werden dann von AutoMod ignoriert (solange sie keine anderen blockierten Begriffe enthalten).

Automatisches Lernen:

- Wenn Moderator*innen eine von AutoMod zurückgehaltene Nachricht ablehnen, wird der problematische Begriff temporär zur Liste der blockierten Begriffe hinzugefügt.
- Wenn Moderator*innen eine Nachricht zulassen, wird der Begriff temporär zur Liste der erlaubten Begriffe hinzugefügt.

Bei wiederholter Aktion verlängert sich die Dauer von einer Stunde über einen Tag und sieben Tage bis hin zu einer permanenten Aufnahme in die jeweilige Liste.

6. Zugriff für Moderator*innen

Deine Moderator*innen spielen eine entscheidende Rolle. Sie können die **AutoMod-Einstellungen** ebenfalls einsehen und anpassen.

- **Zugriff**: Im Chatfenster auf das Zahnrad-Symbol (Einstellungen) klicken und Moderationseinstellungen verwalten auswählen.
- Mod-Ansicht (Mod View): Dieses spezielle Dashboard für Moderator*innen zeigt alle von AutoMod zurückgehaltenen Nachrichten in einem eigenen Widget an, was die Überprüfung erheblich erleichtert.

Durch die Kombination der intelligenten Automatisierung von AutoMod mit deinen benutzerdefinierten Wortlisten und einem engagierten Moderator*innen kannst du eine positive und sichere Community auf Twitch aufbauen und pflegen.

Meta-Plattformen (Facebook & Instagram)

Meta bietet über die Meta Business Suite und direkt auf den Plattformen integrierte Werkzeuge zur automatisierten Moderation von Kommentaren. Das wichtigste native Tool für Facebook ist der Moderations-Assistent (Moderation Assist). Diese Anleitung zeigt dir, wie du diese Tools für Facebook und Instagram effektiv einsetzt.



Teil 1: Facebook-Moderation mit dem Moderations-Assistenten

Der **Moderations-Assistent** ist dein persönlicher, **KI-gestützter Helfer**, der Kommentare auf deiner Facebook-Seite **automatisch** nach von dir **festgelegten Kriterien** verwaltet.

Wichtiger Hinweis: Dieses Tool funktioniert aktuell nur für organische Beiträge, nicht für Werbeanzeigen (Ads).

1. Zugriff auf den Moderations-Assistenten

- 1. Wechsle zu deiner Facebook-Seite.
- 2. Klicke im linken Menü auf das **Professional Dashboard**.
- 3. Scrolle nach unten zum Bereich "**Deine Tools**" und wähle **Moderations-Assistent**.

2. Kriterien für die automatische Moderation festlegen

Im Moderations-Assistenten kannst du verschiedene Regeln hinzufügen, die Kommentare automatisch verbergen. Der Kommentar ist dann nur noch für den Verfasser*innen und dessen Freund*innen sichtbar, nicht aber für die allgemeine Öffentlichkeit. Du kannst ihn später manuell überprüfen.

Klicke auf "Hinzufügen", um neue Kriterien zu erstellen. Diese basieren entweder auf dem*der Autor*in des Kommentars oder dem Inhalt des Kommentars.

Kriterien basierend auf dem*der Autor*in:

- **Neues Konto**: Verbirgt Kommentare von Profilen, die erst vor Kurzem erstellt wurden (z.B. weniger als eine Woche alt).
- Kein Profilbild: Filtert Kommentare von Nutzer*innen ohne Profilbild
 ein häufiges Merkmal von Troll- oder Fake-Konten.
- Keine Freund*in oder Follower*in: Verbirgt Kommentare von isolierten Konten, was ebenfalls auf mangelnde Authentizität hindeuten kann.
- Wiederholter Verstoß: Verbirgt Kommentare von Nutze*innen, deren Kommentare in der Vergangenheit bereits mehrfach von dir oder deinen Admins gemeldet oder gelöscht wurden.

Kriterien basierend auf dem Inhalt:

- Link im Kommentar: Eine der effektivsten Regeln, um Spam und unerwünschte Werbung zu unterbinden. Du kannst dies auch auf Links zu bestimmten Websites beschränken.
- **Bild oder Video im Kommentar**: Verbirgt Kommentare, die Medien enthalten.
- Obszönitäten: Nutzt eine von Facebook vordefinierte Liste von Schimpfwörtern, um Kommentare zu verbergen. Die Verfügbarkeit der Sprachen kann variieren.
- Schlüsselwörter im Kommentar: Hier kannst du deine eigene, durch Kommas getrennte Liste von Wörtern und Phrasen definieren, die automatisch zum Verbergen eines Kommentars führen.

3. Überprüfung der verborgenen Kommentare

Alle vom Moderations-Assistenten verborgenen Kommentare findest du im Aktivitätsprotokoll.

- 1. Gehe zurück zum **Moderations-Assistenten** im **Professional Dashboard**.
- 2. Klicke auf **Aktivitätsprotokoll**.

Hier kannst du die **verborgenen Kommentare** einsehen und **entscheiden**, ob du die **Aktion rückgängig** machen (den Kommentar wieder sichtbar schalten), den Kommentar löschen oder den*die Nutzer*in blockieren möchtest.

Teil 2: Moderation auf Instagram

Die **Moderations-Tools** von **Instagram** sind eng mit denen von Facebook verknüpft und bieten **ähnliche Funktionen**, die du direkt in der **App** einstellen kannst.

1. Zugriff auf die Kommentarsteuerung

- 1. Gehe zu deinem Instagram-Profil.
- 2. Tippe auf das **Menü-Symbo**l (≡) und wähle **Einstellungen** und **Privatsphäre**.
- 3. Scrolle zu "Interaktionen" und tippe auf Kommentare.

2. Wer darf kommentieren?

1. Unter "Kommentare zulassen von" kannst du einschränken, wer deine Beiträge kommentieren darf. Die Optionen sind selbsterklärend und reichen von "Allen" bis zu "Personen, denen du folgst".

3. Erweiterte Kommentarfilterung (Hidden Words)

Dies ist das Kernstück der automatisierten Moderation auf Instagram. Du findest es unter "Verborgene Wörter".

- Antworten auf Vorschläge verbergen: Instagram kann Kommentare verbergen, die seinen vordefinierten Listen mit potenziell anstößigen Begriffen entsprechen. Aktiviere diese Option für einen grundlegenden Schutz.
- Erweiterte Kommentarfilterung: Wenn du dies aktivierst, werden noch mehr Kommentare, die beleidigend sein könnten, automatisch verborgen. Dies gilt auch für Kommentare in deinen Direktnachrichten-Anfragen.
- Benutzerdefinierte Wörter und Ausdrücke verwalten: Hier erstellst du deine eigene, durch Kommas getrennte Liste von Wörtern, Zahlen und Emojis, die du nicht sehen möchtest. Kommentare, die diese Begriffe enthalten, werden automatisch verborgen.

4. Überprüfung der verborgenen Kommentare

Verborgene Kommentare werden in einem separaten Bereich am unteren Rand deines Kommentarbereichs angezeigt. Sie sind als "Verborgene Kommentare anzeigen" gekennzeichnet. Nur du kannst sie sehen und entscheiden, ob du sie genehmigst, löschst oder den*die Nutzer*in meldest.

