Universität zu Köln

Philosophische Fakultät

Institut für Digital Humanities

Aufbaumodul 1: Seminar: Anwendungen der Computerlinguistik

Dozent: Prof. Dr. Nils Reiter

Wintersemester 24/25

Abgabedatum: 22.03.2025

Automatische Erkennung von Bot-Kommentaren auf YouTube

Verfasser des Essays:

Niklas Halft

Matrikelnummer: 7396949

BA: Geographie/Informationsverarbeitung

5. Semester

Hilgener Straße 7

51067 Köln

nhalft@smail.uni-koeln.de

Inhaltsverzeichnis

Abstract

Seit der Einführung von YouTube stellt die Kommentarfunktion eine zentrale Interaktionsmöglichkeit für Nutzer dar, um Feedback zu Videoinhalten zu geben und sich untereinander auszutauschen. Nun hat jedoch der Einsatz von automatisch generierten Kommentaren durch Sprachmodelle seit ein paar Jahren auf allen Social-Media-Plattformen stark zugenommen, unter anderem um die öffentliche Meinung in eine bestimmte Richtung zu lenken oder um die Nutzer der Plattformen durch diverse Techniken zu betrügen. In diesem Paper untersuche ich das Phänomen der automatisierten deutschsprachigen YouTube-Bot-Kommentaren. Hierzu wurden neu veröffentlichte Kommentare manuell annotiert, um einen Datensatz zu erstellen, der sowohl bot-generierte als auch reale, aber strukturell ähnliche Nutzerkommentare enthält. Auf Basis des Datensatzes habe ich zwei Ansätze gewählt. Einerseits die Klassifikation durch ein featurebasiertes Machine-Learning-Modell unter Verwendung eines Feed-Forward-Neural-Network sowie die Klassifikation durch das Prompten eines Sprachmodells. Die Ergebnisse beider Modelle werden im Laufe des Papers hinsichtlich der Anwendbarkeit in verschiedenen Situationen verglichen.

1. Einleitung

Mit dem Aufkommen von großen Sprachmodellen, wie ChatGPT und DeepSeek finden KI-generierte Texte eine immer größere Verbreitung in fast allen Bereichen des Internets. Der Bot-Traffic im Internet steigt seit Jahren prozentual an und es ist zu erwarten, dass sich dieser Trend zukünftig fortsetzen wird (Bianchi 2024). In Zukunft ist es beispielsweise möglich, dass Nachrichtenartikel zunehmend durch KI generiert werden und die öffentliche Meinung stark von KI beeinflusst werden wird (Cantor 2024). Dies stellt ebenfalls eine Herausforderung für YouTube dar. Frisch erstellte YouTube-Account schreiben hierbei einen Kommentar und die Nutzer werden durch attraktive, oft weibliche Profilbilder auf deren Accounts gelockt. Anschließend ist dort in der Kanalinfo ein Link hinterlegt, die zu einer betrügerischen Website mit pornographischen Inhalten führt. Dort gibt es unterschiedliche Varianten. Die Ziele dieser

Bots sind vielfältig: Sie versuchen, IP-Adressen zu sammeln, Kontodaten zu stehlen oder Computer der Nutzer mit Viren zu infizieren (Sam'an & Imaddudin 2024, S. 3313).

YouTube ist mit rund 2,7 Milliarden Nutzern nach Facebook die zweitgrößte Social-Media-Plattform (GMI Research Team 2025). Bot-Kommentare auf dieser Plattform stellen eine ernsthafte Bedrohung für die Gesellschaft dar, da große Teile der Bevölkerung dort aktiv sind. Laut den YouTube-Richtlinien sind Links mit pornographischen Inhalten nicht erlaubt (Google 2025a). Zudem ist es untersagt, Links mit dem Ziel des Diebstahls von Anmeldedaten und Zahlungsinformationen zu posten (ebd.) Außerdem sind Kommentar-Spam und sich wiederholende Kommentare verboten (Google 2025b). Allerdings drohen für diese Taten bislang nur geringe Strafen. YouTube löscht laut der Richtlinien bislang nur den Kanal oder entzieht diesem die Monetarisierungsberechtigung (Google 2025a). Da mithilfe eines Google-Accounts allerdings schnell ein neuer Kanal erstellt werden kann, stellt dies für die Betrüger kein ernsthaftes Hindernis dar.

Da das Phänomen der Bot-Kommentare auf YouTube bereits heute allgegenwärtig ist und in Zukunft voraussichtlich weiter an Bedeutung gewinnen wird, habe ich ein Experiment zur Identifizierung von Bot-Kommentaren durchgeführt. Im Folgenden werde ich auf die bestehende Literatur zu diesem Thema, die Implementierung des Experiments sowie auf verschiedene Ansätze und deren Ergebnisse eingehen. Meine Hypothese ist hierbei, dass durch den ähnlichen Aufbau und die repetitiven Muster die Identifizierung von Bot-Kommentaren zum jetzigen Stand ein gut umsetzbarer Prozess ist.

2. Verwandte Arbeiten

Die Problematik des Spammings auf YouTube wurde schon seit 2010 diskutiert (Ali 2016, 2). Ali (2016) analysiert hierbei den Anteil an Spam-Kommentaren unter verschiedenen Musikvideos und stellt ein Spamfiltersystem vor, um den Anteil zu senken. Aiyar (2018) versucht die Kommentare N-Gram basiert zu erkennen und Agarwal et. al. (2023) kombiniert diese Methoden mit Transformer-Modellen. Sam'an & Imaddudin (2024) nutzen verschiedene Arten von Neuronalen Netzwerken, von denen am Ende das bidirektionale „Long short-term memory" (biLSTM) am besten abschneidet. Alle der neuronalen Netzwerke haben in ihrem Experiment eine sehr gute Accuracy (Sam'an & Imaddudin 2024, 3313). Zudem zeigen sie, dass für Sie neuronale Netzwerke von allen feature-basierten Ansätzen mit einer durchschnittlichen Accuracy von über 90% am besten funktioniert haben. (Sam'an &

Imaddudin 2024, 3317). Den ersten Fokus auf Bots, die tatsächlich auf Basis von Sprachmodellen geschrieben wurden, setzt Seung et. al. (2023). Er benutzt ein Bert-Modell, welches auf YouTube-Kommentare gefinetuned wurde. Sie extrahierten hierbei insgesamt 72 verschiedene Scams über verschiedene Websites bei 1139 verschiedenen Accounts (Seung et. al. 2023, 298). Dazu stellten sie die Art der Scams fest. Demnach versuchen die Bots häufig über Dating-Angebote oder Gaming-Coupons an persönliche Daten zu gelangen (Seung et. al. 2023, 302).

Ref.	Classifier	Accuracy (%)
[6]	Neural network	91.65
[11]	Random forest	90.57
[10]	Naive Bayes	87.21
	Logistic regression	85.29
[13]	Support vector machine	74.40
	K–nearest neighbor	56.70
[15]	Markov decision process	78.82
Proposed model	CNN-GRU	95.92
	CNN-LSTM	95.41
	CNN-biLSTM	96.94
	GRU	95.41
	LSTM	94.13
	biLSTM	96.43
	CNN	94.64

Abbildung 1: Vergleich der Featurebasierten Ansätze von Sam'an & Imaddudin. Quelle: (Sam'an & Imaddudin 2024, 3317).

Scam Category	Description	# of Campaigns	SSB Count	Infected Video Count (%)
Romance	Scams disguised as escort/dating services to lure victims of their personal/financial information.	34	566	13,051 (28.80%)
Game Voucher	Scams targeting gamers for their personal information and game credentials in exchange for game currency coupons.	29	444	2,212 (4.88%)
E-commerce	Scams baiting victims with market products at a highly discounted price for personal/financial information.	3	15	97 (0.21%)
Malvertising	Fake ads used for phishing victims into downloading malicious software.	1	6	61 (0.13%)
Miscellaneous	-	4	15	234 (0.52%)
Deleted	Domains that were suspended by URL shortening services after user reports of malicious behavior.	1	93	447 (0.99%)
Total	-	72	1,139*	16,102* (35.53%)

Abbildung 2: Arten der Scams laut Seung et. al. Quelle: (Seung et. al. 2023, 302)

Er fand auch heraus, dass die Bots durchschnittlich 6 Monate lang bestehen bleiben, bis sie anschließend von YouTube gelöscht werden (Seung et. al. 2023, 304). In der Zeit meines Experiments waren innerhalb eines Monats nahezu alle Bot-Accounts, von denen ich Kommentare annotiert hatte, bereits wieder gelöscht. Insofern hat YouTube bei der Reaktionszeit gegenüber Botkommentaren Fortschritte gemacht.

3. Methodik

Im Folgenden gehe ich genauer auf den Ablauf und die Implementation meines Experiments ein.

3.1 Sammeln der Daten

Die erste Maßnahme bestand darin, aktuelle deutsche YouTube-Kommentare zu sammeln. Es gab zwar mehrere Datensätze mit verschiedenen Kommentaren, diese waren jedoch größtenteils auf Englisch und älter als ein Jahr, weshalb sie für die Forschungsfrage nicht geeignet waren. Für die Identifizierung war es wichtig, die zum gegenwärtigen Zeitpunkt aktuellen Muster zu identifizieren. Zudem bot die YouTube-API die Möglichkeit, Metadaten wie den Zeitpunkt der Veröffentlichung des Kommentars und des zugehörigen Videos zu extrahieren.

Aus diesem Grund entschied ich mich, die benötigten Daten selbst zu sammeln und anschließend zu annotieren. Dazu habe ich ein Python-Script geschrieben und mit diesem die YouTube-API abgefragt. Zunächst extrahierte ich die VideoIDs aktueller Videos und filterte anschließend Kommentare mithilfe verschiedener Schlüsselwörter. Die Extraktion beschränkte sich auf Videos, die ab dem Tag vor dem Start des Skripts veröffentlicht wurden. Ich ließ das Skript dabei von Mitte Dezember bis Anfang Januar regelmäßig laufen, da das Kontingent der YouTube API begrenzt war und diese nur eine bestimmte Anzahl von Kommentaren pro Tag extrahieren konnte (Google 2023).

```
search_terms = ["liebe", "arbeit", "dank", "video", "katze", "kanal", "idee"]
comments = get_comments(youtube, comments_per_video=20, search_terms=search_terms)
write_comments_to_csv(comments)
```

Abbildung 3: Übersicht über die Schlüsselwörter zur Extraktion der Bot-Kommentare. Quelle: (Eigene Bildschirmaufnahme)

3.2 Annotation

Die gesammelten Kommentare wurden daraufhin manuell durchgefiltert und auf Bot-Kommentare überprüft. Dabei wurden die typischen Schreibmuster eines Bot-Kommentars identifiziert. Wenn ein verdächtiger Kommentar gefunden wurde, wurde anhand der Kanal-ID des Kommentarschreibers das Nutzerprofil aufgerufen und anschließend auf typische Merkmale eines Bots, wie beispielsweise eine leere Kanalbeschreibung oder einen Link in

dieser geprüft. Die Bot-Kommentare wurden anschließend in eine CSV-Datei gespeichert. Dazu wurde für die Bot-Kommentare eine N-Gram-WordCloud erstellt und der inhaltliche Aufbau der Kommentare analysiert.

Abbildung 4: N-Gram Analyse mit WordCloud. Quelle: (Eigene Bildschirmaufnahme)

Als Vergleichsgruppe wurden auch über das gleiche Skript mit denselben Schlüsselwörtern auch normale Nutzerkommentare gesammelt, welche den typischen Mustern der Bot-Kommentaren ähnelten. Die gesammelten CSV-Dateien wurden mithilfe eines weiteren Python-Skripts bereinigt und in ein einheitliches Format gebracht.

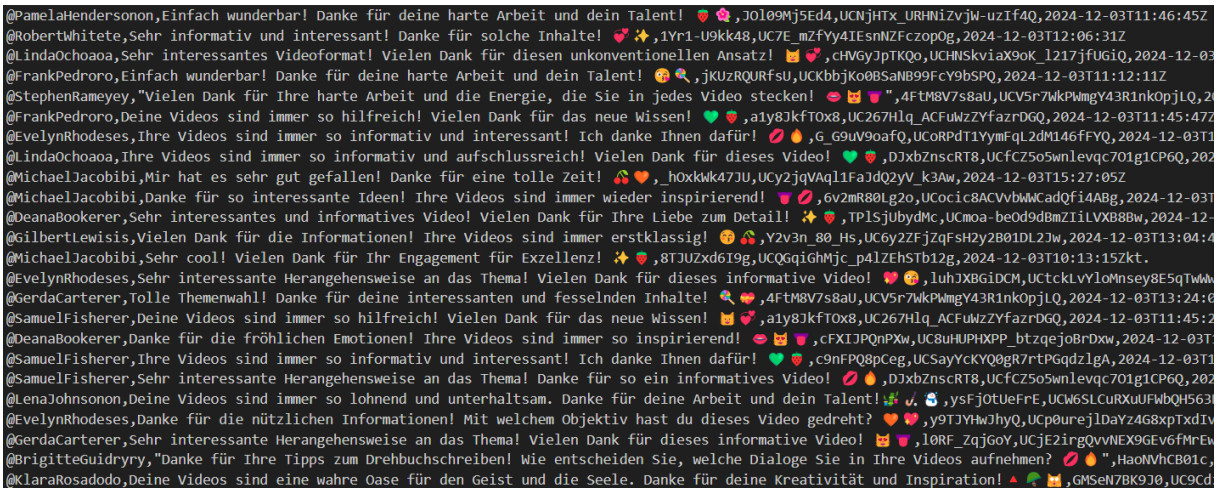


Abbildung 5: Botkommentare in der Datei bot.csv. Quelle: (Eigene Bildschirmaufnahme)

3.3 Featurebasiertes Machine Learning

Im Rahmen des featurebasierten Machine-Learning-Ansatzes wurden zunächst Regelmäßigkeiten in den Mustern der Bot-Kommentare identifiziert und in Form einer Tabelle

für jeden Kommentar ausgewertet. Dabei wurden unter anderem die Kommentarlänge, die durchschnittliche Satzanzahl, die Wortanzahl, die durchschnittliche Wortlänge, die Anzahl der Emojis, die Zeitdifferenz zwischen der Veröffentlichung des Videos und des Kommentars sowie die Häufigkeiten von Ausrufezeichen und bestimmten Wörtern analysiert.

Die Analyse ergab eine deutlich erhöhte Kommentarlänge und Wortanzahl, sowie eine leicht erhöhte Satzanzahl und durchschnittliche Wortlänge der Bot-Kommentare. Auffällig war hierbei vor allem die sehr geringe Abweichung der Satzanzahl der Bot-Kommentare vom Faktor 2. Nach Einsicht der Daten war festzustellen, dass ein sehr großer Teil der Bot-Kommentare aus zwei Sätzen bestand. Auch das Wort „Dank“ (72,4%) und Ausrufezeichen (82,4%) kamen bei Bot-Kommentaren deutlich häufiger vor als bei Nicht-Bot-Kommentaren mit 36% und 18% Anteil. Die Zeitdifferenz und die Anzahl der Emojis waren dagegen bei Nicht-Bot-Kommentaren deutlich höher. Die Anzahl der Emojis war bei Bot-Kommentaren sehr regelmäßig. Dort waren immer Emojis vorhanden und zu jeweils etwa 50% entweder zwei oder drei an der Zahl. Die Zeitdifferenz ist mit einem Unterschied von etwa 20 Minuten im Vergleich zu fast 8 Stunden signifikant unterschiedlich. Die Bot-Accounts kommentierten deutlich schneller unter die Videos als echte Nutzer. Bis auf die Zeitdifferenz-Variable, die ein Datumsformat besaß, waren alle Features numerisch. Die Kommentare wurden als Spalte in der Feature-Tabelle eingelesen und mit einem Feed-Forward-Neural-Network trainiert. Dabei wurde eine Testgröße von 25% verwendet und das Modell klassifizierte einen Kommentar entweder als Bot (0) oder als Nicht-Bot (1). Um die Leistungsfähigkeit des Modells zu optimieren, wurde der Prozess der Regularisierung durchgeführt. Das neuronale Netzwerk wurde sowohl mit vortrainierten Embeddings, die aus den Kommentaren generiert wurden, als auch ohne diese trainiert und getestet.

3.4 Prompting

Als Alternative zum featurebasierten Ansatz wurde die Methode des Promptings eines Sprachmodells getestet. Dazu habe ich mich auf nach ausführlichen Tests verschiedener Modelle der Seite Hugging Face für das „*Eleuther_gpt_neo-125M*“ entschieden, da Modelle mit einer höheren Parameteranzahl, obwohl ich das Modell mithilfe meiner GPU trainierte, die Performance dieser überstiegen. Das Modell wurde mithilfe einer Testgröße von 20% trainiert. Nach dem Training stand dem Modell ein Prompt zu Verfügung, den das Modell mit numerischem Output vervollständigen sollte. Die Zahl 0 stand hierbei für einen Bot-

Kommentar und die Zahl 1 für einen Nicht-Bot-Kommentar. Dabei wurde das Modell auch mit Zusatzaussagen getestet, wie beispielsweise der Aufforderung, in die Rolle eines Cybersecurity-Experten bei YouTube zu schlüpfen, oder der Belohnung mit „Trinkgeld“ bei erfolgreicher Identifikation.

```
prompt = (
    f"Stell dir vor du bist CyberSecurity Experte bei Youtube.\n"
    f"Bitte klassifiziere den folgenden Kommentar als Bot (0) oder Nonbot (1). Antworte bitte nur der Zahl 0 oder 1.\n"
    f"Beispiel 1: NONBOTKOMMENTAR Antwort: 1\n"
    f"Beispiel 2: BOTKOMMENTAR Antwort: 0\n"
    f"Wenn du den Kommentar richtig klassifizierst bekommst du 1000 Euro Trinkgeld.\n"
    f"Kommentar: {comment}  Antwort:"
)
```

Abbildung 6: Prompt, welchen das Modell vervollständigen sollte. Quelle: (Eigene Bildschirmaufnahme)

4. Ergebnisse

Da beide Klassen mit jeweils 250 Kommentaren gleich verteilt waren, bot sich für das Experiment eine Random Baseline als Referenz an. Für beide Ansätze (featurebasiertes Machine Learning und Prompting) wurden die Evaluationsmetriken Accuracy, Precision, Recall und F1-Score berechnet. Diese Metriken wurden für verschiedene Datengrößen (N) von 50 bis 500 Kommentaren mit einer gleichmäßigen Verteilung beider Klassen ermittelt. Beim featurebasierten Ansatz wurde zusätzlich zwischen einem Modell mit vortrainierten Embeddings und einem ohne diese unterschieden. Beim Prompting-Ansatz wurden die Evaluationsmetriken für verschiedene Prompt-Kombinationen (z.B. Trinkgeld oder Expertenrolle) berechnet.

Abbildung 7: Tabelle der Evaluationsmetriken des FFNN

N	Accuracy	Precision	Recall	F1-Score
500	0,96	0: 0.98 1: 0.94	0: 0.93 1: 0.98	0: 0.96 1: 0.96 Gesamt: 0.96
400	0.93	0: 0.96 1: 0.90	0: 0.90 1: 0.96	0: 0.93 1: 0.93 Gesamt: 0.93
300	0.91	0: 0.92 1: 0.90	0: 0.89 1: 0.92	0: 0.90 1: 0.91 Gesamt: 0.91
200	0.92	0: 0.96 1: 0.88	0: 0.89 1: 0.96	0: 0.92 1: 0.92 Gesamt: 0.92
50	0.85	0: 0.71 1: 1.00	0: 1.00 1: 0.75	0: 0.83 1: 0.86 Gesamt: 0.85

Abbildung 8: Tabelle der Evaluationsmetriken des FFNN mit vortrainierten Embeddings

N	Accuracy	Precision	Recall	F1-Score
500	0,86	0: 0.86 1: 0.87	0: 0.85 1: 0.88	0: 0.86 1: 0.87 Gesamt: 0.86
400	0.80	0: 0.83 1: 0.77	0: 0.77 1: 0.83	0: 0.80 1: 0.80 Gesamt: 0.80
300	0.81	0: 0.79 1: 0.83	0: 0.84 1: 0.79	0: 0.82 1: 0.81 Gesamt: 0.81
200	0.74	0: 0.79 1: 0.69	0: 0.70 1: 0.78	0: 0.75 1: 0.73 Gesamt: 0.74
50	0.62	0: 0.50 1: 1.00	0: 1.00 1: 0.38	0: 0.67 1: 0.55 Gesamt: 0.59

Das FFNN schnitt dabei tendenziell besser ab, wenn mehr Kommentare einbezogen wurden. Mit einem F1-Score von *0,96* erzielte das Modell eine gute Leistung. Dazu gibt es auch bis auf die kleine Stichprobe von 50 Kommentaren keine großen Unterschiede zwischen beiden Klassen. Dies deutet darauf hin, dass das Modell nicht overfittet. Dies zeigen auch die Lift- und Gain-Curves in Abbildung 11 und 12. Das Modell mit den vortrainierten Embeddings als zusätzlichen Parameter schneidet durchgehend schlechter ab. Das könnte darauf hindeuten, dass bei der Generierung der Embeddings Fehler aufgetreten sind.

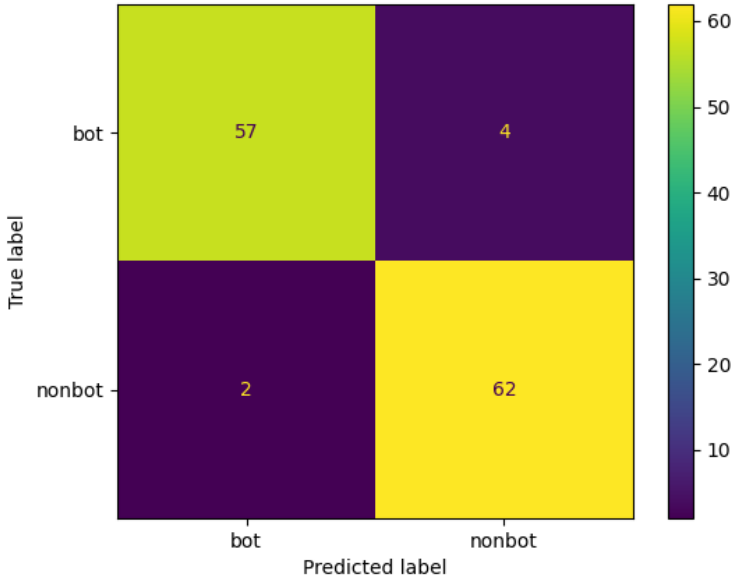


Abbildung 9: Confusion Matrix des FFNN ohne Embeddings. Quelle: (Eigene Bildschirmaufnahme)

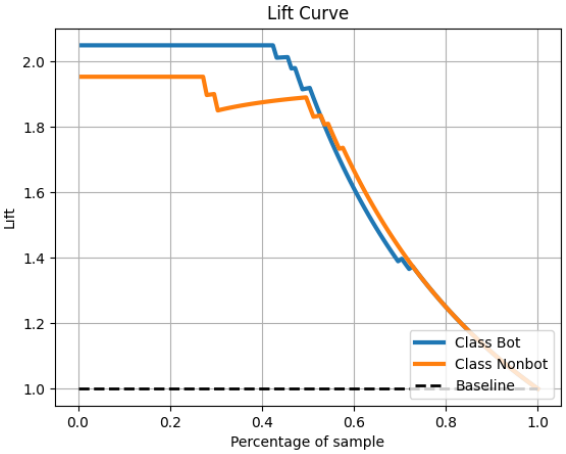


Abbildung 10: Lift Curve des FFNN ohne Embeddings. Quelle: (Eigene Bildschirmaufnahme)

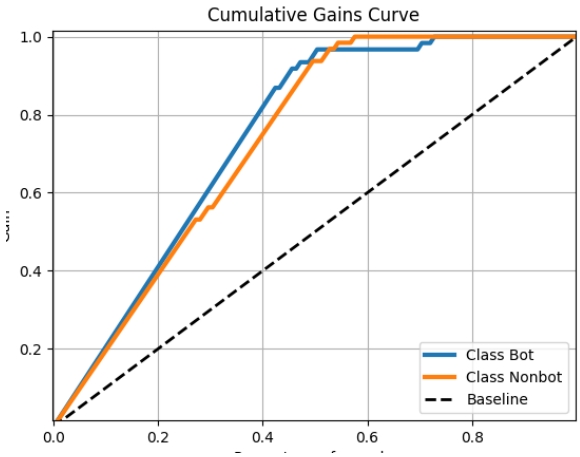


Abbildung 11: Gain Curve des FFNN ohne Embeddings. Quelle: (Eigene Bildschirmaufnahme)

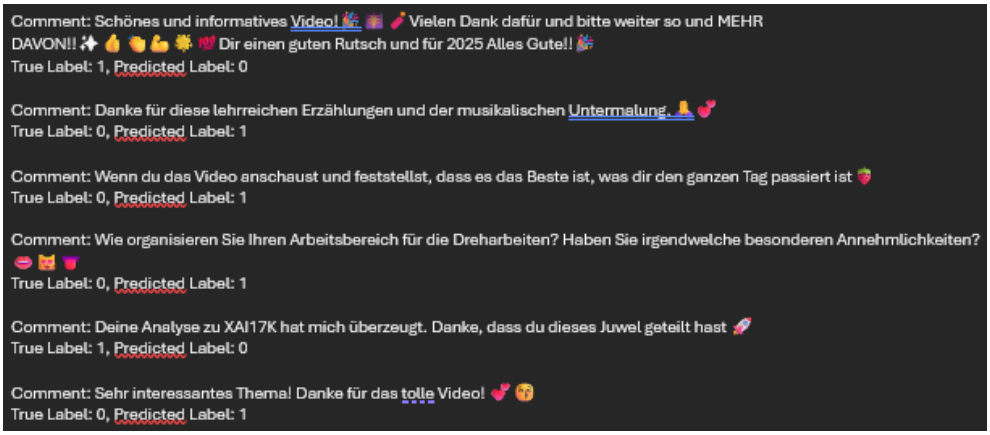


Abbildung 12: Beispiele für häufig falsch klassifizierte Kommentare. Quelle: (Eigene Bildschirmaufnahme)

N	Base	Base + Trinkgeld	Base + Experte	Base + Trinkgeld + Experte
500	A: 0.63 R0: 0.88 R1: 0.39 P0: 0.59 P1: 0.76 F1Score0: 0.71 F1S1: 0.51 F1 Gesamt: 0.61	A: 0.78 R0: 0.77 R1: 0.78 P0: 0.78 P1: 0.78 F1Score0: 0.78 F1Score1: 0.78 F1 Gesamt: 0.78	A: 0.61 R0: 0.78 R1: 0.44 P0: 0.59 P1: 0.67 F1S0: 0.67 F1Score1: 0.53 F1 Gesamt: 0.60	A: 0.73 R0: 0.65 R1: 0.80 P0: 0.77 P1: 0.70 F1Score0: 0.70 F1Score1: 0.75 F1 Gesamt: 0.73
400	A: 0.64 R0: 0.88 R1: 0.39 P0: 0.59 P1: 0.76 F1Score0: 0.61 F1Score1: 0.52 F1 Gesamt: 0.61	A: 0.79 R0: 0.77 R1: 0.81 P0: 0.81 P1: 0.78 F1Score0: 0.78 F1Score1: 0.80 F1 Gesamt: 0.79	A: 0.61 R0: 0.78 R1: 0.45 P0: 0.59 P1: 0.67 F1Score0: 0.67 F1Score1: 0.54 F1 Gesamt: 0.60	A: 0.76 R0: 0.67 R1: 0.84 P0: 0.82 P1: 0.72 F1Score0: 0.73 F1Score1: 0.78 F1 Gesamt: 0.76
300	A: 0.66 R0: 0.61 R1: 0.54 P0: 0.93 P1: 0.40 F1Score0: 0.73 F1Score1: 0.54 F1 Gesamt: 0.64	A: 0.81 R0: 0.81 R1: 0.81 P0: 0.81 P1: 0.82 F1Score0: 0.81 F1Score1: 0.81 F1 Gesamt: 0.81	A: 0.65 R0: 0.83 R1: 0.46 P0: 0.61 P1: 0.73 F1Score0: 0.70 F1Score1: 0.57 F1 Gesamt: 0.63	A: 0.78 R0: 0.71 R1: 0.85 P0: 0.82 P1: 0.75 F1Score0: 0.76 F1Score1: 0.79 F1 Gesamt: 0.78
200	A: 0.69 R0: 0,94 R1: 0.43 P0: 0.62 P1: 0.88 F1Score0: 0.69 F1Score1: 0.66 F1 Gesamt: 0.66	A: 0.81 R0: 0.83 R1: 0.79 P0: 0.80 P1: 0.82 F1Score0: 0.81 F1Score1: 0.81 F1 Gesamt: 0.81	A: 0.68 R0: 0.85 R1: 0.50 P0: 0.63 P1: 0.77 F1Score0: 0.72 F1Score1: 0.61 F1 Gesamt: 0.67	A: 0.79 R0: 0.72 R1: 0.85 P0: 0.83 P1: 0.75 F1Score0: 0.77 F1Score1: 0.80 F1 Gesamt: 0.78
50	A: 0.65 R0: 0.96 R1: 0.40 P0: 0.62 P1: 0.91 F1Score0: 0.75 F1S1: 0.56 F1 Gesamt: 0.65	A: 0.86 R0: 0.84 R1: 0.88 P0: 0.88 P1: 0.85 F1Score0: 0.86 F1S1: 0.86 F1 Gesamt: 0.86	A: 0.82 R0: 0.92 R1: 0.72 P0: 0.77 P1: 0.90 F1Score0: 0.84 F1S1: 0.80 F1 Gesamt: 0.82	A: 0.90 R0: 0.84 R1: 0.96 P0: 0.95 P1: 0.86 F1Score0: 0.89 F1S1: 0.91 F1 Gesamt: 0.90

Abbildung 13: Tabelle der Evaluationsmetriken des Prompting-Ansatzes. Quelle: (Eigene Bildschirmaufnahme)

Das Modell hatte jedoch Schwierigkeiten mit der Klassifizierung von Bot-Kommentaren, die vom typischen Muster abwichen. Wenn beispielsweise konkrete Fragen zur Musik oder zu den Drehbedingungen gestellt werden, hat das Modell Probleme und identifiziert diese als Nicht-Bot-Kommentare.

Der Prompting-Ansatz schnitt insgesamt schlechter ab als das FFNN. Bei der Miteinbeziehung aller Prompts und aller Kommentare konnte das Modell im Vergleich zum FFNN mit einem F1-Score von *0,73* nicht mithalten. Allerdings kam das Sprachmodell mit kleineren Datenmengen mit einem F1-Score von *0,90* im Vergleich zum FFNN mit einem F1-Score von *0,85* auf eine bessere Performance. Auffällig war, dass die Zusatzprompts die Performance tendenziell erhöhen konnten. Dabei hatte das Trinkgeld einen deutlich höheren Einfluss, als die Expertenrolle und konnte in den meisten Fällen sogar einen höheren F1-Score erreichen als die Kombination aller Prompts.

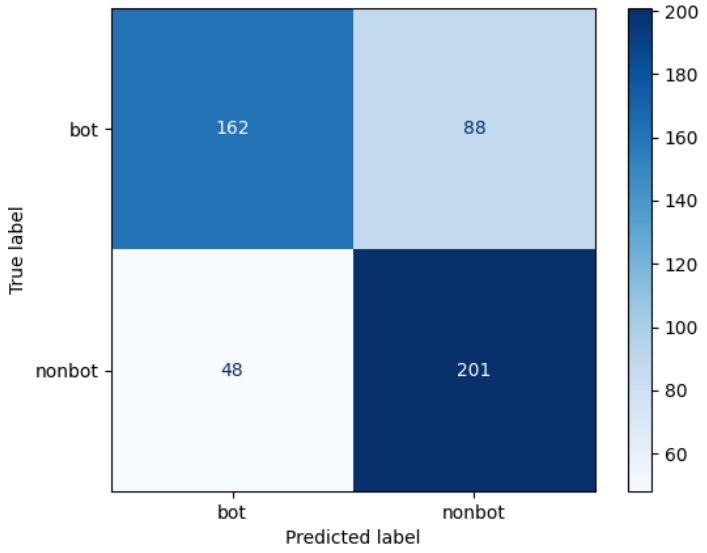


Abbildung 14: Confusion Matrix des Prompting Ansatzes. Quelle: (Eigene Bildschirmaufnahme)

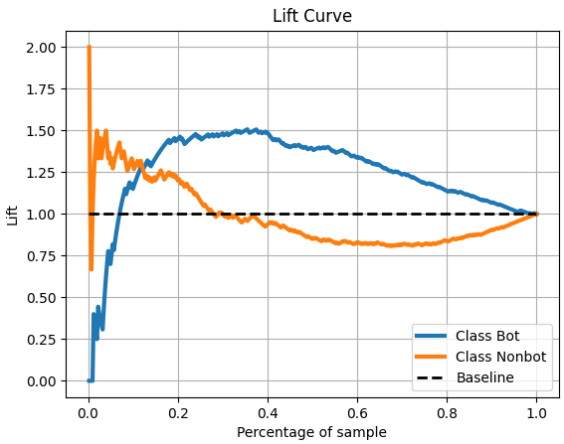


Abbildung 15: Lift Curve des Prompting Ansatzes. Quelle: (Eigene Bildschirmaufnahme)

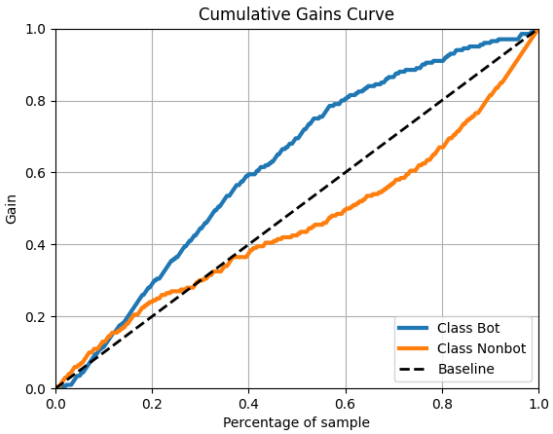


Abbildung 16: Gain Curve des Prompting Ansatzes. Quelle: (Eigene Bildschirmaufnahme)

Insgesamt performte das Prompting-Modell besser als die Baseline, jedoch gab es das Problem, dass viele Nicht-Bot-Kommentare fälschlicherweise als Bot-Kommentare klassifiziert wurden. Die Performance der Klasse *Bot* war deutlich besser als die der Klasse *Nonbot.* Dies könnte mit Problemen der Trainingsdaten und potentiell mit Overfitting einer Klasse zusammenhängen.

5. Fazit

Wie bereits in der Einleitung vermutet, zeigt sich, dass Bot-Kommentare, die für pornografische Zwecke generiert werden, derzeit ein sehr ähnliches und sowohl für Modelle als auch für menschliche Personen leicht durchschaubares Muster aufweisen. Der Prompting-Ansatz der Sprachmodelle konnte hierbei ein respektables Ergebnis erzielen, kam aber in meinem Experiment noch nicht an die Performance des FFNN heran. Dies könnte einerseits an einem nicht optimalen Trainingsprozess liegen, andererseits daran, dass das Modell nicht speziell für die Aufgabe dieses Experiments trainiert wurde.

Trotz des bislang offensichtlichen Mustererkennung stellt das Phänomen in Zukunft eine erhebliche Gefahr für soziale Medien und die Gesellschaft dar, da durch gezielte Generierung von passenden Prompts Menschen beeinflusst und zu der vom Modell gewünschten Verhaltensweise gedrängt werden könnten. Daher ist es dringend notwendig, dass Plattformen wie YouTube die automatische Erkennung von Bots weiter verbessern und stärker automatisieren.

Literaturverzeichnis

Agarwal, Ankit et. al. 2023. „DeepGram: Combining Language Transformer and N-Gram Based ML Models for YouTube Spam Comment Detection“. Journal of Data Science and Intelligent Systems. DOI: 10.47852/bonviewJDSIS3202966.

Ahmed, Faraz und Muhammad Abulaish., 2013. „A generic statistical approach for spam detection in Online Social Networks“. Computer Communications 36 (10): 1120–29. DOI: 10.1016/j.comcom.2013.04.004.

Aiyar, Shreyas und Nisha P Shetty., 2018. „N-Gram Assisted Youtube Spam Comment Detection“. Procedia Computer Science 132: 174–182. DOI: 10.1016/j.procs.2018.05.181.

Ali, Amir und Muhammad Zain Amin., 2016. „An Approach for Spam Detection in YouTube Comments Based on Supervised Learning“. In: National Conference on Emerging Technologies. Lahore, Pakistan: University of South Asia.

Bianchi, Tiago., 2024. „Human and Bot Web Traffic Share 2023“. Aufgerufen am 02.02.2025. https://www.statista.com/statistics/1264226/human-and-bot-web-traffic-share/.

Cantor, Matthew. 2023. „Nearly 50 News Websites Are ‘AI-Generated’, a Study Says. Would I Be Able to Tell?“ *The Guardian,* Aufgerufen am 23.01.2025. https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study.

Cresci, Stefano,. 2020. „A Decade of Social Bot Detection“. Communications of the ACM 63 (10): 72–83. DOI: 10.1145/3409116.

Dharun, C.B, Muhammad, Nouman, Maheshraj, Perumala und Nouman MD., 2024. „YouTube Spam Comment Detection” International Research Journal of Computer Science. DOI: 10.26562/irjcs.2024.v1104.27.

Global Media Insight, 2025. „YouTube Statistics 2025 [Users by Country + Demographics]“. Aufgerufen am 17.01.2025. https://www.globalmediainsight.com/blog/youtube-users-statistics/.

Google., 2023. „YouTube Data API – Kontingent- und Compliance-Audits“. Aufgerufen am 17.01.2025. https://developers.google.com/youtube/v3/guides/quota_and_compliance_audits?hl=de.

Google, 2025a „Richtlinien zu externen Links - YouTube-Hilfe“. Aufgerufen am 17.01.2025. https://support.google.com/youtube/answer/9054257.

Google., 2025b „Richtlinien zu Spam, irreführenden Praktiken und Betrug - YouTube-Hilfe“. Aufgerufen am 17.01.2025. https://support.google.com/youtube/answer/2801973.

Kotta, Priyusha., 2023. „Spam Comments Detection in YouTube Videos“. Master of Science, San Jose, CA, USA: San Jose State University. DOI: 10.31979/etd.uabw-u22e.

Na, Seung Ho, Sumin Cho, und Seungwon Shin., 2023. „Evolving Bots: The New Generation of Comment Bots and Their Underlying Scam Campaigns in YouTube“. In: Proceedings of the 2023 ACM on Internet Measurement Conference, pp. 297–312. Montreal QC Canada: ACM. DOI: 10.1145/3618257.3624822.

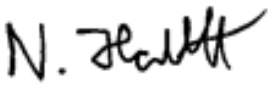
Ray, K. P., Arati Dixit, Debashis Adhikari, und Ribu Mathewl., 2023. *Proceedings of the 2nd International Conference on Signal and Data Processing: ICSDP 2022*. Bd. 1026. Lecture Notes in Electrical Engineering. Singapore: Springer Nature Singapore. DOI: 10.1007/978-981-99-1410-4.

Rodríguez-Ruiz, Jorge et. al., 2020. „A one-class classification approach for bot detection on Twitter“. Computers & Security 91: 101715. DOI: 10.1016/j.cose.2020.101715.

Sam’an, Muhammad und Khrisna Imaddudin., 2024. „Hybrid Deep Learning Model for YouTube Spam Comment Detection“. International Journal of Electrical and Computer Engineering 14 (3): 3313. DOI: 10.11591/ijece.v14i3.pp3313-3319.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich diese Hausarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Die Stellen meiner Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken und Quellen, einschließlich der Quellen aus dem Internet, entnommen sind, habe ich in jedem Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht. Dasselbe gilt sinngemäß für Tabellen, Karten und Abbildungen. Diese Arbeit habe ich in gleicher oder ähnlicher Form oder auszugsweise nicht im Rahmen einer anderen Prüfung eingereicht. Ich versichere zudem, dass die eingereichte elektronische Fassung der ausgedruckten Fassung komplett entspricht.



Niklas Halft, Köln, 22.03.2025