Erkennung von islamistisch-extremistischen Radikalisierungszeichen in Sozialen Medien

Dr. Cruz-Aceves

Working Paper

Stand 2023

(Kanalnamen absichtlich – aus Datenschutzgründen – geschwärzt)

Inhalt

Empirischer Teil

Der empirische Teil ist in zwei Subkategorien geteilt. Einerseits werden die Ergebnissen von der KI-basierte Klassifikation der gesammelten YouTube-Kommentaren präsentiert. Um einen

tiefgründigeren Überblick darüber zu haben, welche Themen in den Kommentaren angesprochen werden, und inwiefern die untersuchten Kanälen und die identifizierten Themen linguistische Merkmalen von islamistisch-radikalen Texten spiegeln, werden im Anschluss Ergebnissen von Themenmodellen SIRA und LIWC präsentiert.

Daten

Die Ergebnissen von SIRA, LIWC und Themenmodellierung beziehen sich auf den bis 3. Juni 2022 gecrawlten Daten, d.h. 258.502 Kommentare auf insgesamt 5957 Videos und 8 YouTube Kanäle gepostet. Das ergibt insgesamt 6.837.585 Wörter. Kommentare enthalten in Durchschnitt 26 Wörter.

Gewalt-Klassifikator

Unter Verwendung von Posts aus Telegramm-Kanälen, Facebook-Seiten und -Profilen aus dem islamistisch-extremistischem Milieu, die als Gewaltaffirmation enthaltend (N=21653) oder keine Gewaltaffirmation enthaltend (N=522) von Pelzer und Uhlenbrock (2021) annotiert wurden[1], entwickelte ich Klassifikatoren mittels verschiedenen Textklassifikationsalgorithmen[2] mit dem Ziel, gewaltaffirmative Äußerungen in den gesammelten YouTube Kommentaren zu erkennen. Den Klassifikator, den ich basiert auf das Bernoulli Naive Bayes (BNB) Algorithmus mit den oben genannten Posts trainiert habe, hat die besten Klassifikationsgüte ergeben[3]. Deswegen habe ich nach den fürs BNB Algorithmus verfügbaren Hyperparameter, die die höchsten Klassifikationsgüte liefern könnten, gesucht. In Tabelle 1 stellen ich dar, die Klassifikationsgüte von BNB-Klassifikatoren, deren Training für verschiedenen Klassifikationsmetriken optimiert wurde.

	Prec-Opt	**F1-Opt**	**Rec-Opt**
precision	0,79	0,53	0,42
recall	0,08	0,79	0,89
f1-score	0,15	0,63	0,57

Tabelle 1. Klassifikationsgüte für Klasse 'beleidigend' von optimisierten (Bernoulli-NB) Klassifikatoren

In einem nächsten Schritt habe ich den trainierten Klassifikator mit den besten Metriken, und zwar, derjenige, den für die Klassifikationsmetrik Trefferquote (aus dem englischen "recall")

[1] Ich danke Dres. Pelzer und Uhlenbrock dafür, dass Sie mit mir ihren annotierten Daten geteilt haben. Weitere Informationen über die annotierten Daten in Pelzer und Uhlenbrock (2021).

[2] Im konkret, Klassifikatoren mit den folgenden Algorithmen wurden trainiert und deren Güte verglichen: Logistic Regression, Decision Trees, Multinomial Naive Bayes, Stochastic Gradient Descent und Random Forest.

[3] In Anlehnung an Pelzer und Uhlenbrock (2021), a) nur die 500 Features mit dem höchsten Chi-2-Wert wurden verwendet, b) die Features wurden gemäß tf – idf gewichtet und c) das Training erfolgte auf allen verbleibenden Token-Unigrams. Da ich eine höhere Klassifikationsgüte mit dem Bernoulli Naive Bayes (BNB) Algorithmus – i.V.z. Logistic Regression, wie Pelzer und Uhlenbrock (2021)—und lemmatisierten Tokens gefunden habe, präsentiere ich Ergebnisse von einem Klassifikator trainiert mit dem BNB Algorithmus.

optimiert wurde (Spalte „Rec-Opt“ von Tabelle 1; entsprechende ROC und Precision-Recall Kurven, s. Appendix 5), YouTube Kommentare aus dem aufgebauten Korpus klassifizieren gelassen. Zum Schluss wurde qualitativ überprüft, inwiefern die als gewalttätig klassifizierten YouTube Kommentaren tatsächlich gewaltaffirmative Äußerungen ausgedruckt haben. Die qualitative Überprüfung der Ergebnisse ergab, dass 30% der gesammelten YouTube Kommentaren korrekt klassifiziert waren, 38% falsch und 33% nachvollziehbar – letzteres bezieht sich z.B. auf Posts, die gewalthaltige Koranverse auflisten (welche dann zu einer Klassifikation als gewaltaffin führen) und bei denen nur aus kontextuellen Details erkennbar ist, dass sie mit islamkritischer Motivation geschrieben wurden und die Autoren also den zitierten Inhalten mutmaßlich ablehnend gegenüberstehen. Auch wenn die Klassifikation somit letztendlich wahrscheinlich nicht korrekt ist, kann diese Schlussfolgerung von einer KI (oder auch unerfahrenen menschlichen Lesern) nicht realistisch erwartet werden.

„Sentiment-based Identification of Radical Authors“ (SIRA)

SIRA wurde entwickelt, um innerhalb eines Korpus mit dschihadistischen Forum-Posts die radikalsten Nutzer zu identifizieren. Dazu wird anhand der Post-Historie eines Nutzers mit Hilfe der Sentimentanalyse-Software SentiStrength berechnet, mit welcher Dauer, Frequenz und Intensität sich der Nutzer negativ zu den hundert korpusweit meistbenutzten Substantiven äußert. Der SIRA-Algorithmus generiert daraus pro Nutzer einer Radikalitätsbewertung von 0 bis 40 Punkten (Scrivens et al. 2017). Denn der Autor vom SentiStrength in privater Kommunikation (4. November 2021) empfohlen hat, keine Bereinigungsstrategie geschweige davon Lemmatisierung vor der Anwendung von SentiStrength durchzuführen, die erste Berechnung von SIRA wird also auf YouTube Kommentare zugreifen, von der lediglich A) HTML-Code-, B) Emojis entfernt wurden und C) für die verschiedene Schreibweisen von den Wörtern Muhammad, Hadith, nasheed, Sunna, und Ramadan homogenisiert wurden.

SIRA liegt ein Verständnis von Radikalisierung zugrunde, das sich auf zunehmenden Antagonismus (und somit zunehmende Negativität) gegenüber den politischen Feinden fokussiert. In einem Korpus, in dem allgemein eine hohe Radikalität angenommen werden kann, ist das ein durchaus sinnvoller Ansatz. In einem Korpus wie dem vorliegenden, in dem Radikalität die Ausnahme ist, werden die Ergebnisse aber durch häufige, negativ besetzte Begriffe wie „Hölle“, „Sünde“ und „das Böse“ dominiert. Je nach Verwendungskontext *können* diese Stellen tatsächlich radikal bzw. extremistisch sein, sehr oft sind sie es aber auch nicht.

Ich dieses Problem teilweise mitigieren, indem ich als Liste der Bezugsbegriffe nicht, wie die Entwickler von SIRA, die hundert meistbenutzen Substantive, sondern eine kuratierte Version verwendet habe, die seltenere, aber tendenziell konfliktbehaftete Begriffe enthält. Perfekt sind die Ergebnisse aber nicht: Von 11 durch SIRA als relevant identifizierten Nutzern waren vier hochrelevant, drei teilweise relevant und drei irrelevant, sprich falsch-positiv. Vor allem aber

muss stets bedacht werden, dass Äußerungen, die inhaltlich radikal sind, aber sentimental positiv geframet werden, von SIRA nicht erkannt werden.

Themenmodellierung

Um einen systematischen Überblick darüber haben zu können, worüber in den beobachteten YouTube Kanälen diskutiert wird, wurden s.g. *strukturierte Themenmodelle* (aus dem englischen *Structural Topic Models*) berechnet (Roberts et al. 2019). Diese Modelle ermöglichen A) Themen von Texten statistisch abzuleiten bzw. Texten in sinnvollen semantischen Kategorien einzuordnen und B) festzustellen, inwiefern die Metadaten die Häufigkeit, in der die resultierenden Themen im Korpus auftauchen (bzw. die Themenprävalenz), beeinflussen.

Nach Vergleichen von Modellen unterschiedlicher qualitativen und quantitativen Güte ich mich für ein 30-Themen Modell[4] entschieden, wobei Themaprävalenz als eine Funktion von den Metadaten *YouTube-Kanal* und *Kommentar-Datum* untersucht wurde. Wie üblich in Forschungsprojekten, die die gleiche analytische Strategie verwendet haben (Nielsen 2017; McDonald et al. 2020; Schwemmer und Jungkunz 2019; Karell und Freedman 2019), nur eine Auswahl von den resultierenden Themen hat sich als relevant für das vorliegende Projekt erwiesen. Die folgende Analyse fokussiert sich also auf die ausgewählten Themen (s. Appendix 1, Appendix 2 und Appendix 3 für Themen-Proportionen, -prävalenz über die Zeit und über die Kanäle für alle 30 Themen).

[4] Davor wurden die Kommentaren als folgend bereinigt: Emojis, HTML-Charaktere, Standard- und spezifisch fürs Projekt erstellte Stoppwörter wurden entfernt, unterschiedliche Schreibweisen von gewissen Wörtern (z.B. Muhammad, Hadith, nasheed) wurden homogenisiert und die verbleibenden Wörtern wurden lemmatisiert. In einem letzten Schritt, Wörter, die in weniger als 0,5 % der gesamten Kommentaren aufgetaucht sind, wurden entfernt, um Berechnungseffizienz zu erzielen, wie in der Literatur empfohlen (Puschmann et al. 2020; Rauchfleisch und Kaiser 2020; Schwemmer und Ziewiecki 2018).

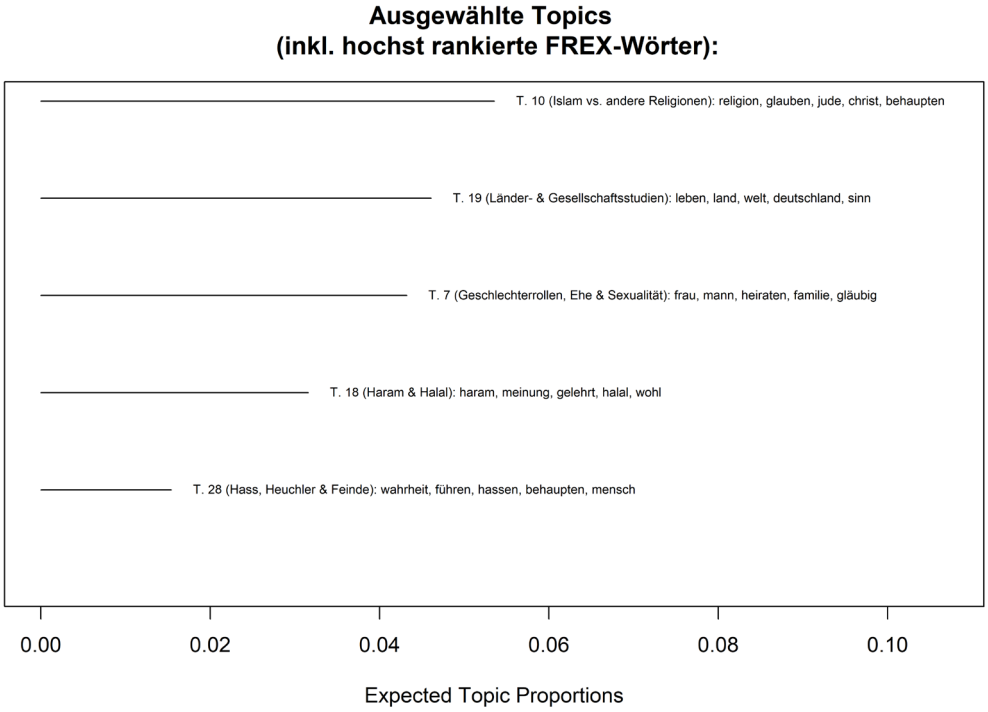


Abbildung 1. Proportionen für ausgewählte Themen inklusive Top FREX-Wörter

Die ausgewählten Themen wurden anhand A) der resultierenden FREX-Wörter[5] und B) der repräsentativsten YouTube-Kommentare (pro Thema) als in Abbildung 1, die die resultierenden Themenproportionen bzw. den geschätzten Anteil der YouTube-Kommentaren, die zu jedem Thema gehört, darstellt, zu sehen. Thema Nummer 10 macht die zweitgrößte Kommentar-Gruppierung (c.f. Appendix 1. Themen-Proportionen inklusive Top Wörter für alle 30 Themen) aus und zwar, ungefähr 5% der analysierten Kommentaren diskutieren rund um „Islam vs. andere Religionen und Glaubenssätze"; das drittgrößte Diskussionsthema (ungefähr 4 % der analysierten Kommentaren) dreht sich um Länder- und Gesellschaftsstudien (Thema Nummer 19); das viertgrößte Thema (Thema 7; ungefähr 4 % der ausgewerteten Kommentaren) diskutiert über „Geschlechterrollen und -verhältnisse, Ehe und Sexualität". Die übrigen zwei Themen (18 und 28), die sich als relevant für das vorliegende Projekt erwiesen haben, gehören nicht zu den meistdiskutierten Fragen, und zwar ungefähr 3 % und 1,5 % der untersuchten Nutzerkommentaren gingen um jeweils „Haram und Halal" und „Hass, Heuchler und Feinde".

Die Themaprävalenz über die Zeit kann man in Abbildung 2 sehen. Außer Thema 19, die Prävalenz von allen für das Projekt relevanten Themen hat in letzter Zeit zugenommen. Interessanterweise die Häufigkeit, in der die untersuchten Kommentaren über alle fünf relevanten Themen diskutiert haben, ist von einer steigenden Phase am Anfang der Beobachtungszeitraum geprägt, vor allem Themen 7 und 18, die steigenden Prävalenzen ununterbrochen seit Januar 2014

[5] FREX steht für häufig-exklusive Wörter (aus dem englischen *frequency-exclusive*). Diese Wörter sind von einem Formel, dass ordnet Wörter ein je nach A) ihrer allgemeinen Häufigkeit im Korpus und B) inwiefern die Wörter exklusiv zu jedem Thema sind, bestimmt (McDonald et al. 2020, S. 972).

bis jeweils Frühjahr 2019 und Sommer 2017 zeigen[6]. Obwohl alle fünf Themen auch kurze sinkende Prävalenzperioden gezeigt haben, die Phasen, in der die jeweilige Themaprävalenz gestiegen ist, sind im Durchschnitt mehr und länger gewesen.

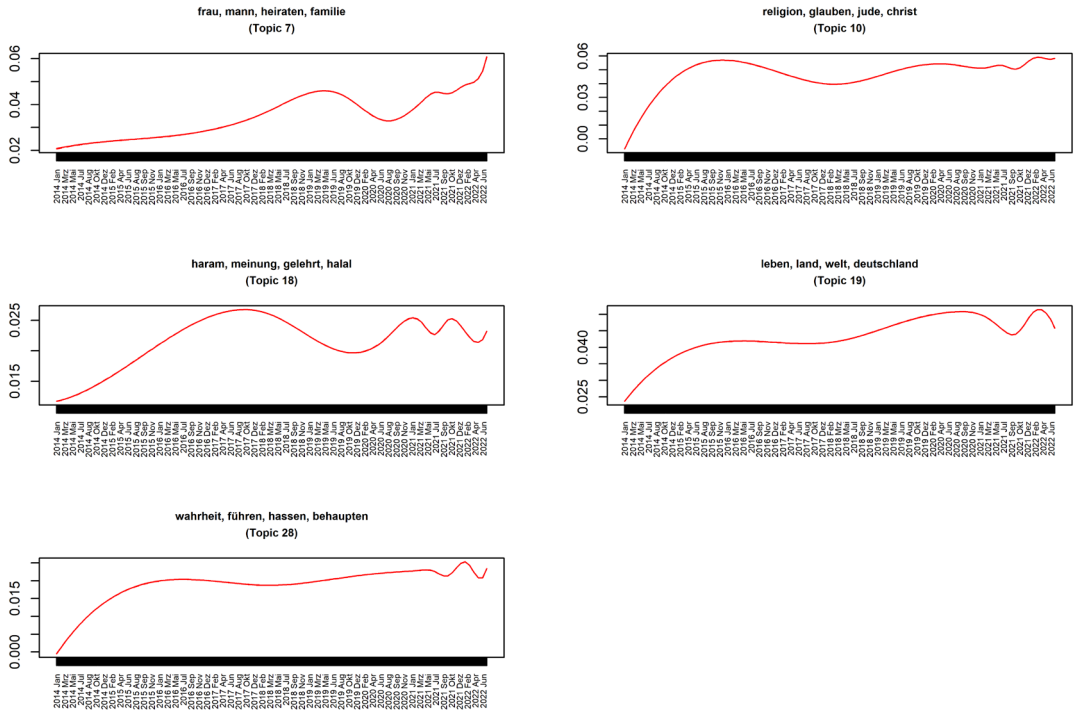


Abbildung 2. Effekt von Kommentar-Veröffentlichungsdatum (glatte „Basis-Spline" Funktion, x-Achse) auf die erwartete Themaproportion (Y-Achse) bzw. Themenprävalenz über die Zeit für ausgewählte Themen.

Die Themaprävalenz über die Kanäle kann man in Abbildung 3 sehen. Die Themen "Geschlechterrollen und -verhältnisse, Ehe" (7) und "Sexualität Haram und Halal" (18) wurden am häufigsten im Kanal [redacted] diskutiert, gefolgt von [redacted] und [redacted]. Die restlichen Kanäle haben diese Themen seltener angesprochen. Die häufigsten Diskussionen über einerseits Hass, Heuchler und Feinde, andererseits Länder- und Gesellschaftsstudien (Themen 28 und 19) wurden im Kanal "[redacted]" beobachtet. Das Thema mit dem sich die meisten Kanälen fast gleichmäßig beschäftigt haben war eine Diskussion über "Islam vs. andere Religionen und Glaubenssätze". Der Kanal "[redacted]" hat am wenigsten dieses Thema angesprochen. Aufgrund der (breiten) Konfidenzintervall der Einschätzung für den Kanal "[redacted]", das bei allen Themen (außer Thema 18) das null enthält, lässt sich nicht ausschließen,

[6] S. Appendix 7. Prävalenz der Themen über die Zeit für ausgewählte Themen (ohne Konfidenzintervall) für eine alternative Darstellung

ob der Kanal tatsächlich weniger über die fünf fürs Projekt als relevant eingestuften Themen diskutiert als die anderen Kanäle.

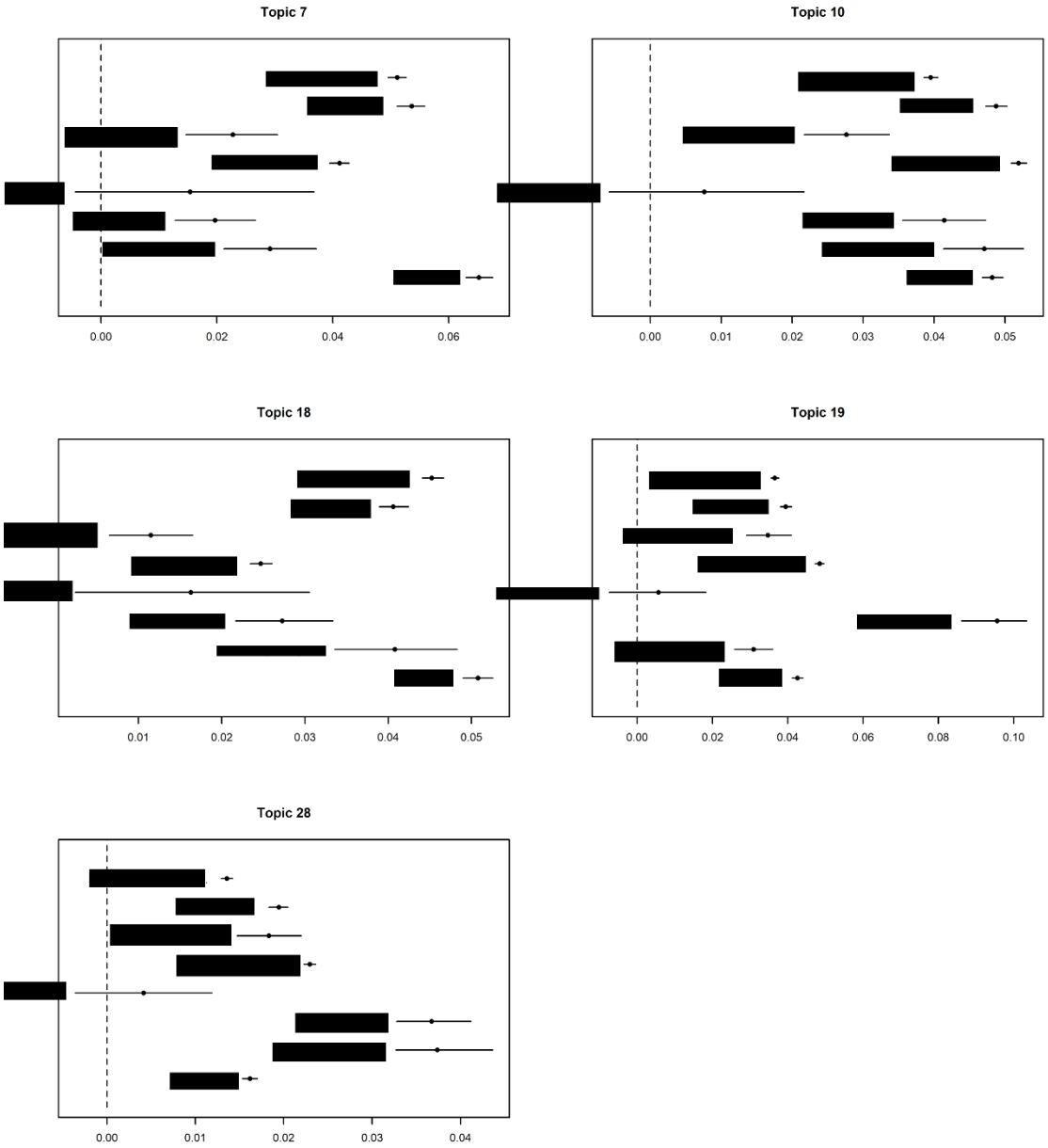


Abbildung 3. Marginal Themaproportion von jedem Kanal für ausgewählte Themen (inkl. 95% Konfidenzintervall). Kanalnamen absichtlich – aus Datenschutzgründen – geschwärzt.

„Linguistic Inquiry and Word Count" (LIWC)

Da bisher es kein Wörterbuch gibt, das islamistische Radikalisierung in schriftlicher Kommunikation gezielt messen kann, habe ich in der Literatur nach möglichen Indikatoren, die bei der Exploration vom aufgebauten Korpus hilfreich sein könnten, gesucht. Vorherige Forschungsprojekten haben diesbezüglich mittels des LIWCs—die dafür entwickelt wurde, um

linguistische Ausdrücken psychologischer Prozesse analysieren zu können. D. h. die Indikatoren davon sollen unter anderem emotionale und/oder kognitive Prozesse schriftlicher Kommunikation spiegeln können—linguistische Merkmalen, die charakteristisch von islamistisch-extremistischen Texten waren, identifiziert (Shrestha 2019; Krippendorff und Bock 2009; Vergani und Bliuc 2015, 2018; Torregrosa et al. 2020; Shrestha et al. 2020). Im konkret handelt es sich um die 12 LIWC-Kategorien[7]: positive-, negative Emotionen, Ärger/Wut, Angst/Furcht, Referenz auf Andere, Tod/Sterben, Erfüllung, Freunde, Gewissheit, Religion, Macht, 1. Person Singular.

Abbildung 4 zeigt ein ähnliches Muster, wie wenn man nach den oben genannten linguistischen Zeichen in sozialen Netzwerken, das auch nicht nur aus radikalen Inhalt bestand, sucht (Shrestha et al. 2020), und zwar, die LIWC Indikatoren im fürs Projekt gesammelte Korpus waren insgesamt wenig präsent (vor allem die Indikatoren „Referenz auf Andere“ und „1. Person Singular“). Die häufigsten Wörter in den acht untersuchten YouTube Kanälen waren nichtsdestotrotz diejenigen, die A) mit Religion und positiven Emotionen verbunden waren (insbesondere in den Kanälen "[redacted]" und „[redacted]“), die B) negative Emotionen ausgedrückt haben (vor allem die Kanäle „[redacted]“ und „[redacted]“, und die C) mit Macht verbunden waren (vor allem die Kanäle „[redacted]“ und „[redacted]“) und die D) Bezug zum LIWC-Indikator „Erfüllung“ hatten (insbesondere die Kanäle „[redacted]„, und "[redacted]“). Weniger ausgeprägt in den ausgewerteten Kanälen waren ansonsten Begriffe, die E) mit den Indikatoren Gewissheit und Ärger/Wut verbunden waren (wobei die Kanäle „[redacted]“ und „[redacted]“ die höchsten Werten dieser Variablen gezeigt haben) und die F) einen Bezug zum Tod, Angst/Furcht, und Freunde hatten, wobei die Kanäle „[redacted]“, „[redacted]“ und „[redacted]“ jeweils die höchsten Werten der letzteren drei Variablen gezeigt haben.

[7] die ursprünglichen Kategorien sind auf Englisch und heißen: Positive emotion, Negative emotion, Anger, Anxiety, Other, Death, Achievement, Friends, Certainty, Religion, Power, 1st pers singular, 3rd pers singular

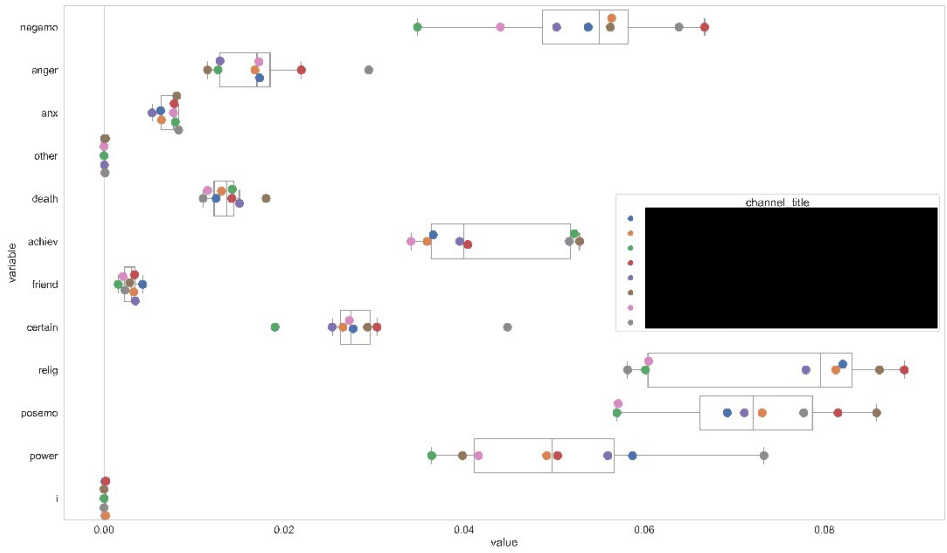


Abbildung 4. Boxplot ausgewählter LIWC-Indikatoren in den untersuchten Kanälen (Kanalnamen absichtlich – aus Datenschutzgründen – geschwärzt)

Wenn man die Kommentare, die die Themenmodellierung in Themen eingeordnet hat--insbesondere in den fünf ausgewählten Themen--mit den o.g. zwölf LIWC-Indikatoren auswertet, bekommt man mindestens zwei interessanten Beobachtungen. Einerseits, Religion ist nicht mehr der ausgeprägtste Indikator, sondern die LIWC-Kategorien bezüglich positiven-, negativen Emotionen und Macht waren die ausgeprägtsten Variablen. Andererseits, die markant hohen Werten (c.f. X-Achsen), die in den fünf ausgewählten Themen eingeordnete Kommentare in den o.g. präsentesten Variablen gezeigt haben, unterstützen- und passen sinnvoll zur semantischen Einordnung, die von der Themenmodellierung resultiert hat. Und zwar, Kommentare ins Thema 10 („Islam vs. andere Religionen und Glaubenssätze“) eingeordnet zeigen eine deutlich größere Proportion an Wörter, die mit Religion verbunden sind; Kommentare rund um „Heuchler und Feinde“ (Thema 28) zeigen ebenso einen auffallend hohen Wert in den Indikatoren für negativen- und positiven Emotionen, sowie Kommentare, die sich mit „Geschlechterrollen und -verhältnisse, Ehe und Sexualität“ (Thema 7) beschäftigt haben, sind von einer unübersehbar höheren Proportion an Wörter mit Bezug auf "Macht" geprägt.

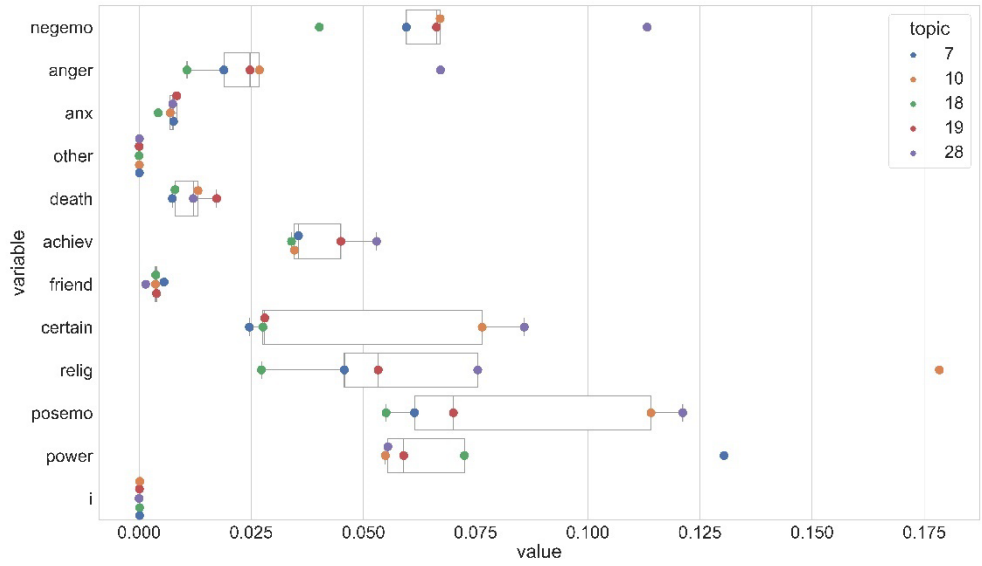


Abbildung 5. Boxplot ausgewählter LIWC-Indikatoren in ausgewählten Themen

Literaturverzeichnis

Karell, Daniel; Freedman, Michael (2019): Rhetorics of Radicalism. In: *Am Sociol Rev* 84 (4), S. 726–753. DOI: 10.1177/0003122419859519.

Krippendorff, K.; Bock, M. A. (2009): Computerized Text Analysis of Al-Qaeda Transcripts. In: James Pennebaker und C. K. Chung (Hg.): The Content Analysis Reader: SAGE Publications.

McDonald, Maura; Porter, Rachel; Treul, Sarah A. (2020): Running as a Woman? Candidate Presentation in the 2018 Midterms. In: *Political Research Quarterly* 73 (4), S. 967–987. DOI: 10.1177/1065912920915787.

Nielsen, Richard A. (2017): Deadly Clerics. Blocked Ambition and the Paths to Jihad. Cambridge: Cambridge University Press (Cambridge Studies in Comparative Politics).

Pelzer, Robert; Uhlenbrock, Mathias (2021): Möglichkeiten und Grenzen der Klassifizierung salafistisch-dschihadistischer Inhalte in sozialen Medien mithilfe von Verfahren maschinellen Lernens. In: Ursula Birsl, Julian Junk, Martin Kahl und Robert Pelzer (Hg.): Inszenieren und Mobilisieren. Rechte und islamistische Akteure digital und analog: Verlag Barbara Budrich.

Puschmann, Cornelius; Ausserhofer, Julian; Šlerka, Josef (2020): Converging on a nativist core? Comparing issues on the Facebook pages of the Pegida movement and the Alternative for Germany. In: *European Journal of Communication* 35 (3), S. 230–248. DOI: 10.1177/0267323120922068.

Rauchfleisch, Adrian; Kaiser, Jonas (2020): The German Far-right on YouTube. An Analysis of User Overlap and User Comments. In: *Journal of Broadcasting & Electronic Media* 64 (3). DOI: 10.1080/08838151.2020.1799690.

Roberts, Margaret E.; Stewart, Brandon M.; Tingley, Dustin (2019): stm. An R Package for Structural Topic Models. In: *J. Stat. Soft.* 91 (2). DOI: 10.18637/jss.v091.i02.

Schwemmer, Carsten; Jungkunz, Sebastian (2019): Whose ideas are worth spreading? The representation of women and ethnic groups in TED talks. In: *Political Research Exchange* 1 (1), S. 1–23. DOI: 10.1080/2474736X.2019.1646102.

Schwemmer, Carsten; Ziewiecki, Sandra (2018): Social Media Sellout. The Increasing Role of Product Promotion on YouTube. In: *Social Media + Society* 4 (3), 205630511878672. DOI: 10.1177/2056305118786720.

Scrivens, Ryan; Davies, Garth; Frank, Richard (2017): Searching for signs of extremism on the web. An introduction to Sentiment-based Identification of Radical Authors. In: *Behavioral Sciences of Terrorism and Political Aggression* 10 (1), S. 39–59. DOI: 10.1080/19434472.2016.1276612.

Shrestha, Amendra (2019): Techniques for analyzing digital environments from a security perspective. In:

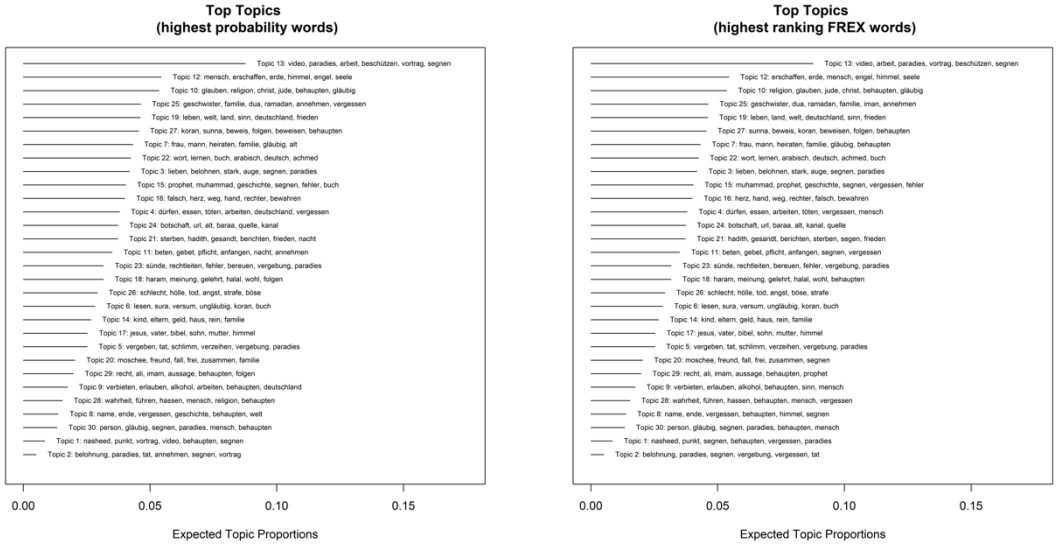
Shrestha, Amendra; Kaati, Lisa; Cohen, Katie (2020): Extreme adopters in digital communities. In: *Journal of Threat Assessment and Management* 7 (1-2), S. 72–84. DOI: 10.1037/tam0000143.

Torregrosa, Javier; Thorburn, Joshua; Lara-Cabrera, Raúl; Camacho, David; Trujillo, Humberto M. (2020): Linguistic analysis of pro-ISIS users on Twitter. In: *Behavioral Sciences of Terrorism and Political Aggression* 12 (3), S. 171–185. DOI: 10.1080/19434472.2019.1651751.

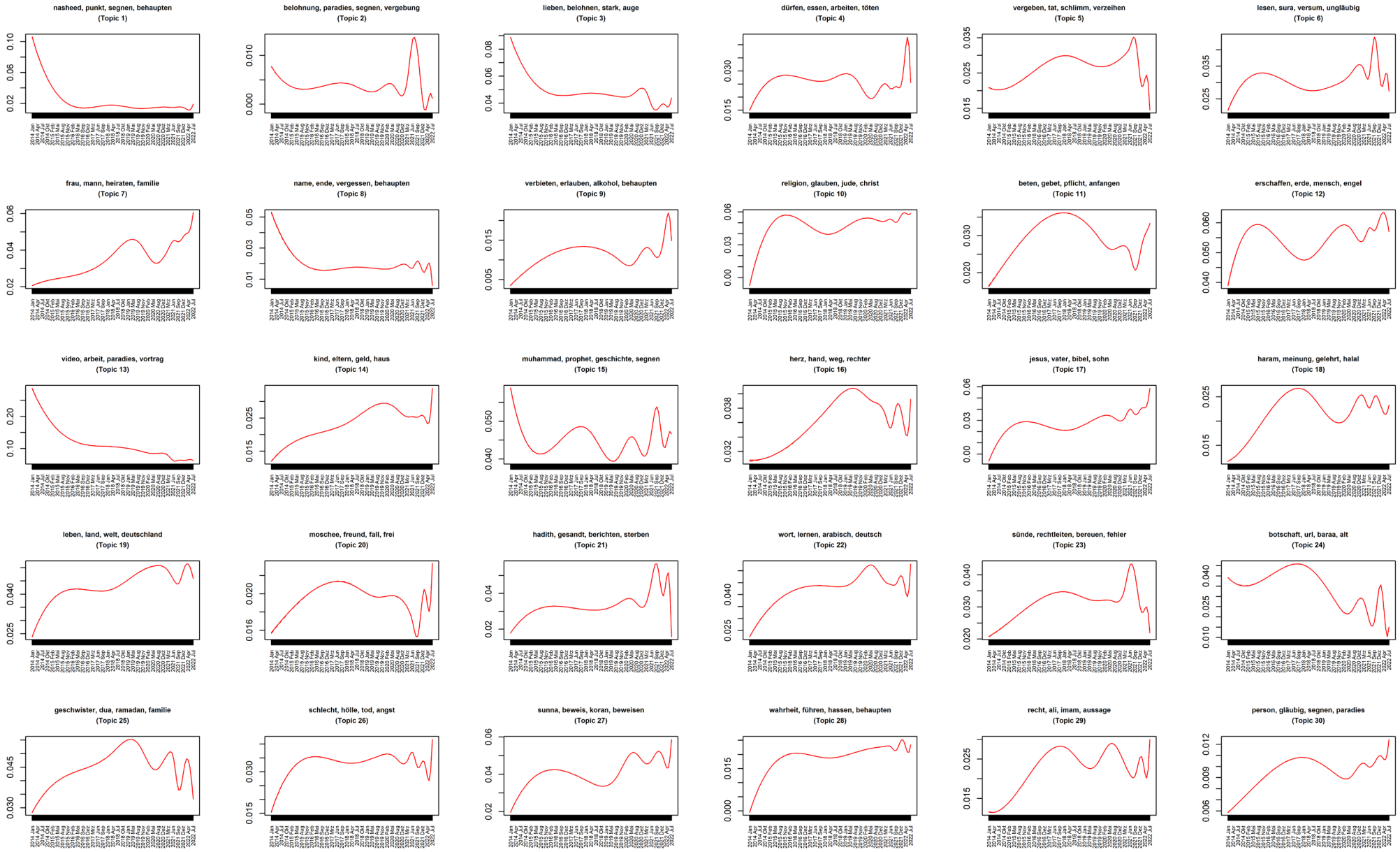
Vergani, Matteo; Bliuc, Ana-Maria (2015): The evolution of the ISIS' language. A quantitative analysis of the language of the first year of Dabiq magazine. In:

Vergani, Matteo; Bliuc, Ana-Maria (2018): The Language of New Terrorism. Differences in Psychological Dimensions of Communication in Dabiq and Inspire. In: *Journal of Language and Social Psychology* 37 (5), S. 523–540. DOI: 10.1177/0261927X17751011.

Appendix

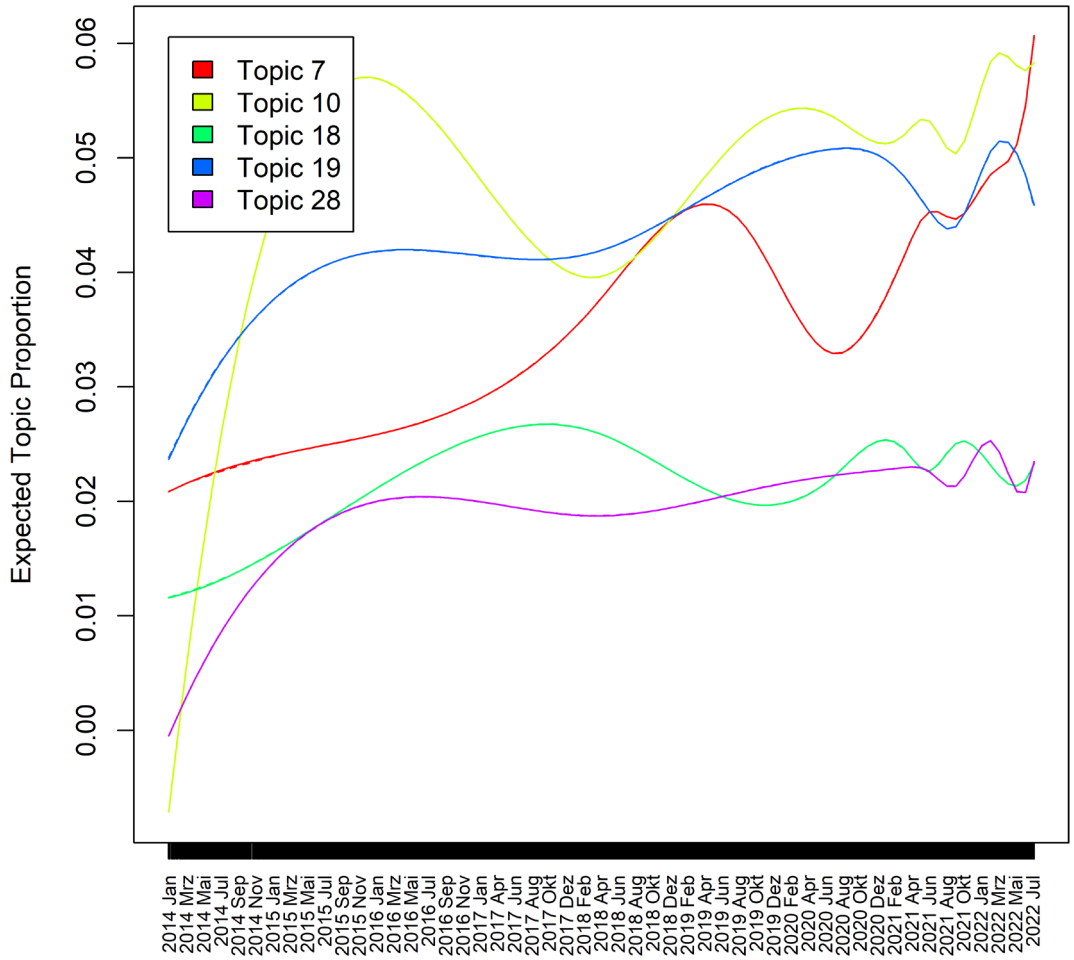


Appendix 1. Themen-Proportionen inklusive Top Wörter für alle 30 Themen

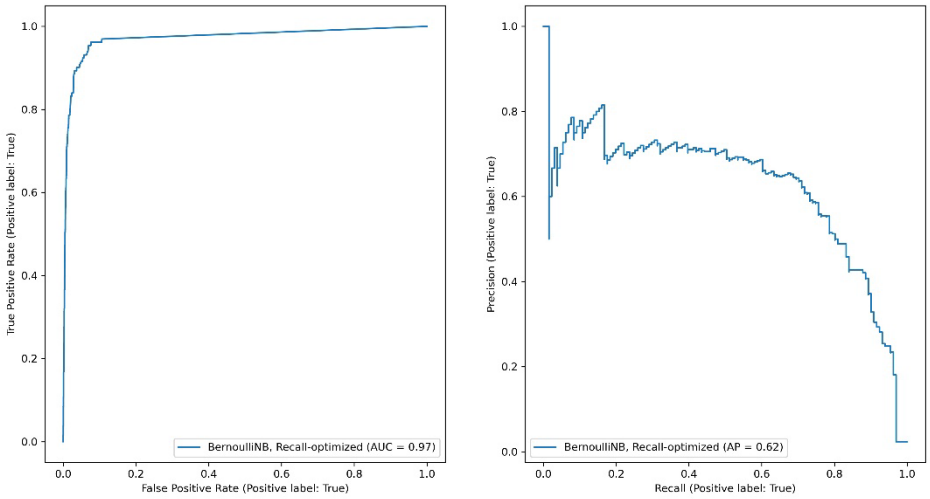


Appendix 2. Effekt von Kommentar-Veröffentlichungsdatum auf die erwartete Themaproportion (Y-Achse) bzw. Themenprävalenz über die Zeit für alle Themen.

Appendix 3. Marginal Themaproportion von jedem Kanal für alle 30 Themen (inkl. 95% Konfidenzintervall)



Appendix 4. Prävalenz der Themen über die Zeit für ausgewählte Themen (ohne Konfidenzinterval



Appendix 5. Receiver Operating Characteristic (ROC, links) und Precision-Recall (rechts) Kurven vom ausgewählten (Recall-optimisiert) Bernoulli-NB Klassifikator. Klasse-Wert 'True' steht für 'beleidigend'