

What do you do with 5 million posts? Versuche zum distant reading religiöser Online-Foren

Pfahler, Lukas

lukas@wandelt-pfahler.de
Technische Universität Dortmund

Elwert, Frederik

frederik.elwert@rub.de
RUB Bochum, Deutschland

Tabti, Samira

Samira.Tabti@ruhr-uni-bochum.de
RUB Bochum, Deutschland

Morik, Katharina

katharina.morik@tu-dortmund.de
Technische Universität Dortmund

Krech, Volker

volkhard.krech@rub.de
RUB Bochum, Deutschland

Einleitung

Religiöse Kommunikation als Teil moderner Gesellschaften findet zunehmend auch über internetbasierte Medien statt. Dabei sind es nicht nur liberale Gruppen, die diese neuen Medien nutzen, sondern gerade auch neo-konservative Gemeinschaften wie etwa Evangelikale oder Salafisten. Vor diesem Hintergrund nehmen wir ein spezielles Segment gegenwärtiger Religiosität in den Blick: Neo-konservative christliche und islamische Bewegungen (etwa Evangelikale oder Anhänger der Salafiyya) haben in den letzten Jahren mit eigenen Online-Foren Kommunikationsplattformen geschaffen, in denen sie jeweils eigene Auslegungen in Theologie und Fragen der Lebensführung diskutieren (Becker 2009: 9, Neumaier 2016).

Bei allen Unterschieden zeichnen sich diese Bewegungen durch zwei Merkmale aus: a) eine Universalisierung von Religion im Sinne einer Ablösung „reiner“ Religion von Kultur und Politik, u. b) eine religiöse Durchdringung aller Lebensbereiche, die sich insbesondere durch eine umfassende Regulierung der Lebensführung ausdrückt (Roy 2010: 57). Die Analyse dieser Online-Communities erlaubt es, Rückschlüsse über die Entwicklung und Verbreitung bestimmter Vorstellungen, aber auch über

die Genese sozialer Strukturen und neuer Autoritäten zu ziehen.

Ein besonderes Augenmerk legen wir auf die diskutierten Inhalte. Themen und ihre zeitliche Entwicklung werden über Topic Models, Keyword-Analysen und ähnliche Verfahren untersucht. Damit lassen sich thematische Konjunktoren und religiöse Traditionseinflüsse identifizieren.

Datenerhebung und -aufbereitung

Als Grundlage unserer Analysen dienen vier Online-Foren: Zwei christliche (jesus.de seit 2009 online/Christianchat.com seit 2012 online) und zwei muslimische (ahlu-sunnah.com von 2008-2016 online/Ummah.com seit 2002 online), wobei jeweils eins überwiegend deutschsprachig und eins überwiegend englischsprachig ist. Mithilfe eines Web-Crawlers wurden erhebliche Teile der Foren heruntergeladen und für die Analyse zur Verfügung gestellt. Aus den erhobenen HTML Daten werden alle Formatierungen entfernt, sodass der reine textuelle Inhalt vorliegt. Standardtechniken zur digitalen Verarbeitung natürlicher Sprache werden angewandt, um den Text weiter zu normalisieren. Dazu gehören Tokenisierung, Konvertierung aller Buchstaben zu Kleinbuchstaben, Entfernen von Sonder- und Satzzeichen, etc. Wir entfernen Wörter, die insgesamt seltener als 10 mal verwendet werden. So erhalten wir insgesamt über 5,52 Mio. Posts in über 260,000 Threads oder mehr als 470 Mio. Wörter an Daten.

Des Weiteren verwenden wir domänenspezifische Ersetzungsregeln, um verschiedene Schreibweisen von Referenzen auf externe Quellen wie Koran und Bibel oder externe religiöse Autoritäten wie Schriftgelehrte zu normalisieren. Dies erlaubt es uns ganze Foren hinsichtlich ihrer religiösen Ausrichtung zu untersuchen, da sich je nach Ausrichtung die vornehmlich referenzierten Gelehrten und Textpassagen unterscheiden.

Methoden

Ein häufiges Problem automatischer Textverarbeitung ist, dass zwei Texte zum selben Thema vollständig verschiedene Wörter verwenden können. Ein zentraler Analyseschritt ist aber das automatische Gruppieren ähnlicher Dokumente, genannt Clustering. Clustering eignet sich zum einen, um einen Überblick über die vorherrschenden Themen in Online-Foren zu verschaffen, andererseits eignet es sich auch als Sampling-Instrument um fokussiert Teilmengen für eine manuelle Inhaltsanalyse auszuwählen. Wie aber erkennt man thematische Ähnlichkeiten, wenn ein Vergleich der Mengen der Wörter nicht ausreicht?

Das populäre Topic-Modeling-Verfahren Latent Dirichlet Allocation (LDA) (Blei et al. 2003) berechnet in einem Schritt latente Repräsentationen und Gruppierungen:

Dokumente werden einem oder mehreren Topics zugewiesen, die Vektoren der Topic-Zugehörigkeiten dienen als latente Repräsentation. Stellt sich bei der Auswertung der Ergebnisse heraus, dass die Anzahl der Themen zu unpassend gewählt wurde, muss die Berechnung mit veränderten Parametern wiederholt werden. Dabei ist nicht garantiert, dass sich genau die Topics vereinigen oder aufspalten, an denen der Anwender festgemacht hat, dass die Anzahl falsch gewählt wurde. Weiterhin ist es rechenaufwendig, die Granularität der Analyse zu verändern, da die volle Berechnung mit den veränderten Parametern wiederholt werden.

Statt LDA-Repräsentationen zu berechnen, verwenden wir die konzeptionell einfacheren Document Embeddings nach Le und Mikolov (2014). Die latente Repräsentation x ist hier ein niedrig-dimensionaler, reellwertiger Parametervektor einer diskreten kategorischen Verteilung über Wörter in einem Dokument $P(w | x, u) \sim \exp(u'x)$. Diese Parametervektoren sowie der Parametervektor der Wortverteilungen u werden so gewählt, dass die Likelihood der Daten maximiert wird. Dieses Optimierungsproblem betrachten wir als Matrix-Faktorisierungsproblem; statt über die x und u zu optimieren, optimieren wir über die Matrix der jeweiligen Skalarprodukte $u'x$.

Mithilfe eines numerischen Optimierungsverfahrens können die latenten Repräsentationen berechnet werden.

In einem zweiten Schritt wird das Clustering der Dokumentenkollektion auf Basis der latenten Repräsentationen berechnet. Hierzu verwenden wir das Agglomerative Hierarchische Clustering. Das Verfahren berechnet einen Cluster-Baum, an dessen Blättern die einzelnen Dokumente liegen; durch sukzessive Vereinigung der zwei ähnlichsten Cluster entsteht ein Baum. Dieser erlaubt uns beliebige Anzahlen von Gruppen zu identifizieren, indem der Baum auf einer festgelegten Höhe abgeschnitten wird. Soll die Clusteranzahl verändert werden, müssen die latenten Repräsentationen der Dokumente nicht neu berechnet werden, es muss lediglich der vollständige Baum anders abgeschnitten werden. Dazu gibt es verschiedene Möglichkeiten: Einzelne Cluster können in ihre zwei Untercluster aufgespalten werden oder andersrum wiedervereinigt werden oder es kann eine andere feste Gesamtanzahl angegeben werden. So kann interaktiv und dynamisch ein Clustering der Dokumente erarbeitet werden; jede Änderung der Cluster-Hierarchie ist in wenigen Sekunden berechnet.

Damit das Clustering interpretiert werden kann, werden Repräsentanten der Cluster berechnet. Wie die Topics bei LDA-Modellen handelt es sich auch hierbei um Wortlisten, die prinzipiell angeben, wie oft die jeweiligen Wörter in jedem Cluster vorkommen.

Besser interpretierbare Ergebnisse werden erzielt, wenn statt der Anzahl der TFIDF-Score (Robertson 2004) der Wörter angegeben wird; dieser setzt die Anzahlen ins Verhältnis zur Anzahl der Dokumente, in denen die jeweiligen Wörter verwendet werden.

Vorläufige Ergebnisse

Das oben vorgestellte Verfahren erlaubt es, die ausgewählten Korpora für die weitere qualitative Feinanalyse gezielt aufzuarbeiten. Die folgende Tabelle zeigt exemplarisch 6 verschiedene Themen, die wir in den Foren-Diskussionen identifiziert haben. In diesen semantischen Feldern finden die unterschiedlichen Diskurse immer im Rahmen religiöser Argumentation und Legitimation statt. Beginnend mit Diskussionen über das Weltgeschehen, besonders im islamischen Raum (Themen 4,5), zur Lebensführung, in diesem Fall Ernährung (Thema 1), über religiöse Rituale oder Handlungen wie Heirat oder Gottesdienst (Thema 2,3) bis zu den unmittelbaren theologischen Diskussionen wie bspw. Diskussionen religiöser Schriften (Thema 6).

Thema 1	Thema 2	Thema 3
halal fleisch enthalten alkohol trinken essen tier blut haraam	moscheen gebetet beten verrichten freitagsgebet raum gebete isha stadt	heirat heiraten wali ehe verheiratet scheidung ehemann vater mahram

Thema 4	Thema 5	Thema 6
staat demokratie gesellschaft krieg ländern gesetze länder regierung staaten	soldaten politik spiegel afghanistan taliban krieg israel regierung usa	sonne wohlgefallen himmel paradies erklärung überlieferung erde sallallahu berichtete

Die Frage der Lebensführung in den neokonservativen Religionsgemeinschaften ist für unsere Forschung zentral. Die Identifizierung von Themenbereichen und Schlüsselwörtern, die mit Lebensführungsfragen in Verbindung stehen, können erste Hinweise darüber geben, was für die salafistisch korrekte Lebensführung besonders stark diskutiert wird. Auch interessiert uns, wie stark diese Themenkomplexe mit religiösem Vokabular durchsetzt sind.

Für verschiedene Bereiche der Lebensführung lassen sich diesbezüglich folgende Beobachtungen machen:

Ernährung und Lebensmittel werden hier besonders im Zusammenhang mit haram/halal (erlaubtes/nicht erlaubtes), also insbesondere aus einer religiösen Perspektive, diskutiert. Daneben finden sich aber auch eher auf Gesundheitsfragen ausgerichtete Diskussionen.

Musik als Thema wird auffällig häufig diskutiert und taucht in unterschiedlichen thematischen Clustern

auf: Einmal im Zusammenhang mit Musikinstrumenten und der Bewertung als haram/halal, aber auch im Themenfeld Medien (gemeinsam mit Film und Fotografie) sowie im Zusammenhang mit negativ konnotierten Verhaltensweisen (Alkohol, Glücksspiel).

Im nächsten Schritt sollen Themen und religiöse Quellen/Autoritäten miteinander in Beziehung gesetzt werden. Dabei interessiert uns die Frage, auf welche Schriftquellen und welche Gelehrten sich die AkteurInnen berufen, wenn sie bestimmte Themen diskutieren.

Am Beispiel ahlusunnah.com zeigt sich, dass bestimmte salafistische Schriftgelehrte eine bedeutendere Rolle spielen als andere (siehe Abbildung 1). Die Zwischenergebnisse lassen erste Rückschlüsse auf die religiösen Richtungen und Referenzpraktiken zu: welche salafistischen Gelehrten finden mehr Zustimmung als andere und welche salafistischen Traditionsschulen lassen sich identifizieren (Wahabiya, Ad-Da'wa As-Salafiya, Madchalia etc.). So sind z.B. die Gelehrten Ibn Taymiyyah und al-Albani wichtige Referenzgrößen. Ibn Taymiyyah (13. Jhd.) als "Vater der Salafiyya" findet große Verehrung bei den puristischen sowie auch bei den politisch-aktivistischen Gruppierungen der Salafiya-Bewegung. Al-Albani (20. Jhd.) dagegen findet eher in der puristischen Bewegung Zuspruch und wird auch kontrovers diskutiert. Vor dem Hintergrund der thematischen Analyse lässt sich dann noch weiter untersuchen, welche Quellen verstärkt für welche Themenkomplexe

Neumaier, Anna (2016): Religion@home? Religionsbezogene Online-Plattformen Und Ihre Nutzung: Eine Untersuchung Zu Neuen Formen Gegenwärtiger Religiosität. Religion in Der Gesellschaft 39.

Pang-Ning, Tan/Michael, Steinbach/Vipin, Kumar (2006): Introduction to data mining

Pfahler, Lukas/Katharina, Morik/Frederik, Elwert/Samira, Tabti/Volkhard, Krech (2017): "Learning Low-Rank Document Embeddings with Weighted Nuclear Norm Regularization" in: Proceedings of the 4th DSAA

Robertson, Stephen (2004) "Understanding inverse document frequency: on theoretical arguments for IDF," in: J. Doc., vol. 60, no. 5, pp. 503–520.

Roy, Olivier (2010): Heilige Einfalt: Über Die Politischen Gefahren Entwurzelter Religionen. München.

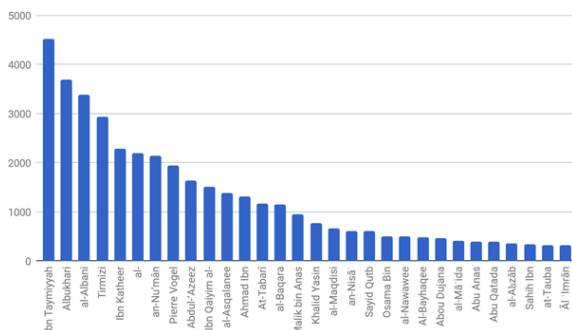


Abbildung : Quellenreferenzen in ahlusunnah.com

Bibliographie

Becker, Carmen (2009): "Gaining Knowledge: Salafi Activism in German and Dutch Online Forums" in: Masaryk University Journal of Law and Technology 3 (1): 79–98.

Blei, David M./Andrew Y. Ng/Michael I. Jordan (2003): "Latent Dirichlet Allocation," J. Mach. Learn. Res., vol. 3, no. 4–5, pp. 993–1022.

LeQuoc V./ Tomas Mikolov (2014): "Distributed Representations of Sentences and Documents" in: International Conference on Machine Learning - ICML 2014, vol. 32, pp. 1188–1196.