

Standortbestimmung mittels ADS-B Sensordaten und maschinellem Lernen

David Fuhry¹

Universität Leipzig
Machine Learning Group
Leipzig, Germany
`df15nocu@studserv.uni-leipzig.de`

Zusammenfassung. Es wird ein Verfahren vorgestellt um mit ADS-B Sensordaten den Breitengrad eines Flugzeugstandortes zu ermitteln. Zum Einsatz kommen Regressionsverfahren des maschinellen Lernens. Der Breitengrad kann mit einer mittleren Genauigkeit von unter 20km bestimmt werden. Für eine produktive Anwendung in der Luftraumüberwachung müssen die eingesetzten Verfahren weiter optimiert werden.

Schlüsselwörter: Maschinelles Lernen, ADS-B, Regression, Positionsbestimmung

1 Einleitung

Verschwörungstheorien sind aktuell häufig im Zentrum der Aufmerksamkeit. Während dieser, aktuelle Aufmerksamkeitsschub inhaltlich auf die aktuelle Pandemie zurückzuführen ist, bietet das Internet und die Verbreitung sozialer Medien das technische Fundament für die schnelle Verbreitung wie sie aktuell stattfindet. Das zunehmende Vernetzung solchen Theorien vorschub leistet ist keine neue Entwicklung, sondern wird schon länger im wissenschaftlichen Diskurs über Verschwörungstheorien beobachtet. Die wissenschaftliche Auseinandersetzung mit Verschwörungstheorien ist trotz des verstärkten Interesses in der jüngsten Vergangenheit bis heute eher verhalten.

Die komplexe Natur von Verschwörungstheorien stellt die wissenschaftliche Auseinandersetzung mit ihnen vor einige Grundlegende Probleme. Das offensichtlichste davon ist sicherlich die Frage was überhaupt eine Verschwörungstheorie ist. Hierzu existiert vor allem in der Wissenssoziologie eine Vielzahl von Literatur, ein klarer Konsens über eine Definition von Verschwörungstheorien fehlt jedoch bislang. Insbesondere für die Arbeit mit Beiträgen aus sozialen Netzen und anderen Online-Quellen schließt sich die Frage an, wie Verschwörungstheoretische Beiträge zuverlässig identifiziert werden können. Der Nutzen solcher Techniken nicht nur innerhalb der Wissenschaft ist naheliegend, so versuchen auch die Betreiber von sozialen Netzwerken daran verschwörungstheoretische Inhalte automatisiert zu erkennen (Beispiel google/youztube). In der Forschung wird dieses Problem klassischerweise mit der Hilfe einer Vielzahl von Hilfskräften gelöst, die die Daten manuell bearbeiten. (Studie mit codierern)

Diese Arbeit stellt einen Ansatz vor, dieses Problem mit Technologien der automatischen Sprachverarbeitung und des maschinellen Lernens zu lösen. Primär werden dabei stilistische Merkmale der untersuchten Texte genutzt und nur sekundär inhaltliche Informationen wie das Verwenden von Schlüsselwörtern etc. Durch das Verwenden solcher Technologien kann zum einen der Ressourcenaufwand drastisch reduziert werden, aber auch die Abhängigkeit von komplexen und damit fehleranfälligen Kriterien für die Klassifizierung lässt sich so umgehen. Auch ist der hier vorgestellte Ansatz durch den Fokus auf stilistische Merkmale verhältnismäßig robust gegenüber neuen inhaltlichen Entwicklungen der Verschwörungstheorien und somit potentiell zukunftssicherer als vergleichbare Ansätze.

Im folgenden soll zunächst ein kurzer Abriss über den aktuellen Stand der wissenschaftlichen Auseinandersetzung mit Verschwörungstheorien allgemein, sowie der Forschung zu linguistischen Merkmalen im speziellen, gegeben werden. Anschließend soll kurz auf die verwendeten Technologien sowie die Datengrundlage dieser Arbeit gegeben werden, bevor konkrete Ergebnisse diskutiert werden.

2 Vorgehen

Der Arbeit zugrunde liegt ein Korpus aus Texten von Internetportalen, die der verschwörungstheoretischen, bzw. Truthther Szene zuzuordnen sind. Diese wurde im Verlauf der letzten Jahre über Webcrawling zusammengestellt. Eine genaue Beschreibung der Zusammensetzung des Korpus wird im weiteren Verlauf dieser Arbeit erfolgen.

Füllmüll?

Um eine strukturierte Analyse mit Methoden des maschinellen Lernens zu ermöglichen wird zunächst ein Vergleichskorpus mit Artikeln die von Form und Thematik möglich ähnlich den betrachteten, verschwörungstheoretischen Texten gelagert sind ohne aber die konspirativen Aspekte zu teilen.

Anschließend werden die kombinierten Korpora gereinigt und mit Verfahren aus der natürlichen Sprachverarbeitung Features extrahiert auf denen Klassifizierungsverfahren arbeiten können. Zum Einsatz kommen sollen dabei sowohl Features die klassischerweise für Textklassifikation herangezogen werden, wie etwa Wortfrequenzen, POS-Tags und Wortdichte [CITATION NEEDED] als auch Features die auf den Erkenntnissen qualitativer Forschung zu Verschwörungstheorien basieren.

Zitieren

Die häufig angesprochene Emotionalität (oder der Mangel an selbiger) [citations needed] von verschwörungstheoretischen Texten etwa soll durch eine Sentiment-Analyse betrachtet werden. Die häufig angesprochene Multimedialität/Fragmentiertheit von verschwörungstheoretischen Inhalten soll durch Analyse der eingebetteten Inhalte wie Bilder, Tweets und YouTube Videos aber auch über die Menge der Verlinkungen untersucht werden.

zitieren

Mit den so erstellten Daten soll abschließend ein Modell trainiert werden, dass zwischen den verschwörungstheoretischen Texten und dem Vergleichskorpus unterscheiden kann. Zum Einsatz kommen soll dazu *LightGBM* (Ke, 2019) eine Implementierung von Gradient Boosting Decision Trees [citation needed]. Dieses Verfahren bietet neben der relativ guten Präzision den Vorteil, dass die Ergebnisse ähnlich einfachen Entscheidungsbäumen verhältnismäßig gut interpretierbar sind, so lässt sich gut der Einfluss einzelner Features auf die Klassifizierung abschätzen.

3 Umsetzung

3.1 Datenbasis

R zitieren so nicht
schon geschehen

Als Grundlage dieser Arbeit wurde ein Korpus aus Texten von insgesamt 7 deutschsprachigen, der verschwörungstheoretiker Szene zuzurechnenden Internetangeboten erstellt. Die Skripte zum automatischen Abruf der Texte nutzen dafür das R-Paket *rvest* (Wickham, 2020) erstellt und extrahieren mittels für jede Seite separat erstellten CSS-Selektoren und XPATH-Querys den jeweiligen Artikeltext. In diesem Vorgang wurden noch kleinere Bereinigungen an den Texten vorgenommen um wiederkehrende, nicht zum Artikeltext gehörende Elemente wie Werbung oder Spendenaufrufe zu entfernen, die Texte selbst wurden jedoch soweit möglich nicht weiter bearbeitet. Weiterhin erfasst wurde zu den Artikeln das (angegebene) Veröffentlichungsdatum, der Artikeltitel sowie die Rubrik so diese Angabe vorhanden waren.

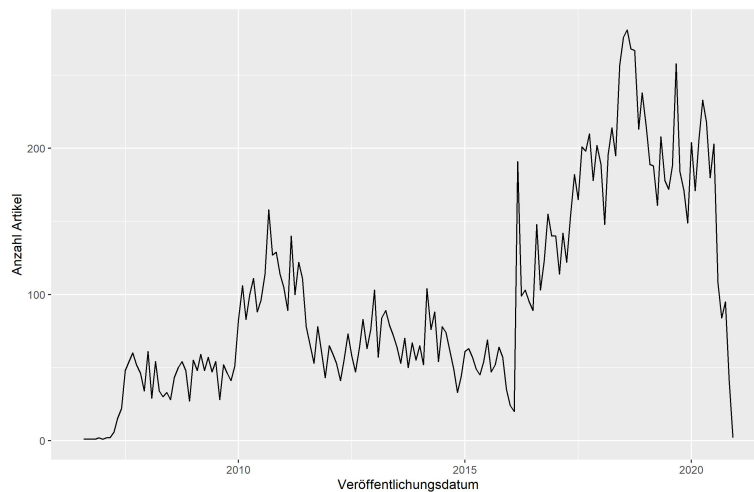


Abb. 1. Anzahl Artikel nach Veröffentlichungsdatum (nach Monat gruppiert)

Wirklich?

Der so erstellte Korpus umfasst insgesamt 16836 Texte und deckt beginnend in 2006 einen Zeitraum von 14 Jahren ab. Wie in Grafik 1 zu erkennen ist, steigt beginnend in 2016 die Zahl der Artikel im Korpus deutlich an. Dies liegt zum einen daran, dass einzelne der betrachteten Angebote die Veröffentlichungsfrequenz erhöht haben, ist aber überwiegend darauf zurückzuführen, dass nicht alle Internetangebote über den gesamten Zeitraum existiert haben sondern erst später aktiv wurden. Genauere Informationen zu den von den einzelnen Angeboten abgedeckten Zeiträumen sowie zu Artikelzahl und Länge finden sich in Tabelle 1.

Wie dort auch ersichtlich ist, unterscheidet sich nicht nur die absolute Zahl der Artikel nach Quelle stark, auch das Veröffentlichungsintervall schwankt von weniger als einem Artikel den Monat bis über 100. Es besteht weiterhin ein negativer Zusammenhang zwischen der Artikellänge und dem Veröffentlichungsintervall ($r = -0.72, p = 0.065$).

r richtiges Zeichen?

	Von	Bis	Mittlere Zeichenzahl	Anzahl Ar- tikel	Artikel/Monat
Alles Schall und Rauch	23.08.2006	02.08.2020	5000	5349	32.0
conrebbi	05.09.2012	20.10.2014	5610	24	1.0
deutschlandpranger	29.10.2016	30.12.2020	7527	118	2.4
fm-tv	31.07.2008	02.11.2018	9158	97	0.8
hinterderfichte	16.01.2010	31.05.2018	4221	1083	10.8
recentr	03.08.2007	05.08.2020	5071	4762	30.5
Watergate.tv	20.05.2016	16.11.2020	2810	5403	101.9

Tabelle 1. Kennzahlen der einzelnen Quellen des Korpus

Auch eine inhaltliche Betrachtung zeigt deutliche Unterschiede zwischen den einzelnen Angeboten. So ist etwa *Alles Schall und Rauch (ASuR)* als ältestes vertretenes Angebot auch einer der "klassischsten" Vertreter des Genres. Hier wird der Nutzer direkt auf der Startseite mit Fragen konfrontiert wie "Was geschau wirklich am 11. September? oder "Was passiert tatsächlich mit dem Klima?" (Freeman, 2021).

Das muss anders

Die Seite bedient sich klassischer Verschwörungstheorien und wird im allgemeinen der Truthther Szene zugeordnet („Alles Schall und Rauch“, 2020). Sie hat laut eigenaussage im Schnitt über 50.000 Zugriffe täglich¹.

Eventuell zitation zu häufigen fragen.

Inhaltlich finden sich sowohl Essay-artige Artikel zu Themen wie 9/11, Bilderberger oder Klimawandel auf der Seite, als auch kürzer gehaltene Meldungen zu aktuellen Ereignissen. Es werden, wie bei fast allen anderen Seiten im Korpus, häufig Inhalte Dritter eingebunden sowie die Artikel durch Bilder oder Grafiken ergänzt.

Ähnlich gelagert sind die Angebote *Hinter der Fichte (HdF)* und *conrebbi*, beide können der Truthther Szene zugerechnet werden (siehe etwa „Conrebbi“, 2020). In beiden Angeboten werden aktuelle Ereignisse kommentiert und verschwörungstheoretisch interpretiert, sowie längere Essays zu typischen Themen verfasst. Im Vergleich zu *ASuR* werden bei beiden Angeboten deutlich mehr Inhalte in Form von Videos und Fremdquellen in Bild und Textform genutzt.

Etwas aus der Reihe fallen die Angebote *fm-tv* sowie *Deutschlandpranger*. Inhaltlich finden sich etwa beim *Deutschlandpranger* recht klassische Themen

¹ Der Besucherzähler der Website ist inzwischen nicht mehr funktional, Angabe übernommen aus (Anonymous, 2020)

wie die Leugnung des Klimawandels (vgl. 2020), sind die dazugehörigen Text überdurchschnittlich lang und voller Werbung für Bücher zu z.T. komplett unabhängigen Themen. Auch sind die Texte selbst teilweise relativ wirr, so enthalten beispielsweise die (sehr ausführlichen) AGB einen Abschnitt zu Strafzahlungen bei "Übersenden eines Statements anstatt einer echten Rechnung (True Bill) des wahren Haftungsgläubigers"(2016).

Die verbleibenden beiden Angebote *recentr* und *Watergate.tv* sind beide der rechten Szene zuzuordnen und verbreiten Verschwörungstheorien in diese Richtung. Beide zeichnet ein eher journalistischer Stil aus, es werden kurze bis mittellange Meldungen veröffentlicht, häufig begleitet von eigenen oder fremden Videos. *Recentr* wurde 2006 als deutscher Ableger von *Inforwars.com* gegründet und bedient sich noch heute einem ähnlichen Konzept. Ähnlich dem Vorbild wird massiv Werbung für den eigenen Shop gemacht in dem diverse Produkte zur Vorbereitung auf apokalyptische Szenarien sowie fragwürdige Nahrungsergänzungsmittel verkauft werden. Inhaltlich ist vor allem interessant, dass sich die Seite (bzw. der Hauptautor Alexander Benesch) z.T. sehr explizit von anderen Teilen der rechten/verschwörungstheoretischen Szene abgrenzt.² So kommt es durchaus vor, dass in einem Artikel zunächst von der implausiblen Sichtweise, verschiedenste Regierungen hätten eine Fake-Pandemie orchestriert"(Benesch, 2021a) gesprochen wird, nur um im nächsten Satz zu erklären:

In Wirklichkeit sind die Methoden der globalen Bevölkerungsreduktion fast ausschließlich legal, simpel und werden schrittweise angewandt [...]. Das Konzept ist so entworfen, um mit möglichst wenig Zwang auszukommen [...]. (Benesch, 2021a)

Die Website *Watergate.tv* (inzwischen Teil von *NEOPresse.com*) ist ein rechtslastiges Nachrichten und Videoportal, das sich verschwörungstheoretischer Narrative bedient. Auf dem Portal finden sich klassische Themen wie Bilderberger (vgl. Redaktion, 2018) aber auch Fake-News und Berichterstattung mit starker pro-russischen Tendenz. In die Öffentlichkeit rückte das Portal als sich das Neo-Magazin-Royale wegen seiner Mitarbeit bei *Watergate.tv* von Hans Meiser trennte (Siehe z.B. 2017).

Kurzes Fazit Korpus

Bei der Erstellung des Vergleichskorpus war es das Ziel Quellen zu wählen die dem Korpus konspirativer Texte möglichst ähnlich sind, ohne die konspirativen Aspekte zu enthalten. Da insbesondere die letzten beiden besprochenen Angebote sich inhaltlich sehr an journalistische Angebote anlehnen, lag es nah ebensolche heranzuziehen. Aus Gründen der Vollständigkeit (Die Angebote sollten möglichst den kompletten Zeitraum des Korpus abdecken) sowie der Zugänglichkeit wurden die Angebote von Spiegel Online sowie der Frankfurter Rundschau ausgewählt. Es wurde für beide Onlineangebote zunächst ein Index aller Artikel die in den selben Zeitraum wie der Korpus fallen erstellt und aus diesem zufällig eine Auswahl von 10000 Artikeln pro Anbieter gezogen.

² So grenzt sich das Portal in Beiträgen inzwischen auch von *Inforwars* Gründer Alex Jones ab (2018) oder spekuliert darüber, ob die Alternative für Deutschland vom britischen Geheimdienst gegründet wurde (Benesch, 2021b)

Die restlichen im Korpus enthaltenen Seiten sind alle mehr oder weniger in Blogform gehalten. Um Inhalte in einer vergleichbaren Form zu finden aber auch dem in der Literatur häufig anzutreffenden Feststellungen der wissenschaftlichen Form von verschwörungstheoretischen Texten Rechnung zu tragen wurde eine Reihe von Wissenschaftsblogs als zweite Komponente im Vergleichskorpus gewählt. Dafür wurden von den beiden Plattformen scienceblogs und scilogs³ ähnlich den journalistischen Angeboten aus dem gesamten Index der Einträge eine Auswahl von jeweils 10000 Artikeln ausgewählt. Genauere Kennwerte zum Vergleichskorpus siehe in Tabelle 2.

	Von	Bis	Mittlere Zei- chenzahl	Anzahl Artikel	Artikel/Monat
Frankfurter Rundschau	03.08.2006	04.08.2020	3670	9959	59.3
scienceblogs	14.02.2007	04.08.2020	3261	8765	54.4
scilogs	29.12.2000	03.08.2020	5206	9533	40.6
Spiegel online	On- 01.08.2006	07.01.2020	4687	8665	53.8

Tabelle 2. Kennzahlen der einzelnen Quellen des Vergleichskorpus

3.2 Datenvorverarbeitung

Zunächst wurden einige Säuberungsschritte auf den Korpus angewendet, um zu verhindern, dass fehlerhafte oder irrelevante Daten mit für die Feature Erstellung herangezogen werden.

Es wurde zunächst Unicode Normalisierung für alle Texte durchgeführt, anschließend wurden mittels des *Google Compact Language Detector 3* (Ooms, 2021) die Sprache aller Text bestimmt nur (überwiegend) deutschsprachige Texte zu betrachten. Da im Korpus an einigen Stellen noch nicht zum Artikeltext gehörige Komponenten aus vielen Sonderzeichen vorhanden waren, wurden alle Wörter entfernt, die nur aus Zeichen bestanden, die unter 1000 Mal im gesamten Korpus vorkamen. Schlussendlich wurden noch kleinere Korrekturen vorgenommen, wie das Zusammenfassen multipler, aufeinanderfolgender Leerzeichen und Unterstriche (die gerne als Trennlinien in den Text benutzt wurden).

Schlussendlich wurden noch alle Text unter 100 Zeichen Länge aus dem Korpus entfernt, da diese für die Klassifizierungsaufgabe nur wenig Informationen enthalten können. Der finale Korpus umfasste 53758 Texte, mit einem Anteil der positiven Klasse (sprich von verschwörungstheoretischen Texten) von etwa 32%.

³ Beide Plattformen sind Anbieter bei denen eine Vielzahl verschiedener Blogs zu finden sind.

8 Fuhry

3.3 Feature Erstellung

Software

- Ke, G. (2019). *lightgbm: Light Gradient Boosting Machine* [R package version 2.3.2]. R package version 2.3.2. <https://github.com/Microsoft/LightGBM>
- Ooms, J. (2021). *cld3: Google's Compact Language Detector 3* [R package version 1.4.1]. R package version 1.4.1. <https://CRAN.R-project.org/package=cld3>
- Wickham, H. (2020). *rvest: Easily Harvest (Scrape) Web Pages* [R package version 0.3.6]. R package version 0.3.6. <https://CRAN.R-project.org/package=rvest>

Online

- Alles Schall und Rauch. (2020). Psiram. Verfügbar 23. Mai 2021 unter https://www.psiram.com/de/index.php/Alles_Schall_und_Rauch
- Anonymous. (2020). *Mein Vater ist der Verschwörungstheoretiker hinter Alles Schall und Rauch*. Vice. Verfügbar 23. Mai 2021 unter <https://www.vice.com/de/article/gyx8aq/mein-vater-der-bekannte-verschwörungstheoretiker>
- Benesch, A. (2018). *Alex Jones: Internierungslager und REX-84 sind jetzt die Rettung Amerikas*. Recentr. Verfügbar 23. Mai 2021 unter <http://recentr.com/2018/06/26/alex-jones-internierungslager-und-rex-84-sind-jetzt-die-rettung-amerikas/>
- Benesch, A. (2021a). *Bevölkerungsreduktion wird schrittweise, mit fast ausschließlich legalen Mitteln betrieben, anders als viele sich das vorstellen*. Recentr. Verfügbar 23. Mai 2021 unter <http://recentr.com/2021/05/10/bevoelkerungsreduktion-wird-schrittweise-mit-fast-ausschliesslich-legalen-mitteln-betrieben-anders-als-viele-sich-das-vorstellen/>
- Benesch, A. (2021b). *Ist die AfD eine britische Schöpfung?* Recentr. Verfügbar 23. Mai 2021 unter <http://recentr.com/2021/06/29/ist-die-afd-eine-britische-schoepfung/>
- Blog AGB's. (2016). Deutschland - Pranger. Verfügbar 23. Mai 2021 unter <http://deutschland-pranger.de/b2evolution/index.php/ueber-diese-seite>
- Conrebbi. (2020). Psiram. Verfügbar 23. Mai 2021 unter <https://www.psiram.com/de/index.php/Conrebbi>
- Conrebbi. (2020). Psiram. Verfügbar 23. Mai 2021 unter <https://www.psiram.com/de/index.php/Conrebbi>
- Freeman. (2021). *Homepage*. Verfügbar 23. Mai 2021 unter <http://alles-schallundrauch.blogspot.com/>
- Redaktion, W. (2018). *+++ Schulz tritt sofort zurück – Bilderberger Scholz gibt jetzt den Ton an +++*. Watergate.tv. Verfügbar 23. Mai 2021 unter <https://wg.neopresse.com/schulz-tritt-sofort-zurueck-bilderberger-scholz-gibt-jetzt-den-ton-an/>
- ZDF bestätigt Trennung von Hans Meiser. (2017). Spiegel Online. Verfügbar 23. Mai 2021 unter <https://www.spiegel.de/kultur/tv/jan-boehmermann-neo-magazin-royale-kuenftig-ohne-hans-meiser-a-1148048.html>