Hamburgisches
WeltWirtschafts
Institut

Reihe Edition HWWI Band 6

Textdaten
Anwendungen und Herausforderungen

Silke Sturm

In:

Neuvermessung der Datenökonomie

herausgegeben von Thomas Straubhaar

Seite 129–156

Hamburg University Press
Verlag der Staats- und Universitätsbibliothek Hamburg
Carl von Ossietzky

Impressum

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen National- bibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

ISSN 1865-7974

ONLINE-AUSGABE
Die Online-Ausgabe dieses Werkes ist eine Open-Access-Publikation und ist auf den Verlagswebseiten frei verfügbar. Die Deutsche Nationalbibliothek hat die Online-Ausgabe archiviert. Diese ist dauerhaft auf dem Archivserver der Deutschen Nationalbibliothek (https://portal.dnb.de) verfügbar.

DOI https://doi.org/10.15460/HUP.HWWI.6.212

ISBN
Print: 978-3-943423-91-4
EPUB: 978-3-943423-94-5

Satz Hamburg University Press

Covergestaltung Hamburg University Press unter Verwendung eines Fotos von Free-Photos auf Pixabay (https://pixabay.com/images/id-768432)

Druck und Bindung Books on Demand (Norderstedt)

Verlag
Hamburg University Press, Verlag der Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky, Hamburg (Deutschland), 2021
https://hup.sub.uni-hamburg.de

Inhalt

Textdaten

Anwendungen und Herausforderungen

Silke Sturm

Einleitung

Textdaten sind die am stärksten wachsende Datenquelle. Seit zwei Jahrzehnten nimmt die Zahl der Haushalte mit dauerhaftem Internetzugang zu, häufig ist dieser durch mobile Daten zusätzlich unabhängig von ihrem Aufenthaltsort.[1] Es werden auf Seiten aller gesellschaftlichen und wirtschaftlichen Akteure Beiträge verfasst, wobei sowohl redaktionell bearbeitete als auch privat verfasste Texte relevant sind. Durch die größere Bedeutung von Textdaten und die sich stetig verbessernde Rechenleistung entwickelt sich der Bereich automatisierter Textanalysen in vielfältigen Fachbereichen dynamisch.

Die Nutzung einer großen Bandbreite von Veröffentlichungen hat Vorteile in Form einer allgemeineren Abdeckung relevanter Themen und Meinungen und damit einer besseren Abschätzbarkeit von Entwicklungen. Allerdings ist durch die Varianz der Kommunikation, welche sich durch Länge, Qualität, Vokabular oder Menge der verschiedenen Themen unterscheidet, die Auswertung schwierig und bedarf eines präzisen Vorgehens. Das Ziel, die Chancen der Auswertung und das Verständnis wirtschaftlicher und politischer Zusammenhänge zu verbessern, ist die treibende Kraft hinter den aktuell steigenden Forschungsanstrengungen.[2]

Dieser Beitrag beschäftigt sich mit der automatisierten Textanalyse und ihren Vor- und Nachteilen. Darüber hinaus widmet er sich einer Betrachtung der Chancen und Herausforderungen Sozialer Medien und politisch-gesellschaftli-

---

[1] Kumar/Das (2013); WorldBank (2019).

[2] Einav/Levin (2014).

cher Texte. Die Anwendbarkeit unstrukturierter Textdaten für Forschungsinteressen wird anhand eines exemplarischen Auszugs aus der politischen Diskussion in der Legislaturperiode 2014 bis 2017 dargestellt.

Texte als Daten

Vor- und Nachteile textbasierter Daten

Der Vorteil strukturierter Daten in Form von Statistiken, Surveys, tabellarischen Aufstellungen oder weiteren quantitativen Erhebungen ist naheliegend. Die Daten liegen bereits in maschinenlesbarer Form vor, können ohne umfangreiche Vorbereitungen genutzt werden und sind weitestgehend vergleichbar. Statistiken ökonomisch und gesellschaftlich relevanter Daten liegen zum Teil in Aggregaten oder mit Zeitverzug vor. Insbesondere makroökonomische Variablen wie zum Beispiel Wachstum, Inflation, Konsum oder Arbeitslosigkeit stehen erst mit zeitlichem Verzug, also als Spätindikatoren zur Verfügung.[3] Es können demnach erst ex-post Veränderungen beziehungsweise vergangene Entwicklungen beobachtet werden. Zur Prognose zukünftiger Entwicklungen werden Forecasts über vorauslaufende Indikatoren wie zum Beispiel Börsenkurs, Bauaktivitäten oder Kredite genutzt.[4]

Um Verzögerungen auszugleichen, werden, etwa im Bereich des Konsumentenverhaltens, aufwendige Befragungen durchgeführt. Diese geben teilweise einen Ex-ante-Eindruck der Verhaltensentwicklung, sind dabei allerdings sehr aufwendig.[5]

Strukturierte, quantitative Daten bilden nur einen kleinen Teil der Realität und der tatsächlich zur Verfügung stehenden Daten ab. Durch die allgemeine Verfügbarkeit des Internets wird davon ausgegangen, dass weltweit bis zu 90 % der generierten Daten in unstrukturierter Form vorliegen.[6] Diese Daten umfassen Text- und Bilddaten aus den verschiedensten Zusammenhängen. Dabei sind Text- und Bilddaten nicht nur für makroökonomische Prognosen relevant, sondern fallen auch in politischen, gesellschaftlichen und unternehmerischen

---

[3] Iyetomi et al. (2020).

[4] Conference Board.

[5] OECD (2017); GfK.

[6] Einav/Levin (2014).

Kontexten in großen Mengen an. Big Data stellt eine große Chance für die Forschung und politische Entscheidungsträger dar, stellt sie aber auch vor die Herausforderung, die Daten aufwendig zu analysieren, Datenschutzrichtlinien zu beachten und eine Reproduzierbarkeit der Ergebnisse sicherzustellen.

Textdaten sind insbesondere auch in der ökonomischen Forschung von Bedeutung. Die Daten werden durch verschiedene Akteure in gesellschaftlichen, sozialen und geschäftlichen Beziehungen generiert. Historische Daten sind vornehmlich offizieller oder redaktioneller Natur, darunter fallen offizielle Reden, journalistische Beiträge, Pressemitteilungen oder Interviews. Durch Soziale Netzwerke, persönliche Blogs oder Websites und Kommentarspalten bilden informelle Informationen einen wachsenden Anteil an der gesamten Datenbasis. Zusätzlich haben sich die Menge der Informationen über einzelne Individuen und deren Aussagekraft für ökonomische und (wirtschafts-)politische Kontexte erhöht.[7]

Zusammengefasst bestehen drei elementare Vorteile der Verwendung textbasierter Daten:

- Die Daten sind ad hoc verfügbar. Es besteht kein Zeitverzug zwischen der Generierung der Daten und ihrer Verwertbarkeit.[8]
- Soziale Medien wie Twitter und Facebook oder Google-Anfragen stellen oft Meinungsäußerungen bewusster oder unbewusster Natur dar.[9]
- Textbasierte Daten sind kostengünstiger als Surveys.

Die zeitgenaue Verfügbarkeit der Daten ist im Vergleich mit traditionellen Statistiken als einer der größten Vorteile zu betrachten. Während für die Verarbeitung und Berechnung von Indizes Daten aus verschiedenen Quellen verarbeitet werden müssen und zum Teil erst verspätet zur Verfügung gestellt werden, sind Textdaten mit ihrer Veröffentlichung rascher verfügbar. Je nach Analysemethode können die Daten damit zeitnah Aufschluss über Veränderungen der Meinung oder der ökonomischen Aktivität liefern. Zusätzlich sind durch Mitteilungen in den Sozialen Medien unmittelbare Äußerungen der Akteure verfügbar. Über unmittelbare Meinungsäußerungen können Entwicklungen nicht nur zeitnah nachverfolgt, sondern auch vorhergesagt werden. Intensiv hat sich die Forschung unter anderem zur Entwicklung des Finanzmarkts nach Äußerun-

---

7 Hu/Liu (2012).

8 Ebenda; Sakaki/Okazaki/Matsuo (2010).

9 Ravi/Ravi (2015).

gen der Zentralbanken entwickelt und kommt zu sehr guten Ergebnissen. Der Forschungsstrang nutzt Sentiments, um zum Beispiel Entwicklungen auf dem Aktienmarkt zu prognostizieren.[10] Diese Methodik kann entsprechend auch auf Mitteilungen von Konsumenten/Parteien/Wählern angewandt werden, um Verhalten und Entwicklungen vorauszusagen. Die Analysemethoden erlauben zusätzlich die Analyse impliziter Inhalte wie zum Beispiel psychologischen Motiven oder Sentiments. Durch die Nutzung impliziter Inhalte werden die unmittelbaren Äußerungen von in-(offiziellen) Akteuren noch besser interpretierbar.

Die Verwendung von Textdaten erfordert umfangreiche Vorbereitungen der Dokumente, die Vorteile der Textdaten müssen den Kosten und dem Aufwand der Auswertung gegenübergestellt werden. Als Dokumente/Textabschnitte zählen dabei einzelne abgeschlossene Texteinheiten, die für die spätere Analysemethode verwendet werden. Je nach verwandter Methodik können dies gesamte Dokumente, Sätze oder Wortgruppen sein. Die Vorbereitung der Textdaten ist wesentlich umfangreicher und zeitaufwendiger als die Vorbereitung von Daten im Rahmen offizieller Statistiken.[11]

Ein weiterer kritischer Aspekt der computerbasierten Textanalyse ist die Reproduzierbarkeit der Ergebnisse. Computerbasierte Textanalyse greift auf Algorithmen mit festgelegten Parametern zurück oder kodiert die Daten innerhalb von überwachten Ansätzen. Die entsprechenden Parameter müssen klar festgehalten und dargestellt werden, um eine Reproduktion zu ermöglichen. Für die händische Kodierung sind klare Richtlinien zur Validität und zur Reliabilität festgeschrieben, unüberwachte Verfahren erfordere eine ähnliche Vorgehensweise. Die Handkodierung eines Subsets der Dokumente kann zum Beispiel dazu beitragen, die Ergebnisse zu validieren.

Es gibt eine wachsende Zahl von Algorithmen zur Auswertung unstrukturierter Textdaten. Eine Erfassung aller Algorithmen ist nicht möglich, sie verteilen sich auf verschiedene Fachbereiche. Die Algorithmen koexistieren und werden je nach Anwendung und Fachbereich eingesetzt. Die Menge der Algorithmen wächst mit den Anwendungsfeldern. Grundsätzlich wird zwischen überwachten und unüberwachten Analysemethoden unterschieden, hinzu kommen Methoden wie Regressionsanalysen, Inhaltsanalysen oder Natural Language Processing (NLP).

---

[10] Tetlock (2007); Bollen/Mao/Zeng (2011).

[11] Gentzkow/Kelly/Taddy (2019).

Da in der ökonomischen Literatur die letztgenannten Methoden nur geringe Anwendung finden, sind sie für den vorliegenden Artikel von untergeordneter Relevanz. Quinn et al.[12] haben eine Liste der Kosten und der benötigten Datengrundlagen erstellt. Mit Ausnahme des Lesens von Texten und des automatisierten Topic Modelings setzen alle Methoden eine sehr gute Kenntnis der zugrundeliegenden Texte voraus. Die Kosten der Präanalysephase sind somit hoch. Die Kodierung der Texte beziehungsweise die Erstellung eines passgenauen Wörterbuchs oder die vollständig händische Kodierung erfordern die Bearbeitung eines Anteils oder des gesamten Textbestands. In Abbildung 1 werden die wichtigsten Algorithmenklassen für die automatisierte Textanalyse dargestellt. Einige ausgewählte Algorithmen werden in der Folge genauer vorgestellt.[13]

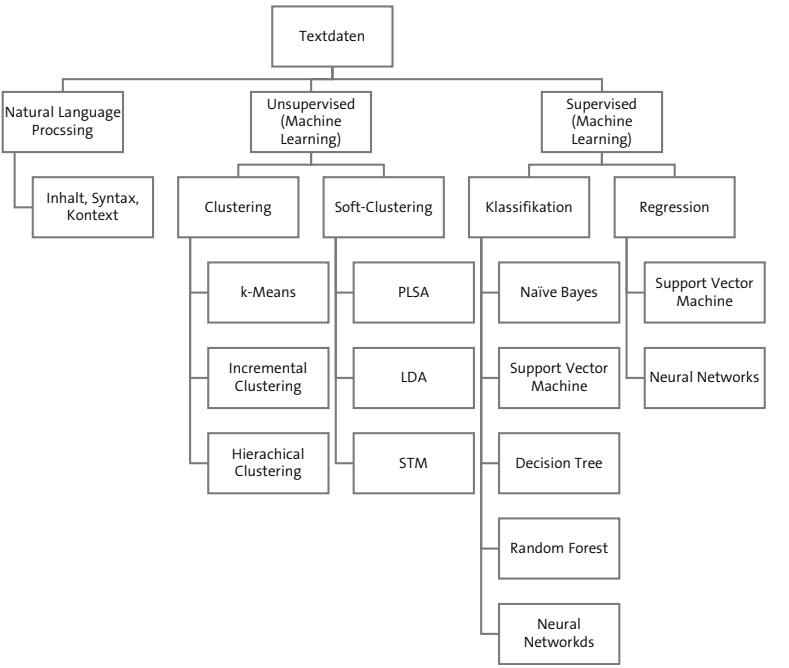

Abb. 1: Schematische Darstellung automatisierten Text Minings
Quelle: eigene Darstellung.

[12] Quinn et al. (2010).

[13] Für eine umfassende Darstellung bieten sich zum Beispiel Murphy (2012), Bishop (2016) und Gentzkow/Kelly/Taddy (2019) an.

Merkmalsextraktion

Für die Auswertung textbasierter Daten ist eine Reduktion der Dimension notwendig. Hierfür werden durch die Filterung unwichtiger Elemente charakteristische Merkmale herausgefunden. Die Extraktion von Merkmalen (*feature extraction*) wird zum Beispiel durch Löschen der Interpunktion erzielt. Zusätzlich werden häufig sogenannte Stop Words aus dem Datensatz entfernt. Stop Words sind üblicherweise häufig genutzte Wörter einer Sprache, wie zum Beispiel die Füllwörter „und", „weil" oder „da". Diese Wörter haben für die Extraktion des Inhaltes kaum Bedeutung, in spezifischen Kontexten kommen weitere Wörter hinzu. Dies können etwa Wörter sein, die zwar nur einer der zu kategorisierenden Gruppen zuzuordnen sind, jedoch nicht zur Beantwortung der Forschungsfrage dienlich sind. Beispielsweise können Politikernamen spezifisch für eine Partei sein, für die Frage des Parteienfokus sind sie jedoch nicht dienlich. Die Nutzung von Stop Words muss für jedes Studiendesign individuell geklärt werden.[14]

Eine Alternative ist die Verwendung von *term frequency – inverse document frequency-(tf-idf-)*Verfahren. Damit gemeint ist die Häufigkeit eines Wortes in einem Dokument, multipliziert mit dem logarithmierten Anteil an Dokumenten, die das Wort enthalten. Im Ergebnis wird ein niedriger tf-idf-Wert für sehr häufige Wörter in allen Dokumenten und sehr seltene Wörter in einem spezifischen Dokument ausgegeben. Anders formuliert werden als Features jene Wörter behalten, die häufig in einem einzelnen Dokument vorkommen, aber selten in anderen Dokumenten.[15]

Als weiterer Schritt wird ein Stemming, das heißt eine Verkürzung der Wörter auf ihren Wortstamm, durchgeführt. Alle beschriebenen Schritte reduzieren die Menge der zu analysierenden Daten. Zwei Aspekte profitieren von den Vorbereitungen des Datensatzes: Zum einen werden die Ergebnisse leichter interpretierbar und zum anderen sinkt die notwendige Rechenleistung.

---

[14] Nenkova/McKeown (2012).

[15] Ebenda.

Überwachte Verfahren

Überwachte oder Klassifizierungsmethoden bringen auf Grundlage eines Trainingsdatensets dem Modell Kriterien bei, die später für unbekannte Daten genutzt werden können, um Vorhersagen über Inhalte oder Verhaltensweisen zu treffen. Dem Trainingsdatenset werden dabei Kategorisierungen zugeordnet, der Algorithmus lernt daraufhin die Charakteristika der Kategorien und ordnet sie den Kategorien zu. In einfachen Ansätzen werden hierzu etwa Kritiken mit einer quantifizierbaren Bewertung als Kategorisierung genutzt. Es ist demnach keine weitere Kodierung durch Experten notwendig (zum Beispiel Produktbewertungen)[16].

Überwachte Methoden gelten insbesondere bei tiefgehenden inhaltlichen sowie psychologischen Textanalysen als präziser. Die Aussagen können vor dem automatisierten Verfahren ähnlich wie bei händischen Verfahren kodiert werden. Durch die Vorkodierung anhand theoriegeleiteter Kriterien verringert sich die Wahrscheinlichkeit einer fehlerhaften Zuordnung. Bei der Auswahl der Algorithmen sind zwei Aspekte zu beachten: Zum einen ist ein sogenanntes Overfitting des Modells möglich. Wenn der Algorithmus das Trainingsdatenset zu kleinteilig analysiert hat, kann es passieren, dass das Ergebnis auf der Basis neuer Daten sich schlechter darstellt. Die detaillierten Charakteristika des Trainingsdatensets müssen nicht im neuen Datensatz vorkommen und können damit nicht vom Algorithmus erfasst werden.[17] Zum anderen werden für überwachte Methoden mehr Daten benötigt, das heißt, es werden ein ausreichend großes, passgenaues Trainingsdatenset und die eigentlich zu untersuchenden Daten benötigt.[18]

Die Trainingsmethode basiert bei den meisten Studien auf Naive Bayes, Entscheidungsbäumen, Support Vector Machines oder regressionsbasierten Klassifizierungen.[19]

---

[16] Liu/Zhang (2012).

[17] Murphy (2012).

[18] Aggarwal (2012).

[19] Für einen umfangreichen Überblick bieten sich Gentzkow/Kelly/Taddy (2019) und Aggarwal/Zhai (2012a) an.

Unüberwachte Verfahren

In Bezug auf unüberwachte Verfahren wird zwischen klassischen Clustering-Ansätzen und dem sogenanntem Soft Clustering unterschieden. Anders als im Falle der überwachten Verfahren benötigen die Algorithmen keine vorkodierten Trainingsdaten. Aus den zur Verfügung stehenden Daten werden Cluster generiert, welche die Struktur der Texte widerspiegeln. Mit dem Ansatz können Informationen aus großen Mengen textbasierter Daten extrahiert werden. Grundsätzlich finden zwei Ansätze Verwendung: das klassisches Clustering und das Soft Clustering.

Beim klassischen Clustering wird die Anzahl der Cluster k definiert und die k Punkte werden randomisiert als Zentrum des Clusters festgelegt. Alle Bestandteile werden dem nächstliegenden Cluster-Zentrum zugeordnet (e*uclidean distance*). Anschließend wird der Mittelwert jedes Clusters berechnet. Dieser bildet sodann das Zentrum der Cluster. Die Methode ist schnell und effizient, kann jedoch etwa durch die präzise Wahl des initialen Zentrums erheblich verbessert werden. Des Weiteren ist die Anzahl der k Cluster nicht immer leicht zu bestimmen.[20]

Im Gegensatz zum deterministischen Ansatz des einfachen Clustering liegt dem probabilistischen Ansatz eine Dichteverteilung zugrunde. Soft-Clustering-Methoden (oder „Topic-Modell“) errechnen die Wahrscheinlichkeit, mit der ein Dokument einem Themencluster zugeordnet werden kann. Ein erster Versuch, probabilistische Aspekte in die Analyse zu integrieren, war die „probabilistic Latent Sementic Analysis“ (pLSA).[21] Dieser Methode liegt kein generatives Modell zugrunde, doch die Latent Dirichlet Allocation (LDA) eliminiert dieses Problem. Der Auswertung in diesem Beitrag liegt die LDA zugrunde.[22] LDA nimmt an, dass sich jedes Dokument aus einer Mischung von Themen zusammensetzt. Jedes Thema besteht aus einer Verteilung des über den Datensatz vorhandenen Vokabulars. Der probabilistische, generative Prozess ist wie folgt definiert:

---

[20] Aggarwal/Zhai (2012b).

[21] Hofmann (2013).

[22] Blei/Ng/Jordan (2003).

Für jedes Thema:

$$\beta k \sim DirV\,(\eta) \tag{1}$$

Für jedes Dokument:

$$\theta_d \sim Dir(\alpha) \tag{2}$$

Für jedes Wort:

$$\begin{aligned} Z_{d,n} &\sim \text{Mult}(\theta_d), \text{with } Z_{d,n} \in \{1, \ldots, K\} \\ W_{d,n} &\sim \text{Mult}\left(\beta_{Z_{d,n}}\right), \text{with } W_{d,n} \in \{1, \ldots, V\} \end{aligned} \tag{3}$$

Die LDA verwendet zwei Dirichlet-Zufallsvariablen: erstens die Themen β_k, welche eine Verteilung über das Vokabular V (mit den Themen 1 bis K) sind, und zweitens die Themenzuordnung pro Dokument θ_d. Die Dirichlet-Verteilung verwendet feste Parameter η und α, die die Themenverteilung über Wörter beziehungsweise die Themenverteilung pro Dokument über Wörter beeinflussen. Die Themenzuweisung pro Wort $Z_{d,n}$ ist die dritte verborgene Variable, welche die zugrunde liegende latente Struktur des Korpus definiert. $W_{d,n}$ ist die beobachtete Variable, die Informationen über das im Korpus verwendete Vokabular enthält.[23] In anderen Worten liefert der Algorithmus eine Zusammenfassung der im Korpus enthaltenen Themen, bestehend aus dem spezifischen Vokabular und der Wahrscheinlichkeit, dass ein Thema einem Dokument zugeordnet werden kann. Seit den ersten Topic-Modells ist eine Vielzahl neuer Modelle mit unterschiedlicher Schwerpunktsetzung, zum Beispiel Themenkorrelation[24], strukturelle Topic-Modelle (STM)[25] oder dynamische Topic-Modelle[26], entstanden.

Politische und gesellschaftliche Texte

Direkte politische oder gesellschaftliche Kommunikation unterscheidet sich durch zwei Merkmale von formalen oder journalistischen Texten. Zum einen beinhalten die schnell verfassten, nicht redaktionell bearbeiteten Kommentare

---

[23] Blei/Lafferty (2009).

[24] Blei/Lafferty (2007).

[25] Roberts et al. (2014).

[26] Blei/Lafferty (2006).

ein vielseitiges Vokabular. Zum anderen steht zwischen dem Verfasser der Texte und den Lesenden kein Interpretationsschritt durch Erläuterungen, Zusammenfassungen oder journalistische Aufarbeitung.[27]

Im Allgemeinen unterscheiden sich Kommentare von politischen Akteur:innen und User:innen in vielerlei Hinsicht von anderen für inhaltsanalytische Ansätze genutzten Texten. Die Texte beinhalten eine wenig formalisierte Sprache mit einem großen und diversen Vokabular, zusätzlich erschweren Rechtschreibungsfehler oder Abkürzungen die Analysen.[28] Des Weiteren sind die Themen vielschichtig, das heißt, es wird sowohl innerhalb einer Nachricht als auch in der Grundgesamtheit die Bandbreite der gesellschaftlichen Kommunikation abgedeckt. Es gibt keine Kodierung dieser Themenbandbreite, sie wird in unkodierter und unstrukturierter Form veröffentlicht. Diese Charakteristika erschweren eine überwachte Analyse dieser Kommunikation. Der Diskurs umfasst eine stetig steigende Menge an Nachrichten. Zudem verändert sich der thematische Schwerpunkt im Zeitverlauf. Während manche Themen an Bedeutung verlieren oder verschwinden, werden andere Themen und damit auch ein erweitertes Vokabular relevant. Sichtbar ist die Wandlung von Themenschwerpunkten ex post, zum Beispiel in handkodierten Parteiprogrammen.[29]

Das Themenspektrum deckt zeitlich und gesellschaftlich die vorhandene Bandbreite ab. Diese Vielseitigkeit legt die Relevanz der Texte nahe. Es werden bevölkerungsrelevante Themen in nicht interpretierter oder gefilterter Form analysiert. Es können Themen erkannt werden, die in Surveys oder anderweitigen theoriegeleiteten Analysemethoden noch nicht abgefragt oder erkannt werden können. Die datengetriebene, explorative Methodik hat demgegenüber den Vorteil, dass sie die vorhandene Kommunikation in umfassender Form betrachtet und auswertet.

Anwendungsbereiche können unter anderem im politischen Kontext (Parteienkommunikation) und im Konsument:innenverhalten zu finden sein. Im politischen Kontext können sowohl spezifische wirtschaftspolitische Themen erkannt werden, als auch die Gesamtkommunikation und ihre Wirkung auf Wähler:innen erfasst werden. Konsument:innen / die Bevölkerung stellen ihre Ansichten, Gedanken und Erfahrungen auf diversen Sozialen Medien dar.

---

[27] Stieglitz/Dang-Xuan (2013); Hong/Nadler (2011).

[28] Hu/Liu (2012).

[29] Volkens et al. (2020).

Vorteile Sozialer Medien

Die Verbreitung und massenhafte Nutzung Sozialer Medien begann Mitte der 2000er-Jahre. Die Nutzung der verschiedenen Sozialen Medien steigt seitdem stetig an. Während sie zu Beginn auf regionale Verknüpfungen und vor allem ein junges Publikum begrenzt war, sind mittlerweile alle Altersgruppen, Geschlechter und gesellschaftlichen Schichten auf den Sozialen Medien miteinander verbunden.[30] Die flächendeckende Verbreitung von Smartphones ermöglicht die orts- und zeitunabhängige Nutzung durch die Konsument:innen/Wähler:innen.[31] Die Verbreitung der verschiedenen Medien unterscheidet sich je nach dem betrachteten Land zum Teil deutlich. Während in den Vereinigten Staaten der Großteil der Bevölkerung mindestens eine (teilweise) öffentliche Plattform nutzt, ist etwa in Deutschland die Nutzung nach wie vor in den jüngeren Altersgruppen weiter verbreitet. Zudem unterscheidet sich die Verbreitung der verschiedenen Plattformen. In Deutschland ist Facebook die am stärksten genutzte Plattform, auf den weiteren Rängen finden sich Twitter und stetig wachsend Instagram.[32] In den Vereinigten Staaten hingegen ist Twitter die meistgenutzte Plattform.[33] Ein weiterer Unterschied im internationalen Vergleich zeigt sich in der Nutzung der Plattformen. Während in den USA schnell eine kommerzielle und politische Nutzung vorangeschritten ist, ist in vielen europäischen Staaten die kommerzialisierte Nutzung der Plattformen ein vergleichsweise neues Phänomen.

Soziale Medien, als relativ neues Medium, stellen neue Herausforderungen an Auswertungsmethoden. Sie können jedoch durchaus in Zeitreihen angewandt werden.[34] Im politischen Kontext ist eine Auswertung seit Beginn der 2010er-Jahre möglich. In den USA werden seit Längerem über Mikrotargeting Wähler:innen gesucht, welche für Politiker:innen ansprechbar sind. In Europa sind derartige Auswertungen nicht mit den Datenschutzbestimmungen vereinbar. Die Anwendung von Mikrotargeting ist bereits länger möglich, da es sich um Einzelpersonen als Untersuchungsobjekte handelt.[35] Eine systematische Analyse zum Beispiel der politischen Kommunikation durch Entscheidungsträger:innen hingegen, bedarf der Aktivität von Parteien und Politiker:innen auf

---

30 Java et al. (2007).

31 Beisch/Schäfer (2020).

32 Ebenda.

33 Hu/Liu (2012).

34 Hu/Liu (2012).

35 Zuiderveen Borgesius et al. (2018); Papakyriakopoulos et al. (2017).

sozialen Kanälen. Die Notwendigkeit wurde von Parteien und Politiker:innen erst stückweise erkannt. Im deutschen Kontext ist die Wahlperiode 2013 bis 2017 die erste flächendeckend auswertbare Periode. Für die Bundestagswahl 2013 ist die kurzfristige Wahlkampfstrategie bereits analysierbar. Für Konsument:innen gelten ähnlich strenge Datenschutzbestimmungen zur Anonymisierung von Einzelpersonen. Die Datennotwendigkeiten für die Analyse von Konsumentscheidungen entsprechen jenen für politisches Mikrotargeting

Die Vorteile der Sozialen Medien liegen auf der Hand. Auf den Sozialen Medien kommunizieren Einzelpersonen und Unternehmen oder Politiker:innen ohne Zeitverzug. Abgesehen von redaktionell verfassten Beiträgen spiegeln die Beiträge die aktuellen Gedanken oder Bedürfnisse ohne Interpretationsschritte oder kognitive Verzerrung wider. Zudem können Konsument:innen/Produzent:innen oder Wähler:innen/Parteien direkt miteinander in Kontakt treten und auf die jeweils andere Seite reagieren. Die stetige Produktion neuer Daten erlaubt eine tagesaktuelle Auswertung von Informationen. Zeitliche Restriktionen und Schlaglichter zu einem einzelnen Zeitpunkt können damit vermieden werden.

Einerseits haben die Sozialen Medien klare Vorteile, doch andererseits gibt es einige Faktoren, die die Analyse behindern. Zum einen sind die Beiträge zum Teil sehr kurz gehalten, zum Beispiel auf Twitter. Zum anderen sind häufig viele verschiedene Themen innerhalb eines Posts zusammengefasst. Eine klare Trennung beziehungsweise Zuordnung der Anteile eines Posts stellt sich teilweise als schwierig dar. Daneben erschwert die Formulierung der Nachrichten die Analyse, die Problematik ergibt sich aus der knappen und schnell verfassten Natur der Beiträge. Dabei werden unterschiedliche Abkürzungen für Wörter verwendet, Umgangssprache genutzt und es treten häufiger Rechtschreibungsfehler auf. Diese Punkte erschweren die systematische Auswertung und erfordern ein ausgeprägtes *pre-processing* der Rohdaten.[36] Abgesehen von offiziellen Accounts besteht die Möglichkeit, dass Personen sich auf Sozialen Medien nicht mit ihrer wahren Identität oder ihrem wahren Standort anmelden. Problematisch sind dabei insbesondere Bots, welche eine große Anzahl an Beiträgen automatisiert verfassen. Um diese Accounts zu filtern und nicht fälschlicherweise in ungewolltem Kontext in Analysen zu integrieren, lassen sich Algorithmen

[36] Hu/Liu (2012).

einsetzen. Bei einem Bewusstsein der Problematik gegenüber lässt sich somit eine fälschliche Integration verhindern.[37]

Anwendungen Sozialer Medien

Die Anwendungen zur textbasierten Auswertung sozialer Medien sind sehr vielschichtig. Die Möglichkeiten und Anwendungen erwachsen aus der hohen Nutzerzahl, der globalen Verfügbarkeit und dem vielfältigen Nutzerkreis. Während für die Analyse und Vorhersage der Börsen/Aktienkurse unter anderem öffentliche Profile und Statements der Zentralbanken und/oder Politiker:innen genutzt werden,[38] stehen bei „social emotion"- und „opinion mining"-Studien private Statements der gesamten Nutzerbasis im Mittelpunkt.[39] Aus der Betrachtung sozialer Emotionen lassen sich über automatisierte Verfahren zusätzlich Sentiment-Lexika entwickeln.[40] Darüber hinaus bieten sich die Sozialen Medien für die Erfassung von Events und Wendepunkten an.[41] Die Betrachtung des Konsument:innenverhaltens (siehe dazu nachfolgend den Beitrag in diesem Band) nutzt zumeist Sentiments und die gesamte Nutzerbasis eines definierten Georaums.[42] Ähnlich wie in diesem Beitrag werden die Sozialen Medien für die Analyse politischer Themen und des Parteiwettbewerbs genutzt.[43]

Eine konkrete Anwendung im Bereich der politischen Kommunikation analysiert das Kommunikationsverhalten deutscher Parteien in der Legislaturperiode 2013 bis 2017. Es wurden die offiziellen Facebook-Seiten der Bundesparteien sowie der jeweiligen Parteivorsitzenden und des Generalsekretärs genutzt (siehe Tabelle 1). Im Zeitraum von Januar 2014 bis Dezember 2017 konnte so die gesamte Wahlperiode inklusive der Koalitionsverhandlungen abgebildet werden. Für die gewählte Periode konnten zwischen 30 und 40 Themen pro Parteien identifiziert werden. Die Auswertung der Facebook-Beiträge erfolgte über

---

37 Zum Beispiel Cai/Li/Zengi (2017).

38 Nguyen/Shirai (2015); Besimi et al. (2019); Bollen/Mao/Zeng (2011); Oliveira/Cortez/Areal (2017).

39 Rao et al. (2014); Bao et al. (2009); Vamshi/Pandey/Siva (2018).

40 Deng et al. (2019); Xie/Li (2012).

41 Xue et al. (2020); Qian et al. (2016).

42 Homburg/Ehm/Artz (2015); Daas/Puts (2014); Pekar/Binner (2017).

43 Joshi/Bhattacharyya/Carman; Takikawa/Nagayoshi (2017); Oliveira et al. (2018); Antonakaki et al. (2017).

Latent Dirichlet Allocation[44], eine unüberwachte Soft-Clustering-Methode. Die Daten wurden auf monatlicher Basis akkumuliert, um eine Aussage über das Kommunikationsverhalten, die Schwerpunkte und Veränderungen zu treffen. Die Parteienkommunikation wird separat ausgewertet, mit dem Ziel kleinere Themen oder Themendifferenzierungen der Parteien zu generieren.

Tab. 1: Parteien und Politker:innen im Datensatz

PARTEI		
AfD	Frauke Petry	Jörg Meuthen
CSU	Horst Seehofer	Andreas Scheuer
CDU	Angela Merkel	Peter Tauber
FDP	Christian Lindner	Nicola Beer
SPD	Sigmar Gabriel	Hubertus Heil
BÜNDNIS 90 / DIE GRÜNEN	Cem Özdemir	Simone Peter
DIE LINKE	Bernd Riexinger	Katja Kipping

Der Algorithmus gibt die prozentualen Anteile der einzelnen Themen innerhalb eines Beitrags und die wichtigsten Wörter eines Clusters aus. Zusätzlich wurde der Sentiment Score erstellt. Der Sentiment Score wird auf einer Skala von *–1* bis *+1*, auf der Grundlage eines Wörterbuchs errechnet. Für das Wörterbuch wurden von Experten Wörter hinsichtlich ihrer positiven oder negativen Konnotation kodiert.[45] Der Sentiment Score für ein Dokument i mit den Sentiment-Wortwerten w und dem Gesamtvokabular v errechnet sich wie in der folgenden Formel dargestellt.

$$S_i = \frac{\sum w_{pos,i} - \sum w_{neg,i}}{\sum v_i}$$

Im Ergebnis konnten signifikante und den Parteien entsprechende Ergebnisse erzielt werden. Das Profil der Parteien, insbesondere der kleinen Parteien, entspricht den Schwerpunktsetzungen in den Wahlprogrammen. Der Vorteil anderen Ansätzen gegenüber ist die monatliche Auswertbarkeit und darauf aufbauend die Möglichkeit von Analysen der Reaktionen von Wähler:innen. Beispielhaft werden an dieser Stelle die Ergebnisse zur Arbeitsmarktpolitik, zur Sozial- und Familienpolitik, zur wirtschaftlichen Entwicklung Deutschlands sowie zu außergewöhnlichen Ereignissen und deren Erkennung betrachtet.

---

[44] Blei/Ng/Jordan (2003).

[45] Remus/Quasthoff/Heyer (2010).

Die Anwendung von Latent Dirichlet Allocation (LDA) gibt als Resultat zwei Ergebnisse aus. Zum einen werden die Themencluster in Form der häufigsten Wörter ausgegeben. Zum anderen wird für jeden Text der zugehörige prozentuale Anteil an einem Thema ausgegeben. Durch die Definition der Parameter kann festgelegt werden, ob viele oder wenige Themen pro Nachricht erwartet werden. Im nachfolgenden Fall wurde, ausgehend von dem tendenziell kurzen Charakter der Facebook-Posts angenommen, dass wenige Themen pro Nachricht kommuniziert werden. Für die Interpretation ist relevant, dass ein Thema, welches in den Clustern nicht auftaucht, nicht zwangsläufig gar nicht behandelt wird. Es wird jedoch in entweder deutlich untergeordneter Relevanz oder in nicht differenziertem Vokabular kommuniziert.

Tab. 2: Codebuch – Auszug zu Sozialpolitik und Budget, Wachstum und Entwicklung

SOZIALPOLITIK

Arbeitsmarkt	Arbeitslosigkeit Löhne Arbeitslosengeld (Hartz IV) Soziale Sicherungssysteme Arbeitsrecht Gewerkschaften Start-ups
Rentenpolitik	Rentenversicherung Altersarmut
Familienpolitik	Work-Life-Balance Mütterrente Elterngeld Kinderarmut
Pflege	
Wohnraum	
Bildungspolitik	

BUDGET, WACHSTUM, ENTWICKLUNG

Budget	Schwarze Null Investitionen
Vermögen	Vermögenssteuer
Wachstum	Wachstumsfaktoren Prosperity
Freihandel	TTIP CETA
Digitalisierung	
Ländliche Entwicklung	

Die resultierenden Wortlisten müssen kodiert werden. Auf Basis eines detaillierten Codebuchs können die Listen Themen zugeordnet werden. Beispielhaft sind in Tabelle 2 die Abschnitte des Codebuchs für die Themenfelder Sozialpolitik, Budget, Wachstum und Entwicklung abgebildet. Das Codebuch wurde in einem explorativen Ansatz entwickelt, basierend auf den Resultaten der LDA-Cluster. Der Vorteil der automatisierten Textanalyse liegt im frühzeitigen Erkennen neuer Themenfelder, daher ist ein explorativer Ansatz der Themengenerierung sinnvoll.

Tab. 3: Themenliste – Arbeitsmarkt

DIE LINKE			
ARBEITSRECHT	GEWERKSCHAFTEN	LÖHNE	ARBEITSLOSIGKEIT
„arbeit“	„beschaeftigt“	„euro“	„hartz“
„gut“	„gewerkschaft“	„mindestlohn“	„sanktion“
„leiharbeit“	„gut“	„million“	„betroff“
„beschaeftigt“	„arbeit“	„jahr“	„abschaff“
„mensch“	„verdi“	„milliard“	„grundrecht“
„leb“	„unterstuetz“	„unternehm“	„jobcent“
„befrist“	„streik“	„zahl“	„sanktionsfrei“
„job“	„loehn“	„ausnahm“	„mindestsicher“
„preka“	„amazon“	„niedrig“	„bundesregier“
"gleich"	„metall“	„fordert“	„andrea“

SPD	CSU	CDU	
LOHNE/ GEWERKSCHAFTEN	ARBEITSLOSIGKEIT/ LÖHNE	LÖHNE	ARBEITSLOSIGKEIT
„mindestlohn“	„loewenstark“	„sich“	„deutschland“
„rent“	„gut“	„stark“	„gut“
„arbeit“	„best“	„deutschland“	„mensch“
„mensch“	„digital“	„inn“	„wirtschaft“
„gut“	„wirtschaft“	„wirtschaft“	„jahr“
„jahr“	„prozent“	„arbeit“	„arbeitslos“
„gesetz“	„arbeitslosenquot“	„bleibt“	„zahl“
„prozent“	„arbeitsmarkt“	„wohlstand“	„arbeit“
„arbeitnehm“	„freistaat“	„arbeitsplaetz“	„prozent“
„andrea“	„top“	„zeit“	„arbeitsmarkt“

Die Parteien diskutieren in Bezug auf die wichtigen wirtschafts- und sozialpolitischen Themenfelder die gleichen Oberthemen. Die Unterscheidung wird in der Diskurssetzung, also den präziseren Unterthemen, evident. Dies wird bei sozialpolitischen Themen deutlich. Tabelle 2 weist die zehn Wörter aus, die im Rahmen von arbeitsmarktbezogenen Themen am häufigsten verwendet wurden. Die zuvor beschriebene Unterteilung der Themen wird hier verdeutlicht. Darüber hinaus zeigt das Agenda-Setting sowohl den Unterschied zwischen Mitte-Links- und Mitte-Rechts-Parteien als auch die Konzentration auf eine kontroverse versus eine ergebnisorientierte Debattenstruktur.

Während in der Linken ein breites Spektrum von Arbeitsmarktfragen unter besonderer Berücksichtigung der sozialen Gerechtigkeit diskutiert wird, konzentriert sich die CDU auf Erfolge und gute Arbeitsmarktbedingungen im Sinne einer niedrigen Arbeitslosigkeit. Die Unterschiede zwischen Mitte-Links- und Mitte-Rechts-Parteien werden auch im Vergleich von CDU und SPD sichtbar. Beide Parteien gingen 2013 eine Koalition ein, wobei sich die SPD auf zentrale sozialdemokratische Themen wie den Mindestlohn konzentriert, der jedoch von der CDU nicht erwähnt wird. Das obige Beispiel zeigt, zu welchen sinnvollen Ergebnissen die Anwendung von Themenmodellen auf die offizielle Social-Media-Kommunikation der Parteien führt. Auffällig ist bei den weiteren Themen der sozialpolitischen Kommunikation auch die Zuordenbarkeit der Themen zu Parteien oder mindestens die Regierungszugehörigkeit oder das politische Spektrum.

So werden zum Beispiel von allen Parteien, CDU und CSU ausgenommen, die Freihandelsabkommen TTIP und CETA behandelt. Während die Linke und die Grünen den Abkommen eher ablehnend gegenüberstehen, kann aus den zehn häufigsten Wörtern bei SPD und AfD keine klare Positionierung abgeleitet werden. Die FDP kommt in den zehn häufigsten Wörtern eher zu einem positiven Urteil. Linke und Grüne diskutieren zudem ihre Kernthemen Umverteilung respektive Umweltschutz im Zusammenhang mit der wirtschaftlichen Entwicklung Deutschlands. Bei den Regierungsparteien und der FDP kommt die Digitalisierung hinzu. Während bei SPD und CDU der Fortschritt im Tenor eher positiv klingt, deutet die Kommunikation der FDP auf kritische Aspekte in Bezug auf die Digitalisierungsstrategie hin. Insgesamt zeichnen die wirtschaftspolitischen Themen, ähnlich wie die sozialpolitischen Themen, das Profil der Parteien und ihrer Ausrichtung nach.

Tab. 4: Themenliste – Budget, Wachstum und Entwicklung

	DIE LINKE		AFD		
	FREIHANDEL	VERMÖGENSTEUER	TTIP	WACHSTUM	SCHWARZE NULL
1	„ttip“	„reich“	„buerg“	„deutschland“	„euro“
2	„ceta“	„deutschland“	„ttip“	„jahr“	„geld“
3	„stopp“	„prozent“	„bargeld“	„zahl“	„jahr“
4	„polit“	„gerecht“	„freiheit“	„prozent“	„kost“
5	„konzern“	„sozial“	„lehnt“	„studi“	„milliard“
6	„freihandelsabkomm“	„reichtum“	„oeffent“	„polit“	„million“
7	„demokrati“	„vermoeg“	„kontroll“	„mensch“	„steuerzahl“
8	„aktion“	„land“	„staat“	„deutsch“	„pro“
9	„mensch“	„einkomm“	„abschaff“	„articl“	„hoeh“
10	„abkomm“	„gesellschaft“	„gross“	„wirtschaft“	„steu“

	BÜNDNIS 90 / DIE GRÜNEN		SPD		
	WACHSTUMS-FAKTOREN	FREIHANDEL	WACHSTUMS-FAKTOREN	TTIP	DIGITALISIE-RUNG
1	„gerecht“	„ttip“	„wirtschaft“	„ttip“	„digital“
2	„sozial“	„ceta“	„unternehm“	„bundesregier“	„frag“
3	„oekolog“	„umwelt“	„deutschland“	„ceta“	„zukunft“
4	„gesellschaft“	„fair“	„deutsch“	„kommission“	„digitalleb“
5	„wirtschaft“	„stopp“	„industri“	„freihandelsab-komm“	„arbeit“
6	„weltoff“	„handel“	„jahr“	„bmwi“	„digitalisier“
7	„umwelt“	„handelsabkomm“	„arbeitsplaetz“	„entscheid“	„diskuti“
8	„zentral“	„usa“	„zukunft“	„oeffent“	„gesellschaft“
9	„arbeit“	„mensch“	„stark“	„debatt“	„leb“
10	„zusammenhalt“	„konzern“	„gut“	„verhandl“	„thema“

	CSU		CDU		
	LANDLICHE ENTWICKLUNG	INVESTITIONEN	SCHWARZE NULL	DIGITALISIE-RUNG	LANDLICHE ENTWICKLUNG
1	„gross“	„euro“	„jahr“	„digital“	„gefuehrt“
2	„polit“	„bund“	„schuld“	„digitalisier“	„bundesregier“
3	„zeit“	„jahr“	„haushalt“	„thema“	„euro“
4	„gemeinsam“	„laend“	„bundestag“	„diskutiert“	„gut“
5	„wichtig“	„milliard“	„schwarz“	„chanc“	„infos“
6	„herausforder“	„zukunft“	„forschung“	„wirtschaft“	„milliard“
7	„bleib“	„schuld“	„bildung“	„frag“	„pfleg“
8	„gut“	„bundesla“	„solid“	„wandel“	„nachricht“
9	„dialog“	„finanz“	„null“	„mittelpunkt“	„internet“
10	„gespraech“	„entlast“	„betont“	„diskuti“	„bund“

	FDP			
	BUDGET	TTIP	DIGITALISIERUNG	START-UPS
1	„steu“	„wirtschaft“	„deutschland“	„bess“
2	„entlast“	„deutsch“	„digitalisier“	„deutschland“
3	„soli“	„ttip“	„digital“	„arbeit“
4	„wissing“	„arbeitsplaetz“	„chanc“	„gruend“
5	„euro“	„gut“	„bildung“	„ide“
6	„politikdierech-nenkann“	„gerad“	„braucht“	„unternehm“
7	„kalt“	„unternehm“	„endlich“	„buerokrati“
8	„progression“	„handwerk“	„brauch“	„fordert“
9	„buerg“	„mindestlohn“	„infrastruktur“	„gerad“
10	„geld“	„wohlstand“	„agenda“	„braucht“

Nachdem die Themencluster codiert sind, kann der prozentuale Anteil der Kommunikation an den Themenclustern ermittelt werden. Um eine aussagekräftige Verteilung zu erzielen, werden hierfür die Oberthemen Sozialpolitik und Budget, Wachstum und Entwicklung in der Darstellung kumuliert. Für eine präzisere Auswertung und konkrete Anwendungen im Bereich der Wählerreaktion können die detaillierteren Themenfelder wie Arbeitsmarkt oder Familienpolitik genutzt werden.

In Abbildung 2 wird die kumulierte Kommunikation der Bundesparteien zu sozialpolitischen Themen (zur Ausrichtung der Debatte siehe Tabelle 4) dargestellt. Exemplarisch zeigen die sozialpolitischen Themen, dass die Parteien bezüglich der kommunizierten Themen starke Schwerpunkte setzen. Über den gesamten Zeitraum hinweg bestreitet die Linke das Thema in ausgeprägter Form. Die regierungsbeteiligten Parteien SPD und CDU folgen mit einer gleichbleibend hohen Beteiligung am Themenspektrum, verweisen dabei jedoch hauptsächlich auf arbeitsmarktbezogene Themen und im Fall der CDU auf Erfolge der Regierung.

Am rechten Rand des politischen Spektrums wird der Akzent auf migrationsbezogene Themen gesetzt und als Wahlkampfthema im September 2017 weiter intensiviert. Abbildung 3 zeigt zwei Phänomene deutlich auf. Zum einen ist ein Anstieg der Kommunikation, ausgelöst durch ein externes Ereignis, zu sehen (Anstieg der Zuwanderung Geflüchteter 2015). Zum anderen kann, ähnlich wie in der sozialpolitischen Kommunikation, die Vereinnahmung eines Themas durch einzelne Parteien beobachtet werden: Die Kommunikation der CSU und der AfD über zuwanderungsrelevanten Themen nimmt zu.

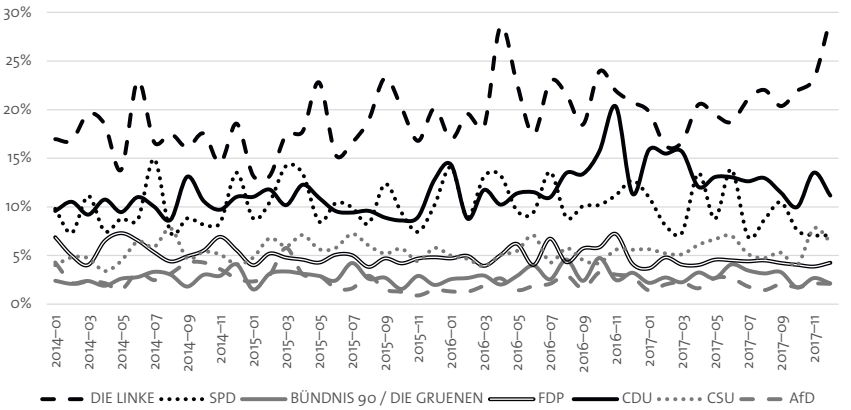

Abb. 2: Monatlich kumulierte Kommunikation – Sozialpolitik
Quelle: eigene Darstellung.

Ebenfalls zu beobachten ist in der sozialpolitischen Kommunikation, dass die Vereinnahmung eines Themas mit einer problemorientierten Darstellung der Thematik einhergeht. Während insbesondere das linke Parteienspektrum die Integration und die Fluchtursachen thematisiert, werden von CSU und AfD verstärkt die Konsequenzen und problematischen Aspekte der Zuwanderung dargestellt. Des Weiteren ist bei der AfD zu beobachten, dass, ähnlich wie bei der Partei DIE LINKE, das Thema in vielen Varianten kommuniziert wird. Als interessant zu vermerken ist zudem, dass der Wechsel an der AfD-Parteispitze (5. Juli 2015) und damit die Neuausrichtung deutlich zu erkennen ist, Euro-kritische Themen werden zu migrationskritischen Themen. Dies ist umso bezeichnender, als über den gesamten Zeitraum dieselben Personen in die Analyse einbezogen waren.

Die Analyse der Relevanz einzelner Themen zeigt drei zu erwartende und wichtige Phänomene der politischen Kommunikation. Erstens: Kleine und Oppositionsparteien verfolgen die Strategie, einzelne Themen für sich zu vereinnahmen und den Diskurs zu bestimmen.[46] Zweitens können externe Effekte mit politischer Durchschlagskraft in der öffentlichen Kommunikation zuverlässig erkannt werden. Drittens werden Strategiewechsel, innerparteiliche Auseinandersetzungen und wahlkampfstrategische Überlegungen evident.

[46] Die Ergebnisse werden von verschiedenen Studien zur strategischen Ausrichtung von Parteien unterstützt (siehe zum Beispiel Dragu/Fan [2016]; Glazer/Lohmann [1999]).

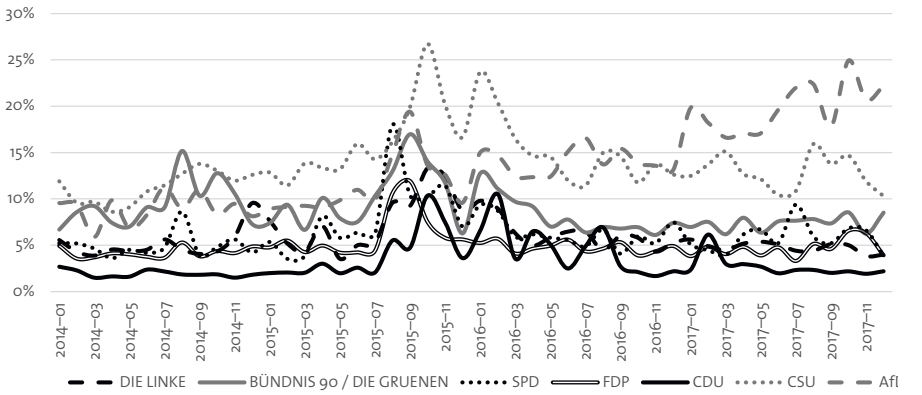

Abb. 3: Monatlich kumulierte Kommunikation – Migration
Quelle: eigene Darstellung.

Um eine empirische Auswertung der Daten zu ermöglichen, muss der Tenor der Parteienkommunikation quantifiziert werden. Verschiedene Ansätze können hierfür verwandt werden. Die einfachste Möglichkeit ist die Verwendung eines wörterbuchbasierten Verfahrens zur Berechnung der Sentiments einzelner Wörter. Ein Problem dieser Methodik liegt in der Ausgestaltung des Wörterbuchs. Gerade Wortneuschöpfungen mit einem stark positiven oder negativen Charakter sind in den Wörterbüchern selten enthalten. Einen Eindruck von der Kommunikation spiegelt die Methode jedoch gut wider. Exemplarisch sind an dieser Stelle die Oppositionsparteien (Die Linke, Bündnis 90 / die Grünen, FDP, AfD) sowie die Regierungsparteien dargestellt. Oppositionsparteien fallen durch ihre weitestgehend negativen Sentiment-Werte auf (Abbildung 4). Die Regierungsparteien hingegen sind überwiegend im positiven Wertebereich zu finden. Dies deutet auf eine sehr unterschiedliche Darstellung der Lage in verschiedenen Themenbereichen und die Hervorhebung der Erfolge auf Seiten der Regierungsparteien hin.

Weitere Ansätze zur Analyse beinhalten maschinelle Ansätze zum NLP oder psychologische Ansätze zur Generierung eines psychologischen Motivs. Diese Methodik wird im letzten Beitrag in diesem Band genauer betrachtet.

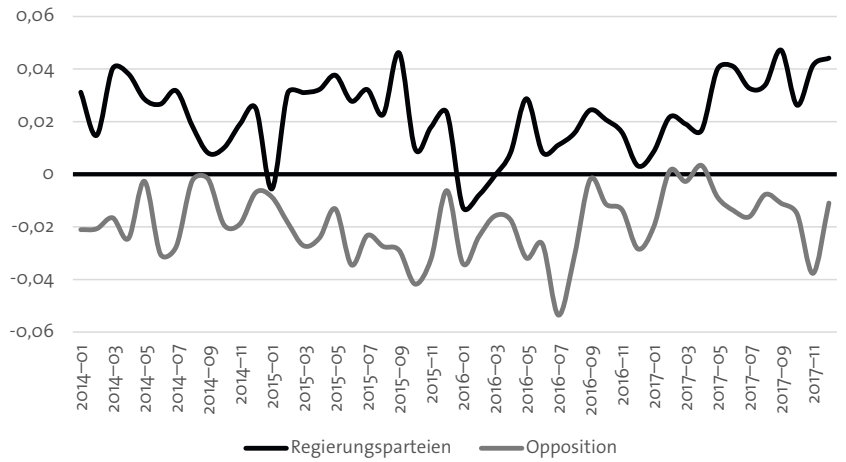

Abb. 4: Kumulierte Sentiment Scores – Regierungs- und Oppositionsparteien
Quelle: eigene Darstellung.

Die Ergebnisse zeigen die vielfältigen Möglichkeiten auf, mit denen textbasierte Verfahren in der politischen Analyse eingesetzt werden können. Es können frühzeitig relevante Themen erkannt und ihre Wirkung auf Wähler:innen und Konsument:innen untersucht werden. Die Methodik erlaubt in weitestgehend automatisierter Form, die Profile von Parteien zu analysieren. Dies lässt sich auf vielfältige weitere Themenfelder übertragen. Neben politischer Kommunikation und wirtschaftspolitischen Aspekten können Trends der gesamtwirtschaftlichen Entwicklung ausgewertet werden. Textdaten sollten daher in Forschung und Wirtschaft einen höheren Stellenwert erhalten. Sie können quantitative Ergebnisse unterstützen beziehungsweise sogar deren Aussagegehalt erhöhen.

Schlussfolgerungen

Wie gezeigt wurde, erfordert die Verwendung unstrukturierter Textdaten komplexe Auswertungsmethoden. Die Methoden können sich je nach Zielstellung unterscheiden. Gemeinsam ist ihnen allen das Ziel, mit möglichst geringen händischen Kodierungen große Textmengen zu analysieren. Während bei überwachten Strategien die Kodierung eines Trainingsdatensets nötig ist, ist bei un-

überwachten Strategien eine nachträgliche Kodierung der Cluster notwendig. Wie im Abschnitt „Texte als Daten“ gezeigt, haben beide Methoden Vorteile.

Gerade für den in diesem Beitrag erläuterten Ansatz zur Analyse politischer Kommunikation bietet sich das unüberwachte Clusteringverfahren an. Die explorative Strategie kann kann dazu dienen, neue Themen zu erkennen und so gesellschaftliche Entwicklungen für spätere Analyseverfahren aufzudecken. Das ermöglicht es, in Fore- und Nowcasts Entwicklungen abzubilden,[47] welche in Surveys noch nicht als implementierte Frage integriert sind. Neben Clusteringmethoden können über Sentimentanalysen oder präzisere psychologische Tools Informationen zu Stimmungen extrahiert werden.

Verschiedene Veröffentlichungen zeigen bereits den Nutzen unstrukturierter Daten in empirischen Verfahren. So konnten Entwicklungen an Aktienmärkten, die Reaktion auf Zentralbankkommunikation und positive Zusammenhänge zwischen Konsument:innenverhalten und Kommunikation in den Sozialen Medien identifiziert werden.

Die Ergebnisse und die vielfältigen Anwendungsfelder zeigen das Potenzial unstrukturierter Daten auf. Die Entwicklung zeigt in Richtung eines sich stetig ausweitenden Datenpools, eine Auswertung scheint daher ein vielversprechender und notwendiger Weg zu sein. Da die Methoden noch nicht gleichwertig validiert sind, bleibt das Feld ausgesprochen dynamisch, und es ist mit zunehmend besseren Modellen und Ergebnissen zu rechnen.

Literatur

Aggarwal, Charu C. (2012): An Introduction to Text Mining. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, S. 1–10.

Aggarwal, Charu C. / Zhai, ChengXiang (2012a): A Survey of Text Classification Algorithms. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, S. 163–222.

Aggarwal, Charu C. / Zhai, ChengXiang (2012b): A Survey of Text Clustering Algorithms. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, S. 77–128.

[47] Für weitere Informationen zu Fore- und Nowcasts siehe den vorhergehenden Beitrag in diesem Band.

Antonakaki, Despoina / Spiliotopoulos, Dimitris / V Samaras, Christos / Pratikakis, Polyvios / Ioannidis, Sotiris / Fragopoulou, Paraskevi (2017): Social media analysis during political turbulence. in: PloS one, Vol. 12, Nr. 10 (2017), e0186836.

Bao, Shenghua / Xu, Shengliang / Zhang, Li / Yan, Rong / Su, Zhong / Han, Dingyi / Yu, Yong (2009): Joint Emotion-Topic Modeling for Social Affective Text Mining. in: Data Mining, 2009. ICDM, 2009 Ninth IEEE International Conference, S. 699–704.

Beisch, Natalie / Schäfer, Carmen (2020): Internetnutzung mit großer Dynamik: Medien, Kommunikation, Social Media. in: Media Perspektiven, Nr. 9 (2020), S. 462–481.

Besimi, Adrian / Dika, Zamir / Shehu, Visar / Selimi, Mubarek (2019): Applied Text-Mining Algorithms for Stock Price Prediction Based on Financial News Articles. in: Managing Global Transitions, Vol. 17, Nr. 4 (2019), S. 335–351.

Bishop, Christopher M. (2006): Pattern recognition and machine learning. 1. Aufl., New York 2006.

Blei, David M. / Lafferty, John D. (2006): Dynamic topic models. in: ICML 2006. Proceedings, twenty-third International Conference on Machine Learning. hrsg. von Moore, Andrew / Cohen, William W., New York 2006, S. 113–120.

Blei, David M. / Lafferty, John D. (2007): A correlated topic model of Science. in: The annals of applied statistics Vol. 1, Nr. 1 (2007), S. 17–35.

Blei, David M. / Lafferty, John D. (2009): Topic Models. in: Text mining. Classification, clustering, and applications. hrsg. von Srivastava, Ashok N. / Sahami, Mehran, Boca Raton, Fla., 2009, S. 71–93.

Blei, David M. / Ng, Andrew Y. / Jordan, Michal I. (2003): Latent dirichlet allocation. in: Journal of Machine Learning Research, Vol. 3, Nr. 0 (2003), S. 993–1022.

Bollen, Johan / Mao, Huina / Zeng, Xiaojun (2011): Twitter mood predicts the stock market. in: Journal of Computational Science, Vol. 2, Nr. 1 (2011), S. 1–8.

Cai, Chiyu / Li, Linjing / Zengi, Daniel (2017): Behavior enhanced deep bot detection in social media. in: 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), S. 128–130.

Conference Board: Leading Economic Index. elektronisch veröffentlicht unter der URL: https://conference-board.org/data/bcicountry.cfm?cid=4, 02.11.2020.

Daas, Piet J. H. / Puts, Marco J. H. (2014): Social media sentiment and consumer confidence, Frankfurt, 2014.

Deng, Dong / Jing, Liping / Yu, Jian / Sun, Shaolong / Ng, Michael K. (2019): Sentiment lexicon construction with hierarchical supervision topic model. in: IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 27, Nr. 4 (2019), S. 704–718.

Dragu, Tiberiu / Fan, Xiaochen (2016): An Agenda-Setting Theory of Electoral Competition. in: The Journal of Politics, Vol. 78, Nr. 4 (2016), S. 1170–1183.

Einav, Liran / Levin, Jonathan (2014): Economics in the age of big data. in: Science, Vol. 346, Nr. 6210 (2014), S. 715–723.

Gentzkow, Matthew / Kelly, Bryan / Taddy, Matt (2019): Text as Data. in: Journal of Economic Literature, Vol. 57, Nr. 3 (2019), S. 535–574.

GfK: GfK-Konsumklima MAXX. elektronisch veröffentlicht unter der URL: https://www.gfk.com, 02.11.2020.

Glazer, Amihai / Lohmann, Susanne (1999): Setting the Agenda: Electoral Competition, Commitment of Policy, and Issue Salience. in: Public Choice, Vol. 99, Nr. 3/4 (1999), S. 377–394.

Hofmann, Thomas (2013): Probabilistic Latent Semantic Analysis, in: Machine Learning, Vol. 42, Nr. 1–2, S. 177–196.

Homburg, Christian / Ehm, Laura / Artz, Martin (2015): Measuring and Managing Consumer Sentiment in an Online Community Environment. in: Journal of Marketing Research, Vol. 52, Nr. 5 (2015), S. 629–641.

Hong, Sounman / Nadler, Daniel (2011): Does the early bird move the polls? in: dg.o 2011 : the proceedings of the 12th annual International Digital Government Research Conference : Digital Government Innovation in Challenging Times : University of Maryland, College Park, Maryland, USA, June 12–15, 2011. hrsg. von Luna Reyes, Luis F. / Chun, Soon A. / Bertot, John, 2011, S. 182.

Hu, Xia / Liu, Huan (2012): Text Analytics in Social Media. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, S. 385–414.

Iyetomi, Hiroshi / Aoyama, Hideaki / Fujiwara, Yoshi / Souma, Wataru / Vodenska, Irena / Yoshikawa, Hiroshi (2020): Relationship between Macroeconomic Indicators and Economic Cycles in U.S. in: Scientific reports, Vol. 10, Nr. 8420 (2020), S. 1–12.

Java, Akshay / Song, Xiaodan / Finin, Tim / Tseng, Belle (2007): Why we twitter. in: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis – WebKDD/SNA-KDD '07. hrsg. von Zhang, Haizheng / Mobasher, Bamshad / Giles, Lee / McCallum, Andrew / Nasraoui, Olfa / Spiliopoulou, Myra / Srivastava, Jaideep / Yen, John, New York, New York, USA 2007, S. 56–65.

Joshi, Aditya / Bhattacharyya, Pushpak / Carman, Mark (2016): Political Issue Extraction Model: A Novel Hierarchical Topic Model That Uses Tweets By Political And Non-Political Authors. in: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Stroudsburg, PA, USA, S. 82–90.

Kumar, P. / Das, T. K. (2013): BIG Data Analytics: A Framework for Unstructured Data Analysis. in: International journal of engineering and technology, Vol. 5 (2013), S. 153–156.

Liu, Bing / Zhang, Lei (2012): A Survey of Opinion Mining and Sentiment Analysis. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, S. 415–463.

Murphy, Kevin P. (2012): Machine learning. A probabilistic perspective, Cambridge, MA 2012.

Nenkova, Ani / McKeown, Kathleen (2012): A Survey of Text Summarization Techniques. in: Mining Text Data. hrsg. von Aggarwal, Charu C. / Zhai, ChengXiang, New York, NY 2012, 43–76.

Nguyen, Thien H. / Shirai, Kiyoaki (2015): Topic modeling based sentiment analysis on social media for stock market prediction. in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. hrsg. von Zong, Chengqing / Strube, Michael, 2015, S. 1354–1364.

OECD (2017): Leading indicators. Consumer Confidence Index 2017.

Oliveira, Lucas / Vaz de Melo, Pedro / Amaral, Marcelo / Pinho, José Antônio (2018): When Politicians Talk About Politics: Identifying Political Tweets of Brazilian Congressmen. in: Proceedings of the International AAAI Conference on Web and Social Media, Vol. 12, Nr. 1 (2018), S. 664–667.

Oliveira, Nuno / Cortez, Paulo / Areal, Nelson (2017): The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. in: Expert Systems with Applications, Vol. 73 (2017), S. 125–144.

Papakyriakopoulos, Orestis / Shahrezaye, Morteza / Thieltges, Andree / Medina Serrano, Juan Carlos / Hegelich, Simon (2017): Social Media und Microtargeting in Deutschland. in: Informatik-Spektrum, Vol. 40, Nr. 4 (2017), S. 327–335.

Pekar, Viktor / Binner, Jane (2017): Forecasting Consumer Spending from Purchase Intentions Expressed on Social Media. in: Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Stroudsburg, PA, USA 2017, S. 92–101.

Qian, Shengsheng / Zhang, Tianzhu / Xu, Changsheng / Shao, Jie (2016): Multi-Modal Event Topic Model for Social Event Analysis. in: IEEE Transactions on Multimedia, Vol. 18, Nr. 2 (2016), S. 233–246.

Quinn, Kevin M. / Monroe, Burt L. / Colaresi, Michael / Crespin, Michael H. / Radev, Dragomir R. (2010): How to Analyze Political Attention with Minimal Assumptions and Costs. in: American Journal of Political Science, Vol. 54, Nr. 1 (2010), S. 209–228.

Rao, Yanghui / Li, Qing / Wenyin, Liu / Wu, Qingyuan / Quan, Xiaojun (2014): Affective topic model for social emotion detection. in: Neural Networks, Vol. 58 (2014), S. 29–37.

Ravi, Kumar / Ravi, Vadlamani (2015): A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. in: Knowledge-Based Systems, Vol. 89 (2015), S. 14–46.

Remus, Robert / Quasthoff, Uwe / Heyer, Gerhard (2010): SentiWS – A Publicly Available German-language Resource for Sentiment Analysis. in: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta 2010.

Roberts, Margaret E. / Stewart, Brandon M. / Tingley, Dustin / Lucas, Christopher / Leder-Luis, Jetson / Gadarian, Shana K. / Albertson, Bethany / Rand, David G. (2014): Structural topic models for open-ended survey responses. in: American Journal of Political Science, Vol. 58, Nr. 4 (2014), S. 1064–1082.

Sakaki, Takeshi / Okazaki, Makoto / Matsuo, Yutaka (2010): Earthquake shakes Twitter users. in: Proceedings of the 19th international conference on World wide web – WWW '10, New York, New York, USA 2010, S. 851.

Stieglitz, Stefan / Dang-Xuan, Linh (2013): Social media and political communication: a social media analytics framework. in: Social network analysis and mining, Vol. 3, Nr. 4 (2013), S. 1277–1291.

Takikawa, Hiroki / Nagayoshi, Kikuko (2017): Political polarization in social media: Analysis of the „Twitter political field" in Japan. in: 2017 IEEE International Conference on Big Data (Big Data), Piscataway 2017, S. 3143–3150.

Tetlock, Paul C. (2007): Giving Content to Investor Sentiment: The Role of Media in the Stock Market. in: The Journal of Finance, Vol. 62, Nr. 3 (2007), S. 1139–1168.

Vamshi, Krishna B. / Pandey, Ajeet Kumar / Siva, Kumar A. P. (2018): Topic Model Based Opinion Mining and Sentiment Analysis. in: 2018 International Conference on Computer Communication and Informatics. January 04-06, 2018, Coimbatore, India. hrsg. von Informatics, I. C. o. C. C. a., Piscataway, NJ, 2018, S. 1–4.

Volkens, Andrea / Burst, Tobias / Krause, Werner / Lehmann, Pola / Matthieß, Theres / Merz, Nicolas / Regel, Sven / Weßels, Bernhard / Zehnter, Lisa / Wissenschaftszentrum Berlin Für Sozialforschung (WZB; 2020): Manifesto Project Dataset.

WorldBank (2019): Individuals using the internet (% of population). elektronisch veröffentlicht unter der URL: https://data.worldbank.org/indicator/IT.NET.USER.ZS, 5.1.2021.

Xie, Rui / Li, Chunping (2012): Lexicon construction: A topic model approach. in: 2012 International Conference on Systems and Informatics. ICSAI 2012: Yantai, Shandong, China, 19–20 May 2012, Piscataway, NJ 2012, S. 2299–2303.

Xue, Feng / Hong, Richang / He, Xiangnan / Wang, Jianwei / Qian, Shengsheng / Xu, Changsheng (2020): Knowledge-Based Topic Model for Multi-Modal Social Event Analysis. in: IEEE Transactions on Multimedia, Vol. 22, Nr. 8 (2020), S. 2098–2110.

Zuiderveen Borgesius, Frederik / Möller, Judith / Kruikemeier, Sanne / Ó Fathaigh, Ronan / Irion, Kristina / Dobber, Tom / Bodo, Balazs / Vreese, Claes H. de (2018): Online political microtargeting: promises and threats for democracy. in: Utrecht Law Review, Vol. 14, Nr. 1 (2018), S. 82–96.

Verfasserinnen und Verfasser

Islam , Zahurul, Professor an der NORDAKADEMIE Hochschule der Wirtschaft

Karaman Örsal, Deniz Dilan, Dr. rer. pol. (Humboldt-Universität zu Berlin), Universität Hamburg und außerplanmäßige Professorin an der Leuphana Universität Lüneburg.

Maass, Christina Heike, M. Sc. in Economics, Universität Hamburg

Roth, Felix, Privatdozent für Volkswirtschaftslehre an der Universität Hamburg und Leiter des Projekts GLOBALINTO im Rahmen der Horizon-2020-Forschungsförderung der Europäischen Kommission

Scheffer, Niklas, cand. rer. pol. (Universität Potsdam), Universität Hamburg, Institut für Computer Aided Psychometric Text Analysis (CAPTA)

Schneider, Henrique, Professor für Volkswirtschaftslehre an der Nordakademie, Hochschule der Wirtschaft, in Elmshorn und stellvertretender Direktor des Schweizerischen Gewerbeverbands sgv in Bern, Schweiz

Straubhaar, Thomas, Professor für Volkswirtschaftslehre, insbesondere Internationale Wirtschaftsbeziehungen der Universität Hamburg

Sturm, Silke, M. Sc (Universität Bayreuth), Universität Hamburg

Vöpel, Henning, Hamburgisches WeltWirtschaftsinstitut (HWWI) und Professor der Hamburg School of Business Administration (HSBA)

Hamburgisches
WeltWirtschafts
Institut

Reihe Edition HWWI
herausgegeben von Thomas Straubhaar

In der Edition HWWI (ISSN 1865-7974) erscheinen abgeschlossene, umfangreiche Projektergebnisse sowie Dissertationen zu Forschungsthemen, die vom HWWI bearbeitet werden . Folgende Titel sind bisher erschienen:

Band 1: Thomas Straubhaar (Hg.): Bedingungsloses Grundeinkommen und Solidarisches Bürgergeld – mehr als sozialutopische Konzepte, 2008.
ISBN 978-3-937816-47-0, DOI https://doi.org/10.15460/HUP.HWWI.1.69.

Band 2: Martin-Peter Büch et al. (Hg.): Sportfinanzierung – Spannungen zwischen Markt und Staat, 2009.
ISBN 978-3-937816-53-1, DOI https://doi.org/10.15460/HUP.HWWI.2.70.

Band 3: Martin-Peter Büch et al. (Hg.): Zur Ökonomik von Spitzenleistungen im internationalen Sport, 2012.
ISBN 978-3-937816-87-6, DOI https://doi.org/10.15460/HUP.HWWI.3.122.

Band 4 : Martin-Peter Büch et al. (Hg.): Sport und Sportgroßveranstaltungen in Europa – zwischen Zentralstaat und Regionen, 2012.
ISBN 978-3-937816-88-3, DOI https://doi.org/10.15460/HUP.HWWI.4.123.

Band 5: Seçil Paçacı Elitok, Thomas Straubhaar (eds.): Turkey, Migration and the EU: Potentials, Challenges and Opportunities, 2012.
ISBN 978-3-937816-94-4, DOI https://doi.org/10.15460/HUP.HWWI.5.118.

Band 6: Thomas Straubhaar (Hg.): Neuvermessung der Datenökonomie, 2021.
ISBN (Print) 978-3-943423-91-4, (Epub) 978-3-943423-94-5,
DOI https://doi.org/10.15460/HUP.HWWI.6.212.

Die Online-Ausgaben der Reihe sind frei zugänglich als Open-Access-Publikation erschienen. Die Printversion kann über den Buchhandel oder direkt beim Verlag (https://hup.sub.uni-hamburg.de) bezogen werden.