HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTERARBEIT

Herr
Andy Ludwig

Untersuchung der Themendynamik in sozialen Netzen am Beispiel deutschsprachiger Texte auf Twitter

2021

Fakultät **Angewandte Computer- und Biowissenschaften**

MASTERARBEIT

Untersuchung der Themendynamik in sozialen Netzen am Beispiel deutschsprachiger Texte auf Twitter

Autor:
Andy Ludwig

Studiengang:
Cybercrime/Cybersecurity

Seminargruppe:
CY19wC-M

Erstprüfer:
Prof. Dr. rer. nat. Dirk Labudde

Zweitprüfer:
Dr. rer. nat. Michael Spranger

Mittweida, 2021

Bibliografische Angaben

Ludwig, Andy: Untersuchung der Themendynamik in sozialen Netzen am Beispiel deutschsprachiger Texte auf Twitter, 105 Seiten, 78 Abbildungen, Hochschule Mittweida, University of Applied Sciences, Fakultät Angewandte Computer- und Biowissenschaften

Masterarbeit, 2021

Satz: LaTeX

Referat

Die vorliegende wissenschaftliche Abschlussarbeit behandelt die Untersuchung von Themenentwicklungen in deutschsprachigen Texten. Dazu wurden Twitterdaten von Bundestagsparteien analysiert. Über verschiedene Vorverarbeitungsschritte wurde eine LDA an das Problem angepasst. Mittels verschiedener Distanz- und Ähnlichkeitsmaße wurde eine Beschreibung der Themendynamik durchgeführt. Weiterhin wurden verschiedene Rahmenbedingungen erprobt, die zu einer Verbesserung der Ergebnisse führten.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Zeit der Anfertigung dieser Abschlussarbeit unterstützt und motiviert haben.

Vor allem danke ich Dr. Michael Spranger für die Bereitstellung des interessanten Themas, die intensive Betreuung und Beratung. Bedanken möchte ich mich auch bei Prof. Dr. Dirk Labudde für die Unterstützung und konstruktiven Besprechungen.

Weiterhin möchte ich dem gesamten Arbeitskreis danken, für die herzliche Aufnahme und das angenehme Arbeitsklima sowie die Unterstützung in dieser Zeit.

1 Einleitung

Einen besonderen Stellenwert in der modernen Kommunikation nehmen soziale Netzwerke ein. Ungefähr ein Drittel aller Menschen sind in einem sozialen Netzwerk registriert [Leskovec et al., 2009]. Durch die Nutzung dieser Dienste fallen enorme Mengen an Daten an, die eine sehr große Konzentration von Informationen beinhalten. Über computergestützte automatisierte Verfahren wird es möglich, das bestehende Wissen auszuwerten und aufzubereiten.

Der Kurznachrichtendienst Twitter bietet eine Plattform, auf der die Nutzenden über sogenannte Tweets eigene Inhalte teilen können. Der Dienst wird dabei von einer Vielzahl unterschiedlicher Gruppen genutzt, neben Privatpersonen unter anderem Unternehmen, öffentliche Institutionen und Parteien.

Im Umfeld der forensischen Untersuchungen stellt die Analyse der Textdaten von sozialen Netzwerken einen zentralen Aspekt dar. Dabei kann die große Datenmenge im Hinblick auf strafrechtlich relevante Sachverhalte untersucht werden. Neben einer Reihe etablierter Verfahren wie die Identifizierung von Meinungsführenden in sozialen Netzen können neue Ansätze weitere Informationen liefern.

Im Rahmen dieser wissenschaftlichen Abschlussarbeit sollte die Dynamik von Themen untersucht werden. Neben der statistischen und graphischen Aufbereitung der deutschsprachigen Daten sollten Themen extrahiert und deren Verlauf charakterisiert werden. Dafür sollte LDA als Verfahren zur automatischen Themendetektion eingesetzt werden, die mit verschiedenen Parameterkonfigurationen an das vorliegende Problem angepasst werden sollte.

Die extrahierten Themen sollten mittels verschiedener Abstands- und Ähnlichkeitsmaße miteinander verglichen werden, um deren Potential zur Beschreibung von Themenähnlichkeiten abzuschätzen. Weiterhin sollte eine Gruppierung der Ausgangsdaten in verschiedene Zeitintervalle erprobt werden, um den Einfluss auf die Themenextraktion zu untersuchen. Als weiteres Kriterium sollte der thematische Zusammenhang zwischen Tweets und Retweets analysiert werden.

Aus den extrahierten Themen sollte der Übergang zur Beschreibung von Themendynamiken erfolgen. Dafür sollten die Ähnlichkeiten von Themen zu verschiedenen Zeitpunkten betrachtet werden. Dies sollte sowohl innerhalb eines Autors als auch autorenübergreifend untersucht werden. Gleiche Themen können sich dabei zu verschiedenen Zeitpunkten in ihren Unterthemen unterscheiden. Weiterhin können Inhalte wiederholt aufgegriffen werden, sodass das Vorkommen von zyklisch wiederkehrenden Themen analysiert werden sollte. Abschließend sollte die Eignung von Hashtags zur Beschreibung von Themen untersucht werden.

2 Theoretischer Hintergrund

In diesem Kapitel soll ein Überblick über die theoretischen Hintergründe der angewandten Methoden und Modelle gegeben werden. Zu Beginn folgt eine allgemeine Beschreibung der Themenevolution in Textdaten. Im Weiteren werden Vorgehensweisen diskutiert, mit denen Themen aus Textdaten extrahiert werden können. Abschließend werden verschiedene Abstands- und Ähnlichkeitsmaße vorgestellt, die zur Charakterisierung der Themendynamik dienen.

2.1 Dynamik und Evolution von Themen in Textdaten

Durch eine immer größer werdende Menge an Textdaten sind computergestützte automatisierte Verfahren zur Datenanalyse unablässig. Vor allem soziale Netzwerke stellen eine große Menge an Daten bereit, die für forensische Untersuchungen genutzt werden können. Bereits zur Anwendung kommen Verfahren zur Analyse von Nachrichtendiensten und deren Inhalten sowie sozialen Netzwerken, um automatisch mögliche fallrelevante Spuren zu extrahieren [Spranger et al., 2016, Spranger et al., 2017]. Darüber hinaus können in Netzwerken und Gruppen Meinungsführende identifiziert werden [Spranger et al., 2018]. Ein weiterer Ansatz für die Informationsgewinnung in der Forensik ist die Untersuchung von Themen und deren Entwicklung in sozialen Netzwerken. Dadurch sollen größere Zusammenhänge und Strukturen in Textdaten aufgedeckt werden. Diese können zur Prädiktion von strafrechtlich relevanten Sachverhalten genutzt werden.

Zentrale Anwendungsfälle der Textdatenanalyse in sozialen Medien im Umfeld der Themenevolution sind die unter anderem die Identifizierung von Kernaspekten bzw. Themen, das Verhalten der Nutzenden und deren Einfluss auf die Themendynamik sowie die Beeinflussung der Profile untereinander. Bei den genannten Aspekten ist vor allem der zeitliche Verlauf von Bedeutung.

Ein grundsätzliches Problem der Analyse von Textdaten sozialer Medien ist die begrenzte Anzahl an verwendeten Zeichen. Dadurch müssen Techniken angewandt werden, die entweder aus sehr kurzen Texten Informationen extrahieren können oder die Texte zu einer größeren Menge miteinander kombinieren. Zur ersten Kategorie zählen beispielsweise Ansätze, die über die Analyse der Texte hinaus weitere Aspekte wie Autoreninformationen einbeziehen [Cai et al., 2014]. Beim Zusammenfassen der Texte zu größeren Gruppen wird versucht, eine ausreichend große Menge an Textdaten zu generieren, auf denen klassische Methoden angewandt werden können [Malik et al., 2013].

Die Themendynamik bzw. -evolution kann über verschiedene Ansätze beschrieben werden, da keine einheitliche Definition des Problems existiert. Ein in der Literatur zentraler

Aspekt ist das Verfolgen von Themen, nachdem diese in den Dokumenten isoliert wurden [Allan, 2009]. So soll der zeitliche Verlauf eines Themas charakterisiert werden. Eigenschaften, wie die spezifische Entwicklung des Hauptthemas, die Entwicklung von Subthemen oder der Anteil des Themas im Vergleich zu anderen Themenschwerpunkten, sind häufig behandelte Inhalte [Cai et al., 2014]. So können Themen zu spezifischen Zeitpunkten unter anderem entstehen, vergehen, konstant über die Zeit erhalten bleiben oder sich aufspalten [Abulaish, 2018].

Eng mit den beschriebenen Ansätzen verbunden ist eine Datenaufbereitung und Visualisierung. So ist die Darstellung der Themenentwicklung über verschiedene Zeitintervalle ein wichtiger Aspekt, damit große Datenmengen betrachtet werden können. Einfache Statistiken über das gemeinsame Auftreten von Termen oder Hashtags bieten meist einen engen Fokus und schaffen somit keine komplexe Themenbetrachtung. [Malik et al., 2013]

Somit können Ansätze zur Themendynamik Anwendung in der Identifizierung von temporär wichtigen Themen finden. Dadurch ist eine Analyse von Trends möglich. Weiterhin können sich gegenseitig beeinflussende Größen erkannt werden, die eine Vorhersage zur Dynamik von Themen und Diskursen ermöglichen. Auch eine Ausweitung auf die Rationalität oder Emotionalität in Unterhaltungen ist möglich.

2.2 Themen in Textdaten

Damit die Dynamik von Kernaspekten in Textdaten untersucht werden kann, müssen die Themen pro Dokument identifiziert werden. Dafür werden in diesem Abschnitt grundlegende Methodiken zu Sprach- und Themenmodellen vorgestellt. Anschließend folgen Ansätze zur Detektion von Themen und deren Verteilungen in Dokumenten.

2.2.1 Themenmodelle und Themenextraktion

Für die folgende Beschreibung der Modelle ist eine Formalisierung des Problems notwendig. Die Grundlage bildet eine Menge von N Dokumenten D, die in einer Sammlung C zusammengefasst werden: $C = \{D_1, ..., D_N\}$. Dabei besteht jedes Dokument D_i aus Wörtern w, die wiederum Teil des gesamten Vokabulars $V = \{w_1, ..., w_M\}$ sind.

Sprachmodell

Für die Charakterisierung und Beschreibung von Texten und deren Wortverteilungen im Umfeld der computergestützten Analysen können **statistische Sprachmodelle** herangezogen werden. Ein Sprachmodell ordnet einer Abfolge von Termen einen individuellen Wahrscheinlichkeitswert zu, sodass eine Wahrscheinlichkeitsverteilung für Wortsequenzen entsteht. Je nach Kontext können Abfolgen von Termen verschiedene Wahrscheinlichkeiten besitzen. [Zhai and Massung, 2016]

Im Gegenteil dazu weist das **Unigram-Sprachmodell** einzelnen Termen eine individuelle Wahrscheinlichkeit zu, sodass die Abfolge von Termen keinen Einfluss auf deren Wahrscheinlichkeitswert hat. In Gleichung (2.1) ist der Zusammenhang von Termen und deren Kombination dargestellt. [Zhai and Massung, 2016]

$$p(w_1, ..., w_M) = \prod_{i=1}^{M} p(w_i) \tag{2.1}$$

Dabei existiert die Bedingung, dass die Summe aller Wahrscheinlichkeiten in einem Unigram-Modell θ eins ergibt (Vgl. Gleichung (2.2)).

$$\sum_{w \in V} p(w) = 1 \tag{2.2}$$

In Tabelle 2.1 ist eine fiktive Wortverteilung eines Unigram-Sprachmodells dargestellt.

Tabelle 2.1: Beispiel einer Termverteilung in einem Dokument $p(w|\theta_{Dok})$

Dokument: Nachrichtenartikel	
...	
Aktuell	0.05
Bericht	0.03
Opfer	0.009
...	
versuchen	0.003
...	
Summe:	1

Bei der Betrachtung eines alleinigen Sprachmodells entsprechen die jeweiligen Termwahrscheinlichkeiten der relativen Häufigkeit im Dokument.

Themenmodell

Die Eigenschaften des Unigram-Sprachmodells können nicht nur auf die Termverteilung der gesamten Dokumente angewandt werden, sondern auch auf die Verteilung von Wörtern in einem Thema. Inhaltlich besteht ein Dokument aus einer Anzahl von k Themen $\{\theta_1, ..., \theta_k\}$. Der Anteil des Themas θ_j in Dokument D_i wird über $\pi_{i,j}$ beschrieben. Somit ergibt sich eine Themenabdeckung pro Dokument D_i über die jeweiligen Einzelanteile der Themen mit $\{\pi_{i1}, ..., \pi_{ik}\}$. Die Wahrscheinlichkeiten aller Themen in einem Dokument summieren sich zu eins (siehe Gleichung (2.3)).

$$\sum_{j=1}^{k} \pi_{i,j} = 1 \tag{2.3}$$

Themen können somit als eine Art Hauptgedanke des Textes beschrieben werden. Abhängig von dem betrachteten Intervall an Termen, können Themen für Sätze, Dokumente und Sammlungen von Dokumenten bestimmt werden. Im einfachsten Fall wird ein Thema durch ein Wort charakterisiert. In diesem Fall gibt π die relative Häufigkeit eines Terms an der Gesamtwortmenge an. Bei diesem Vorgehen können Probleme wie Wortsinnambiguität oder fehlende Komplexität in Themen resultieren. Durch Beschreibung der Kernaspekte mittels Wortverteilungen lassen sich diese vermeiden.

Bei **probabilistischen Themenmodellen** wird ein Thema durch eine Wortverteilung beschrieben, in der jeder Term eine individuelle Wahrscheinlichkeit für das Auftreten in dem Thema besitzt. Somit gibt π den Anteil eines Themas, repräsentiert durch eine Menge an Termen, am Dokument an. In Tabelle 2.2 ist eine fiktive Wortverteilung eines Themas dargestellt.

Tabelle 2.2: Beispiel einer Termverteilung in einem Thema $p(w|\theta_{Topic})$

Thema: Forensik	
Beweis	0.20
Spur	0.15
Tatort	0.08
...	
Computer	0.03
...	
Summe:	1

Eine Unterschied in den Themenmodellen ist die Anzahl an Themen, die in einem Dokument vorkommen. Sogenannte Mix-Modelle erlauben mehrere Themen in einem Dokument, Einzelthemenmodelle beschreiben dagegen Dokumente mit nur einem Thema.

In Mix-Modellen wird die Anzahl an Themen k als Prior-Parameter vorausgesetzt. [Anandarajan et al., 2019]

Ein Themenmodell kann als eine Art der Dimensionsreduktion angesehen werden. Eine Dokument-Term-Matrix dient als Input und wird wie Abbildung 2.1 in zwei Wahrscheinlichkeitsverteilungen überführt. Die erste gibt die Verteilung der Terme pro Thema an, die zweite die Verteilung der Themen über die Dokumente [Anandarajan et al., 2019].

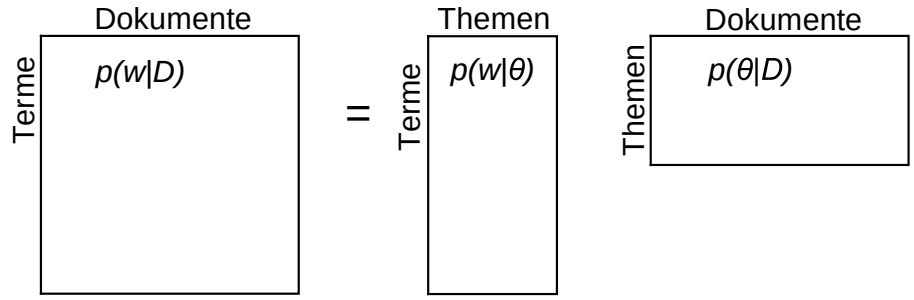

Abbildung 2.1: Schematische Darstellung der Dimensionsreduktion durch die Anwendung von Themenmodellen

Zusätzlich folgt die Betrachtung der bisher genannten Inhalte als **generative Modelle**. Liegen beispielsweise Unigram-Modelle vor, können die darin enthaltenen Terme mit der assoziierten Wahrscheinlichkeit dazu genutzt werden, Texte zu erzeugen. Als Input werden die Dokumentensammlung C, das damit verbundene Vokabular V und die Anzahl an Themen k übergeben. Das Ergebnis sind erzeugte Dokumente mit einer individuellen Themenabdeckung π der k Themen und eine Beschreibung der jeweiligen Themen θ selbst. Diese enthalten eine spezifische Termverteilung.

Die Wahrscheinlichkeiten für das Erzeugen zweier unterschiedlicher Dokumente D_1, D_2 unter Einbeziehung eines Sprachmodells θ sind verschieden: $p(D_1|\theta) \neq p(D_2|\theta)$. Eine höhere Wahrscheinlichkeit erzielt das Dokument, das mehrere Terme beinhaltet, die einen hohen Wahrscheinlichkeitswert im Sprachmodell θ liefern. [Zhai and Massung, 2016]

Themenextraktion

Das Finden von Themen bzw. dem zugrundeliegenden Sprachmodell in Texten ist ein zentraler Anwendungsfall der Themenmodelle und deren generativen Eigenschaften. Die Themenmodellierung stellt einen unüberwachten maschinellen Lernansatz dar. Zu einem gegeben Dokument D wird dasjenige Modell θ gesucht, welches die zugehörigen Termwahrscheinilchkeiten pro Wort $p(w|\theta)$ abbildet. Dazu ist eine Parameterschätzung notwendig, die in vielen Fällen über eine Maximum-Likelihood-Schätzung (ML-Schätzung) berechnet wird [Hofmann, 1999, Zhai and Massung, 2016].

Die zugrundeliegende Likelihood-Funktion bei einem Thema bzw. Unigram-Modell ist in Gleichung (2.4) beschrieben. Dabei setzt sich der Wahrscheinlichkeitswert für ein Dokument unter einem gegebenen Modell aus den jeweiligen Einzelwahrscheinlichkeiten jedes Terms unter θ zusammen. Der Exponent $c(w_i, D)$ gibt die Häufigkeit eines Wortes im Dokument am.

$$p(D|\theta) = p(w_1|\theta)^{c(w_1,D)} \cdot \ldots \cdot p(w_M|\theta)^{c(w_M,D)} = \prod_{i=1}^{M} p(w_i|\theta)^{c(w_i,D)} \tag{2.4}$$

Die Maximum-Likelihood-Schätzung bestimmt eine Reihe von verschiedenen Modellen $(\hat{\theta}_1, \ldots, \hat{\theta}_M)$, von denen dasjenige Modell gesucht wird, welches die folgende Gleichung (2.5) optimiert. [Zhai and Massung, 2016]

$$(\hat{\theta}_1, \ldots, \hat{\theta}_M) = \arg\max_{\hat{\theta}_1,\ldots,\hat{\theta}_M} p(D|\theta) = \arg\max_{\hat{\theta}_1,\ldots,\hat{\theta}_M} \prod_{i=1}^{M} p(w_i|\theta)^{c(w_i,D)} \tag{2.5}$$

Besteht das Dokument nur aus einem Thema bzw. wird ein Unigram-Sprachmodell gesucht, ist der Wahrscheinlichkeitswert für $p(w_i|\theta)$ durch Gleichung (2.6) gegeben.

$$p(w|\hat{\theta}) = \frac{c(w,D)}{|D|} \tag{2.6}$$

Komplexer wird die Bestimmung, wenn ein Mix-Modell betrachtet wird. Im Falle von zwei Modellen wird häufig die Menge der Füll- bzw. Stoppwörter in einem eigenständigen Hintergrundmodell θ_B zusammengefasst. Somit besteht die Betrachtung mittels Mix-Modellen aus einem Hintergrundmodell θ_B und einem Themenmodell θ_d. Wie in Gleichung (2.7) am Beispiel von zwei Modellen zu sehen, muss pro Term die Wahrscheinlichkeit einbezogen werden, aus welcher Wortverteilung dieser Term stammt.

$$p(w) = p(\theta_1)p(w|\theta_1) + p(\theta_2)p(w|\theta_2) \tag{2.7}$$

Pro betrachtetem Dokument muss die Summe über alle verwendeten Modelle eins ergeben (siehe Gleichung (2.8)).

$$p(\theta_d) + p(\theta_B) = 1 \tag{2.8}$$

Weiterhin muss die Bedingung gelten, dass die Summe über alle Wahrscheinlichkeitswerte pro Term und Thema wie in Gleichung (2.9) eins ergibt.

$$\sum_{i=1}^{M} p(w_i|\theta_d) = \sum_{i=1}^{M} p(w_i|\theta_B) = 1 \tag{2.9}$$

Auch die Parameterschätzung wird durch die Verwendung eines Mix-Modells angepasst. Zugrunde liegt im Falle von zwei Themen die Parametermenge
$\Lambda = (p(w|\theta_1), p(w|\theta_2), p(\theta_1), p(\theta_2))$. In Gleichung (2.10) ist die Likelihood-Funktion unter Verwendung von zwei verschiedenen Unigram-Modellen dargestellt.

$$p(d|\Lambda) = \prod_{i=1}^{|D|} p(x_i|\Lambda) = \prod_{i=1}^{M} [p(\theta_d)p(w_i|\theta_d) + p(\theta_B)p(w_i|\theta_B)]^{c(w,d)} \tag{2.10}$$

Zusammenfassend ergibt sich aus den genannten Bedingungen in Gleichung (2.8), (2.9) und der Likelihood-Funktion in Gleichung (2.10) die allgemeine ML-Schätzung zu Gleichung (2.11).

$$\Lambda^* = \arg\max_{\Lambda} p(D|\Lambda) \tag{2.11}$$

Eine Möglichkeit, die ML-Schätzung zu berechnen, ist der Expectation-Maximization-Algorithmus (EM-Algorithmus) [Asuncion et al., 2009]. Dieser Ansatz konvergiert bei lokalen Maxima der Likelihood $p(d|\theta)$. Es wird kein globales Maximum gefunden, weswegen der Ansatz wiederholt mit verschiedenen Startparametern durchgeführt werden muss [Zhai and Massung, 2016]. Die Berechnung erfolgt dabei in zwei Schritten [Hofmann, 2001]:

1. Expectation-Step (E-Step): die a-posteriori-Wahrscheinlichkeiten der unbekannten Variablen werden in Abhängigkeit der aktuellen Parameterwerte berechnet

2. Maximization-Step (M-Step): die im E-Step berechneten Wahrscheinlichkeiten werden benutzt, um die Parameter anzupassen und somit die Schätzung zu verbessern

2.2.2 Ansätze zur Themenextraktion - PLSA und LDA

Es existieren verschiedene Ansätze zur Extraktion mehrerer Themen in einem Dokument. Weit verbreitet sind der grundlegende Ansatz Probabilistic Latent Semantic Analysis (PLSA) nach [Hofmann, 2001] und Latent Dirichlet Allocation (LDA) nach [Blei et al., 2002]. Dabei kann LDA als Weiterentwicklung der PLSA angesehen werden, da die grundlegenden Elemente Gemeinsamkeiten aufweisen.

PLSA ermöglicht die Zuordnung der Terme eines Dokumentes zu mehreren Themen. Ein zentraler Bestandteil ist die Bestimmung der jeweilige Themenabdeckung pro Dokument. Dabei werden Themen als Wortverteilungen betrachtet und jeder Term eines Dokumentes kann durch ein probabilistisches Modell generiert werden. PLSA ist somit die Erweiterung eines Mix-Modells mit zwei Unigram-Modellen zu einem Mix-Modell mit mehreren Themen. [Zhai and Massung, 2016]

Unter Annahme, dass Füll- bzw. Stoppwörter in einem Hintergrundmodell zusammengefasst sind, ergeben sich folgende Wahrscheinlichkeiten der Themen und Terme in einem Modell:

1. Die Wahrscheinlichkeit, dass ein Term aus den Hintergrundmodell θ_B stammt, entspricht λ_B.
2. Alle weiteren Themenmodelle $\theta_1, .., \theta_k$ besitzen die kumulierte Wahrscheinlichkeit $1 - \lambda_B$.
3. Die individuelle Wahrscheinlichkeit eines Themas in Dokument D entspricht $p(\theta_i) = (1 - \lambda_B)\pi_{D,i}$ und es gilt die folgende Bedingung in Gleichung (2.12).

$$\sum_{i=1}^{k} \pi_{D,i} = 1 \tag{2.12}$$

4. Pro Term ist die individuelle Wahrscheinlichkeit im Hintergrundmodell durch $\lambda_B p(w|\theta_B)$ und in allen anderen Themenmodellen durch $(1 - \lambda_B)\pi_{D,i} p(w|\theta_i)$ gegeben. Dabei gilt die Bedingung in Gleichung (2.13).

$$\sum_{i=1}^{M} p(w_i|\theta_j) = 1, \forall j \in [1,k] \tag{2.13}$$

Der Generierungsprozess besteht somit aus der Wahl des Themenmodells und daran anschließend aus der Wahl eines Terms aus dem selektierten Themenmodell. Die

grundlegende Funktionsweise ist in Abbildung 2.2 dargestellt. Dabei wird als Generierungsprozess für ein Dokument $D \in N$ ein Thema θ_z gewählt und daraus ein Term $w \in M$ abgeleitet. Die Anzahl der Themen ist ein Input-Parameter des Algorithmus. Der Generierungsprozess bezieht sich lediglich auf die Terme eines Dokuments. Mittels PLSA können keine neuen Dokumente erstellt werden, weshalb es nicht allgemein als generatives Modell bezeichnet wird.

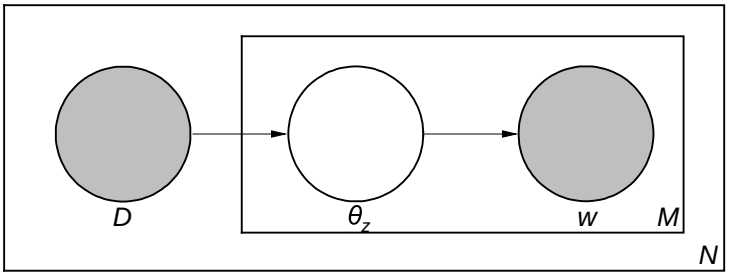

Abbildung 2.2: Schematische Darstellung PLSA

Aus den genannten Anpassungen bei PLSA, im Vergleich zu den bisherigen Ansätzen, folgt auch eine Aktualisierung der Likelihood-Funktion. Die Menge der unbekannten Parameter lautet: $\Lambda = (\{\pi_{D,j}\}, \{\theta_j\}, \forall j \in [1,k])$. Die Wahrscheinlichkeit für ein Wort w in Dokument D ist in Gleichung (2.14) gegeben. In Gleichung (2.15) und (2.16) folgen die Anpassungen für ein Dokument und für die Sammlung von Dokumenten. Die Einführung des Logarithmus ist auf den Umstand zurückzuführen, dass ein arithmetischer Unterlauf vermieden werden soll [Zhai and Massung, 2016].

$$p_d(w) = \lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{D,j} p(w|\theta_j) \tag{2.14}$$

$$\log p(d) = \sum_{w \in V} c(w,d) \log[\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{D,j} p(w|\theta_j)] \tag{2.15}$$

$$\log p(C|\Lambda) = \sum_{d \in C} \sum_{w \in V} c(w,d) log[\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{D,j} p(w|\theta_j)] \tag{2.16}$$

Zur Optimierung der Likelihood-Funktion kann erneut eine ML-Schätzung verwendet werden. In Gleichung (2.17) ist die Optimierung dargestellt, es gelten die Bedingungen wie in Gleichung (2.12) und (2.13).

$$\Lambda^* = \arg\max_{\Lambda}(C|\Lambda) \tag{2.17}$$

Ein Nachteil dieses Ansatzes ist, dass bei der Ableitung der Themen kein weiteres Wissen einbezogen wird. Somit sind vor allem keine Informationen über die Verteilung der Themen pro Dokument und die Verteilung der Terme pro Thema a priori vorhanden.

Neben einer nutzergesteuerten PLSA, die zusätzliches Wissen einbezieht, löst **LDA** die Probleme des fehlenden a-priori Wissens. Zusätzlich kann LDA als komplett generatives Modell betrachtet werden. Dabei generiert LDA zusätzliche Informationen aus Multinomial- und Dirichlet-Verteilungen und ergänzt somit die PLSA-Grundlagen. LDA wird somit zu einer Bayes'schen Version von PLSA. [Zhai and Massung, 2016]

Die Dichtefunktion der Dirichlet-Verteilung ist in Gleichung (2.18) dargestellt, wobei Γ die Gammafunktion bezeichnet.

$$p(x|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_i x_i^{\alpha_i - 1} \tag{2.18}$$

In Abbildung 2.3 ist die grundlegende Funktionsweise von LDA dargestellt. Darin ist als zentrales Element der Ansatz der PLSA sichtbar, zusätzlich sind die neuen Verteilungen zu erkennen. Der Parameter α symbolisiert die a-priori Dirichlet-Verteilung, die die Wahrscheinlichkeiten der Themen in einem Dokument steuert. Daraus resultiert eine Multinomialverteilung für die Themenverteilung pro Dokument π. Somit ergeben sich die individuellen Themen θ_z und weiterhin die Terme w pro Thema und Dokument. Hinzu kommt bei LDA der Parameter β, der eine a-priori Dirichlet-Verteilung für die Wahrscheinlichkeiten der Terme in einem Thema steuert. Aus dieser folgt eine Multinomialverteilung θ der Terme pro Thema.

Der generierende Prozess kann durch folgende Schritte zusammengefasst werden [Blei and Lafferty, 2009]:

1. Für jedes Thema:
   a) Wähle eine Termverteilung $p(\vec{\theta}_k) \sim Dirichlet(\vec{\beta})$
2. Für jedes Dokument:
   a) Wähle eine Themenverteilung $\vec{\pi}_D \sim Dirichlet(\vec{\alpha})$
   b) Für jedes Wort:
      i. Wähle ein Thema $\theta_i \sim Multinomial(\vec{\pi}_D)$ mit $\theta_i \in [1,...,k]$
      ii. Wähle ein Wort $w_i \sim Multinomial(\vec{\theta}_k)$ mit $w_i \in V$

Somit wird der Algorithmus über die Parameter α und β gesteuert. Dabei ist $\vec{\alpha} = (\alpha_1,...,\alpha_k)$ mit $\alpha_i > 0$ und $\vec{\beta} = (\beta_1,...,\beta_M)$ mit $\beta_i > 0$. Die Themenabdeckung pro Dokument $\vec{\pi}_D$ ist durch die a-priori Dirichlet-Verteilung gegeben durch: $\vec{\pi}_D = Dirichlet(\vec{\alpha})$. Die Termverteilung pro Thema $\vec{\theta}_i$ ist durch die a-priori Dirichlet-Verteilung gegeben

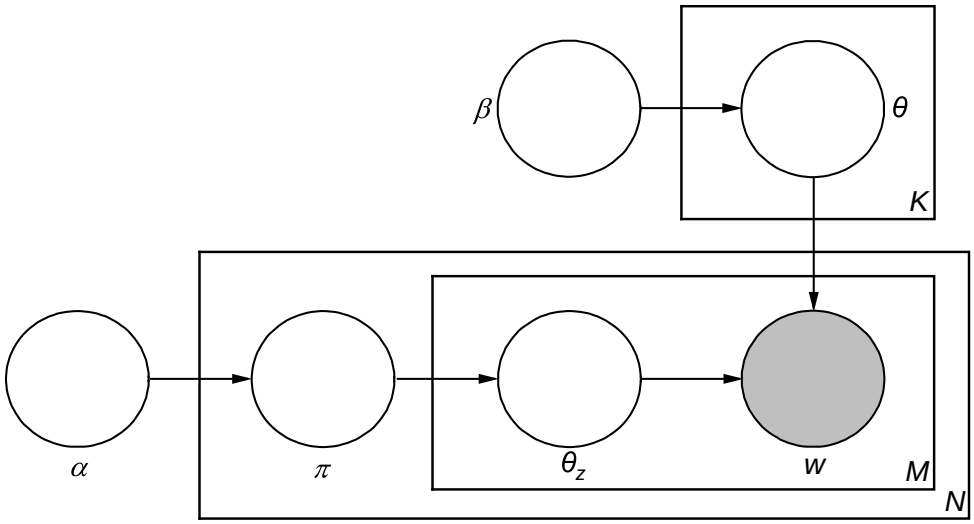

Abbildung 2.3: Schematische Darstellung LDA (nach [Zhu et al., 2016])

durch: $p(\vec{\theta}_i) = Dirichlet(\vec{\beta})$. Obwohl α und β Vektoren der Dimension k sind, werden in den meisten Fällen Skalare übergeben. Der Grund liegt in fehlendem Wissen darüber, wie beispielsweise einzelne Verteilungen in Dokumenten real aussehen. Somit steht der Skalar für α und β stellvertretend für alle k, M Werte der jeweiligen Vektoren.

Die Wahl der Parameter α und β bestimmt maßgeblich das Resultat der Themenextraktion. Wie in Abbildung 2.4 zu erkennen, beeinflussen die Parameter die Dirichlet-Verteilung. Je größer diese gewählt werden, desto zentrierter ist die Werteverteilung. Am Beispiel der Themenabdeckung bewirkt ein hohes α eine zentrierte Verteilung der Themen im Dokument. Wird dagegen eine differenziertere Themenverteilung angestrebt, muss der Parameter kleiner gewählt werden. Ein ähnliches Verhalten liefert der β: je kleiner β, desto weniger Worte besitzen eine hohe Wahrscheinlichkeit und sind somit charakteristisch pro Thema.

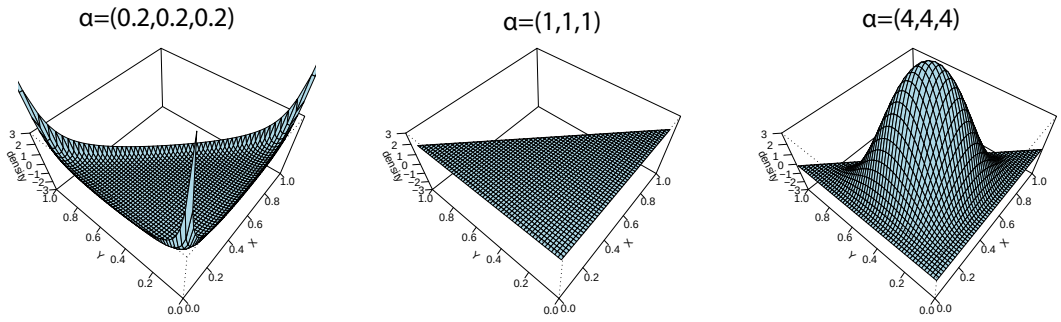

Abbildung 2.4: Beispiele verschiedener Dirichlet-Verteilungen; links: $\alpha = 0.2$, mittig: $\alpha = 1$, rechts: $\alpha = 4$

Eine Übertragung dieses Verhaltens der Dirichlet-Verteilung auf die Term- und Themenwahl kann durch Visualisierung eines Simplex erfolgen. Eine einfache Darstellung erfolgt mittels eines Dreiecks (2-Simplex) wie in Abbildung 2.5. Darin sind drei Themen mit drei Termen dargestellt. Die Lage der Themen-Punkte beschreibt die Zugehörigkeit bzw. Wahrscheinlichkeit, ein Term in diesem Thema zu beobachten. Die Lage der Dokumente in den Themen verdeutlicht die jeweiligen Anteile der Themen pro Dokument. [Blei et al., 2002]

Abschließend können auch für LDA die Likelihood-Funktionen bestimmt werden. Diese unterscheiden sich durch die Aufnahme der Parameter α und β. Die Wahrscheinlichkeit für ein Wort ist dabei unverändert wie in Gleichung (2.19). Die Modifikationen zur Bestimmung der Dokumentwahrscheinlichkeit ist in Gleichung (2.20) und der Dokumentensammlung in Gleichung (2.21) dargestellt.

$$p_d(w|\{\theta_j\},\{\pi_{D,j}\}) = \sum_{j=1}^{k} \pi_{D,j} p(w|\theta_j) \tag{2.19}$$

$$\log p(d|\vec{\alpha},\{\theta_j\}) = \int \sum_{w\in V} c(w,D) \log \left[\sum_{j=1}^{k} \pi_{D,j} p(w|\theta_j) \right] p(\vec{\pi}_D|\vec{\alpha}) d\vec{\pi}_D \tag{2.20}$$

$$\log p(C|\vec{\alpha},\vec{\beta}) = \int \sum_{d\in C} \log p(D|\vec{\alpha},\theta_j) \prod_{j=1}^{k} p(\theta_j|\vec{\beta}) d\theta_1,...,\theta_k \tag{2.21}$$

Weiterhin kann an dieser Stelle zur Optimierung der Parameter eine ML-Schätzung verwendet werden [Blei et al., 2002]. Die Optimierungsfunktion ist in Gleichung (2.22) dar-

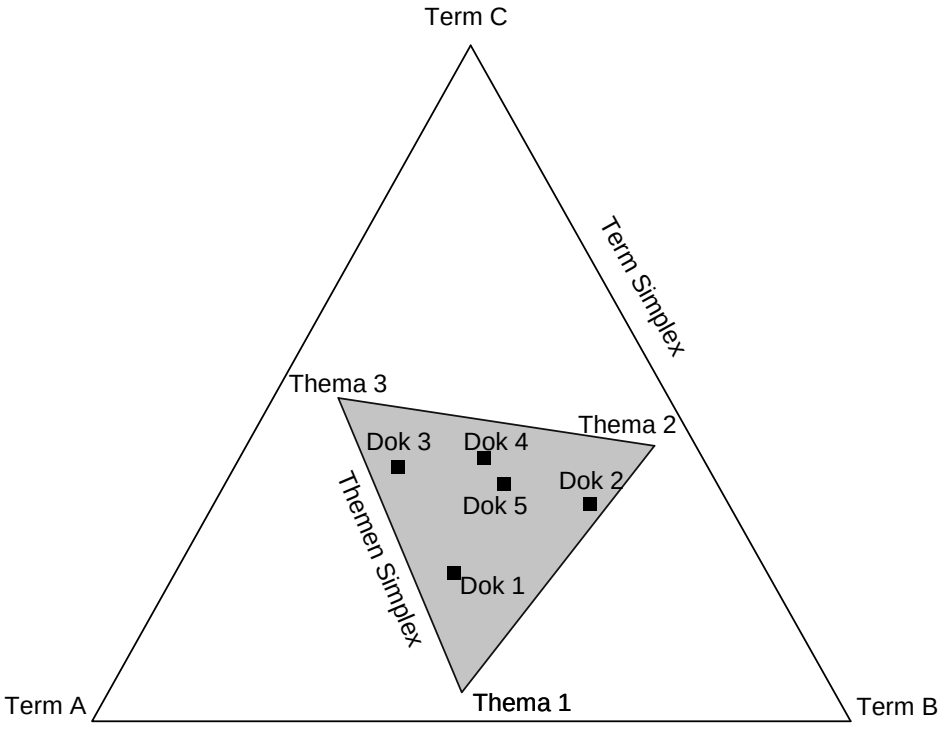

Abbildung 2.5: Darstellung eines 2-Simple zur Funktionsweise von LDA

gestellt. Dabei muss nun die Wortverteilung $\{\theta_i\}$ und die Themenabdeckung $\{\pi_{D,j}\}$ a-posteriori bestimmt werden. Die Berechnung folgt in Gleichung (2.23) unter Anwendung der Bayes'schen Regel. Diese komplexe Berechnung kann beispielsweise mittels Collapsed Gibbs Sampling durchgeführt werden. Diese Methode stellt eine Art Markov-Ketten-Monte-Carlo Verfahren dar (MCMC-Verfahren). Weiterhin wurden Anpassungen für das Collapsed-Gibbs-Sampling entwickelt, die eine signifikante Zeitersparnis bei der Berechnung liefern. [Zhai and Massung, 2016, Griffiths and Steyvers, 2004, Porteous et al., 2008]

$$(\vec{\alpha},\vec{\beta}) = \arg\max_{\vec{\alpha},\vec{\beta}} \log p(C|\vec{\alpha},\vec{\beta}) \tag{2.22}$$

$$p(\{\theta_i\},\{\pi_{D,j}\}|C,\alpha,\beta) = \frac{p(C|\{\theta_i\},\{\pi_{D,j}\})p(\{\theta_i\},\{\pi_{D,j}\}|\alpha,\beta)}{p(C|\alpha,\beta)} \tag{2.23}$$

2.3 Abstände und Ähnlichkeiten

Dieser Abschnitt gibt eine Übersicht zu verschiedenen Abstands- und Ähnlichkeitsmaßen. Dabei werden Distanzen in metrischen Räumen, Ähnlichkeiten von Objekten und Wahrscheinlichkeitsverteilungen betrachtet.

Ein Distanzmaß $dist : X \times X \to \mathbb{R}$ besitzt dabei die Eigenschaft, dass gleiche Objekte keinen Abstand aufweisen: $dist(x,x) = 0,\ x \in X$. Unterschiedliche Objekte weisen einen Wert verschieden von Null auf: $dist(x,y) > 0,\ x,y \in X,\ x \neq y$. Ein Distanzmaß kann zusätzlich normiert werden, sodass gilt: $dist(x,y) \leq 1$.

Ein Ähnlichkeitsmaß $sim : X \times X \to \mathbb{R}$ hingegen ordnet identischen Objekten einen Wert von eins zu: $sim(x,x) = 1,\ x \in X$. Werden verschiedene Objekte miteinander verglichen, liegt der Wert unter eins: $sim(x,y) < 1,\ x,y \in X,\ x \neq y$. Somit besitzen ähnliche Objekte einen geringen Abstand und eine hohe Ähnlichkeit.

Ähnlichkeits- und Distanzfunktionen können dabei ineinander umgerechnet werden. In Gleichung (2.24) ist das Überführen einer Ähnlichkeitsfunktion in ein Distanzmaß dargestellt. Die entgegengesetzte Umrechnung erfolgt durch eine Normierung der Distanzen mit dem größten Abstand auf ein Intervall $[0,1]$ in Gleichung (2.25).

$$dist(x,y) = 1 - sim(x,y) \tag{2.24}$$

$$sim(x,y) = 1 - \frac{dist(x,y)}{max\{dist(a,b)|a,b \in X\}} \tag{2.25}$$

2.3.1 Abstandsmaße

Die Distanz zwischen zwei Objekten im Raum kann formal über eine Metrik definiert werden. Wird eine Menge X betrachtet, ist eine Metrik auf dieser Menge eine Abbildung $d : X \times X \to \mathbb{R},\ (x,y) \mapsto d(x,y)$. [Forster, 2017]

Dafür müssen folgende Eigenschaften erfüllt sein:

1. Positive Definitheit: $d(x,y) = 0 \Leftrightarrow x = y,\ x,y \in X$
2. Symmetrie: $d(x,y) = d(y,x),\ \forall x,y \in X$
3. Dreiecksungleichung: $d(x,z) \leq d(x,y) + d(y,z),\ \forall x,y,z \in X$

Eine Metrik erfüllt im Unterschied zu allgemeinen Distanzmaßen die Eigenschaft der Symmetrie. Das erste Distanzmaß, das zugleich die Grundlage bzw. eine Verallgemei-

nerung weiterer Maße darstellt, ist die **Minkowski-Distanz** in Gleichung (2.26) [Sartorius, 2019].

$$d(X,Y) = \left[\sum_{d=1}^{D} |X_d - Y_d|^C\right]^{\frac{1}{c}} \tag{2.26}$$

D repräsentiert die Dimensionalität der Daten und c wird als Minkowski-Konstante bezeichnet. Je nach Wahl des Wertes für c ergeben sich verschiedene Distanzmaße. Für $c = 1$ folgt die **Manhattan-Distanz** (L^1-Norm) in Gleichung (2.27), mit $c = 2$ die **euklidische Distanz** (L^2-Norm) in Gleichung (2.28). Bei diesen Metriken geben die Ergebnisse den Abstand im Raum an, je größer das Ergebnis, desto größer die Distanz zwischen X und Y.

$$d(X,Y) = \sum_{d=1}^{D} |Xd - Y_D| \tag{2.27}$$

$$d(X,Y) = \sqrt{\sum_{d=1}^{D} (X_d - Y_d)^2} \tag{2.28}$$

Eine Anpassung der euklidischen Distanz ist über die Einbeziehung der Standardabweichung s_k möglich. Damit werden Einflüsse einzelner Ausreißer minimiert. Die Anpassung ist in Gleichung (2.29) dargestellt.

$$d(X,Y) = \sqrt{\sum_{d=1}^{D} \frac{(X_d - Y_d)^2}{s_k^2}} \tag{2.29}$$

2.3.2 Ähnlichkeitsmaße

Ein Maß zur Bestimmung der Ähnlichkeit zweier Vektoren ist die **Kosinus-Ähnlichkeit** in Gleichung (2.30). Dabei wird der Winkel zwischen zwei Vektoren $\vec{X} = (x_1, ..., x_n)$ und $\vec{Y} = (y_1, ..., y_n)$ im Raum bestimmt. Das Ergebnis liegt im Intervall $[-1, 1]$. Ein Ergebnis von -1 tritt bei entgegengesetzten Vektoren auf. Sind die Vektoren dagegen identisch, resultiert ein Ergebnis von 1. [Aggarwal, 2018]

$$sim_{Kosinus}(\vec{X}, \vec{Y}) = \frac{\sum_{d=1}^{D} x_d y_d}{\sqrt{\sum_{d=1}^{D} x_d^2} \sqrt{\sum_{d=1}^{D} y_i^2}} \tag{2.30}$$

Die **Jaccard-Ähnlichkeit** beruht auf dem Vergleich von zwei Mengen. Wie in Gleichung (2.31) zu erkennen, werden zwei Sets miteinander verglichen. Abhängig von der Anzahl der übereinstimmenden Elemente ergibt sich die Ähnlichkeit.

$$sim_{Jaccard}(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.31}$$

Aus Gleichung (2.31) folgt in anderer Darstellung der **Tanimoto-Koeffizient** in Gleichung (2.32). Eine Anpassung für den Vergleich binärer Daten ist in Gleichung (2.33) dargestellt. Dabei bezeichnen N_X und N_Y die Anzahl von 1-Bits pro Vektor und N_{XY} steht für die Anzahl an übereinstimmenden 1-Bits.

$$sim_{Tanimoto}(X,Y) = \frac{XY}{||X||^2 + ||Y||^2 - XY} \tag{2.32}$$

$$sim_{Tanimoto-Binary}(X,Y) = \frac{N_{XY}}{N_X + N_Y - N_{XY}} \tag{2.33}$$

Eine weitere Maßzahl zur Berechnung von Ähnlichkeiten ist der **Pearson-Korrelationskoeffizient** in Gleichung (2.34) [Zhai and Massung, 2016].

$$sim_{Pearson}(X,Y) = \frac{\sum_{d=1}^{D}(x_d - \bar{x})(y_d - \bar{y})}{\sqrt{\sum_{d=1}^{D}(x_d - \bar{x})^2} \cdot \sqrt{\sum_{d=1}^{D}(y_i - \bar{y})^2}} \tag{2.34}$$

2.3.3 Vergleich von Zeichenketten

Für den Vergleich von Zeichenabfolgen gibt es verschiedene Ansätze, die eine Ähnlichkeit bzw. Distanz numerisch beschreiben. Der **Hamming-Abstand** bestimmt den Unterschied zwischen zwei Abfolgen gleicher Länge anhand der Anzahl unterschiedlicher Stellen [Hamming, 1950]. Der Hamming-Abstand findet u.a. Anwendung in dem Vergleich zweier Binärfolgen zur Detektion von Übertragungsfehlern. In Gleichung (2.35) ist die Differenzbestimmung anhand von Binärdaten beschrieben. Durch das Erweitern des Alphabets kann die Logik auf den Vergleich von Strings angewandt werden.

$$d_{Hamming}(X,Y) = \sum_{d=1}^{D} |x_d - y_d| \tag{2.35}$$

Die **Levenshtein-Distanz**, auch Edit-Distanz genannt, bestimmt den Unterschied zwischen Strings durch die Anzahl an Editieroperationen, die notwendig sind, um einen

String in einen anderen zu überführen [Levenshtein, 1966]. Dabei können auch Strings verarbeitet werden, die eine unterschiedliche Länge besitzen. Die Editieroperationen Einfügen und Löschen sind zugelassen. In Gleichung (2.36) ist die Distanzbestimmung beschrieben, wobei LCS für longest common subsequence (längste gemeinsame Teilzeichenkette) steht.

$$d_{Levenshtein}(X,Y) = |X| + |Y| - 2|LCS(X,Y)| \tag{2.36}$$

2.3.4 Vergleich von Wahrscheinlichkeitsverteilungen

Neben dem Vergleich von Vektorrepräsentationen von Termen bzw. die Symbolähnlichkeit, können Termverteilungen direkt als Wahrscheinlichkeitsverteilungen einander gegenübergestellt werden. Dabei wird die Distanz zwischen Termverteilungen und den jeweiligen assoziierten Wahrscheinlichkeiten berechnet.

Ein weit verbreitetes Maß ist die **Kullback-Leibler-Divergenz** (KL-Divergenz) in Gleichung (2.37). Dabei stehen P und Q für zwei diskrete Wahrscheinlichkeitsverteilungen.

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{2.37}$$

Die KL-Divergenz ist asymmetrisch: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, was zu verschiedenen Ergebnissen bei unterschiedlicher Argumentreihenfolge führt. Deshalb kann einerseits die symmetrische KL-Divergenz in Gleichung (2.38) [Zong et al., 2021], andererseits die gemittelte KL-Divergenz in Gleichung (2.39) [Huang, 2008] bestimmt werden. In letzterer Version bedeutet: $\pi_1 = \frac{P}{P+Q}$, $\pi_2 = \frac{Q}{P+Q}$ und $M = \pi_1 P + \pi_2 Q$.

$$D_{KL_{Sym}}(P,Q) = D_{KL}(P||Q) + D_{KL}(Q||P) \tag{2.38}$$

$$D_{KL_{Avg}}(P,Q) = \pi_1 D_{KL}(P||M) + \pi_2 D_{KL}(Q||M) \tag{2.39}$$

Ein weiteres Maß zum Vergleich von Wahrscheinlichkeitsverteilungen ist die sogenannte **Jensen-Shannon-Divergenz** (JS-Divergenz) in Gleichung (2.40). Diese Berechnung kann als eine symmetrische und geglättete Version der KL-Divergenz betrachtet werden.

$$D_{JS}(P||Q) = \frac{1}{2}\sum P_s \log \frac{P_s}{\frac{1}{2}(P_s+P_g)} + \frac{1}{2}\sum P_g \log \frac{P_g}{\frac{1}{2}(P_s+P_g)} \tag{2.40}$$

Eine angepasste Form der Gleichung (2.40) ermöglicht das Vergleichen von Wahrscheinlichkeitsverteilungen auf Ebene der einzelnen Einträge x. In Gleichung (2.41) ist die Anpassung dargestellt. [Arun et al., 2010]

$$D_{JS-Single}(x) = \frac{1}{2} \cdot P_s(x) \log \frac{2P_s(x)}{P_s(x)+P_g(x)} + P_g(x) \log \frac{2P_g(x)}{P_s(x)+P_g(x)} \tag{2.41}$$

Eine letzte Maßzahl zur Berechnung von Ähnlichkeiten von Wahrscheinlichkeitsverteilungen ist die sogenannte **Hellinger Distanz** in Gleichung (2.42). [Blei and Lafferty, 2009]

$$D_{Hellinger}(P||Q) = \sum_{k=1}^{K} (\sqrt{P_k} - \sqrt{Q_k})^2 \tag{2.42}$$

2.4 Literaturübersicht - Ansätze und Methoden

Zur Vervollständigung des theoretischen Hintergrundes werden in diesem Abschnitt Ansätze aus der Literatur vorgestellt. Eine breite Übersicht zu verschiedenen Themenmodellen auf Basis von LDA und deren Anwendung gibt das Review [Jelodar et al., 2019]. Eine zweite Übersicht zu Ansätzen der Themenevolution in Kombination mit probabilistischen Themenmodellen liefert das Review [Zhou et al., 2017]. In [Allan, 2009] wurde in mehreren Publikationen das Gebiet des Topic Detection and Tracking behandelt.

Für die Evolution von Themen müssen grundlegende Rahmenbedingungen gewählt werden [Zhou et al., 2017]:

1. die Betrachtung der Zeit:
    a) diskrete Zeitintervalle: die Daten werden in Zeitabschnitte eingeteilt
    b) kontinuierliche Zeitintervalle: keine Untergliederung der Zeiten
    c) online Modell: die Zeit wird durch einen Online-Stream bestimmt
2. die Anzahl der Themen
    a) im Evolutionsprozess wird immer eine konstante Themenanzahl bestimmt
    b) im Evolutionsprozess wird eine variable Themenanzahl bestimmt
3. die genutzte Datengrundlage im Corpus

a) nur Textdaten

b) zusätzliche Informationen neben Textdaten wie bspw. Autoreninformationen

In [Malik et al., 2013] wurde der Fokus neben der Analyse der Themendynamik auf die visuelle Aufbereitung der Ergebnisse gelegt. Die Autoren entwickelten ein eigenes interaktives Tool (TopicFlow) zur graphischen Datenaufbereitung. Darin wurden die zeitlichen Verläufe und Beziehungen der einzelnen Themen pro Zeitabschnitt zusammengefasst. Für die Analyse von Themendynamiken wurden Twitterdaten (pro Datensatz ca. 1500 Tweets) genutzt, deren Themen für die graphische Darstellung in diskreten Zeitfenstern stündlich ermittelt wurden. Weiterhin wurde eine LDA-Implementierung angewandt, die eine fixe Anzahl von 15 Themen pro Zeitfenster nutzte. Auf jedes Zeitfenster wurde die Themenextraktion separat angewandt. Die Ähnlichkeit zwischen den Zeitintervallen wurde mittels Kosinus-Ähnlichkeit berechnet.

Ein zweiter Ansatz, der die Analyse von Trends und Themen in Twitter anhand von graphischer Datenaufbereitung analysiert, stammt von [Sopan et al., 2012]. Das Ziel der Autoren war eine Analyse von Personenbeteiligungen an akademischen Konferenzen über soziale Medien, vorrangig Twitter. Die Identifizierung der Themen über die Zeit wurde dabei hauptsächlich durch eine Analyse der Hashtags umgesetzt. Dabei wurde der Verlauf einzelner Tweets mit konkreten, vorher festgelegten Hashtags über die Zeit abgebildet. Zusätzlich wurden die Anteile an Tweets, die von konkreten Nutzern stammen, gegen alle im Zeitbereich vorkommenden Tweets dargestellt.

Ein weiterer Ansatz war Topics Over Time von [Wang and McCallum, 2006]. Die Autoren nannten als generellen Nachteil des LDA-Ansatzes eine fehlende zeitliche Betrachtung. Dieses Problem tritt allgemein auf, wenn LDA zur Themenextraktion genutzt wird. Bei der Anwendung von LDA auf einen Datensatz, der einen großen zeitlichen Bereich abdeckt, können Terme zu Themen zugeordnet werden, die zeitlich in keinem Zusammenhang stehen. Somit werden Themen beschrieben, deren Wörter aus unterschiedlichen zeitlichen Bereichen stammen und somit nicht zusammengehören. Die Autoren lösten das Problem nicht durch eine Diskretisierung der Zeit. Sie nutzten eine kontinuierliche Zeitverteilung und zur Themenextraktion wurde der Zeitstempel des Dokuments als Parameter einbezogen. Durch dieses Vorgehen ist es den Autoren gelungen, die Themenabdeckung über die Zeit exakter zu gestalten. Ist ein Thema mit einem historischen Ereignis assoziiert, dann ist dieses Thema nur zu dem jeweiligen Zeitpunkt an Dokumenten beteiligt.

Zusätzlich existieren Ansätze, die eine Änderung an den zugrundeliegenden Techniken zur Themenextraktion beschreiben. [Cai et al., 2014] nutzten MLDA, einen neuen Ansatz, der Ungenauigkeiten der Themenextraktion in den kurzen Texten der sozialen Netzen minimieren soll. Dieser Ansatz bezieht neben den Textdaten die Beziehung von Dokumenten untereinander, Hashtags und Autorenbeziehungen mit in die Themenextraktion ein. Als Basis wird LDA genutzt. Darauf aufbauend wird die Themenevolution in

diskreten Zeitfenstern betrachtet. Als Ähnlichkeitsmaß nutzten die Autoren die Kullback-Leibler-Divergenz. In [Ramage et al., 2009] wurde unter anderem die Nutzung von Tags in Kombination mit LDA beschrieben.

[Liu et al., 2009] entwickelten einen Ansatz zur Themenextraktion, der neben der Textdatenanalyse auch die Eigenschaften des Netzwerkes mit einbezieht. Für den Vergleich von zwei Termverteilungen nutzten die Autoren die Kullback-Leibler-Divergenz. Die Hellinger-Distanz wurde u.a. in [Blei and Lafferty, 2009] zum Vergleich von generierten Dokumenten benutzt.

[Yan et al., 2013] beschrieben das Biterm Topic Model für kurze Texte. Das Ziel war das Vermeiden von Fehlern bei der Themenextraktion zu kurzer Texte. Zur Textgenerierung wurden sogenannte Term co-occurences (das gemeinsame Auftreten von Wörtern) genutzt. Die Neuerung bestand darin, dass diese Termkombinationen nicht auf Dokumentenebene gesucht werden, sondern in der gesamten Dokumentensammlung. Zur Berechnung der Ähnlichkeit von Themen bzw. von Erzeugten Dokumenten aus dem Generierungsprozesse, nutzten die Autoren die Jensen-Shannon-Divergenz.

In der Veröffentlichung von [Abulaish, 2018] wurde unter anderem der Themenübergang zwischen zwei Themen beschrieben. Dabei können folgende Fälle auftreten:

1. Emergenz: ein Thema tritt nur einmal zu einem Zeitpunkt t_i auf
2. Persistenz: ein Thema ist in zwei aufeinanderfolgenden Zeitschritten t_i, t_{i+1} vorhanden
3. Konvergenz: Zwei Themen in Zeitschritt t_i werden im nächsten Zeitschritt t_{i+1} zusammengefasst
4. Divergenz: aus einem Thema zu Zeitpunkt t_i werden in Zeitschritt t_{i+1} zwei individuelle Themen
5. Auslöschung: Ein Thema aus Zeitschritt t_i ist in Zeitschritt t_{i+1} nicht mehr vorhanden

Neben dem Finden von ähnlichen Themen und den möglichen genannten Übergangen, beschrieben [Zhu et al., 2016] die Möglichkeit, durch die charakteristischsten Wörter einer Termverteilung die Unterthemen zu definieren. Somit können für ähnliche Themen Differenzierungen bestimmt werden, die sich in unterschiedlichen Subthemen widerspiegeln.

Eine anwendungsorientierte Betrachtung des Problems wurde u.a. in [Signorini et al., 2011] beschrieben. Darin versuchten die Autoren durch die Analyse von Twitterdaten einerseits die öffentliche Stimmung in Bezug auf H1N1 (Schweinegrippe) zu verfolgen, andererseits die Ausbreitung der Krankheiten selbst abzuschätzen. Dabei wurden Tweets analysiert und die Inhalte und Stimmungen über eine Zeitspanne hinweg ver-

folgt. In [Lau et al., 2012] wurde versucht, Trends und Themen über eine definierte Zeit zu identifizieren. Das besondere hierbei war eine eingebaute Updatefunktion für neue Daten, damit der Ansatz kontinuierlich angewandt werden konnte. Somit sollten fortlaufend neue Trends und Themen identifiziert werden.

3 Daten und Methoden

In diesem Kapitel werden die zugrunde liegenden Daten und die angewandten Methoden vorgestellt. Einleitend folgt eine Vorstellung des sozialen Netzwerkes Twitter und die damit verbundene Datengewinnung. Anschließend werden alle vorhandenen Daten zusammengefasst und graphisch aufbereitet. Abschließend folgen die verwendeten Methoden für die Themenbeschreibung und -dynamik.

3.1 Der Kurznachrichtendienst Twitter

Twitter ist ein Mikrobloggingdienst aus dem Jahr 2006[1]. Aktuell belaufen sich die Mitgliedszahlen auf 192 Millionen täglich aktiver Nutzer [Twitter, 2021]. Es können sogenannte **Tweets** veröffentlicht werden, die die jeweilige Kurznachricht beinhalten. Dabei existiert eine maximale Zeichenanzahl von 280 Symbolen. Vor dem Jahr 2017 waren lediglich 140 Symbole zulässig. [Kuri, 2017]

Über die Funktion des **Folgens** können Nutzer die Aktivitäten anderer Mitglieder nachvollziehen. Über sogenannte **Retweets** können Nutzer die Tweets von anderen Personen teilen und diese somit den eigenen Folgenden präsentieren.

Jeder Nutzer kann auf Posts anderer Personen reagieren, indem er sie liked, retweeted oder kommentiert. Eine weitere Besonderheit in Tweets ist die Verwendung von **Hashtags**. Dabei wird ein Begriff mittels eines Doppelkreuzes (#) hervorgehoben und als relevant gekennzeichnet. Weiterhin können direkt Personen über sogenannte **Mentions** angesprochen werden, indem ein @-Symbol vor den entsprechenden Namen im Tweet gesetzt wird.

Jedem Tweet wird eine eindeutige **Tweet-ID** zugewiesen, was eine spätere Identifizierung ermöglicht. Über die **Twitter-API** können die Inhalte des Kurznachrichtendienstes beispielsweise für Bildungs- und Forschungszwecke genutzt werden. Durch deren Nutzung stehen unter anderem folgende Informationen pro Tweet zur Verfügung:

1. Tweet-ID
2. Zeitstempel
3. Autoren-ID
4. Conversation-ID
5. Nachricht des Tweets

[1] https://twitter.com/jack/status/20

6. verwendete Hashtags
7. verwendete Urls
8. erwähnte Personen
9. Anzahl der Likes, Retweets und Kommentare
10. Mediakeys (Foto, Video oder Gif)
11. ...

Nachfolgend umfasst der Begriff Tweet sowohl Tweets als auch Retweets. In Fällen, in denen explizit die Retweets relevant sind, wird darauf hingewiesen.

3.2 Datenüberblick und Visualisierungen

Die Datengrundlage bildeten die Tweets der Profile der Parteien des deutschen Bundestages dar. Nachfolgend wurden CDU/CSU, AfD und Die Linke als Querschnitt aller Parteien betrachtet und vorgestellt. Zum Vergleich wurden die gleichen Parteien aus den Landtagen Niedersachsen und Sachsen-Anhalt betrachtet. Zusätzlich wurde die Tagesschau als neutrales und unparteiisches Profil untersucht.

Die Profile der Parteien und deren Tweets wurden auf Grund der weitestgehend gleichbleibenden Inhalte durch die politische Orientierung genutzt. Auch die Charakteristiken der Posts stellten durch die hohe öffentliche Aufmerksamkeit eine annähernd konstante Größe dar. Die Parteien veröffentlichten über den betrachteten Zeitraum regelmäßig Inhalte, die durch ihre jeweiligen Standpunkte geprägt wurden. Somit konnte die Annahme getroffen werden, dass behandelte Inhalte und Themen wiederkehrend aufgegriffen wurden.

Die Verwendung von Tweets der Tagesschau diente unter anderem zur Ermittlung von Abhängigkeiten der Profile. Dadurch konnte untersucht werden, ob Themen zuerst von Parteien oder von Nachrichtenkanälen behandelt wurden. Dafür musste die Ähnlichkeit von Themen über die Profile hinweg ermittelt werden.

3.2.1 Vorstellung der betrachteten Profile

In diesem Abschnitt werden die Charakteristika der Bundestagsparteien CDU/CSU, AfD und Die Linke sowie die der Tagesschau vorgestellt. In Anhang A.1 sind die Inhalte der Landtagsparteien zusammengestellt. Zur Vereinfachung der Visualisierung von Nutzerreaktionen, wurden nur eigene Inhalte (Tweets) der Parteien betrachtet.

CDU/CSU

Das erste betrachtete Profil war die CDU/CSU.[2] Alle Tweets lagen im Zeitraum vom 12.06.2009 bis 26.06.2021. Während dieser Zeit wurden insgesamt 28.539 Tweets gepostet. In Abbildung 3.1 ist der Verlauf aller Tweets über den gesamten Zeitraum dargestellt. Dabei war zu erkennen, dass Twitter seit der Profilerstellung regelmäßig genutzt wurde. Eine stärkere Nutzung trat ca. ab dem Jahr 2015 ein. In den Jahren 2009-2014 wurden im Schnitt 15 Tweets pro Woche veröffentlicht, ab 2015 waren es 72.

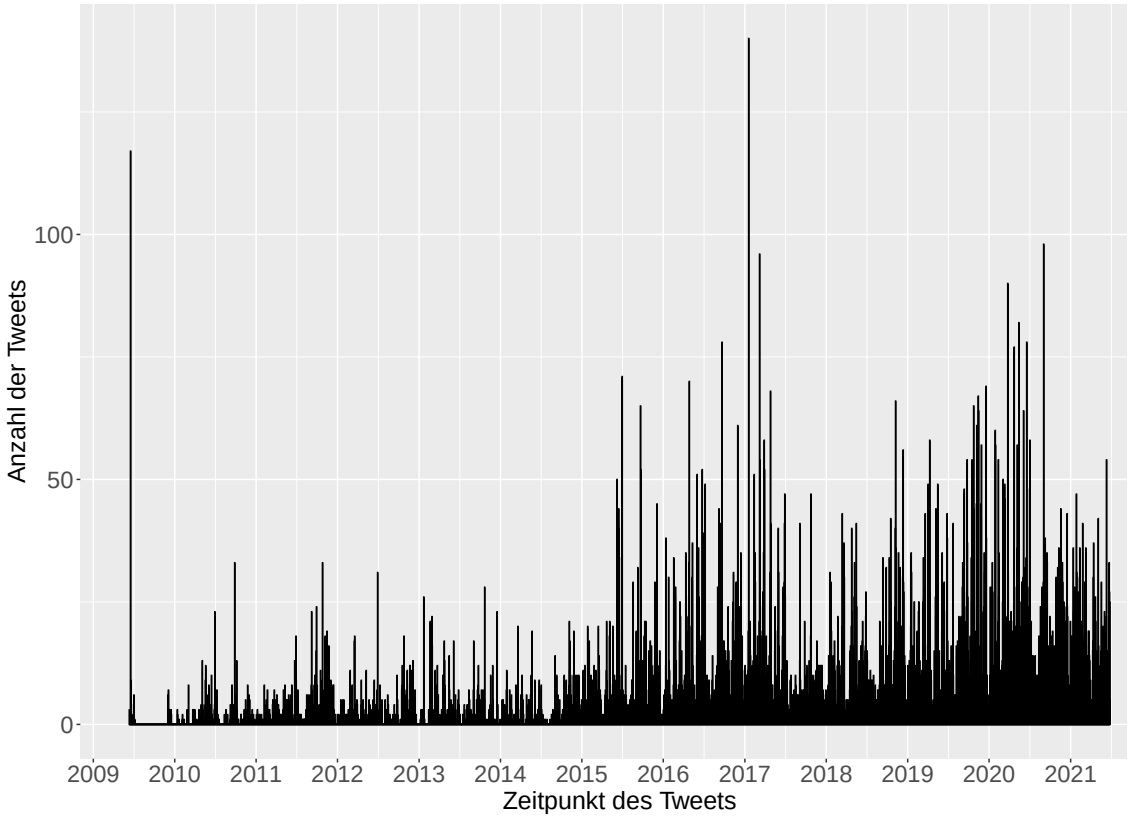

Abbildung 3.1: Darstellung der Anzahl an veröffentlichten Tweets der CDU/CSU seit 2009

[2] https://www.twitter.com/cducsubt

Dabei betrug der Anteil an Tweets 16.017 und an Retweets 10.336. Pro Woche wurden im Schnitt 48 und pro Monat 202 Tweets veröffentlicht. Die Partei verwendete am häufigsten (8.890) einen Hashtag pro Tweet, in 7.083 Tweets wurde kein Hashtag verwendet. In Abbildung 3.2 sind die 10 häufigsten Hashtags dargestellt.

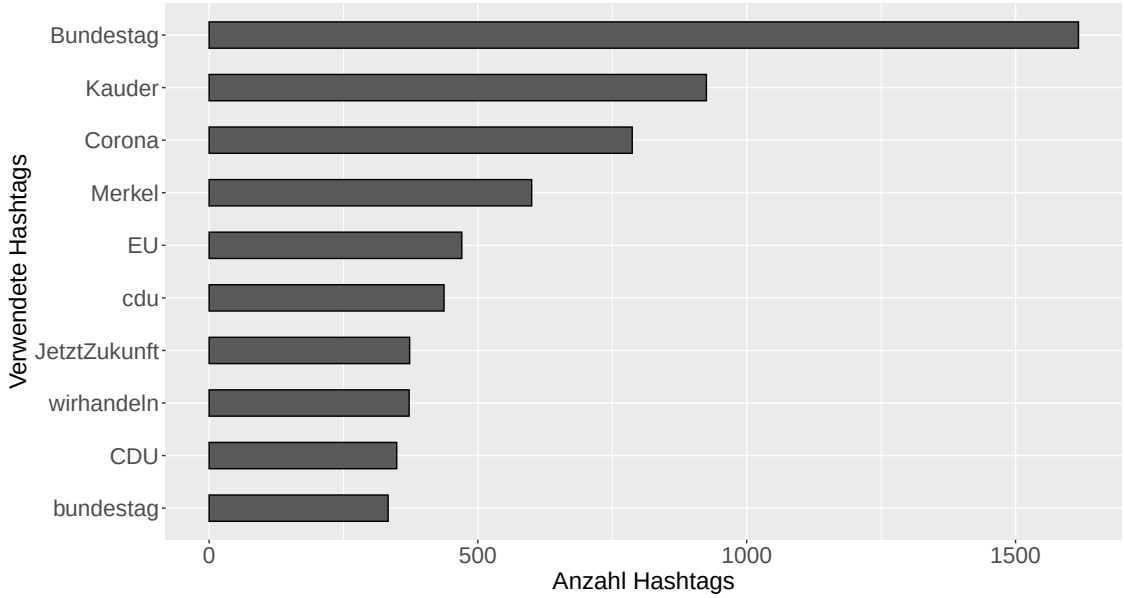

Abbildung 3.2: Darstellung der am häufigsten verwendeten Hashtags der CDU/CSU

Weiterhin sind die Nutzerreaktionen auf eigene Inhalte der Partei dargestellt. In den Abbildungen 3.3, 3.4 und 3.5 sind die Likes, Retweets und Kommentare auf alle Tweets visualisiert.

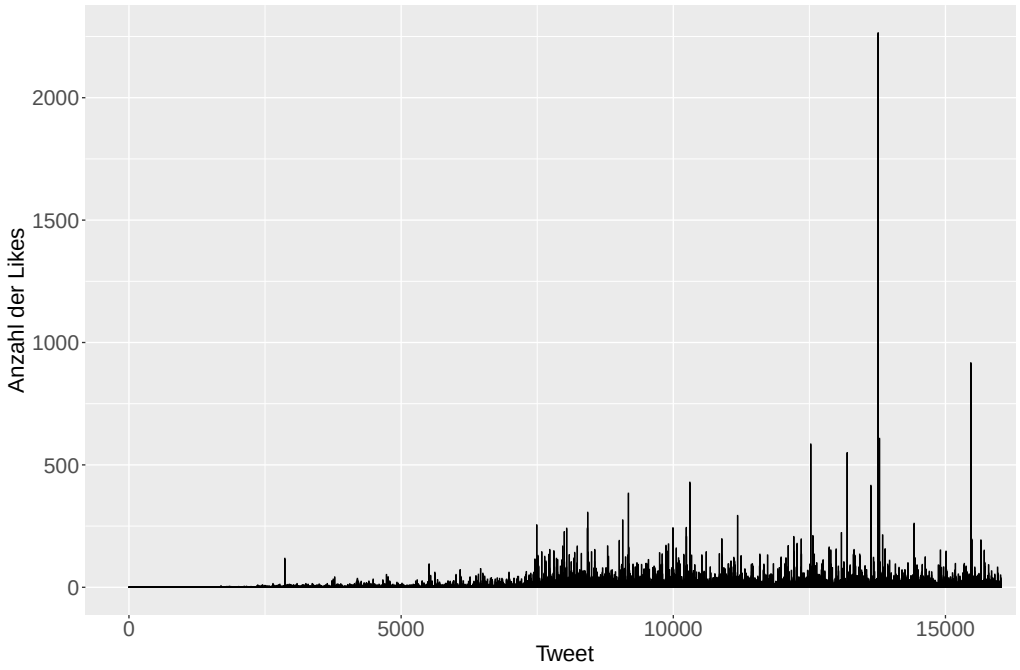

Abbildung 3.3: Darstellung der Likes als Nutzerreaktion über alle Tweets der CDU/CDU

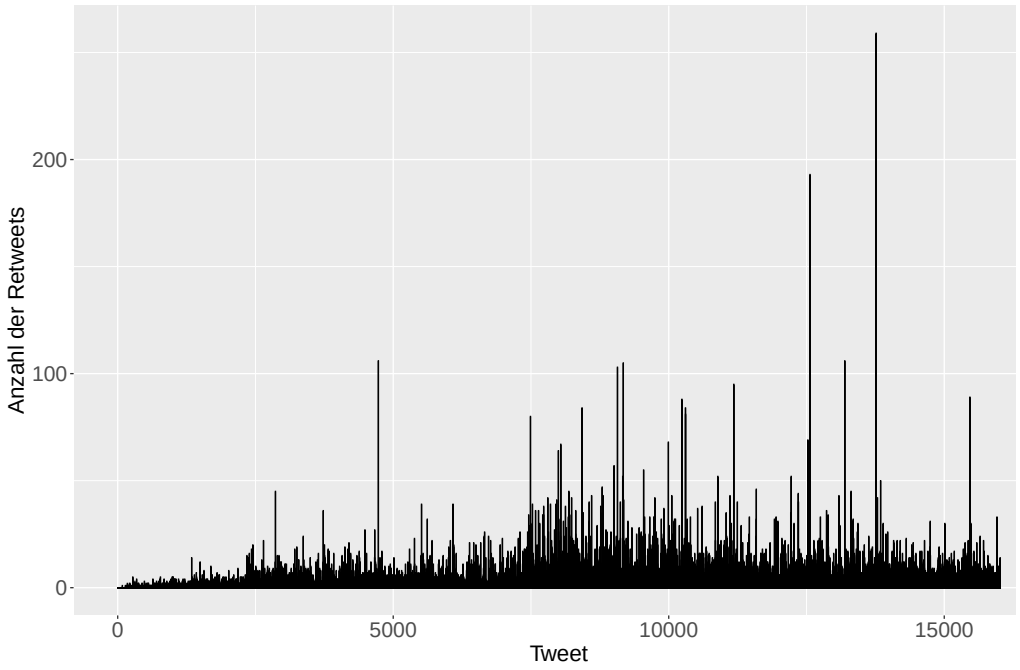

Abbildung 3.4: Darstellung der Retweets als Nutzerreaktion über alle Tweets der CDU/CDU

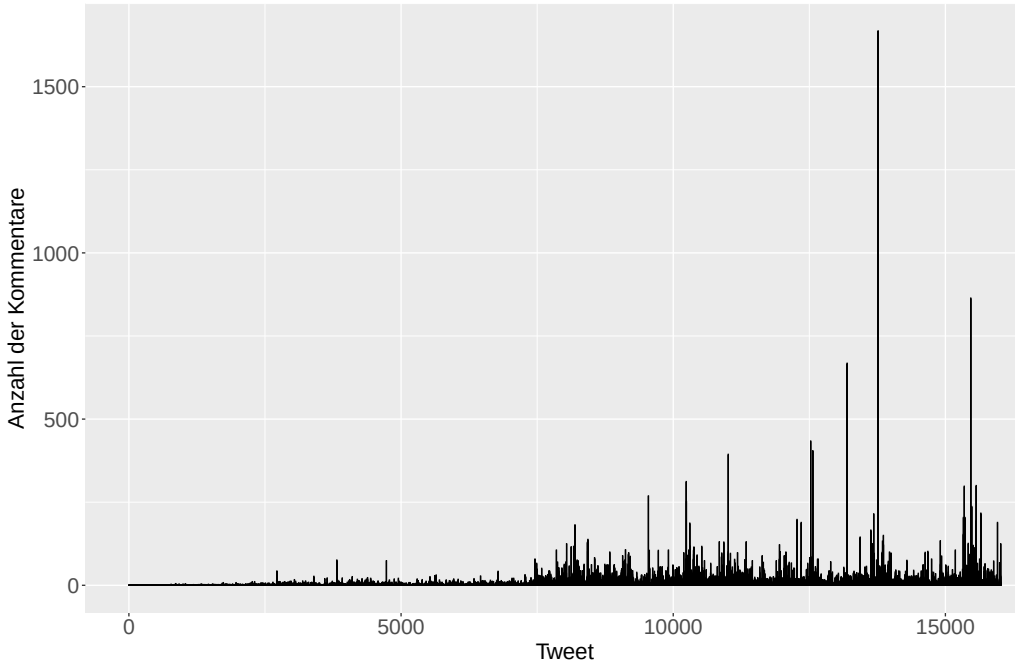

Abbildung 3.5: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der CDU/CDU

AfD

Die AfD[3] hat auf Grund der späteren Parteigründung ein kürzeres Zeitfenster, in dem sie Tweets veröffentlicht hat. Im Zeitraum vom 04.11.2012 bis 26.06.2021 hat die Partei insgesamt 3.257 Tweets gepostet. Somit hat die AfD ca. 9 mal weniger Tweets abgesetzt als die CDU/CSU. In Abbildung 3.6 ist der Verlauf der Tweets über die gesamte Zeit dargestellt. Auffällig ist dabei vor allem der kurzzeitige Anstieg in den veröffentlichten Tweets im September 2017, der inhaltlich auf die Bundestagswahl 2017 zurückzuführen ist. Eine kontinuierliche Nutzung des Twitterprofils begann im März 2020.

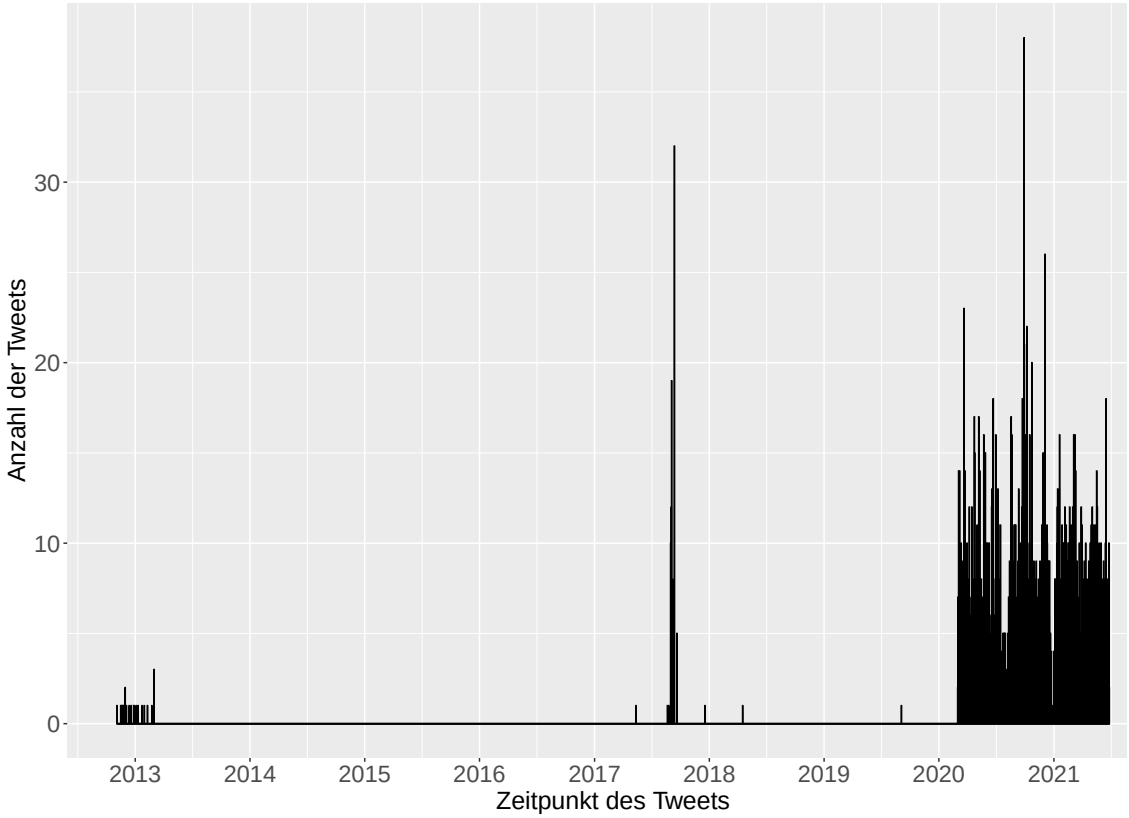

Abbildung 3.6: Darstellung der Anzahl der veröffentlichten Tweets der AfD seit 2012

Die Partei veröffentlichte durchschnittlich 34 Tweets pro Woche und 120 pro Monat. Insgesamt lag der Anteil an Tweets bei 1.267 und an Retweets bei 1.593, sodass sich Tweets und Retweets im Umfang ähneln. Die 10 häufigst verwendeten Hashtags sind in Abbildung 3.7 dargestellt. Die meisten Tweets (1.140) wurden ohne Hashtag gepostet, unter allen Tweets mit Hashtags wurde ein Tag am meisten genutzt (508).

[3] https://www.twitter.com/AfD

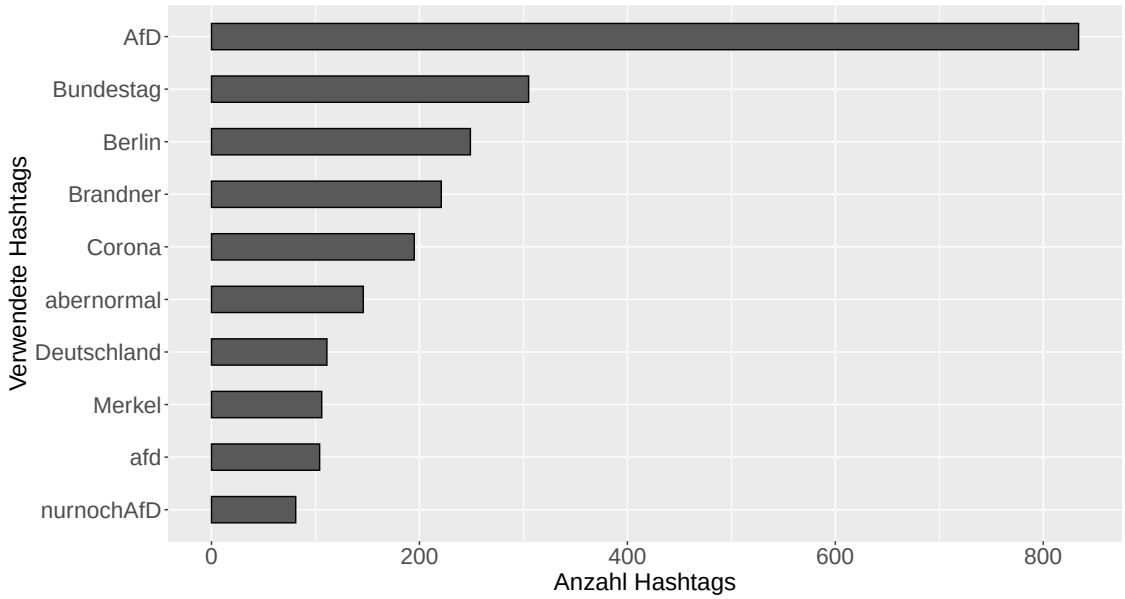

Abbildung 3.7: Darstellung der am häufigsten verwendeten Hashtags der AfD

Die Nutzerreaktionen auf die eigenen Inhalte der Partei sind in den folgenden Grafiken visualisiert. Die Likes sind in Abbildung 3.8, die Retweets in Abbildung 3.9 und die Kommentare in Abbildung 3.10 dargestellt.

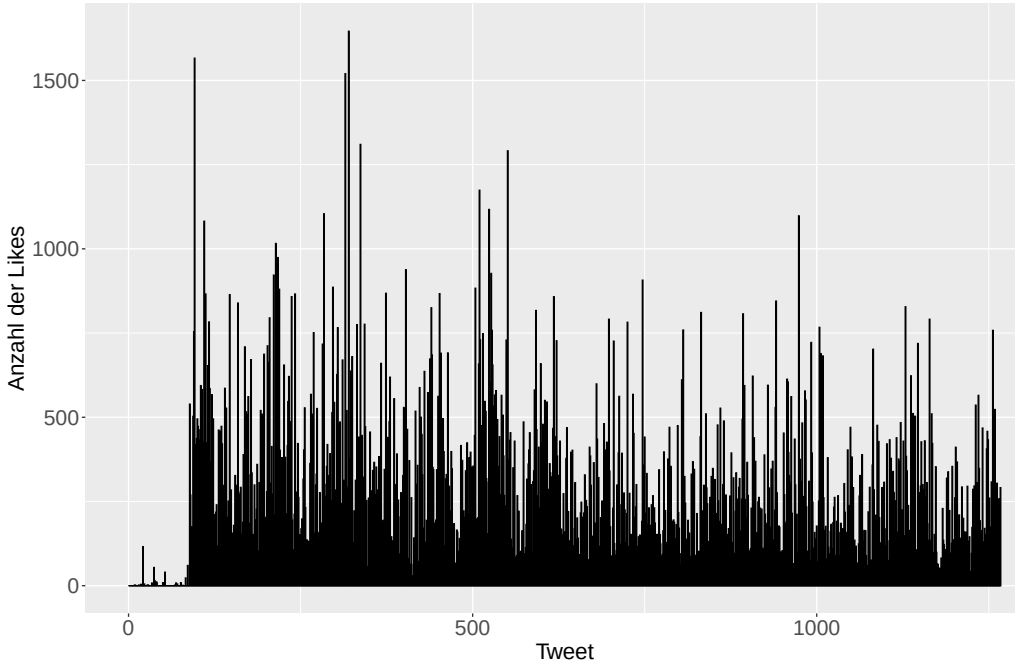

Abbildung 3.8: Darstellung der Likes als Nutzerreaktion über alle Tweets der AfD

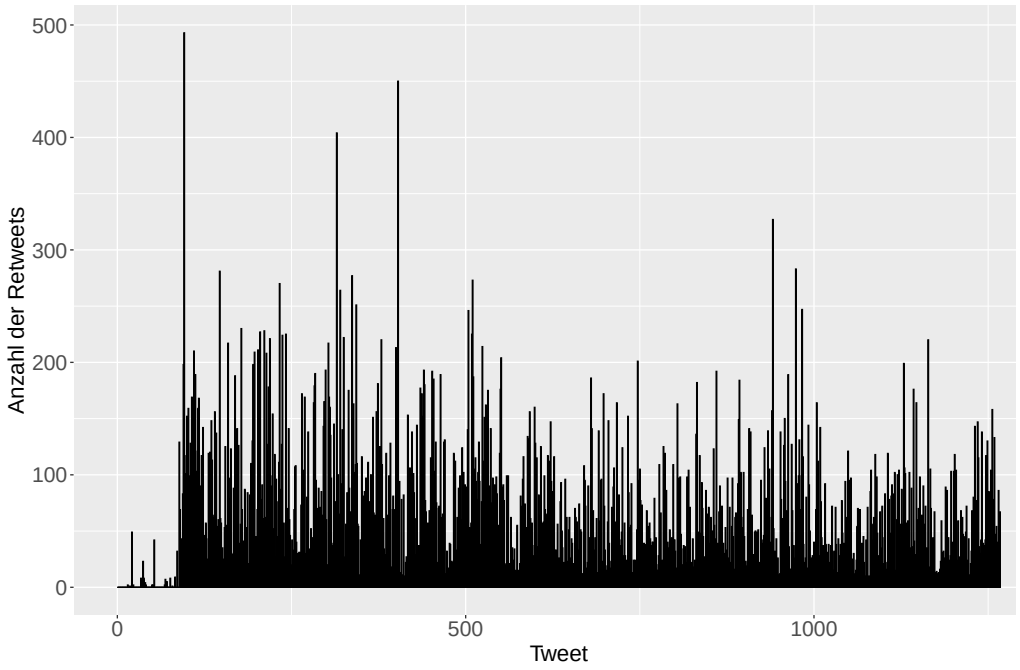

Abbildung 3.9: Darstellung der Retweets als Nutzerreaktion über alle Tweets der AfD

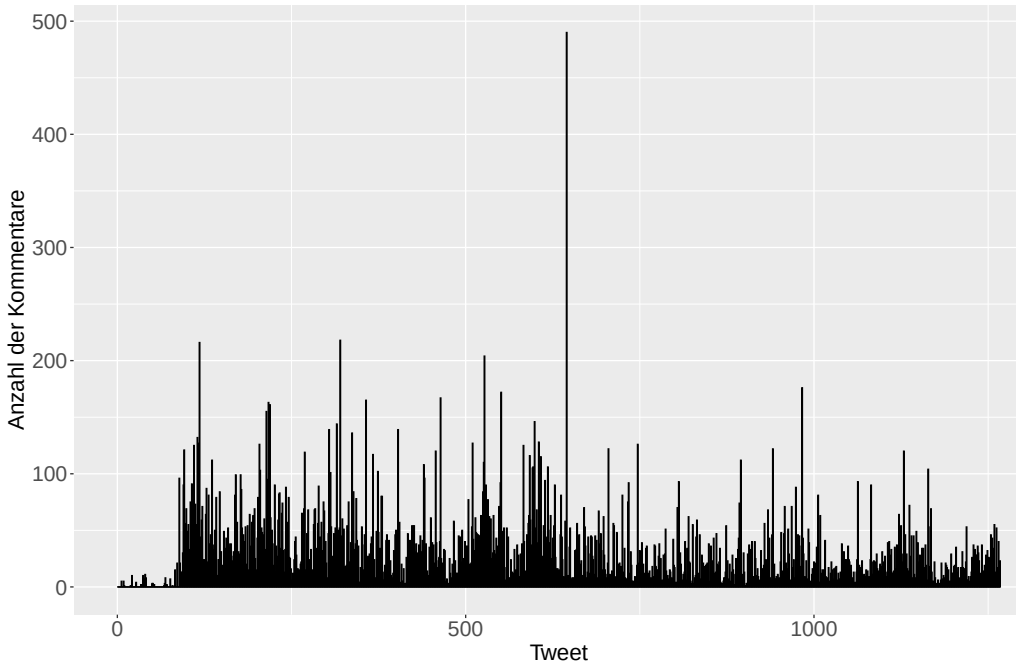

Abbildung 3.10: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der AfD

Die Linke

Die Linke[4] hat im Zeitraum vom 04.06.2009 bis 26.06.2021 32.056 Tweets veröffentlicht und somit die meisten Tweets im Vergleich aller drei Parteien. In Abbildung 3.11 sind die Tweets über die Zeit dargestellt. Auffällig ist auch hier, dass Die Linke ab 2015 durchschnittlich mehr Tweets veröffentlicht hat. Bis 2015 waren es 18 Tweets pro Woche, ab 2015 84. Hinzu kommen einzelne, sich deutlich vom Durchschnitt abhebende Peaks. Die Partei hat mehrmals zwischen 150 und 250 Tweets pro Tag abgesetzt.

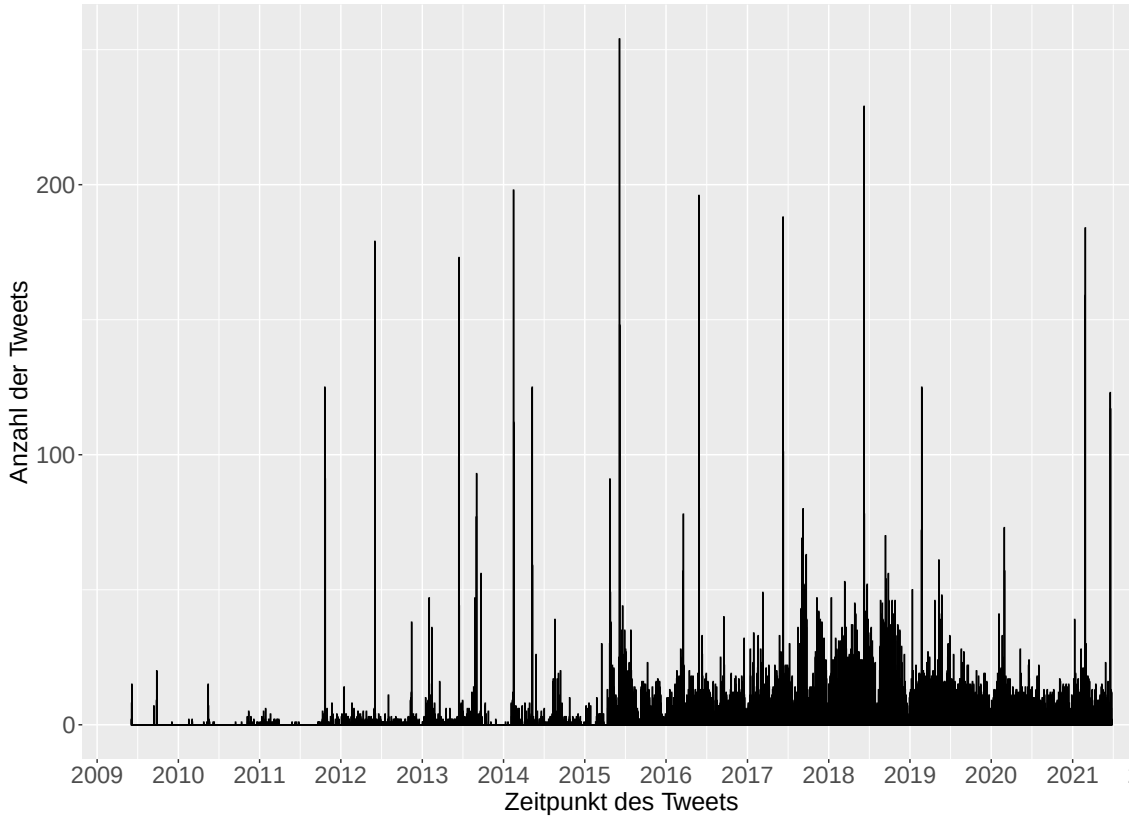

Abbildung 3.11: Darstellung der Anzahl der veröffentlichten Tweets der Linken seit 2009

Im Durchschnitt hat Die Linke pro Woche 60 und pro Monat 239 Tweets gepostet. Die Anzahl an Tweets setzte sich aus 9.523 Tweets und 19.769 Retweets zusammen, sodass der Großteil aus geteiltem Inhalt anderer Profile besteht. In Abbildung 3.12 sind die zehn häufigsten Hashtags zusammengestellt. Die Partei verwendete am häufigsten einen Hashtag pro Tweet (8.749), 6.984 Tweets wurden ohne Hashtag veröffentlicht.

[4] https://twitter.com/dieLinke

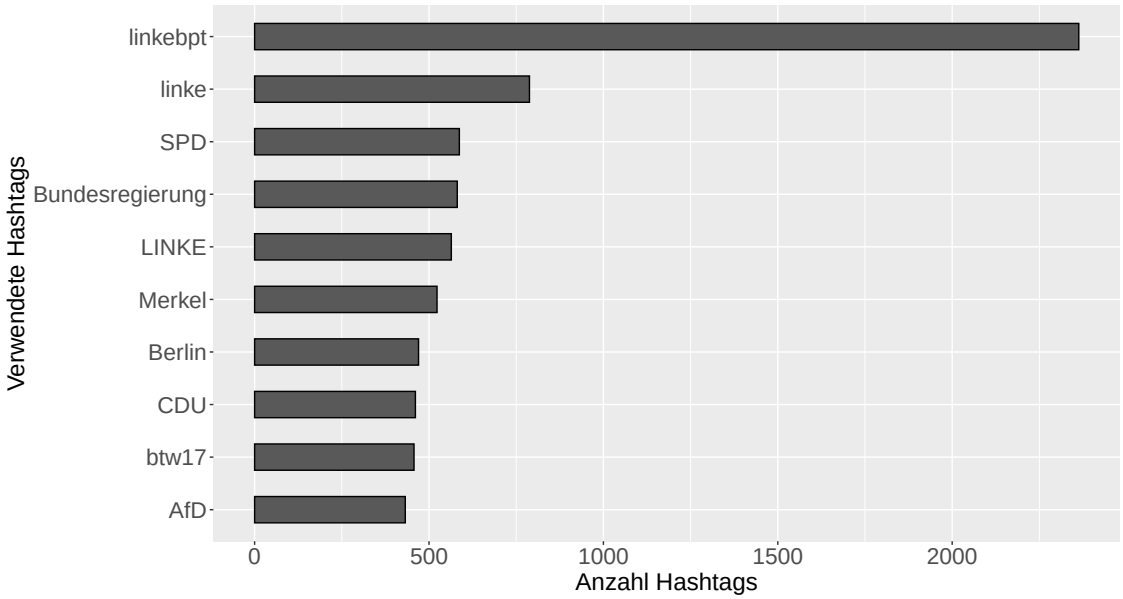

Abbildung 3.12: Darstellung der am häufigsten verwendeten Hashtags der Linken

In den Abbildungen 3.13, 3.14 und 3.15 sind die Nutzerreaktionen auf die eigenen Inhalte der Partei dargestellt. Dabei sind Likes, Retweets und Kommentare visualisiert.

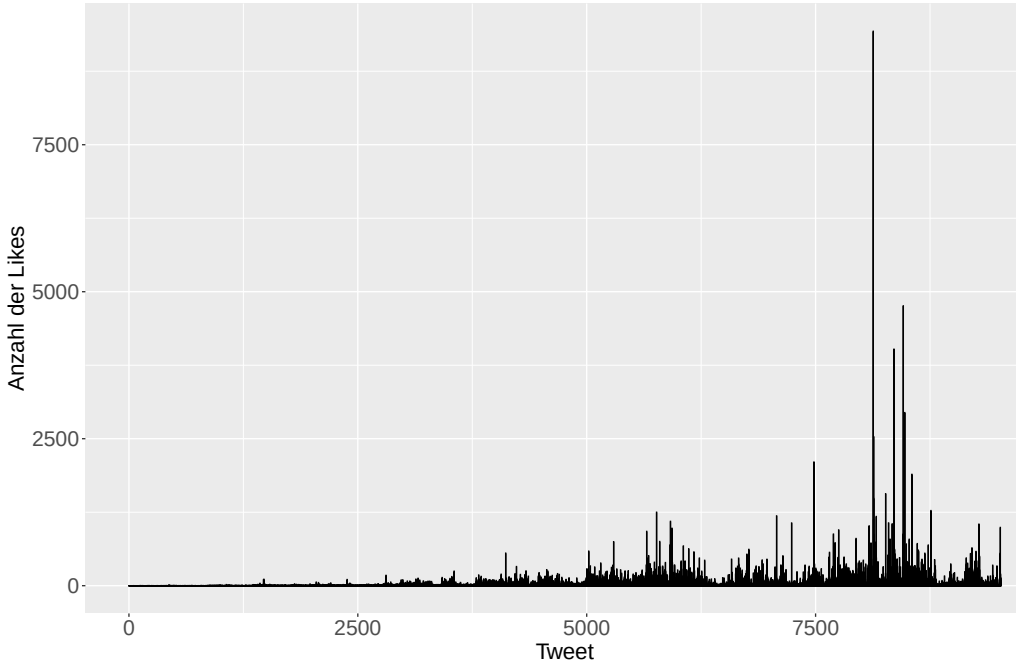

Abbildung 3.13: Darstellung der Likes als Nutzerreaktion über alle Tweets der Linken

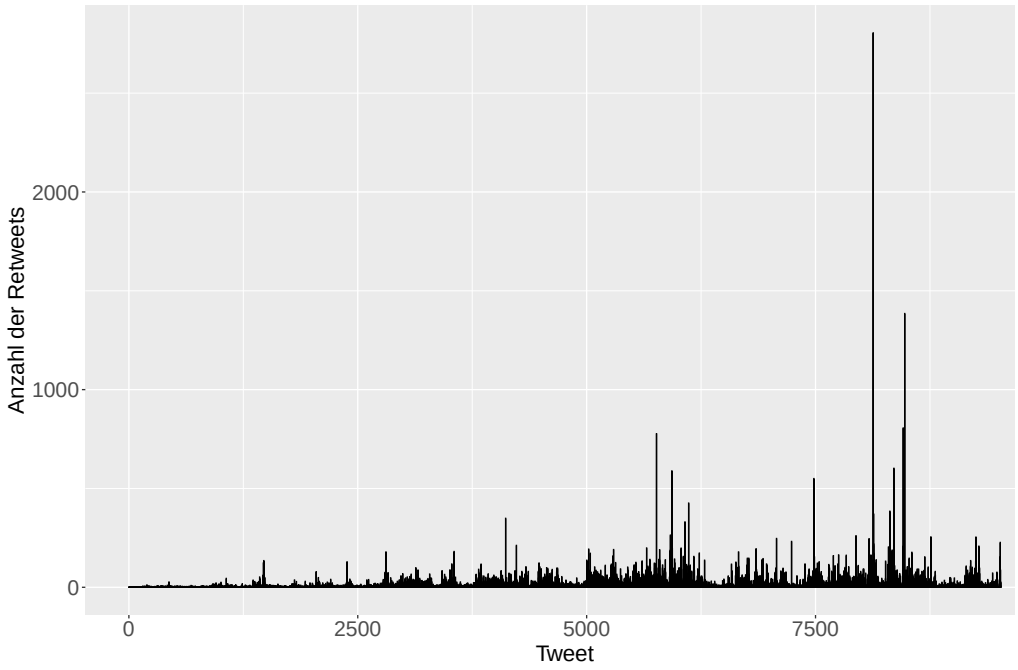

Abbildung 3.14: Darstellung der Retweets als Nutzerreaktion über alle Tweets der Linken

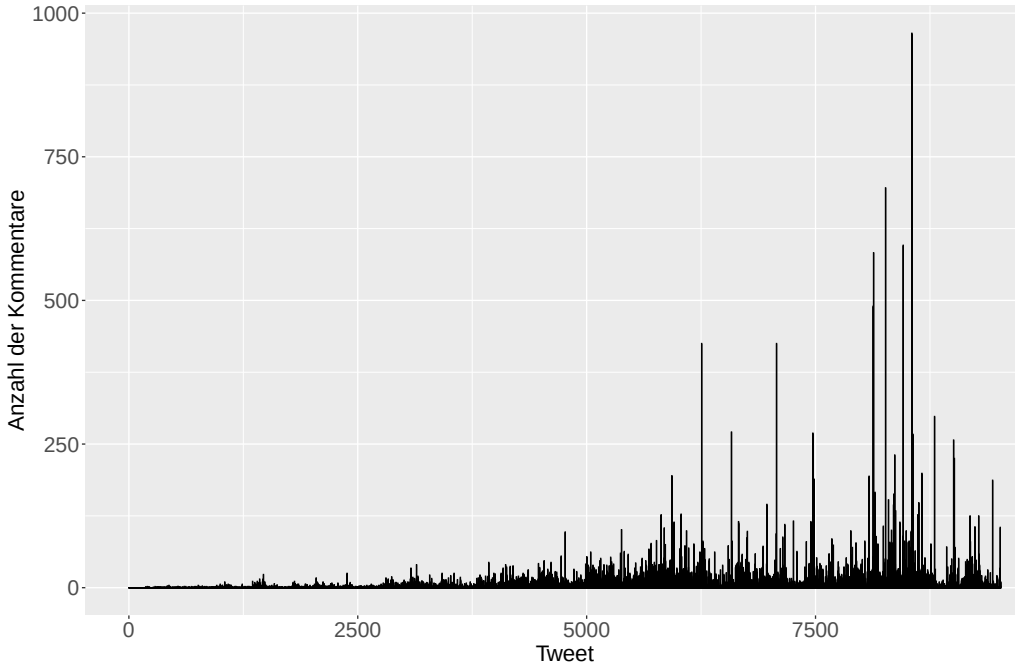

Abbildung 3.15: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der Linken

Die Tagesschau

Die Tagesschau[5] stellt das letzte Twitterprofil dar. Insgesamt wurden über den Zeitraum vom 03.05.2007 bis 26.06.2021 191.868 Tweets gepostet. Somit hat die Tagesschau den längsten Zeitraum der Nutzung und die meisten Tweets veröffentlicht. In Abbildung 3.16 sind die Tweets gegen die Zeit aufgetragen.

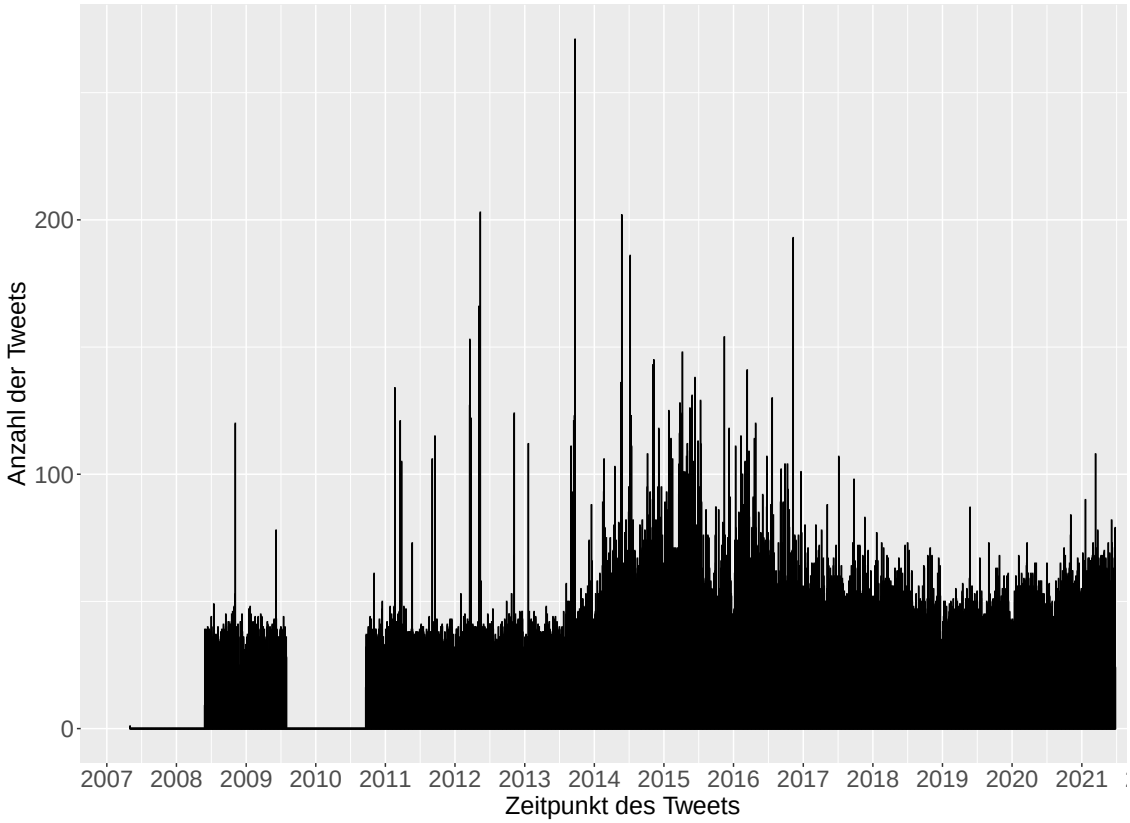

Abbildung 3.16: Darstellung der Anzahl der veröffentlichten Tweets der Tagesschau seit 2007

Die Tagesschau hat eine deutlich höhere Anzahl an Tweets pro Woche mit 302 und pro Monat mit 1.314. Auch das Verhältnis von Tweets zu Retweets unterschied sich von dem der Parteien. Die Anzahl an Tweets mit 168.399 überstieg deutlich die Anzahl an Retweets mit 21.470. Die Tagesschau verwendete am häufigsten zwei Hashtags pro Tweet (57.309), die Anzahl an Posts ohne Hashtags betrug 54.349. In Abbildung 3.17 sind die zehn häufigsten Hashtags dargestellt.

[5] https://twitter.com/Tagesschau

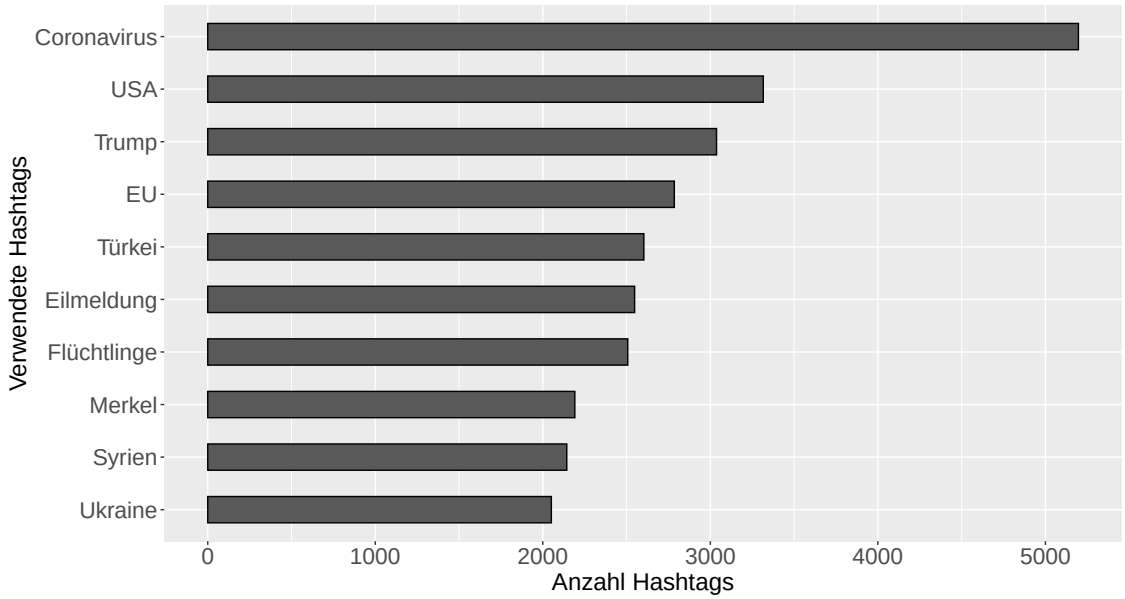

Abbildung 3.17: Darstellung der am häufigsten verwendeten Hashtags der Tagesschau

Bei der Tagesschau sind abschließend die Reaktionen der Nutzer auf die eigenen Inhalte des Profils visualisiert. In den Abbildungen 3.18, 3.19 und 3.20 sind die Likes, Retweets und Kommentare dargestellt.

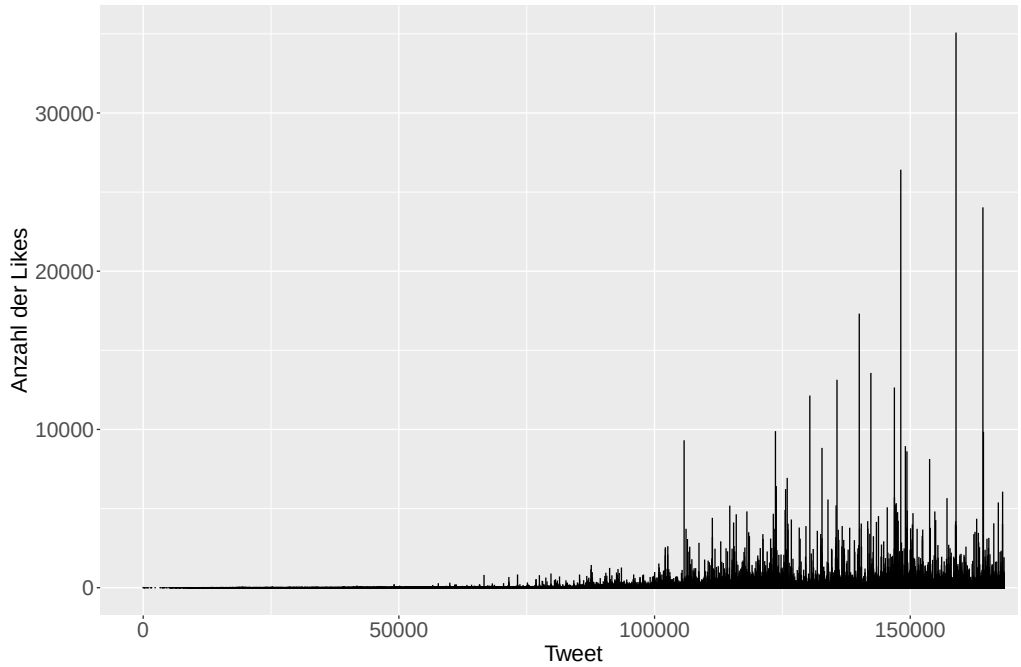

Abbildung 3.18: Darstellung der Likes als Nutzerreaktion über alle Tweets der Tagesschau

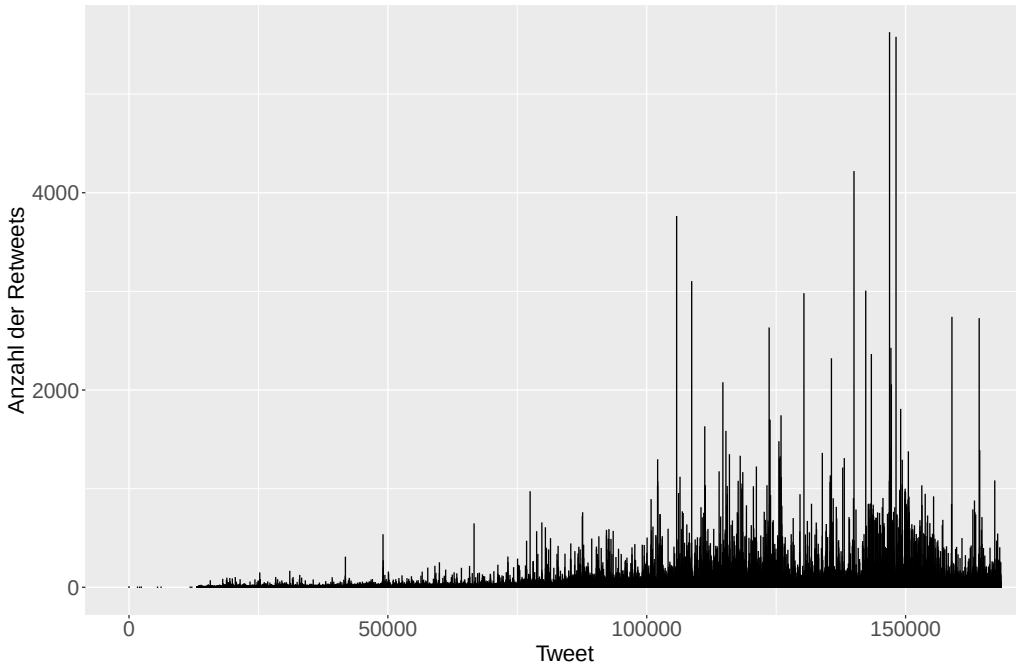

Abbildung 3.19: Darstellung der Retweets als Nutzerreaktion über alle Tweets der Tagesschau

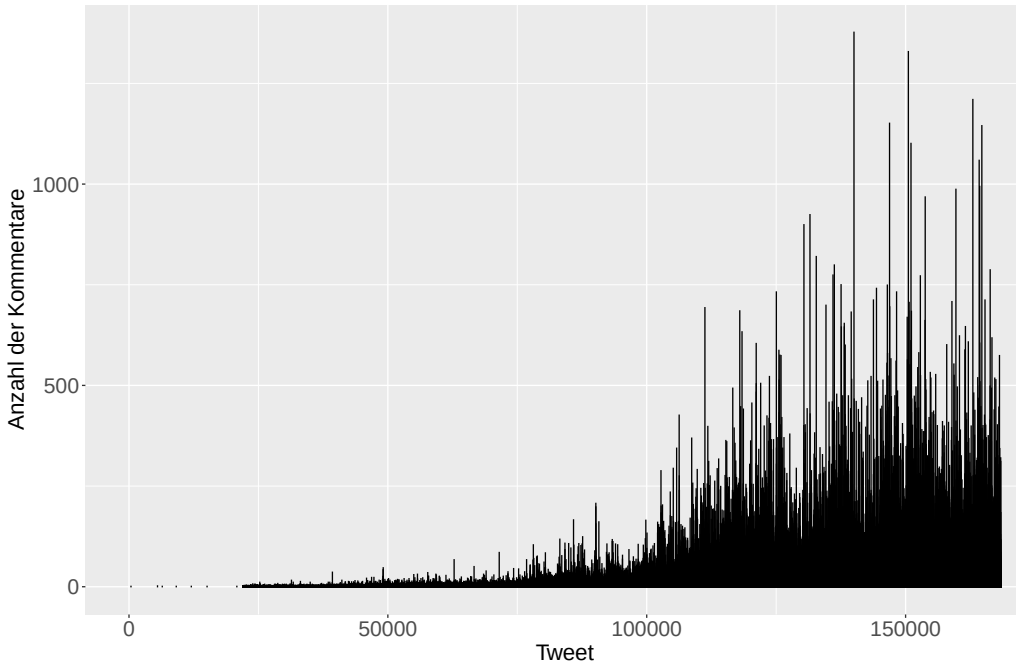

Abbildung 3.20: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der Tagesschau

3.2.2 Vergleich der Profile

Die Profile der Parteien und der Tagesschau stellten die Grundlage für die Untersuchung der Themendynamik und -evolution dar. Die Eigenschaften der jeweiligen Profile und deren Texte der Tweets haben Einfluss auf das Ergebnis. Deshalb sind in der nachfolgenden Tabelle 3.1 alle Informationen der Parteien zusammengefasst. Die Diskrepanz zwischen der Gesamtzahl der Posts und der Summe von Tweets und Retweets ist dadurch zu erklären, dass neben Tweets und Retweets weitere Möglichkeiten wie bspw. direkte Antworten (Replies) existieren.

Tabelle 3.1: Zusammenfassung der Eigenschaften der betrachteten Profile

Profil	Beginn	Ende	# Tweets	# Woche	# Monat	Tweets/ Retweets
CDU/CSU	12.06.09	26.06.21	28.539	48	202	16.017/10.336
AfD	04.11.12	26.06.21	3.257	34	120	1.267/1.593
Die Linke	04.06.09	26.06.21	32.056	60	239	9.523/19.769
Tages-schau	03.05.07	26.06.21	191.868	302	1.314	168.399/21.470

Ein deutlicher Unterschied zwischen der CDU/CSU und der Linken ist der Anteil der eigenen Inhalte. Die CDU/CSU veröffentlichte mehr Tweets, Die Linke deutlich mehr Retweets. Bei der AfD verhielten sich beide Maßzahlen ungefähr gleich. Das jüngste Profil stammt von der AfD, sie haben am wenigsten Inhalte veröffentlicht. Die Tagesschau hat die meisten Tweets und zugleich die meisten eigenen Inhalte veröffentlicht.

Für eine Extraktion der Themen aus Texten spielt der Umfang der zugrundeliegenden Daten eine wichtige Rolle. Deshalb sind in Tabelle 3.2 die durchschnittliche Anzahl an Wörtern pro Tweet zusammengefasst. Dabei wurde zusätzlich in Tweets und Retweets unterschieden. In Hinblick auf die zur Verfügung stehende Anzahl an Symbolen, ist weiterhin die verwendete Symbolanzahl in der Tabelle dargestellt.

Tabelle 3.2: Durchschnittliche Anzahl an Wörtern und Symbolen pro Profil

Profil	Gesamt Wörter	Tweets Wörter	Retweets Wörter	Gesamt Symbole	Tweets Symbole	Retweets Symbole
CDU/CSU	20,8	19,0	23,6	173,7	163,7	191,1
AfD	25,6	21,2	29,4	221,6	191,8	254,7
Die Linke	22,3	16,5	24,7	181,5	159,0	200,8
Tagesschau	10,4	9,5	17,2	98,3	93,1	138,1

Über alle Profile hinweg ist zu erkennen, dass die Anzahl an Wörtern und die Menge an Zeichen in Retweets höher ist. Die Tagesschau unterscheidet sich durch eine deutlich geringere Anzahl an Termen pro Tweet. Unter den Parteien sind die meisten Terme und Zeichen bei der AfD zu finden.

3.3 Angewandte Methoden und Vorgehen

Nachfolgend wird das Vorgehen von der Textvorverarbeitung bis zur Beschreibung der Themendynamik erläutert.

3.3.1 Vorverarbeitung der Texte

Ein notwendiger Schritt für die Analyse von Textdaten ist die Vorverarbeitung der Texte. In der nachfolgenden Aufzählung sind die wesentlichsten Schritte zusammengefasst:

1. alle Buchstaben in Kleinschreibung umgewandelt
2. alle Umlaute entfernt (zur Vermeidung von Termdifferenzierungen durch unterschiedliche Schreibweisen)
3. alle Urls wurden entfernt
4. alle Symbole, die keine Buchstaben sind, wurden entfernt (Satzzeichen, Zahlen etc.)
5. alle Terme, die weniger als 3 Symbole besitzen, wurden entfernt
6. alle Füll- bzw. Stoppwörter wurden entfernt
7. Tweets, die nach der Vorverarbeitung keinen Inhalt besaßen (z.B. nur geteilte Links) wurden entfernt

In Tabelle 3.3 sind die Anzahlen an Wörtern und Symbolen nach den Vorverarbeitungsschritten zusammengefasst. Im Vergleich zu Tabelle 3.2 ist der deutlich reduzierte Umfang zu erkennen. Die Anzahl an Termen hat bei den Parteien um ca. die Hälfte abgenommen. Der Umfang der Tagesschau hat weniger stark abgenommen, wobei die Tagesschau von vornherein weniger Terme pro Tweet benutzte.

Tabelle 3.3: Durchschnittliche Anzahl an Wörtern und Symbolen pro Profil nach Anwendung aller Vorverarbeitungsschritte

Profil	Gesamt Wörter	Tweets Wörter	Retweets Wörter	Gesamt Symbole	Tweets Symbole	Retweets Symbole
CDU/CSU	9,9	9,3	11,2	98,6	93,1	109,0
AfD	12,4	10,6	14,4	120,5	101,0	140,6
Die Linke	10,5	9,3	11,6	102,2	91,4	112,6
Tagesschau	6,2	5,9	8,6	57,3	54,7	77,6

3.3.2 Themenextraktion mittels LDA

Die Themen wurden mittels LDA aus den Tweets extrahiert. Dafür wurden verschiedene Parameterkonfigurationen gewählt und der Ansatz auf unterschiedlichen Teilmengen der Daten angewandt. Grundsätzlich stellen Tweets einen Grenzfall für die Themenextraktion dar, da der Umfang an Wörtern sehr gering ist. Sind zu wenig Terme vorhanden, kann die LDA nicht zuverlässig Themen bestimmten. Das Problem wurde durch eine Gruppierung der Tweets umgangen, das genaue Vorgehen ist in Abschnitt 3.3.3 beschrieben.

Die Textdaten wurden auf zwei verschiedene Arten vorverarbeitet. Nach dem Anwenden der Schritte aus Abschnitt 3.3.1, wurden einerseits nur Terme, andererseits Terme und Bigramme zur Themenextraktion genutzt. Bigramme können durch die Information, welche Terme häufig mit anderen zusammen auftreten, eine höhere Genauigkeit in den Themenbeschreibungen liefern [Jurafsky and Martin, 2020].

Damit die Themen keine homogene Verteilung pro Dokument, sondern eine differenzierte Verteilung besitzen, wurde $\alpha = 0.3$ gewählt. Das gleiche Verhalten sollte für die Termverteilung pro Thema gelten, weshalb $\beta = 0.1$ genutzt wurde. Zur Schätzung der a-posteriori Wahrscheinlichkeiten von Term- und Themenverteilungen wurde das Collapsed-Gibbs-Sampling angewandt.

Die Anzahl an Themen wurde für jede Extraktion individuell bestimmt. Dabei wurde lediglich das Maximum der zu bestimmenden Anzahl an Themen festgelegt. In Abhängigkeit des gewählten Zeitfensters der Gruppierung (siehe 3.3.3) wurden für Wochen maximal 15 und für Monate maximal 50 Themen bestimmt. Zur genauen Bestimmung der Themenanzahl wurden verschiedene Optimierungsfunktionen genutzt [Arun et al., 2010, Cao et al., 2009, Deveaud et al., 2014, Griffiths and Steyvers, 2004]. Für jede Funktion wurde das Optimum bestimmt. Aus den Ergebnissen wurde über den Median die resultierende Themenanzahl gewählt.

Weiterhin wurde die Extraktion sowohl auf dem Datensatz der eigenen Posts (Tweets) als auch auf den Retweets separat durchgeführt. Darüber sollten Unterschiede zwischen den eigenen Inhalten und den geteilten Inhalten deutlich werden. Unter dem Gesichtspunkt, dass nur dann Inhalte als Retweets geteilt werden, wenn diese mit der allgemeinen Meinung des jeweiligen Profils harmonieren, wurde die Themenextraktion auch auf der Kombination beider Datensätze durchgeführt. In allen Datensätzen sind die verwendeten Hashtags Teil der Tweets und werden mit zur Themenextraktion genutzt.

3.3.3 Methodisches Vorgehen zur Beschreibung der Themendynamik

Zeitintervalle

Damit die Dynamik und Evolution von Themen beurteilt werden kann, muss die Themenextraktion Ergebnisse zu verschiedenen Zeitpunkten liefern. Dieser Umstand, in Kombination mit dem Problem zu weniger Daten, hat zur diskreten Zeiteinteilung aller Daten in Wochen und Monate geführt. Dabei wurden alle Daten eines Profils, die in einer Woche bzw. einem Monat lagen, zu einer Dokumentensammlung zusammengefasst. Mittels LDA wurden auf jeder Sammlung separat Themen extrahiert.

Die Erprobung von Wochen und Monaten adressiert einerseits eine genauere Beschreibung und Verfolgung von Themen, andererseits das Zusammenfügen einer ausreichenden Menge an Tweets, damit zuverlässig Themen gefunden werden. In beiden Fällen entstand eine Art Zeitstrahl, dessen Ursprung das Datum der Profilerstellung war. Zu jedem diskreten Zeitpunkt t_i sind alle Themen zusammengefasst, die in diesem Zeitabschnitt gefunden wurden. In Abbildung 3.21 ist dieses Prinzip schematisch dargestellt. Dabei stellt n_{max} die maximale Themenanzahl pro Zeitabschnitt dar.

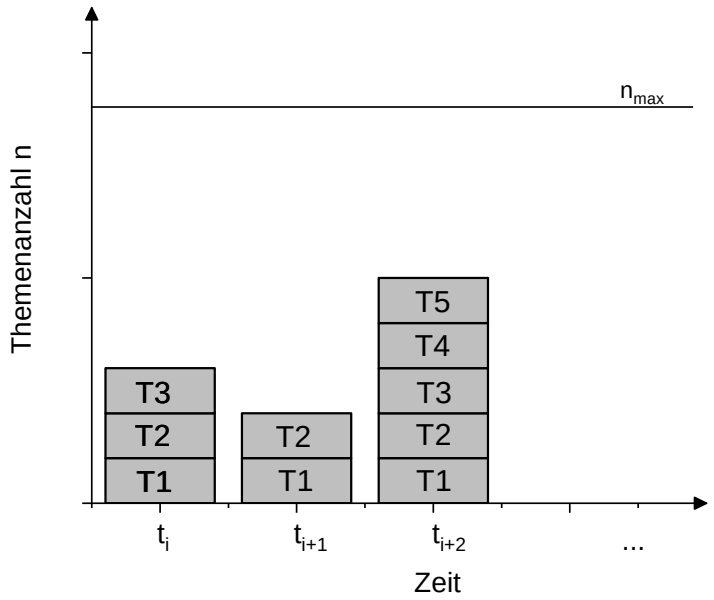

Abbildung 3.21: Schematische Darstellung der zeitlichen Diskretisierung der Daten und den ermittelten Themen

Vergleich von Themen zu verschiedenen Zeitpunkten

Für die Beschreibung der Evolution und Dynamik von Themen wurden die Daten vom Ursprung des Zeitstrahls aus in positiver Richtung miteinander verglichen. Dabei wurden direkte Vergleiche von zwei benachbarten Zeitintervallen t_i und t_{i+1} durchgeführt. Zusätzlich wurden Themen zu einem beliebigen Zeitpunkt t_j mit allen Themen aller Zeitpunkte verglichen. Damit sollen Themenverläufe erkannt werden, die nicht direkt aufeinanderfolgen.

Der Vergleich der Themen wurde mittels den vorgestellten Distanzmaßen in Abschnitt 2.3 durchgeführt. Dabei existierten zwei verschiedene Möglichkeiten, wie Themen miteinander verglichen wurden:

1. Vergleich von Vektorrepräsentationen der Terme
2. Vergleich von Wahrscheinlichkeitsverteilungen der Termverteilungen pro Thema

Für die Vektorrepräsentationen existieren verschiedene Möglichkeiten. Die Daten wurden einerseits durch einen Binärvektor, andererseits durch die Realwerte im Raum platziert. Die Vektoren wurden dann mittels Distanz- bzw. Ähnlichkeitsmaßen miteinander verglichen. Der Umfang an Termen, die in die Berechnung einbezogen wurden, variierte. Bei der Distanz zwischen Wahrscheinlichkeitsverteilungen wurden alle Terme der Verteilung berücksichtigt. Bei dem Vergleich mittels Vektorähnlichkeiten wurden lediglich die 30 wahrscheinlichsten Terme benutzt werden.

Durch die Bestimmung von ähnlichen Themen kann deren Übergang und Verhalten graphisch aufbereitet und zusammengefasst werden. In Abbildung 3.22 ist eine schematische Darstellung dieses Vorgehens gezeigt.

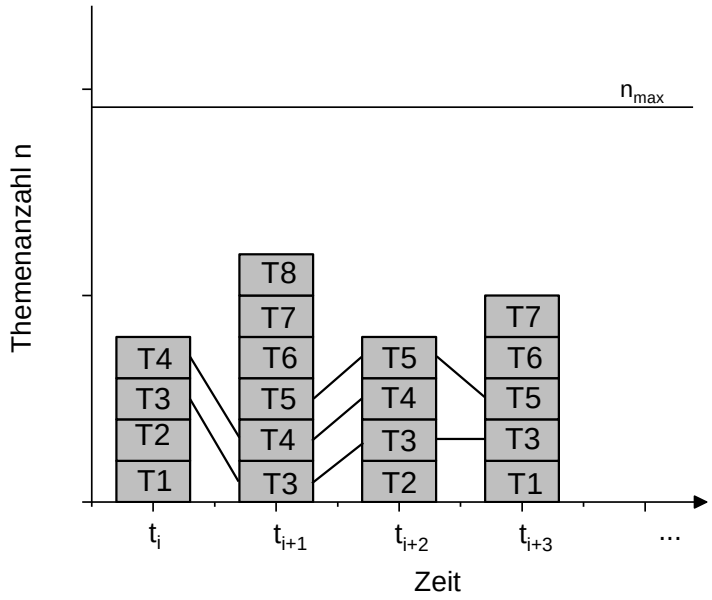

Abbildung 3.22: Schematische Darstellung der Identifizierung von ähnlichen Themen in direkt angrenzenden Zeitintervallen

Themenbeschreibung mittels Hashtags

Grundsätzlich bestand die Möglichkeit, die Themen anhand von Hashtags zu beschreiben. Dadurch könnte der Aufwand der Themenextraktion verringert werden. Das ermöglicht eine schnellere Verarbeitung der Daten und eine exakte Bestimmung von aktuellen Inhalten. Dafür wurde der Verlauf einzelner Hashtags analysiert.

Themenintensität

Neben dem Beschreiben der Dynamik von Themen, war eine Analyse des Verlaufs eines Themas von Interesse. Dafür wurde der Anteil eines Themas pro Zeitpunkt bestimmt und dieser über die Zeit hinweg beobachtet [Zhu et al., 2016]. Dadurch konnten Verläufe von Themen beschrieben werden. Dabei wurde nicht nur das Auftreten des Themas betrachtet, sondern wie viele Dokumente der jeweiligen Sammlung dieses Thema behandelten.

Durch das Bestimmen der Themenintensität über die Zeit konnte der Verlauf charakterisiert werden. Themen können unter anderem konstant vertreten sein, einmalig auftauchen oder zyklisch wiederkehren.

4 Auswertung und Diskussion

In diesem Kapitel werden die Resultate der unterschiedlichen Methoden vorgestellt. Dabei werden zu Beginn die verschiedenen Rahmenbedingungen verglichen: Abstands- und Ähnlichkeitsmaße, Textkorpora und Zeitdiskretisierungen. Zusätzlich werden die Themen in eigenen Inhalten und Retweets miteinander verglichen. Danach folgt die Auswertung der Themendynamik und eine erste Visualisierung von Themenübergängen. Abschließend werden Themenintensitäten und die Beziehungen von Profilen untereinander analysiert.

4.1 Bewertung der Maßzahlen zum Vergleich von Themen

Für die Beurteilung, welche Maßzahlen für den Vergleich von Themen am geeignetsten sind, wurden extrahierte Themen der Tagesschau miteinander verglichen. Diese stammen vom Juni 2021 und wurden durch eine wöchentliche Themenextraktion ermittelt. Die erste Gruppe von Themen wurde in der Woche vom 14.06.2021 bis zum 20.06.2021 bestimmt und ist in Tabelle 4.1 dargestellt. Die zweite Gruppe beschreibt Themen aus der Woche vom 21.06.21 bis 26.06.2021 und ist in Tabelle 4.2 zu sehen. In beiden Fällen sind nicht alle gefundenen Themen der jeweiligen Woche dargestellt. Die Themen werden in den Tabellen durch die fünf häufigsten Terme beschrieben.

Tabelle 4.1: Beispiele von extrahierten Themen der Tagesschau mit den jeweils fünf charakteristischsten Termen der Woche vom 14.06.2021 bis 20.06.2021

Thema 1	Thema 2	Thema 3	Thema 4	Thema 5
marktbericht,	usa,	gruene,	coronavirus,	deutschland,
dax,	china,	gruenen,	corona,	euro,
boerse,	biden,	parteitag,	liveblog,	frankreich,
dowjones,	nato,	app,	maskenpflicht,	eriksen,
curevac	putin	bundeswehr	delta	deutsche

Tabelle 4.2: Beispiele von extrahierten Themen der Tagesschau mit den jeweils fünf charakteristischsten Termen der Woche vom 21.06.2021 bis 26.06.2021

Thema 1	Thema 2	Thema 3	Thema 4	Thema 5
usa,	liveblog,	wahlprogramm,	ungarn,	marktbericht,
biden,	coronavirus,	unioin,	uefa,	dax,
italien,	corona,	spd,	russland,	boerse,
amazon,	spahn,	csu,	euro,	dowjones,
vatikan	delta	bundestagswahl	deutschland	inflation

Thema eins und vier aus Woche eins weisen hohe Ähnlichkeiten zu den Themen zwei und fünf aus Woche zwei auf. Auch einzelne Terme der anderen Themen stimmen überein, was ebenfalls für eine thematische Ähnlichkeit spricht. Folgende Ähnlichkeiten wurden berechnet:

1. Thema 1 (Woche 1) $\rightarrow$ Thema 5 (Woche 2)
2. Thema 2 (Woche 1) $\rightarrow$ Thema 1 (Woche 2)
3. Thema 3 (Woche 1) $\rightarrow$ Thema 3 (Woche 2)
4. Thema 4 (Woche 1) $\rightarrow$ Thema 2 (Woche 2)
5. Thema 5 (Woche 1) $\rightarrow$ Thema 4 (Woche 2)
6. Thema 1 (Woche 1) $\rightarrow$ Thema 2 (Woche 2)

Die verschiedenen Maße aus Abschnitt 2.3 wurden auf die beschriebenen Themen und deren Termverteilungen angewandt. In den nachfolgenden Abschnitten sind die Ergebnisse zusammengefasst.

Abstands- und Ähnlichkeitsmaße

In Tabelle 4.3 sind die Ergebnisse der Themenvergleiche mittels verschiedenen Maßzahlen zusammengefasst.

Die Manhattan- und euklidische Distanz wurden mit binarisierten Werten berechnet. Diese wurden aus den jeweiligen Termwahrscheinlichkeiten mittels Ceiling-Funktion bestimmt. Für den Vergleich wurden die 30 wahrscheinlichsten Wörter verwendet. Weiterhin wurden die Werte auf das Intervall $[0, 1]$ über den größten Abstand normalisiert. Durch die Subtraktion der Ergebnisse von eins stehen hohe Werte für eine hohe Ähnlichkeit.

Tabelle 4.3: Ergebnisse des Themenvergleichs unter Anwendung verschiedener Abstands- und Ähnlichkeitsmaße

Methode	1→5	2→1	3→3	4→2	5→4	1→2
Manhattan	0,20	0,07	0,07	0,40	0,20	0
Euklid	0,10	0,03	0,03	0,22	0,10	0
Cosinus	0,90	0,46	0,10	0,85	0,31	0
Jaccard	0,82	0,26	0,05	0,73	0,18	0
Tanimoto	0,40	0,11	0,03	0,40	0,12	0
Pearson	0,874	0,549	0,795	0,915	0,84	0

Die Ergebnisse liegen nahe beieinander, wodurch nur schwer eine Beurteilung für die Unterschiedlichkeit getroffen werden kann. Obwohl die Themen eine hohe Ähnlichkeit aufweisen, spiegeln die Ergebnisse das nicht wider. Die höchste Ähnlichkeit von 0,4 (Manhattan) bzw. 0,22 (Euklid) sind für die hohe Ähnlichkeit der Themen nicht ausreichend.

Die Kosinus-Ähnlichkeit wurde mit den jeweiligen Wahrscheinlichkeitswerten auf den häufigsten 30 Termen bestimmt. Die Ergebnisse decken einen großen Bereich ab und liefern sehr gute Ähnlichkeiten für die Themen 1→5 und 4→2. Die Jaccard-Ähnlichkeit liefert ähnlich gute Ergebnisse wie die Cosinus-Ähnlichkeit. Dabei fielen die Resultate im allgemeinen etwas geringer aus.

Die Ergebnisse des Tanimoto-Koeffizienten sind sehr gering, wodurch ähnliche Themen nicht ausreichend hohe Resultate liefern. Dadurch sind die Themenvergleiche nahe beieinander, was Abstufungen erschwert. Der letzte Vergleich mittels Pearson-Korrelation lieferte sehr hohe Werte, die eine hohe Ähnlichkeit aller Themen impliziert.

Zusammenfassend lieferte die Kosinus-Ähnlichkeit die besten Werte. Die Ergebnisse sind über das gesamte Intervall verteilt, was eine differenzierte Betrachtung und Kategorisierung von Ähnlichkeiten zulässt. Außerdem wurden den Themen, die eine hohe Ähnlichkeit besitzen, eine hohe Ähnlichkeit zugewiesen.

Vergleich von Wahrscheinlichkeitsverteilungen

Für die Beurteilung von Ähnlichkeiten mittels Wahrscheinlichkeitsverteilungen wurden alle Terme pro Thema benutzt. Die Summe über alle Wahrscheinlichkeiten der Terme eines Themas muss eins ergeben. In Tabelle 4.4 sind die Ergebnisse der Themenvergleiche zusammengefasst.

Die Ergebnisse aller Maße beschreiben die Ähnlichkeit der Termverteilungen. Die Er-

Tabelle 4.4: Ergebnisse des Themenvergleichs unter Betrachtung der Wahrscheinlichkeitsverteilungen der Terme pro Thema

Methode	1→5	2→1	3→3	4→2	5→4	1→2
Kullback-Leibler	3,19	3,79	4,03	2,71	3,74	4,54
Jensen-Shannon	0,45	0,56	0,57	0,42	0,55	0,61
Hellinger	1,58	1,75	1,78	1,53	1,74	1,83

gebnisse der Kullback-Leibler Divergenz und der Hellinger Distanz sind nicht normiert, weshalb kleinere Ergebnisse für ähnlichere Termverteilungen stehen. Die Jensen-Shannon Divergenz liefert Ergebnisse im Intervall $[0, 1]$, wobei auch hier gilt, dass kleinere Werte für eine höhere Ähnlichkeit steht.

In den Resultaten aller Maße sind ähnliche Themen zu erkennen, dennoch liegen alle Ergebnisse nahe beieinander, was Abstufungen in den Ähnlichkeiten erschwert. Dieser Umstand trifft vor allem auf die Hellinger Distanz zu. Die Kullback-Leibler Divergenz bietet den größten Bereich der Ergebnisse.

Im Vergleich der Maßzahlen zur Charakterisierung von Verteilungen mit den im vorherigen Abschnitt behandelten allgemeinen Abstands- und Distanzmaßen, erreichen die komplexeren Berechnungen der Verteilungen keine besseren Ergebnisse. Hinzukommt die aufwändigere Berechnung, die eine Beeinträchtigung der Performanz bei häufigen Iterationen bedeutet.

Zusammenfassend bieten die Vergleiche der gesamten Termverteilungen keine Vorteile. Auch wenn die Berechnungen theoretisch präzisere Ergebnisse liefern, sind die Termverteilungen jenseits der ersten 30 Terme nicht so charakteristisch über die gesamte Menge, dass die Ergebnisse besser sind. In allen Berechnungen wurde neben der Kosinus-Ähnlichkeit die KL-Divergenz als Vergleich mitgeführt.

4.2 Vergleich der Textkorpora: Terme und Bigramme

Der Vergleich der Korpora von Termen und Bigrammen wird exemplarisch an extrahierten Themen der Tagesschau geführt. In Tabelle 4.5 ist ein Ausschnitt der Themen aus der Woche vom 07.06.2021 bis 13.06.2021 dargestellt. Darin wurden alle Themen durch die fünf charakteristischsten Terme veranschaulicht. Für die Extraktion dieser Themen wurde eine LDA mit einzelnen Termen (ohne Bigramme) als Input genutzt genutzt. In Tabelle 4.6 sind die extrahierten Themen des gleichen Zeitraums mit Termen und Bigrammen dargestellt.

Bei beiden Extraktionen sind sehr ähnliche Terme unter den charakteristischsten Wör-

Tabelle 4.5: Beispiele von extrahierten Themen ohne Bigramme der Tagesschau mit den jeweils fünf charakteristischsten Termen der Woche vom 07.06.2021 bis 13.6.2021

Thema 1	Thema 2	Thema 3	Thema 4	Thema 5
usa,	marktbericht,	coronavirus,	sachsen,	gruene,
biden,	dax,	liveblog,	anhalt,	paretitag,
china,	boerse,	corona,	ltwsa,	baerbock,
merkel,	dowjones,	bundestag,	ltwsa,	gruenen,
gipfel	anleger	inzidenz	wahl	red

Tabelle 4.6: Beispiele von extrahierten Themen mit Bigrammen der Tagesschau mit den jeweils fünf charakteristischsten Termen der Woche vom 07.06.2021 bis 13.6.2021

Thema 1	Thema 2	Thema 3	Thema 4	Thema 5
usa,	marktbericht,	coronavirus,	sachsen,	parteitag,
biden,	dax,	liveblog,	anhalt,	gruene,
china,	boerse,	corona,	sachsen_anhalt,	baerbock,
gipfel,	marktbericht_ boerse,	liveblog_ coronavirus,	ltwsa,	rente,
inflation	boerse_dax	europa	ltwlsa	kommentar

tern. Über die jeweils fünf dargestellten Terme hinaus unterscheiden sich die Termverteilungen nur minimal in den Anordnungen. Auffällig ist, dass die Bigramme Kombinationen aus häufig vorkommenden Termen sind. Der Hintergrund dafür liegt in dem gemeinsamen Auftreten gleichwahrscheinlicher Terme. Treten die relevantesten Terme eines Themas häufig in Kombination auf, werden die dadurch gebildeten Bigramme charakteristischer für das jeweilige Thema. Somit ergeben sich Bigramme, die seltener neue Terme in die oberen Positionen der Verteilungen bringen.

Die Beschreibung der Themen wird durch die Nutzung von Bigrammen nicht ungenauer. Die berechneten Ähnlichkeiten fallen jedoch schlechter aus, da in einigen Fällen nicht identische Bigramme gebildet wurden. Somit verschlechterte sich der Vergleich der Themen unter der Nutzung von Bigrammen.

Die Verwendung von Termen und Bigrammen in Kombination führt zu einer sehr viel größeren Menge an Objekten in den Verteilungen pro Thema. So waren bei dem Ansatz ohne Bigramme in einem Thema 1.640 Terme vorhanden. Unter Verwendung von Bigrammen erhöhte sich die Zahl auf 4.017. Dieser Umstand verursachte eine schlechtere Performanz, was vor allem beim Vergleich großer Datenmengen relevant war.

Ein letzter Nachteil der Bigramme wird deutlich, wenn weniger Texte für die Themenextraktion zur Verfügung stehen. In diesen Fällen führen Bigramme dazu, dass die charak-

teristischsten Einträge häufig Kombinationen der Terme sind, die einen geringen Wahrscheinlichkeitswert besitzen. Durch das Fehlen von einschlägigen Termen pro Thema, werden die Irrelevanten in verschiedenen Kombinationen als quasi-relevant betrachtet. Somit wird das Resultat verfälscht und liefert keine besseren Ergebnisse als die Methode der einzelnen Terme ohne Bigramme.

Zusammenfassend stellt die Themenbeschreibung mit Termkombinationen in Form von Bigrammen keine Verbesserung dar. Der Vergleich von Themen mit Bigrammen fällt schlechter aus und ist in großen Datenmengen weniger performant. Somit wurden für die weiteren Betrachtungen Themenmodelle mit einzelnen Termen ohne Bigramme verwendet.

4.3 Vergleich der Zeitintervalle

Wie in Kapitel 3.3.3 beschrieben, wurden die Extraktionen in diskreten Zeitintervallen wöchentlich und monatlich durchgeführt. Dabei ist die Anzahl an Texten pro Zeitpunkt von großer Bedeutung, da nur bei ausreichend großer Menge zuverlässig Themen extrahiert werden können. Neben dem Umfang der Dokumentensammlung sind die extrahierten Themen von Relevanz. Wenn die Texte pro Monat analysiert werden, können einzelne Themen und deren Verlauf bzw. Übergänge unter Umständen nicht ausreichend gut detektiert werden.

Sowohl der Umfang der Texte als auch die gefundenen Themen werden im Folgenden kurz beschrieben. Die CDU/CSU stellte größtenteils genügend Datenmengen zur Verfügung, weshalb am Beispiel der zwei Monate April und Mai 2021 und den dazugehörigen Wochen die gefundenen Themen verglichen wurden.

Das untersuchte Zeitfenster der CDU/CSU liegt zwischen Anfang April und Ende Mai 2021. Dabei wurden Themen betrachtet, die in Verbindung mit Presseinformationen standen. In Tabelle 4.7 sind die extrahierten Themen des monatlichen Zeitfensters in Bezug auf Presseinformationen zusammengefasst.

Tabelle 4.7: Extrahierte Themen mit Bezug zu Presseinformationen pro Monat mit den zehn charakteristischsten Termen der CDU/CSU

Monat	Thema Presseinformation
April	presseinfo, luczak, mathias, middelberg, tschipanski, stephan, stracke, sicherheitsgesetz, wichtige, zentrale
Mai	presseinfo, juergenhardt, antjetillmann, jowadephul, gewalt, israel, stegemannalbert, angriffe, hamas, fguentzler

Für den Monat April 2021 standen 315 und für Mai 2021 293 Texte zur Verfügung. Die extrahierten Themen der wöchentlichen Analyse sind in Tabelle 4.8 dargestellt. Auf Grund der geringeren Datenmenge werden jeweils die fünf häufigsten Terme genannt. Die Anzahl an Posts ist pro Woche angegeben.

Tabelle 4.8: Extrahierte Themen mit Bezug zu Presseinformationen pro Woche mit den fünf charakteristischsten Termen der CDU/CSU

Woche	# Posts	Thema Presseinformation
28.03.-02.04.	33	presseinfo, peterweissmdb, einsatz, stegemannalbert, wasserstoff
04.04.-10.04.	32	presseinfo, union, unterstuetzung, juergenhardt, fdp
11.04.-17.04.	135	presseinfo, ramadan, bauen, wuensche, ramadankareem
18.04.-24.04.	94	presseinfo, mathias, middelberg, video, sicherheitsgesetzte
25.04.-01.05.	57	pandemie, presseinfo, innovation, impfung, gemeinsam
02.05.-08.05.	99	presseinfo, ulrichlange, menschen, pandemie, baulandmobilisierungsgesetz
09.05.-15.05.	33	presseinfo, interview, denken, impfturbo, stephan
16.05.-22.05.	97	presseinfo, bundestag, antjetillmann zukunft, trend
24.05.-29.05.	48	presseinfo, juergenhardt, staatsterrorismus, wichtige, frauen

Zwischen den extrahierten Themen pro Monat und pro Woche sind deutliche Unterschiede zu erkennen. Ein Teil der Terme aus den monatlichen Extraktionen kann den verschiedenen Wochen zugeordnet werden. In den Wochen wurde differenzierter über Inhalte in Bezug auf Presseinformationen gesprochen. Somit gehen detaillierte Informationen in monatlicher Betrachtung der Daten verloren.

Trotzdem ist die monatliche Analyse anwendbar, wenn die Kenntnis über das Auftreten von Themen in Bezug auf Presseinformationen ausreichend ist. Eine Beurteilung, welches Extraktionsfenster besser geeignet ist, kann nur im Kontext des Anwendungsfalles getroffen werden. Durch monatliche Auswertung kann die Dynamik des übergeordneten Themas der Presseinformationen nachvollzogen werden. Das Auftreten und Verschwinden der Thematik ist nachvollziehbar. Wird hingegen eine Auswertung von genaueren Inhalten angestrebt, muss ein kleineres Zeitintervall gewählt werden. Dieser Umstand wird besonders dann relevant, wenn Themen und Unterthemen differenziert betrachtet werden.

Je kleiner ein Zeitfenster gewählt wird, desto detaillierter können kurzzeitige Schwankungen in Themen detektiert werden. Auch das Aufkommen von neuen Themen wird genauer, je kleiner das Zeitfenster ist. Dafür sind jedoch ausreichend große Datenmengen nötig, da andernfalls keine Themen mittels LDA extrahiert werden können.

Das wöchentliche Zeitfenster adressiert den Ansatz möglichst genauer Themenbeschreibungen, die eine detaillierte Beschreibung der Dynamik erlauben. Bei der Anwendung dieses Zeitfensters ergaben sich Probleme bei Profilen, die nicht ausreichend Daten zur Verfügung stellten. Vor allem bei älteren Tweets der CDU/CSU und der Linken sowie bei der AfD. Auch die Analyse der Landtagsparteien lieferte oft schlecht Ergebnisse durch die geringe Zahl an Posts.

Die Analyse der Daten in Monaten versuchte das Problem der zu wenigen Daten auszugleichen, wohingegen Themenänderungen gröber aufgelöst wurden. Bei beiden Versionen wurden Zeitpunkte ignoriert, an denen von vornherein zu wenige Daten verfügbar waren. In der Abfolge von Themenextraktionen entstanden zusätzlich Unterbrechungen, wenn keine Tweets gepostet wurden.

4.4 Nutzung von Tweets und Retweets

Der Grund für die Unterscheidung von Tweets und Retweets zur Themenextraktion liegt in der Datenquelle der Texte. Tweets sind Inhalte, die von dem jeweiligen Profil selbst erstellt und formuliert wurden. Retweets hingegen wurden von unterschiedlichen anderen Profilen erstellt. Somit sind die Daten heterogener durch die Nutzung von Retweets. Dieser Umstand kann sich in den extrahierten Themen widerspiegeln.

Am Beispiel des Profils der Linken wurden Tweets und Retweets in den Wochen vom 02.05.2021 bis 12.06.2021 verglichen. Aus Kapitel 3.2.1 ist bekannt, dass bei der Linken das Verhältnis von Tweets zu Retweets sehr unausgeglichen ist. Die Partei hat einen überwiegenden Teil der Inhalte retweeted. In Tabelle 4.9 ist die Anzahl der Tweets und Retweets in den jeweiligen Wochen zusammengestellt. Zusätzlich ist die Anzahl der extrahierten Themen angegeben.

Durch den Vergleich der sechs Wochen wird deutlich, dass die Anzahl der Tweets nicht ausreicht, damit zuverlässig Themen extrahiert werden können. Die minimale Themenanzahl, die durch den Algorithmus bestimmt werden kann, liegt bei zwei, sodass diese überwiegt. Somit ist die Nutzung von Tweets und Retweets in Kombination notwendig, um Themenuntersuchungen durchzuführen und kontinuierlich die Entwicklung der Inhalte nachzuvollziehen.

Tabelle 4.9: Vergleich von Tweets und Retweets im Zeitraum vom 02.05.2021 bis 12.06.2021 sowie den extrahierten Themen

Woche	# Tweets	# Themen Tweets	# Retweets	# Themen Retweets
02.05.2021-08.06.2021	9	3	39	8
09.05.2021-15.06.2021	9	2	36	9
16.05.2021-22.06.2021	8	2	38	8
23.05.2021-29.06.2021	10	2	42	8
30.05.2021-05.06.2021	6	2	31	8
06.06.2021-12.06.2021	7	2	43	9

In Tabelle 4.10 sind exemplarisch die Themen der Tweets der Woche vom 09.05.2021 bis 15.06.2021 zusammengefasst. In Tabelle 4.11 sind die Themen des selben Zeitraumes der Retweets aufgeführt. In beiden Fällen sind die fünf häufigsten Terme angegeben.

Tabelle 4.10: Extrahierte Themen der Linken in der Woche vom 09.05.2021 bis 15.06.2021 aus ausschließlich eigenen Inhalten

Thema	Terme pro Thema
Thema 1	personal, tagderpflege, pflegekraefte, pflege, pflegenotstand
Thema 2	janine, wissler, machtdaslandgerecht, dietmarbartsch, gemeinsam

Tabelle 4.11: Extrahierte Themen der Linken in der Woche vom 09.05.2021 bis 15.06.2021 aus ausschließlich geteilten Inhalten

Thema	Terme pro Thema
Thema 1	corona, land, ausschließt, kuerzen, sozial
Thema 2	interview, linken, nsu, hausbesetzungen, dietmarbartsch
Thema 3	dielinke, janine, wissler, btw, susannehenning
Thema 4	loehne, petition, aufkommen, kosten, pandemie
Thema 5	amazon, makeamazonpay, jahren, greift, studie
Thema 6	israel, angesichts, antisemitismus, entschieden, form
Thema 7	zweistellig, legal, peanuts, aufbruch, auseinander
Thema 8	spahn, gewinnt, auftragsvergabe, besserdielinke, branchen
Thema 9	tagderpflege, pflege, endlich, bundesweiten, deutliche

Durch den direkten Vergleich der zwei Wochen ist sichtbar, dass alle Themen der Tweets in den Retweets vorhanden sind. Auch wenn die geteilten Posts nicht direkt von der Partei verfasst wurden, entsprechen die Themen denen der eigenen Inhalte. Somit führt die Nutzung von Tweets und Retweets zu einem größerem Datensatz, der die Extraktion von Themen verbessert. Dabei werden wesentliche Inhalte nicht verfälscht, was zusätzlich für die Einbeziehung heterogener Daten spricht.

4.5 Auswertung der extrahierten Themen

Die Themenextraktion und der Vergleich der Themen wurde für jedes Profil individuell durchgeführt. Pro Profil wird eine Übersicht zu den gefundenen Themen pro Zeitintervall vorgestellt. Danach werden einzelne Zeitabschnitt detaillierter betrachtet und diskutiert.

Die Daten der Profile wurden in wöchentlichen und monatlichen Intervallen betrachtet, wobei sowohl Tweets als auch Retweets einbezogen wurden. Dabei wurden die Daten vom Datum der Profilerstellung an in die jeweiligen Zeitintervalle eingeteilt. Pro Zeitintervall wurden alle darin extrahierten Themen betrachtet. In Abbildung 4.1 ist die gefundene Anzahl an Themen pro wöchentlichem Zeitintervall der CDU/CSU dargestellt.

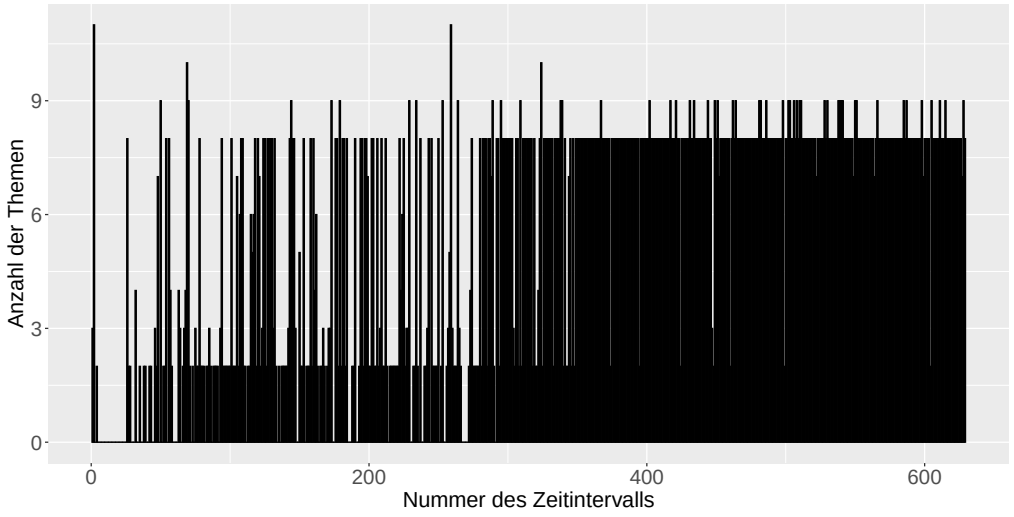

Abbildung 4.1: Darstellung der Anzahl extrahierter Themen pro wöchentlichem Zeitabschnitt der CDU/CSU

Dabei ist der deutliche Anstieg der Twitternutzung ab 2015 zu erkennen (siehe Abschnitt 3.2.1). Da zwei Themen die minimale Anzahl darstellt, sind die Themen in der ersten Hälfte der Zeitintervalle weniger verlässlich. Ab ca. der Hälfte ist die deutlich stärkere

Nutzung von Twitter in den extrahierten Themen zu erkennen. In diesem Bereich wurden konstant acht Themen pro Woche ermittelt.

In Abbildung 4.2 sind die wöchentlich extrahierten Themen der AfD dargestellt.

Abbildung 4.2: Darstellung der Anzahl extrahierter Themen pro wöchentlichem Zeitabschnitt der AfD

Wie in dem Verlauf der Posts in Abschnitt 3.2.1 zu erkennen, begann die kontinuierliche Twitternutzung ab März 2020. Ab diesem Zeitpunkt wurden konstant acht Themen extrahiert. Bis auf einen kurzen Anstieg im September 2017 sind vor März 2020 zu wenig Tweets vorhanden, um die Themendynamik ausreichend gut nachzuvollziehen.

Der Verlauf der extrahierten Themen der Linken ist in Abbildung 4.3 dargestellt.

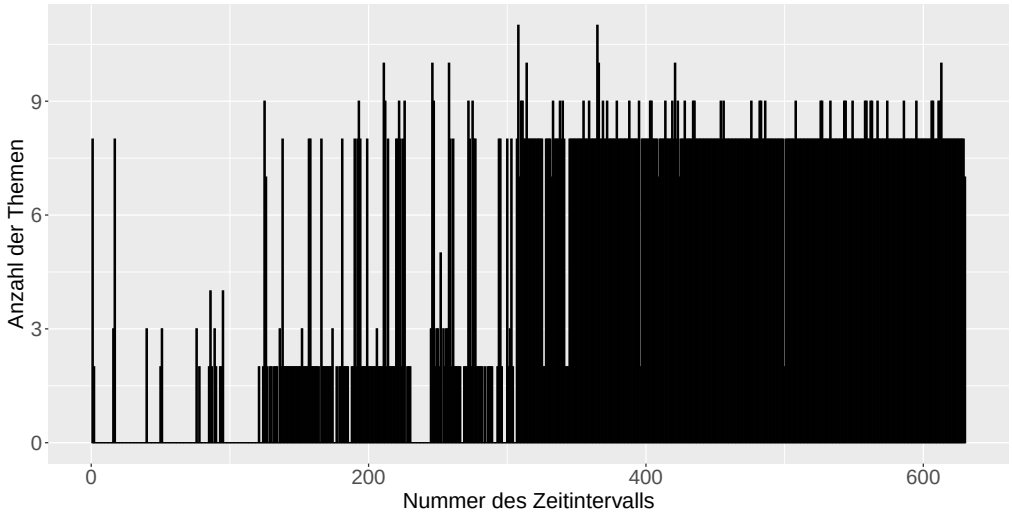

Abbildung 4.3: Darstellung der Anzahl extrahierter Themen pro wöchentlichem Zeitabschnitt der Linken

Der Verlauf der Themen ähnelt dem der CDU/CSU. Die Linke nutzte Twitter verstärkt ab dem Jahr 2015, weshalb ab diesem Zeitraum deutlich mehr Themen pro Woche extrahiert wurden. Auch die häufig auftretenden einzelnen Peaks des Tweetverlaufs (siehe Abschnitt 3.2.1) sind in den Themen zu erkennen.

Der letzte Verlauf der Themenanzahl der Tagesschau ist in Abbildung 4.4 dargestellt.

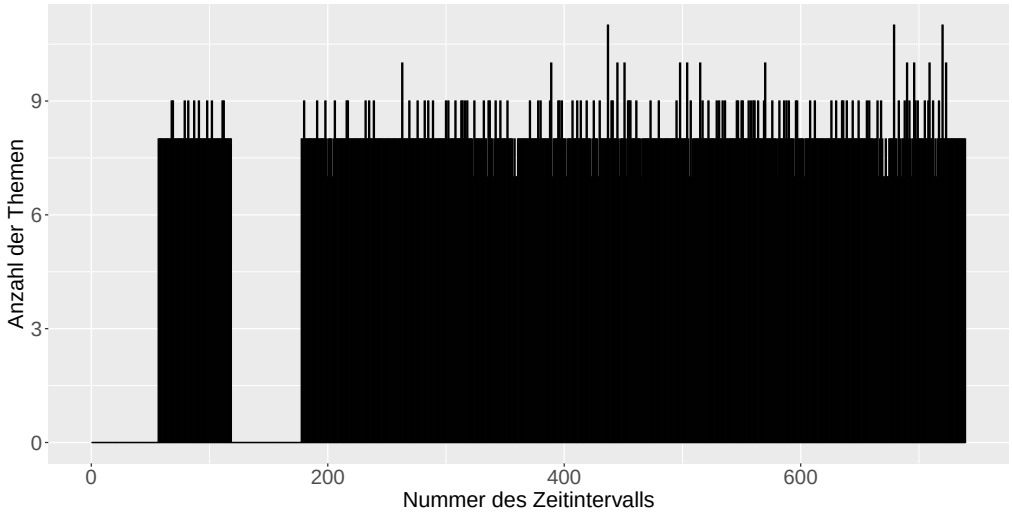

Abbildung 4.4: Darstellung der Anzahl extrahierter Themen pro wöchentlichem Zeitabschnitt der Tagesschau

Aus den Daten der Tagesschau wurden konstant acht Themen extrahiert. Zwei Zeitabschnitte, in denen keine Tweets gepostet wurden, stellen keine Themen zur Verfügung. Dieser Datensatz bietet den längsten zusammenhängenden Datenkomplex. Je länger der zusammenhängende Teil an ausreichenden Themen ist, desto besser können Themendynamiken und -entwicklungen nachvollzogen werden. In Anhang A.2 sind die monatlichen Themenanzahlen aller Profile zusammengefasst.

4.5.1 Identifizierung ähnlicher Themen in benachbarten Zeitabschnitten

Der Verlauf von Themen in direkt aneinander angrenzenden Zeitabschnitten ist ein starker Indikator für die Entwicklung und Dynamik von Themen. Deshalb wurde in allen Datensätzen der Verlauf der Themen anhand ähnlicher Kernaspekte untersucht. Im Zeitraum vom 02.05.2021 bis 26.06.2021 wurden wochenweise alle Themen eines Zeitabschnittes mit allen Themen der angrenzenden Zeitabschnitte verglichen.

Die visuelle Darstellung der Themenübergänge wurde mittels angepassten Sankey-Diagrammen durchgeführt. Diese erlauben eine graphische Aufbereitung von Mengenflüssen. Angewandt auf die Themenübergänge sind ähnliche Themen in angrenzenden Zeitabschnitten miteinander verbunden. Dabei ist die Breite der Verbindung direkt proportional zu der Ähnlichkeit. Die vertikale Anordnung der Themen ist dabei nicht relevant, sie entspricht dem Output der LDA. In Abbildung 4.5 ist die Themendynamik der CDU/CSU dargestellt.

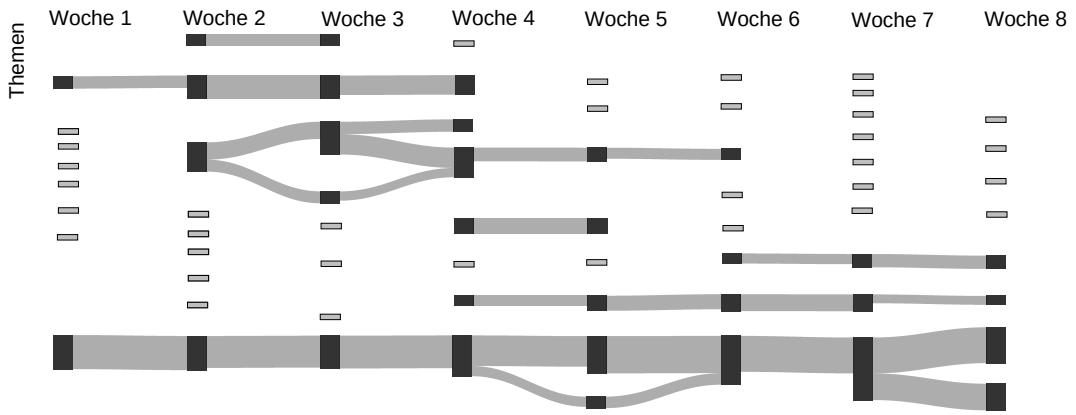

Abbildung 4.5: Schematische Darstellung des zeitlichen Verlaufs von Themen durch wöchentliche Analyse der Tweets der CDU/CSU

Ein großer Anteil aller Themen tritt isoliert pro Woche auf. Wenn Ähnlichkeiten vorlagen, waren diese zu gering, als dass von einer Fortsetzung bzw. Entwicklung des Themas gesprochen werden kann. Neben der Beobachtung einzelner Themen sind alle in Abschnitt 2.4 beschriebenen Themenübergänge vorhanden. Das dritte Thema in Woche drei Spaltet sich zur Folgewoche auf und die Inhalte werden in zwei verschiedenen Themen behandelt. Nach dem dieser Themenkomplex in Woche zwei begann, endet er in Woche sechs, da kein Thema in Woche sieben eine ausreichend hohe Ähnlichkeit zu den Inhalten besaß.

Die markanteste Dynamik ist in Thema acht Woche eins zu erkennen. Über den gesamten Zeitraum hinweg ist dieses Thema präsent und weist wöchentlich sehr hohe Ähnlichkeiten auf. Inhaltlich wurden in diesem Strang die Tweets mit Bezug zu Presseinformationen zusammengefasst. Das Thema wurde in Abschnitt 4.3 bereits vorgestellt. Durch die Visualisierung mittels angepasster Sankey-Diagramme ist dieser Verlauf schnell und eindeutig zu erkennen, was eine Bewertung von Themen und deren Verlauf vereinfacht.

Der Fluss der Themen der AfD im selben Zeitraum ist in Abbildung 4.6 dargestellt.

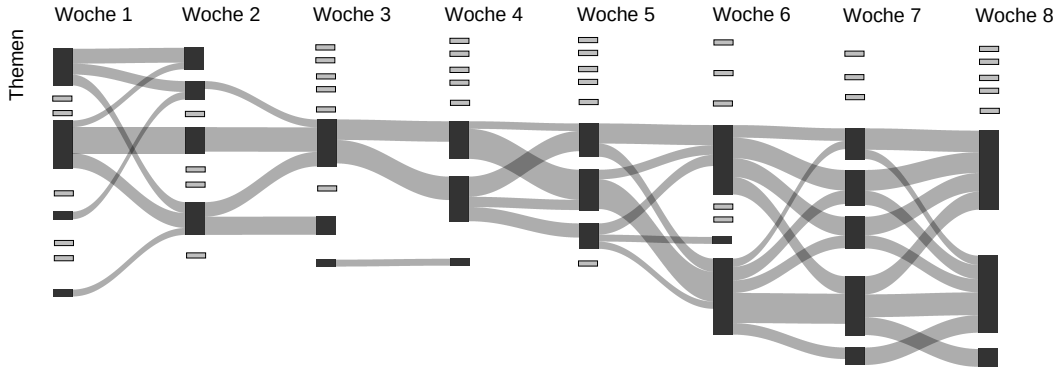

Abbildung 4.6: Schematische Darstellung des zeitlichen Verlaufs von Themen durch wöchentliche Analyse der Tweets der AfD

Auch bei der AfD existieren pro Woche mehrere Themen, die in dem jeweiligen Zeitabschnitt isoliert auftraten. Weiterhin sind viele Verknüpfungen in den Themen zu erkennen. Besonders Thema vier in Woche eins spaltet sich ab Woche drei in immer mehr Themen auf und wird vorgesetzt.

Die zum Teil sehr hohe Ähnlichkeit zwischen den Themen wird bei der AfD durch die Wahl spezieller Hashtags beeinflusst. Hashtags wie #afd und #abernormal wurden sehr

oft an verschiedene Inhalte angefügt. Somit ergaben sich zum Teil hohe Ähnlichkeiten zwischen Themen, die neben den Hashtags wenig Gemeinsamkeiten aufwiesen. Hier könnte eine detailliertere Unterscheidung in Haupt- und Unterthemen hilfreich sein, wie sie in Abschnitt 4.5.3 beschrieben wurde.

Abbildung 4.7 zeigt die Entwicklung der Themen über den definierten Zeitbereich der Linken.

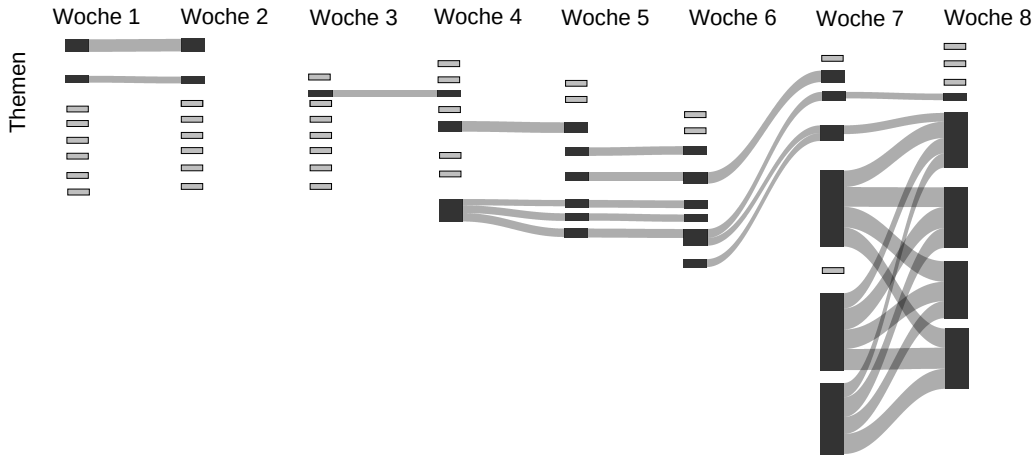

Abbildung 4.7: Schematische Darstellung des zeitlichen Verlaufs von Themen durch wöchentliche Analyse der Tweets der Linken

Im Vergleich zu den bisher vorgestellten Themendynamiken weist Die Linke sehr viel weniger Themenähnlichkeiten zwischen den den Zeitintervallen auf. Bis Woche vier dominierten die Einzelthemen deutlich, nur wenige Themen besaßen Ähnlichkeiten in benachbarten Intervallen.

Ab Woche vier und fünf begann ein Verlauf von Themen, der sich bis Woche acht fortsetzte. Besonders auffällig sind die Ähnlichkeiten zwischen den Themen in Woche sieben und acht. Inhaltlich sind diese Ähnlichkeiten auf eine sehr hohe Anzahl an Tweets mit Wahlbezug zurückzuführen. In diesen Wochen wurden viele Texte mit den Hashtags #dielinke, #sozialklimagerecht und #linkebpt im Kontext der anstehenden Bundestagswahl gepostet.

Abschließend ist die Themenentwicklung der Tagesschau in Abbildung 4.8 dargestellt.

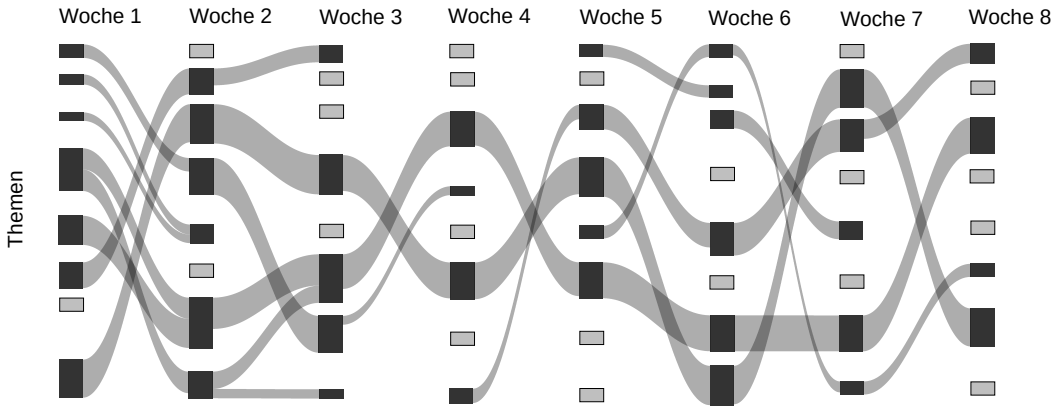

Abbildung 4.8: Schematische Darstellung des zeitlichen Verlaufs von Themen durch wöchentliche Analyse der Tweets der Tagesschau

Dabei ist zu erkennen, dass der Großteil der Themen miteinander in Verbindung stand. Einzelne Themen besitzen keine Ähnlichkeit zu anderen Themen, was bedeutet, dass die Inhalte nur in diesem Zeitabschnitt behandelt wurden. Ein Beispiel ist in Woche zwei Thema eins zu finden. Die fünf charakteristischsten Terme lauteten: fdp, spd, scholz, lindner, bundestagswahl.

Auch das Entstehen und Enden von Themen ist zu erkennen. In Woche sechs entsteht Thema drei und bleibt bis Woche sieben erhalten. Die fünf wahrscheinlichsten Terme des Themas aus Woche sechs lauten: gruene, parteitag, baerbock, gruenen, red. Die Terme des damit assoziierten Themas aus Woche sieben lauten: gruene, gruenen, parteitag, app, bundeswehr.

Die markantesten Themendynamiken sind jedoch die zu den Themen Corona und Börse. Exemplarisch wird der Verlauf des Themas zu Corona in Abbildung 4.9 hervorgehoben.

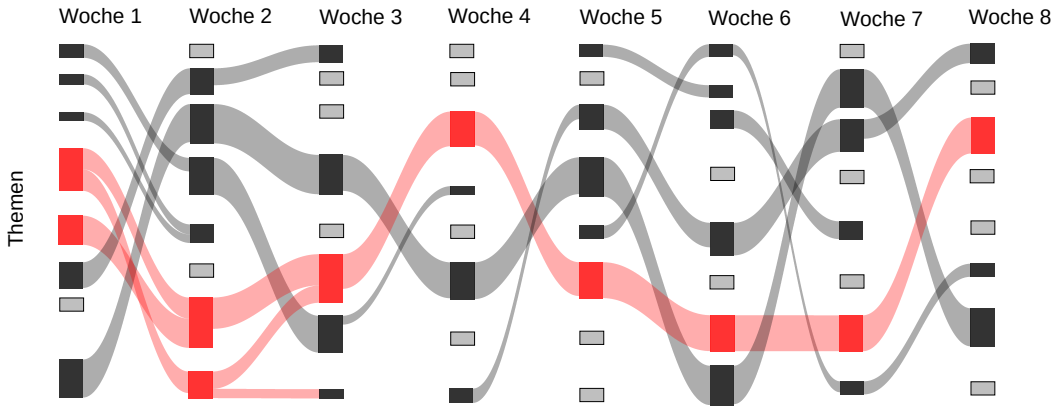

Abbildung 4.9: Themendynamik der Tagesschau mit Hervorhebung des Themas Corona

Durch den markierten Verlauf werden die individuellen Themendynamiken deutlich. Das Thema spaltet sich von Woche eins zu Woche zwei auf, danach vereinen sich die Stränge wieder zu einem Thema. Über den gesamten Zeitraum der acht Wochen ist dieses Thema vorhanden und die Ähnlichkeiten sind mittels Kosinus-Ähnlichkeit deutlich nachweisbar.

Für den Vergleich der wöchentlichen und monatlichen Analyse folgt in Abbildung 4.10 die graphische Darstellung der Themendynamiken der Monate April, Mai und Juni der Posts der Tagesschau.

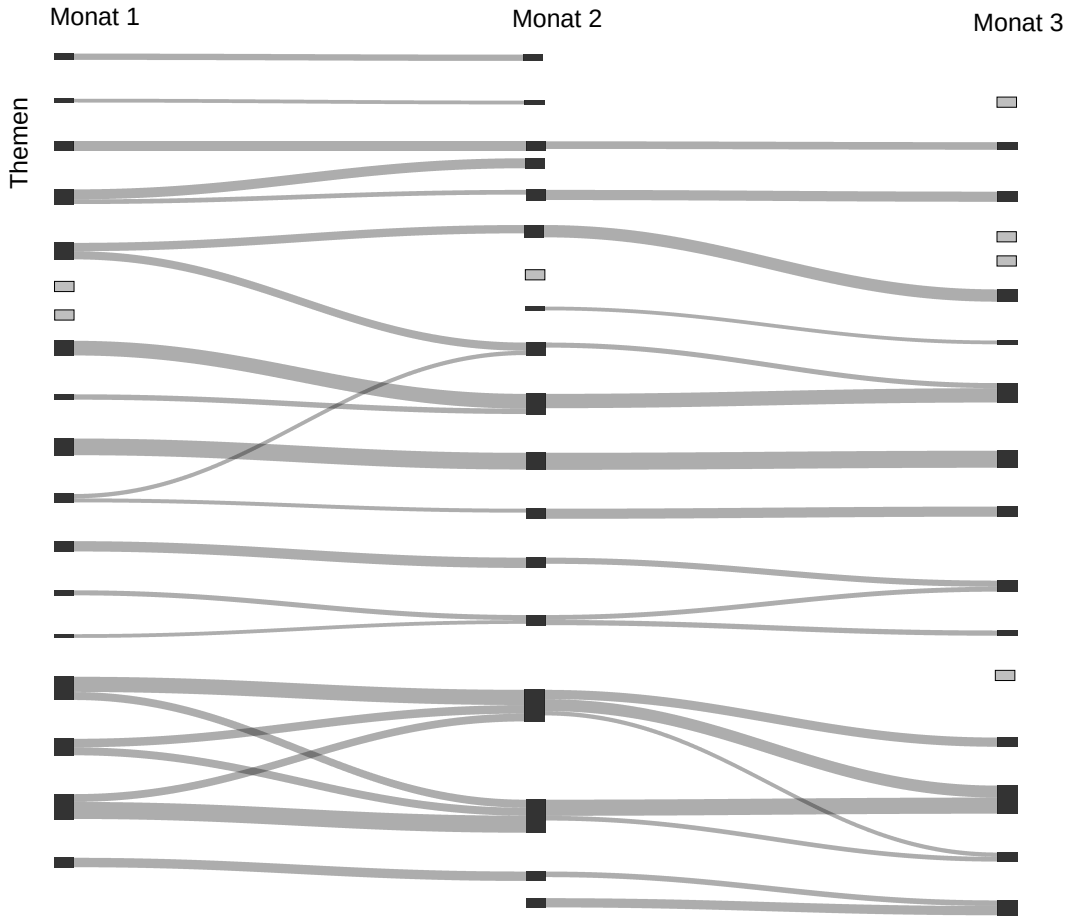

Abbildung 4.10: Schematische Darstellung des zeitlichen Verlaufs von Themen

Durch den größeren Zeitraum pro Zeitschritt waren wesentlich mehr Themen vorhanden, die mit den Angrenzenden verglichen wurden. Besonders auffällig ist auch hier, dass der überwiegende Teil der Themen über mehrere Zeitschritte vorhanden waren. Der Großteil der Kernaspekte aller Tweets blieb somit über die Zeit annähernd konstant und es wurden weniger neue Themen unter den Posts angesprochen.

Somit kann in der monatlichen Betrachtungsweise gezeigt werden, dass die Schwerpunkte über die Zeit hinweg erkannt und verfolgt werden können. Eine detailliertere Betrachtung ist hingegen bei der Wochenanalyse möglich. Je nach notwendiger Granularität der Themen und deren Entwicklung über die Zeit, sind beide Zeitfenster geeignet für die Beschreibung der Dynamik von Themen.

4.5.2 Beschreibung des Verlaufs einzelner Themen über den gesamten Zeitraum

Ein weiterer wichtiger Aspekt ist die Nachverfolgung von Themen über den gesamten Zeitraum seit der Profilerstellung. Deshalb wurden in diesem Abschnitt exemplarisch einzelne Themen und deren Ähnlichkeit zu allen Themen aller Zeitintervalle desselben Profils bestimmt. Somit kann ein konstantes bzw. wiederkehrendes Auftreten einzelner Themen nachgewiesen werden.

Die Aufbereitung der Ergebnisse bezieht den Aspekt der Themenintensität ein. Der Nachweis, dass ein Thema über die gesamte Zeit hinweg Ähnlichkeiten zu anderen besitzt, stellt die Basis der Themenverfolgung dar. Zusätzlich kann pro Zeitabschnitt die Anzahl der Dokumente betrachtet werden, die dieses Thema behandeln. Somit wird neben dem Wiederfinden von Themen auch dessen Relevanz pro Zeitabschnitt quantifiziert.

Die Themenintensität bestimmt somit den relativen Anteil aller Dokumente, die das gesuchte Thema behandeln. Somit liegt der Wertebereich der folgenden Graphiken im Intervall $[0, 1]$, da entweder kein Dokument das Thema behandelt oder alle Dokumente den Kernaspekt besitzen. Für das Profil der Linken wurde exemplarisch folgendes Thema betrachtet: sozialklimagerecht, linkebpt, wahlprogramm, susannehennig, bodoramelow. In Abbildung 4.11 ist der wöchentliche Verlauf der Themenähnlichkeiten dargestellt. Darin sind die jeweiligen Themenintensitäten pro Zeitintervall abgebildet.

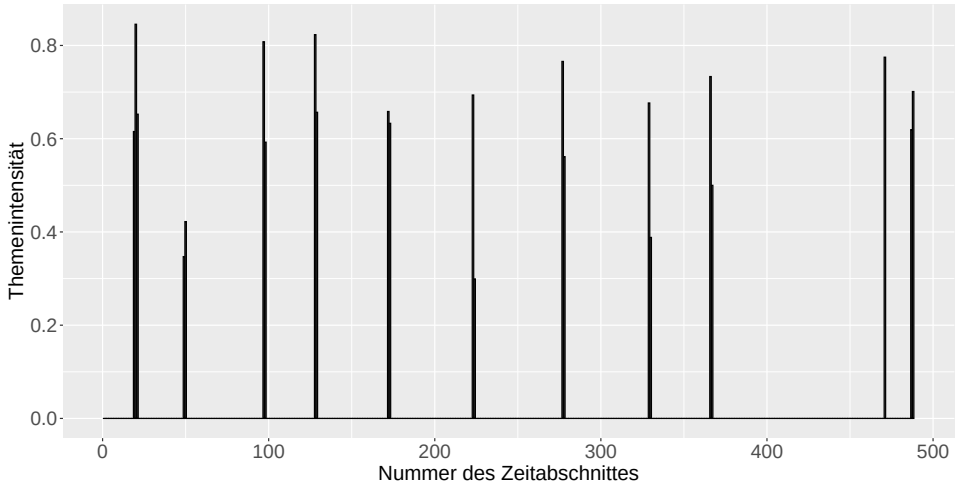

Abbildung 4.11: Darstellung der Themenintensität eines definierten Themas der Linken über den gesamten Zeitraum seit der Profilerstellung

Der Verlauf der Themenintensität zeigt den wiederkehrend hohen Anteil des Themas an allen Dokumenten des jeweiligen Zeitabschnittes. Somit ist das ausgewählte Thema sowohl ein zyklisches Thema als auch ein sehr relevantes in den betreffenden Intervallen. Das Thema wurde über den gesamten Zeitraum der Posts wiederholt aufgegriffen.

Im Gegensatz dazu steht die Behandlung des Themas COVID-19 bei der Tagesschau. Dieses Thema steht exemplarisch für kein regelmäßig wiederkehrenden Aspekt. Der Verlauf des Themas (coronavirus, corona, liveblog, maskenpflicht, delta) ist in Abbildung 4.12 dargestellt.

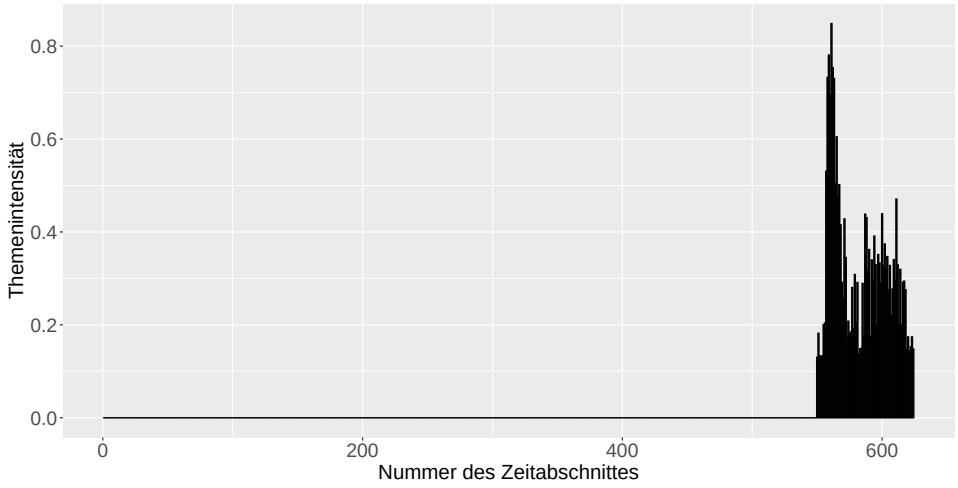

Abbildung 4.12: Darstellung der Themenintensität eines definierten Themas der Tagesschau über den gesamten Zeitraum seit der Profilerstellung

Dieser Verlauf unterscheidet sich deutlich von dem aus Abbildung 4.11. Seit März 2020 wird dieses Thema konstant behandelt. Zu Beginn war der Anteil des Themas an allen Tweets mit teilweise über 80 % sehr hoch. Über die Zeit hinweg nahm der Anteil ab und das Thema wurde in weniger Tweets angesprochen. Dabei gab es keine Woche, an dem dieses Thema nicht behandelt wurde.

Durch die vorgestellte Auswertung war es möglich, einzelne Themen isoliert mit allen Themen der gesamten Zeit zu vergleichen. Dadurch konnten Themen detektiert werden, die zu verschiedenen Zeitpunkten von den Autoren angesprochen wurden. Die Art des Auftretens ließ sich durch die Kurvenverläufe der Themenintensität beschreiben. So konnten sowohl zyklische als auch konstante Themen nachgewiesen werden. Die Quantifizierung der Themenanteile in den Dokumenten pro Zeitabschnitt erlaubte zusätzlich eine Bewertung der Relevanz eines Themas.

4.5.3 Beschreibung von Über- und Unterthemen

In den bisherigen Betrachtungen von Themenähnlichkeiten steht eine hohe Kosinus-Ähnlichkeit für sehr nahe bzw. beinahe identische Themen. Dabei wurde durch den Parameter β bei LDA die Termverteilung so eingestellt, dass wenige Wörter eine hohe Wahrscheinlichkeit für das jeweilige Thema besitzen. Dieser Umstand hat zur Folge, dass durch den Vergleich von Themen, eine Übereinstimmung der relevantesten Terme eine hohe Themenähnlichkeit bedeutet.

Dieses Verhalten ist für die Beschreibung von Themendynamiken von Vorteil, damit ähnliche Themen leichter erkannt werden. Durch diesen Aspekt lassen sich die Themen in Über- und Unterthemen einteilen. Durch die Betrachtung der beispielsweise fünf relevantesten Terme lassen sich ähnliche Themen finden, durch die weitere Betrachtung der folgenden Terme lassen sich die Unterschiede in den ähnlichen Themen bestimmen. Die weiteren, unter Umständen weniger relevanten Terme, geben Auskunft über leichte Unterschiede in verwandten Inhalten.

Durch die Betrachtung von weiteren Termen pro Thema lässt sich die Entwicklung von Unterthemen nachvollziehen. Dieses Vorgehen ist am Beispiel des Themas zur Börse der Tagesschau exemplarisch verdeutlicht. In Tabelle 4.12 sind die häufigsten 10 Terme der Themen mit Bezug zur Börse dargestellt. Es wurden die fünf Wochen vom 02.05.2021 bis 05.06.2021 betrachtet. Dabei sind die vier obersten Terme gesondert dargestellt.

Über alle Wochen hinweg sind die vier Terme marktbericht, dax, boerse und dowjones die charakteristischsten. Durch diese Übereinstimmung in den ersten Termen erreichten die Vergleiche mittels Kosinus-Ähnlichkeit sehr hohe Werte. Weiterhin wurden die Themenähnlichkeiten ohne die ersten vier Terme berechnet, was deutlich niedrigere Ergebnisse lieferte. Diese Werte erlauben eine Beurteilung der Ähnlichkeit der Unterthemen.

Tabelle 4.12: Extrahierte Themen der Tagesschau in den Wochen vom 09.05.2021 bis 15.06.2021 zum Thema Börse

Woche 1	Woche 2	Woche3	Woche 4	Woche 5
marktbericht, dax, boerse, dowjones,	marktbericht, dax, boerse, dowjones,	marktbericht, dax, boerse, dowjones,	marktbericht, dax, dowjones, boerse,	marktbericht, dax, boerse, dowjones,
schottland, snp, sturgeon, mai, grossbritannien, boersen	inflation, zinsen, myanmar, covid, anleger, eilmeldung	bitcoin, zinsen, bayer, inflation, anleger, bundesliga	deutschland, inflation, arbeitsmarkt, dollar, nasdaq, namibia	inflation, oelpreis, anleger, aktien, dollar, marktberichtdax

In Abbildung 4.13 sind die Ähnlichkeiten zwischen den Themen graphisch dargestellt, wobei sowohl die aller Terme als auch die der Unterthemen visualisiert sind.

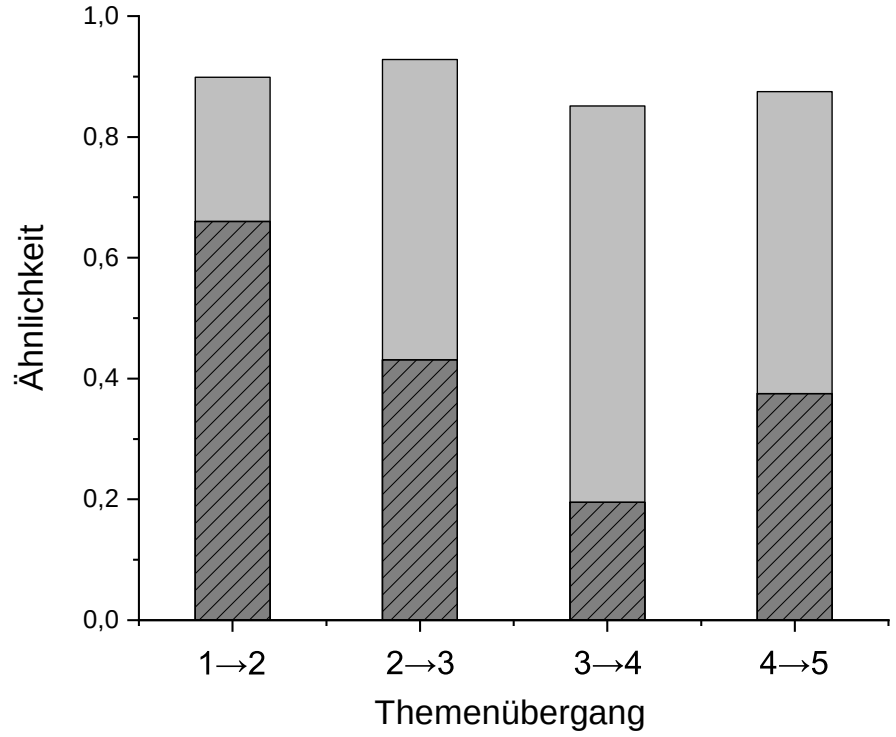

Abbildung 4.13: Darstellung der Ähnlichkeit von Themen (grau) und deren Subthemen (schraffiert) der Tagesschau

Die Ähnlichkeiten in den Unterthemen der Woche zwei und drei sowie Woche vier und fünf sind höher als die der anderen Themen. Durch diese Charakteristiken lassen sich neben den Hauptthemen auch die Unterthemen identifizieren und charakterisieren. Zusammenfassend kann die Betrachtung von Unterthemen weitere Informationen über die inhaltliche Entwicklung der Texte liefern und die Beschreibung der Dynamik konkretisieren. Je nach Anwendungsfall kann durch eine differenzierte Termauswertung pro Thema die notwendige Granularität eingestellt werden.

4.5.4 Betrachtung von Hashtags

Die Beschreibung von Hauptgedanken und deren Verläufe in Texten kann anhand von Tags erfolgen. Die Nutzung der Hashtags bei Twitter soll unter anderem Schlagworte des Posts hervorheben und beispielsweise Treffer bei einer Suche nach bestimmten Inhalten verbessern. Somit kann der Schluss gezogen werden, dass die Betrachtung von Hashtags ein Hilfsmittel zur Beschreibung der Themendynamik ist.

Dabei ergeben sich mehrere Probleme, die die alleinige Nutzung von Hashtags zur Charakterisierung von Themenverläufen ausschließen. Eine erste Schwierigkeit ist die Beschreibung eines Themas ausschließlich mit Hashtags. Dabei werden unter anderem komplexe Themen mit einem Wort beschrieben, was durch Ambiguitäten oder der fehlenden Komplexität Schwierigkeiten hervorbringt (siehe dazu Abschnitt 2.2.1). Ein Thema durch ein einzelnen Term zu umschreiben ist in vielen Fällen nicht ausreichend. Zusätzlich sind häufig Hashtags sehr ubiquitär einsetzbar. Das bedeutet, dass die Nutzung von Hashtags oft nicht dem Zweck dient, dass dadurch das angesprochene Thema konkretisiert wird. Im Umfeld der Parteien ist dieser Umstand beispielsweise durch die Hashtags der Parteinamen gegeben. Oftmals werden die Parteinamen als Hashtag an Tweets angefügt, was zu keiner inhaltlichen Konkretisierung führt.

Weiterhin besitzen viele veröffentlichte Tweets keinen Hashtag (siehe dazu die Statistiken in Abschnitt 3.2.1). Dadurch können Verläufe von Hashtags nur unter den Tweets betrachtet werden, die diese beinhalten. Auch besitzen Hashtags keine einheitliche Schreibweise. So existieren für die Landtagswahl in Sachsen-Anhalt unter anderem die Hashtags #ltwsa und #ltwlsa. Ohne die Texte der Tweets einzubeziehen, müssten alle Schreibformen von inhaltlich ähnlichen Hashtags bekannt sein, damit der Verlauf wahrheitsgetreu abgebildet werden kann.

Dennoch kann der Verlauf von Hashtags Aufschluss zu aktuellen Trends und Schwerpunkten geben. Dazu ist in Abbildung 4.14 exemplarisch der Verlauf des Hashtags #linkebpt der Linken dargestellt.

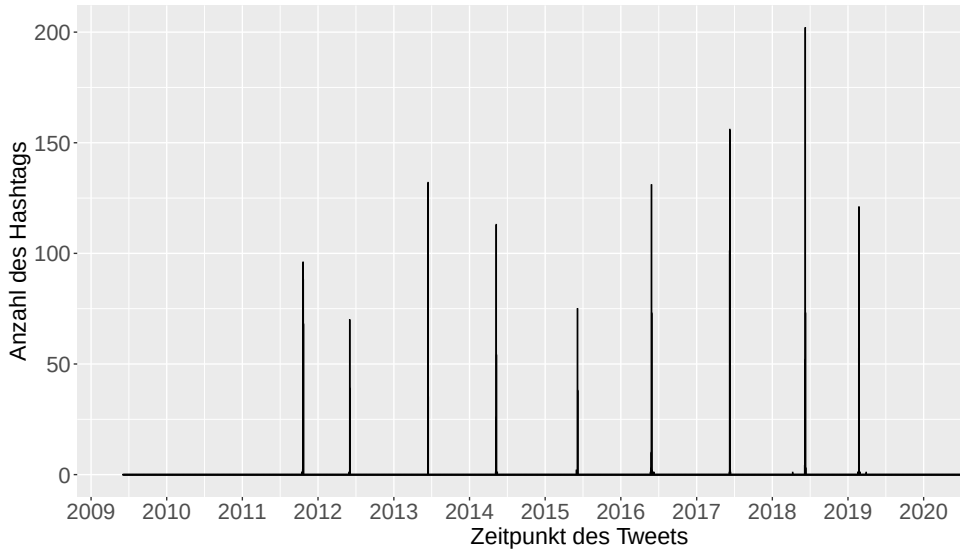

Abbildung 4.14: Darstellung des Hashtags #linkebpt der Linken über den gesamten Zeitraum seit der Profilerstellung

Ähnlich wie in Abbildung 4.11 ist der zyklische Verlauf des Themas zu erkennen. Im Kontext der Erkennung von Verläufen einzelner Schlagworte ist eine Auswertung der Hashtags anwendbar. Durch den Wegfall der LDA und der direkten Auswertung der Hashtags ist ein deutlicher Performanzgewinn vorhanden.

In vielen Fällen sind Hashtags durch ihre häufige Nutzung unter den charakteristischsten Termen von Themen. Die im vorangegangenen Kapitel beschriebenen Themen mit Bezug zur Börse der Tagesschau besaßen alle die gleichen vier relevantesten Terme. Dabei sind #marktbereicht, #boerse und #dax verwendete Hashtags der Tagesschau.

Obwohl Hashtags einen engen Fokus bieten und keine Interpretationen und Charakterisierungen von Themen erlauben, ist deren Analyse zusätzlich zur Themenextraktion anwendbar. Unter Betrachtung der genanten Punkte kann eine Auswertung der Hashtags keine Themenextraktion und deren Beschreibung der Themendynamik ersetzen.

4.6 Profilübergreifende Themendynamik

Der Vergleich von Themen und deren Ähnlichkeiten über verschiedene Profile hinweg ermöglicht eine Isolierung von relevanten Inhalten, die über die Standpunkte bzw. Schwerpunkte der jeweiligen Parteien hinausgehen. Dadurch können Ereignisse und Themen bestimmt werden, die unabhängig von sonstigen Inhalten der jeweiligen Parteien diskutiert wurden.

Zur Veranschaulichung wurden die letzten 100 Wochen ausgewertet und alle Parteien individuell mit der Tagesschau verglichen. Pro Woche wurden alle extrahierten Themen der Parteien mit denen der Tagesschau abgeglichen. Somit sollen unparteiische Inhalte erkannt werden, die von den Parteien aufgegriffen wurden. In Abbildung 4.15 ist der Verlauf der Dynamik von ähnlichen Themen der Tagesschau zur CDU/CSU dargestellt. Dabei wird die höchste Ähnlichkeit zwischen zwei Themen pro Woche dargestellt.

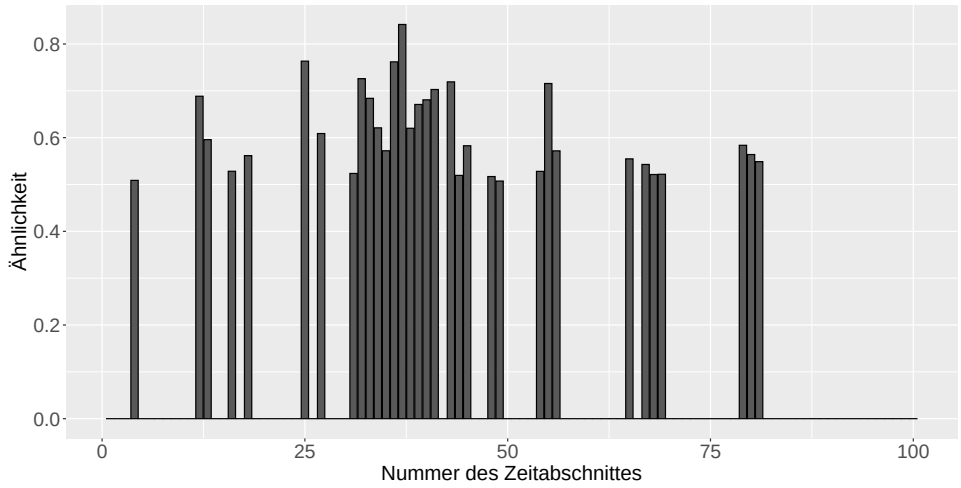

Abbildung 4.15: Darstellung der Ähnlichkeit der CDU/CSU und der Tagesschau in den letzten 100 Wochen

Die CDU/CSU behandelte über den betrachteten Zeitraum häufig ähnliche Inhalte wie die Tagesschau. Dabei sind zyklische Wiederholungen in den Häufigkeiten zu erkennen. Trotz verschiedener Autoren sind die jeweiligen Ähnlichkeiten sehr hoch.

In Abbildung 4.16 ist der Verlauf der Themenentwicklung der Tagesschau zur AfD dargestellt.

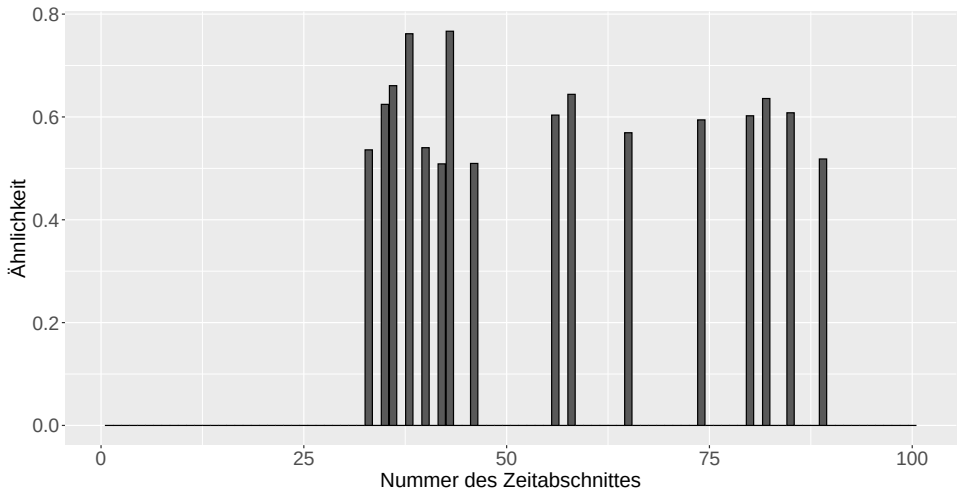

Abbildung 4.16: Darstellung der Ähnlichkeit der AfD und der Tagesschau in den letzten 100 Wochen

Auf Grund der späteren Nutzung von Twitter sind in den ersten 32 Wochen keine Inhalte der AfD vorhanden. Erst ab März 2020 veröffentlichte die Partei regelmäßig Inhalte. Dabei traten in den Inhalten weniger Gemeinsamkeiten mit der Tagesschau als bei der CDU/CSU auf. Eine Regelmäßigkeit im Aufgreifen gleicher Inhalte ist zu erkennen.

Abschließend folgt in Abbildung 4.17 der Vergleich von der Tagesschau mit der Linken.

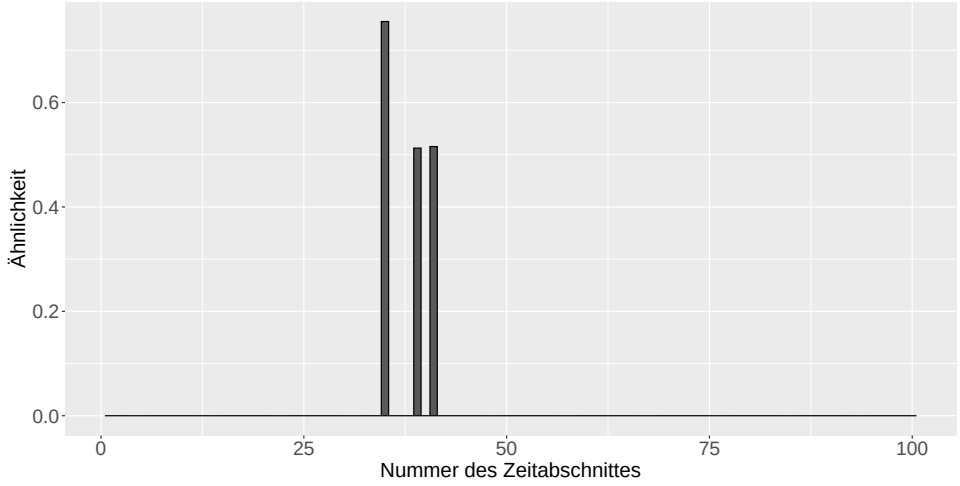

Abbildung 4.17: Darstellung der Ähnlichkeit der Linken und der Tagesschau in den letzten 100 Wochen

Im Unterschied zu den vorherigen Parteien zeigen sich hier kaum Schnittmengen. Über den Zeitraum der letzten 100 Wochen hat Die Linke in drei Wochen dieselben Themen wie die Tagesschau aufgegriffen.

Zusammenfassend zeigt dieser Vergleich der Profile, dass die Themen unabhängig vom Autor wiedergefunden werden können. Somit können über verschiedene Profile hinweg Gemeinsamkeiten in den angesprochenen Inhalten gefunden werden.

5 Zusammenfassung und Ausblick

Bisherige Untersuchungen von Themen und deren Dynamik beschränken sich größtenteils auf englischsprachige Texte. Deswegen ist eine Anpassung dieser Ansätze und Techniken auf deutschsprachige Texte notwendig. Nach erfolgreicher Adaption könnte dieses Vorgehen im forensischen Rahmen für eine Prädiktion von Themenverläufen eingesetzt werden.

Im Rahmen dieser wissenschaftlichen Abschlussarbeit wurde der Fokus auf die statistische und graphische Datenaufbereitung sowie die Untersuchung von Themendynamiken gelegt. Die Extraktion von Themen aus Tweets von Bundestagsparteien wurde dabei mittels LDA durchgeführt. Dabei wurden Parameter festgelegt, die das Verfahren für die vorliegenden Daten optimierten. Die Wahl der a-priori Parameter α und β sorgte für eine differenzierte Themen- und Termverteilung. Die Themenanzahl k wurde für jeden Iterationsschritt individuell durch eine Kombination verschiedener Optimierungsfunktionen bestimmt.

Als weiterer maßgeblicher Faktor für die Resultate der Themenextraktionen wurden verschiedene Vorverarbeitungen der Ausgangsdaten erprobt. Im Vergleich von Termen und Bigrammen ergab sich, dass die Nutzung von einzelnen Termen die besten und performantesten Ergebnisse lieferten. Weiterhin wurden verschiedene Maßzahlen zur Beurteilung der Ähnlichkeiten der extrahierten Themen verglichen. Dabei stellte sich heraus, dass die Kosinus-Ähnlichkeit die besten Ergebnisse unter den angewandten Maßen lieferte. Somit wurde dieses Maß mit der Kullback-Leibler-Divergenz für die weitere Betrachtung von Themenähnlichkeiten verwendet.

Ein weiterer Schritt stellte die Auswahl geeigneter Zeitfenster für die Gruppierung der Tweets dar. Dadurch sollten einerseits zu wenige Daten durch Zusammenfügen ausgeglichen werden, andererseits sollte trotzdem eine ausreichende Granularität gewährleistet sein. Dafür wurden wöchentliche und monatliche Gruppierungen untersucht. Beide Zeitfenster erlaubten eine Extraktion und spätere Verfolgung von Themen. Durch das kleinere Zeitfenster bei wöchentlichen Analysen konnten Themen jedoch genauer beschrieben werden.

Zusätzlich wurde der Einfluss der Datenquelle auf die detektierten Themen untersucht. Tweets und Retweets lieferten keine grundsätzlich verschiedenen Themen, was zu einer kombinierten Nutzung führte. Die Anzahl an eigenen Inhalten ist bei manchen Profilen zu gering, als dass daraus zuverlässig Themen extrahiert werden können. Somit bietet die kombinierte Nutzung den Vorteil, kontinuierlicher Themen zu bestimmen.

Nachdem die beschriebenen Versuche durchgeführt wurden, konnten die Ergebnisse genutzt werden, um für alle Profile in jedem Zeitintervall die Themen zu bestimmen.

Dieser fortlaufende Extraktionsprozess ermöglichte die Beschreibung der Themendynamik. Das Verhalten der Themen wurde durch deren Ähnlichkeit in den verschiedenen Zeitintervallen beschrieben. Dazu wurden mittels angepassten Sankey-Diagrammen die Themenflüsse jedes Profils visualisiert.

Neben der Beschreibung von Themenverläufen innerhalb eines Autors wurde die Verbindung zwischen den Profilen untersucht. Dafür wurden über einen definierten Zeitraum die extrahierten Themen verglichen und graphisch dargestellt. Dabei wurde zusätzlich die Themenintensität genutzt. Diese Maßzahl kann die Relevanz eines Themas anhand der Anzahl derjenigen Dokumente beschreiben, die das Thema behandelten.

Weitere Untersuchungen waren die Bestimmung von Haupt- und Unterthemen. Diese Unterscheidung konnte anhand des Rankings der relevantesten Terme pro Thema getroffen werden. Durch die Charakterisierung von Unterthemen konnte ein detaillierterer Themenverlauf beschrieben werden, der in Haupt- und Unterthemenentwicklung differenziert. Daran schloss die Beurteilung der Eignung von Hashtags zur Themenbeschreibung. Diese stellen eine Möglichkeit der Erweiterung der LDA dar, sind aber nicht als alleinige Themenbeschreibung zulässig. In Fällen, in denen die Autoren einzelne Hashtags sehr häufig einsetzen, definierten die Hashtags Oberthemen in den Extraktionen.

Diese Abschlussarbeit wurde im Rahmen umfangreicher Untersuchungen und Bearbeitungen erstellt. Über den Rahmen der Arbeit hinaus werden die gewonnen Erkenntnisse angewandt. Im Hinblick auf die kommende Bundestagswahl im September 2021 werden die Themenverläufe aller aktuellen Bundestagsparteien inhaltlich untersucht. Weiterhin werden die Erkenntnisse dazu genutzt, das prädiktive Potential durch das Erlernen von Themenverläufen zu nutzen. Dieser Ansatz kann unter anderem durch eine Anwendung von Hidden Markov Modellen umgesetzt werden.

Die gewonnen Erkenntnisse können weiterhin mit Sentimentanalysen kombiniert werden. Dadurch kann der Standpunkt des Autors beurteilt werden. Außerdem ermöglicht dieser Ansatz eine Identifizierung von kritischen Themenentwicklungen. Somit könnten toxische Strömungen frühzeitig erkannt werden.

Zusammenfassend war es möglich, einen ersten Ansatz zur Beschreibung von Themenverläufen in deutschsprachigen Texten zu etablieren. Dieser kann nun in den weiteren Schritten ausgebaut werden, um die genannten Anwendungen zu erproben.

Anhang A: Anhang

A.1 Übersicht zu den Landtagsparteien

CDU - Sachsen-Anhalt

1. Zeitraum: 28.12.2011-24.06.2021[6]
2. Anzahl Tweets: 3.244
3. Tweets/Retweets: 1.469/1.525
4. Durchschnitt Woche: 7; Monat: 28
5. Häufigste Anzahl an Hashtags pro Tweet: 0 (906); Tweets mit einem Hashtag: 744
6. Abbildung A.1 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.2 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.3, A.4 und A.5 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

[6] https://www.twitter.com/CDUlsa

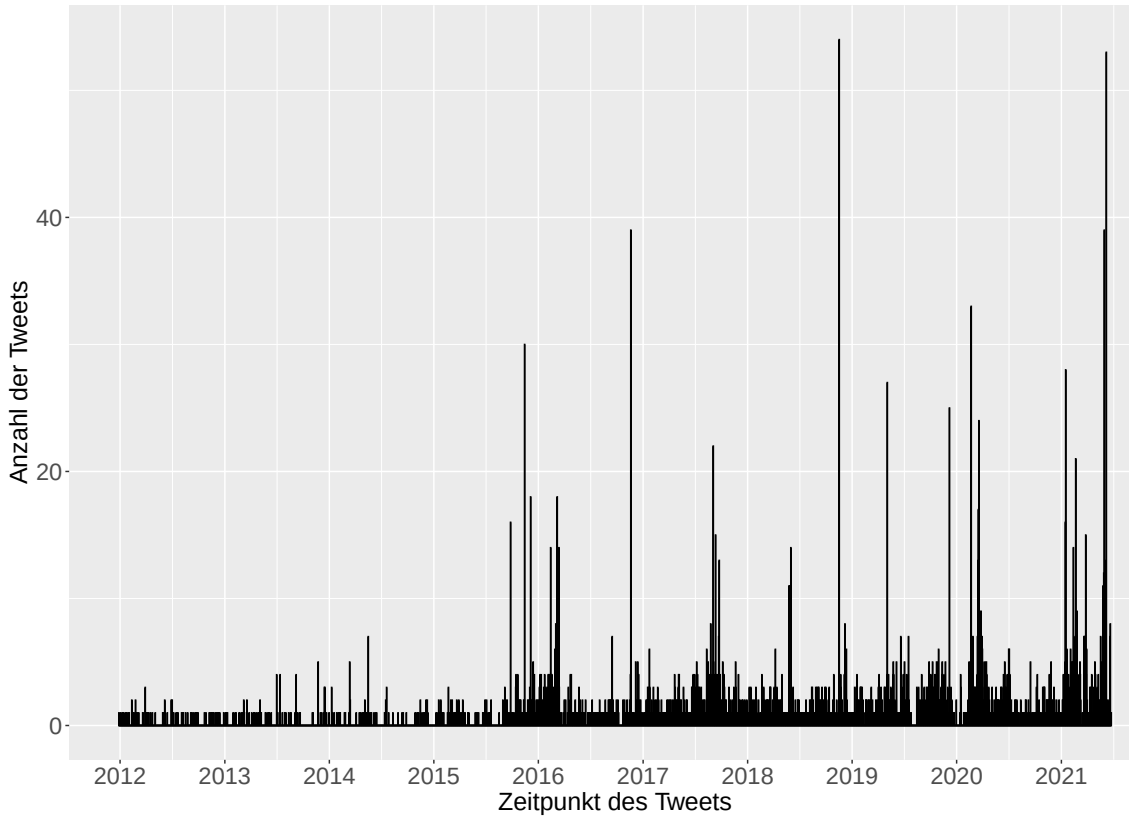

Abbildung A.1: Darstellung der Anzahl an veröffentlichten Tweets der CDU Sachsen-Anhalt seit 2011

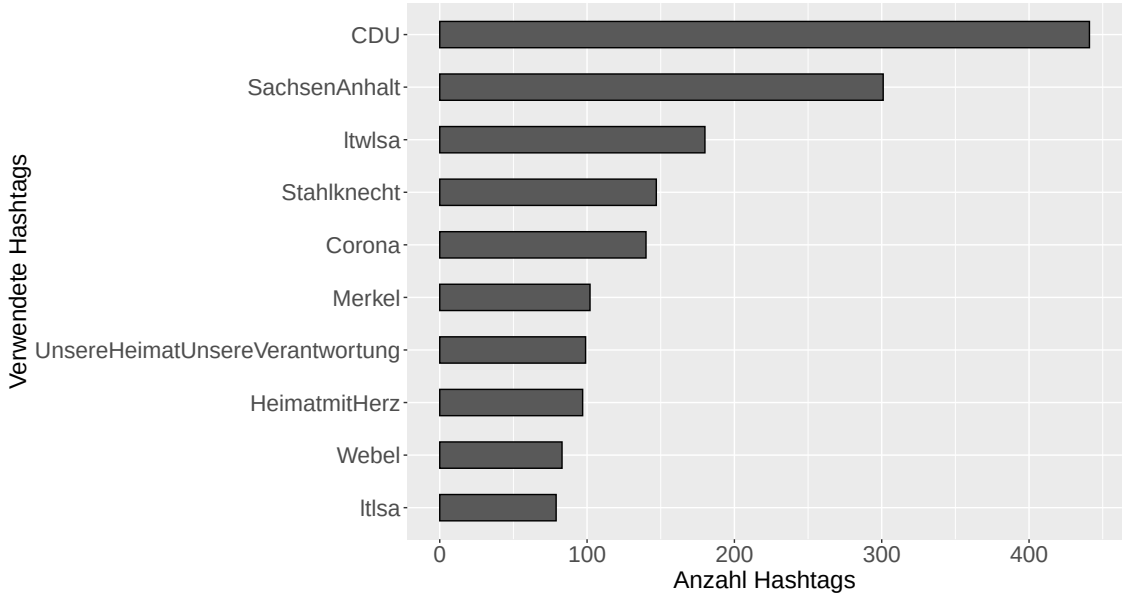

Abbildung A.2: Darstellung der am häufigsten verwendeten Hashtags der CDU Sachsen-Anhalt

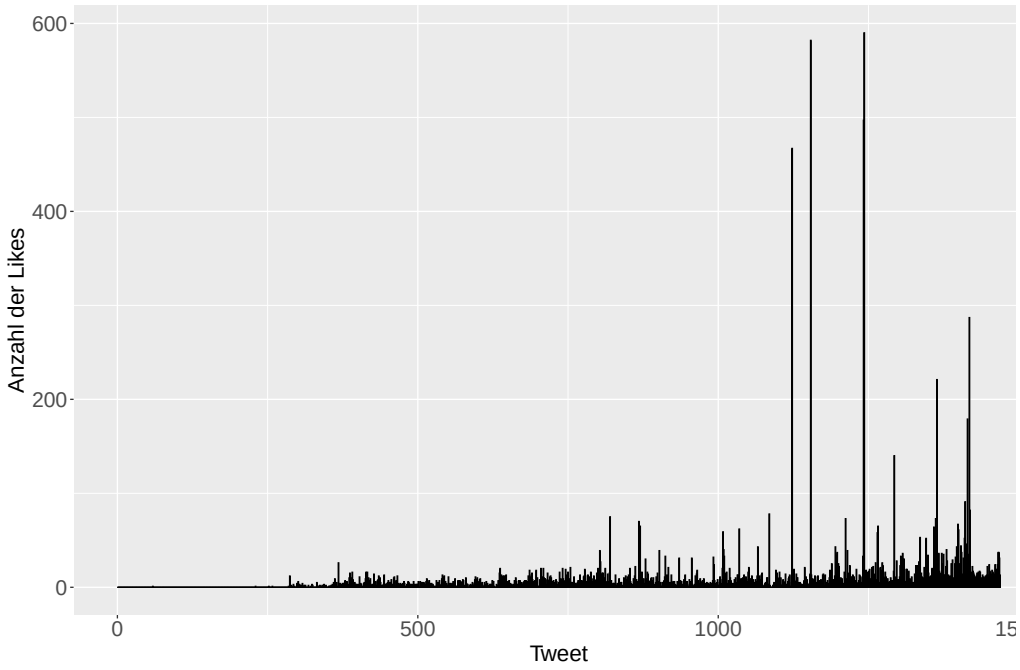

Abbildung A.3: Darstellung der Likes als Nutzerreaktion über alle Tweets der CDU Sachsen-Anhalt

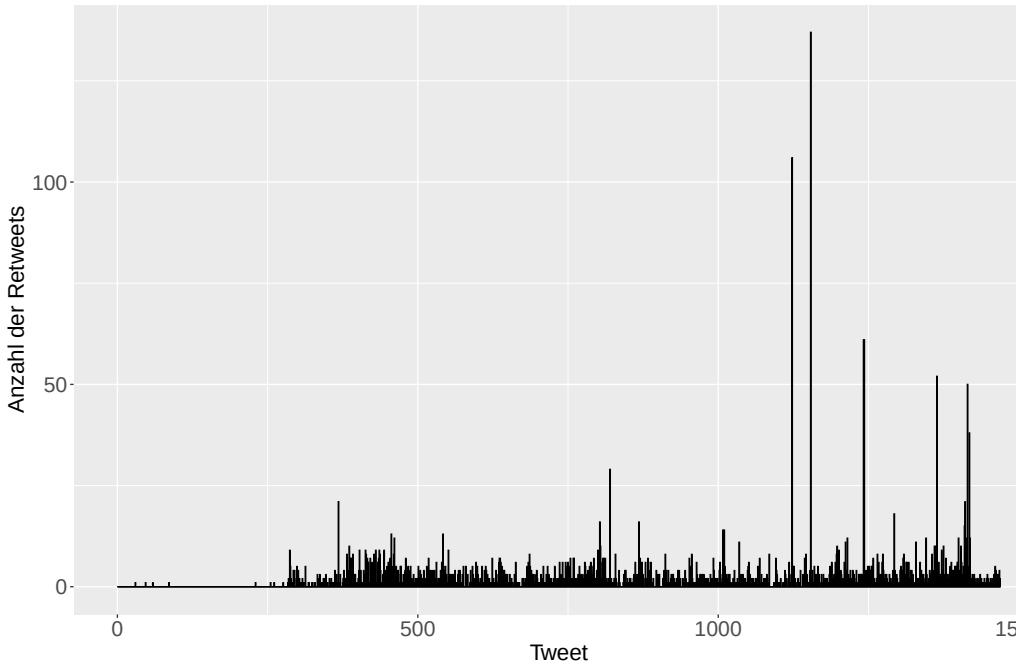

Abbildung A.4: Darstellung der Retweets als Nutzerreaktion über alle Tweets der CDU Sachsen-Anhalt

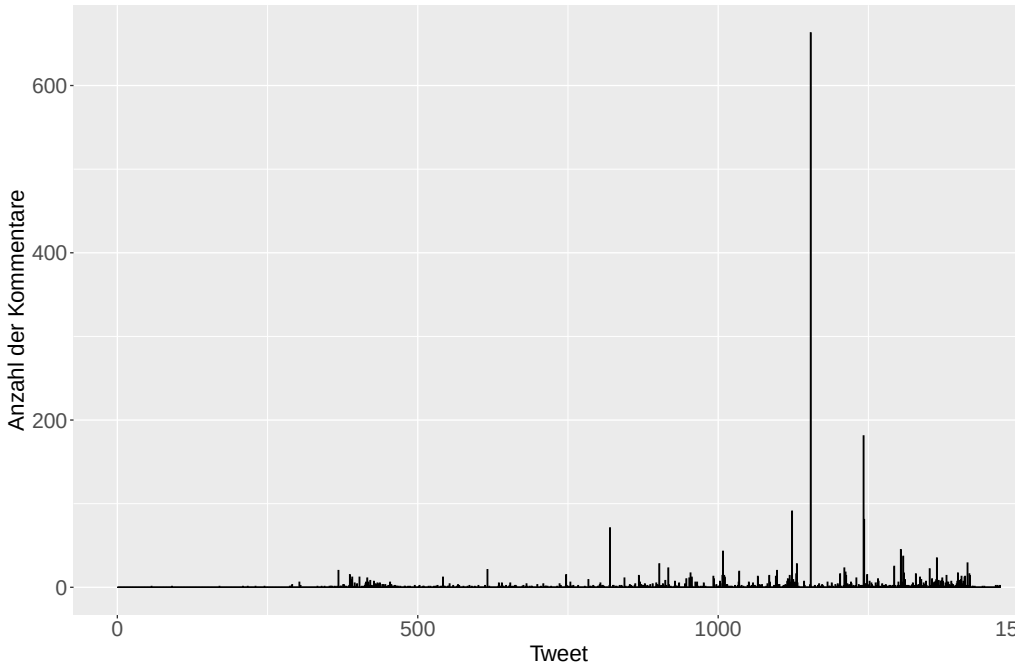

Abbildung A.5: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der CDU Sachsen-Anhalt

CDU - Niedersachsen

1. Zeitraum: 27.08.2008-26.06.2021[7]
2. Anzahl Tweets: 2.475
3. Tweets/Retweets: 1.412/843
4. Durchschnitt Woche: 5; Monat: 17
5. Häufigste Anzahl an Hashtags pro Tweet: 1 (831); Tweets mit keinem Hashtag: 630
6. Abbildung A.6 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.7 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.8, A.9 und A.10 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

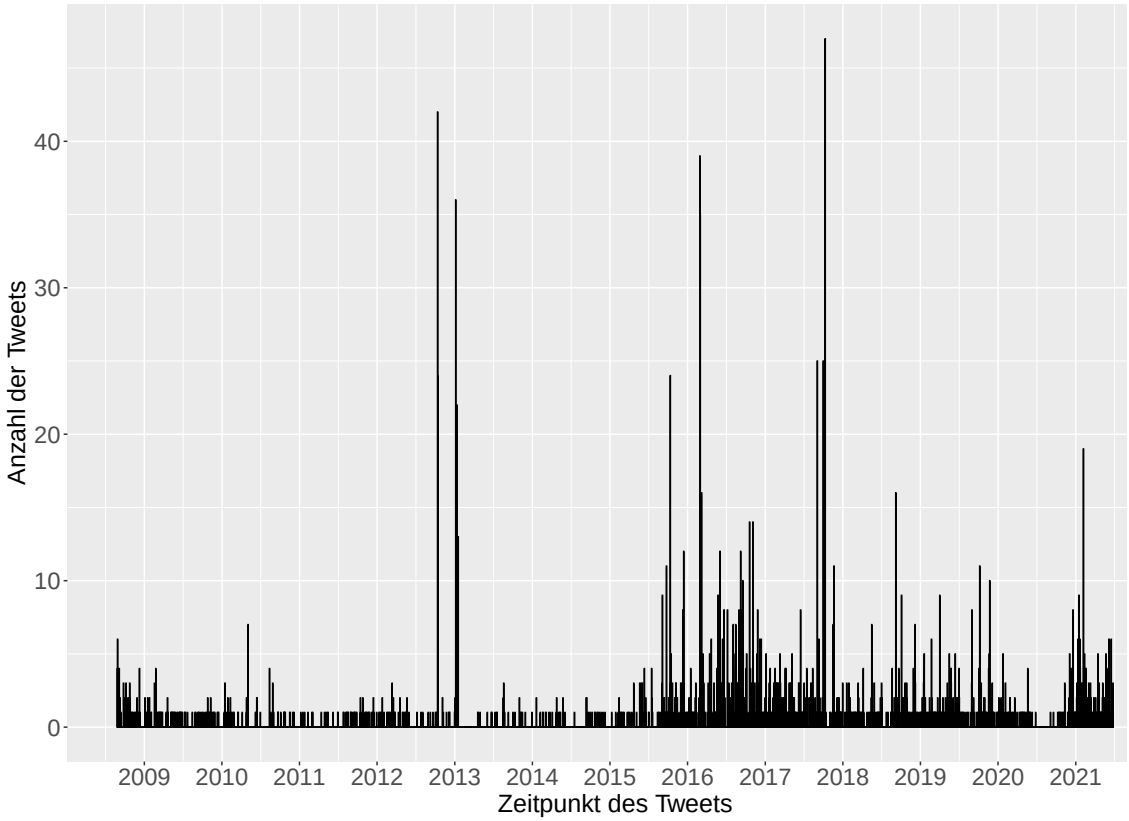

Abbildung A.6: Darstellung der Anzahl an veröffentlichten Tweets der CDU Niedersachsen seit 2008

[7] https://www.twitter.com/CDUNds

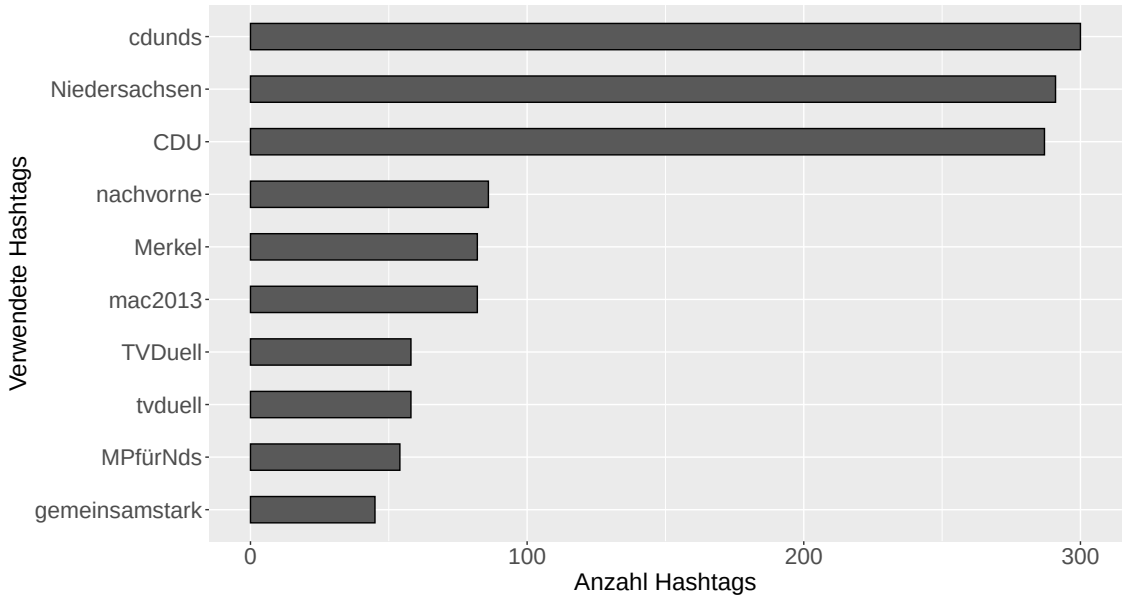

Abbildung A.7: Darstellung der am häufigsten verwendeten Hashtags der CDU Niedersachsen

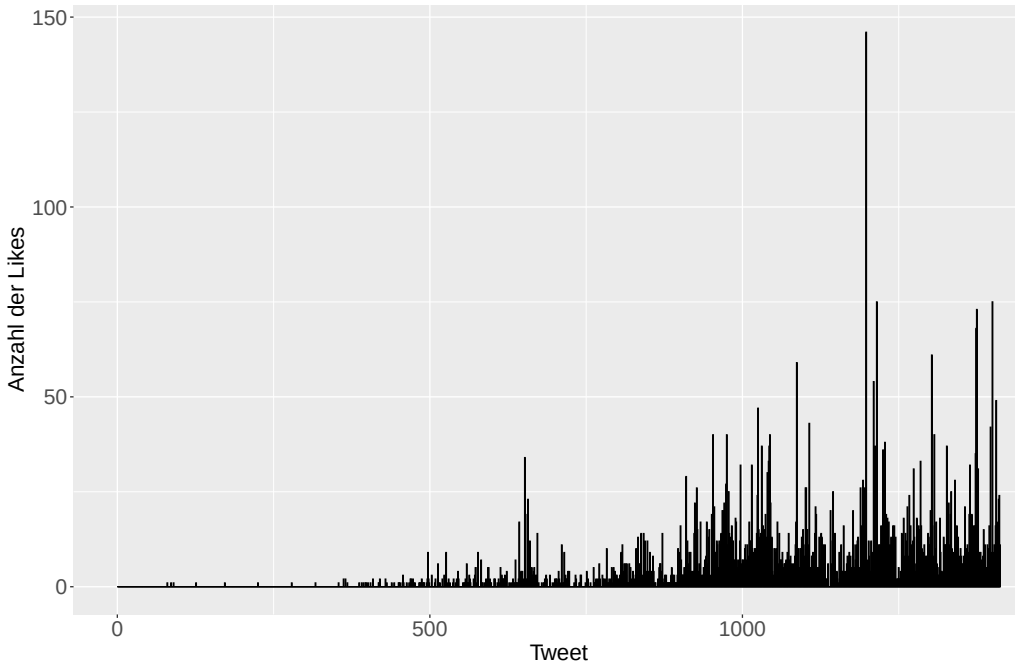

Abbildung A.8: Darstellung der Likes als Nutzerreaktion über alle Tweets der CDU Niedersachsen

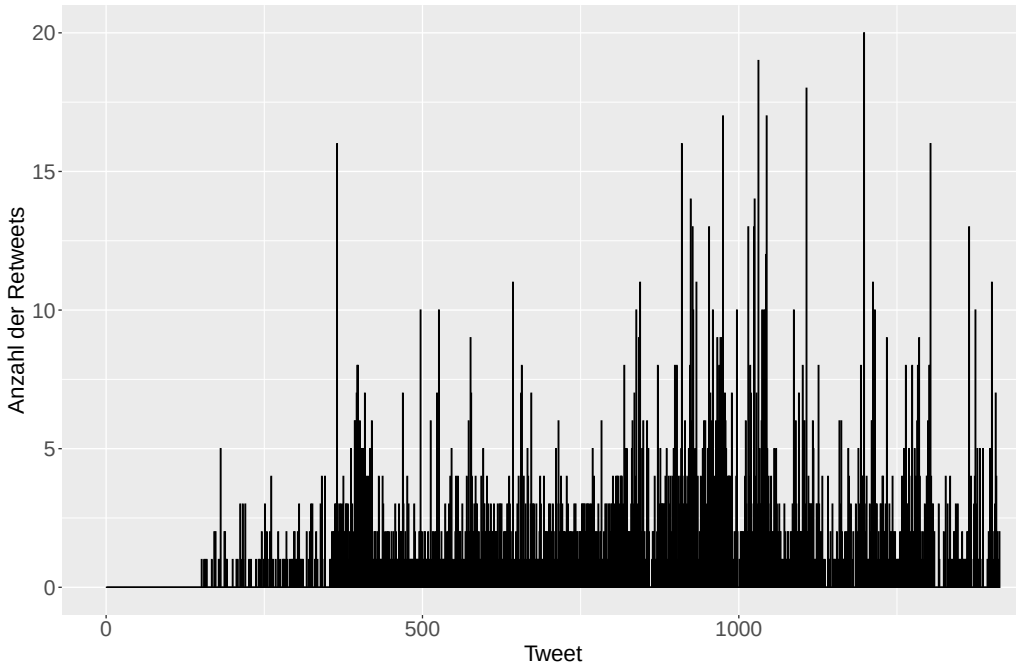

Abbildung A.9: Darstellung der Retweets als Nutzerreaktion über alle Tweets der CDU Niedersachsen

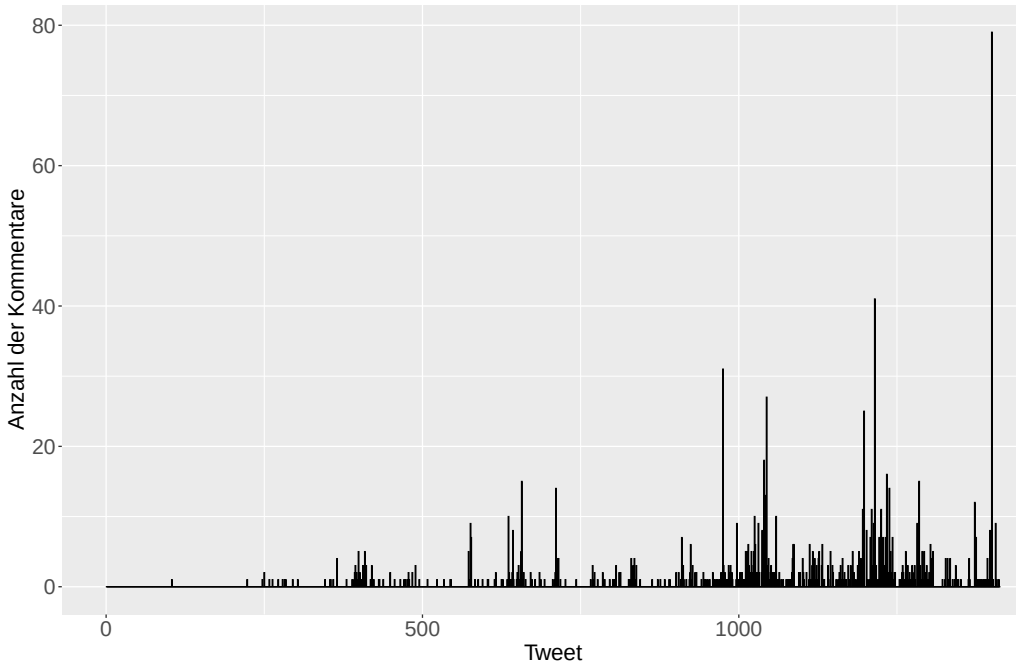

Abbildung A.10: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der CDU Niedersachsen

AfD - Sachsen-Anhalt

1. Zeitraum: 20.03.2016-26.06.2021[8]
2. Anzahl Tweets: 3.888
3. Tweets/Retweets: 103/3.780
4. Durchschnitt Woche: 17; Monat: 69
5. Häufigste Anzahl an Hashtags pro Tweet: 2 (582); Tweets mit keinem Hashtag: 380
6. Abbildung A.11 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.12 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.13, A.14 und A.15 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

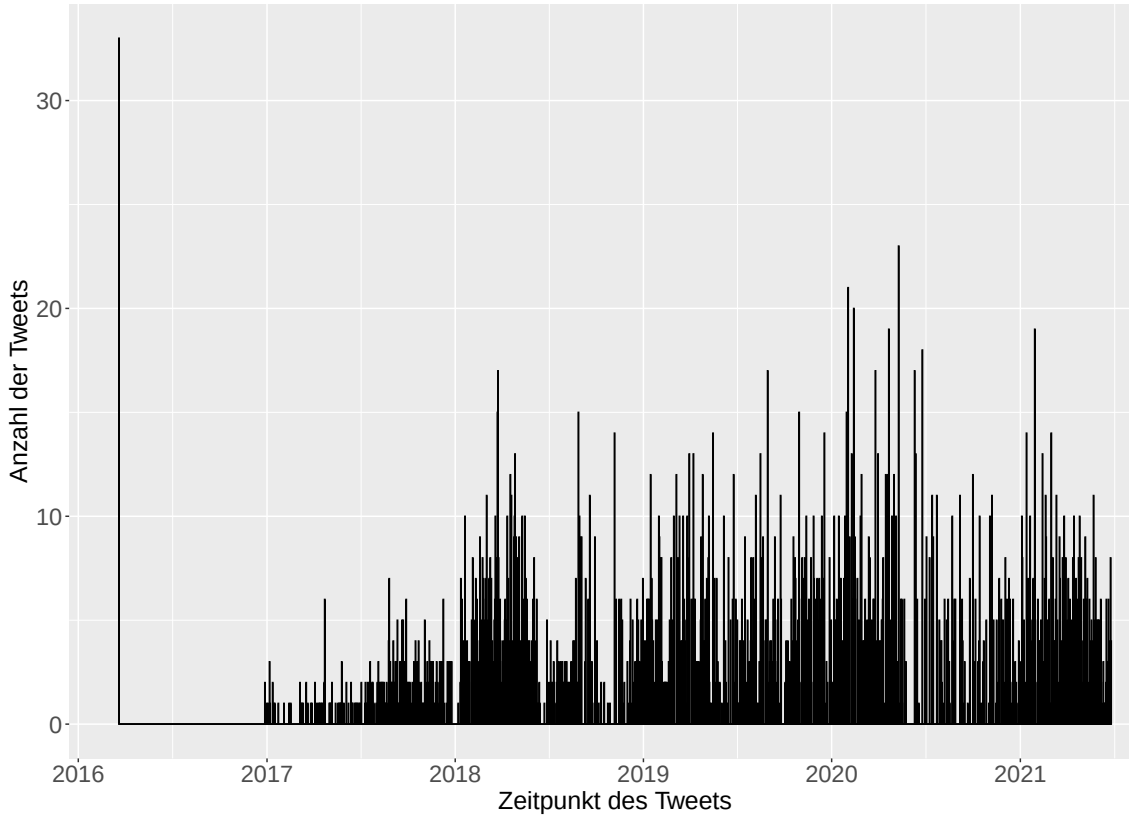

Abbildung A.11: Darstellung der Anzahl an veröffentlichten Tweets der AfD Sachsen-Anhalt seit 2016

[8] https://www.twitter.com/afd_lsa

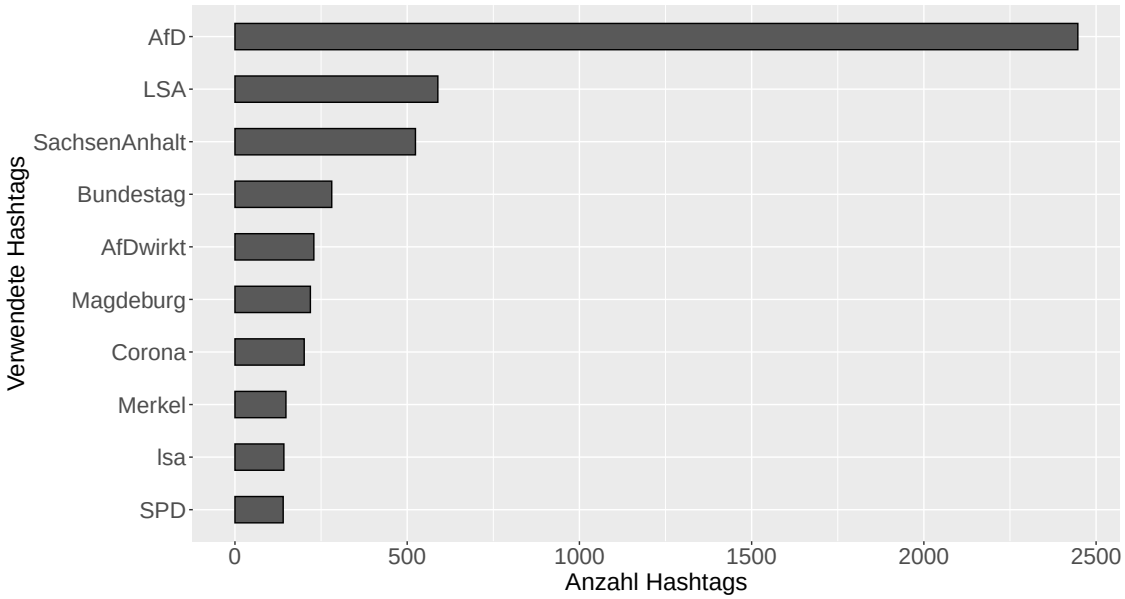

Abbildung A.12: Darstellung der am häufigsten verwendeten Hashtags der AfD Sachsen-Anhalt

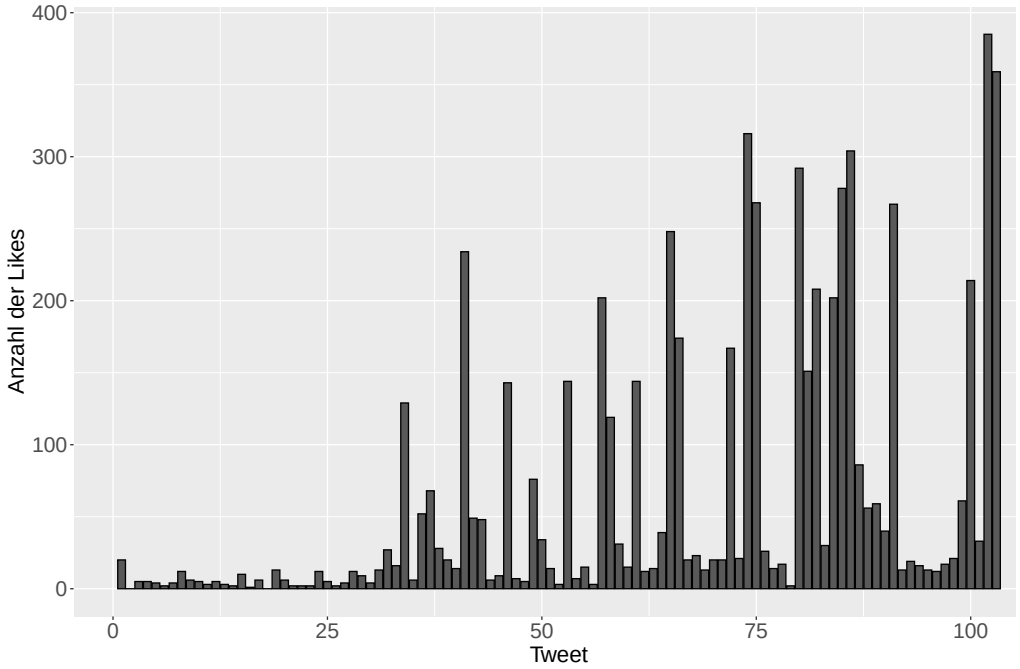

Abbildung A.13: Darstellung der Likes als Nutzerreaktion über alle Tweets der AfD Sachsen-Anhalt

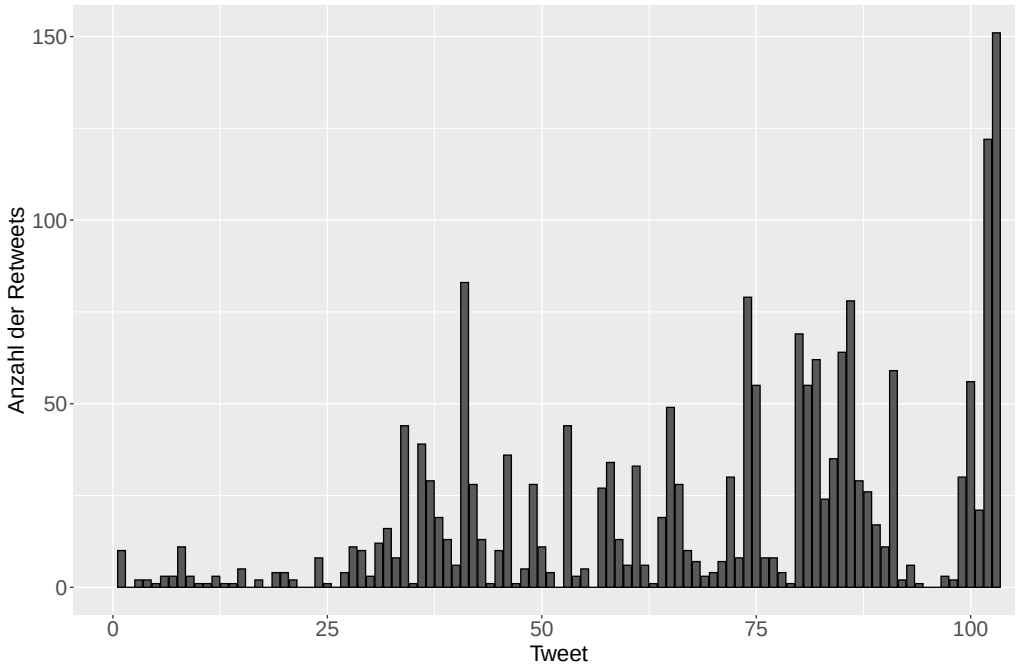

Abbildung A.14: Darstellung der Retweets als Nutzerreaktion über alle Tweets der AfD Sachsen-Anhalt

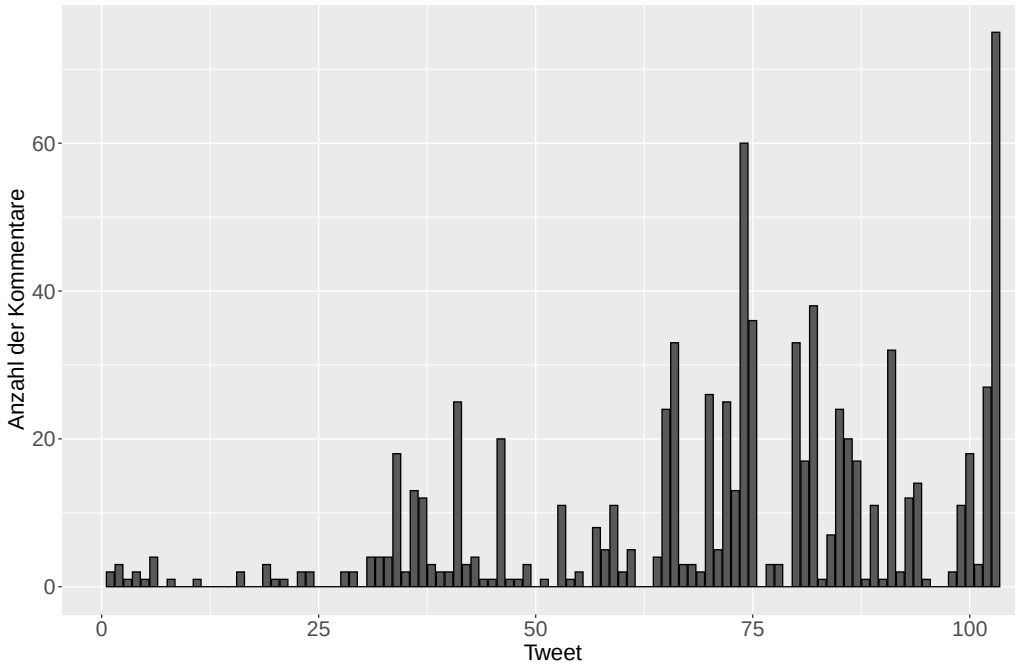

Abbildung A.15: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der AfD Sachsen-Anhalt

AfD - Niedersachsen

1. Zeitraum: 01.01.2019-06.06.2021[9]
2. Anzahl Tweets: 659
3. Tweets/Retweets: 143/201
4. Durchschnitt Woche: 7; Monat: 24
5. Häufigste Anzahl an Hashtags pro Tweet: 0 (247); Häufigste Anzahl unter Tweets mit Hashtags: 3 (105)
6. Abbildung A.16 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.17 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.18, A.19 und A.20 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

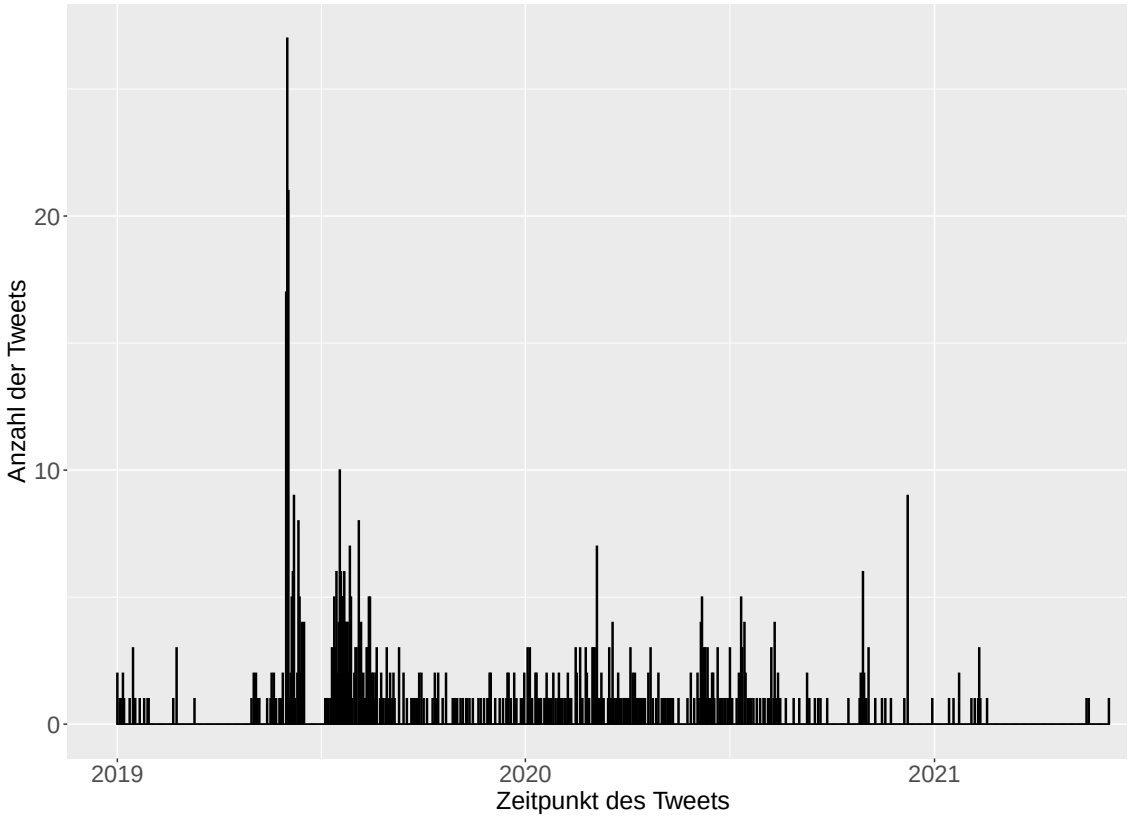

Abbildung A.16: Darstellung der Anzahl an veröffentlichten Tweets der AfD Niedersachsen seit 2019

[9] https://www.twitter.com/afdnds

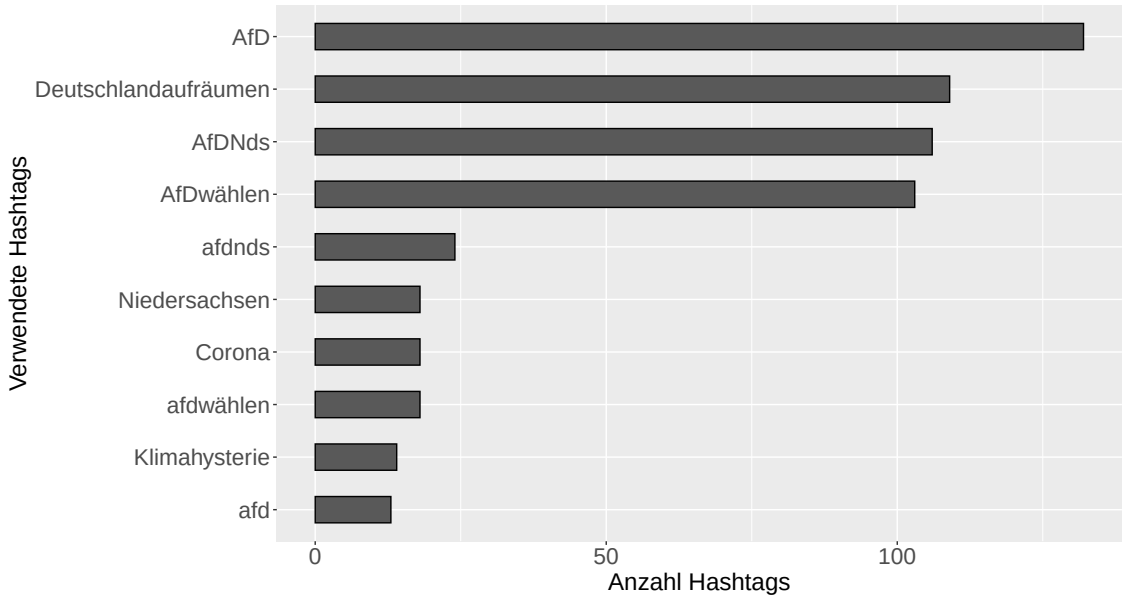

Abbildung A.17: Darstellung der am häufigsten verwendeten Hashtags der AfD Niedersachsen

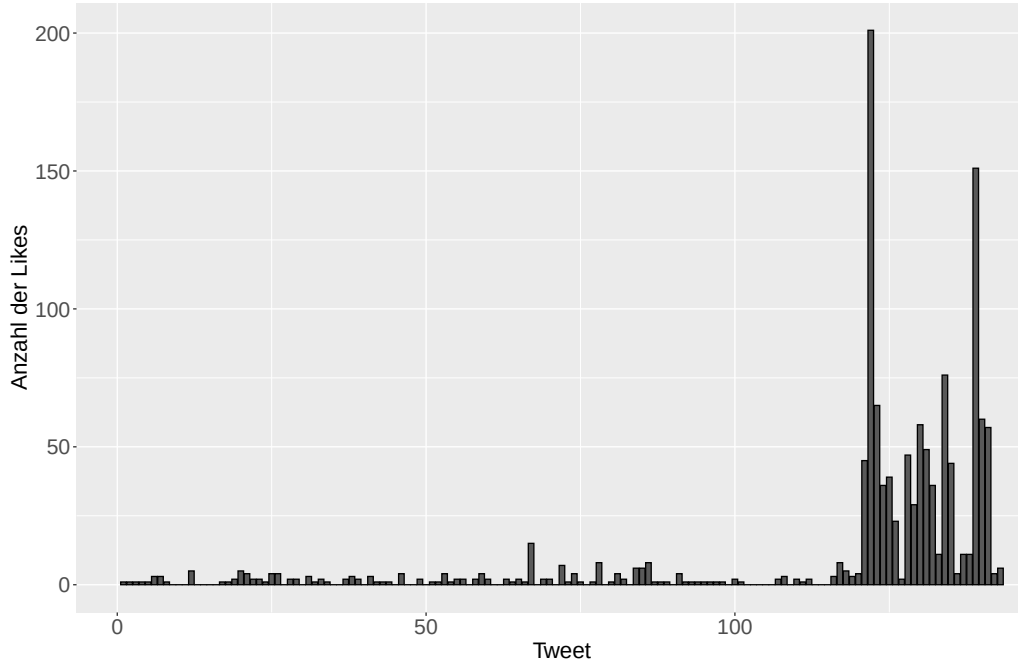

Abbildung A.18: Darstellung der Likes als Nutzerreaktion über alle Tweets der AfD Niedersachsen

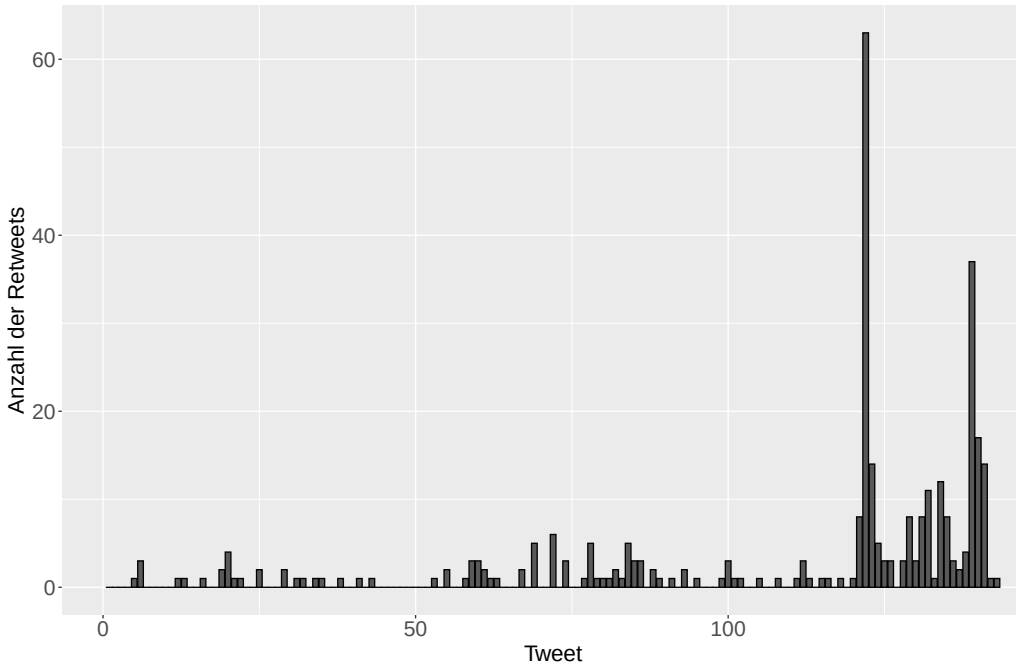

Abbildung A.19: Darstellung der Retweets als Nutzerreaktion über alle Tweets der AfD Niedersachsen

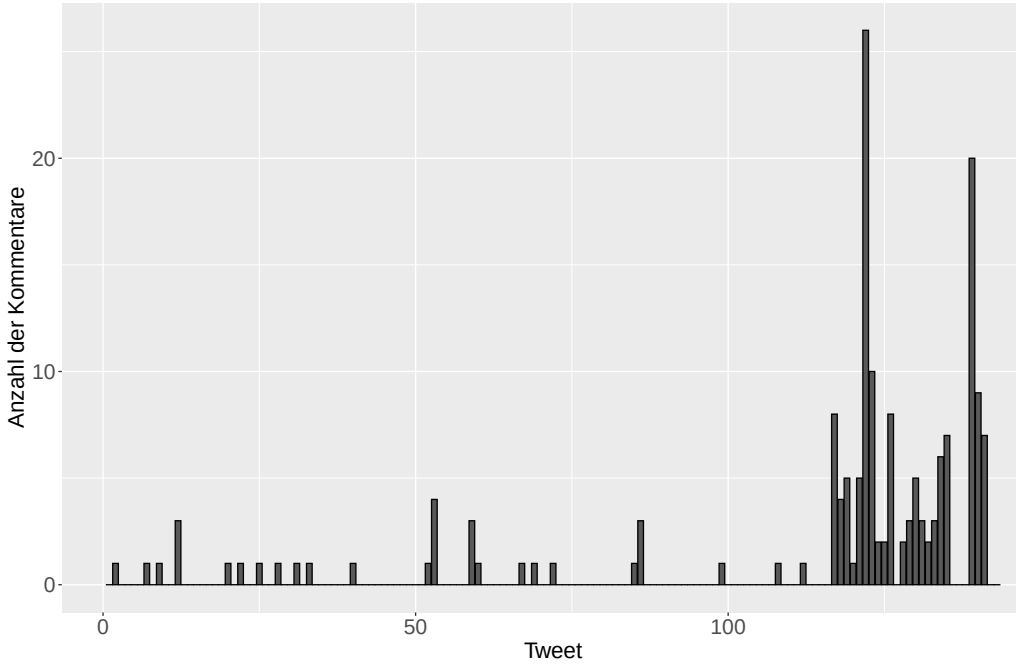

Abbildung A.20: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der AfD Niedersachsen

Die Linke - Sachsen-Anhalt

1. Zeitraum: 20.07.2012-24.06.2021[10]
2. Anzahl Tweets: 6.573
3. Tweets/Retweets: 2.357/3.612
4. Durchschnitt Woche: 16; Monat: 61
5. Häufigste Anzahl an Hashtags pro Tweet: 0 (2.432); Häufigste Anzahl unter Tweets mit Hashtags: 1 (2.134)
6. Abbildung A.21 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.22 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.23, A.24 und A.25 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

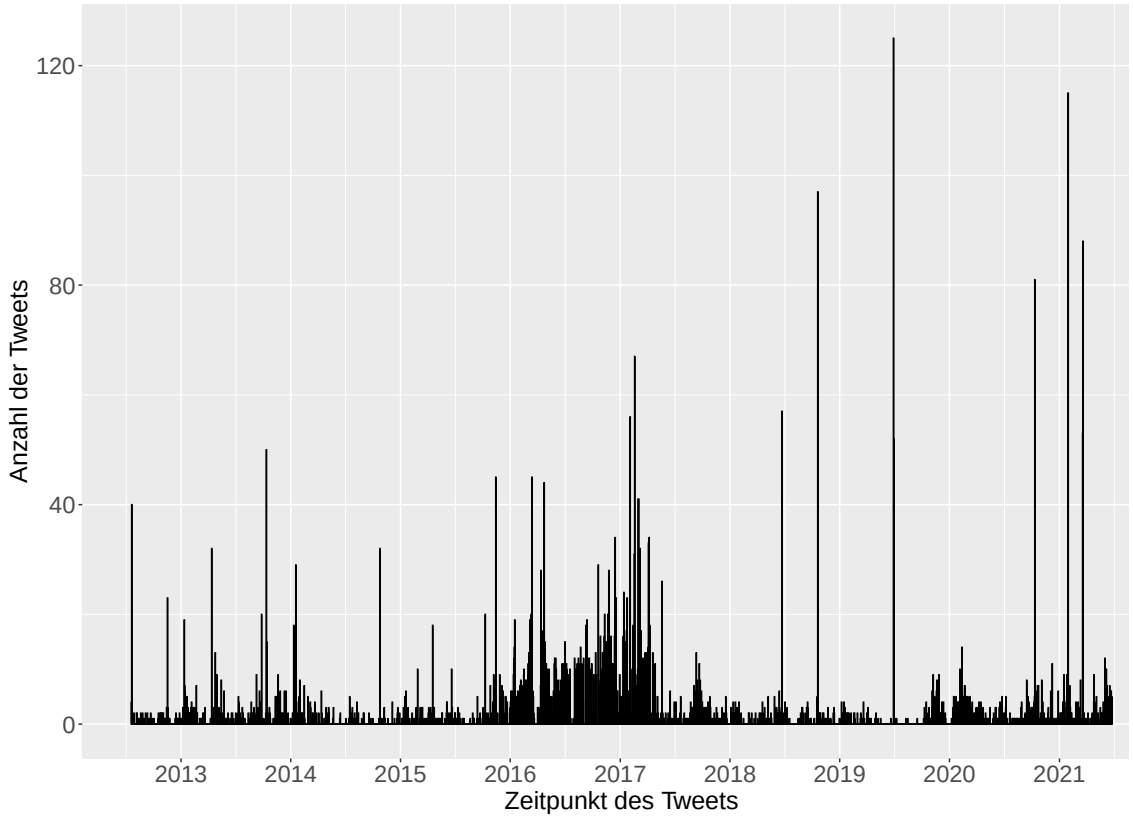

Abbildung A.21: Darstellung der Anzahl an veröffentlichten Tweets der Linken Sachsen-Anhalt seit 2012

[10] https://www.twitter.com/dielinkelsa

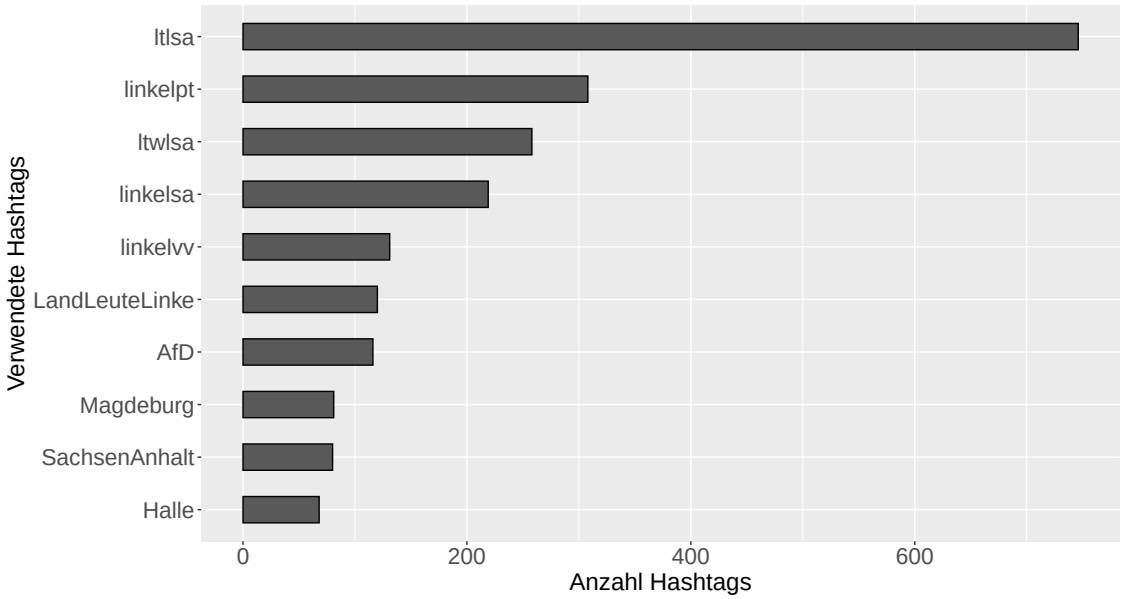

Abbildung A.22: Darstellung der am häufigsten verwendeten Hashtags der Linken Sachsen-Anhalt

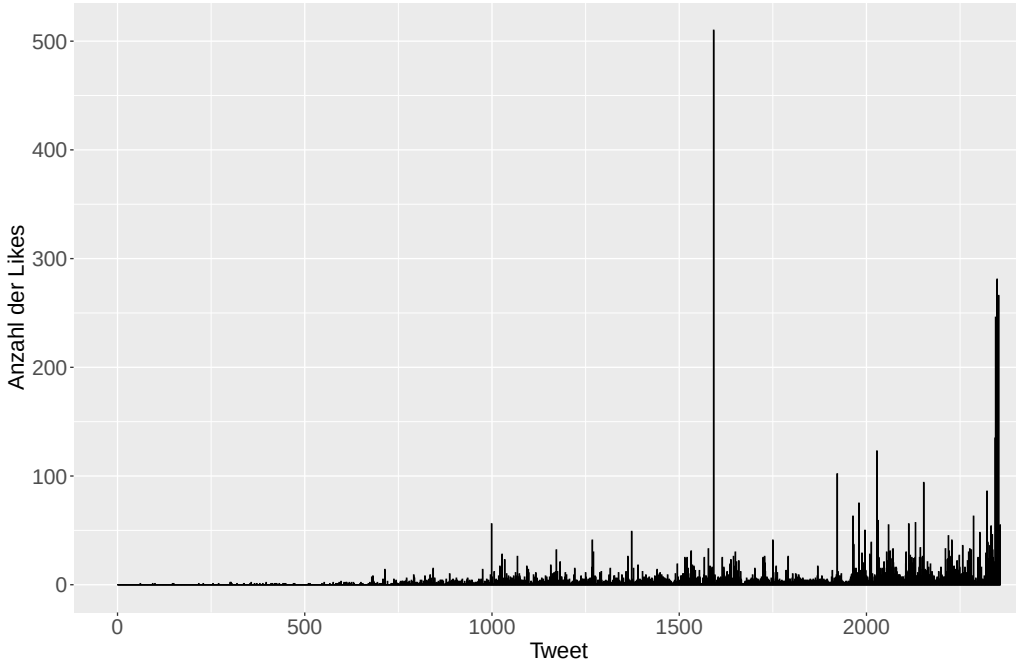

Abbildung A.23: Darstellung der Likes als Nutzerreaktion über alle Tweets der Linken Sachsen-Anhalt

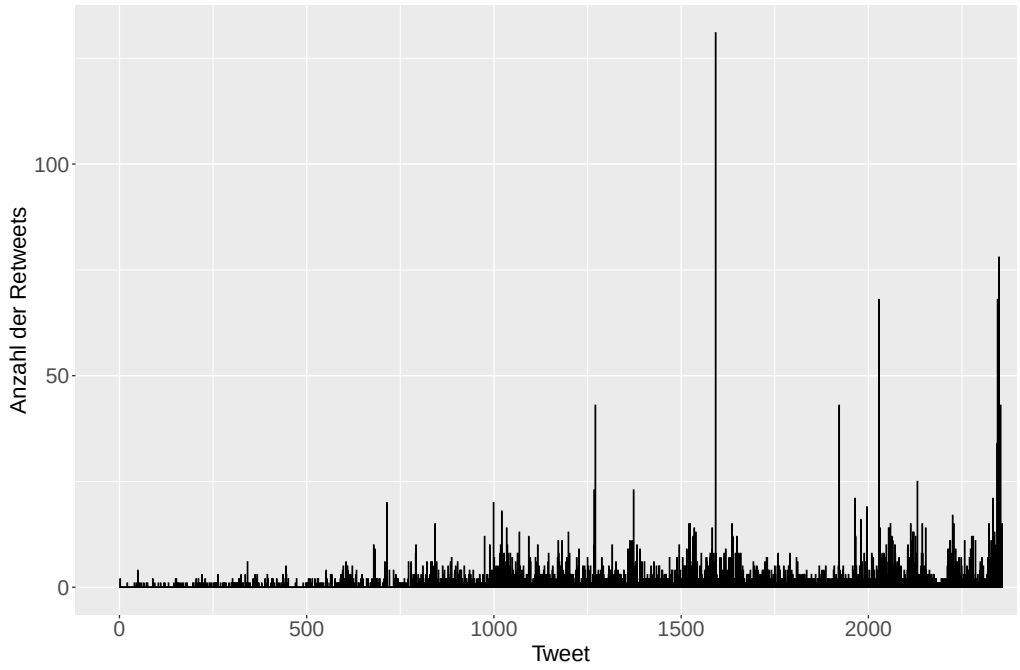

Abbildung A.24: Darstellung der Retweets als Nutzerreaktion über alle Tweets der Linken Sachsen-Anhalt

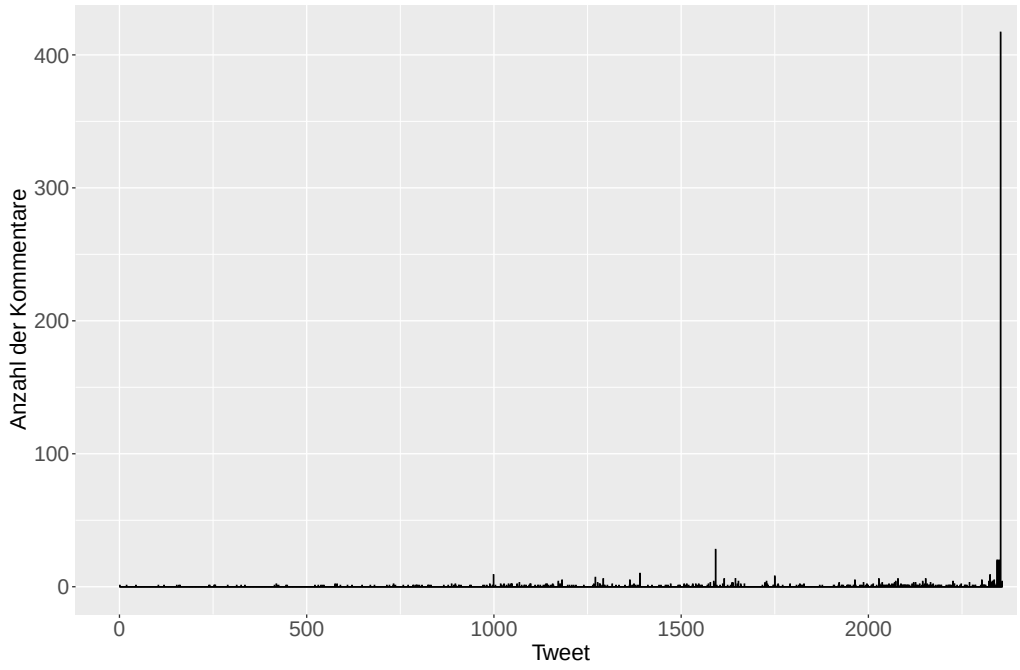

Abbildung A.25: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der Linken Sachsen-Anhalt

Die Linke - Niedersachsen

1. Zeitraum: 06.12.2012-26.06.2021[11]
2. Anzahl Tweets: 4.494
3. Tweets/Retweets: 397/3.959
4. Durchschnitt Woche: 18; Monat: 70
5. Häufigste Anzahl an Hashtags pro Tweet: 0 (1.192); Häufigste Anzahl unter Tweets mit Hashtags: 1 (1.038)
6. Abbildung A.26 zeigt den Verlauf der Tweets über die Zeit
7. Abbildung A.27 zeigt die zehn häufigsten Hashtags
8. Abbildungen A.28, **??** und A.30 zeigen die Likes, Retweets und Replies auf die eigenen Inhalte der Partei

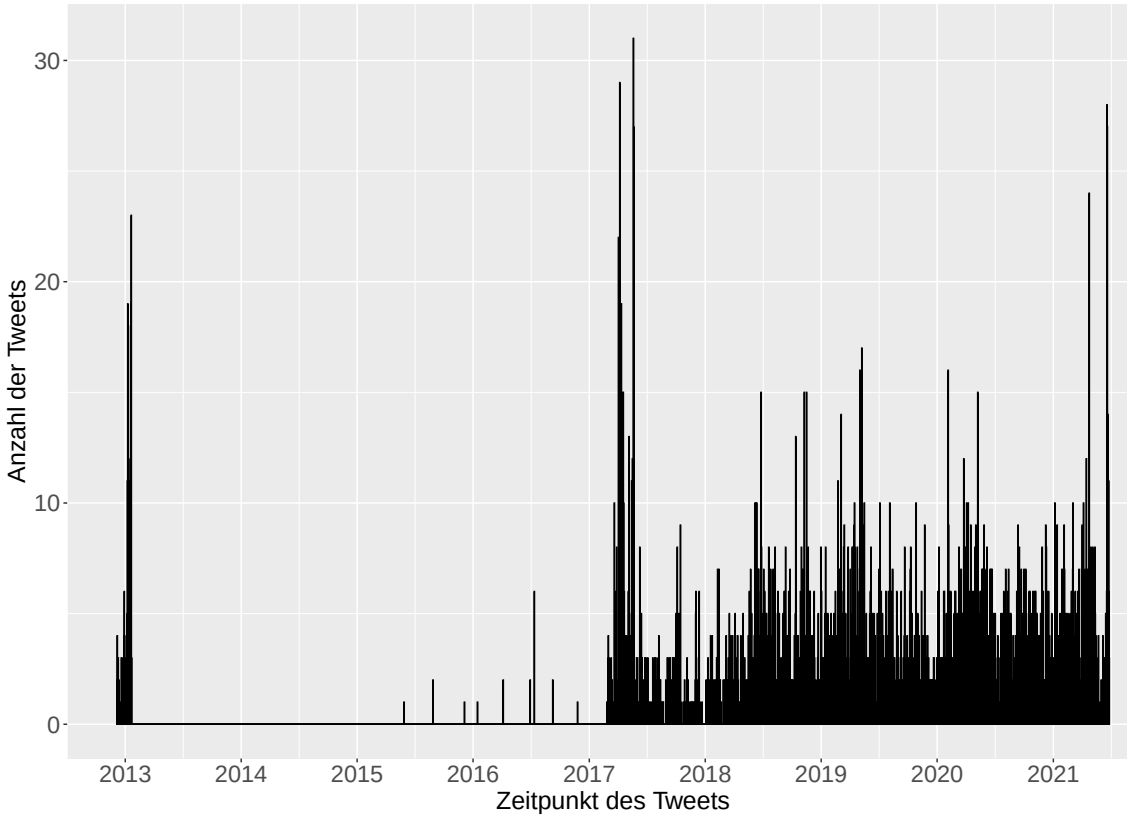

Abbildung A.26: Darstellung der veröffentlichten Tweets der Linken Niedersachsen seit 2012

[11] https://www.twitter.com/die_linke_nds

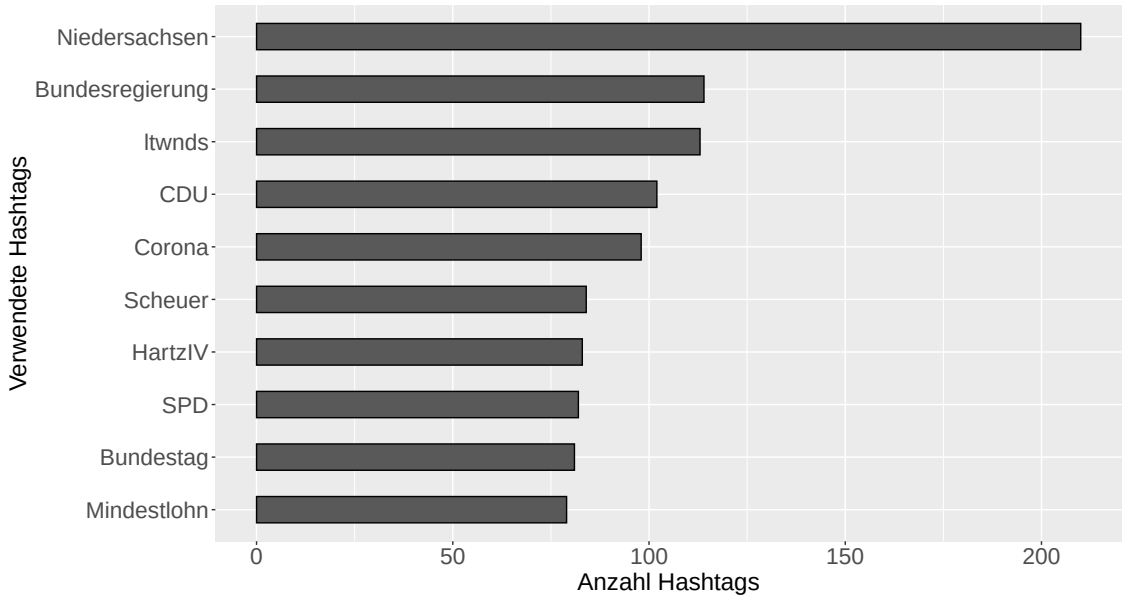

Abbildung A.27: Darstellung der am häufigsten verwendeten Hashtags der Linken Niedersachsen

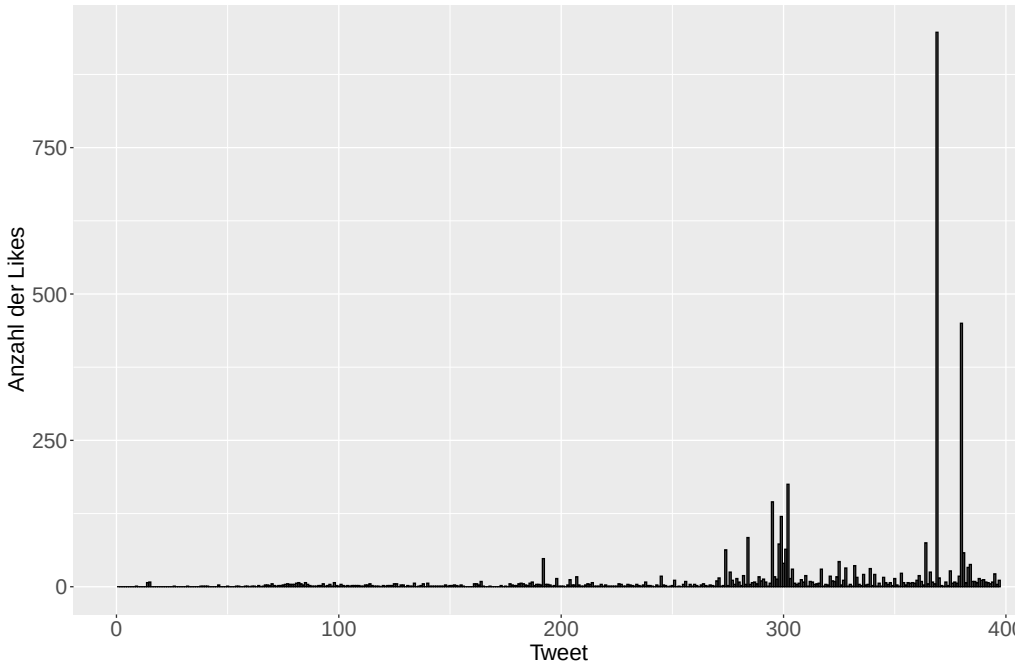

Abbildung A.28: Darstellung der Likes als Nutzerreaktion über alle Tweets der Linken Niedersachsen

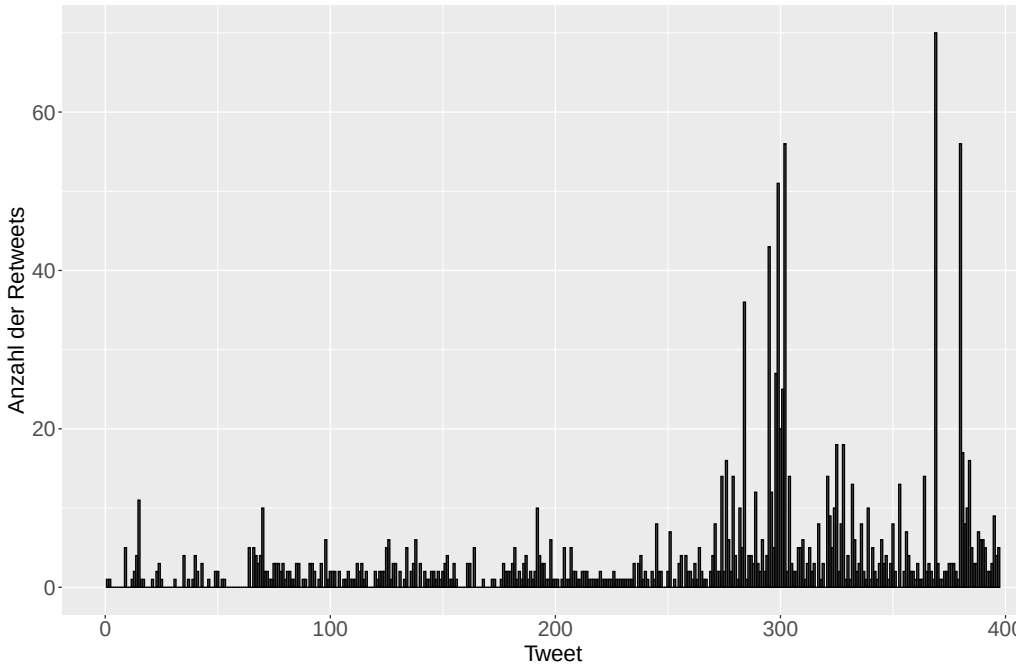

Abbildung A.29: Darstellung der Retweets als Nutzerreaktion über alle Tweets der Linken Niedersachsen

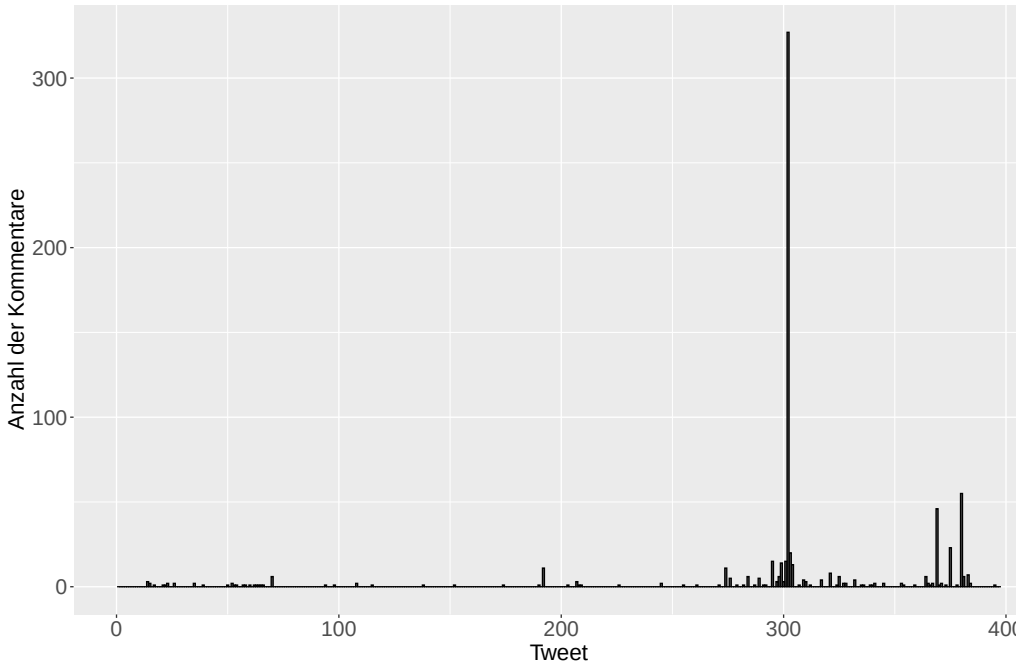

Abbildung A.30: Darstellung der Kommentare als Nutzerreaktion über alle Tweets der Linken Niedersachsen

A.2 Anzahl extrahierter Themen pro Monat aller Profile

CDU/CSU

In Abbildung A.31 ist der Verlauf der Anzahl an extrahierten Themen der CDU/CSU pro Monat dargestellt.

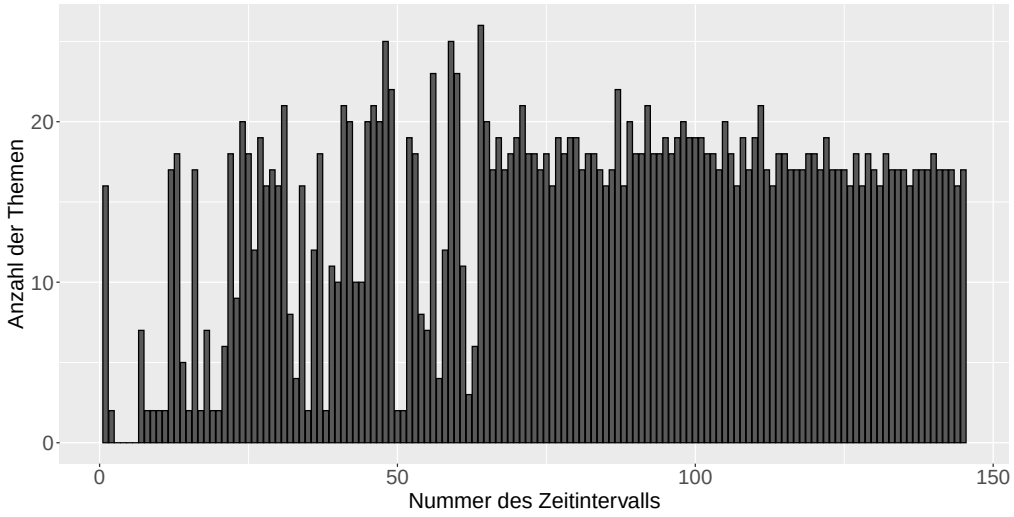

Abbildung A.31: Darstellung der Anzahl extrahierter Themen pro monatlichem Zeitabschnitt der CDU/CSU

AfD

In Abbildung A.32 ist der Verlauf der Anzahl an extrahierten Themen der AfD pro Monat dargestellt.

Abbildung A.32: Darstellung der Anzahl extrahierter Themen pro monatlichem Zeitabschnitt der AfD

Die Linke

In Abbildung A.33 ist der Verlauf der Anzahl an extrahierten Themen der Linken pro Monat dargestellt.

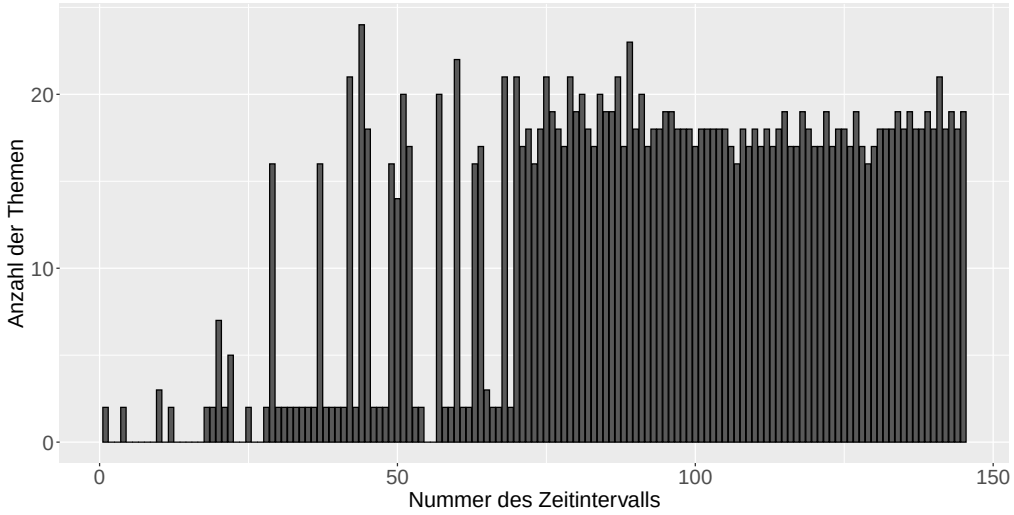

Abbildung A.33: Darstellung der Anzahl extrahierter Themen pro monatlichem Zeitabschnitt der Linken

Die Tagesschau

In Abbildung A.34 ist der Verlauf der Anzahl an extrahierten Themen der Tagesschau pro Monat dargestellt.

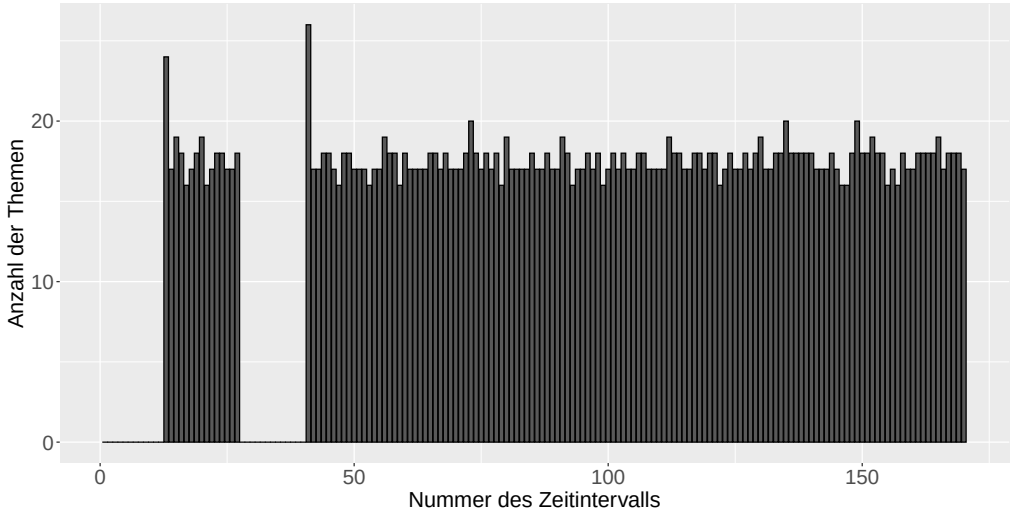

Abbildung A.34: Darstellung der Anzahl extrahierter Themen pro monatlichem Zeitabschnitt der Tagesschau

Literaturverzeichnis

[Abulaish, 2018] Abulaish, M. (2018). Modeling topic evolution in twitter: An embedding-based approach. *IEEE Access*, 6:64847–64857.

[Aggarwal, 2018] Aggarwal, C. C. (2018). Machine Learning for Text.

[Allan, 2009] Allan, J. (2009). *Search : Topic Detection and Tracking : Event-Based Information Organization*, volume 26. Kluwer Academic Publishers, Boston [u.a.].

[Anandarajan et al., 2019] Anandarajan, M., Hill, C., and Nolan, T. (2019). *Practical Text Analytics Maximizing the Value of Text Data*. Springer, Cham, 1 edition.

[Arun et al., 2010] Arun, R., Suresh, V., Veni Madhavan, C. E., and Narasimha Murthy, M. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In Zaki, M. J., Yu, J. X., Ravindran, B., and Pudi, V., editors, *Advances in Knowledge Discovery and Data Mining*, pages 391–402, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Asuncion et al., 2009] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009*, (Ml):27–34.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Topic Models. *Text Data Mining*, pages 145–162.

[Blei et al., 2002] Blei, D. M., Ng, A. Y., and Jordan, M. T. (2002). Latent dirichlet allocation. *Advances in Neural Information Processing Systems*, (July).

[Cai et al., 2014] Cai, G., Peng, L., and Wang, Y. (2014). Topic Detection and Evolution Analysis on Microblog. *Intelligent Information Processing*, VII:61–77.

[Cao et al., 2009] Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, 72:1775–1781.

[Deveaud et al., 2014] Deveaud, R., SanJuan, E., and Bellot, P. (2014). Accurate and effective Latent Concept Modeling for ad hoc information retrieval. *Document Numerique*, 17(1):61–84.

[Forster, 2017] Forster, O. (2017). *Analysis 2*.

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific

topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(SUPPL. 1):5228–5235.

[Hamming, 1950] Hamming, R. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal*, XX1X(April, 1950).

[Hofmann, 1999] Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 51(2):50–57.

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1-2):177–196.

[Huang, 2008] Huang, A. (2008). Similarity measures for text document clustering. *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*, (April):49–56.

[Jelodar et al., 2019] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., and Zhao, L. (2019). *Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey*, volume 78.

[Jurafsky and Martin, 2020] Jurafsky, D. and Martin, J. H. (2020). Speech and Language Processing: An introduction to natural language processing. *SPEECH and LANGUAGE PROCESSING An Introduction to Natural Language Processing Computational Linguistics and Speech Recognition*, pages 1–18.

[Kuri, 2017] Kuri, J. (2017). Twitter verdoppelt maximale Länge der Tweets auf 280 Zeichen, https://www.heise.de/newsticker/meldung/Twitter-verdoppelt-maximale-Laenge-der-Tweets-auf-280-Zeichen-3883047.html, Aufgerufen am 22.08.2021.

[Lau et al., 2012] Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models: Twitter trends detection topic model online. *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, 2(December):1519–1534.

[Leskovec et al., 2009] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 497–505.

[Levenshtein, 1966] Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10:845–848.

[Liu et al., 2009] Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link lda: Joint

models of topic and author community. *ACM International Conference Proceeding Series*, 382(January).

[Malik et al., 2013] Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., and Shneiderman, B. (2013). TopicFlow. pages 720–726.

[Porteous et al., 2008] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 569–577, New York, NY, USA. Association for Computing Machinery.

[Ramage et al., 2009] Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, pages 248–256.

[Sartorius, 2019] Sartorius, G. (2019). *Erfassen, Verarbeiten und Zuordnen multivariater Messgrößen*.

[Signorini et al., 2011] Signorini, A., Segre, A. M., and Polgreen, P. M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS ONE*, 6(5).

[Sopan et al., 2012] Sopan, A., Rey, P. J., Butler, B., and Shneiderman, B. (2012). Monitoring academic conferences: Real-time visualization and retrospective analysis of backchannel conversations. *Proceedings of the 2012 ASE International Conference on Social Informatics, SocialInformatics 2012*, (February 2016):62–69.

[Spranger et al., 2016] Spranger, M., Heinke, F., Appelt, L., Puder, M., and Labudde, D. (2016). MoNA: Automated Identification of Evidence in Forensic Short Messages. *International Journal On Advances in Security*, 9:14–24.

[Spranger et al., 2018] Spranger, M., Heinke, F., Siewerts, H., Hampl, J., and Labudde, D. (2018). Opinion Leaders in Star-Like Social Networks: A Simple Case?

[Spranger et al., 2017] Spranger, M., Siewerts, H., Hampl, J., Heinke, F., and Labudde, D. (2017). SoNA: A Knowledge-based Social Network Analysis Framework for Predictive Policing. *International Journal On Advances in Intelligent Systems*, 10.

[Twitter, 2021] Twitter (2021). Q4 and Fiscal Year 2020: Letter to Shareholders.

[Wang and McCallum, 2006] Wang, X. and McCallum, A. (2006). Topics over time. page 424.

[Yan et al., 2013] Yan, X., Guo, J., Lan, Y., and Cheng, X. (2013). A biterm topic model for short texts. In Schwabe, D., Almeida, V. A. F., Glaser, H., Baeza-Yates, R., and Moon, S. B., editors, *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1455. International World Wide Web Conferences Steering Committee / ACM.

[Zhai and Massung, 2016] Zhai, C. and Massung, S. (2016). Text data management and analysisa practical introduction to information retrieval and text mining.

[Zhou et al., 2017] Zhou, H., Yu, H., and Hu, R. (2017). Topic evolution based on the probabilistic topic model: a review. *Frontiers of Computer Science*, 11(5):786–802.

[Zhu et al., 2016] Zhu, M., Zhang, X., and Wang, H. (2016). A LDA Based Model for Topic Evolution: Evidence from Information Science Journals. 58(Msota):49–54.

[Zong et al., 2021] Zong, C., Xia, R., and Zhang, J. (2021). *Text Data Mining*. Springer, Singapore, 1 edition.

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 26. August 2021