Studie: Abbildung von Vielfalt und Regionalität in Suchmaschinen

Autor: Thomas Müller, Werner Bol, Nina Stromberger, Koen Loogman, Jens Duhme
Kontakt: thomas2.mueller@atos.net / jens.duhme@atos.net

Version: 1.1.1
Dokumentendatum: 21.01.2022
Dokumentennummer:

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Atos – der Partner für die Studie Abbildung von Vielfalt und Regionalität in Suchmaschinen

ATOS AUF EINEN BLICK

Die Landesanstalt für Kommunikation Baden-Württemberg (LFK) hat Atos in einem Ausschreibungsverfahren als Partner für die vorliegende Studie ausgewählt. An dieser Stelle möchten wir uns für das entgegengebrachte Vertrauen bedanken und einige Informationen zu unserem Unternehmen darlegen.
Atos ist ein weltweit führender Anbieter für die digitale Transformation mit 107.000 Mitarbeitern und einem Jahresumsatz von über 11 Milliarden Euro.
Als europäischer Marktführer für Cybersecurity sowie Cloud und High Performance Computing bietet die Atos Gruppe maßgeschneiderte, ganzheitliche Lösungen für sämtliche Branchen in mehr als 70 Ländern. Als Pionier im Bereich nachhaltiger Dienstleistungen und Produkte arbeitet Atos für seine Kunden an sicheren, dekarbonisierten Digitaltechnologien. Atos ist eine SE (Societas Europaea) und an der internationalen Börse Euronext Paris sowie in den Aktienindizes CAC 40 ESG und CAC Next 20 notiert. Unsere Firmenzentralen sind in Bezons (Frankreich, nahe Paris) und München.
Besonders erwähnenswert ist die jahrelange vertrauensvolle Zusammenarbeit als weltweiter IT-Partner der Olympischen und Paralympischen Spiele.
Das Ziel von Atos ist es, die Zukunft der Informationstechnologie mitzugestalten. Fachwissen und Services von Atos fördern Wissensentwicklung, Bildung sowie Forschung in einer multikulturellen Welt und tragen zu wissenschaftlicher und technologischer Exzellenz bei. Weltweit ermöglicht die Atos Gruppe ihren Kunden und Mitarbeitern sowie der Gesellschaft insgesamt, in einem sicheren Informationsraum nachhaltig zu leben, zu arbeiten und sich zu entwickeln.
Die folgende Abbildung zeigt unsere weltweite Aufstellung und lokale Standorte in Deutschland:

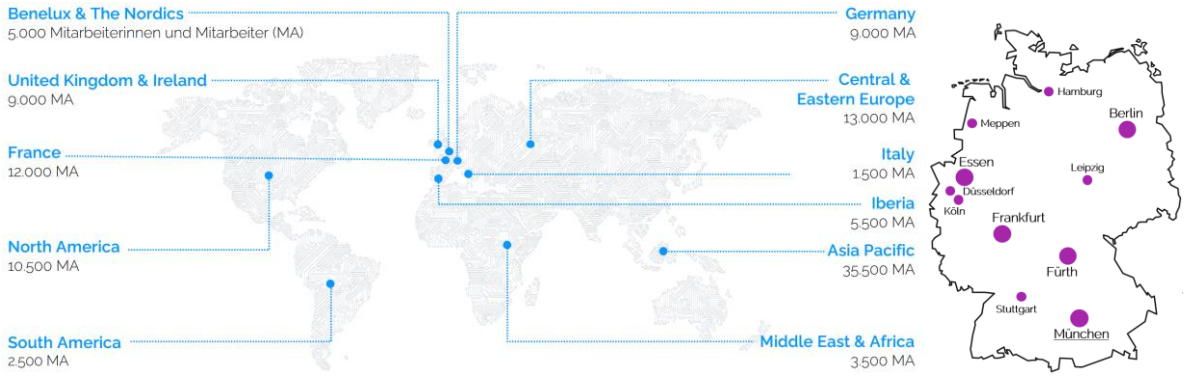


Abbildung 1 Atos weltweit und lokal

Um Kundenanforderungen optimal zu unterstützen hat Atos folgende Märkte definiert:

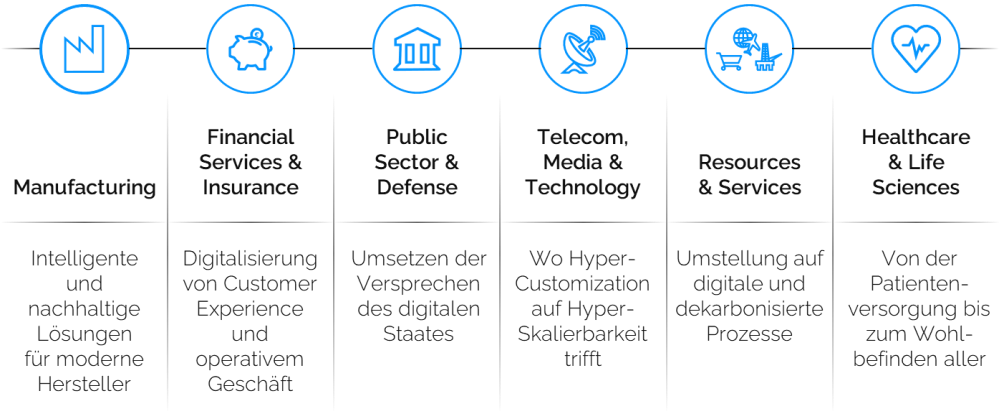


Abbildung 2 Atos – Märkte

DIE UNTERNEHMENSBEREICHE

Die Marktsegmente werden von Atos je nach Fachlichkeit aus unterschiedlichen Unternehmensbereichen heraus bedient. Dem Bereich Digital ist die Einheit ARAA zugehörig, deren Zuständigkeit sich insbesondere auf die Themenbereiche Automation, Robotics, AI (=Artificial Intelligence) & Analytics erstreckt.
Häufig umfassen die Aktivitäten dabei die Automatisierung von Prozessen, analytische Auswertungen u.a. zur Entscheidungsunterstützung. ARAA berät und implementiert/integriert maßgeschneiderte IT Systeme und setzt dabei eigene Lösungen oder Lösungen von Software-Partnern ein. Der Fokus liegt auf der Verarbeitung unstrukturierter Daten (nicht identifizierbare Datenstruktur, z.B. Dokumente, Mails und Web Inhalte) und strukturierter Daten (in normalisierter Form vorliegende Daten, z.B. Zahlen in Datenbanken).
Die folgende Darstellung visualisiert die Aufgabenbereiche des ARAA Teams:

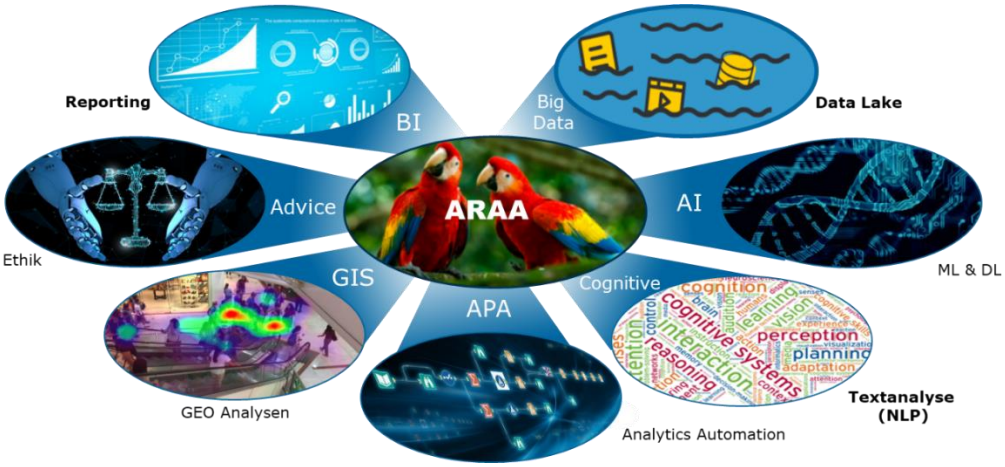


Abbildung 3 Atos - Aufgabenbereiche ARAA

Im Kontext der Studie sollen einige Teilgebiete hervorgehoben werden:

- Big Data: Zentrales Thema hierbei ist das Erfassen und Verarbeiten von großen Datenmengen – hier der Data Lake mit der großen Menge der Schlagzeilen.
- Reporting und Datenanalyse – hier wurde im Rahmen der Studie die Aufbereitung und Darstellung der Schlagzeilen vorgenommen.
- Textanalytics und NLP (Natural Language Processing) – in diesem Bereich lassen sich beispielsweise die WordClouds, Sentiment Analysen und die Identifikation von Schlagwörtern zuordnen.

Zusammenfassung

Die Landesanstalt für Kommunikation Baden-Württemberg (LFK) ist die Aufsichtsbehörde für den privaten Rundfunk und für Telemedien.

Anlass für diese von der LFK in Auftrag gegebene Studie war die Novellierung des Rundfunkstaatsvertrags der Länder zu einem Medienstaatsvertrag (MStV). Der MStV bezieht Suchmaschinen und soziale Netzwerke erstmals in die Vielfaltsregulierung ein und verpflichtet sie zu Transparenz und Diskriminierungsfreiheit.
Die LFK möchte daher mehr über die Wirkung der eingesetzten Auswahlmechanismen dieser Anbieter auf die Vielfalt der medialen Angebote, und insbesondere regionalen Inhalte, erfahren und Erkenntnisse sammeln, ob und wie Anomalien der Schlagzeilenergebnisse erkennbar sind.

Zur Durchführung der Studie wurden exemplarisch die Schlagzeilenfunktion von google.de genutzt und die Einträge im Schlagzeilenfenster nach Eingabe einer Reihe von unterschiedlichen Suchbegriffen über einen Zeitraum von 4 Wochen ausgewertet. Der Beobachtungszeitraum (Sonntag, 12.09.2021 00:00 bis Sonntag, 10.10.2021 23:59) stand in direktem Zusammenhang mit der Bundestagswahl 2021. Anhand einer Vielzahl von Fragestellungen wurde untersucht, wie sich die Auswahl von Suchbegriffen und die Wahl der Abfragestandorte und -geräte auf die zurückgelieferten Schlagzeilen auswirkt.

Im Rahmen der technischen Umsetzung wurde in einer Cloud Umgebung eine parametrisierbare Lösung aufgebaut, mit der zyklisch 1 Mal pro Stunde die von Google bereitgestellten Schlagzeilen für eine bestimmte Menge an Suchbegriffen abgefragt wurden. Mittels Einsatz eines Scraping Dienstes mit Proxy Funktionalität wurde eine neutrale, nicht personalisierte Durchführung der Abfragen sichergestellt. Die Simulation von unterschiedlichen Standorten (Stuttgart und Frankfurt) sowie Abfragegeräten (Desktop und Mobil) wurde durch geeignete Parametereinstellungen bewerkstelligt. Die zurückgelieferten Daten bestanden aus den von Google bereitgestellten HTML Seiten, aus denen im Rahmen eines Nachverarbeitungsprozesses die relevanten Informationen zu den Schlagzeilen (u.a. Schlagzeilentext, Datums- und Quellinformationen) extrahiert und in einer Datenbank abgespeichert wurden. Nach Bereinigung und Qualitätssicherung der verarbeiteten Daten folgte die statistische Auswertung. Auf Basis einer umfangreichen und im Rahmen der Studie gemeinsam mit der LFK erarbeiteten Frageliste wurde eine Auswertesystematik festgelegt, die auch zur Gliederung der Ergebnisdarstellung genutzt wurde.

Die wesentlichen Ergebnisse der statistischen Auswertung lassen sich wie folgt zusammenfassen:

- Der gewählte Abfragestandort (Stuttgart oder Frankfurt) hat keinen signifikanten Einfluss auf die Auswahl der auf die Suche zurückgelieferten Schlagzeilen. Ähnliches gilt für die simulierten Abfragegeräte, auch hier konnte kein signifikanter Einfluss festgestellt werden.
- Bzgl. des Ursprungs der zurückgelieferten Schlagzeilen fällt die starke Ungleichverteilung signifikant auf: vergleichsweise wenige Quellen sind für sehr viele Schlagzeilen verantwortlich, wohingegen die überwiegende Anzahl der Quellen nur wenige Meldungen liefert.
- Ein ähnliches Bild ergibt sich bei der Betrachtung der Schlagzeilen in Bezug auf die unterschiedlichen Mediengattungen.
- Lorenzkurven und Gini Koeffizienten sind geeignete Werkzeuge zur Beschreibung derartiger Ungleichverteilungen.
- Es ergaben sich eine schwach positiv ausgeprägte Tendenz hinsichtlich des Stimmungsgehalts der Schlagzeilen. Eine negative Grundhaltung z.B. in Bezug auf bestimmte Suchbegriffe war nicht zu beobachten.
- Die Erwartung, dass Schlagzeilen unter den ersten 10 Positionen deutlich länger als unter den ersten 3 gehalten werden, wurde durch die Untersuchungen bestätigt.
- Häufig auftretende Schlagwörter können sehr übersichtlich in WordClouds visualisiert werden.
- Auch bei den Schlagzeilen von Quellen aus Baden-Württemberg spielt der Abfrageort kaum eine Rolle. Im Durchschnitt gesehen finden sich diese Schlagzeilen auf mittleren Positionen im Schlagzeilenfenster, eine Bevorzugung der schlagzeilenträchtigeren Baden-Württemberg Quellen gegenüber den Quellen mit nur wenigen Schlagzeilen fällt kaum ins Gewicht. Die

dominierenden Orte in den Schlagzeilen der Baden-Württemberg Quellen sind (erwartungsgemäß) Stuttgart und Karlsruhe.

1 Projekthintergrund und -durchführung

1.1 Zielsetzung des Projekts

Medienintermediäre wie z.B. Google oder Facebook erstellen zwar keine eigenen Nachrichteninhalte, können aber durch ihre große Reichweite die öffentliche Meinungsbildung in erheblichem Umfang bspw. durch eine zielgerichtete Bereitstellung oder Auswahl von Informationen, beeinflussen. Als Wirtschaftsunternehmen ist zu vermuten, dass sich die Auswahl unter Umständen stark an den wirtschaftlichen Interessen der Intermediären orientiert. Charakterisieren lässt sich diese Problematik treffend durch die Hypothese der sogenannten Filterblasen, bei denen sich die durch Algorithmen ausgewählten neuen Nachrichteninhalte stark an den bisherigen Interessen und Suchbegriffen der Nutzer anlehnen oder das Konzept der Echokammern, das die Verstärkung von Meinungen durch die einseitige Präsentation von Nachrichten und Meldungen beschreibt. Angesichts von mehr als 3 Milliarden Facebook Nutzern weltweit und über 200 Millionen Google Suchanfragen pro Stunde wird deutlich, welches Einflusspotenzial in diesem Umfeld theoretisch besteht.

Vor diesem Hintergrund wurde im Rahmen des Studienprojekts „Abbildung von Vielfalt und Regionalität in Suchmaschinen" untersucht, inwieweit Google bei der Bereitstellung Informationen in Form von Schlagzeilen, mittels der Google Suchfunktion, einen neutralen Standpunkt einnimmt, bzw. Nachrichten möglicherweise nach bestimmten Gesichtspunkten verstärkt ausgewählt und im besonderen Maße in das Blickfeld des Nutzers gerückt werden.

1.2 Signifikante Ergebnisse

An dieser Stelle sollen einige signifikante Ergebnisse der Studie zusammengefasst werden. Hierbei wird unterschieden zwischen den einzelnen Untersuchungsbereichen, die im Detail in den Abschnitten 2.1 - 2.7 beschrieben werden und übergreifenden Erkenntnissen:

- Übergreifende Ergebnisse
  - Der gewählte Abfragestandort (Standort Stuttgart oder Frankfurt) hat keinen signifikanten Einfluss auf die Auswahl der auf die Suche zurückgelieferten Schlagzeilen
  - Ähnliches gilt für die simulierten Abfragegeräte (Mobile oder Desktop), auch diese haben (insbesondere in der Summe über die Quellen) keinen signifikanten Einfluss auf die Auswahl der auf die Suche zurückgelieferten Schlagzeilen.
- Quellen
  - Die Verteilung der Quellen bzgl. der auf die Anfragen zurückgelieferten Schlagzeilen folgt einer typische „Long Tail" Charakteristik: wenige der insgesamt über 500 erfassten Quellen liefern einen Großteil der Schlagzeilen, wohingegen der weit überwiegende Anteil der Quellen nur wenige Schlagzeilen im Google Schlagzeilenfenster positionieren kann.
  - Eine Betrachtung der ersten 3 oder der ersten 10 Einträge im Google Schlagzeilenfenster ändert nur wenig an der Reihenfolge der Top Schlagzeilenquellen.
  - Bei der Berechnung der Sentimente der Schlagzeilen für unterschiedliche Quellen ergaben sich jeweils leicht positive Werte ohne große Unterschiede zwischen den Quellen untereinander. Dies spricht für eine ausgewogene Berichterstattung mit einer wohlwollenden Grundtendenz.

- Mediengattungen
  - Bei den Mediengattungen ist eine deutliche Ungleichverteilung bzgl. der Anzahl der pro Gattung zurückgelieferten Schlagzeilen zu beobachten, wobei die Gattung Zeitung die anderen Bereiche klar dominiert.

- Suchbegriffe

- Bei den Suchbegriffen fiel eine Ausgewogenheit bzgl. der Anzahl der zurückgelieferten Schlagzeilen insbesondere für die 3 Kanzlerkandidaten auf. Eine Bevorzugung eines Kandidaten zu Lasten der anderen Kandidaten war nicht erkennbar.
- Auch die Sentimente der den einzelnen Suchbegriffen zugeordneten Schlagzeilen zeigten keine Auffälligkeiten in Bezug auf eine mögliche Besser- oder Schlechterstellung. Zudem konnte keine zeitliche Abhängigkeit im Verlauf der Sentimente festgestellt werden.
- Im Hinblick auf die von ausgewählten Quellen für die unterschiedlichen Suchbegriffe zurückgelieferten Schlagzeilen zeigten sich deutliche Unterschiede zwischen thematisch breit aufgestellten Quellen (z.B. Der Spiegel, Tagesschau, Die Welt) und sehr fokussiert (etwa regional) berichtenden Quellen (z.B. ka-news). So kann die Annahme getroffen werden, dass Suchbegriffe mit regional Bezug auch deutlich eher von eher regional aufgestellten Quellen mit entsprechenden Nachrichtentreffern beantwortet werden, während thematisch breit aufgestellten Quellen naturgemäß weniger hohe Ränge bei regionalen Themen belegen.

- Schlagzeilen
  - Bzgl. des mittleren Alters der Schlagzeilen konnten signifikante Unterschiede zwischen den Bundestagswahl-bezogenen Meldungen (durchschnittliches Alter vor der Wahl: ~15 Stunden) im Vergleich mit den Schlagzeilen aller Suchbegriffe (Durchschnittsalter vor der Wahl ~25 Stunden) beobachtet werden. In beiden Fällen steigt die Aktualität kurz nach der Wahl für kurze Zeit an.
  - Hinsichtlich der mittleren Verweildauer in unter den ersten 10 („Top10") bzw. ersten 3 Schlagzeilen („Top3") im Schlagzeilenfenster ergeben sich erwartungsgemäß deutlich längere Präsenzzeiten in der Gruppe Top10, die z.B. bei der Tagesschau mit knapp 80 Stunden fast doppelt so lange im Vergleich zum Top3 Schlagzeilenfenster derselben Quelle sind. Allerdings ist unter den 10 Quellen mit den meisten Schlagzeilen kein direkter Zusammenhang zwischen der Anzahl der Schlagzeilen und der Verweildauer in den Top10 bzw. Top3 Schlagzeilenfenstern zu erkennen.
  - Bei der Untersuchung des zeitlichen Verlaufs des Rangs einer Schlagzeile wurde der Fokus auf besonders langlebige Schlagzeilen gelegt und bei einigen ausgewählten Beispielen näher betrachtet. Häufig starteten die Schlagzeilen bei einem niedrigen Rang zu Beginn des Beobachtungszeitraums, der sich mit der Zeit dann verschlechterte. Jedoch wurden auch Fälle beobachtet, in denen sich der Rang nach einer gewissen Zeit (unerwartet) wieder verringerte, mit der Folge, dass die Schlagzeile mit einer höheren Wahrscheinlichkeit vom Nutzer wahrgenommen wird.
  - Zur Bewertung möglicher textueller Ähnlichkeiten zwischen den Schlagzeilen wurde aus Gründen der Handhabbarkeit (Vermeidung eines zu hohen Rechenaufwands) auf einen Entitäts-basierten Vergleich zwischen den Schlagzeilen zurückgegriffen. Anhand einiger Beispiele wurden für eine hohe Anzahl an übereinstimmenden Entitäten ähnliche Schlagzeilen identifiziert. Dies funktionierte jedoch nur bei einem hohen Grad an Übereinstimmung der Entitäten und zeigte dabei deutlich die begrenzte Nutzbarkeit des angewendeten Verfahrens auf.

- Schlagwörter
  - Zur Analyse der in den Schlagzeilen auftretenden Schlagwörter wurden die Entitäten in den Schlagzeilen bestimmt und in WordClouds entsprechend der Häufigkeit des Auftretens der jeweiligen Entitäten dargestellt. Neben den Suchbegriffen selbst und zugehörigen Ableitungen (z.B. Kanzlerkandidaten mit Vor- und Nachnamen oder nur Nennung des Nachnamens, Verwendung Ein-/Mehrzahl, …) konnten gut erklärbare Zusammenhänge aufgezeigt werden. Die Darstellung in Form von WordClouds erwies sich als ausgesprochen nützlich um einen kompakten Überblick über den textuellen Inhalt einer großen Menge an Schlagzeilen zu erhalten.
  - Es konnte beobachtet werden, dass die häufigsten Schlagwörter (die meist auch dem Suchbegriff oder den direkten Ableitungen entsprachen) in der Regel durchgängig im Beobachtungszeitraum präsent waren.

- Lokalisierungen Baden-Württemberg
  - Setzt man den Anteil der Quellen mit Baden-Württemberg Bezug ins Verhältnis zur Gesamtanzahl an Quellen, die über dem Beobachtungszeitraum Schlagzeilen geliefert haben, so liegt dieser Anteil bei knapp 23%. Dabei ist zu berücksichtigen, dass 11 der 17 ausgewählten Suchbegriffe einen Bezug zu Baden-Württemberg aufweisen und sich dies auf den Anteil der Baden-Württemberg Quellen auswirkt. So liefern einige der Baden-Württemberg Quellen sehr hohe Anzahlen an Schlagzeilen, jedoch überwiegend zu Baden-Württemberg Suchbegriffen (siehe Abschnitt 2.3.4).
  - Der durchschnittliche Rang (d.h. die Position im Schlagzeilenfenster) von Quellen aus Baden-Württemberg liegt bei etwa 6. Dabei gibt es kaum systematische Unterschiede zwischen Baden-Württemberg Quellen mit einer hohen bzw. eher niedrigen Anzahl an zurückgelieferten Schlagzeilen. Quellen mit einer hohen Anzahl an Schlagzeilen werden hinsichtlich der Positionierung im Schlagzeilenfenster also kaum bessergestellt als die „kleineren" Quellen.
  - Die Baden-Württemberg Quellen mit den meisten Schlagzeilen unter den ersten 3 Platzierungen im Schlagzeilenfenster sind die ka-news, Stuttgarter Nachrichten und Stuttgarter Zeitung, die bezüglich der Anzahl der Schlagzeilen auf einem vergleichbaren Niveau liegen. Allerdings ist die Positionierung der KA-News stark durch den Suchterm „Karlsruhe" bedingt.
  - Wie bei der Betrachtung der Schlagzeiten aller Quellen spielt die Wahl des Abfrageortes (Stuttgart oder Frankfurt) auch bei den Baden-Württemberg Quellen keine signifikante Rolle.
  - In Baden-Württemberg Quellen tauchen oft die Namen von Orten aus diesem Bundesland auf. Dominant dabei ist das Auftreten der Städte Stuttgart und Karlsruhe.
  - Bei der Untersuchung der am häufigsten auftretenden Schlagworten in Schlagzeilen von Baden-Württemberg Quellen dominierten die Begriffe „Baden-Württemberg", „Stuttgart", „Karlsruhe" und „Daimler". Wird die Schlagwortverteilung für alle Quellen zugrunde gelegt, so treten die 3 zuerst genannten Begriffe etwas weniger in Erscheinung. Einzig das Schlagwort „Daimler" erscheint in der überregionalen Betrachtung einen höheren relativen Anteil zu besitzen.
  - Generell lässt sich aus den Ergebnissen schlussfolgern, dass Quellen mit Baden-Württemberg-Bezug auch vorrangig über Themen mit regionalem-Bezug berichten.

- Ungleichverteilung
  - Ein signifikantes Ergebnis bei der Untersuchung der Fragestellungen zu den Quellen ist die deutliche Ungleichverteilung in Bezug auf die Anzahl der zurückgelieferten Schlagzeilen („wenige Quellen mit vielen Schlagzeilen, viele Quellen mit wenigen Schlagzeilen"). Bei der Berechnung der die Ungleichverteilung beschreibenden Lorenzkurven und Gini Koeffizienten ist zu beachten, dass die resultierenden Teilmengen der Schlagzeilen durch die Wahl z.B. der Suchbegriffe (alle <> einzelne oder Gruppen von Suchbegriffen), des betrachteten Zeitraums (Tage, Woche, Gesamtzeitraum) usw. nicht zu klein werden, da andernfalls die resultierenden Gini Koeffizienten stark schwanken. Als vorteilhaft hat sich die Betrachtung von allen Suchbegriffen, oder bei Auswertungen für einzelne Suchbegriffe, die Wahl von zumindest Wochenzeiträumen erwiesen.

1.3 Ausblick

Bei der intensiven Beschäftigung mit den Fragestelllungen der Studie „Vielfalt und Regionalität von Suchmaschinen" wurden auch Ideen hinsichtlich weitergehender Fragestellungen diskutiert. Als besonders lohnenswert erscheinen dabei die nachfolgend genannten Themenbereiche.

Typ	Beschreibung	Nutzen	Potential
weitergehende Auswertung	Zusammenfassung von Schlagzeilen eines Suchbegriffs und gemeinsame Berechnung der Sentimente	Verlässlichere Sentimentberechnung	mittel
	Erweiterung der Sentimentbetrachtung (Nutzung von sowohl Score als auch Magnitude Werten)	Verlässlichere Sentimentberechnung	mittel
	Weitergehende Untersuchung der zeitlichen Entwicklung des Rangs von Schlagzeilen („erneute Verringerung des Ranges nach einer gewissen Zeit"). Zum Beispiel: (Gesamter Datenumfang) wie oft folgten die Schlagzeilen dem "klassischen" / erwartbaren Verlauf, wie oft kam es zu unerwarteten Verlauf.	Detailklärung eines unerwarteten Effekts	hoch
	Einbeziehung der zu den Schlagzeilen gehörenden vollständigen Nachrichtentexte der Quelle in die Textanalyse	Erhebliche Aufweitung des Fokus von den sehr kurzen Schlag-zeilentexten, Nachrichtentexte haben vermutlich ebenfalls einen Einfluss auf die Schlagzeilen-auswahl der Such-maschine Verbesserung der Sentiment- und Schlagwortbestimmung	sehr hoch
	Nutzung von NLP Verfahren für die Bestimmung der textuellen Ähnlichkeit	Verbesserte Ermittlung von semantisch ähnlichen Schlagzeilen	hoch
	Kap 2.5.2: Verweildauer von Schlagwörtern in Abhängigkeit vom Suchbegriff → Ausblendung von zu definierenden Begriffen (erwartbare Ergebnisse, z.B. zu den Suchbegriffen selbst)	Erhöhung der Aussagekraft	mittel
weitergehende Fragestellungen	Nach welchen Kriterien kommen Meldungen in das Google Schlagzeilenfenster (Einfluss des Schlagzeilen- und Meldungstextes)?	Detaillierte Untersuchung der Auswahlkriterien der Suchmaschine	
	Betrachtung weiterer Suchmaschinen, Social Media Platt-formen (Facebook, Twitter, Instagram, …), aktuell erscheint insbesondere eine Betrachtung der Social Media Plattformen interessant.	Ergänzung zu den auf Google fokussierten Untersuchungen	

	Erweiterung/Veränderung der untersuchten Suchbegriffe	Ergänzung des derzeit starken BaWü Fokus	
	Suchbegriff-neutrale Suche, z.B. auf news.google.de	Welche Suchbegriffe sind überhaupt aktuell?	
	Ergänzende Messung von Ungleich-verteilungen mittels Entropie als Erweiterung zu Lorenzkurve und Gini-Koeffizient	Erhöhung der Aussagekraft	
übergreifende Ansätze	Automatisierung der Auswertung und fortlaufende Aufzeichnung und Analyse	Aufhebung Fokussierung auf 4 Wochenzeitraum	sehr hoch
	Aufbau eines Daten Pools	Ermöglichung nachgelagerter Analysen beliebiger Art, Aufbau einer umfassenden Datenbasis auch in Hinblick auf Vergleichbarkeit über die Zeit (etwa Anzahl der Quellen, Ungleichverteilung)	sehr hoch
	Implementierung eines Dashboards	zeitaktuelle Darstellung von vergangenen und aktuellen Nachrichtenlage	sehr hoch

1.4 Rahmenparameter der Studie

1.4.1 Numerischen Randbedingungen und Kennzahlen

In Folgenden eine Übersicht der Randbedingen und Parameter der Studie

ZEITRAUM DER UNTERSUCHUNG

Der Analysezeitraum der Studie war von Sonntag, 12.09.2021 00:00 bis Sonntag, 10.10.2021 23:59 (29 Tage).

TERMIN DER WAHL

Der Termin der Wahl war Sonntag der 26.09.2021.

ANZAHL DER SUCHBEGRIFFE

Es wurden 17 Suchbegriffe abgefragt.

ZAHL DER ABFRAGEN PRO ZEITRAUM (Z.B. TAG ODER STUNDE)

Es wurde täglich zu jeder vollen Stunde je 4 Abfragen pro Suchbegriff gestellt, mit den Parametern: Ort der Anfrage (Frankfurt und Stuttgart) und Anfragegerät (Mobile und Desktop).

ANZAHL DER ANFRAGEN

29 (Tage) x 24 (Stunden) x 17 (Suchbegriffe) x (4 Abfrageparameter) ergibt eine Summe von 47.328 Abfragen.

ZAHL DER SCHLAGZEILEN

Es wurden 340.468 Schlagzeilen (nach Aufbereitung der Daten) erfasst.
Hinweis: Nicht jede Abfrage hat Daten geliefert und nicht immer alle 10 möglichen Ränge.

ZAHL DER QUELLEN IN TOP10 UND TOP3

Es wurden 545 unterschiedliche Quellen in den Top 10 und 400 unterschiedliche Quellen in den Top3 festgestellt.

1.4.2 Die Schlagzeilenfunktion von Google

Suchmaschinen, insbesondere Google als unangefochtener Marktführer, spielen eine zentrale Rolle als „Zugangspunkt" zu journalistisch-redaktionellen Inhalten im Internet. Mit der Einführung des „Schlagzeilenelements" umfasst die Google Suche ein Element, das ausschließlich auf journalistisch-redaktionell gestaltete Angebote fokussiert und optisch sowie inhaltlich abgesetzte Ergebnisse liefert (Abbildung 1).

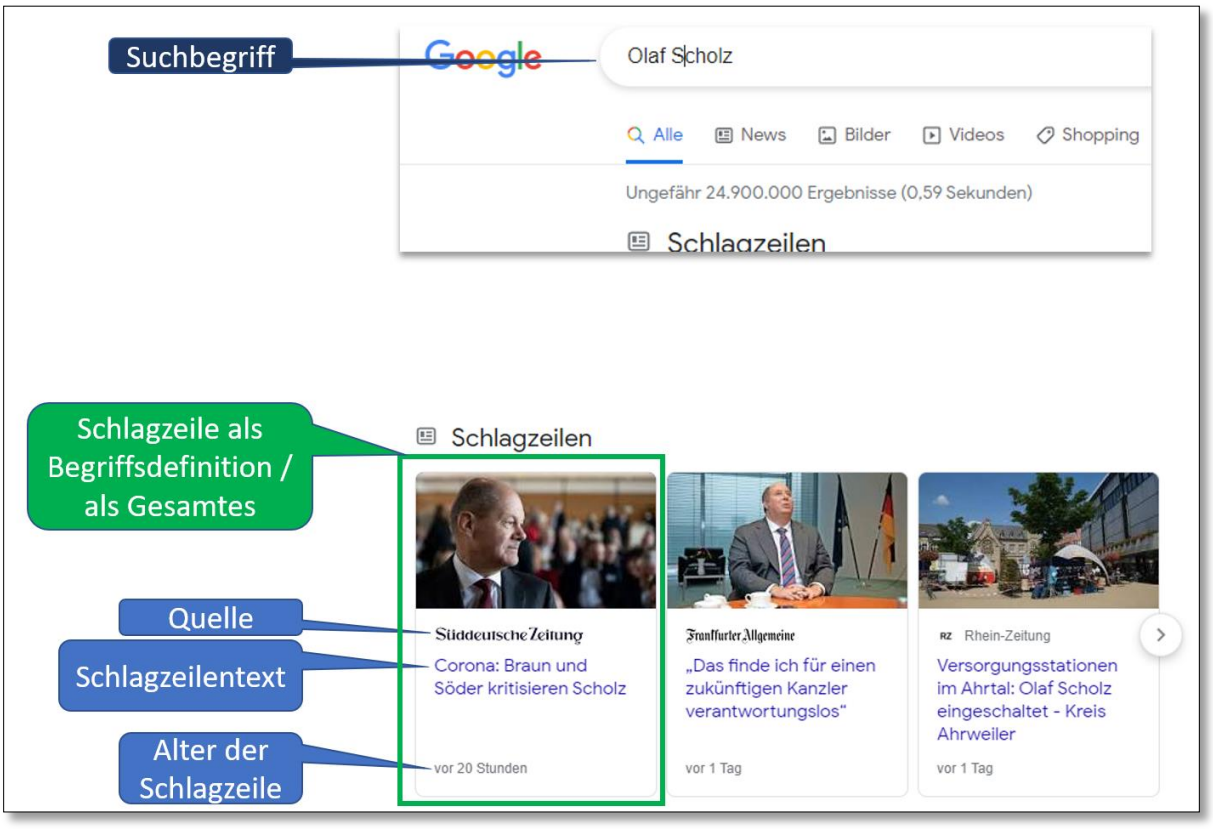


Abbildung 4: Google Suche und Anzeige von Treffern in der Schlagzeilen Box

Das Schlagzeilenelement kanalisiert die Aufmerksamkeit auf einzelne journalistische Angebote und beeinflusst maßgeblich deren Auffindbarkeit. Durch die Beschränkung der Ergebnisse auf zunächst

maximal drei sichtbare Treffer besteht ein Diskriminierungspotential einzelner Angebote und/oder Anbieter. Weitere Treffer lassen sich erst durch Betätigung des „Weiter"-Buttons (rechts in Abbildung 4) anzeigen. Darüber hinaus haben die automatisierten Selektions- und Rankingprozesse Auswirkungen auf die Vielfalt innerhalb der im Schlagzeilenfenster dargestellten Nachrichteninhalte.

Google präsentiert im Rahmen seiner Suchergebnisse bei einer Vielzahl von Suchtermen sogenannte „Schlagzeilen", journalistisch-redaktionelle Einzelbeiträge, die auf den eingegebenen Suchbegriff passen. Diese Schlagzeilen werden dabei nicht bei jeder Suchanfrage und auch nicht immer in gleichem Umfang dargestellt. Zum Ursprung, der Zusammensetzung und Auswahl der dargestellten Quellen und Nachrichteninhalte veröffentlicht Google indes keine Informationen.

1.4.3 Begriffsdefinitionen

Im Rahmen der Studie sind eine Reihe von Begriffen von zentraler Bedeutung, die der Klarstellung halber im Folgenden kurz erläutert werden. Dabei sei auch auf die Abbildung 4 verwiesen.

- **Schlagzeile:** Überbegriff für die Schlagezeile als Gesamtheit (Quelle, Schlagzeilentext, Alter, URL…)
- **Suchbegriff**: Der Suchbegriff (Beispiel: Klimawandel) beschreibt den Term, der nach Aufruf der Google Suchfunktion unter google.de in das Suchfenster eingegeben wird und für den die zurückgelieferten Schlagzeilenelemente im Schlagzeilenfenster analysiert werden sollen. Im Kontext der Studie wurden insgesamt 17 Suchbegriffe (siehe Kapitel 1.4.5 Suchbegriffe) festgelegt, deren resultierende Schlagzeilenelemente erfasst und im Folgenden ausgewertet wurden.
- **Quelle:** Die Quelle (Beispiel: *„ZDF"*) ist der Ersteller bzw. der Lieferant des jeweiligen Schlagzeilenelements. Als Synonym wurden in der Studie auch die Begriffe Editor, oder Herausgeber bzw. Hersteller verwendet. Hat ein Ersteller mehrere URLs, z. *zdfheute-stories-scroll.zdf.de* und *www.zdf.de*, so werden die Schlagzeilen beider URLs (entsprechend der Liste Zuordnung_Gattung_LFK_21-10-27.xlsx) der Quelle *„ZDF"* zugeordnet.
- **Rang:** (engl.: rank) Im Google-Schlagzeilenfenster werden mehrere Ergebnisse nebeneinander angezeigt (zum Zeitpunkt der Durchführung der Studie maximal drei auf einmal), wobei durch Betätigen des „Weiter" Buttons ein Blättern in der Gesamttrefferliste möglich ist. Der erste Eintrag erhält den Rang 1, der rechts darauf Folgende den Rang 2 usw.. Zum Zeitpunkt der Durchführung der Studie wurden bei Nutzung des Desktop Abfrage Modus (siehe Kapitel 1.5.1 Erfassung der Schlagzeilenelemente) maximal zehn Schlagzeilen im Schlagzeilenfenster angezeigt. Ein Schlagzeilenelement mit einem Rang zwischen 1 und 3 wird entsprechend unmittelbar angezeigt und ist damit auf den ersten Blick für den Nutzer sichtbar. Für den Zugriff auf Schlagzeilenelemente ab Rang 4 muss der „Weiter"-Button ein- bzw. mehrmals betätigt werden, so dass davon auszugehen ist, dass diese Schlagzeilen von vielen Nutzern mit einer geringeren Wahrscheinlichkeit wahrgenommen werden.
  Vereinbarungsgemäß wurden nur die Ränge 1-10 berücksichtigt (Ränge > 10 können i.d.R. nur im Abfragemodus Mobile auftreten).
- **Schlagzeilen:** Hierunter sind die Titel der jeweiligen Schlagzeilenelemente zu verstehen, die Google zu einem Suchbegriff liefert. Es handelt sich hierbei um kurze vollständige oder unvollständige Sätze, die den Inhalt der eigentlichen Nachricht charakterisieren. Im Rahmen der Studie wurden entsprechend der Beauftragung die Schlagzeilen ausgewertet, eine Betrachtung der Inhalte der Nachrichtentexte ist nicht erfolgt.
- **Mediengattungen:** Hierunter ist die Zuordnung einer Quelle (etwa Stuttgarter Nachrichten) zu einer Gattung (Zeitung) und Untergattung (Zeitung lokal/regional) zu sehen. Folgende Gattungen und Untergattungen wurden im Rahmen der Studie definiert:

Gattung	Untergattung
Corporate Publishing	Corporate Publishing
	Corporate Publishing NGO
	Parteien
Online	Online
	Online Int.
	Online lok./reg.
	Öffentl.-rechtl. Online
Red./Agentur	Red./Agentur
regÖA	regÖA

Gattung	Untergattung
Rundfunk	TV int.
	TV int. publ.
	TV privat
	TV öffentl.-rechtl.
	Öffentl.-rechtl.
	Radio int.
	Radio int. publ.
	Radio privat
	Radio öffentl-rechtl.
Zeitschrift	Zeitschrift
	Zeitschrift int.
	Zeitschrift lok./reg.
Zeitung	Zeitung int.
	Zeitung überreg.
	Zeitung lok./reg.
	Wochenzeitung

Tabelle 1 Mediengattungen

Hinweise: Regierungsamtliche Öffentlichkeitsarbeit (regÖA) stellt in Deutschland eine Sonderform der Öffentlichkeitsarbeit oder auch Unternehmenskommunikation dar. Als regÖA werden alle Informationsmaßnahmen amtlicher Stellen, wie zum Beispiel Behörden, Ministerien, aber auch staatlich finanzierte Einrichtungen bezeichnet.
Generell gilt, dass alle in den Schlagzeilen genannten Quellen Internetverweise, also Onlinequellen sind. Die Differenzierung der Gattung erfolgt nach dem erkennbaren Hintergrund der Quellen. Ist bspw. ein Zeitschriftentitel in der URL genannt oder explizit in der Marke enthalten, so wird dies der Gattung Zeitschriften zugeschlagen (bspw. *Spiegel* und *Spiegel-online* werden der Gattung Zeitschrift zugeordnet). Die Online-Angebot der *Tagesschau* wird mit explizitem Bezug auf die Sendung bspw. der Gattung TV öffentl.-rechtlich zugeordnet. Ist eine Zuordnung zu einem Printtitel, einer konkreten Sendung oder einem Programm nicht möglich, wird dies der Gattung Online zugeordnet.

- Die Zuordnungen der Quellen zu den Gattungen und Untergattungen erfolgte über eine vom Auftraggeber zur Verfügung gestellte Medienliste (Siehe Kapitel 1.4.6 Medienlisten).
- **BW-Bezug:** Einige Quellen haben gemäß einer durch die LFK bereitgestellten Liste (Medienliste Bezug BW_Bezug, siehe 1.4.6 Medienlisten) einen Bezug zum Land Baden-Württemberg.
- **Standort (Gesamt):** In Verbindung mit der Lokation besteht ein Standort „Gesamt" stets aus allen Daten zum Standort „Stuttgart" vereint mit allen Daten zum Standort „Frankfurt".
- **Alter der Schlagzeile:** Das Alter beschreibt die Aktualität der von Google gelieferte Schlagzeilenelemente zum Zeitpunkt der Abfrage und wird von Google für jede Schlagzeile in der Form „vor x Minuten/Stunden/Tagen/Wochen/Monaten" angegeben (siehe Abbildung 4 → Alter der Schlagzeile). Diese Angabe entspricht der Differenz zwischen dem Zeitpunkt des Auslesens und dem ebenfalls ausgelesenem Veröffentlichungsdatum der Schlagzeile.
- **Veröffentlichungsdatum:** Diese Information wird durch die Schlagzeilenfunktion von Google bereitgestellt. Es ist dabei zu beachten, dass das bei einer Abfrage übermittelte Veröffentlichungsdatum einer Schlagzeile nicht zwangsweise das *Erst*veröffentlichungsdatum sein muss.

Beispielsweise durch Aktualisierungen der Schlagzeile kann sich ein neues Veröffentlichungsdatum ergeben.

- **Top:** Bei einigen Darstellungen wird der Ausdruck „Top" verwendet, z.B. im Zusammenhang mit Schlagzeilen oder Quellen. In Verbindung mit Schlagzeilen bezieht sich „Top" auf den Rang einer Schlagzeile. Beispielsweise ist der Ausdruck „Top3" so zu interpretieren, dass die Schlagzeilen der Ränge 1 bis 3 bewertet werden. Hingegen bedeutet „Top" in Verbindung mit einer Quelle, dass die stärksten Quellen, gemessen an ihrer Anzahl an Schlagzeilen, bewertet werden.
- **Lorenzkurve:** Die Lorenzkurve ist eine Möglichkeit, die Gleich- oder Ungleichverteilung einer Messgröße zu beschreiben. Sie zeigt mit einer Kurve unterhalb der Diagonalen (vom Nullpunkt bis zum Maximalwert), wie viele Einheiten (z.B. die Zahl der Quellen im Verlauf auf der X-Achse) wie viel Wert (Anzahl an Schlagzeilen der jeweiligen Quelle im Verlauf auf der Y-Achse) besitzen. Dabei markiert die Winkelhalbierende die Kurve der absoluten Gleichverteilung. Ein flacher Verlauf einer Lorenzkurve zeigt, dass viele Einheiten (Quellen) nur sehr wenig Wert (Anzahl Schlagzeilen) besitzen. Ein abschließender sehr steiler Anstieg sagt aus, dass sehr wenige Einheiten sehr viel Wert besitzen. In dem Fall ist die Ungleichverteilung sehr hoch.
- **Gini-Koeffizient:** Er drückt den Grad einer Ungleichverteilung (siehe Lorenzkurve) mit einer einzigen Zahl zwischen 0 (absolute Gleichverteilung) und 1 (absolute Ungleichverteilung) aus. Neben dem Koeffizienten wird auch immer der Wert N angegeben, welcher die Anzahl der betrachteten Daten anzeigt (identisch mit der Anzahl N auf der X-Achse der Lorenzkurve).
- **Schlagwort/ Entität:** Im Allgemeinen bestehen Texte aus einer Vielzahl von Wörtern. Einige dieser Wörter, die keine Füllwörter sind, können als Schlagwörter definiert werden, die den Text oder Teile davon charakterisieren. Im Rahmen einer Entitätsanalyse können Schlagwörter als Bestandteil von sogenannten Entitäten ermittelt werden. Eine Entität beinhaltet mehrere Eigenschaften, unter anderem den Entitätswert (hier: das Schlagwort selbst in Textform) sowie einen Entitätstypen (z.B. „Zahl", „Person", „Ort", „Datum", „Organisation" usw.). Beispielsweise wird das Schlagwort „Olaf Scholz" dem Entitätstypen „Person" oder das Schlagwort „Stuttgart" dem Typen „Ort" zugeordnet.
- **Entitätsanalyse:** Dies ist eine Disziplin des „Natural Language Processing" (NLP), bei der in vorgegebenen Texten Entitäten identifiziert werden. Eine Entität besteht unter anderem aus einem textuellen Schlagwort sowie einem Entitätstypen (siehe auch „Schlagwort/Entität"). Im Rahmen dieser Ausarbeitung wurden die Titel aller Schlagzeilen einer separaten Entitätsanalyse unterzogen und somit die Schlagwörter bzw. Entitäten jeder Schlagzeile identifiziert. Für die Identifikation von Entitäten in Text gibt es verschiedene statistische und auf KI (Künstlicher Intelligenz) basierende Verfahren, die in Form von z.B. Python Bibliotheken oder APIs einfach genutzt werden können. Für diese Studie wurde die Entitätsanalysefunktionen aus der Google Cloud NLP API angewendet.
- **Sentimentanalyse:** Hierbei handelt es sich ebenfalls um eine Disziplin aus dem Bereich NLP. Vorgegebene Texte, im Fall der Studie erneut die Titel der Schlagzeilen, werden daraufhin analysiert, ob sie einen überwiegend positiven, negativen oder eher neutralen (Wert ~0) Stimmungswert ausdrücken. Eine Schlagzeile über wirtschaftliche Probleme oder einem Vorfall mit Verletzen wäre eher negativer Natur. Erfolgreiche Ereignisse dagegen eher positiver Natur. Die Berechnung von Sentimenten für sehr kurze Texte (hier die Titel der Schlagzeilen) ist schwierig, da nur wenige Worte für die Bewertung zur Verfügung stehen. Zudem besteht die Herausforderung, bestimmte sprachliche Stilmittel wie z.B. Sarkasmus zu erkennen und dem richtigen Sentimentwert zuzuordnen. Darüber hinaus kommen falsch-positive Bewertungen vor, wie z.B. Berichte über „positive Corona-Tests bei Spielern des VfB Stuttgart". Bei der Bestimmung von Sentimenten werden üblicherweise zwei Werte ermittelt: Ein „score" und ein „magnitude" Wert. Ersteres drückt das Gesamtbild des Textes aus (positiv, negativ, neutral). Letzteres gibt an, wie stark bzw. wie oft Emotionen in dem Text vertreten sind. Aufgrund der Kürze der Texte werden im Rahmen dieser Analysen nur die score-Werte betrachtet. Ein score kann einen beliebigen Wert zwischen -1 und 1 annehmen. Der Wert -1 bedeutet, dass der Text sehr stark negativ geprägt ist. Ein Wer von 1 dagegen, dass der Text sehr stark positiv gestimmt ist. Ein Wert von 0 bedeutet, dass der Text in Bezug auf die Stimmung als neutral bewertet wird. Ein neutraler Text kann dadurch entstehen, dass kaum Emotionen enthalten sind oder, dass negative wie positive Emotionen sich gegenseitig ausgleichen („mixed"). Eine derartige

Mischung würde man dann durch einen hohen Wert der magnitude erkennen. Ein emotionsloser Text hätte dagegen eines sehr geringen magnitude-Wert. Wie bei der Entitätsanalyse gibt es auch zur Berechnung von Sentimenten zahlreiche Bibliotheken und APIs. Für diese Studie wurde die Sentimentanalyse aus der Google Cloud NLP API angewendet.

- **Box-Plot:** Eine Darstellung zur statistischen Verteilung von Werten. Mittels eines Rechtecks (Box) wird exakt die Hälfte (50%) der verteilten Daten abgebildet mit oberem und unterem Schwellwert (oberes und unteres Quartil) dieser Box. Auslaufende Linien („Antennen") zeigen den Abstand zu Minimum und Maximum Werte außerhalb der Box. Etwaige Ausreißer werden erkannt und punktuell eingeblendet. Der Fokus der Betrachtung liegt auf der Box und ihrer relativen Position auf der Y-Achse. Auch die Größe der Box gibt Aufschluss darüber, wie viele Daten in dem Rechteck enthalten sind. Ein Strich (als Median bezeichnet) zeigt den Mittelwert der Daten aus der Box. Ein Punkt (x) zeigt den Mittelwert der gesamten Daten (innerhalb und außerhalb der Box).

1.4.4 Zeitraum

Da aus vergangenen Wahlen in anderen Ländern teilweise erhebliche Versuche der Einflussnahme über z.B. soziale Medien bekannt sind, wurde der Fokus in der vorliegenden Studie auf die deutsche Bundestagswahl 2021 gelegt. Der Analysezeitraum erstreckt sich entsprechend dem Projektauftrag auf eine Zeitspanne von 2 Wochen vor und 2 Wochen nach der Bundestagswahl:

Analysezeitraum der Studie: Sonntag, 12.09.2021/00:00 bis Sonntag, 10.10.2021/23:59.

Bei einer wöchentlichen Darstellung wurden die Daten derart selektiert, dass Daten einer Woche von Montag bis Sonntag enthalten sind. Der Tag der Bundestagswahl fällt damit auf den Sonntag der zweiten Woche. Der Start-Tag des Betrachtungszeitraums, der 12.09., soll nicht als separate Woche dargestellt werden und ist somit in wöchentlichen Darstellungen nicht enthalten. Diese Sonderregelung tritt nur bei „2.2.1 Anzahl/Häufigkeit je Mediengattung (#12)" sowie der Darstellung zu Gini-Koeffizienten (siehe Kapitel 2.7 Untersuchungen zur Berechnung der Ungleichverteilung) auf.

1.4.5 Suchbegriffe

In Abstimmung mit dem Auftraggeber wurden Suchbegriffe genutzt, die eine Auswahl von prominenten Persönlichkeiten aus der Bundes- und Landes-Politik, der Bundestagswahl, aktuelle Themen und Begrifflichkeiten in Zusammenhang mit dem Land Baden-Württemberg enthalten.
Hinweis: Bedingt durch die hohe Anzahl der Suchbegriffe mit Bezug zu Baden-Württemberg (11 von 17) ist eine signifikante Auswirkung in Bezug auf Quellen und ermittelte Schlagzeilen aus Baden-Württemberg in der Analyse und den statistischen Auswertungen anzunehmen.
Im Einzelnen wurden folgende Begriffe in die Untersuchung mit einbezogen:

SUCHBEGRIFFE MIT BUNDESTAGSWAHLBEZUG

- Annalena Baerbock
- Armin Laschet
- Olaf Scholz
- Kanzlerkandidaten
- Koalition

SUCHBEGRIFFE MIT BEZUG ZU BADEN-WÜRTTEMBERG / SPITZENKANDIDATEN DER PARTEIEN IN BADEN-WÜRTTEMBERG FÜR DIE BUNDESTAGSWAHL

Spitzenkandidaten der Parteien in Baden-Württemberg für die Bundestagswahl:

- Alice Weidel
- Bernd Riexinger
- Franziska Brantner

- Michael Theurer
- Saskia Esken
- Wolfgang Schäuble

Weitere Begriffe:

- Baden-Württemberg
- Bosch
- Daimler
- Karlsruhe
- Stuttgart

SUCHBEGRIFF MIT ÜBERGEORDNETEM BEZUG

- Klimawandel

Die Reihenfolge in dieser Aufzählung stellt keine Wertung oder Priorität dar. Es wurden immer alle Suchbegriffe einmal pro Stunde abgefragt und, soweit Treffer vorhanden, in einer Datenbank persistiert.

1.4.6 Medienlisten

Bei der Auswertung der erhobenen Schlagzeilendaten wurde u.a. eine Betrachtung hinsichtlich der Zugehörigkeit der Schlagzeilenquellen zu bestimmten Mediengattungen gefordert. Da alle Schlagzeilen auf ein Onlinemedium also ein Medienangebot im Internet verweisen, wurde zur weiteren Einordnung der Hintergrund des Mediums herangezogen, also ob bspw. explizit auf eine Sendung, ein Programm oder ein Titel einer Zeitung Bezug genommen wird. Der folgende Screenshot zeigt einen Auszug, die vollständige Liste ist im Kapitel „3.1 Listen/Datenquellen, die für die Studie genutzt wurden" beigefügt.

1	Quelle	URL	Gattung	Untergattung	BW Bezug	Top50
20	Ariva	Zuordnung_Gattung_LFK_21-10-27	Zeitschrift	Zeitschrift		
21	Arseblog News	arseblog.news	Online	Online Int.		
22	arte	www.arte.tv	Rundfunk	TV int publ.		
23	arte Mediathek	www.ardmediathek.de	Rundfunk	TV int publ.		
24	Ärzte Zeitung	www.aerztezeitung.de	Corporate Publishing	Corporate Publishing		
25	Augsburger Allgemeine	www.augsburger-allgemeine.de	Zeitung	Zeitung lok./reg.		
26	Auto Motor und Sport	www.auto-motor-und-sport.de	Zeitschrift	Zeitschrift	1	
27	Autoactu.com	www.autoactu.com	Online	Online Int.		
28	Autobild	www.autobild.de	Zeitschrift	Zeitschrift		
29	Autobild	amp.autobild.de	Zeitung	Zeitung überreg.		
30	autoevolution.com	www.autoevolution.com	Online	Online Int.		
31	Automobil Industrie	www.automobil-industrie.vogel.de	Zeitschrift	Zeitschrift		
32	Automobilwoche.de	www.automobilwoche.de	Zeitschrift	Zeitschrift		
33	Automotive News	www.autonews.com	Zeitschrift	Zeitschrift int.		
34	Automotive News Europe	europe.autonews.com	Zeitung	Zeitung int.		
35	Autovisie	www.autovisie.nl	Online	Online Int.		
36	AutoWeek	www.autoweek.nl	Online	Online Int.		
37	Baden Online	www.bo.de	Zeitung	Zeitung lok./reg.	1	
38	baden.fm	www.baden.fm	Rundfunk	TV privat	1	
39	badenTV	www.baden-tv.com	Rundfunk	TV privat	1	
40	Baden-Württemberg	www.baden-wuerttemberg.de	regÖA	reg./ÖA	1	32
41	Badische Neueste Nachrichten	bnn.de	Zeitung	Zeitung lok./reg.	1	12
42	Badische Zeitung	www.badische-zeitung.de	Zeitung	Zeitung lok./reg.	1	31
43	Badisches Tagblatt	www.badisches-tagblatt.de	Zeitung	Zeitung lok./reg.	1	
44	Bankier.pl	www.bankier.pl	Online	Online int.		
45	BASIC thinking	www.basicthinking.de	Online	Online		
46	Baublatt	www.baublatt.ch	Zeitschrift	Zeitschrift int.		
47	Bayerischer Rundfunk	www.br.de	Rundfunk	Öffentl. -rechtl.		37
48	BB Heute	www.bbheute.de	Zeitung	Zeitung lok./reg.		
49	BBC	www.bbc.com	Rundfunk	TV int publ.		
50	BBC	www.bbc.co.uk	Rundfunk	TV int.		
51	Beinsports	www.beinsports.com	Online	Online Int.		

Tabelle 2 Zuordnung von Gattungen und Untergattungen zu den Quellen

1.5 Technische Umsetzung

Die Durchführung der Studie „Abbildung von Vielfalt und Regionalität in Suchmaschinen" gliederte sich in 2 Phasen, die sich direkt aneinander anschlossen:

Phase 1: Erfassung von Schlagzeilenelementen über einen Zeitraum von 4 Wochen
Phase 2: Auswertung der erfassten Daten entsprechend der vorgegebenen Fragestellungen

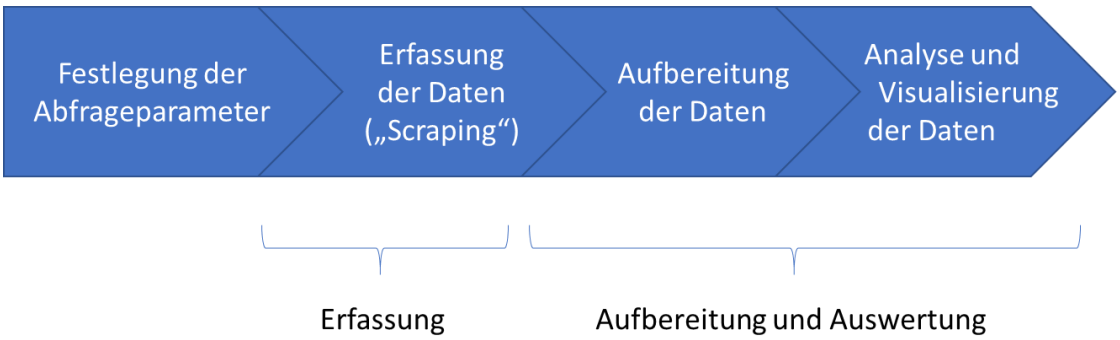


Abbildung 5: Phasen bei der Durchführung der Studie

Zunächst wurden die Rahmenparameter (etwa Suchbegriffe) definiert. Danach erfolgte die Datensammlung mittels der „Scraping Lösung" durch stündliche Abfragen aller vereinbarten Suchbegriffe während der Laufzeit von 4 Wochen. Die Ergebnisse der Abfragen wurden in einer Google Cloud Projektumgebung gesammelt und persistiert. Im weiteren Verlauf wurden die Ergebnisdaten konsolidiert, aufbereitet und qualitätsgesichert. Im Anschluß daran wurde auf Basis des Einsatzes von Google BigQuery Auswertungen und einer Natural Language Processing (NLP) Verarbeitung eine Erstanalyse der Datenbestände vorgenommen. Die Erzeugung der finalen Auswerteergebnisse erfolgte durch Datenexport nach MS Excel. Die MS Excel Umgebung wurde auch für die grafische Aufbereitung der Analyseergebnisse genutzt. Eine Ausnahme bilden dabei die WordClouds (siehe zum Beispiel Kapitel 2.5.1 Zusammenhang zwischen Schlagwörtern und Suchbegriffen (#5)), für die auf spezifische Python Bibliotheken zurückgegriffen wurde.

1.5.1 Erfassung der Schlagzeilenelemente

Im Vorfeld der Implementierung eines Verfahrens zur Erfassung („Scraping") der Schlagzeilen-elemente wurden zunächst mögliche Lösungsansätze wie z.B.

- der Aufbau einer Erfassungs- und Auswertelösung in einer virtuellen oder on premises Server Umgebung,
- die Nutzung einer Cloud Umgebung,
- die Erfassung der Daten durch Nutzung einer eigenprogrammierten Lösung auf Basis von z.B. Selenium,
- oder die Nutzung eines Scraping Dienstleisters

betrachtet und im Hinblick auf folgende Anforderungen geprüft:

- Konfigurierbarkeit
- Simulation unterschiedlicher Anfragebedingungen (z.B. Desktop, Mobil Abfragen, Abfragen eines fiktiven Nutzers aus Stuttgart bzw. außerhalb Stuttgarts)
- Zuverlässigkeit und Skalierbarkeit der Lösung
- Vermeidung der Detektion und infolgedessen eines möglichen Blockens der Abfragen durch die Suchmaschine
- Kosten- und Implementierungsaufwand

Nach Abwägen der Eigenschaften der einzelnen Lösungsansätze hinsichtlich der genannten Anforderungen fiel die Entscheidung zugunsten einer Google Cloud-basierten Lösung, die zur Durchführung der Abfragen einen kommerziellen Scraping Dienst (Anbieter: „ScrapingBee" (https://www.scrapingbee.com)) nutzt.

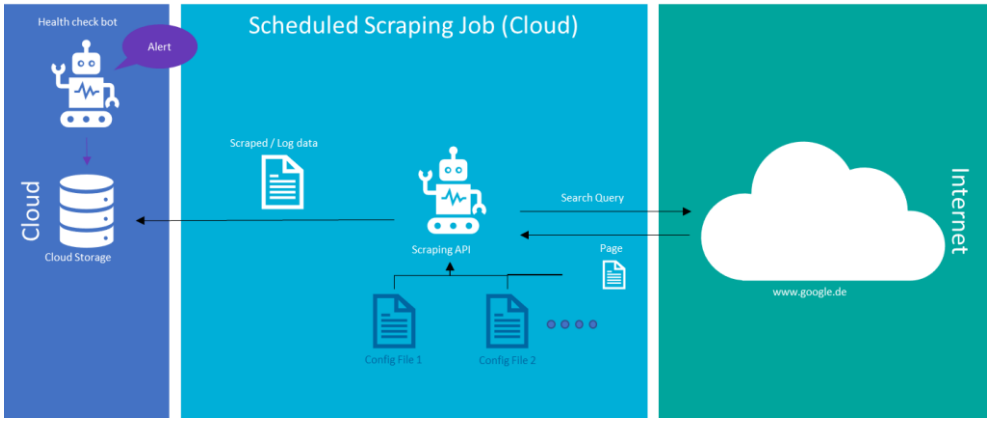


Abbildung 6: Grundprinzip bei der Datenerfassung

Unter Nutzung einer Programmierschnittstelle (API) des Scraping Dienstes erzeugt der für die Studie entwickelte Bot Anfragen („SearchQuery") an Google, die die menschlichen Suchanfrage an google.de simulieren. Die Anfragen können dabei so konfiguriert werden, dass unterschiedliche Abfragebedingungen berücksichtigt werden. Dies sind im Einzelnen

- der gewünschte Suchbegriff (Beispiel: „Klimawandel"),
- der Ort, von dem aus die Abfrage gesendet wurde (Beispiele „Stuttgart" oder „Frankfurt"),
- die Information, ob die Anfrage von einem mobilen oder einem Desktop Endgerät stammt (Beispiel „Mobile" oder „Desktop").

Das Ergebnis wird vom Scraping Dienst in Form einer HTML Seite zurückgeliefert, die im Folgenden weiter verarbeitet und zusammen mit den jeweiligen Aufrufparametern in einem Speicherbereich der für das Projekt angelegten Google Cloud Umgebung abgelegt wird. Weitere Funktionen zur Überwachung des Scraping Vorgangs („Health Check Bot") ergänzen die Lösungsarchitektur.

Zur Veranschaulichung sind in der nachfolgenden Abbildung die einzelnen Funktionsblöcke und Unterblöcke auf einer weiteren Detaillierungsstufe dargestellt.

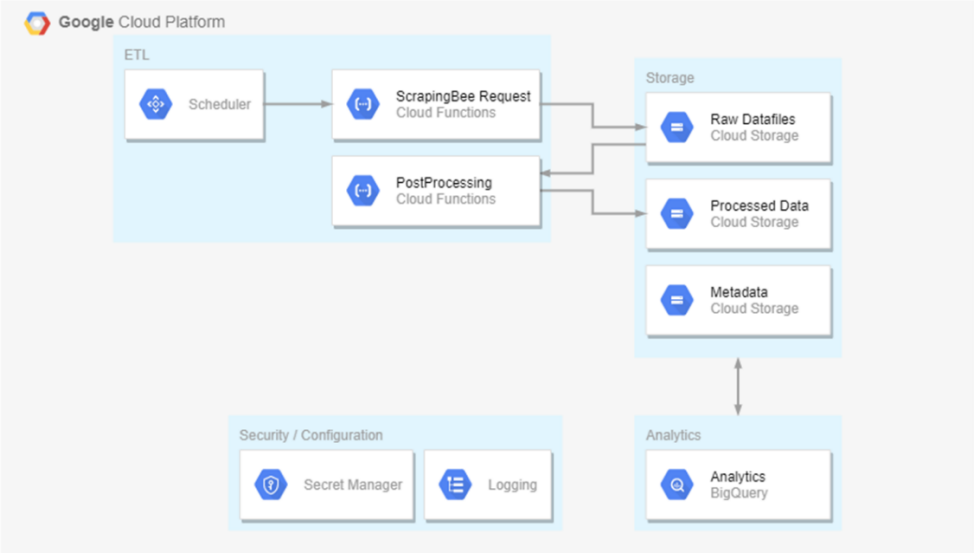


Abbildung 7: Funktionsblöcke der Google Cloud Umgebung

Dies sind insbesondere der ETL-, der Storage-, der Analytics- und der Security/Configuration-Block, deren Funktion in der nachfolgenden Tabelle kurz beschrieben ist.

Block	Erläuterung
ETL („Extract-Load-Transform“)	Erzeugung der Abfrage Requests und Versenden an google.de Entgegennahme der zurückgelieferten Daten Zeitliche Steuerung (Scheduling) der Abfragen
Storage	Unterschiedliche Speicherbereiche in der Google Cloud Umgebung für Rohdaten sowie die verarbeiteten Daten
Analytics	Zusammenfassung der Auswertefunktionen zur Analyse der erfassten Daten
Security/Configuration	Aufzeichnung von Statusdaten, um die einwandfreie Funktion der Datenerfassungs- und Auswerte Lösung fortlaufend überwachen und dokumentieren zu können

Tabelle 3 Funktionen der Erfassung („Scraping“)

Zusammen mit den Eingabeparametern Suchterm, Abfrageort und Abfragegerät (Device Typ) zeigt die nächste Tabelle eine Zusammenstellung der wesentlichen Datenfelder, die aus der auf die Abfrage zurückgelieferten html Seite im Rahmen der Nachverarbeitung extrahiert werden.

Nr	Datenfeld	Datentyp	Beispiel	Bemerkungen
1	**Zeitpunkt** der Anfrage	<TIMESTAMP>	'2021-08-06T10:41:00.020911'	String Isoformat (datetime.now().isoformat())
2	ID des Servers	<Abfragekonfig>	1, 2, ….	Verweis auf Anfrage-Konfiguration
3	**Suchterm**	<STRING>	Olaf Scholz	
4	URL der Schlagzeile	<STRING>	https://www.stuttgarter-zeitung.de/...	
5	Anbietername (**Quelle**)	<STRING>	Stuttgarter Zeitung	
6	**Rang** jeder URL	<INT64>	3 (von 1…10)	Position im Schlagzeilenfenster
7	Gesamtanzahl Schlagzeilen	<INT64>	1…10 (Maximum)	Gesamtanzahl der Schlagzeilen im Schlagzeilenfenster
8	**Aktualität** (Veröffentlichungsdatum)	<STRING>	vor 2 Stunden	Berechnung des Veröffentlichungszeitpunkts aus dem Alter und dem Abfragezeitpunkt
9	**Schlagzeile**	<STRING>	So könnte die Stammelf von Ramon Gehrmann aussehen	
10	**Ort** der Anfrage	<STRING>	Stuttgart	Stuttgart oder Frankfurt
11	**Device** Typ	<STRING>	Mobile	Mobile oder Desktop
12	Header der Anfrage	<STRING>	…	Wiederholbarkeit der Anfrage

Tabelle 4 Datenfelder der Erfassung („Scraping")

Damit stehen für die spätere statistische Auswertung insbesondere die Datenfelder

- Zeitpunkt der Abfrage
- Suchterm/-begriff
- Für jede zurückgelieferte Schlagzeile
  - Quelle
  - Rang
  - Alter der Schlagzeile
  - Schlagzeile
- Ort der Anfrage
- Anfragegerät

zur Verfügung.

1.5.2 Aufbereitung der Daten

Bei der Analyse der Rohdaten ist aufgefallen, dass es bei der Datenerfassung zu einer Reihe von Besonderheiten gekommen ist, die im Folgenden kurz beschrieben werden sollen.

- Suchbegriffe ohne Schlagzeilenfenster
  Für einige Suchbegriffe liefert die Google Schlagzeilenfunktion nur eine geringe Anzahl an Schlagzeilenelementen zurück, wobei vereinzelt das Schlagzeilenfenster komplett in der Ergebnisseite fehlen kann. In diesem Fall finden sich auch keine Schlagzeilenelemente in der Ergebnisliste.

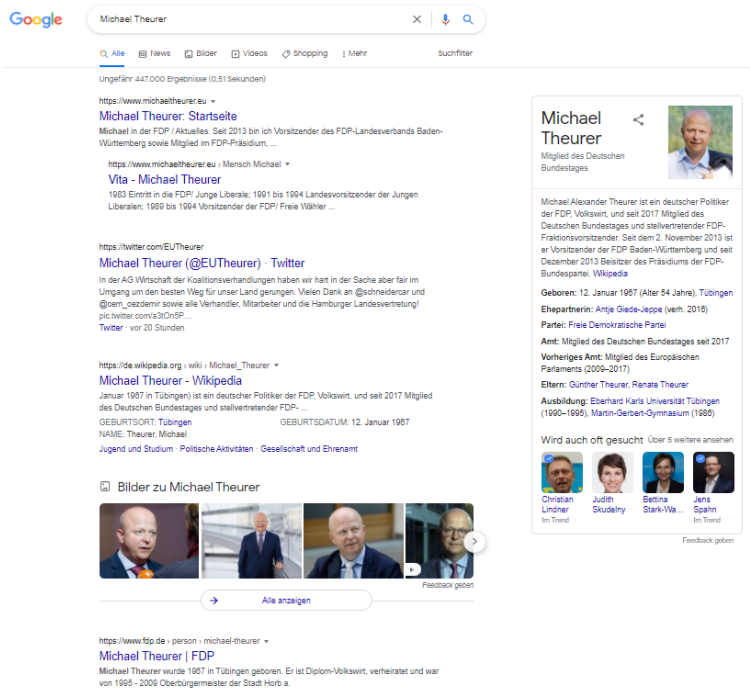


Abbildung 8: Fehlen des Schlagzeilenfensters

- Twitter Meldungen anstelle von Schlagzeilen
  Ein weiterer, ebenfalls selten beobachteter Effekt war, dass anstelle des Schlagzeilenfensters lediglich ein Fenster mit Twitter Meldungen als Ergebnis auf die Suchanfrage zurückgeliefert wurde. Auch in diesem Fall war keine Extraktion von Schlagzeilen möglich.

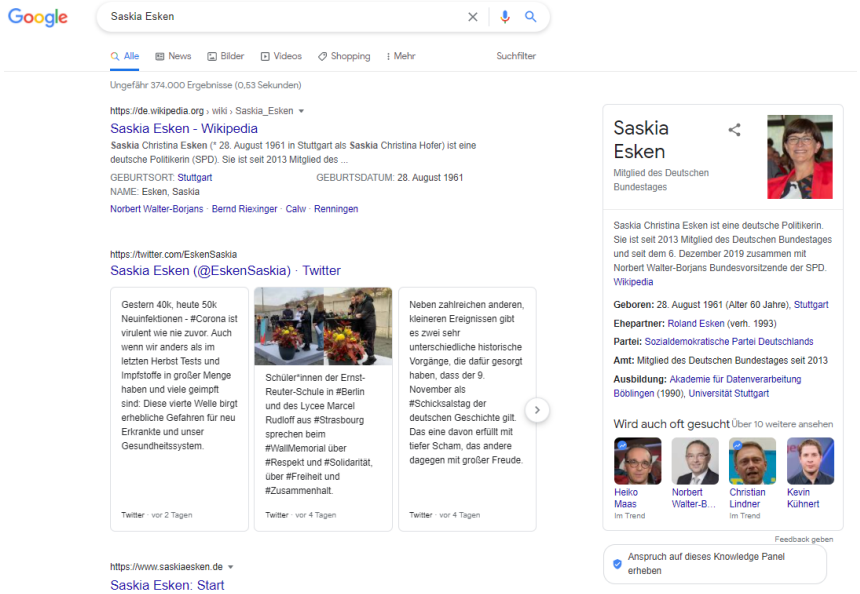


Abbildung 9: Fenster mit Twitter Meldungen anstelle des Schlagzeilenfensters

- Mehrfaches Auftreten von identischen Schlagzeilen innerhalb eines Abfragezyklus
  Bei einigen Abfragezyklen wurden Schlagzeilen teils mehrfach zurückgeliefert, d.h. für die gleiche Schlagzeile der gleichen Quelle innerhalb eines bestimmten stündlichen Abfragezeitraums ergaben sich mehrere Schlagzeilenelemente. Die Ursache dieses Phänomens wird auf Seiten des verwendeten Scraping Dienstes vermutet. Um zu verhindern, dass der genannte Effekt zu einer übermäßigen Gewichtung der betreffenden

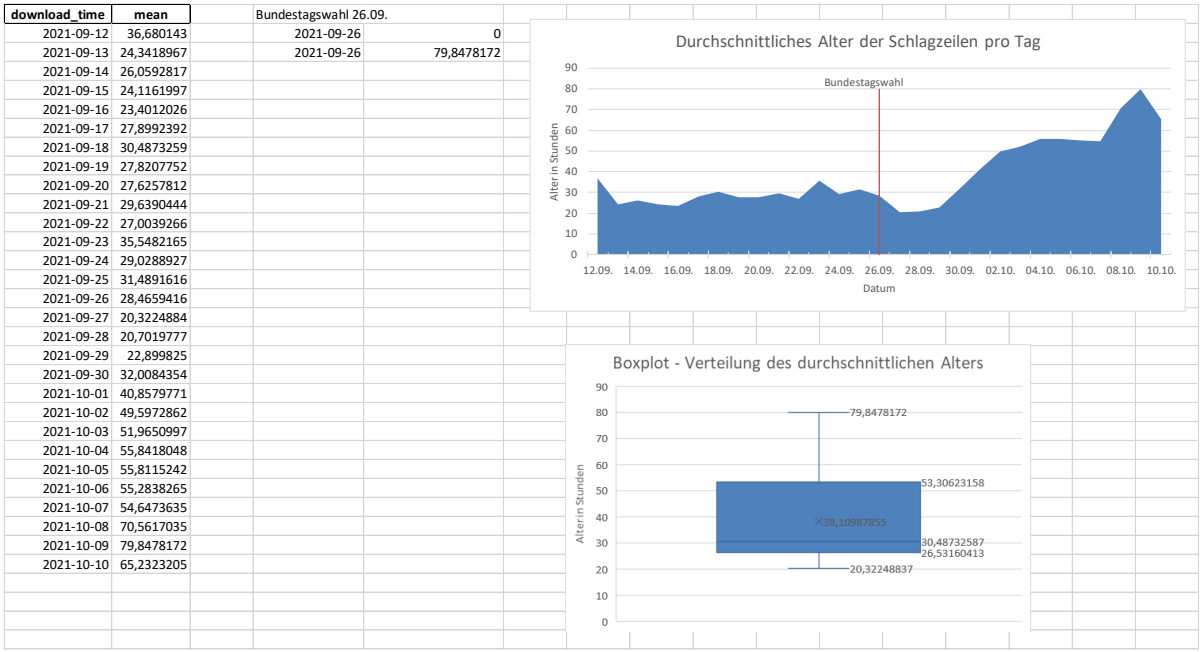
Schlagzeilen in der Auswertestatistik führt wurden im Rahmen der Datenvorverarbeitung in diesem Fall die mehrfachen (identischen) Schlagzeilen entfernt und nur die erste eindeutige Schlagzeile beibehalten.

- Fehlender Titel der Schlagzeilen: Schlagzeilen, bei denen der Titel nicht angegeben ist aus der Schlagzeilenfunktion von Google, können nicht verarbeitet werden und sind daher aus der Betrachtung der Analyse ausgeschlossen worden.
- Fehlende Quelle der Schlagzeile: Für manche Schlagzeilen wurde aus der Schlagzeilenfunktion von Google nur ein Punkt („.“) als Quelle überliefert. Für den Großteil dieser Datensätze konnte die Quelle durch eine Betrachtung der URL der Schlagzeilennachricht nachträglich aufgelöst und hinzugefügt werden. Dafür wurden andere Schlagzeilen mit demselben Hostnamen der URL herangezogen und die Quelle der herangezogenen Daten genutzt.
- Schlagzeilen mit identischem Hostnamen Youtube: Einige Schlagzeilen waren mit einem Link zu der Seite Youtube hinterlegt. Dadurch konnte nicht für alle Fälle eine eindeutige Gattung zugewiesen werden. Eine manuelle Korrektur dieser Zuweisung wurde, soweit nachvollziehbar, durchgeführt.
- Quellen mit unterschiedlicher Schreibweise: Im Fall der Tagesschau existierten einige Datensätze, bei denen die Quelle in Kleinschreibweise angegeben war. Für die korrekte Verarbeitung der Daten, wurde dieser Quellenname angepasst.

Durch die umfangreiche Analyse der Besonderheiten sowie durch die getroffenen Korrekturmaßnahmen wurde die Datenqualität der Rohdaten deutlich verbessert, sodass dies einen signifikanten Mehrwert für die weitere Analyse der Daten darstellt.

1.5.3 Analyse und Visualisierung der Daten

Im letzten Schritt wurden die vorverarbeiteten Daten aus der Google Cloud heraus exportiert, in MS Excel importiert, dort unter Nutzung von MS Excel Funktionen weiter aufbereitet und schließlich mittels der entsprechenden Grafikfunktionen visualisiert.



download_time	mean
2021-09-12	36,680143
2021-09-13	24,3418967
2021-09-14	26,0592817
2021-09-15	24,1161997
2021-09-16	23,4012026
2021-09-17	27,8992392
2021-09-18	30,4873259
2021-09-19	27,8207752
2021-09-20	27,6257812
2021-09-21	29,6390444
2021-09-22	27,0039266
2021-09-23	35,5482165
2021-09-24	29,0288927
2021-09-25	31,4891616
2021-09-26	28,4659416
2021-09-27	20,3224884
2021-09-28	20,7019777
2021-09-29	22,899825
2021-09-30	32,0084354
2021-10-01	40,8579771
2021-10-02	49,5972862
2021-10-03	51,9650997
2021-10-04	55,8418048
2021-10-05	55,8115242
2021-10-06	55,2838265
2021-10-07	54,6473635
2021-10-08	70,5617035
2021-10-09	79,8478172
2021-10-10	65,2323205

Bundestagswahl 26.09.	
2021-09-26	0
2021-09-26	79,8478172

Abbildung 10: Beispiel einer Datenauswertung und -visualisierung in MS Excel

Der Vorteil des gewählten Vorgehens liegt darin, dass die im Rahmen der Studie erzeugten Auswertungen damit nachvollziehbar sind und durch den Auftraggeber nachträglich Änderungen an der Ergebnisvisualisierung vorgenommen werden können.

Einen Sonderfall stellen die WordClouds dar, die mit Hilfe von eigens entwickelten Python Skripten und dem Einsatz von Grafikbibliotheken wie Mathplotlib und WordCloud erzeugt wurden. Auf Details hierzu wird in dem entsprechenden Kapitel näher eingegangen.

2 Auswertungen der erfassten Daten

Ausgangspunkt für die statistische Auswertung der im Rahmen der Studie „Abbildung von Vielfalt und Regionalität in Suchmaschinen" erfassen Daten waren zunächst 8 Fragestellungen, die seitens des Auftraggebers im Vorfeld der Studie formuliert wurden:

1. Wie **vielfältig** sind die Ergebnisse der Google-Schlagzeilen in ihrer Gesamtheit und für jeden Suchterm? Welche Strukturen und Gesetzmäßigkeiten sind hierbei zu erkennen? Welche **Faktoren** sind hierbei zu identifizieren?
2. Wie **verändert sich die Vielfalt** bzw. welche Veränderung erfahren die sie beschreibenden Parameter im **wöchentlichen Wechsel**?
3. Wie **verteilen** sich die **Schlagzeilen** auf die verschiedenen **Mediengattungen** in Abhängigkeit von **Suchterm** und **Standort** des Servers?
4. Welchen **Anteil** an den Google-Schlagzeilen haben **Domains** aus Baden-Württemberg (z.B. swr.de, stuttgarter-zeitung.de)?
5. Wenn man nur die Sets betrachtet, in denen Unternehmen aus Baden-Württemberg vorkommen, welchen **Rang** nehmen diese Unternehmen im Durchschnitt ein?
6. Welche **Unterschiede** zeigen sich, wenn man nicht alle 10 Schlagzeilen eines Sets nach obigen Fragestellungen untersucht, sondern nur die **jeweils ersten drei Schlagzeilen**?
7. Welche **Unterschiede** ergeben sich bei den zwei Servern? Welches Maß an **Regionalisierung** ist daraus abzuleiten?
8. Der Anbieter soll im Rahmen seines Angebots prüfen, ob eine **Clusteranalyse** geeignet ist, Zusammenhänge in den Daten zu erkennen und diese ggfs. anbieten.

Im weiteren Studienverlauf und als Ergebnis intensiver Diskussionen wurde daraus ein Fragenkatalog mit mehr als 20 Detailfragen entwickelt (siehe Kapitel 2 Auswertungen der erfassten Daten), der daraufhin die Grundlage für die Durchführung der Auswertungen bildete.

Da sich in der Reihenfolge der Fragen weitgehend der zeitliche Verlauf der Diskussion widerspiegelt, war die Zusammenfassung von inhaltlich miteinander zusammenhängenden Fragen zunächst nicht gegeben. Aus diesem Grund wurde eine Ordnungsstruktur entwickelt, die die Fragen auf Basis der Konzepte

- Quellen
- Mediengattungen
- Suchbegriffe
- Schlagzeilen
- Schlagwörter
- Lokalisierungen Baden-Württemberg

zuordnet und eine gute Orientierung innerhalb des Fragenkatalogs gewährleistet. Der vollständige und thematisch geordnete Fragenkatalog ist in der nachfolgenden Tabelle wiedergegeben:

Kategorie	Nr	Fragestellung	Erläuterung/Beispiel
Quelle	2	Welche Quellen hatten die meisten Schlagzeilen?	Beispiel: Der Spiegel hat unverhältnismäßig viele Ergebnisse
	4	Wie ist die Verbindung zwischen Quellen und Rang der Schlagzeilen?	Beispiel: Besonders häufig sind in den Top 3 Schlagzeilen die Quellen "Welt" und "Spiegel" zu finden Ziel: Erkennung der Wichtigkeit einzelner Quellen
	8	Welche Sentimente haben die Schlagzeilen, gruppiert nach Quellen?	Beispiel: Schlagzeilen der "BILD" sind sehr sentimentbelastet. Anm.: Die Auswahl der Quellen kann begrenzt sein
Mediengattungen	12	Welche Mediengattungen hatten die meisten Schlagzeilen?	Beispiel: "Zeitung" hat unverhältnismäßig viele Ergebnisse

Kategorie	Nr	Fragestellung	Erläuterung/Beispiel
	23	Wie ist die Verteilung der Mediengattung in Bezug auf Suchterme?	Beispiel: "Olaf Scholz" bekommt vor allem Schlagzeilen aus der Mediengattung "Zeitung".
Suchbegriffe	1	Welcher Suchbegriff erreichte die meisten Schlagzeilen?	Beispiel: Die meisten Schlagzeilen gibt es für den Suchbegriff "Armin Laschet", auf Platz 2 "Olaf Scholz", ...
	7	Welcher Suchbegriff hat überwiegend welche Sentimente?	Beispiel: Suchbegriff "Janine Wissler" hatte in Woche KW 39 überwiegend negative Schlagzeilen
	10	Wie entwickelt sich das Sentiment der Schlagzeilen für die Suchbegriffe?	Beispiel: Für den Suchbegriff "Armin Laschet" wurden die Schlagzeilen zunehmend negativer
	3	Wie ist die Verbindung zwischen Quellen und Suchbegriff?	Beispiel: Besonders häufig wurden Schlagzeilen über "Armin Laschet" vom "Spiegel" erstellt.
Schlagzeilen	16	Wie aktuell sind die Schlagzeilen?	Beispiel: Für Tag X waren die Schlagzeilen durchschnittlich 4h "alt".
	17a	Wie langlebig sind die Schlagzeilen in den Top3 bzw. Top10 in Bezug auf Quellen und Mediengattungen?	Beispiel: Schlagzeile X ist durchgängig über 3 Tage hinweg aufgetaucht. Anm: Gruppierung nach Mediengattung oder Quelle ist möglich.
	17b	Wie langlebig sind die Schlagzeilen in Bezug auf die Suchbegriffe?	
	19	Verändert sich der Rang einer Nachricht über der Zeit?	Beispiel: Schlagzeilen gehen durchschnittlich nach 5 Stunden einen Rang weiter nach hinten. Gruppierung nach Schlagzeile, Bewertung der Ränge/Zeiten; Filter nach Anzahl der Auftritte
	20	Welche Schlagzeilen sind textuell ähnlich?	
Schlagwörter	5	Welche Schlagwörter kommen in Verbindung mit den Suchbegriffen besonders häufig vor?	Beispiel: Für den Suchbegriff "Annalena Baerbock" kommt das Schlagwort "Klimawandel" besonders häufig vor. Anm.: Schlagwortzuordnung erfolgt über die Google Cloud NLP API.
	6	Gibt es für Suchbegriffe Schlagwörter, die sich besonders lange im Schlagzeilenfenster halten?	Beispiel: Das Schlagwort "Cum Ex" hat sich für den Suchbegriff "Olaf Scholz" am längsten gehalten.
	11	Welche Schlagwörter befinden sich am häufigsten in den Top 3 Schlagzeilen bei Google?	Beispiel: Das Schlagwort "E-Bike" taucht häufig bei den Top 3 Schlagzeilen auf
	22	(Vorausgesetzt: Liste mit Schlagworten) Wie oft kommen spezifische Schlagworte in den Schlagzeilen vor?	Das Schlagwort "Triell" kommt 507 mal vor
Lokalisierungen Baden-Württemberg	13	Welchen Anteil an den Google-Schlagzeilen haben Quellen aus Baden-Württemberg (z.B. swr.de, stuttgarter-zeitung.de)?	Beispiel: "swr.de" gehören 5% der Schlagzeilen an

Kategorie	Nr	Fragestellung	Erläuterung/Beispiel
	14	Welchen Rang nehmen Quellen aus BaWü bei dem Rang der Schlagzeilen pro Suchanfrage ein?	Beispiel: "swr.de" ist durchschnittlich auf Rang 6 der Suchanfragen
	14a	Welche Quellen aus BaWü schaffen es auf Platz 1-3?	
	14b	Verteilung der Ränge komplett, nicht nur Durchschnitt als Histogramm (nur BaWü) auch nach Alle / STG und Fran als Abfragestandort	
	18	Wie schneiden die BaWü-Quellen untereinander ab in Bezug auf die Häufigkeit?	Beispiel: Quelle A kommt doppelt so häufig vor wie Quelle B in Schlagzeilen.
	21	In Schlagzeilen von Quellen aus BaWü: ist ein Bezug zu Ortschaften in BaWü gegeben?	Beispiel: In Schlagzeilen mit Domain "swr.de" kommt häufig die Ortschaft "Gerlingen" vor
	9	Welche Quellen stehen mit welchen Schlagwörtern in Verbindung?	Beispiel: "BaWü News" berichtet besonders häufig über Schlagzeilen mit dem Schlagwort "Tempolimit"
Berechnung von Ungleichverteilungen	24	Lorenz-Kurve und Gini-Koeffizient können für ein Unzahl von Kombinationen aus verwendetem Suchterm oder Gruppen von Suchtermen, Rang (nur die ersten drei oder alle 10) und Dauer (Tag, Woche, ganzer Zeitraum) berechnet werden. Welche der Kombinationen sind aussagekräftig?	Beispiel: Gini Koeffizienten können tage-, wochen- oder monatsweise berechnet werden.

Tabelle 5 Kategorisierung

Die nachfolgenden Kapitel orientieren sich an der beschriebenen Ordnungsstruktur und dokumentieren die durchgeführten Auswertungen sowie die resultierenden Ergebnisse. Der Grundaufbau in jedem der Kapitel wurde dabei aus Gründen der besseren Übersichtlichkeit stets gleich gewählt und beschreibt nach der zugrundeliegenden Fragestellung und der angewendeten Abfragetechnik die Ergebnisse der Analyse.

2.1 Quellen

2.1.1 Anzahl Schlagzeilen (#2)

AUFGABENSTELLUNG DER FRAGE:

Untersucht wurde, welche Quellen wie viele Schlagzeilen generiert haben und wie die Quellen dabei im Vergleich zueinanderstehen.

ABFRAGETECHNIK

Die Menge aller Schlagzeilen des Erfassungszeitraums von 4 Wochen wurde nach den Quellen gruppiert und die Anzahlen der zugehörigen Schlagzeilen erhoben. Die Anzahlen der Quellen werden in absteigender Reihenfolge dargestellt.
Für die Berechnung der Lorenzkurve wurde für alle Datensätze eine kumulierende Summe über die Anzahl der Schlagzeilen der Quellen gebildet und jeweils durch die Gesamtanzahl der Schlagzeilen geteilt. Die Daten wurden der Anschaulichkeit halber in umgekehrter Reihenfolge abgebildet (von groß nach klein).
Die Auswertung erfolgte zusätzlich nach Abfragegeräten (Desktop und Mobil).

ERGEBNIS DER ANALYSE:

Die schnell abnehmende Kurve in dem folgenden Diagramm zeigt, dass wenige Quellen den größten Teil der Schlagzeilen generiert haben und im Umkehrschluss die meisten Quellen nur einen geringen Anteil der Schlagzeilen hervorgebracht haben. Man spricht in diesem Fall von einer sogenannten Long Tail Verteilung, die aufgrund ihrer Charakteristik eine deutlich Ungleichverteilung aufweist.

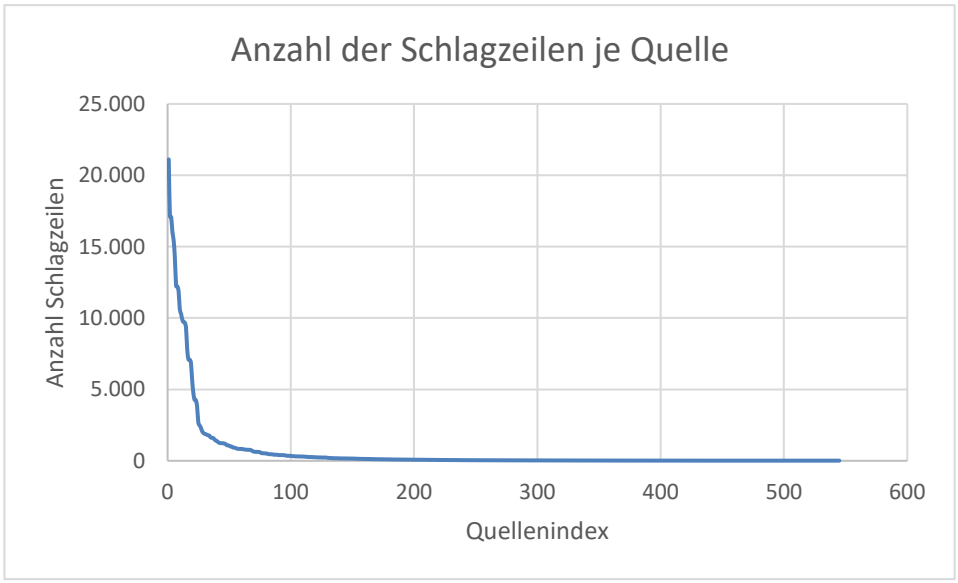


Abbildung 11: Anzahl der Schlagzeilen je Quelle für den Gesamtzeitraum

Eine Darstellung nach Abfragegeräten zeigt keinen signifikanten Unterschied:

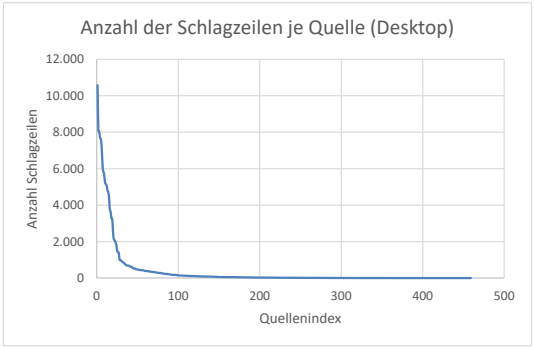


Abbildung 12: Anzahl der Schlagzeilen je Quelle für den Gesamtzeitraum - Desktop

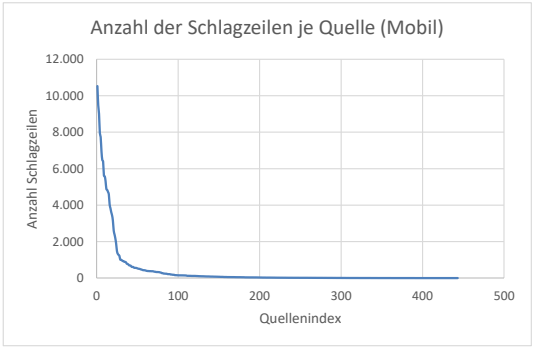


Abbildung 13: Anzahl der Schlagzeilen je Quelle für den Gesamtzeitraum - Mobile

Von den zuvor erwähnten wenigen Quellen mit den meisten Schlagzeilen, werden die 20 Quellen mit den meisten Schlagzeilen in der nachstehenden Darstellung detaillierter betrachtet. Diese zeigt die jeweilige Quelle mit ihrer absoluten Anzahl an Schlagzeilen. Die Quelle „Spiegel" aus der Gattung Zeitschriften zeigt einen klaren Vorsprung gegenüber den übrigen Quellen.

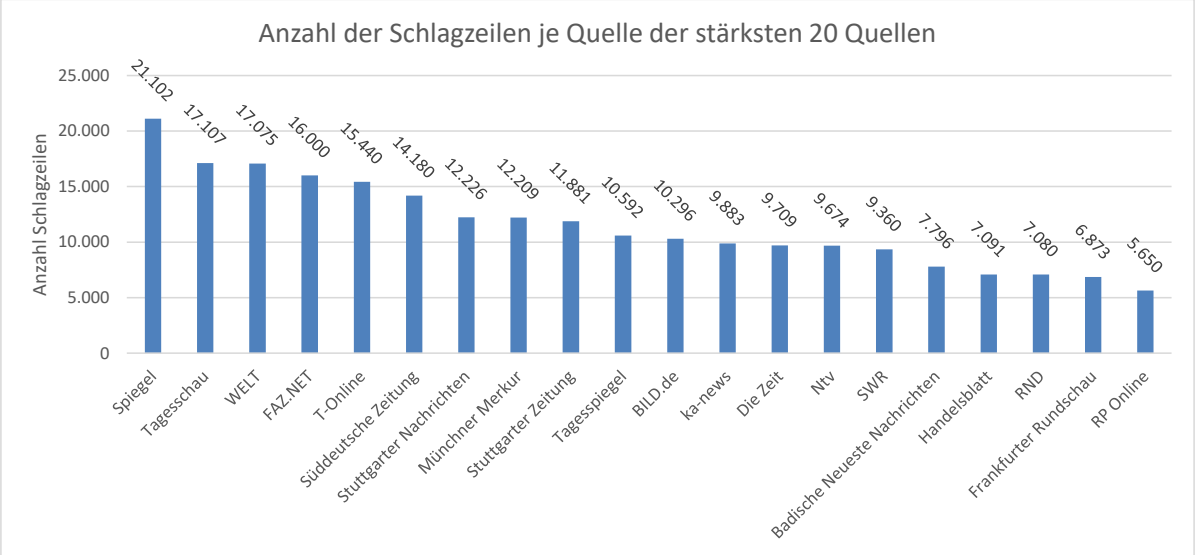


Abbildung 14: Anzahl der Schlagzeilen je Quelle für die 20 „Top Quellen"

Die in der Abbildung erkennbare Ungleichverteilung kann charakterisiert werden durch die sogenannte Lorenzkurve und den hieraus berechneten Gini-Koeffizienten. Die Lorenzkurve weist einen sehr flachen beginnenden Verlauf auf und deutet daher auf ein hohes Maß an Ungleichverteilung hin. Der Gini-Koeffizient drückt dies entsprechend mit einer hohen Zahl nahe 1 aus.

N = 545 Gini-Koeffizient = 0,900

Aus der folgenden Lorenzkurve mit der Darstellung „Schlagzeilen über Quellen" lässt sich ablesen, dass etwa 80% der Quellen nur 5% der Schlagzeilen ausmachen und gleichzeitig ca. 12 % der Quellen etwa 90% der Schlagzeilen ausmachen.

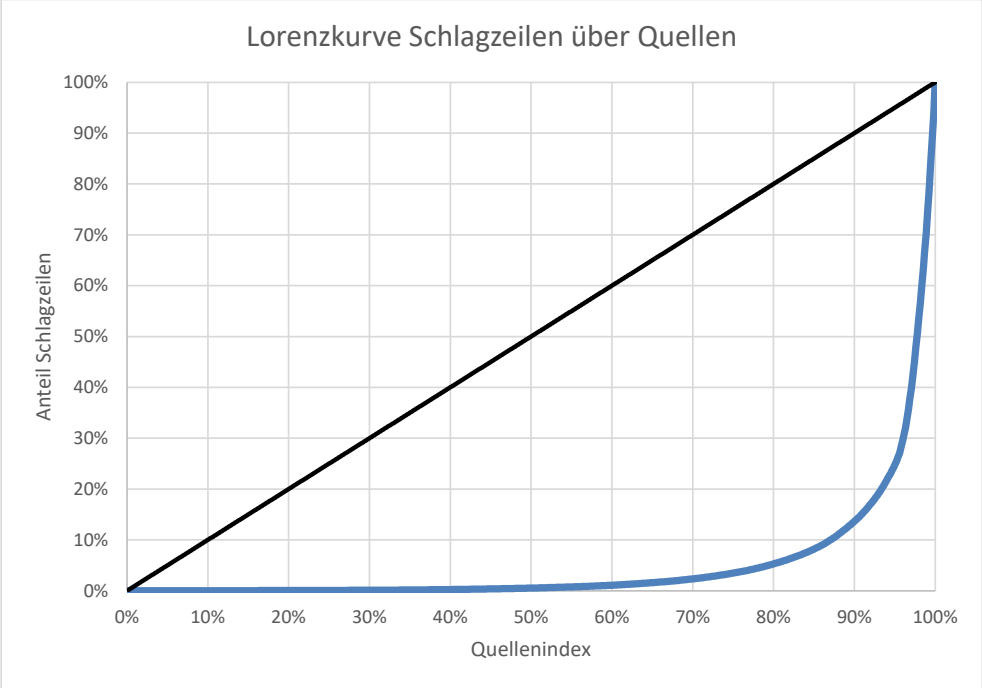


Abbildung 15: Lorenzkurve und Gini-Koeffizient für die Verteilung „Anzahl der Schlagzeilen je Quelle"

Auch hier zeigt eine Darstellung nach Abfragegeräten keinen signifikanten Unterschied:

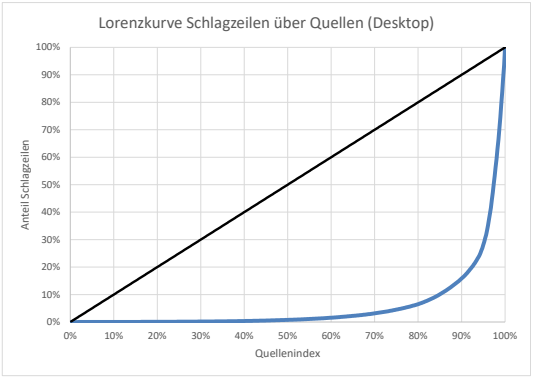


Abbildung 16: Lorenzkurve und Gini-Koeffizient für die Verteilung „Anzahl der Schlagzeilen je Quelle" - Desktop

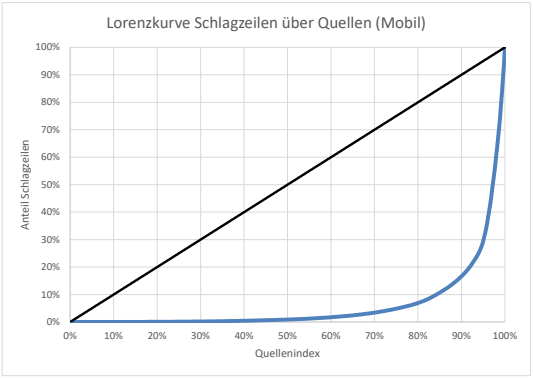


Abbildung 17: Lorenzkurve und Gini-Koeffizient für die Verteilung „Anzahl der Schlagzeilen je Quelle" - Mobile

Aus der folgenden Darstellung lässt sich eine relativ gleichmäßige Verteilung über die Gerätetypen (Mobile und Desktop) ableiten, mit leichten Abweichungen je nach Quelle:

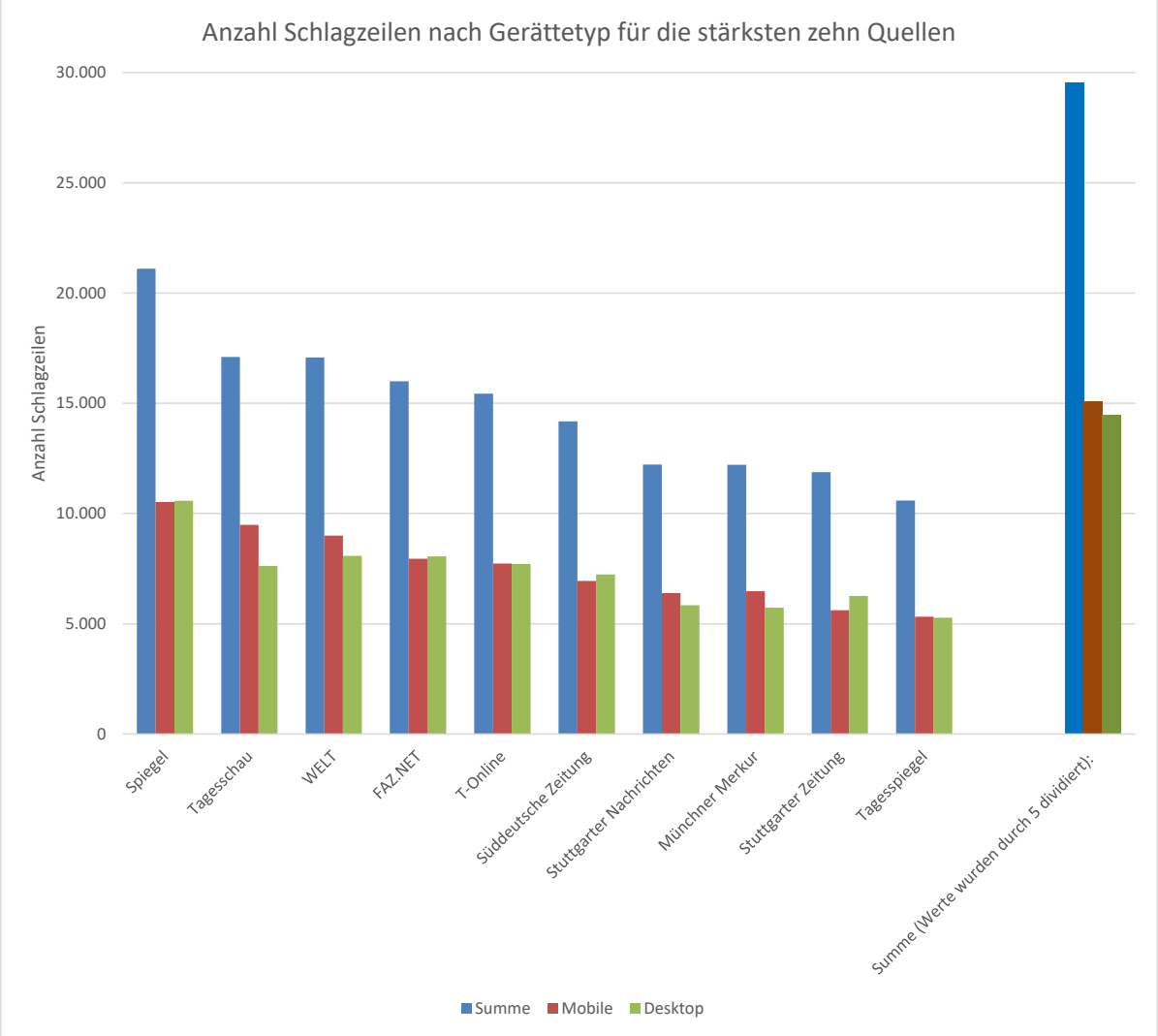


Abbildung 18 Anzahl Schlagzeilen nach Gerätetyp für die 10 stärksten Quellen

2.1.2 Rang der Schlagzeilen (#4)

AUFGABENSTELLUNG DER FRAGE:

Bei dieser Frage wird analysiert, wie die Verbindung zwischen der Anzahl der Schlagzeilen der Quellen und dem Rang der Schlagzeilen ist (bei Betrachtung der Ränge 1-3 bzw. 1-10).

ABFRAGETECHNIK

Es wird der prozentuale Anteil der Schlagzeilen je Quelle ermittelt, und zwar für die Varianten „Betrachtung der Ränge 1-3" und „Betrachtung der Ränge 1-10". Der jeweilige Anteil der stärksten Quellen wird in einem Pie-Chart dargestellt.

Eine Lorenzkurve und ein Gini-Koeffizient wurden anhand der Daten für Schlagzeilen der Ränge 1-3 berechnet. Eine Lorenzkurve für Schlagzeilen der Ränge 1-10 wurde bereits in Zusammenhang mit Frage #2 dargestellt.

ERGEBNIS DER ANALYSE:

Zu erkennen ist, dass das Medium „Spiegel" auch bei Betrachtung der Top Ränge 1-3 den größten Anteil ausmacht. Viele der weiteren Medien platzieren sich ähnlich in beiden Varianten. Wobei aber beispielsweise „ka-news" in der Betrachtung der Ränge 1-10 deutlich abfällt (2,90 zu 4,14%).

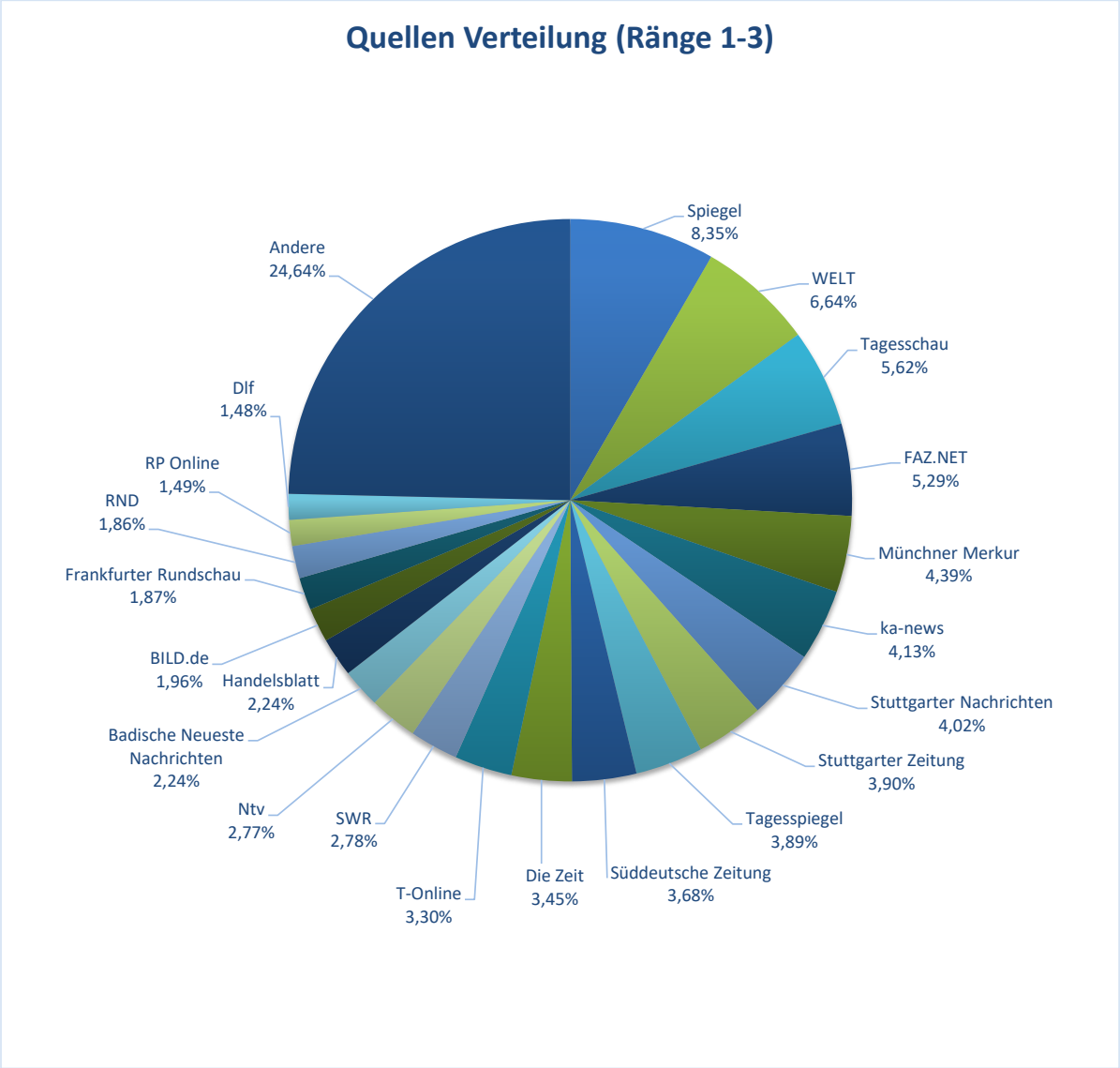


Abbildung 19: Schlagzeilen je Quelle bei Betrachtung der Ränge 1-3

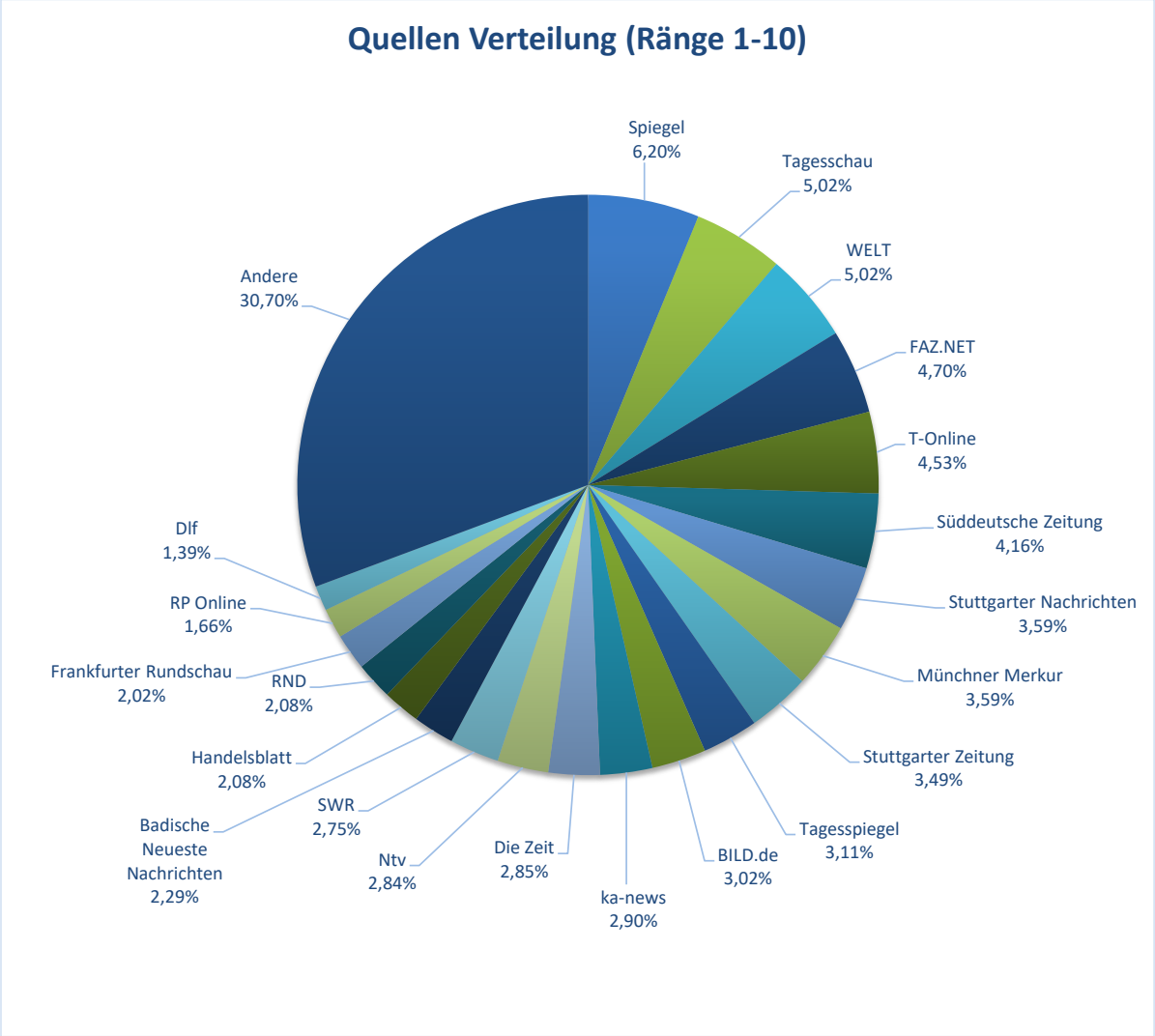


Abbildung 20: Schlagzeilen je Quelle bei Betrachtung der Ränge 1-10

Die Lorenzkurve und der Gini-Koeffizient für Schlagzeilen der Ränge 1-3 unterscheiden sich nur geringfügig von der Darstellung für Schlagzeilen der Ränge 1-10. Auch in diesem Bereich ist eine große Ungleichverteilung festzustellen.

N = 400 Gini-Koeffizient = 0,897

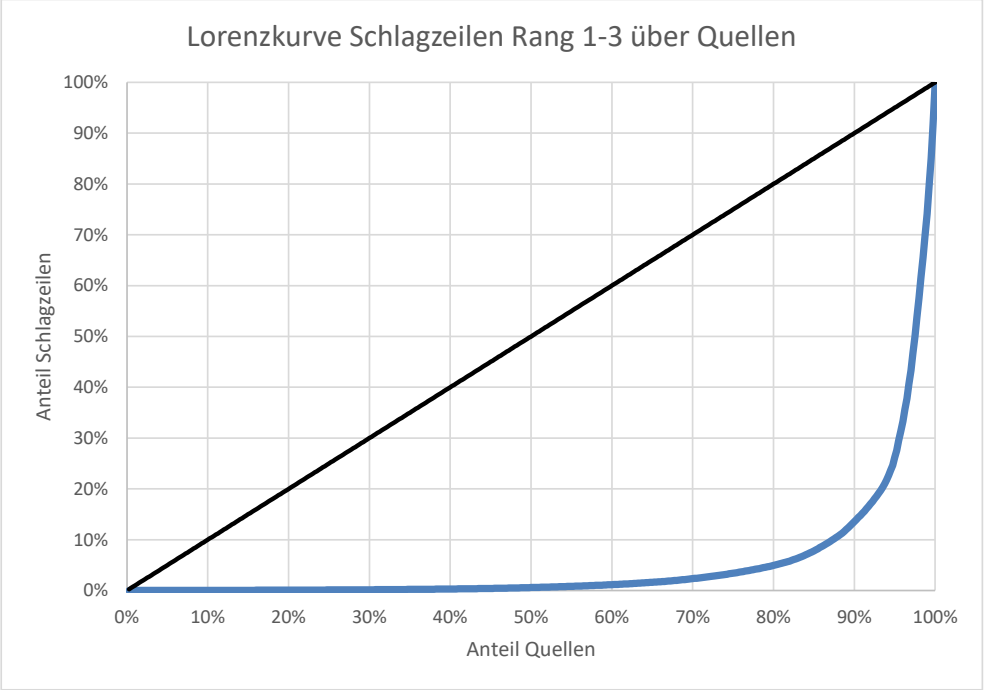


Abbildung 21: Lorenzkurve und Gini-Koeffizient für die Verteilung „Anzahl der Schlagzeilen je Quelle" für die Ränge 1-3

2.1.3 Sentimente der Schlagzeilen (#8)

AUFGABENSTELLUNG DER FRAGE:

Im Zusammenhang mit dieser Frage wird analysiert, ob sich die Quellen hinsichtlich der durchschnittlichen Sentiment Bewertung ihrer Schlagzeilen unterscheiden. Nähere Informationen über Sentimente können in der Begriffsdefinition nachgelesen werden.

ABFRAGETECHNIK

Der Datenbestand wurde nach allen Gattungen gruppiert. Aus jeder Gruppe wurden, gemessen an der Anzahl der Schlagzeilen, die jeweils drei stärksten Quellen je Gattung ermittelt. Zu jeder dieser Quellen wurden alle Schlagzeilen gesammelt. und die Sentimente berechnet. Über die Sentiment-werte aller Schlagzeilen pro ausgewählter Quelle wird dann eine Verteilung der Sentimente in Form einer separaten BoxPlot-Darstellung abgebildet.

ERGEBNIS DER ANALYSE:

Insgesamt ist für allen ausgewählten Quellen eine sehr ähnliche Verteilung der Sentimente zu beobachten. Bei den meisten Darstellungen liegt der Kasten in einem minimal positiven Bereich von 0 bis 0,3. Lediglich die Quelle „Katholisch.de" zeigt ausschließlich Sentimente mit einem konstanten Wert von -0,4 und weist somit eine leicht negative Tendenz auf. Dies liegt daran, dass hier nur sehr wenige Schlagzeilenelemente vorliegen, die sich alle auf denselben Titel beziehen, sodass die gesamte Verteilung auf dem Sentiment nur eines einzigen Schlagzeilentitels beruht.

Ein Beispiel für eine Schlagzeile mit einem Sentimentwert von 0,7 ist die folgende:
„VfB Stuttgart - Bayer Leverkusen 1:3: Dank Blitzstart und Wirtz! Zehn Leverkusener gewinnen in Stuttgart"

Ein Beispiel für eine Schlagzeile mit einem Sentimentwert von -0,6 ist die folgende:
„Chipkrise - Daimler lässt die Kunden büßen"

Im Folgenden werden für alle Gattungen jeweils die drei stärksten (in Bezug auf Anzahl) der Schlagzeilen hinsichtlich ihres Sentimentwertes dargestellt:

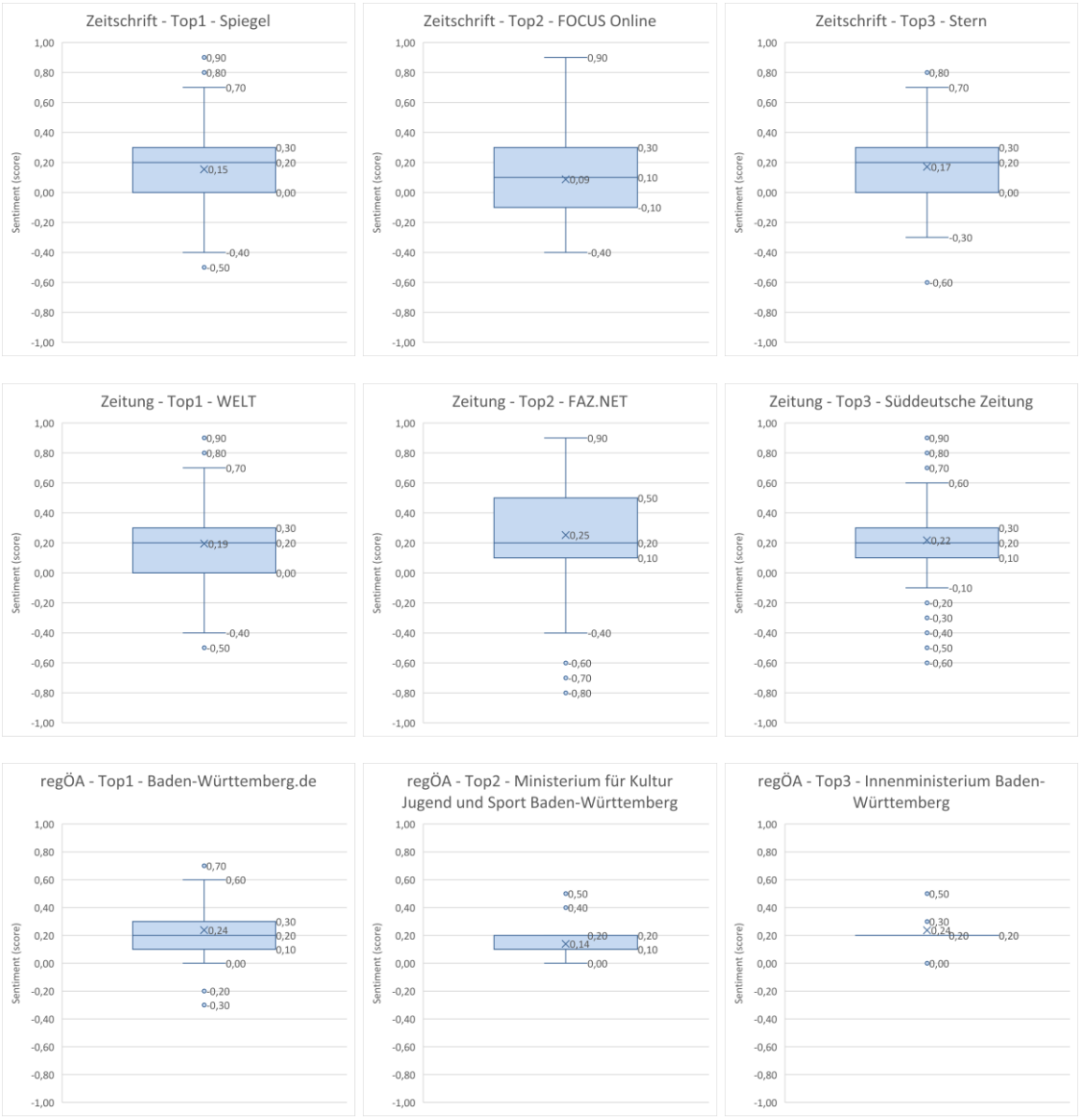


Abbildung 22 Sentimentanalyse für ausgewählte Quellen unterschiedlicher Mediengattungen

2.2 Mediengattungen

2.2.1 Anzahl/Häufigkeit je Mediengattung (#12)

AUFGABENSTELLUNG DER FRAGE:

Diese Frage befasst sich mit der Analyse der Anzahl an Schlagzeilen je Mediengattung, wobei zusätzlich der zeitliche Verlauf für jede der 4 betrachteten Kalenderwochen mitberücksichtigt werden soll.

ABFRAGETECHNIK

Zunächst wird eine Differenzierung in die vier Wochen des Gesamtbetrachtungszeitraums vorgenommen. Die Schlagzeilen werden hierbei nach ihrer jeweiligen Hauptgattung und nicht wie in Abschnitt 2.1 nach der jeweiligen Quelle gruppiert und die Anzahlen der Schlagzeilen ermittelt. Außerdem werden für alle Mediengattungen die jeweils drei stärksten Quellen, nach der Anzahl ihrer Schlagzeilen in dieser Gattung tabellarisch ausgegeben. Abschließend wird die Lorenzkurve und der Gini-Koeffizient für die Verteilung der Schlagzeilen in Abhängigkeit von den Mediengattungen gezeigt.

ERGEBNIS DER ANALYSE:

Die Gattung „Zeitung" sticht hierbei über den gesamten Zeitraum sehr deutlich hervor. Die Mediengattungen „Zeitschrift", „Online" und „Rundfunk" sind im Vergleich zur Gattung „Zeitung" zwar nur gering vertreten, liegen untereinander aber in demselben Zahlenbereich. Die übrigen Gattungen „Corporate Publishing", regÖA" sowie „Red./Agentur" weisen insgesamt sehr niedrige Anzahlen an Schlagzeilen auf.
Ein eindeutiger Trend im Verhältnis der einzelnen Wochen lässt sich bei Betrachtung aller Gattungen gleichermaßen kaum ausmachen. Bei der zahlenmäßig überlegenen Gattung „Zeitung" zeigt sich jedoch ein leichter Trend der zunehmenden Anzahl an Schlagzeilen bis in Woche 3 und eine darauffolgende Abnahme in Woche 4.

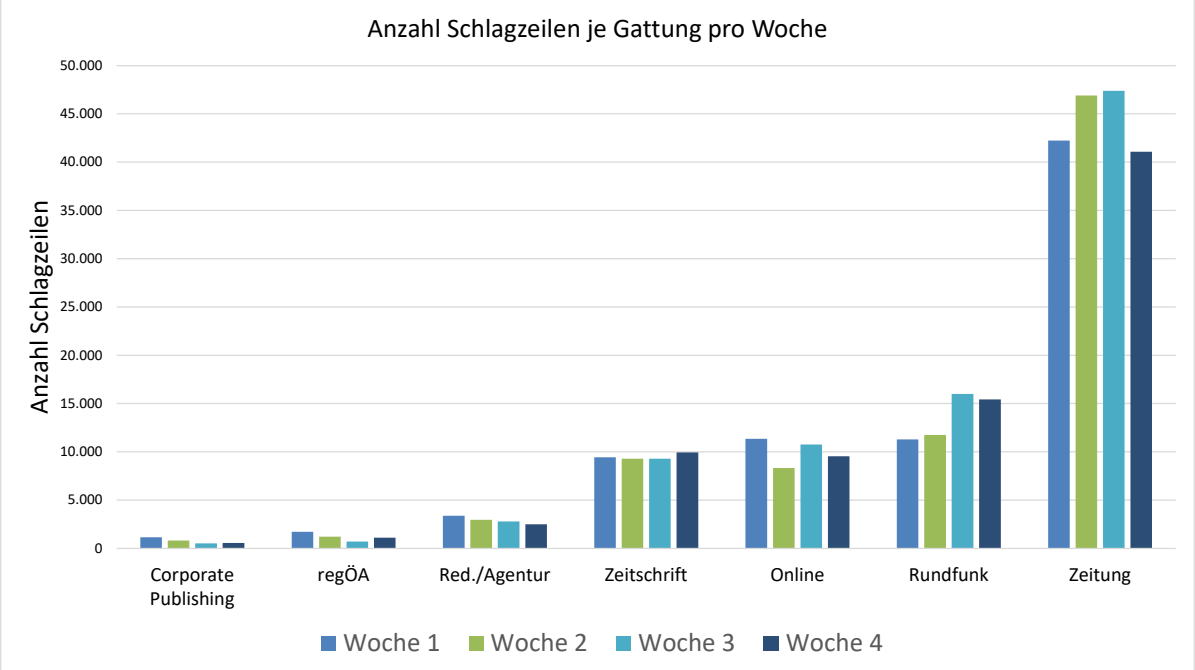


Abbildung 23: Anzahl Schlagzeilen je Mediengattung (Betrachtungszeitraum: 4 Wochen)

Es folgt eine detailliertere Betrachtung der einzelnen Mediengattungen. Zu sehen sind jeweils die Top3 Quellen mit der größten Anzahl an Schlagzeilen in der entsprechenden Mediengattung. Die Ansicht in tabellarischer Form wird zusätzlich in Abhängigkeit vom Abfrageort Stuttgart und Frankfurt ausgegeben. Die Mediengattungen werden im Gegensatz zur vorherigen Darstellung in absteigender Reihenfolge gelistet.
Zur erkennen ist, dass die Unterscheidung der Abfragestandorts nahezu keine unterschiedlichen Ergebnisse liefert. In den meisten Fällen bleibt die Platzierung der Quellen gleich.

Top 3 je Mediengattung (Gesamt)

Platz	Gattung	Quelle	Anzahl
1	Zeitung	WELT	17.075

2	Zeitung	FAZ.NET	16.000
3	Zeitung	Süddeutsche Zeitung	14.180
1	Rundfunk	Tagesschau	17.107
2	Rundfunk	Ntv	9.674
3	Rundfunk	SWR	9.360
1	Online	T-Online	15.440
2	Online	ka-news	9.883
3	Online	Heise	1.628
1	Zeitschrift	Spiegel	21.102
2	Zeitschrift	FOCUS Online	4.237
3	Zeitschrift	Stern	2.081
1	Red./Agentur	RND	7.080
2	Red./Agentur	Presseportal	4.258
3	Red./Agentur	Correctiv	337
1	regÖA	Baden-Württemberg.de	1.809
2	regÖA	Ministerium für Kultus, Jugend und Sport Baden-Württemberg	510
3	regÖA	Innenministerium Baden-Württemberg	484
1	Corporate Publishing	AfD.de	913
2	Corporate Publishing	VfB Stuttgart	774
3	Corporate Publishing	Katholisch.de	309

Tabelle 6 Top 3 je Mediengattung (Gesamt)

Top 3 je Mediengattung (Stuttgart)

Platz	Gattung	Quelle	Anzahl
1	Zeitung	WELT	8.539
2	Zeitung	FAZ.NET	7.908
3	Zeitung	Süddeutsche Zeitung	7.050
1	Rundfunk	Tagesschau	8.515
2	Rundfunk	Ntv	4.636
3	Rundfunk	SWR	4.404
1	Online	T-Online	7.637
2	Online	ka-news	4.800
3	Online	Heise	739
1	Zeitschrift	Spiegel	10.566
2	Zeitschrift	FOCUS Online	2.120
3	Zeitschrift	Stern	1.021
1	Red./Agentur	RND	3.539
2	Red./Agentur	Presseportal	2.056
3	Red./Agentur	Reuters	170
1	regÖA	Baden-Württemberg.de	1.502

2	regÖA	Innenministerium Baden-Württemberg	425
3	regÖA	Ministerium für Finanzen Baden-Württemberg	384
1	Corporate Publishing	AfD.de	451
2	Corporate Publishing	VfB Stuttgart	379
3	Corporate Publishing	Daimler Media	168

Tabelle 7 Top 3 je Mediengattung (Stuttgart)

Top 3 je Mediengattung (Frankfurt)

Platz	Gattung	Quelle	Anzahl
1	Zeitung	WELT	8.536
2	Zeitung	FAZ.NET	8.092
3	Zeitung	Süddeutsche Zeitung	7.130
1	Rundfunk	Tagesschau	8.592
2	Rundfunk	Ntv	5.038
3	Rundfunk	SWR	4.956
1	Online	T-Online	7.803
2	Online	ka-news	5.083
3	Online	Heise	889
1	Zeitschrift	Spiegel	10.536
2	Zeitschrift	FOCUS Online	2.117
3	Zeitschrift	Der Aktionär	1.129
1	Red./Agentur	RND	3.541
2	Red./Agentur	Presseportal	2.202
3	Red./Agentur	Correctiv	174
1	regÖA	Baden-Württemberg.de	307
2	regÖA	Ministerium für Kultus, Jugend und Sport Baden-Württemberg	291
3	regÖA	Stadt Stuttgart	144
1	Corporate Publishing	AfD.de	462
2	Corporate Publishing	VfB Stuttgart	395
3	Corporate Publishing	Katholisch.de	158

Tabelle 8 Top 3 je Mediengattung (Frankfurt)

Die nachfolgende Lorenzkurve wurde für die Verteilung der Schlagzeilen in Abhängigkeit von den 7 Hauptmediengattungen berechnet (wie sie bei Abbildung 23 bereits dargestellt wurden). Im Vergleich zu den Lorenzkurven bei Betrachtung der Quellen (Abschnitt 2.1) ergibt sich ein deutlich niedrigerer Grad an Ungleichverteilung, der auch durch den niedrigeren Gini-Koeffizienten von 0,563 gekennzeichnet ist. Dennoch ist auch dieser Wert weiterhin ein deutlicher Indikator für eine weiterhin hohe Ungleichverteilung.

N = 7
Gini-Koeffizient = 0,563

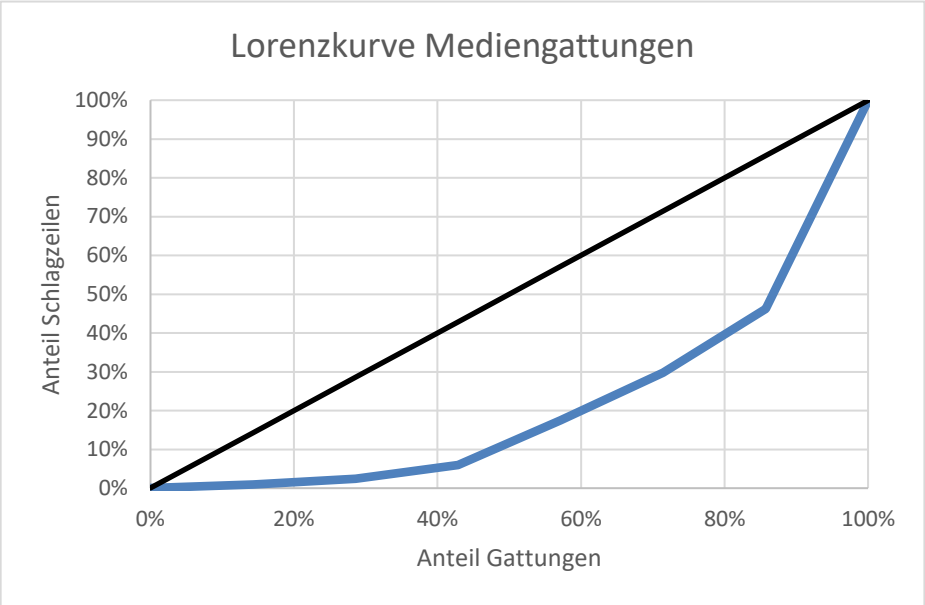


Abbildung 24: Lorenzkurve und Gini-Koeffizient für die Verteilung „Anzahl der Schlagzeilen je Mediengattung"

2.2.2 Anzahl/Häufigkeit Schlagzeilen für Mediengattungen in Bezug auf Suchbegriffe (#23)

AUFGABENSTELLUNG DER FRAGE:

Im Rahmen dieser Frage wird die Häufigkeitsverteilung von Schlagzeilen pro Mediengattung in Bezug auf die einzelnen Suchbegriffe untersucht.

ABFRAGETECHNIK

Für alle Suchterme wird der Anteil an Schlagzeilen für alle Mediengattungen berechnet und in einem gestapelten Balkendiagrammen dargestellt. Ein zweites Balkendiagramm, skaliert auf 100% für jeden Balken zeigt die Anteile zu jeder Gattung. Die Reihenfolge der Suchterme wird in absteigender Reihenfolge der Anzahl Schlagzeilen (siehe auch Tabelle 6) dargestellt.

ERGEBNIS DER ANALYSE:

Wie zuvor bereits erkannt, weist die Mediengattung „Zeitung" die größten Anteile auf. Bei den Suchbegriffen Karlsruhe und Bosch ist der Anteil der Gattung Online jedoch ähnlich groß wie zur Gattung Zeitung. Lediglich für den Begriff Baden-Württemberg zeigt sich ein nennenswerter Anteil aus der Gattung regÖA. Für die Begriffe Bernd Riexinger, Franziska Brantner und Michael Theurer werden in allen Gruppen nur sehr wenige Anteile (indiziert auf 100%) gezeigt.
Bei den übrigen Gattungen fallen folgende Besonderheiten auf

- Namen der Kanzlerkandidaten: weitgehend vergleichbare Verteilung der Präsenz in unterschiedlichen Mediengattungen der einzelnen Suchbegriffe

- Wolfgang Schäuble: prozentual hoher Anteil an Gattung „Zeitung" im Vergleich zu anderen Politikern
- Bei Bernd Riexinger stammt der deutlich überwiegende Teil der (insgesamt im Vergleich zu den anderen Suchbegriffen geringen Anzahl an) Schlagzeilen aus der Gattung Rundfunk.

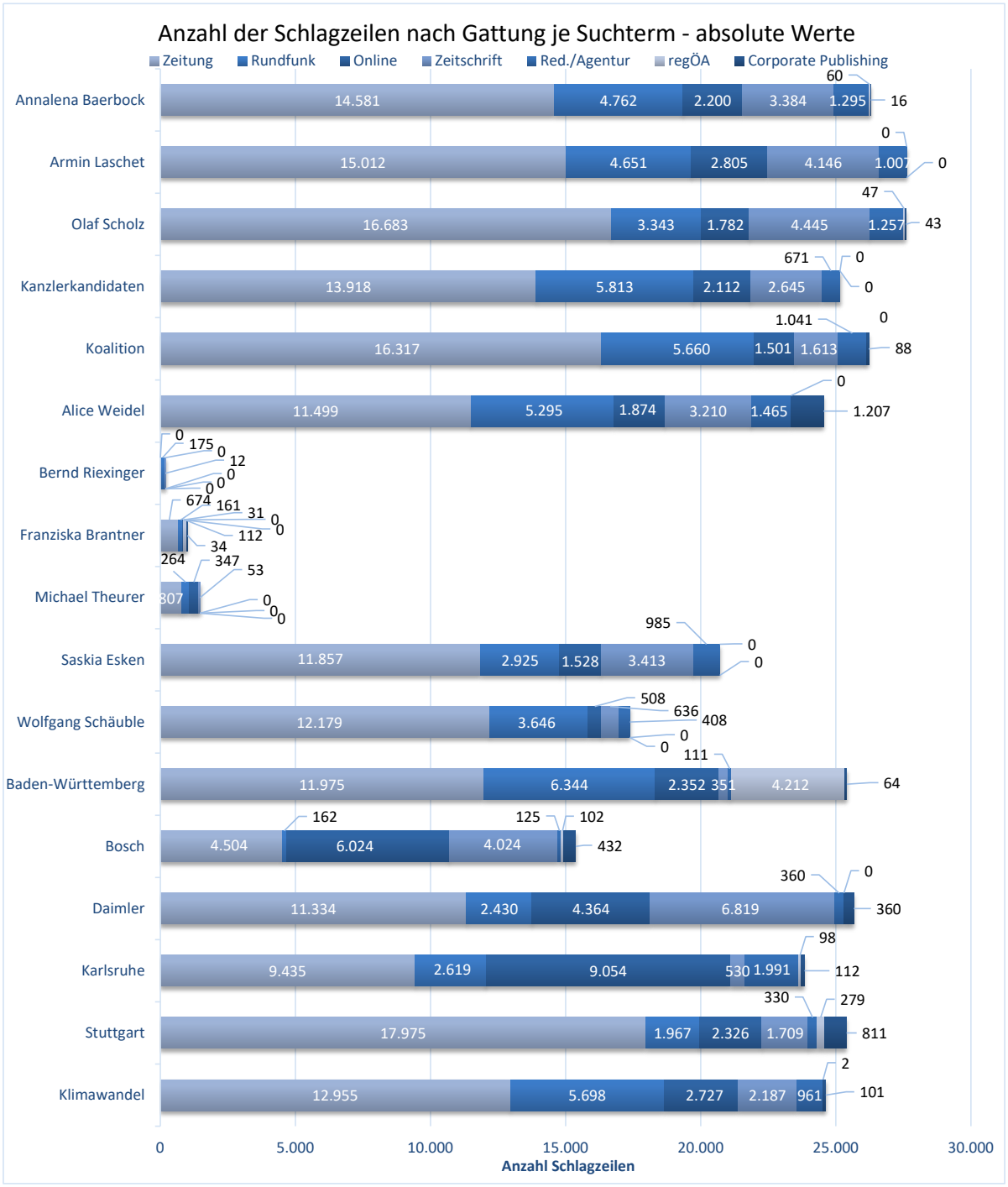


Abbildung 25: Anzahl der Schlagzeilen nach Gattung je Suchterm

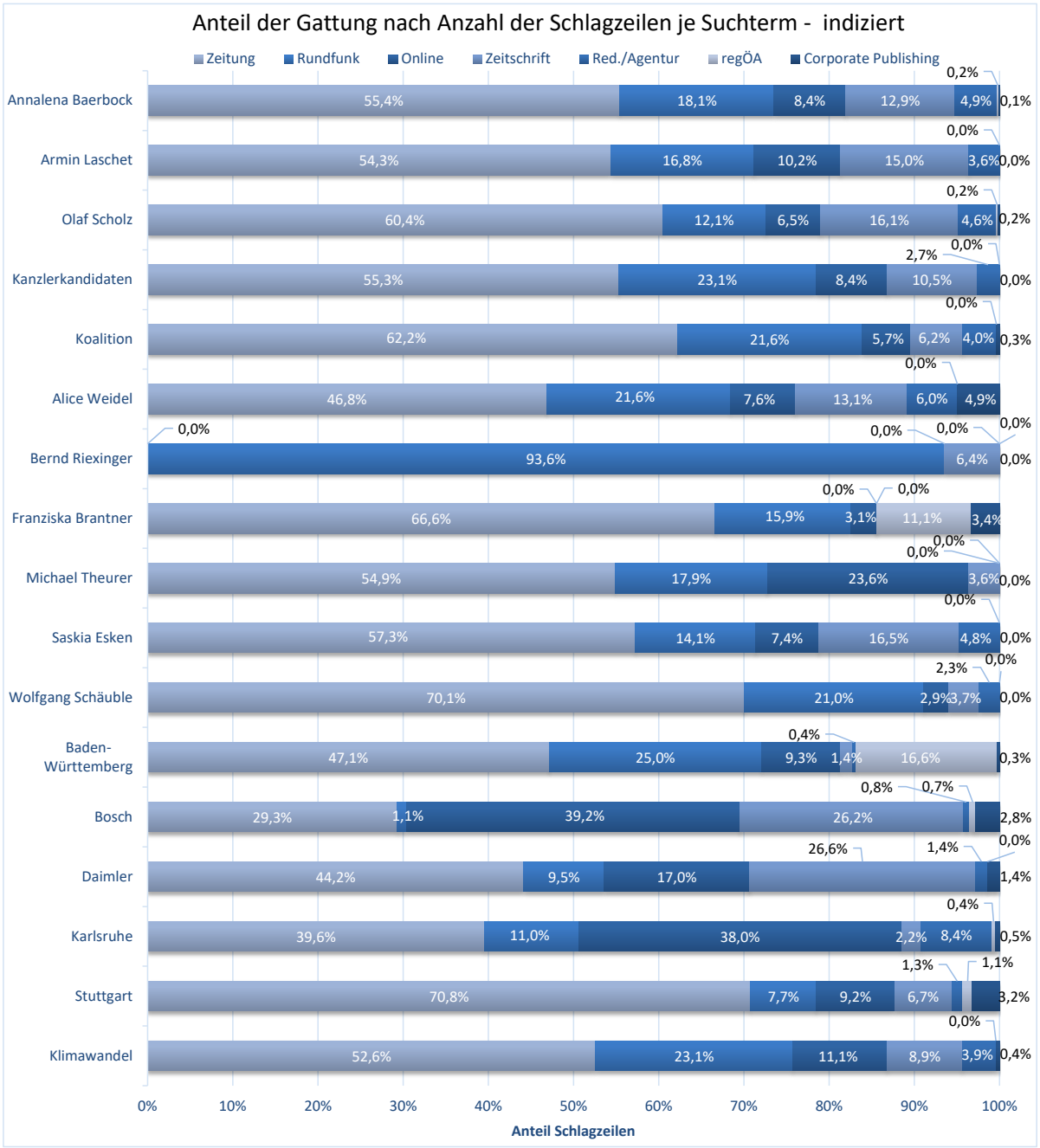


Abbildung 26: Anteil der Gattung nach Anzahl der Schlagzeilen je Suchterm - indiziert

2.3 Suchbegriffe

2.3.1 Anzahl/Häufigkeit Schlagzeilen (#1)

AUFGABENSTELLUNG DER FRAGE:

Gegenstand der Frage aus diesem Abschnitt ist die Untersuchung der Anzahl bzw. des Anteils der Schlagzeilen in Abhängigkeit vom jeweiligen Suchbegriff unter Berücksichtigung einer möglicherweise bestehenden Standortabhängigkeit.

ABFRAGETECHNIK

Die Anzahl der Schlagzeilen wurde nach den Suchtermen gruppiert und prozentuale Anteile gebildet. Die Abfrage wurde für alle Schlagzeilen und solchen aus den Abfragestandorten Frankfurt und Stuttgart durchgeführt.

ERGEBNIS DER ANALYSE:

Insgesamt sind die Schlagzeilen über alle Suchterme und dabei insbesondere für die drei Kanzlerkandidaten relativ gleich verteilt. Lediglich für die Suchbegriffe Michael Theurer, Franziska Brantner und Bernd Riexinger fallen die Anteile signifikant ab.
Bei dem Vergleich über die Abfragestandorte Gesamt, Stuttgart und Frankfurt können nahezu keine Unterschiede festgestellt werden.

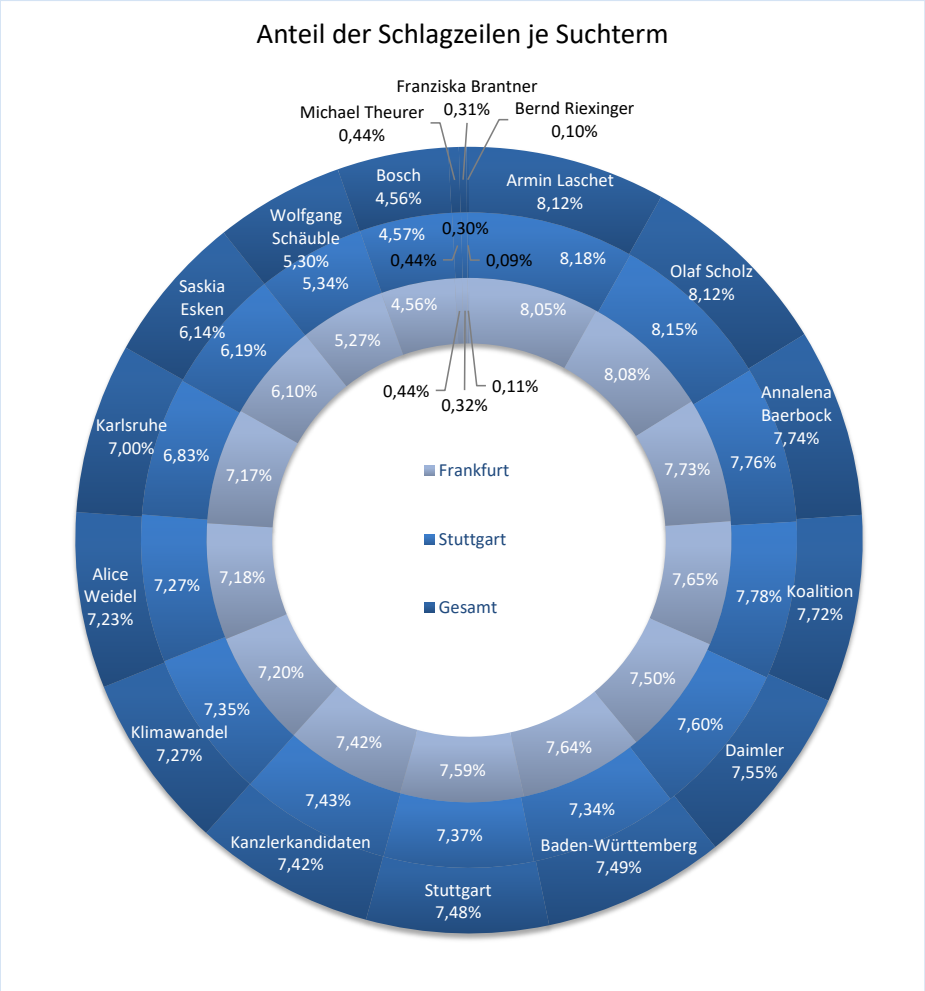


Abbildung 27: relativer Anteil Schlagzeilen pro Suchbegriff für unterschiedliche Abfragestandorte

2.3.2 Sentimente in Abhängigkeit vom Suchbegriff (#7)

AUFGABENSTELLUNG DER FRAGE:

In diesem Abschnitt wird der Frage nachgegangen, ob die Suchbegriffe mit jeweils spezifischen Sentimenten, d.h. Stimmungen in Verbindung gebracht werden können.

ABFRAGETECHNIK

Für alle Schlagzeilen des Abfragezeitraums wurden die Sentimente der Schlagzeilentexte bestimmt. Diese Sentimente werden zur Beantwortung der Frage nach den Suchbegriffen gruppiert und in separaten Betrachtungen zu Boxplot-Darstellung verarbeitet. Pro Suchbegriff wird somit ein eigener Boxplot erzeugt. Die Reihenfolge der Suchterme wird in Form einer thematischen Gruppierung dargestellt. Es wurden alle Sentimente berücksichtigt (auch Analyse einer Schlagzeile mit neutralem Sentiment).

ERGEBNIS DER ANALYSE:

Die meisten der Boxplot-Darstellungen zeigen sehr ähnliche Ergebnisse. Die Sentimente liegen zum großen Teil in einem eher neutralen Bereich, mit minimaler Tendenz zum Positiven bei Werten in der Box zwischen 0 und 0,4. Insbesondere werden stark positive und leicht negative Werte als Ausreißer (als Punkte dargestellt) gewertet, da sie nur in geringen Mengen vorkommen.
Zudem lässt sich feststellen, dass bei den Kanzlerkandidaten die Sentimente nahezu identisch sind, auch die negativen/positiven Abweichungen sind nahezu gleich.
Der Suchbegriff Michael Theurer sticht etwas heraus, da hierfür keine negativen Sentimente berechnet wurden und der obere Grenzwert des Kastens bis 0,5 reicht. Insgesamt gibt es dafür aber auch sehr wenige Datensätze zu diesem Begriff. Die Schlagzeilen sind also nicht übermäßig positiv oder negativ, sondern hauptsächlich neutral gehalten. Die methodischen Grenzen, sowie die Limitierung durch die kurzen Schlagzeilentexte, der Sentimentanalyse wurden im Rahmen der Begriffsdefinition bereits erläutert.

SUCHBEGRIFFE MIT BUNDESTAGSWAHLBEZUG

- Annalena Baerbock
- Armin Laschet
- Olaf Scholz
- Kanzlerkandidaten
- Koalition

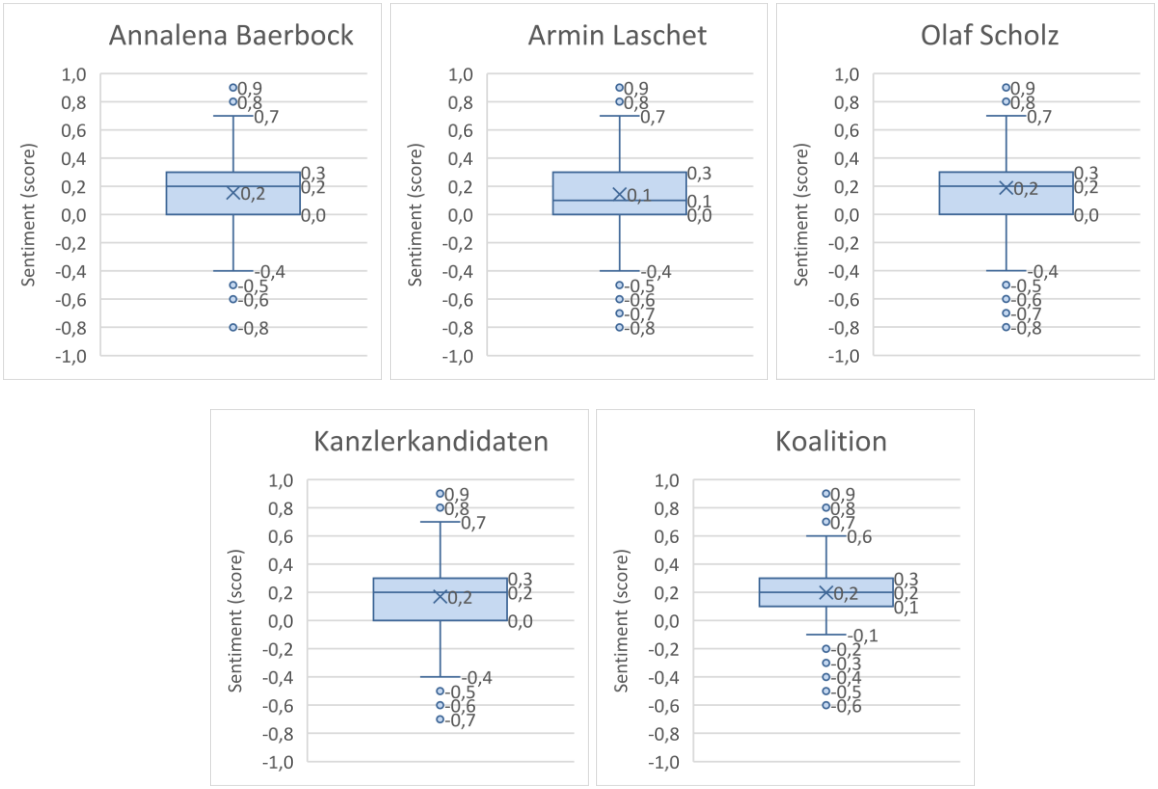


Abbildung 28 Sentimente - Suchbegriffe mit Bundestagswahlbezug

SUCHBEGRIFFE: SPITZENKANDIDATEN DER PARTEIEN IN BADEN-WÜRTTEMBERG FÜR DIE BUNDESTAGSWAHL

- Alice Weidel
- Bernd Riexinger
- Franziska Brantner
- Michael Theurer
- Saskia Esken
- Wolfgang Schäuble

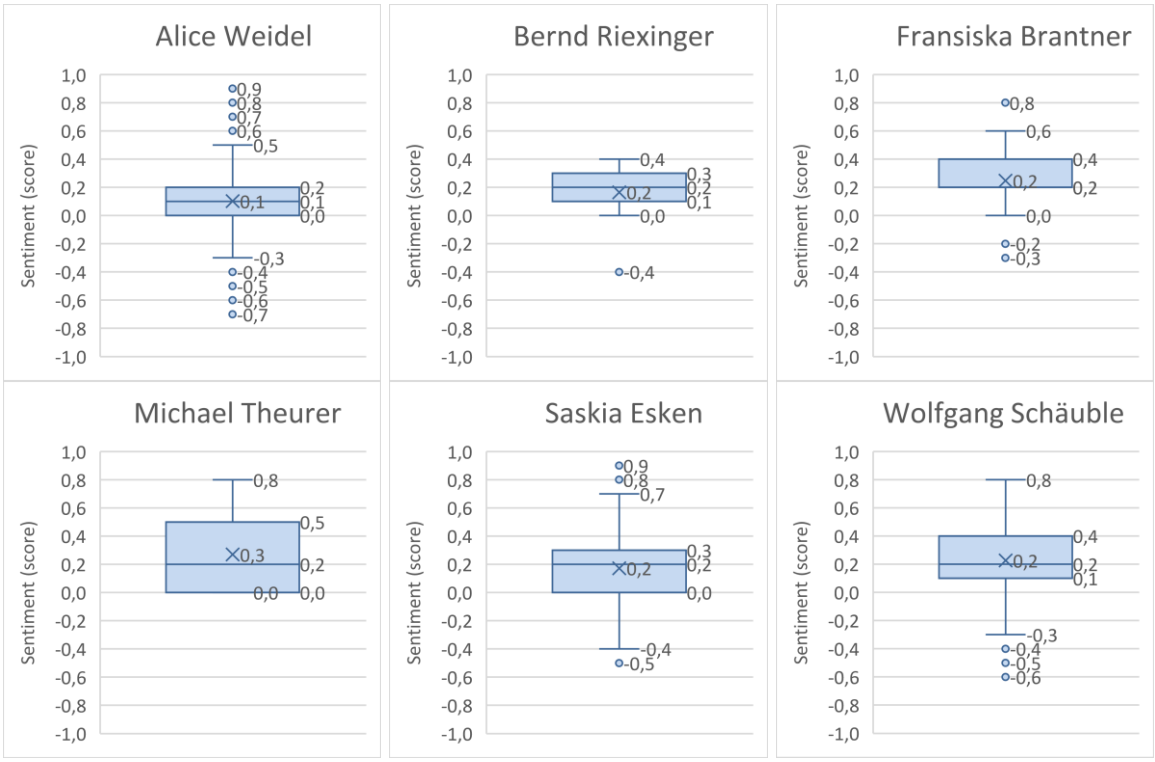


Abbildung 29 Sentimente – Suchbegriffe: Spitzenkandidaten für BaWü

SUCHBEGRIFFE MIT BEZUG ZU BADEN-WÜRTTEMBERG

- Baden-Württemberg
- Bosch
- Daimler
- Karlsruhe
- Stuttgart

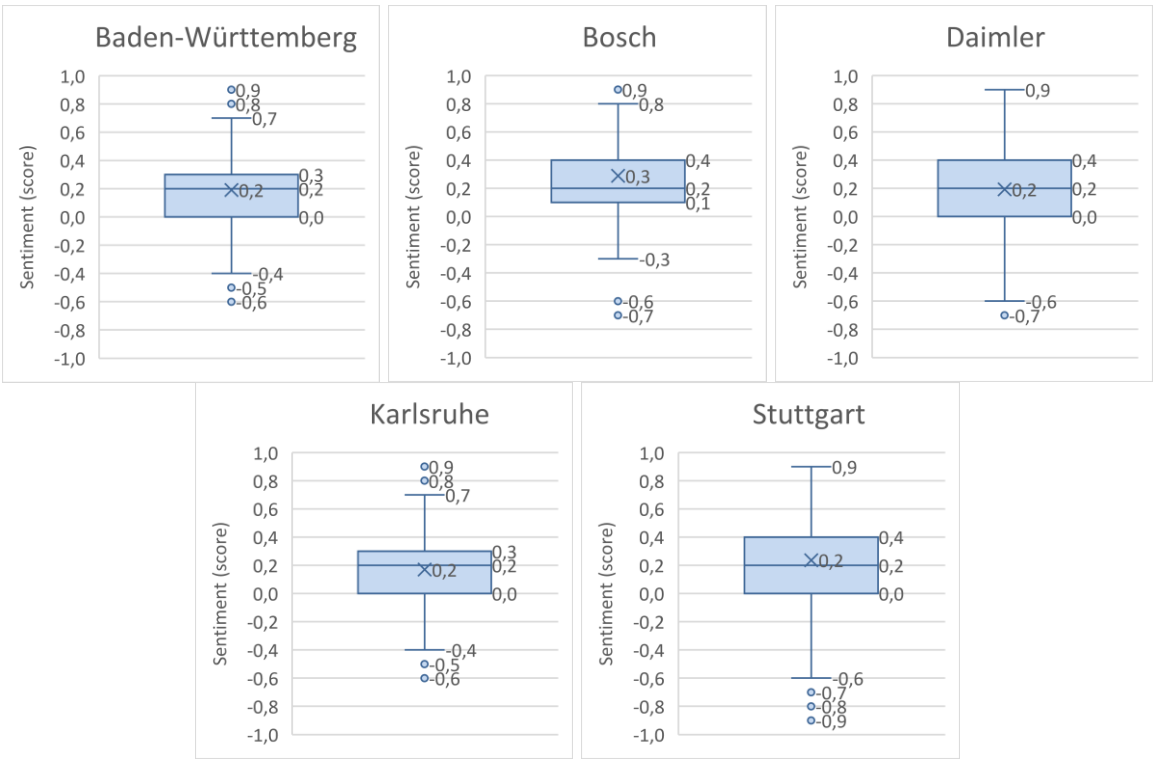


Abbildung 30 Sentimente - Suchbegriffe mit Bezug zu BaWü

SUCHBEGRIFF MIT ÜBERGEORDNETEN BEZUG

- Klimawandel

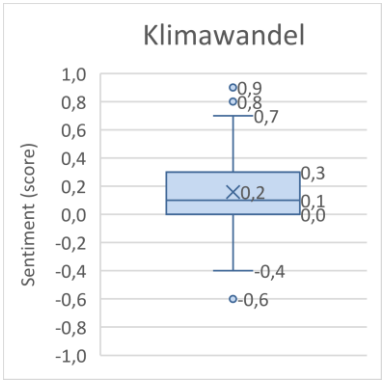


Abbildung 31 Sentimente - Suchbegriffe mit übergeordnetem Bezug

2.3.3 Sentiment im zeitlichen Verlauf (#10)

AUFGABENSTELLUNG DER FRAGE:

In Ergänzung zur Fragestellung aus Abschnitt 2.3.2 wird in diesem Abschnitt die zeitliche Abhängigkeit der für die Suchbegriffe ermittelten Sentimente betrachtet.

ABFRAGETECHNIK

Im Rahmen der Sentimentanalyse wurde bereits für alle Titel der Schlagzeilen die Sentimente berechnet. Die Schlagzeilen wurden bei dieser Fragestellung nach den Suchbegriffen gruppiert und pro Tag des Betrachtungszeitraums ein Mittelwert über den Score der Sentimente gebildet. Die so entstehenden Daten werden als Linie über den Verlauf der Tage des Zeitraums abgebildet.

ERGEBNIS DER ANALYSE:

Die Suchbegriffe werden neben einer Gesamtdarstellung auch in drei Diagrammen aufgeteilt gemäß der Einteilung nach Bundes-bezogenen Suchbegriffen, Namen der Spitzenkandidaten aus Baden-Württemberg sowie alle übrigen Begriffe (vgl. Einteilung der Suchbegriffe gemäß Kapitel 1.4.5.)

Aus der Betrachtung der Diagramme ergeben sich folgende Beobachtungen

- Die Sentimente verlaufen im leicht positiven Bereich mit (geschätzten) Mittelwert um 0,2 in einer Spanne von theoretisch +1 bis zu -1.
- Deutliche, jedoch scheinbar zufällige Schwankungen der Sentimente im zeitlichen Verlauf sind zu beobachten.
- Insbesondere scheint es keinen pauschalen Bezug zur Bundestagswahl 2021 zu geben und mögliche zeitliche Trends sind ebenfalls nicht erkennbar.

Eine mögliche Interpretation diese Beobachtungen ist, dass die Berichterstattung in Bezug auf Stimmungen zumindest im Schlagzeilentitel sehr ausgeglichen ist. Allerdings sei dabei auch an dieser Stelle nochmals darauf hingewiesen, dass die Schlagzeilenmeldungstexte bei der Berechnung der Sentimente nicht mit einbezogen wurden und eine Bestimmung von Sentimenten für kurze Texte (hier: Schlagzeilentitel) nicht unproblematisch ist. Vor diesem Hintergrund erscheint eine zusätzliche Betrachtung der Schlagzeilenmeldungstexte als ein vielversprechendes Thema für eine weitergehende Untersuchung (siehe Kapitel 1.3 Ausblick)

Ausblick

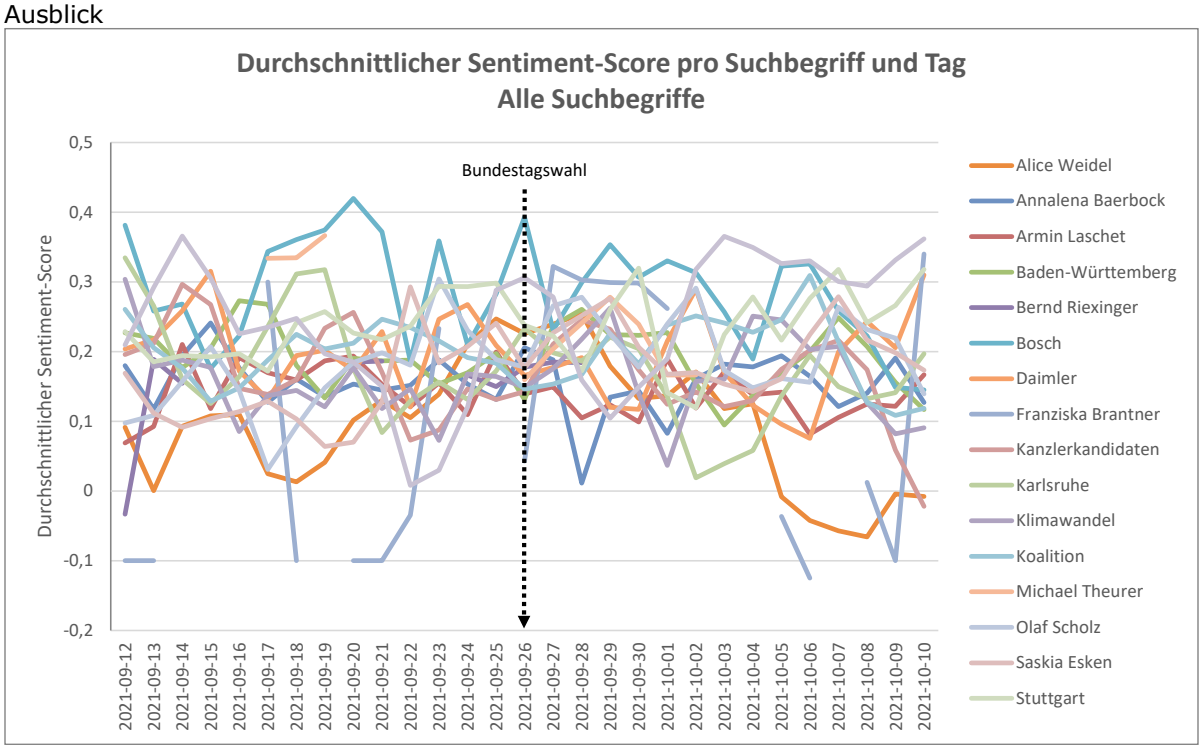


Abbildung 32 Durchschnittlicher Sentiment-Score pro Suchbegriff und Tag

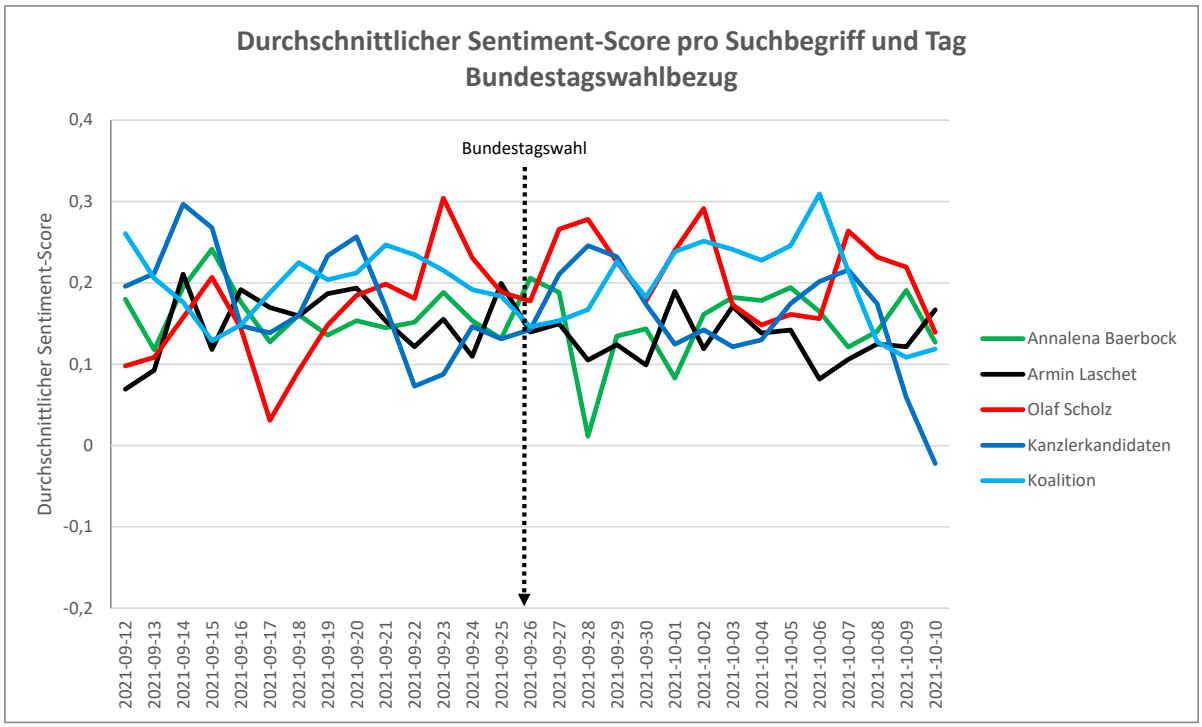


Abbildung 33 Durchschnittlicher Sentiment-Score - Bundestagswahlbezug

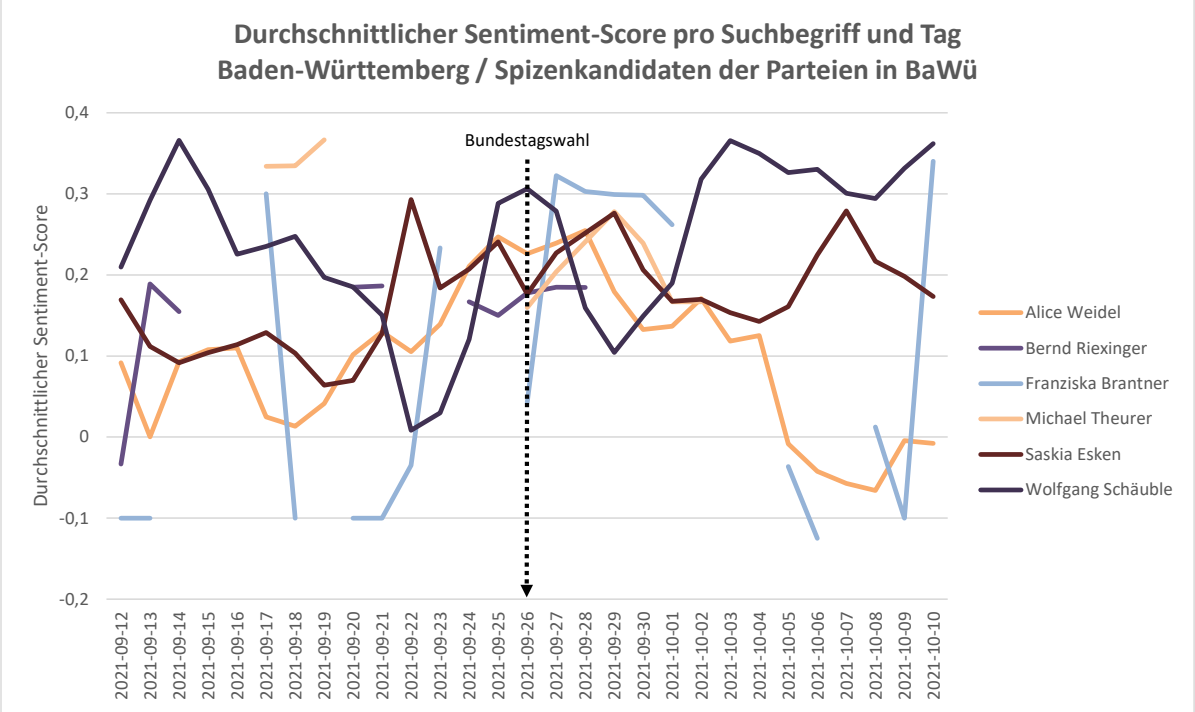


Abbildung 34 Durchschnittlicher Sentiment-Score - Spitzenkandidaten der Parteien in BaWü

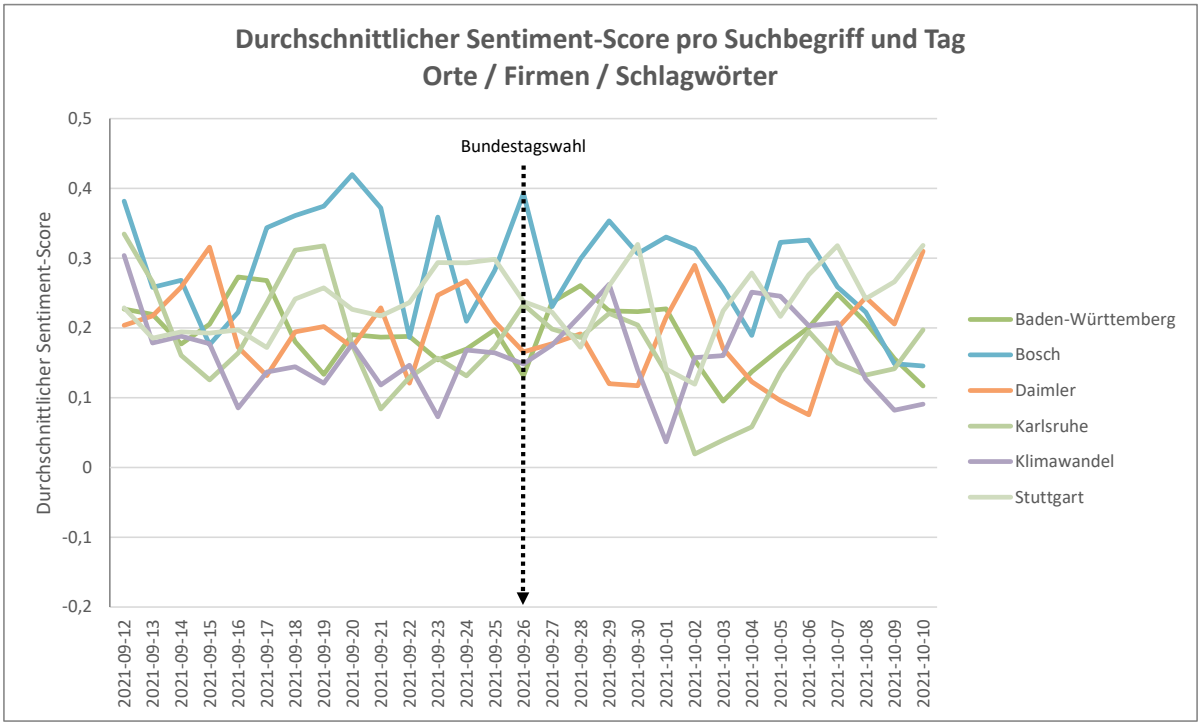


Abbildung 35 Durchschnittlicher Sentiment-Score - Orte / Firmen / Schlagwörter

2.3.4 In welchen Quellen kommen best. Suchbegriffe häufig vor? (#3)

AUFGABENSTELLUNG DER FRAGE:

Abschnitt 2.3.4 beschäftigt sich mit der Frage, ob bestimmte Suchbegriffe bevorzugt in bestimmten Quellen zu Schlagzeilen führen, d.h. wie hoch der der Anteil der Schlagzeilen in Abhängigkeit von Quelle und Suchbegriff ist.

ABFRAGETECHNIK

Bei dieser Abfrage musste für Suchbegriffe und Quellen ermittelt werden, wie viele Schlagzeilen vorliegen.
Da dies aufgrund der hohen Anzahl an Quellen zu sehr vielen Kombinationsmöglichkeiten führt, erfolgte die Untersuchung unter Berücksichtigung der 20 Kombinationen aus Quelle und Suchbegriff mit den meisten Schlagzeilen. Aus Grundlage der identifizierten Kombinationen wurden alle in den Kombinationen enthaltenen Quellen als auch alle enthaltenen Suchbegriffe separat entnommen.
Beispielsweise ergaben sich die folgenden drei Kombination aus Quelle und Suchbegriff mit den meisten Schlagzeilen:

Quelle	Suchbegriff
ka-news	Karlsruhe
Stuttgarter Zeitung	Stuttgart
Stuttgarter Nachrichten	Stuttgart

Tabelle 9 Top3 Kombinationen Quelle und Suchbegriff (Beispiel)

Aus dem vorherigen Beispiel separat entnommen würden sich als Beispiel folgende Quellen und Suchbegriffe ergeben:

Separat entnommene Quellen	Separat entnommene Suchbegriffe
ka-news	Karlsruhe
Stuttgarter Zeitung	Stuttgart
Stuttgarter Nachrichten	

Tabelle 10 Separat entnommene Quellen und Suchbegriffe (Beispiel)

Für die Darstellung unten wurde nun für alle separat entnommenen Quellen und Suchbegriffe die Anzahl Schlagzeilen gezählt (kartesisches Produkt) und tabellarisch dargestellt. Andere Quellen und Suchbegriffe, die nicht Teil mindestens einer dieser Kombinationen war, sind daher nicht in dieser Abbildung vertreten.
Die tabellarische Darstellung zeigt neben der absoluten Anzahl an Schlagzeilen auch einen Balken in der jeweiligen Zelle mit dem relativen Anteil an den Schlagzeilen.

ERGEBNIS DER ANALYSE:

Folgende Auffälligkeiten sind zu beobachten

- Über die Kanzlerkandidaten (Suchbegriffe Annalena Baerbock, Armin Laschet, Olaf Scholz) wird vor allem im Spiegel berichtet, wobei Armin Laschet deutlich mehr Schlagzeilen verursacht als die anderen beiden Kandidaten.
- Auffällig ist auch das zahlmäßig hohe Auftreten von Schlagzeilen bei Alice Weidel.
- Regionale Quelle wie z.B. ka-news und die Badische Neueste Nachrichten haben sehr hohen Anzahl an Schlagzeilen für den (regionalen) Begriff Karlsruhe
- Die Stuttgarter Nachrichten und die Stuttgarter Zeitung erscheinen im Vergleich hierzu etwas breiter aufgestellt zu sein, wobei bei diesen Quellen signifikant Schlagzeilen zu den Suchbegriffen Baden-Württemberg und Stuttgart auftreten.
- Quellen wie z.B. die Tagesschau, T-Online, der Spiegel oder die Welt berichten über eine breite Vielzahl an Themen (d.h. Suchbegriffen).

Quelle	Annalena Baerbock	Armin Laschet	Olaf Scholz	Kanzler-kandidaten	Alice Weidel	Saskia Esken	Wolfgang Schäuble	Baden-Württemberg	Bosch	Karlsruhe	Stuttgart
Badische Neueste Nachrichten	39		66		2	32	681	595	263	5.300	30
FAZ.NET	1.394	1.658	2.073	516	1.573	2.628	318	229	824	144	159
Münchner Merkur	2.377	1.145	673	1.309	695	1.404	1.102	300	1.235	48	6
RP Online	305	855	429	2.688	148	245	223	96		1	21
SWR	5	6	116	25	242	262	393	4.001		1.890	1.263
Spiegel	2.573	3.739	3.194	1.603	2.890	1.370	332	134	1.184	119	483
Stuttgarter Nachrichten	109	8	103	171	120	267	100	2.150	222	331	5.878
Stuttgarter Zeitung	93	17	95	18	354	23	152	2.095	160	219	7.099
T-Online	1.898	2.460	1.573	996	1.550	251	139	779	2.774	142	919
Tagesschau	1.609	2.248	1.151	2.327	2.068	1.564	2.286	207	1	258	59
WELT	2.242	2.319	2.966	783	2.170	1.720	888	161	6		313
ka-news	17	1	2					767		8.736	2

Tabelle 11 Anzahl und relativer Anteil Schlagzeilen pro Suchbegriff und Quelle

In einer weiteren Betrachtung werden die Anzahlen der Schlagzeilen für die 10 Anzahl-stärksten Quellen (gemessen an der Gesamtzahl aller Schlagzeilen) und die Namen der Kanzlerkandidaten sowie von Alice Weidel als Suchbegriff dargestellt. Der Übersichtlichkeit halber wurden die Werte als Linien dargestellt. Dabei fällt auf, dass

- der Spiegel, die Süddeutsche Zeitung, WELT und FAZ.NET eine relativ ausgeglichene Anzahl an Schlagzeilen für alle 4 PolitikerInnen liefern, wobei durchaus Ausreißer erkennbar sind.
- die Stuttgarter Nachrichten und Stuttgarter Zeitung, im Verhältnis zu den anderen Quellen, kaum Schlagzeilen zu den 4 PolitikerInnen liefern – oder sie sich nicht in den vorderen Rängen platzieren konnten.

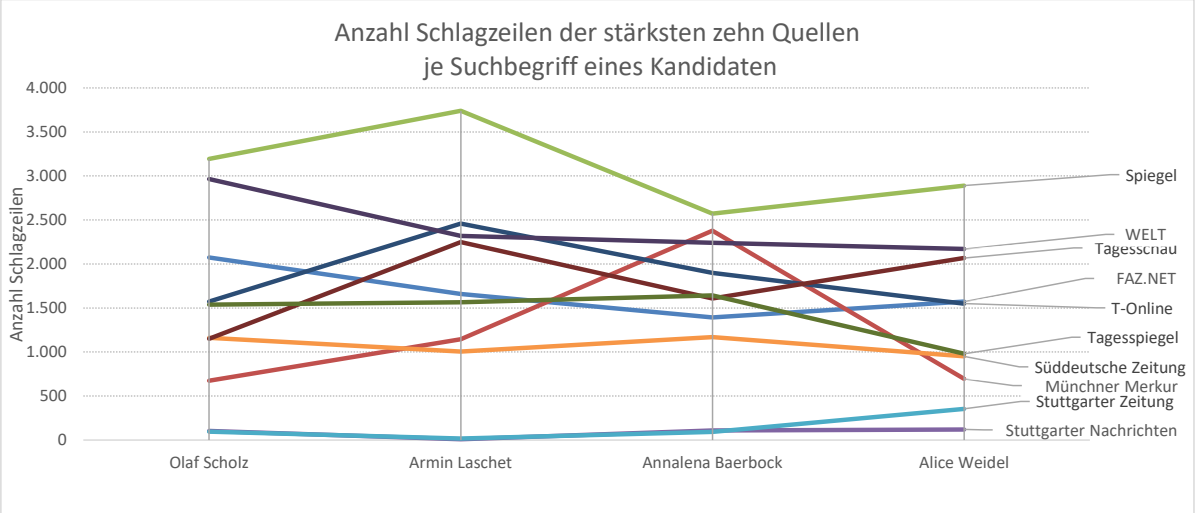


Abbildung 36: Anzahl Schlagzeilen pro Suchbegriff für ausgewählte Quellen

2.4 Schlagzeilen

2.4.1 Aktualität/Verteilung Alter (#16)

AUFGABENSTELLUNG DER FRAGE:

In der vorliegenden Frage wird die Aktualität der Schlagzeilen (d.h. das Alter der Veröffentlichung (siehe auch Begriffsdefinition Veröffentlichungsdatum) untersucht, wobei insbesondere der Blick auf den zeitlichen Verlauf des 4-wöchentlichen Betrachtungszeitraumes gerichtet wird.

ABFRAGETECHNIK

Gemessen wurde das Alter einer Schlagzeile, indem die Differenz zwischen dem Zeitpunkt des Abfragens der Schlagzeile und dem Veröffentlichungsdatum der betreffenden Schlagzeile berechnet wurde. Für alle Schlagzeilen, gruppiert nach den einzelnen Tagen des Betrachtungszeitraumes, wurde ein Durchschnittswert des Alters gebildet.

ERGEBNIS DER ANALYSE:

Das Bereichsdiagramm in Abbildung 37 zeigt das durchschnittliche Alter der Schlagzeilen an den Tagen des Beobachtungszeitraums
Besonders gut zu erkennen ist, dass die Aktualität der Nachrichten unmittelbar nach der Bundestagswahl deutlich zunimmt, was dadurch begründet sein könnte, dass viele neue Nachrichten veröffentlicht werden. In den zwei Wochen vor der Wahl war die Aktualität ungefähr gleichbleibend mit einem mittleren Alter von ca. 25 Stunden. Einige Tage nach der Wahl (ab ca. 30.9.) steigt das Durchschnittsalter jedoch rapide an, sodass es im Vergleich zu der Zeit von vor der Wahl beinahe sogar doppelt so hoch liegt. Zum Ende des Betrachtungszeitraumes kommt es noch einmal zu einem weiteren deutlichen Anstieg des Alters der Schlagzeilen.

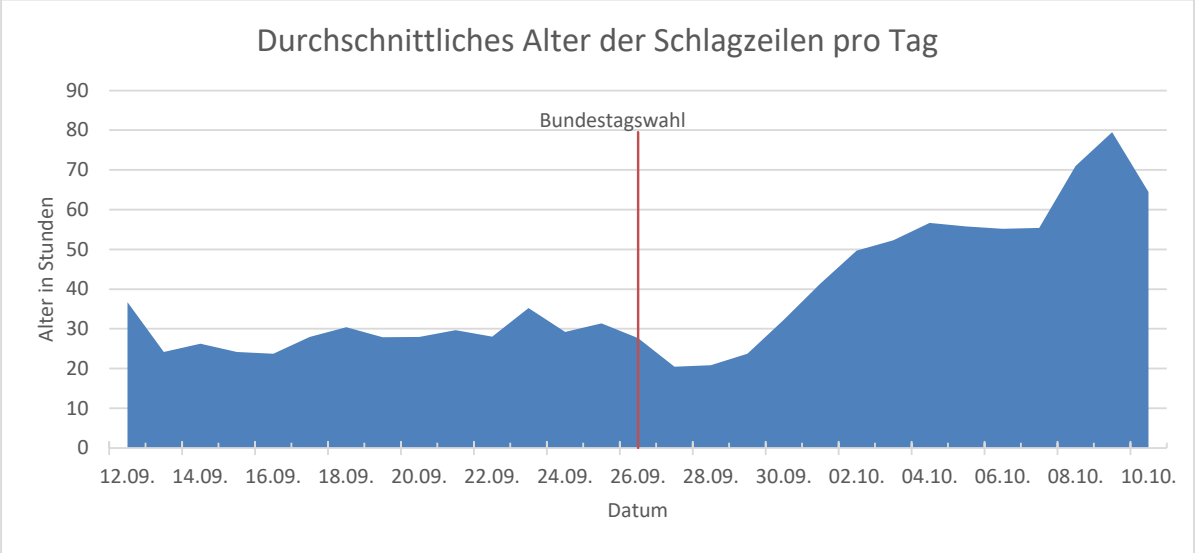


Abbildung 37: Zeitlicher Verlauf des durchschnittlichen Alters der Schlagzeilen

Im Vergleich dazu soll folgend noch eine erneute Darstellung, jedoch nur mit Schlagzeilen mit den Suchbegriffen „Kanzlerkandidaten", „Koalition", „Armin Laschet", „Olaf Scholz" und „Annalena Baerbock", abgebildet werden. Über den gesamten Zeitraum hinweg kann eine sehr ähnliche Ausprägung des durchschnittlichen Alters erkannt werden. Es zeigt sich, bis auf eine kleine Ausnahme, kein andersartiger Trend als wie zuvor bereits beschrieben. Auch hier steigt die Aktualität der Schlagzeilen kurz nach der Bundestagswahl und nimmt im weiteren Verlauf weiter ab. Lediglich ab dem 04.10. zeigt sich hier eine erneute Senkung des durchschnittlichen Alters der Schlagzeilen. In der vorherigen Grafik war hier lediglich ein Abflachen zu erkennen. Zum Ende des Zeitraums hin schießt das Alter, wie in der Darstellung zuvor, stark in die Höhe.

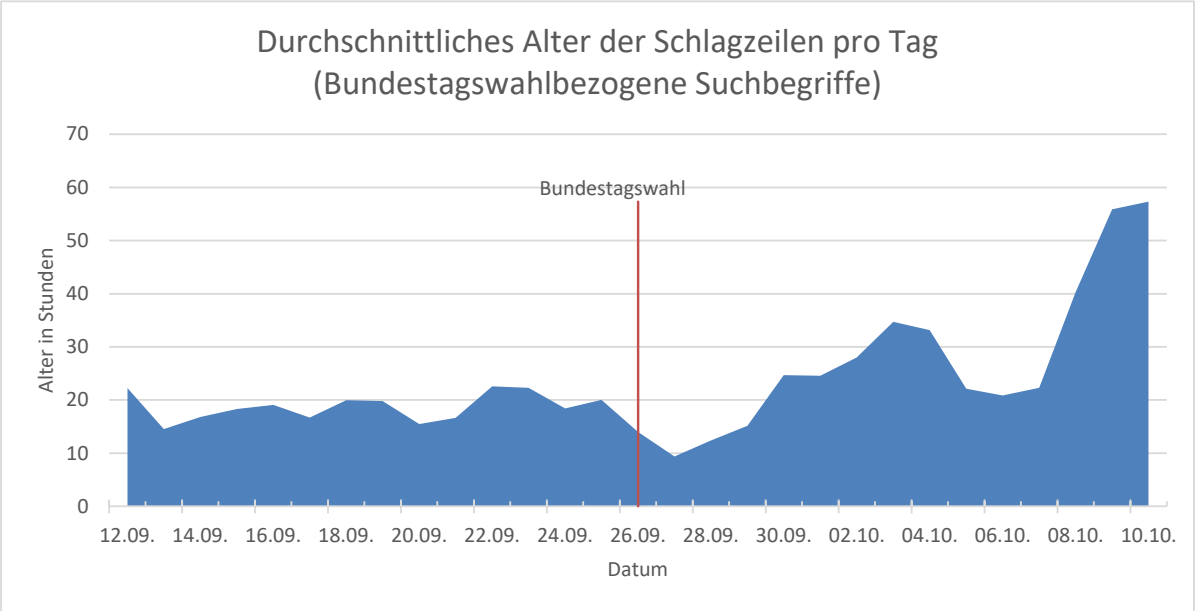


Abbildung 38 Zeitlicher Verlauf des durchschnittlichen Alters der Schlagzeilen (Bundestagswahlbezogene Suchbegriffe)

Mit dem folgenden Boxplot kann die Verteilung des Durchschnittalters der Schlagzeilen (Gesamte Betrachtung ohne Einschränkung auf Kanzlerkandidaten) noch einmal separat betrachtet werden. Es zeigt den Minimum- und Maximal-Wert an den Enden der Linien sowie den Löwenanteil der

Schlagzeilen in der Box dazwischen. Die meisten Schlagzeilen waren für einen Tag betrachtet also zwischen 27 und 53,7 Stunden alt. An einem Tag waren die Schlagzeilen im Schnitt aber nie jünger als 20,4 Stunden oder älter als 79,5 Stunden.

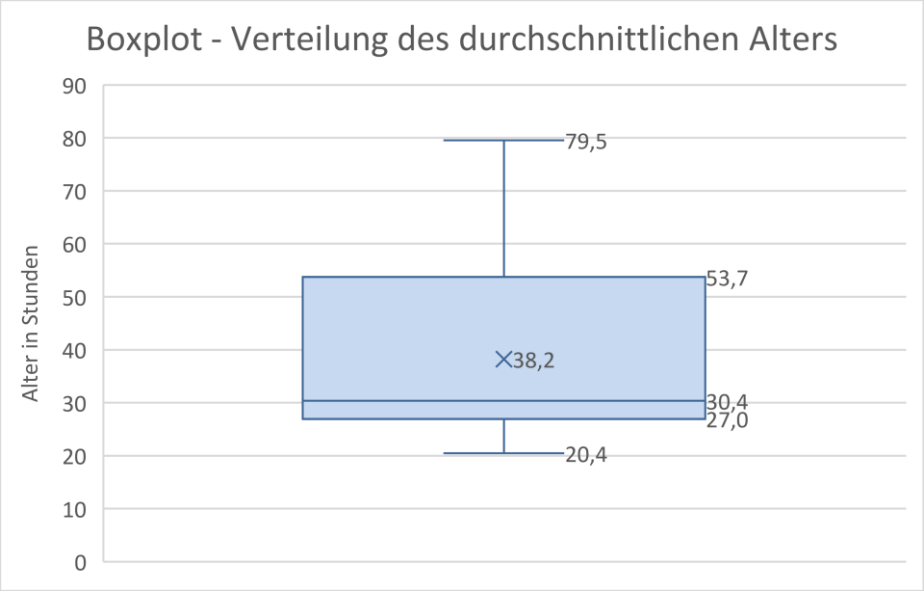


Abbildung 39: Zeitlicher Verlauf des durchschnittlichen Alters der Schlagzeilen

2.4.2 Langlebigkeit/Verweildauer

2.4.2.1 Verweildauer abhängig von der Quelle? (#17a)

AUFGABENSTELLUNG DER FRAGE:

Es soll geprüft werden, wie lange (in Stunden) eine Schlagzeile wiederholt auftritt, d.h., wie lange sich die jeweilige Schlagzeile in dem Schlagzeilenfenster „halten" kann.

ABFRAGETECHNIK

Für das Auftreten einer Schlagzeile wird der Zeitpunkt des Ladens ermittelt. Das Lesen von Schlagzeilen wurde stündlich ausgeführt. Für jedes Finden einer Schlagzeile während einer dieser Abfragen wird angenommen, dass die Schlagzeile genau eine Stunde lang verweilte. Zusammengerechnet über alle Vorkommen, ergibt sich eine Verweildauer. Tritt sie beispielsweise zwei Mal am Vormittag und einmal am Nachmittag auf, ergibt das eine gesamte Verweildauer von drei Stunden. Über mehrere Tage hinweg kann sich dieser Wert entsprechend kumulieren.
Gebildet wurden daraufhin Durchschnittswerte über alle Schlagzeilen, gruppiert nach einer Quelle oder einer Gattung. Die Ergebnisse werden in absteigender Reihenfolge nach dem durchschnittlichen Rang dargestellt.

ERGEBNIS DER ANALYSE:

Für eine Bildung eines Durchschnittswertes über eine Gruppierung von Quellen werden nur die stärksten zehn Quellen (gemessen an ihrem Anteil an Schlagzeilen an der betrachteten Gesamtdatenmenge) dargestellt. In dem folgenden Liniendiagramm zeigt die blaue Linie die durchschnittliche Verweildauer für Schlagzeilen der Ränge 1-3 und die rote Linie die Schlagzeilen der Ränge 1-10.
Zu sehen ist, dass die Tagesschau im Durchschnitt Schlagzeilen mit der größten Verweildauer herausbringt. Über die weiteren Quellen nimmt dieser Wert stetig ab. Auch die Diskrepanz zwischen der Betrachtung der Ränge 1-10 sowie 1-3 schrumpft dabei deutlich. Das bedeutet, dass sich Schlagzeilen der stärksten Quellen insgesamt weniger in den Rängen 1-3 halten können und in

untere Ränge abrutschen. An dieser Stelle sei auf Frage #19 hingewiesen, bei welcher genau dieser Verlauf für eine exemplarische Auswahl an konkreten Schlagzeilen dargestellt wird.

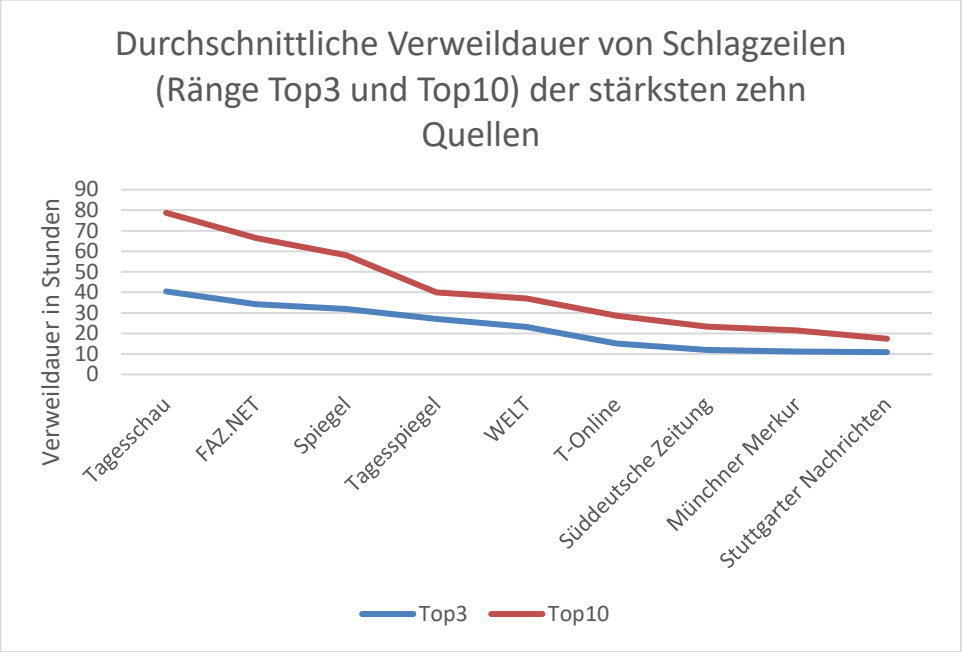


Abbildung 40: Durchschnittliche Verweildauer von Schlagzeilen

Bei der Betrachtung der Gattung liegen die Daten weit weniger unterschiedlich stark ausgeprägt vor. Allerdings kann man erkennen, dass die Gattung Zeitschrift die größten Verweildauern hervorbringt, während die Gattung Zeitung schnelllebiger zu sein scheint. Die Differenz zwischen der Betrachtung der Ränge 1-10 und 1-3 ist in etwa gleichbleibend im Verhältnis zueinander. Lediglich die Gattung Corporate Publishing zeigt für die Ränge 1-3 deutlich kürzere Verweildauern für die Ränge 1-3, so dass Schlagzeilen hier früher abrutschen in den Rängen als bei anderen Gattungen.

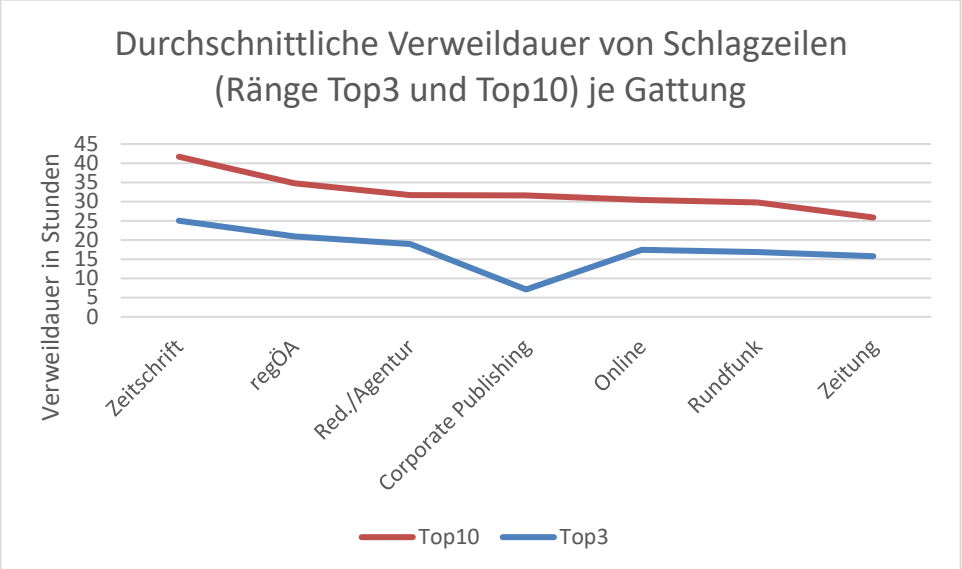


Abbildung 41: Zeitlicher Verlauf des durchschnittlichen Alters der Schlagzeilen

2.4.2.2 Zeitpunkt des Auftretens abhängig von der Quelle (#17b)

AUFGABENSTELLUNG DER FRAGE:

An dieser Stelle soll das durchschnittliche Alter von Schlagzeilen im Kontext von Quellen und Suchtermen betrachtet werden.

Das Alter einer Schlagzeile ergibt sich wie zuvor bereits beschrieben aus der Differenz zwischen dem Zeitpunkt des Abrufens und dem Veröffentlichungsdatum. Für eine Gruppierung nach Quellen werden nur die zehn stärksten Quellen betrachtet. Bei den Suchbegriffen wurde die Darstellung der Übersichtlichkeit in zwei Diagramme aufgeteilt.

ERGEBNIS DER ANALYSE:

Das folgende Diagramm zeigt eine Verlaufslinie des durchschnittlichen Alters der Schlagzeilen gruppiert nach einer der zehn stärksten Quellen.
Man kann gut erkennen, dass einige Linien über den gesamten Zeitraum hinweg flach verlaufen, betroffen sind Stuttgarter Zeitung, Stuttgarter Nachrichten und der Münchener Merkur. Diese Quellen zeigen entsprechend jeden Tag eine eher gleichmäßige Performance (d.h. geringe Verweildauer) in Bezug auf die Aktualität ihrer Schlagzeilen.
Bei den übrigen Quellen sind unterschiedliche starke Ausschläge und deutlich längere Verweildauern zu erkennen. Besonders auffallend ist jedoch, dass direkt nach der Bundestagswahl ein signifikantes Abfallen des Alters der Schlagzeilen zu erkennen ist. Einige Verläufe zeigen zudem eine starke Erhöhung der Aktualität zum Termin der Wahl hin und eine erneute starke Abnahme im Zeitraum nach der Wahl.

Zur besseren Lesbarkeit wurde diese Abbildung zusätzlich in zwei Teilabbildungen dargestellt.

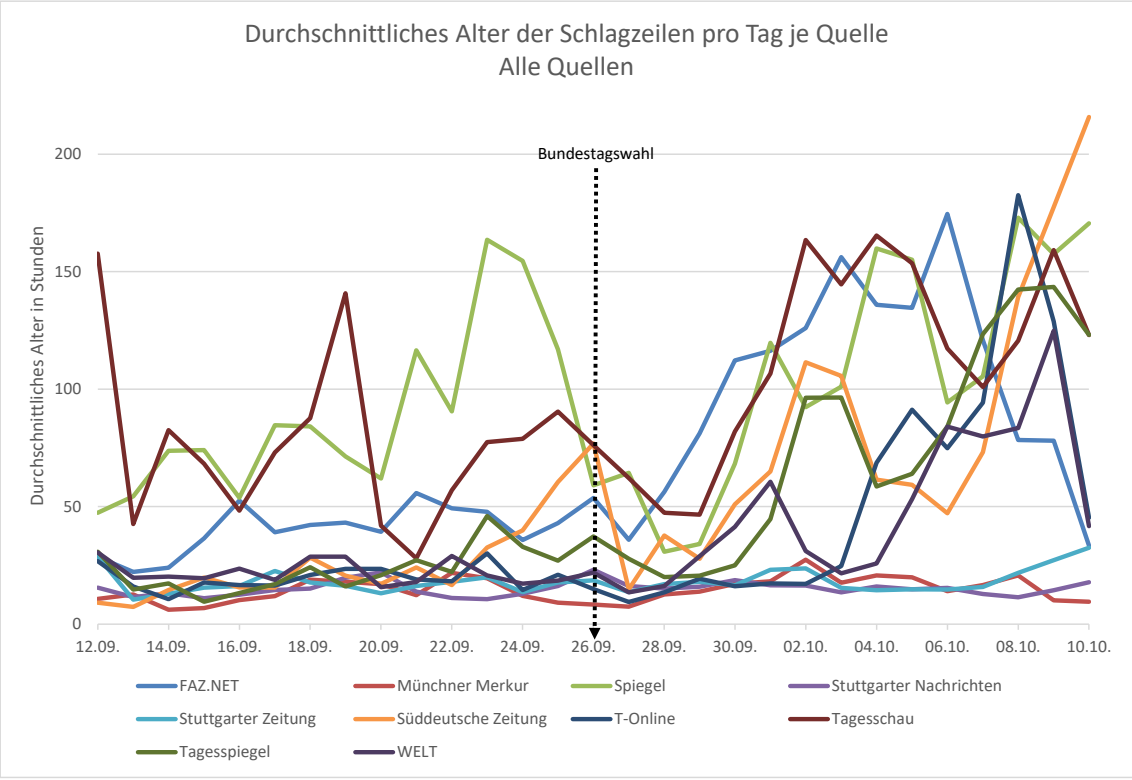


Abbildung 42: Durchschnittliches Alter der Schlagzeile / Quelle – Alle Quellen

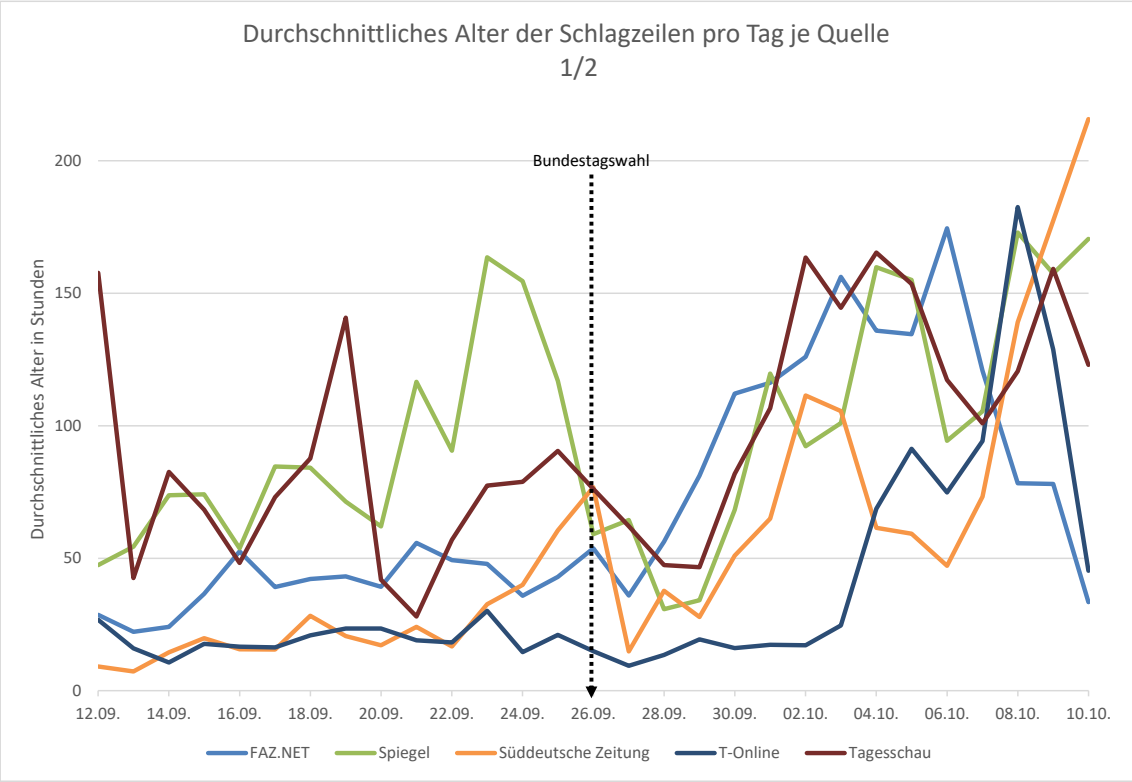


Abbildung 43: Durchschnittliches Alter der Schlagzeile / Quelle – 1/2

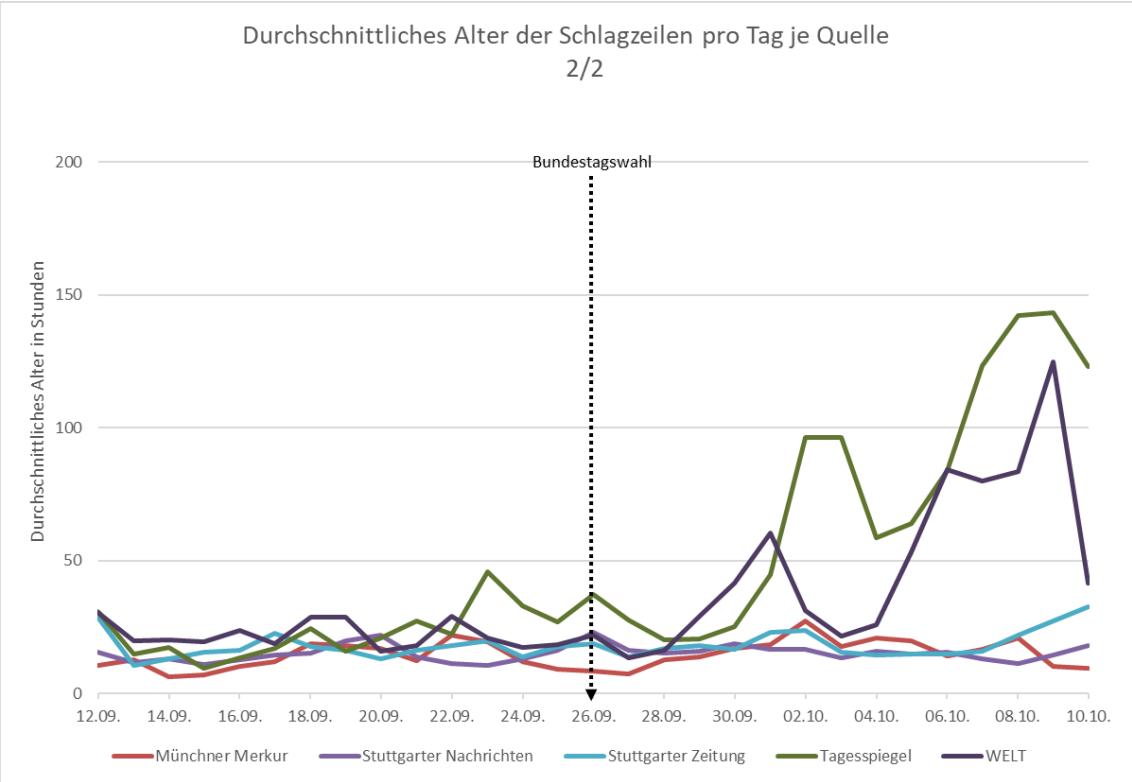


Abbildung 44: Durchschnittliches Alter der Schlagzeile / Quelle – 2/2

Die Darstellung - in Abhängigkeit nach den Suchbegriffen - wurde der Übersicht halber zusätzlich in mehrere getrennten Darstellungen aufgeteilt:

- Bundestagswahlbezug
- Baden-Württemberg Personen
- Orte / Firmen / Schlagwörter.

Der Suchbegriff Bernd Riexinger ist nicht vertreten. Zu diesem Suchbegriff existieren insgesamt nur wenige Schlagzeilen und für diese wurde von Google kein Veröffentlichungsdatum mitgeliefert. Eine Verarbeitung dieser Daten ist im Rahmen dieser Fragestellung also nicht sinnvoll möglich.

Bei allen Suchbegriffen kann wieder ein Anstieg des durchschnittlichen Alters der Schlagzeilen zum Ende des Untersuchungszeitraumes festgestellt werden. Für den Begriff Wolfgang Schäuble zeigt sich ein kurzfristiger starker Anstieg im Alter von Schlagzeilen zum Zeitpunkt der Bundestagswahl. Sprich am Tag der Wahl gab es wenig aktuelle Schlagzeilen zu diesem Suchbegriff. Für den Suchbegriff Alice Weidel ist direkt zu Beginn des Untersuchungszeitraums ein sehr starker Rückgang des Alters der Schlagzeilen zu sehen. Während des Betrachtungszeitraums um die Wahl herum sind Schlagzeilen zu diesem Suchbegriff eher jung. Zum Ende hin verfällt dieser Effekt jedoch wieder sehr stark und die Kurve steigt sogar auf einen noch höheren Alterswert als zu Beginn des Verlaufs. Für die Suchterme Franziska Brantner und Michael Theurer werden nur Teilstriche dargestellt, da keine weiteren Daten zu anderen Tagen vorliegenden.

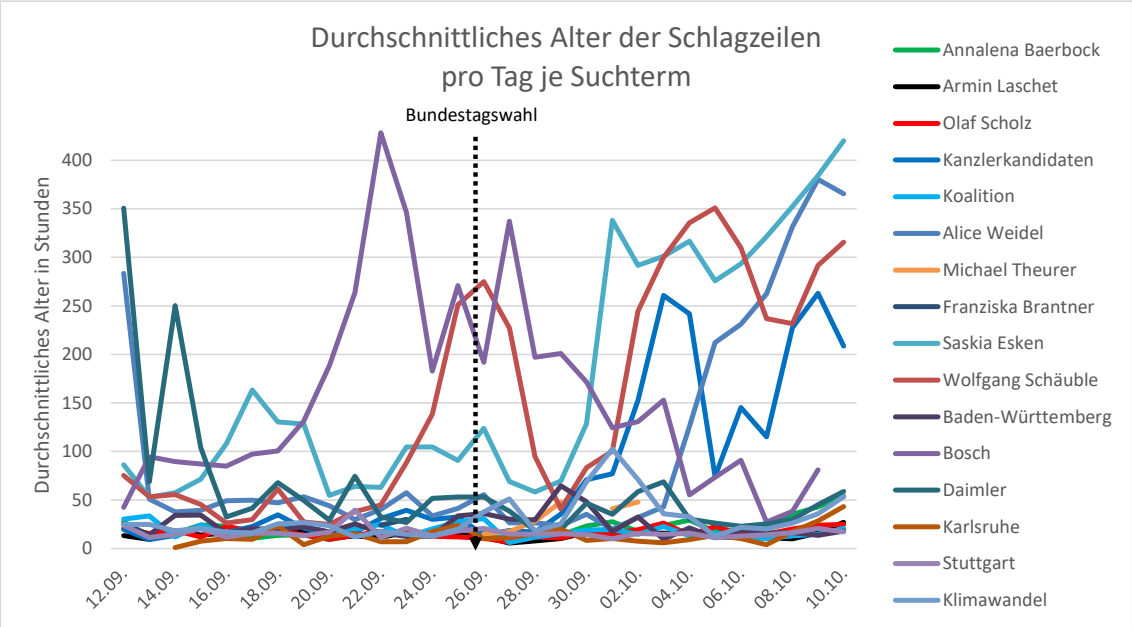


Abbildung 45 Durchschnittliches Alter der Schlagzeile / Gesamt

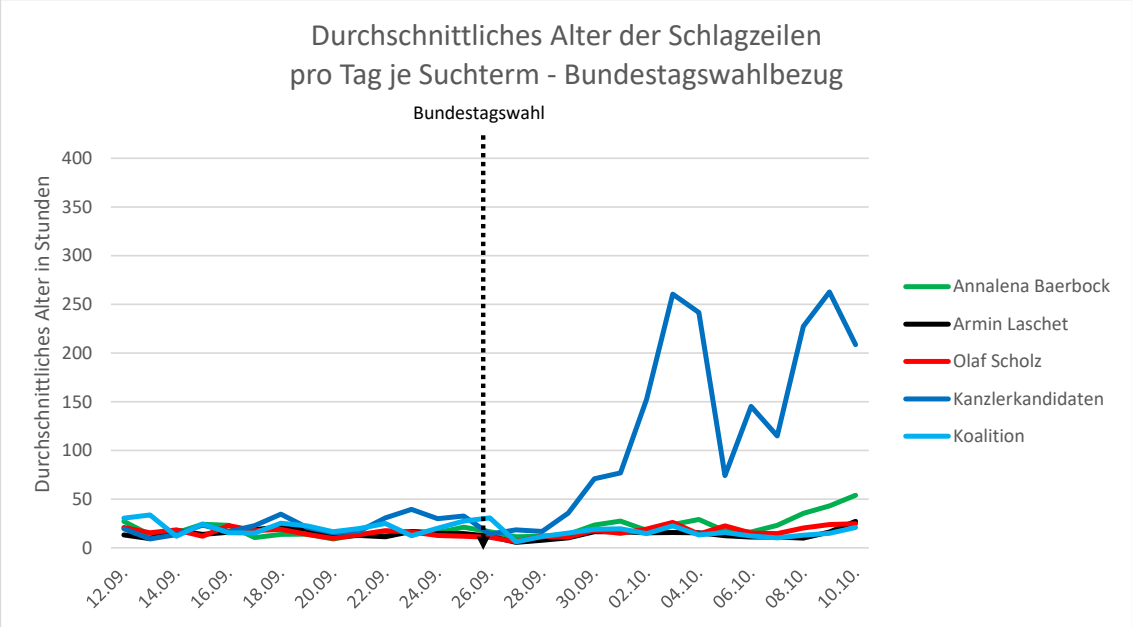


Abbildung 46: Durchschnittliches Alter der Schlagzeile / Bundestagswahlbezug

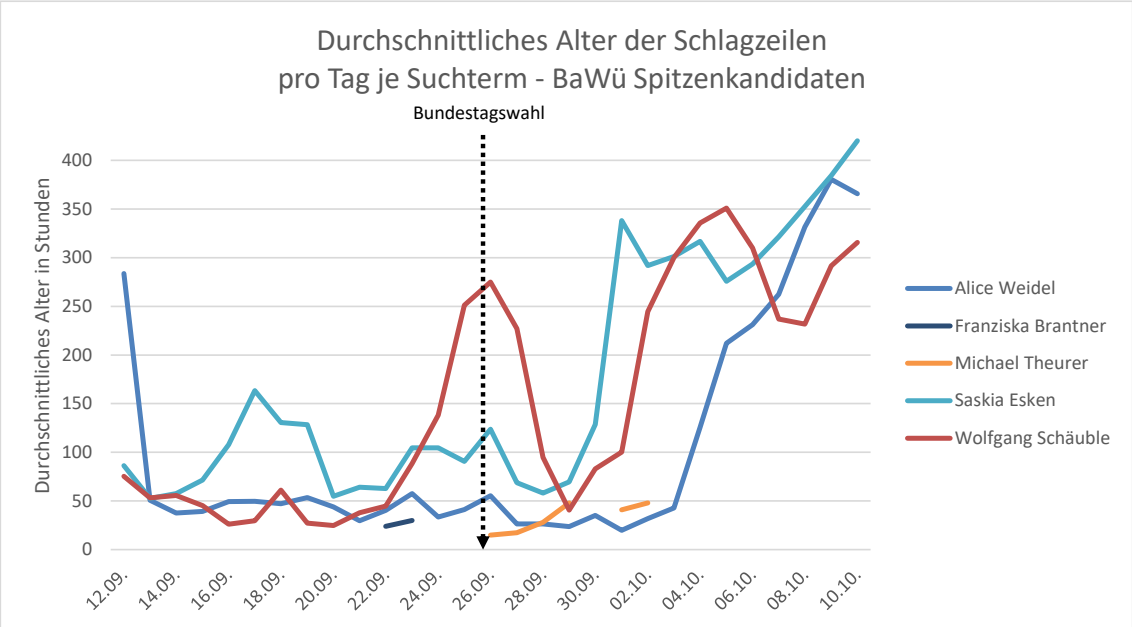


Abbildung 47: Durchschnittliches Alter der Schlagzeile / BaWü Spitzenkandidaten

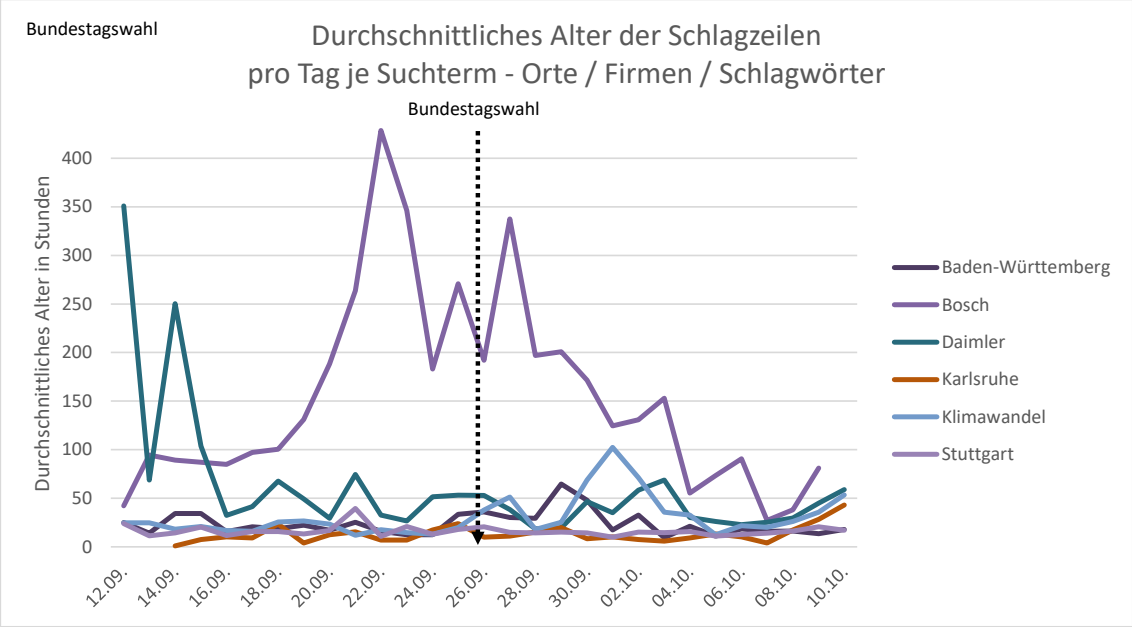


Abbildung 48: Durchschnittliches Alter der Schlagzeile / Orte / Firmen / Schlagwörter

Beispielhaft am Suchbegriff Bosch lässt sich folgende Aussage treffen: Zu Beginn der Analyseperiode war der Suchbegriff eher relativ „jung" wurde jedoch im Laufe der Tage immer „älter" und war kurz vor der Wahl als kurz danach am „ältesten". Folgend wurde das durchschnittliche Alter wieder unter 100 Stunden und damit etwas „jünger". Es könnte so gedeutet werden, dass der Suchbegriff „Bosch" für eine gewisse Zeitspanne (um die Wahl herum) kein aktuelles/aktualisiertes Thema war.

2.4.3 Veränderung Rang über der Zeit (#19)

AUFGABENSTELLUNG DER FRAGE:

Beobachtet werden soll die Veränderung des Rangs einer Schlagzeile über den Betrachtungszeitraum hinweg.

ABFRAGETECHNIK

Zur Darstellung werden hierbei nur die zehn langlebigsten Schlagzeilen berücksichtigt, um eine sinnvolle Visualisierung zu ermöglichen. Dazu wird das Lade-Datum von Schlagzeilen gruppiert nach der Schlagzeile und dem Tag des Ladens. Tritt eine Schlagzeile an einem Tag mehrmals auf, z.B. in mehreren der stündlichen Abfragen oder zu unterschiedlichen Suchbegriffen, so wurde die Schlagzeile dennoch nur ein einmal stellvertretend für den gesamten Tag betrachtet. Dabei wurde aus allen Schlagzeilen des Tages das Minimum das Rangs der Schlagzeile als stellvertretender Wert verwendet. Die Ergebnisse werden pro Schlagzeile in separaten Liniendiagrammen dargestellt. Im Titel der Diagramme steht der Text der Schlagzeile sowie anschließend in Klammern die zugehörige Quelle. Die Reihenfolge der Grafiken entspricht der Reihenfolge nach dem Kriterium der Lebensdauer gemessen an der Anzahl der stündlichen Vorkommen. Diese Anzahl ist aus Gründen der Übersichtlichkeit in den Grafiken nicht abgebildet.

ERGEBNIS DER ANALYSE:

Auf der Y-Achse der Diagramme wird der Rang dargestellt. Wie beschrieben werden nur Schlagzeilen der Ränge 1-10 betrachtet. Wenn für eine Schlagzeile die Linie unterbrochen

dargestellt wird, bedeutet das, dass diese Schlagzeile nicht mehr die Ränge 1-10 erreichte, sondern entweder gar nicht mehr auftrat oder auf einen (numerisch) höher liegenden Rang abgefallen ist.

Bei einigen Kurvernverläufen lässt sich ein Einfluss der Bundestagswahl feststellen, da der Rang der Schlagzeile zu diesem Datum hin einer Änderung unterlag.

Hinweise:

- Im zeitlichen Verlauf verändert sich das Veröffentlichungsdatum einer Schlagzeile unter Umständen mehrfach.
  Beispiel: Es könnte ein Veröffentlichungsdatum wie folgt abgerufen werden, wobei der Link zur Nachricht unverändert ist:
  Initialer Abruf (Tag1, minimaler Rang): „vor 2 Tagen"
  Nächster Abruf (Tag2, minimaler Rang): „vor 3 Tagen"
  Nächster Abruf (Tag3, minimaler Rang): „heute"
  Nächster Abruf (Tag4, minimaler Rang): „vor 4 Tagen""
  Der Grund für dieses Verhalten lässt sich mit den zugrundeliegenden Daten nur vermuten, so könnte die Quelle (oder Google) das Datum verändert haben.
- Ferner ist zu berücksichtigen, dass für einen Tag immer das Minimum aller Ränge des Tages gezogen wurde.
  Beispiel: Am 10.10. gab es zwei Schlagzeilen. Eine mit Rang 3 und eine mit Rang 1. Gezeigt wird hier also Rang 1.

Bei der nachstehenden Schlagzeile ist ein anfänglicher niedriger (d.h. eine gute Positionierung) Rang zu erkennen, der zur Wahl hin abfällt und dann leicht alterniert. Später ist der Verlauf unterbrochen. Zum Ende des Betrachtungszeitraums tritt die Schlagzeile jedoch wieder auf und arbeitet sich wieder stetig auf einen niedrigen Rang vor. Letzteres kann technisch gesehen zwei Ursachen haben: Der Bericht hinter der Schlagzeile wurde im damals aktuellen zeitlichen Kontext aktualisiert und erneut herausgebracht, andernfalls könnte es sein, dass andere Schlagzeilen zu dem Suchbegriff stärker in den Hintergrund rückten, wodurch diese Schlagzeile indirekt begünstigt wurde.

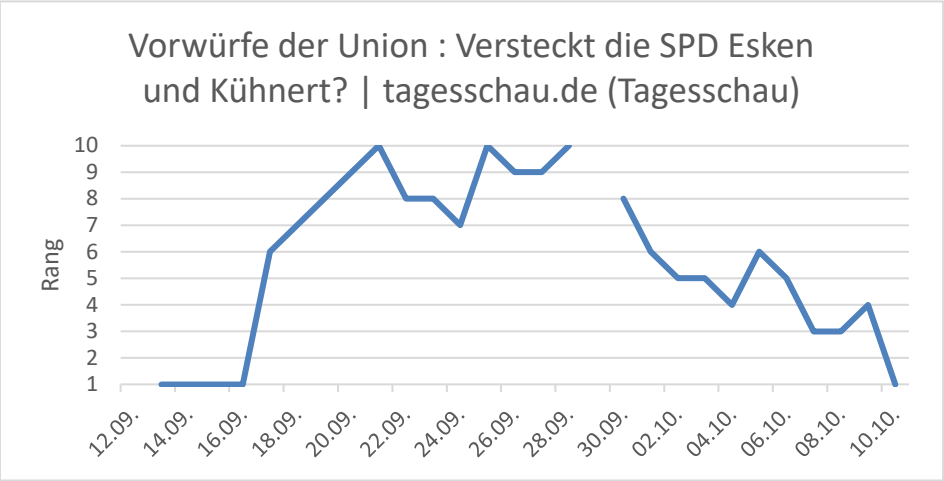


Abbildung 49: Zeitlicher Verlauf des Rangs - Schlagzeile 1

Bei der zweiten Schlagzeile sind mehrere deutliche Schwankungen zu erkennen. Der Verlauf beginnt und endet wie zuvor mit eher besseren (niedrigeren) Rängen.

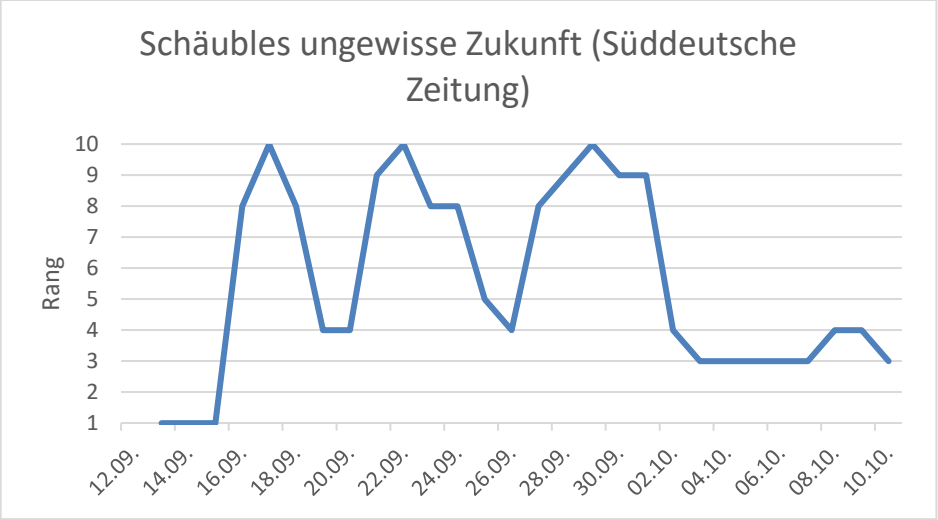


Abbildung 50: Zeitlicher Verlauf des Rangs - Schlagzeile 2

Bei diesem Verlauf der Schlagzeile, vermeintlich ohne Bezug zur Bundestagswahl, scheint sich um den Zeitraum der Wahl eine Verbesserung (geringerer Rang) zu zeigen mit anschließender Verschlechterung des Rangs. Beginn und Ende des Verlaufs liegen in schlechteren (höheren) Rängen. Dies ist ein entgegengesetzter Verlauf zu den meisten übrigen Schlagzeilen.

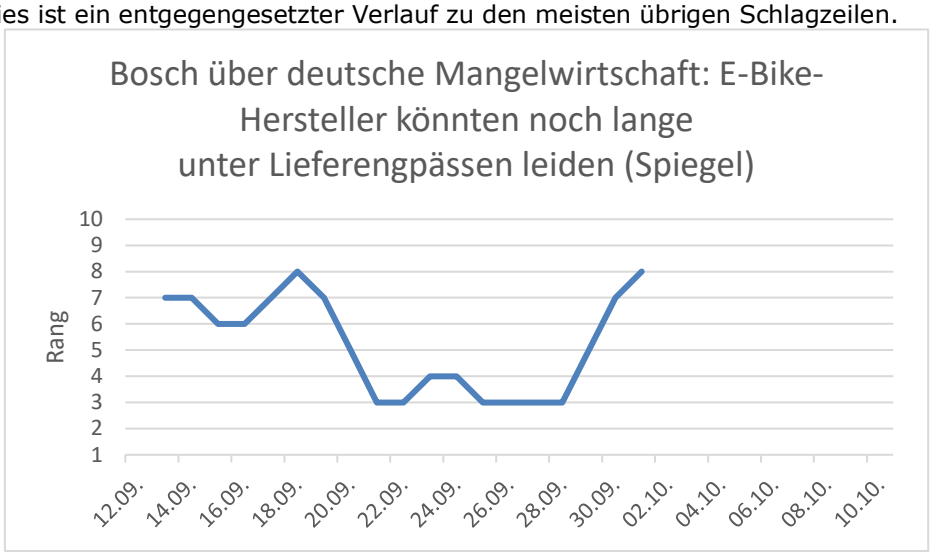


Abbildung 51: Zeitlicher Verlauf des Rangs - Schlagzeile 3

Wie bereits bei der Schlagzeile zu Beginn ist wieder ein anfänglicher typischer Verlauf zu erkennen mit einer kurzen Unterbrechung des Auftretens und einem Wiederaufleben der Schlagzeile zum Ende des Betrachtungszeitraumes.

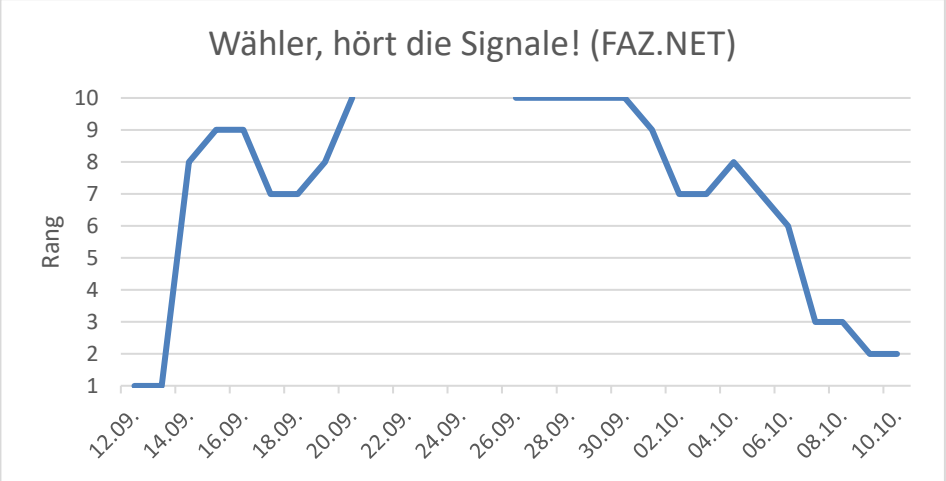


Abbildung 52: Zeitlicher Verlauf des Rangs - Schlagzeile 4

Auch bei dieser Schlagzeile, welche inhaltlich eher keinen Bezug zur Bundestagswahl hat, zeigt sich eine Verbesserung zum Tag der Wahl hin und eher schlechtere Ränge bei Beginn und Ende des Verlaufs. Erneut ein gegenläufiger Trend gegenüber anderen Schlagzeilen mit politischem Bezug.

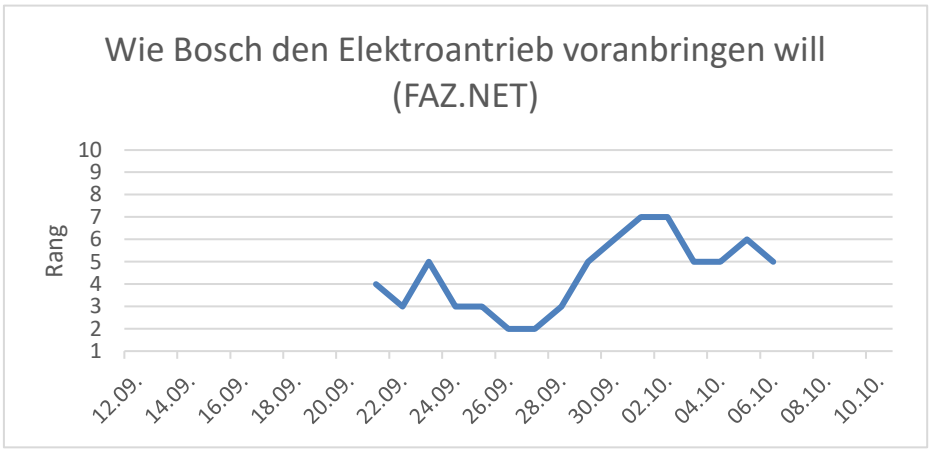


Abbildung 53: Zeitlicher Verlauf des Rangs - Schlagzeile 5

Erneut zwei politische Beispiele für Schlagzeilen, welche im späteren Verlauf wieder auf niedrige (bessere) Ränge zurückkehren:

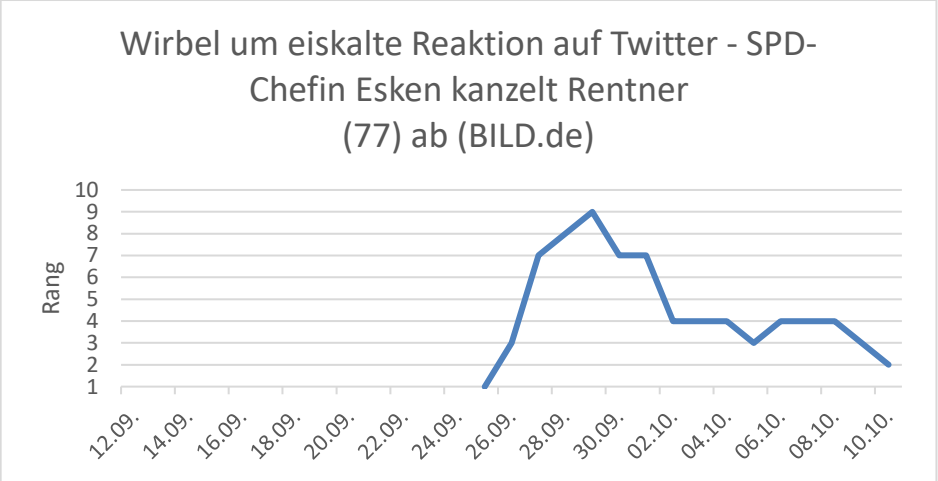


Abbildung 54: Zeitlicher Verlauf des Rangs - Schlagzeile 6

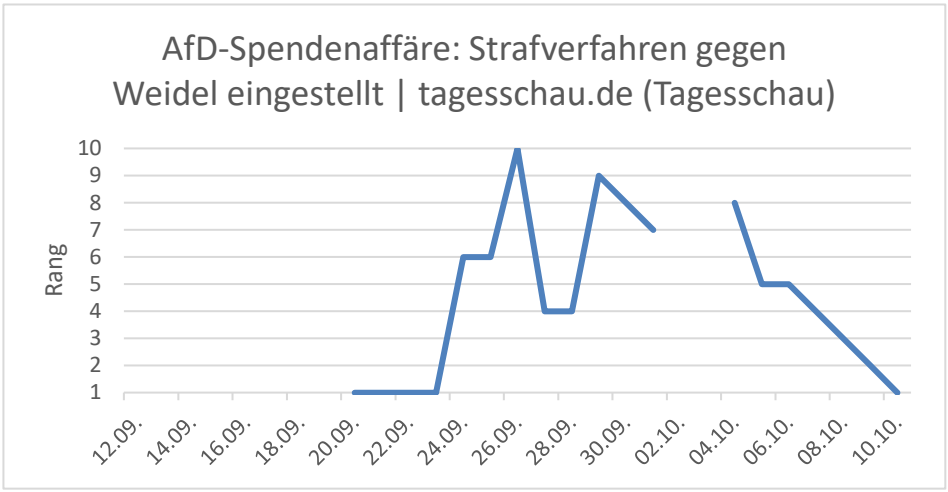


Abbildung 55: Zeitlicher Verlauf des Rangs - Schlagzeile 7

Die folgende Schlagzeile der FAZ zeigt einen typischen Verlauf, wie man ihn zunächst annehmen würde. Eine Schlagzeile tritt auf, mit einem niedrigen (guten) Rang aufgrund ihrer Aktualität und einem schlechteren Rangverlauf mit zunehmendem Alter der Schlagzeile. Ein Wiederaufleben tritt im betrachteten Zeitraum nicht wieder auf.

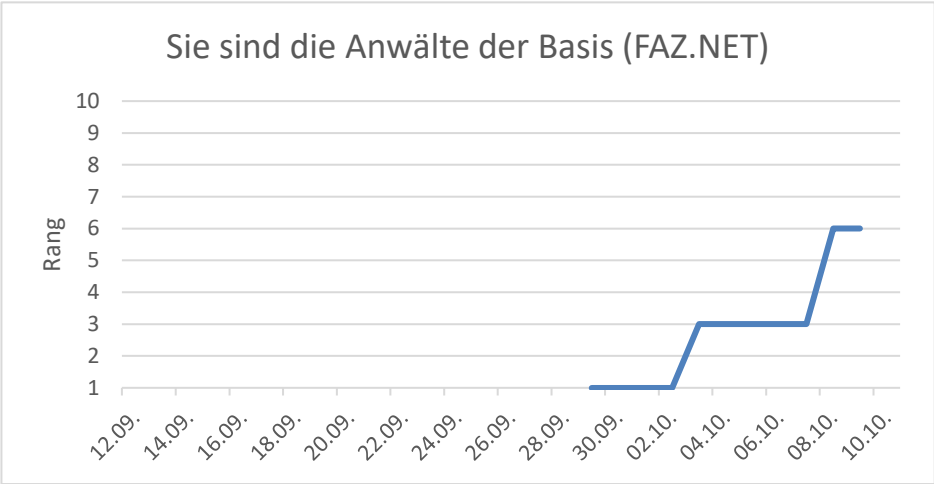


Abbildung 56: Zeitlicher Verlauf des Rangs - Schlagzeile 8

Die nächsten Schlagzeilen mit dem klaren inhaltlichen Bezug zu dem Schlagwort „Schäuble" zeigt, wie die Schlagzeile zu „Schäuble" zuvor einen Zusammenhang mit dem Tag der Wahl, eine anschließende Abschwächung und endet jedoch mit einem deutlich verbesserten (niedrigeren) Rangverlauf.

Abbildung 57: Zeitlicher Verlauf des Rangs - Schlagzeile 9

Diese Schlagzeile zeigt einen ähnlichen Trend wie viele vorherige, einem Start in niedrigen Rängen, einem Abfallen und schließlich einem erneuten Wiederaufleben zurück in niedrigere Ränge.

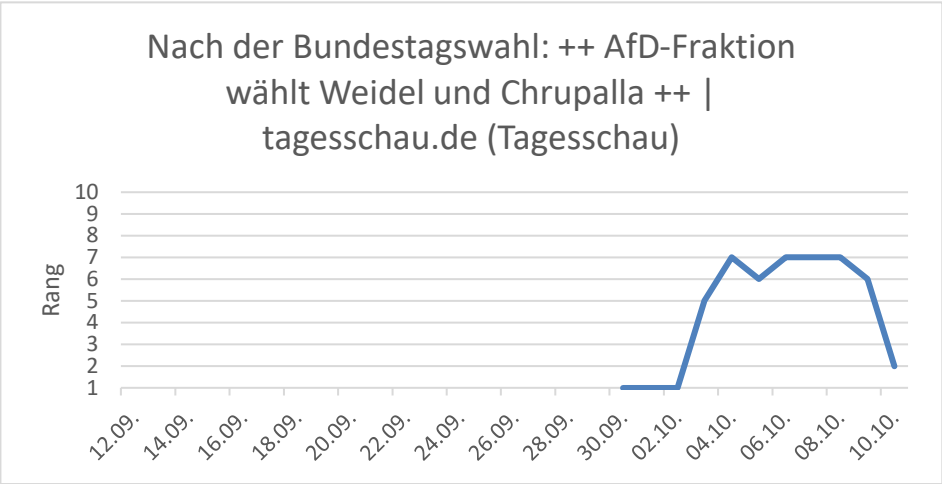


Abbildung 58: Zeitlicher Verlauf des Rangs - Schlagzeile 10

2.4.4 Textuelle Ähnlichkeit (#20)

AUFGABENSTELLUNG DER FRAGE:

Im Rahmen dieser Aufgabenstellung soll geprüft werden, ob es Schlagzeilen gibt, die textuell sehr ähnlich sind.

ABFRAGETECHNIK

Eine Möglichkeit zur Beantwortung dieser Fragstellung ist der Einsatz einer Entitätsanalyse. Diese wurde bereits im Kapitel 0

Begriffsdefinitionen näher beschrieben. Im Wesentlichen werden hierbei aus den Schlagzeilen Begriffe bestimmt, die einem der Standard-Entitätstypen wie z.B. Personen, Orte, Organisationen o.Ä. entsprechen und es wird gezählt, wie groß die Anzahl identischer Entitäten beim Vergleich aller Schlagzeilen untereinander ist.
Bei einer vollständigen Betrachtung aller der insgesamt ca. erfassten 340.000 Schlagzeilen ergäbe sich dabei die Notwendigkeit, insgesamt 340.000x340.000 Kombinationen (mit jeweils mehreren Entitäten) zu prüfen, was einen enormen Rechenaufwand bedeutet. Aus diesem Grund wurde ein vereinfachtes Verfahren gewählt, bei dem die Anzahl der untersuchten Schlagzeilen und damit die Zahl der Vergleiche deutlich reduziert wird.

Für die Durchführung der vereinfachten Analyse wurde wie folgt vorgegangen:

- Die Schlagzeilen wurden nach ihrer Häufigkeit im gesamten Betrachtungszeitraum über alle Parameter hinweg sortiert.
- Es werden 2 Mengen an Schlagzeilen gebildet:
  - Menge A: die häufigsten 1.000 Schlagzeilen werden als Such-Schlagzeilen festgelegt
  - Menge B: die häufigsten 1.000 Schlagzeilen werden als Vergleichs-Schlagzeilen festgelegt
- Für die Such-Schlagzeilen (Menge A) wurden alle Entitäten mit den Entitäten aller Vergleichs-Schlagzeilen (Menge B) abgeglichen und dabei die Schnittmenge identischer Entitäten gebildet.
- Die Anzahl der Elemente in den Schnittmengen wurde gezählt.
- Aus allen Abgleichen wurden im Folgenden nur die Kombinationen mit der höchsten Anzahl aus der jeweiligen Schnittmenge weiter betrachtet.
- Zusätzlich wurde dieses Verfahren auf die Suchbegriffe mit Bundestagswahlbezug umgesetzt. Die Menge aller Schlagzeilen zu einem bestimmten Suchbegriff bildet dabei sowohl Such- als auch Vergleichsmenge.

ERGEBNIS DER ANALYSE:

Gezeigt werden die Schlagzeilen mit der höchsten Ähnlichkeit in Bezug auf die Entitäten. Wie die Daten zeigen, existiert keine höhere Ähnlichkeit als in neun gemeinsamen Schlagwörtern (bei den Kandidaten). Viele der gefundenen Schlagzeilen mit einem hohen Ähnlichkeitsmaß sind bzgl. der Wortwahl nahezu identisch und unterscheiden sich nur in Nuancen. Bei geringeren Ähnlichkeitsmaß werden in der Regel unterschiedliche Inhalte thematisiert, so dass aus semantischer Sicht nur selten eine inhaltliche Ähnlichkeit zu beobachten ist.
Ferner kann festgestellt werden, dass auch teilweise nur kleine Änderungen zu einer neuen Schlagzeile führen (Beispielsweise „2021" weggelassen oder auch nur Umstellung des Textes wie „China und Russland" zu „Russland und China" – ohne Auswirkung auf die Ähnlichkeit)

Folgend ein Bespiel, gleiche Textstellen sind „grün" (= gleicher Text) und sichtbare Unterschiede „gelb" hinterlegt:

Such-Schlagzeile	Ähn-lich-keit	Vergleichs-Schlagzeile
Wahl 2021: ARD-Mann Zamperoni führt skurriles AfD-Weidel-Interview - Zuschauer nennen ihn „Gewinner des Abe…	7	Wahl: ARD-Mann Zamperoni führt skurriles Interview mit Alice Weidel - Zuschauer nennen ihn „Gewinner des Ab…
SPD-Chefin Saskia Esken geht Rentner an: „Das muss reichen…" - Wirbel in social Media	6	Saskia Esken (SPD-Chefin) geht Rentner an: „Das muss reichen…" - Wirbel in social Media → Hinweis: Nahezu gleicher Text, aber Reihenfolge im Text geändert

Tabelle 12 Textuelle Ähnlichkeit Beispiel

In Folgenden ein Auszug aus der „Top 1.000" Tabelle. Die vollständige Tabelle ist im Anhang beigefügt:

Such-Schlagzeile	Ähn-lich-keit	Vergleichs-Schlagzeile
SPD-Politikerin als Bundestagspräsidentin im Gespräch: Aydan Özoğuz könnte Nachfolgerin von Schäuble werden	7	SPD-Politikerin als Bundestagspräsidentin im Gespräch: Aydan Özoğuz könnte Nachfolgerin von Schäuble werden ...
Wahl 2021: ARD-Mann Zamperoni führt skurriles AfD-Weidel-Interview - Zuschauer nennen ihn „Gewinner des Abe...	7	Wahl: ARD-Mann Zamperoni führt skurriles Interview mit Alice Weidel - Zuschauer nennen ihn „Gewinner des Ab...
SPD-Politikerin als Bundestagspräsidentin im Gespräch: Aydan Özoğuz könnte Nachfolgerin von Schäuble werden ...	7	SPD-Politikerin als Bundestagspräsidentin im Gespräch: Aydan Özoğuz könnte Nachfolgerin von Schäuble werden
Robuster Umgang mit China und Russland: Baerbock will eine neue Außenpolitik, Scholz nicht	7	Robuster Umgang mit Russland und China: Baerbock will eine neue Außenpolitik, Scholz nicht
... (weitere Schlagzeilen mit Ähnlichkeit 7 finden sich im Anhang)	...	...
Wahl 2021: ARD-Moderator Zamperoni führt skurriles AfD-Weidel-Interview - Zuschauer nennen ihn „Gewinner de...	6	Wahl: ARD-Mann Zamperoni führt skurriles Interview mit Alice Weidel - Zuschauer nennen ihn „Gewinner des Ab...
SPD-Chefin Saskia Esken geht Rentner an: „Das muss reichen..." - Wirbel in social Media	6	Saskia Esken (SPD-Chefin) geht Rentner an: „Das muss reichen..." - Wirbel in social Media
Spitzenkandidatin: Alice Weidel kündigt Klagen der AfD gegen 2G-Regel an	6	AfD-Spitzenkandidatin Weidel kündigt Klagen gegen 2G-Regel an
... (weitere Schlagzeilen finden sich im Anhang)	...	...
Wahl der AfD-Fraktionsspitze: Weidel schrammt an einer Niederlage vorbei	5	"Wahl an die Spitze einer radikalen Truppe: Weidel und Chrupalla führen die AfD-Fraktion
Spahn konfrontiert im TV Dreyer mit wüster Behauptung zu dem Verbleib Eskens	5	Spahn konfrontiert im TV Dreyer mit wüster Behauptung zu Eskens Verbleib
Nach Bundestagswahl: Weidel und Chrupalla neue AfD-Fraktionsspitze	5	"Nach der Bundestagswahl: ++ AfD-Fraktion wählt Weidel und Chrupalla ++ |

Tabelle 13 Textuelle Ähnlichkeit Top 1.000

In Folgenden ein Auszug aus der Tabelle „Annalena Baerbock"
Die vollständige Tabelle ist im Anhang beigefügt:

Such-Schlagzeile	Ähnlichkeit	Vergleichs-Schlagzeile
Bundestagswahl 2021: Wieso Laschet, Scholz und Baerbock nicht auf jedem Wahlzettel stehen – Fakten zur Bundestagswahl	*9*	Bundestagswahl 2021: Wieso Laschet, Scholz und Baerbock nicht auf jedem Wahlzettel stehen – Fakten zur Wahl
TV-Triell zwischen Baerbock, Laschet und Scholz - So sehen Sie die Sendung live im TV und im Live-Stream	*8*	Triell: Laschet, Baerbock und Scholz - So sehen Sie die Sendung live im TV oder Live-Stream
Bundestagswahl 2021 in Brandenburg: Olaf Scholz vor Annalena Baerbock - Ergebnisse im Wahlkreis 61 von Potsdam bis Teltow-Fläming	*8*	Bundestagswahl Potsdam Ergebnis: Hochrechnung, Prognose, Ergebnisse – Olaf Scholz bekommt Direktmandat, Baerbock auf Platz 2
… (weitere Schlagzeilen finden sich im Anhang)	…	…
Habeck als Retter fatales Signal: Warum Baerbock doch Vizekanzlerin werden könnte	*5*	Habeck als Retter fatales Signal: Warum Baerbock doch Vizekanzlerin werden könnte - n-tv.de
Grünen-Knall nach Bundestagswahl: Habeck übernimmt für Baerbock - Lindner äußerte sich vielsagend	*5*	Bundestagswahl: Sondierungen von Grüne und FDP! Baerbock, Habeck, Lindner in erstem Statement mit Fingerzeig
Koalitionsvorsondierungen: Erstes Treffen von Grünen und FDP endet mit Selfie	*5*	Vor-Sondierungen über mögliche Koalition: Erstes Treffen von Grünen und FDP endet mit Selfie

Tabelle 14 Textuelle Ähnlichkeit Annalena Baerbock

In Folgenden ein Auszug aus der Tabelle „Armin Laschet"
Die vollständige Tabelle ist im Anhang beigefügt:

Such-Schlagzeile	Ähnlichkeit	Vergleichs-Schlagzeile
Bundestagswahl 2021 Live: Laschet will Jamaika-Koalition - Scholz benennt Verhandlungsteam	*8*	Bundestagswahl 2021 Live: Scholz benennt Verhandlungsteam – Laschet will Jamaika-Koalition
„Ist Maaßen ein Nazi?" Armin Laschet: Kinder nehmen CDU-Chef in die Mangel	*8*	Armin Laschet: Kinder nehmen CDU-Chef in die Mangel - „Ist Maaßen ein Nazi?"

CDU-Hammer: Abgeordnete fordert Laschets Rücktritt - Ex-Minister denkt laut über Kanzler Söder oder Merz nach	*8*	CDU-Hammer: Abgeordnete fordert Laschet-Rücktritt - Ex-Minister denkt laut über Kanzler Söder oder Merz nach
… (weitere Schlagzeilen finden sich im Anhang)	…	…
Bundestagswahl 2021 Umfrage: So erfolgreich wäre die Union mit Söder statt Laschet gewesen	*5*	Bundestagswahl: Plus 2! Wende-Hoffnung für Laschet durch neue Forsa-Umfrage
Bundestagswahl: Söder stürzt sich für Laschet ins letzte Gefecht - CSU-Chef gibt erstes Ziel bereits auf	*5*	Laschet am Ende und Söder am Ziel? Was der CSU-Chef nun plant
Armin Laschet Nachfolger: Hendrik Wüst - der neue Ministerpräsident für NRW	*5*	NRW-Ministerpräsident: Laschet will Wüst als seinen Nachfolger vorschlagen

Tabelle 15 Textuelle Ähnlichkeit Armin Laschet

In Folgenden ein Auszug aus der Tabelle „Olaf Scholz"
Die vollständige Tabelle ist im Anhang beigefügt:

Such-Schlagzeile	Ähn-lich-keit	Vergleichs-Schlagzeile
Bundestagswahl 2021 News - Scholz wirbt um FDP zur Bildung einer Ampel-Koalition	*8*	Bundestagswahl 2021 News: Scholz wirbt um FDP zur Bildung einer Ampel-Koalition
Bundestagswahl 2021 Live: Laschet will Jamaika-Koalition - Scholz benennt Verhandlungsteam	*8*	Bundestagswahl 2021 Live: Scholz benennt Verhandlungsteam – Laschet will Jamaika-Koalition
Bundestagswahl 2021 live: Kostet Laschet der Wahllokal-Fauxpas die Wahl? Insa-Chef legt sich fest	*8*	Bundestagswahl live: Scholz oder Laschet - Kostet der Wahllokal-Fauxpas die Wahl? Insa-Chef legt sich fest
… (weitere Schlagzeilen finden sich im Anhang)	…	…
Allemagne : Olaf Scholz, le succès surprise du favori à la course à la chancellerie	*5*	Allemagne : Olaf Scholz conforte sa position de favori dans la course à la chancellerie
„People want the Christian Democratic Union in the opposition": Scholz überrascht mit Pressekonferenz auf Englisch	*5*	„People want the CDU in the opposition": Scholz überrascht mit Pressekonferenz auf Englisch
Die Bundestagswahl 2021 live: Scholz gewinnt Direktduell im Wahlkreis deutlich gegen Baerbock	*5*	Bundestagswahl 2021: So schneiden Baerbock und Scholz in ihrem Potsdamer Wahlkreis ab

Tabelle 16 Textuelle Ähnlichkeit Olaf Scholz

Eine Alternative zu dem angewendeten Verfahren des Ähnlichkeitsvergleichs basierend auf der Zählung von Entitäten besteht im Einsatz von NLP (Natural Language Processing) Algorithmen, bei

denen die zu vergleichenden Schlagzeilen entsprechend ihrer semantischen Bedeutung in einen höher-dimensionalen Vektorraum transformiert werden (z.B. Verfahren wie Word2Vec oder Sentence2Vec). Das Transformationsverfahren wird dabei so gewählt, dass semantisch ähnliche Schlagzeilen im resultierenden Vektorraum in einer räumlichen Nachbarschaft zu liegen kommen. Durch die Untersuchung von Clustern im Vektorraum können dann semantisch ähnliche Schlagzeilen identifiziert werden. Mit einem derartigen Verfahren können semantisch ähnliche, aber bzgl. der Wortwahl völlig unterschiedliche Begriffe wie z.B.

Auto und Fahrzeug
Klimawandel und Erderwärmung
Strafverfahren und Prozess
…

identifiziert und miteinander in Beziehung gesetzt werden. Nachteil des Verfahrens ist, dass die Parameter des Transformationsalgorithmus zunächst in einer Trainingsphase unter Nutzung einer ausreichenden Anzahl von Trainingsbeispielen berechnet werden müssen. Aufgrund des hiermit verbundenen Aufwands und des begrenzten Scopes der Studie sind derartige Untersuchungen einer nachfolgenden Betrachtung überlassen.

2.5 Schlagwörter

2.5.1 Zusammenhang zwischen Schlagwörtern und Suchbegriffen (#5)

AUFGABENSTELLUNG DER FRAGE:

In diesem Kapitel wird der Frage nachgegangen, ob es in den Schlagzeilen zu einem bestimmten Suchbegriff einen signifikanten Zusammenhang zwischen herausragenden Schlagwörtern und dem Suchbegriff gibt.

ABFRAGETECHNIK

Der erste Schritt bei der Bearbeitung dieser Aufgabenstellung ist die Ermittlung der Schlagwörter einer Schlagzeile. Hierzu wird wie in Abschnitt 1.4.5 beschrieben eine Entitätsanalyse auf Basis der Google NLP API genutzt.
Im zweiten Schritt erfolgt die Visualisierung der gefundenen Schlagworte unter Nutzung von sogenannten WordClouds. In einer WordCloud können Begriffe in Verbindung mit der Häufigkeit des Auftretens dieser Begriffe dargestellt werden, wobei ein Begriff in umso größerer Schriftgröße erscheint, je häufiger er im Verhältnis zu den anderen Begriffen vorkommt. Besonders lange Wörter wie „Bundestagswahl" nehmen also naturgemäß mehr Fläche ein als kurze Schlagwörter wie „SPD", falls beide Wörter gleichermaßen häufig vorkämen und somit dieselbe Schriftgröße erhielten. Die Positionierung und die Farbgebung einzelner Wörter innerhalb der Darstellung sowie die Ausrichtung des Textes sind dabei rein zufällig erzeugt.
Für die Darstellung in Form von WordClouds muss also nach Bestimmung der Schlagwörter mittels der Entitätsanalyse die Häufigkeit der Schlagwörter für den jeweiligen Suchbegriff ermittelt werden.

ERGEBNIS DER ANALYSE:

Zur besseren Einordnung der Ergebnisse soll zunächst eine Hilfstabelle erzeugt werden, die die Anzahl der Schlagzeilen für den jeweiligen Suchbegriffen auflistet. Es wurden dabei alle Schlagzeilen der Ränge 1-10 über den gesamten Beobachtungszeitraum berücksichtigt.

Suchbegriff	Anzahl Schlagzeilen
Annalena Baerbock	26.366
Armin Laschet	27.634
Olaf Scholz	27.630
Kanzlerkandidaten	25.275
Koalition	26.272
Alice Weidel	24.603
Bernd Riexinger	344
Franziska Brantner	1.056
Michael Theurer	1.486
Saskia Esken	20.918
Wolfgang Schäuble	18.060
Baden-Württemberg	25.509
Bosch	15.535
Daimler	25.702
Karlsruhe	23.844
Stuttgart	25.472
Klimawandel	24.762

Tabelle 17 Suchbegriffe und Anzahl Schlagzeilen

Es wurden für alle Suchbegriffe je eine WordCloud erzeugt, jedoch werden nur einige exemplarisch an dieser Stelle dargestellt.

Die WordClouds zeigen auf interessante und anschauliche Art und Weise die Verteilung der Schlagwörter. Folgende Sachverhalte sind besonders auffallend:

- Der jeweilige Suchbegriff (und direkte Ableitungen davon, z.B. „Olaf Scholz" und „Scholz") nimmt in aller Regel den größten Raum in der betreffenden WordCloud ein. Ein Indiz dafür, dass die Schlagzeilen in aller Regel auch den Suchbegriff selbst enthalten und das Anzeigen eines Schlagzeilenobjekts weniger über den Inhalt des dahinterstehenden Artikels gesteuert wird. Eine Ausnahme hiervon besteht lediglich in der WordCloud des Suchbegriffs „Koalition".
- Wenig überraschend ist, dass die zugehörigen Partei-Kürzel bei Suchbegriffen von Kanzlerkandidaten sowie der eigene Name des Kandidaten prägnant auftreten.
- Einige der WordClouds, z.B. die zu Bernd Riexinger zeigen noch viele freie Flächen auf, die bei anderen Suchbegriffen insbesondere durch große prangende Schlagworte gefüllt sind. Dies liegt unter anderem daran, dass zu diesem Suchbegriff insgesamt nur wenige Schlagzeilen (siehe Tabelle 17) und damit nur wenige Schlagworte vertreten sind.
- Einige der WordClouds zeigen ein relativ starkes Abfallen der Häufigkeit der jeweiligen Schlagworte, mit sehr wenigen dominanten Hauptschlagwörtern und vielen deutlich selteneren Unterschlagwörtern. Beispiele hierfür die die Suchbegriffe „Baden-Württemberg", „Saskia Esken" oder „Klimawandel". Dahingegen ist in den meisten anderen WordClouds der Effekt zu beobachten, dass es neben dem Hauptschlagwort noch eine ganze Reihe von weiteren Schlagworten mit häufigem Auftreten gibt.

Hinweis: Zur besseren Lesbarkeit werden im Folgenden nur die WordClouDs der Suchbegriffe mit direktem Bundestagswahlbezug dargestellt. Alle übrigen WordCloud-Darstellungen können im Anhang (siehe Kapitel 3.3 Vollständige Darstellung der WordClouds) eingesehen werden.

2.5.1.1 WordCloud Suchbegriff Annalena Baerbock

Abbildung 59 WordCloud Suchbegriff Annalena Baerbock

2.5.1.2 WordCloud Suchbegriff Armin Laschet

Abbildung 60 WordCloud Suchbegriff Armin Laschet

2.5.1.3 WordCloud Suchbegriff Olaf Scholz

Abbildung 61 WordCloud Suchbegriff Olaf Scholz

2.5.1.4 WordCloud Suchbegriff Kanzlerkandidaten

Abbildung 62 WordCloud Suchbegriff Kanzlerkandidaten

2.5.1.5 WordCloud Suchbegriff Koalition

Abbildung 63 WordCloud Suchbegriff Koalition

2.5.2 Verweildauer von Schlagwörtern in Abhängigkeit vom Suchbegriff (#6)

AUFGABENSTELLUNG DER FRAGE:

Gegenstand dieser Untersuchung war die Frage, ob es in Abhängigkeit von den einzelnen Suchbegriffen Schlagwörter gibt, die sich besonders lange in der Schlagzeilen Box halten.

ABFRAGETECHNIK

Als Ausgangslage dient erneut die Entitätsanalyse. Gruppiert nach den Schlagwörtern kann die Häufigkeit gemessen werden, mit der Schlagzeilen bei den stündlichen Suchen wiederkehrend gefunden werden konnte.
Pro Suchbegriff wurden die Schlagwörter ermittelt, die am längsten während des insgesamt 29-tägigen Beobachtungszeitraums vorkommen (in Tagen). Einige der Schlagwörter treten an jedem Tag über den gesamten Betrachtungszeitraum hinweg auf (d.h. an insgesamt 29Tagen).
Für jeden Suchbegriff wurden die drei zeitlich am längsten vorkommenden Schlagwörter bestimmt. Waren mehrere Schlagwörter gleich lange verfügbar, so wurden diejenigen mit dem häufigsten Auftreten insgesamt ausgewählt.

ERGEBNIS DER ANALYSE:

Die meisten der häufigsten Schlagworte werden über den gesamten Zeitraum gefunden. Lediglich für die Suchbegriffe (Personen) Franziska Brantner, Michael Theurer und Bernd Riexinger kommen die Begriffe seltener und weniger dauerhaft vor. Erneut sind die andauerndsten Schlagwörter für Suchbegriffe von Personen der eigene Name oder die Partei dieser Person.

TOP1			
Suchbegriff	Schlagwort	Tage	Vorkommen
Annalena Baerbock	Baerbock	29	12.283
Armin Laschet	Armin Laschet	29	10.849
Olaf Scholz	Olaf Scholz	29	12.285
Kanzlerkandidaten	Kanzlerkandidaten	29	8.302
Koalition	Koalition	29	11.874
Alice Weidel	AfD	29	14.337
Bernd Riexinger	Bernd Riexinger	11	332
Franziska Brantner	Europe	14	29
Michael Theurer	FDP	10	1.076
Saskia Esken	SPD	29	7.573
Wolfgang Schäuble	Schäuble	29	7.519
Baden-Württemberg	Baden-Württemberg	29	14.727
Bosch	Bosch	29	13.078
Daimler	Daimler	29	21.524
Karlsruhe	Karlsruhe	29	12.512
Stuttgart	Stuttgart	29	11.471
Klimawandel	Klimawandel	29	17.679

Tabelle 18 Verweildauer in Abhängigkeit von Schlagwort und Suchbegriff TOP 1

Top2			
Suchbegriff	Schlagwort	Tage	Vorkommen
Annalena Baerbock	Annalena Baerbock	29	5.507
Armin Laschet	Laschet	29	9.171
Olaf Scholz	Scholz	29	8.941
Kanzlerkandidaten	2021	29	7.951
Koalition	Bundestagswahl	29	5.283
Alice Weidel	Alice Weidel	29	8.283
Bernd Riexinger	Linken	11	51
Franziska Brantner	European Council on Foreign Relations	14	29
Michael Theurer	Theurer	9	496
Saskia Esken	Esken	29	7.120
Wolfgang Schäuble	Wolfgang Schäuble	29	5.830
Baden-Württemberg	Corona	29	2.989
Bosch	Elektroantrieb	23	824
Daimler	Aktie	29	2.925
Karlsruhe	Karlsruher	29	3.534
Stuttgart	VfB Stuttgart	29	6.643
Klimawandel	Folgen	25	967

Tabelle 19 Verweildauer in Abhängigkeit von Schlagwort und Suchbegriff TOP 2

Top3			
Suchbegriff	Schlagwort	Tage	Vorkommen
Annalena Baerbock	Grünen	29	5.092
Armin Laschet	CDU	29	5.903
Olaf Scholz	SPD	29	5.394
Kanzlerkandidaten	Bundestagswahl	29	6.897
Koalition	2021	29	4.849
Alice Weidel	Weidel	29	8.172
Bernd Riexinger	Die Linke	10	104
Franziska Brantner	European Green Deal	14	29
Michael Theurer	Fraktionsvize	7	124
Saskia Esken	Saskia Esken	29	5.678
Wolfgang Schäuble	CDU	29	2.531
Baden-Württemberg	BW	29	2.632
Bosch	Amazon	19	2.475
Daimler	BMW	29	1.987
Karlsruhe	POL-KA	29	1.744
Stuttgart	Stuttgarter	28	925
Klimawandel	Klimawandels	21	1.569

Tabelle 20 Verweildauer in Abhängigkeit von Schlagwort und Suchbegriff TOP 3

2.5.3 Häufigste Schlagwörter unter den Top 3 (#11)

AUFGABENSTELLUNG DER FRAGE:

In der Frage aus Abschnitt 2.5.3 wird untersucht, welche die häufigsten aus den Schlagzeilen extrahierten Schlagwörter der Ränge 1-3 in einer Gesamtdarstellung sind. Es soll hierbei keine Unterscheidung nach Suchbegriffen vorgenommen werden.

ABFRAGETECHNIK

Wie zuvor wird erneut eine WordCloud zur Darstellung der Ergebnisse genutzt. Eine Gruppierung der Daten nach einem Suchbegriff (wie in Abschnitt 2.5.1 Zusammenhang zwischen Schlagwörtern und Suchbegriffen (#5)) erfolgt nicht. Die Wörter stammen aus der Gesamtmenge der Schlagzeilen, lediglich gefiltert nach Schlagzeilen der Ränge 1-3.

ERGEBNIS DER ANALYSE:

Mehrere Schlagworte treten mit ähnlicher Schriftgröße auf. Bis auf dem Wort „Bundestagswahl", welches am größten auftritt, gibt es keine eindeutige Tendenz von semantisch oder linguistische ähnlichen Begriffen.

Abbildung 64: Schlagwörter unter den Top 3

2.6 Lokalisierungen Baden-Württemberg

Im Gegensatz zu den in den Abschnitten 2.1-2.5 untersuchten Fragestellungen wird abschließend in Kapitel 2.6 der Fokus der Untersuchungen auf die Region Baden-Württemberg gerichtet.

Zu beachten ist hierbei auch, dass sich dieses spezielle Interessensgebiet auch in der Wahl der Suchbegriffe widerspiegelt: so haben 11 der insgesamt 17 untersuchten Suchbegriffe einen klaren Baden-Württemberg Bezug.

2.6.1 Prozentualer Anteil von Quellen aus Baden-Württemberg (#13)

AUFGABENSTELLUNG DER FRAGE:

In diesem Abschnitt wird der Frage nach dem Anteil an Schlagzeilen von Quellen aus Baden-Württemberg im Vergleich zu allen Schlagzeilenquellen nachgegangen.

ABFRAGETECHNIK

Ein näherer Fokus auf Quellen mit Bezug zu Baden-Württemberg wird durch einen Vergleich mit allen Daten der Betrachtung gesetzt. Die Vorgehensweise zur Feststellung des Bezugs von Quellen zu Baden-Württemberg wurde bereits in den Begriffsdefinitionen erläutert.

ERGEBNIS DER ANALYSE:

Mit dem folgenden einfachen Kreisdiagramm ist zu erkennen, dass Medien mit Bezug zu Baden-Württemberg knapp ein Viertel der gesamten Schlagzeilen der erhobenen Analysemenge ausmachen.
Eine weitere Unterteilung nach den Abfragestandorten Stuttgart und Frankfurt zeigt, dass diese in etwa gleich große Anteile an den Quellen aus Baden-Württemberg haben. Es gibt also keine Anzeichen für eine regionale Bevorzugung der Anteile.

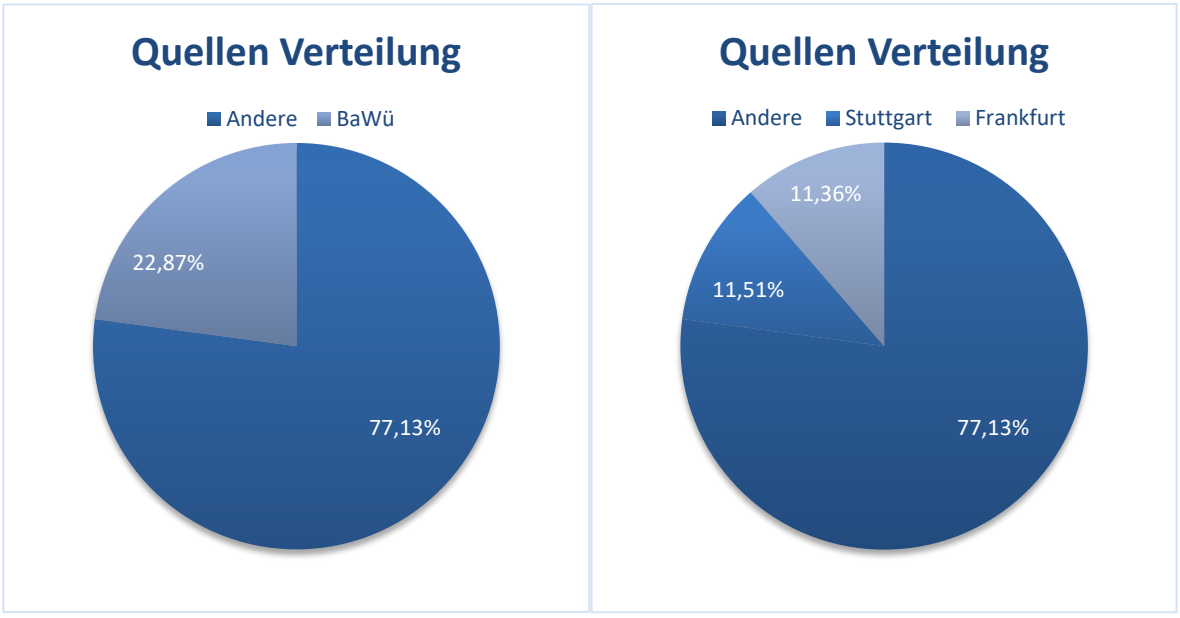


Abbildung 65 Quellen Verteilung BaWü und Lokationen

Von diesem Anteil an Quellen mit Bezug zu Baden-Württemberg wird der jeweilige Anteil der stärksten zehn Quellen in der nachstehenden Grafik präsentiert. Hierbei kann ein Abfall nach den ersten fünf Quellen erkannt werden. Anschließend sind nur noch Anteile unter einem Prozentpunkt vertreten, mit weiter fallender Tendenz der übrigen (nicht repräsentierten) Quellen.

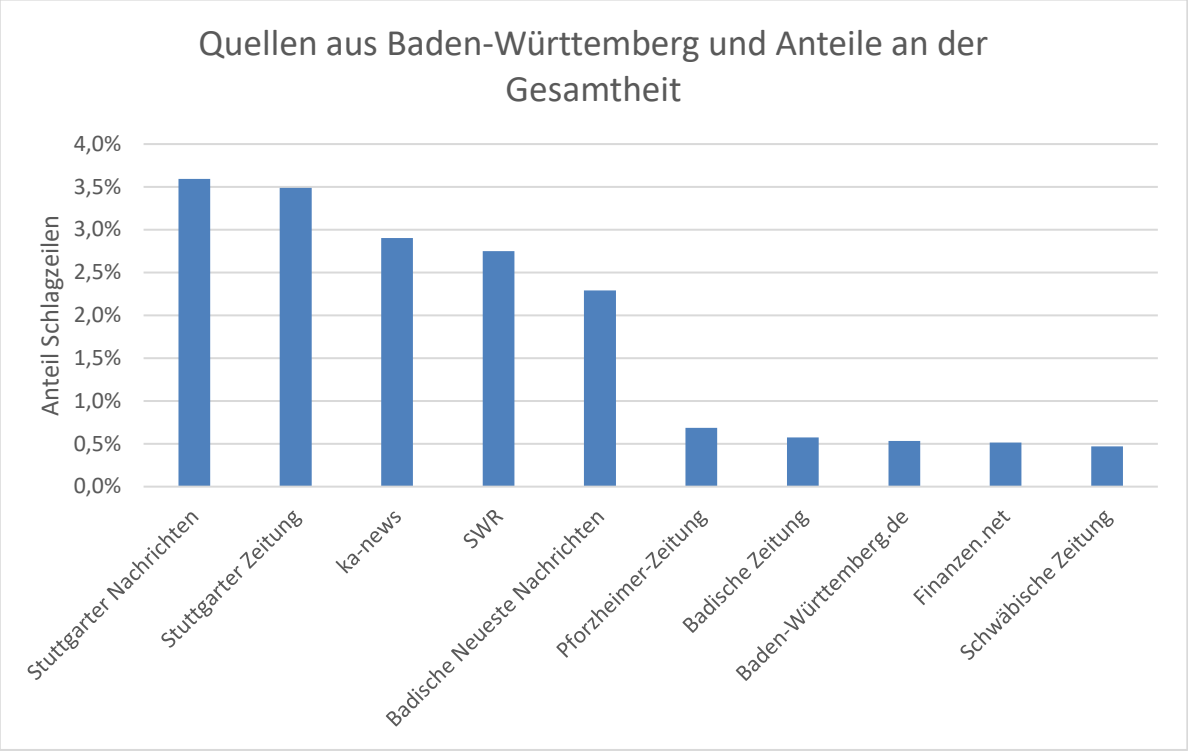


Abbildung 66 Anteil der Quellen aus BaWü

Abschließend wird eine detailliertere Betrachtung auf die Anteile an Schlagzeilen nach Suchbegriffen geboten. Aus allen Schlagzeilen, ohne weitere Einschränkung, wurden die Schlagzeilen von Quellen mit Bezug zu Baden-Württemberg gewertet. Diese Daten wurden jedoch nach den Suchbegriffen gruppiert und die Information von welcher Quelle die Schlagzeile stammt wird nicht weiter berücksichtigt. Die Menge der im folgenden Säulendiagramm dargestellten Anteile entspricht dem Anteil „BaWü" aus dem linken Kreisdiagramm aus Abbildung 65. Die Säulen sind in absteigender Reihenfolge sortiert dargestellt.

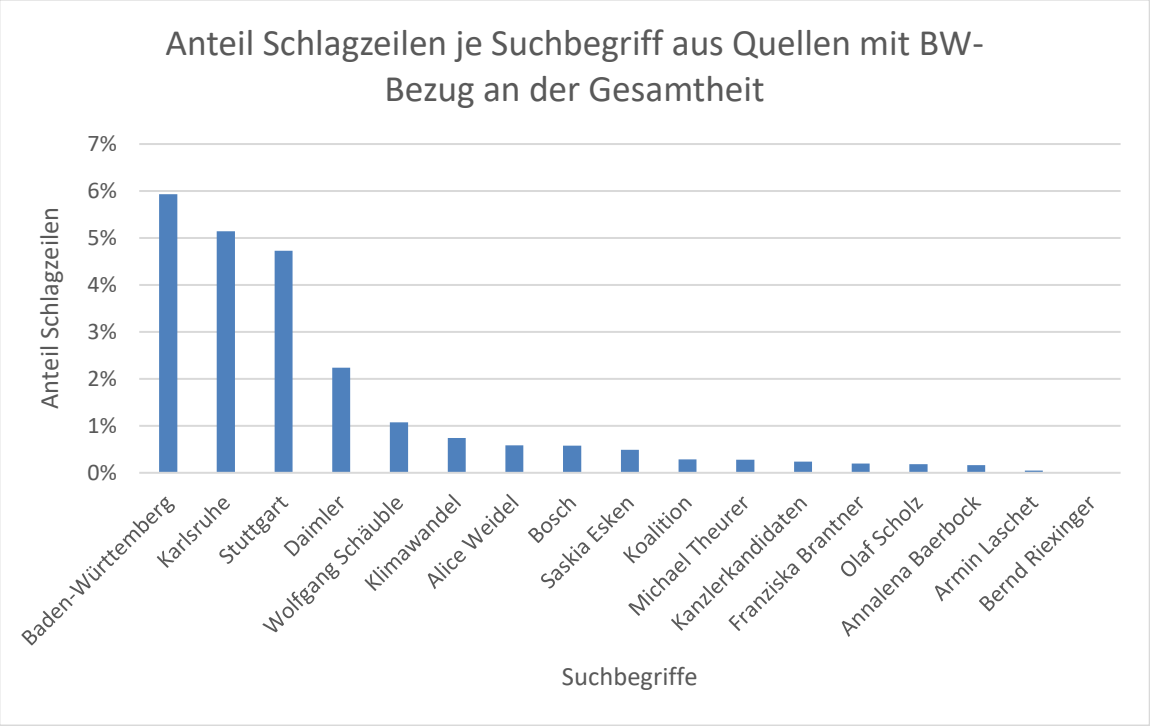


Abbildung 67 Anteil Schlagzeilen je Suchbegriff aus Quellen mit BaWü-Bezug an der Gesamtheit

2.6.2 Rang innerhalb der Gesamtrangverteilung (#14)

AUFGABENSTELLUNG DER FRAGE:

Beantwortet werden soll die Frage, welchen Platz Quellen aus Baden-Württemberg bei dem Rang der Schlagzeilen pro Suchanfrage einnehmen.

ABFRAGETECHNIK

Für Quellen mit Baden-Württemberg-Bezug wurde über alle zugehörigen Schlagzeilen ein durchschnittlicher Rang gebildet. Die Daten werden aufsteigend sortiert und als Säulen dargestellt. Bei dieser Betrachtung ist es zunächst ohne Einfluss, ob es sich dabei um eine starke, etablierte Quelle handelt oder um einen eher kleinen Herausgeber. Zur besseren Einordnung nach diesem Kriterium wurden zusätzlich zum durchschnittlichen Rang auch die absoluten Anzahlen der Schlagzeilen (pro 1.000 Stück) als danebenstehende Säulen bereitgestellt. Beide Werte (der Rang als auch die Schlagzeilen je 1.000 Stück) teilen sich dieselbe Skalierung der Y-Achse.

ERGEBNIS DER ANALYSE:

Die durchschnittlich (numerisch) kleineren Ränge erzielen vor allem kleinere Herausgeber. Größere Quellen platzieren sich eher im vorderen Mittelfeld. Viele weitere kleine Herausgeber liegen aber auch auf den hinteren Platzierungen. Eine klare Aussage, dass kleine Quellen ihre Schlagzeilen prinzipiell auf besseren Rängen platzieren, kann nicht getroffen werden. Im Umkehrschluss kann an dieser Stelle keine Bevorzugung größerer, etablierter Quellen festgestellt werden. Zu beachten ist an dieser Stelle, dass der Fokus hier auf Quellen mit Baden-Württemberg-Bezug liegt. Eine ähnliche Darstellung könnte für eine Gesamtbetrachtung ohne Einschränkung nach einem Baden-Württemberg-Bezug ebenfalls von Interesse sein.
In der Frage #18 (siehe Kapitel 2.6.3 Verhältnis des Vorkommens von BaWü Quellen (#18)) kann diese Darstellung mit einer Sortierung nach der Häufigkeit der Schlagzeilen wiedergefunden werden, was eine andere Blickweise ermöglicht.

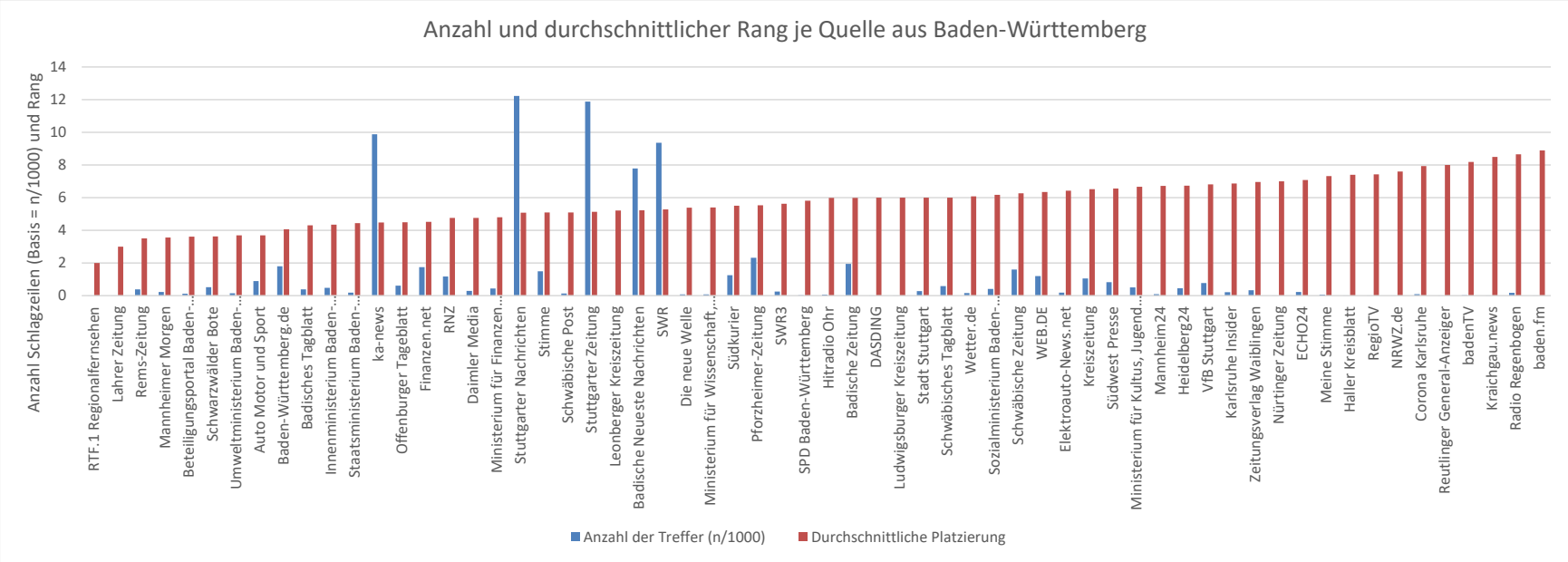


Abbildung 68 durchschnittlicher Rang und Anzahl der Quellen aus BaWü

2.6.3 Verhältnis des Vorkommens von BaWü Quellen (#18)

AUFGABENSTELLUNG DER FRAGE:

Es soll beantwortet werden, wie das Verhältnis zwischen den Quellen mit Baden-Württemberg-Bezug in Kontext der Gesamtanzahl Schlagzeilen ist.

ABFRAGETECHNIK

Dies ist eine anders fokussierte Darstellung der Grafik aus Frage #14. Für Quellen mit Baden-Württemberg-Bezug wird die Anzahl der Schlagzeilen gezählt und nach diesen Werten absteigend sortiert. Die Werte werden in Säulen abgebildet. Parallel wird eine zweite Säule jeweils daneben gezeigt, die den durchschnittlichen Rang dieser Schlagzeilen beschreibt. Beide Werte (der Rang als auch die Schlagzeilen je 1.000 Stück) teilen sich dieselbe Skalierung der Y-Achse.

ERGEBNIS DER ANALYSE:

In dieser Darstellung zeigt sich optisch zunächst eine eher zufällige Verteilung der Ränge in Abhängigkeit der Anzahl der Schlagzeilen. Zuvor wurde bereits darauf hingewiesen, dass nur fünf Quellen den Großteil der Schlagzeilen ausmachen und übrigen Quellen mit einer abnehmenden Tendenz nur schwach zur Anzahl der Schlagzeilen beitragen. Diese Tendenz lässt sich hier sehr gut bestätigen, ebenso wie die Dominanz der fünf Quellen.

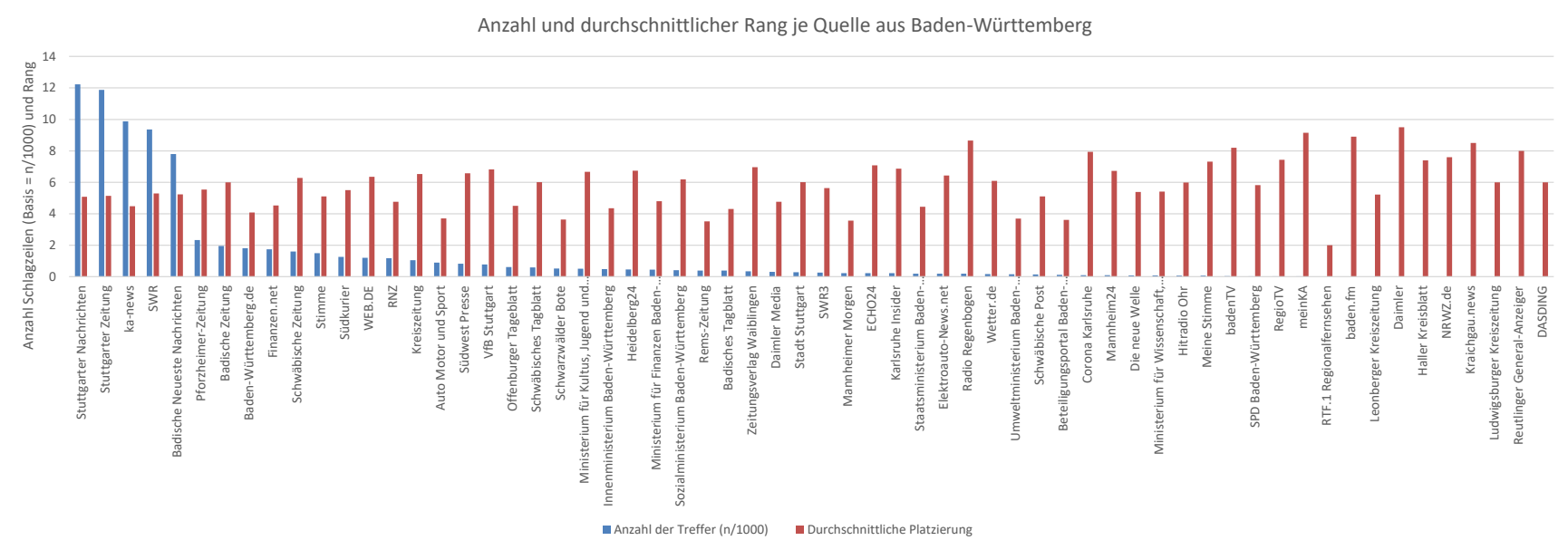


Abbildung 69 Anzahl der Quellen aus BaWü und durchschnittlicher Rang

2.6.4 Quellen aus Baden-Württemberg auf Rängen 1-3? (#14a)

AUFGABENSTELLUNG DER FRAGE:

Bei dieser Frage wird untersucht, welche der Quellen aus Baden-Württemberg auf den Rängen 1-3 zu finden sind.

ABFRAGETECHNIK

Neben der Filterung der Schlagzeilen nach den Rängen 1-3 sowie der Gruppierung der Schlagzeilen nach Quellen mit Baden-Württemberg-Bezug wird auch parallel eine Unterscheidung zwischen den Abfragestandorten Stuttgart und Frankfurt vorgenommen. Für die Darstellung werden lediglich die zehn stärksten Quellen herangezogen, gemessen am Gesamtbestand. Für diese zehn Quellen wurden dann zusätzlich die Anteile für die Standortunterscheidung hinzugefügt. Die Summe der beiden Werte aus den Standorten beträgt ebenso viel wie der Gesamtwert.

ERGEBNIS DER ANALYSE:

Während in Frage #13 (Betrachtung der Ränge 1-10, siehe Kapitel 2.6.1 Prozentualer Anteil von Quellen aus Baden-Württemberg (#13)) die ka-news nur Platz drei belegten, positionieren sie sich bei der Betrachtung der Ränge 1-3 nun ganz vorne. Auch hier bleiben die ersten fünf Quellen stark vertreten, wonach nur noch Quellen mit geringeren Anteilen vorkommen. Bei der Unterscheidung zwischen den Abfragestandorten Stuttgart und Frankfurt zeigen sich nahezu keine Unterschiede, mit Ausnahme der Quelle „Baden-Württemberg.de", welche größtenteils am Standort Stuttgart ermittelt wurde.

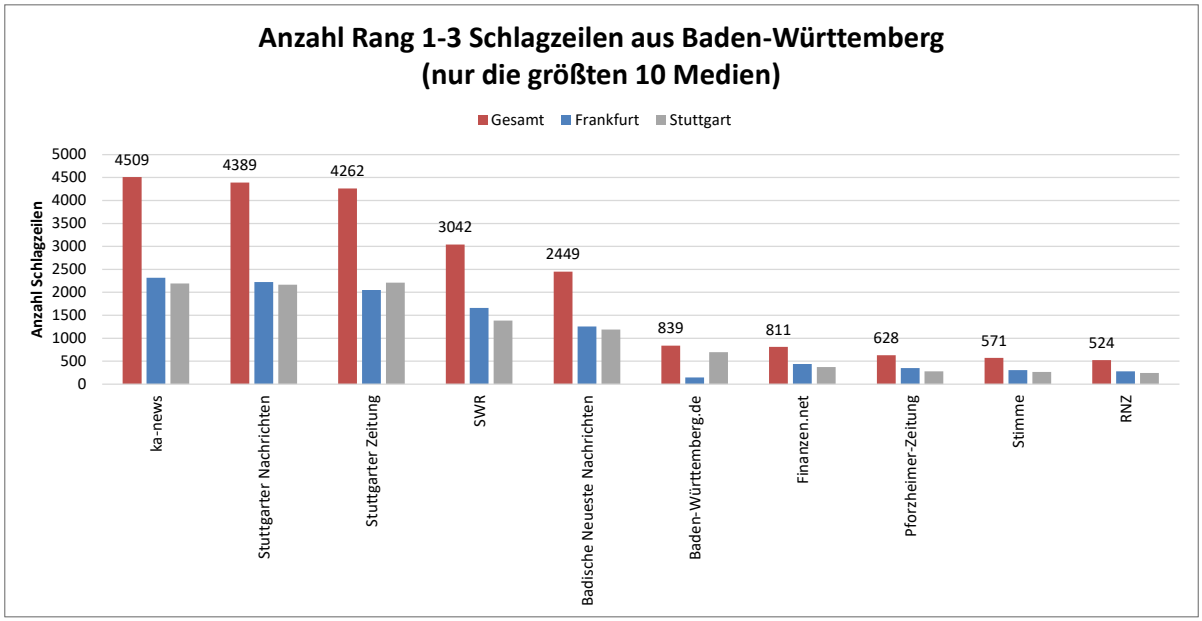


Abbildung 70 Anzahl Schlagzeilen Baden-Württemberg auf Rang 1-3

2.6.5 Rang innerhalb der Gesamtrangverteilung, abhängig von Abfrageort (#14b)

AUFGABENSTELLUNG DER FRAGE:

Es soll analysiert werden, wie sich die Verteilung der Ränge (abhängig von Lokation) darstellt.

ABFRAGETECHNIK

Zunächst erfolgt die Abfrage mit den durchschnittlichen Rängen je Quelle mit Baden-Württemberg-Bezug, sowie mit Unterscheidung nach den Abfragestandorten Stuttgart und Frankfurt. Danach erfolgt eine Darstellung der Verteilung der absoluten Ränge (also ohne Durchschnittsbildung je Quelle) und ohne Standortunterscheidung.
Für die Darstellungen werden sogenannte Box-Plots verwendet.

ERGEBNIS DER ANALYSE:

Bei der Betrachtung der drei Boxplots zu der Verteilung der durchschnittlichen Ränge nach Gesamt, Stuttgart und Frankfurt können nur sehr geringe Unterschiede festgestellt werden. In Frankfurt sind die Ränge durchschnittlich nur minimal schlechter gestellt als in Stuttgart, was sich schließlich in der Gesamtbetrachtung widerspiegelt. Eine signifikante Bevorzugung bestimmter Lokationen lässt sich somit nicht feststellen. Der Gros der Ränge nimmt im Schnitt Werte zwischen 4,77 und 6,86 ein, was eine sehr mittige Darstellung abbildet bei Wertebereichen von 1-10.

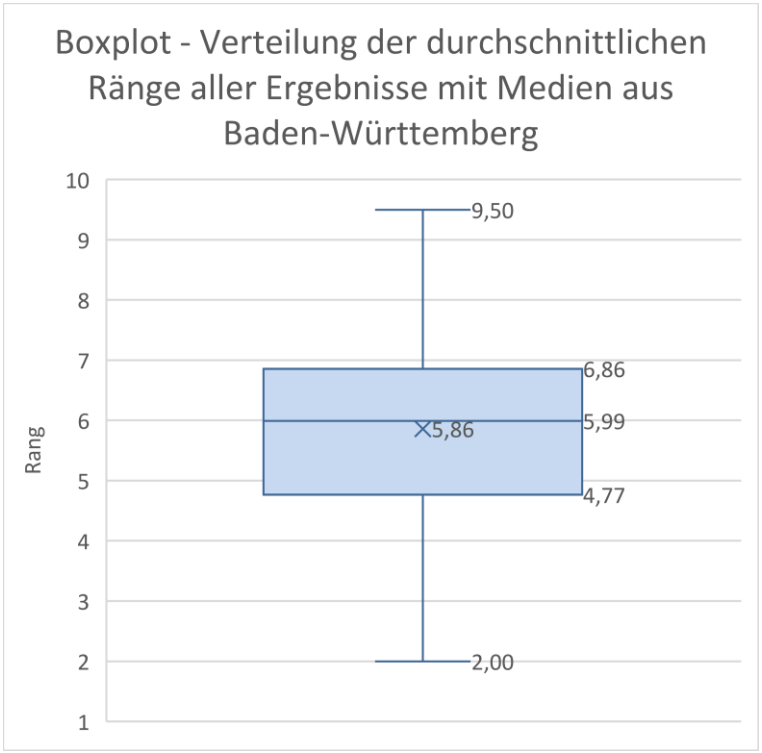


Abbildung 71 Abbildung 62 BoxPlot – Durchschnittliche Ränge aller Ergebnisse von Medien aus BaWü

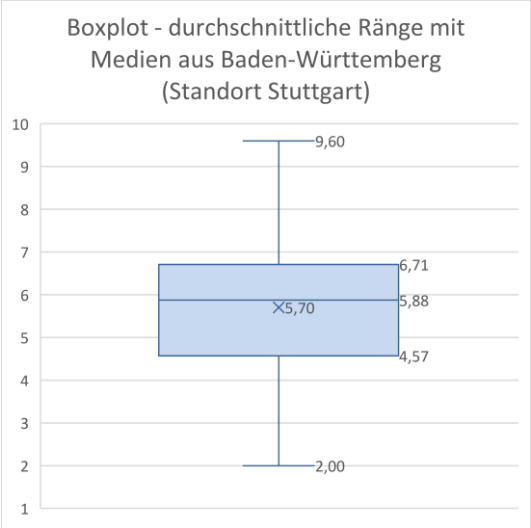


Abbildung 72 Abbildung 63 BoxPlot - Durchschnittliche Ränge aller Ergebnisse von Medien aus BaWü / Standort STG

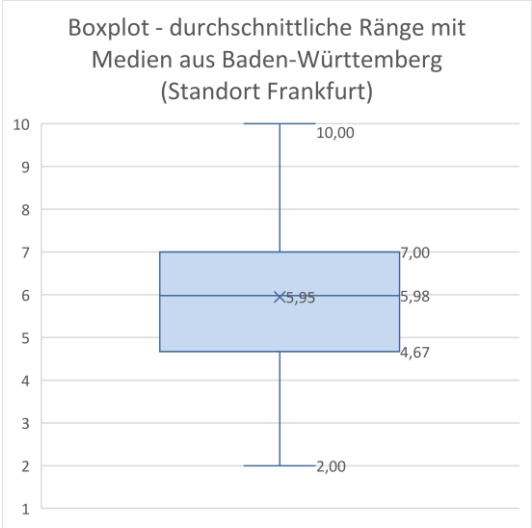


Abbildung 73 Abbildung 64 BoxPlot - Durchschnittliche Ränge aller Ergebnisse von Medien aus BaWü / Standort FRA

Auch durch Betrachtung der absoluten Rang-Werte ohne Durchschnittsbildung ergibt sich kein signifikant anderes Bild. Lediglich vergrößert es die Rangverteilung des Gros zwischen 3 und 8. Der Median von 5 deutet eine sehr leichte Tendenz zu niedrigeren Rängen.
Aus einer sehr leichten Tendenz könnte man schließen, dass Schlagzeilen aus Baden-Württemberg eher niedrigere Ränge aufweisen. Eine weitere Unterscheidung nach den Standorten wurde hier nicht vorgenommen, da sie keine zusätzlichen Erkenntnisse liefert.

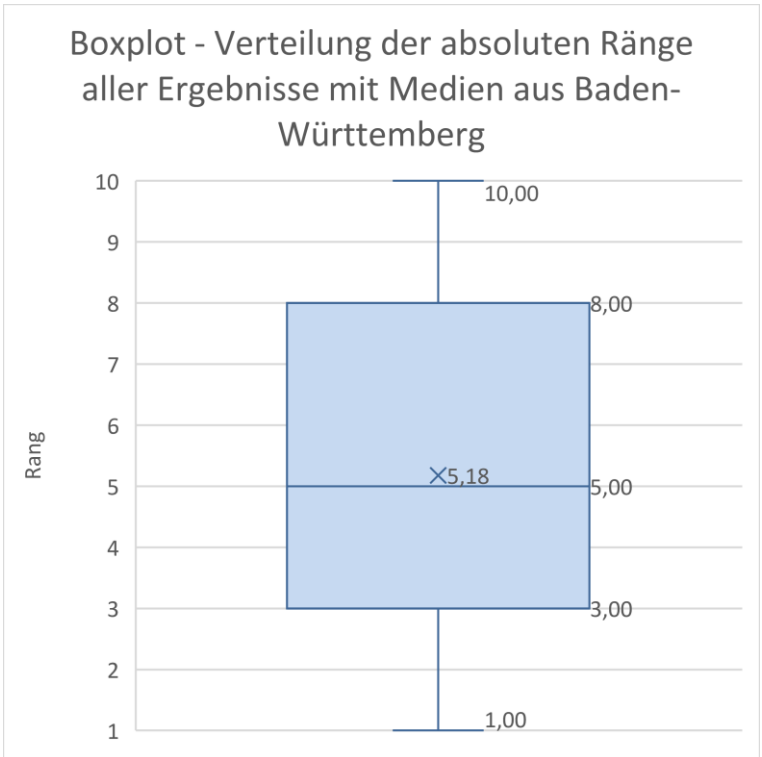


Abbildung 74 BoxPlot – absolute Ränge aller Ergebnisse von Medien aus BaWü

2.6.6 Vorkommen von "BaWü-Orten" in Schlagzeilen (#21)

AUFGABENSTELLUNG DER FRAGE:

In diesem Kapitel wird der Fragestellung nachgegangen, ob in Schlagzeilen von Quellen aus Baden-Württemberg ein besonderer Bezug zu Ortschaften aus Baden-Württemberg gegeben ist.

ABFRAGETECHNIK

Alle Schlagzeilen wurden auf solche mit einer Quelle mit Bezug zu Baden-Württemberg gefiltert. Für diese Schlagzeilen wurden die Schlagwörter, die im Rahmen der Entitätsanalyse ermittelt wurden, näher betrachtet. Eine kundenseitig bereitgestellte Liste mit Ortschaften aus Baden-Württemberg wurde zum Vergleich genutzt. Weiter wurde nun abgeglichen, ob eine dieser Ortschaften namentlich mit einem der Schlagwörter aus den betrachteten Schlagzeilen übereinstimmt. Für jede Übereinstimmung wurde gezählt, wie oft diese für eine bestimmte Quelle und ein bestimmtes Schlagwort vorliegt. Sprich die übereinstimmenden Daten wurden nach Quelle und Schlagwort gruppiert und die Anzahlen gewertet.

ERGEBNIS DER ANALYSE:

Für die Quellen Finanzen.net sowie Baden-Württemberg.de konnten keine Vorkommen von Ortschaften in Titeln von Schlagzeilen gefunden werden. Beispielsweise gibt es bei Baden-Württemberg.de sehr viele Schlagzeilen zu insbesondere Corona-Regelungen, jedoch ist bei keiner der Schlagzeilen der Name einer Ortschaft aus Baden-Württemberg vertreten.
Die Stuttgarter Nachrichten haben in all ihren Schlagzeilen insgesamt 2.934 mal das Wort Stuttgart erwähnt. Dagegen tritt das Wort Karlsruhe mit 281 Vorkommen deutlich seltener. In Zeile zwei zeigt sich, dass die Stuttgarter Zeitung mit 4.542 Wortvorkommen deutlich häufiger das Wort Stuttgart in den Titeln der Schlagzeilen verwendet.

In der tabellarischen Darstellung sowie in dem Säulendiagramm werden die stärksten zehn Quellen in absteigender Reihenfolge nach ihrer Anzahl Schlagzeilen aufgelistet. Die in den Darstellungen angegeben Anzahlen beziehen sich hingegen nicht auf die Anzahl Schlagzeilen, sondern auf die Anzahl Vorkommen von Ortschaften in den Titeln der Schlagzeilen. Siehe dazu auch Frage 2.6.1 Prozentualer Anteil von Quellen aus Baden-Württemberg (#13).

Top1_Quelle	Top1 Ort	Top1 Anzahl	Top2 Ort	Top2 Anzahl	Top3 Ort	Top3 Anzahl
Stuttgarter Nachrichten	Stuttgart	3001	Karlsruhe	298	Calw	203
Stuttgarter Zeitung	Stuttgart	4786	Karlsruhe	184	Weilheim	105
ka-news	Karlsruhe	2674	Pforzheim	140	Eggenstein-Leopoldshafen	58
SWR	Karlsruhe	1579	Stuttgart	712	Pforzheim	254
Badische Neueste Nachrichten	Karlsruhe	3432	Rastatt	619	Ettlingen	316
Pforzheimer-Zeitung	Karlsruhe	265	Stuttgart	158	Pforzheim	71
Badische Zeitung	Offenburg	515	Bad Säckingen	21	-	0
Baden-Württemberg.de	-	0	-	0	-	0
Finanzen.net	-	0	-	0	-	0
Schwäbische Zeitung	Stuttgart	103	Karlsruhe	70	Offenburg	70

Tabelle 21 Vorkommen von "BaWü-Orten" in Schlagzeilen

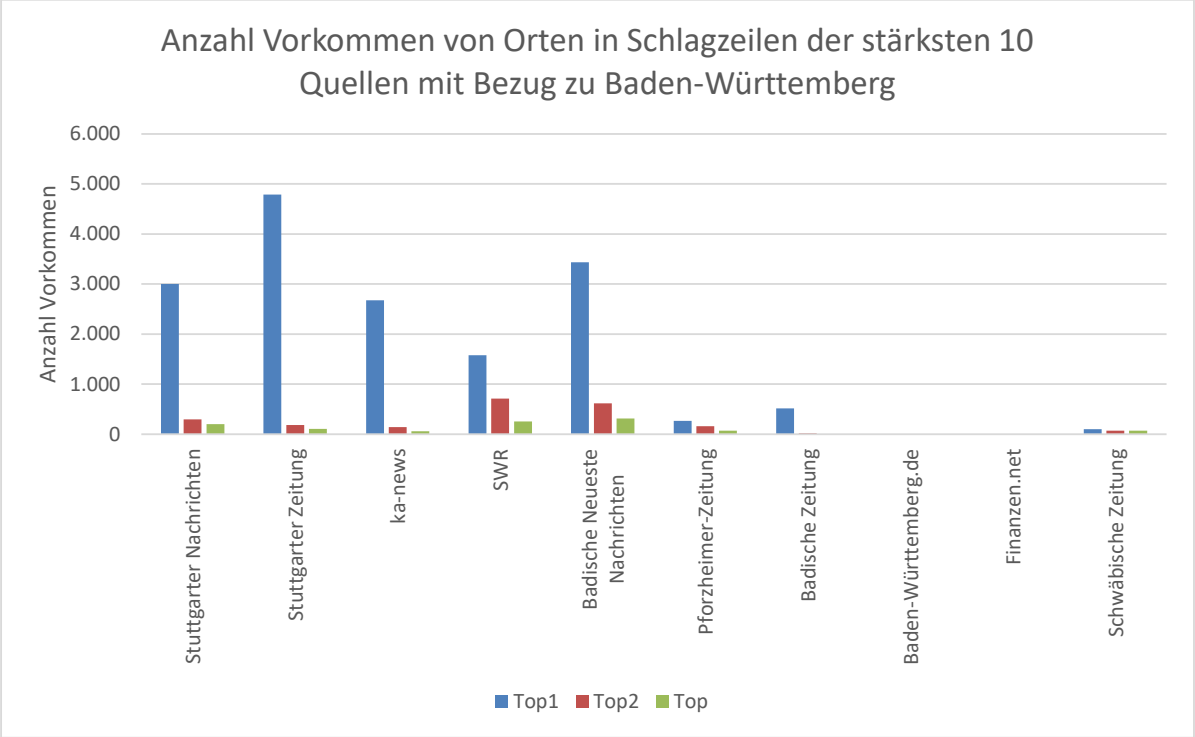


Abbildung 75 Anzahl Vorkommen von Orten in Schlagzeilen der stärksten 10 Quellen mit Bezug zu BaWü (Orte Top1-3 siehe Tabelle 21)

Vorherige Darstellung betrachtete nur die 10 stärksten Quellen. Die nachstehende Grafik zeigt jedoch die 20 häufigsten Kombinationen einer Quelle mit dem Namen einer Ortschaft aus Baden-Württemberg (sortiert nach der Anzahl Vorkommen, ohne Berücksichtigung der stärksten zehn Quellen).
Quellen, die durch diese neue Darstellung nun ebenfalls Beachtung finden sind RNZ, Offenburger Tageblatt, Mannheimer Morgen, Südkurier, Stadt-Stuttgart sowie Radio Regenbogen. Wie zuvor bereits festgestellt, enthalten die beiden Quellen Finanzen.net sowie Baden-Württemberg.de keine Erwähnungen von Ortschaften, sodass diese beiden Quellen in der folgenden Darstellung nicht enthalten sind.

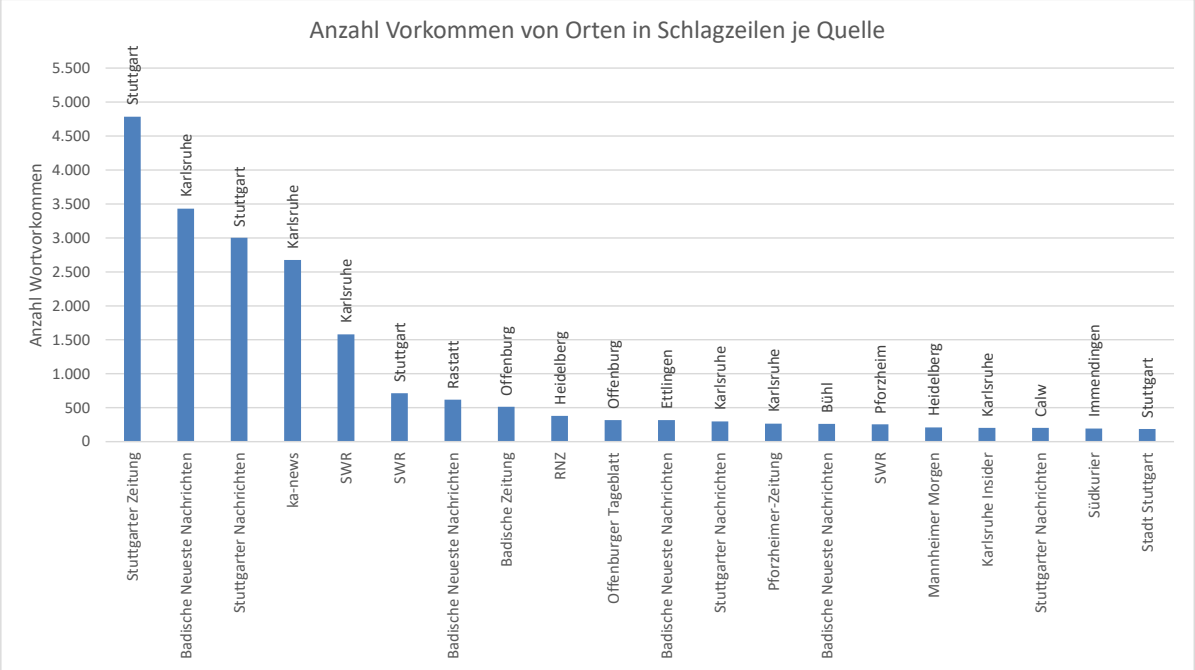


Abbildung 76 Anzahl Vorkommen von BaWü Orten in Schlagzeilen je Quelle

2.6.7 Welche Schlagwörter treten am häufigsten auf (#9)

AUFGABENSTELLUNG DER FRAGE:

Eine Darstellung soll zeigen, welche Schlagwörter am häufigsten auftreten für eine Betrachtung insgesamt, sowie für Schlagzeilen von Quellen mit Baden-Württemberg-Bezug.

ABFRAGETECHNIK

Zuerst wird eine WordCloud über alle Schlagzeilen ohne Einschränkung erzeugt. Danach eine weitere WordCloud berechnet für Schlagzeilen, die von Quellen stammen, welche einen Bezug zu Baden-Württemberg aufweisen.

ERGEBNIS DER ANALYSE:

Die WordCloud über alle Schlagzeilen zeigt einen Fokus auf den Begriff „Bundestagswahl" und einige Namen von Kandidaten oder Parteien. In der WordCloud aus der Frage #11 (siehe Kapitel 2.5.3 Häufigste Schlagwörter unter den Top 3 (#11)), welche nur die Schlagzeilen der Ränge 1-3 betrachtete, können sehr ähnliche Wortgrößen entdeckt werden.

Abbildung 77 WordCloud von Schlagwörtern über alle Schlagzeilen ohne Einschränkung

Eine deutliche Unterscheidung zeigt hingegen die WordCloud mit Baden-Württemberg-Bezug. Große Schlagwörter sind hier „Baden-Württemberg", „Karlsruhe", „Stuttgart" und „Daimler". Namen von Personen und Parteien erscheinen dagegen kleiner als in der Darstellung zuvor. Daraus lässt sich schlussfolgern, dass Quellen mit Baden-Württemberg-Bezug auch vorrangig über Themen mit regionalem-Bezug berichten.

Abbildung 78 WordCloud von Schlagwörtern über alle Schlagzeilen mit BaWü Bezug

2.7 Untersuchungen zur Berechnung der Ungleichverteilung

Ein signifikantes Ergebnis bei der Untersuchung der Fragestellungen zu den Quellen (siehe Abschnitt 2.1) war die deutliche Ungleichverteilung in Bezug auf die Anzahl der zurückgelieferten Schlagzeilen: so liefern wenige Quellen den überwiegenden Anteil an Schlagzeilen, wohingegen eine Vielzahl an Quellen im Vergleich dazu eine sehr niedrige Zahl an Schlagzeilen beisteuert. Treffend beschrieben wird das Phänomen der Ungleichverteilung durch die zugehörige Lorenzkurve und den Gini Koeffizienten.

Allerdings gibt es bei der Berechnung in Bezug auf die Wahl

- der Suchterme (alle Suchterme, Gruppen von Suchtermen, einzelne Suchterme),
- des betrachteten Zeitraums (Tage, Wochen, Gesamtzeitraum) oder
- des Schlagzeilenfensters (nur die ersten 3 oder alle 10 Schlagzeilen)

eine erhebliche Anzahl an Freiheitsgraden.

2.7.1 Allgemeine Beobachtungen

AUFGABENSTELLUNG DER FRAGE:

Einleitend wird untersucht, welchen Einfluss die Anzahl der Schlagzeilen und Quellen auf den zeitlichen Verlauf des Gini Koeffizienten ausüben.

ABFRAGETECHNIK

Für jeden Tag des Beobachtungszeitraum werden die zurückgelieferten Schlagzeilen in Teilmengen aufgeteilt und für jede Teilmenge der zugehörige Gini Koeffizient hinsichtlich der Quellen berechnet und gemeinsam mit den jeweiligen Werten der Anzahl der Schlagzeilen und der Quellen in einem Diagramm dargestellt.

In verallgemeinerter Form läßt sich die Vorgehensweise, wie in Abbildung 79 dargestellt, beschreiben.

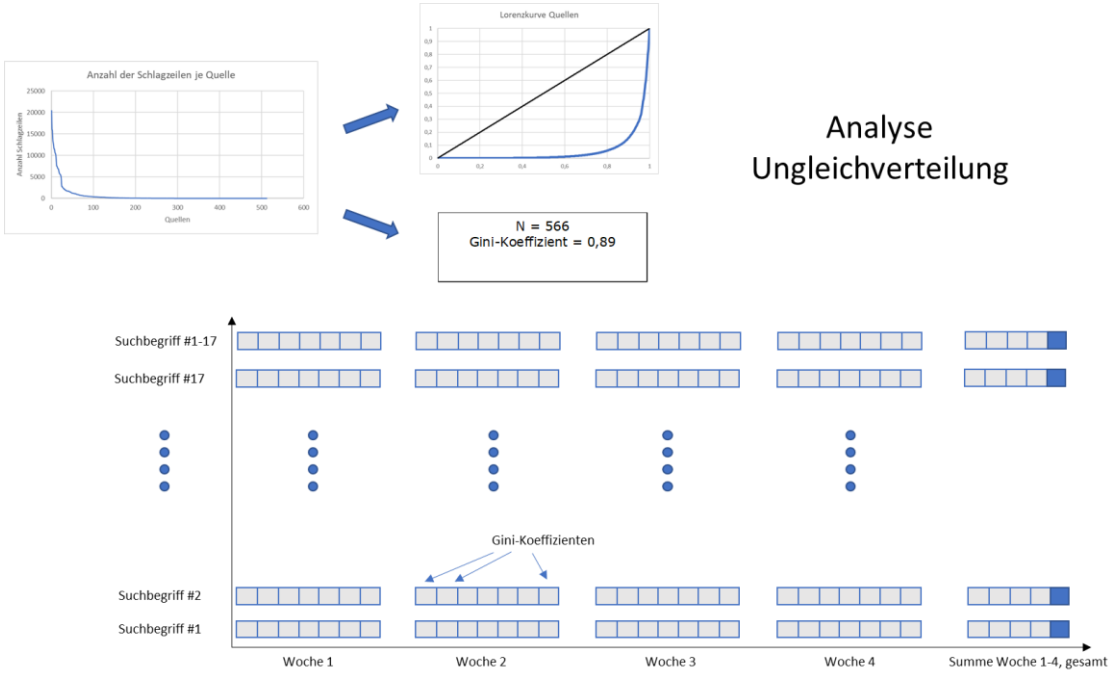


Abbildung 79 Vorgehen bei der Untersuchung der Ungleichverteilung mit Hilfe von Lorenzkurve und Gini Koeffizienten

ERGEBNIS DER ANALYSE:

Die zeitlichen Verläufe von Gini Koeffizient, Anzahl Schlagzeilen und Quellen sind in der Abbildung 80 dargestellt. In einigen Zeitbereich zeigt sich eine Korrelation zwischen der Kurvenverläufen (abnehmende Anzahl von Schlagzeilen und Quellen führt auch zu einem abnehmenden Gini Koeffizienten und umgekehrt), allerdings gibt es auch Bereiche, in denen sich z.B. ein starkes Abnehmen der Anzahl an Schlagzeilen nicht auf den Gini Koeffizienten auswirkt (Beispiel: 13.9., hier liegt vermutlich eine Kompensation durch den Anstieg der Anzahl der Quellen vor).

Einflüssen der Anzahl von Schlagzeilen und Quellen können also nur zeitweise ein Ab- oder Zunehmen des Gini Koeffizienten erklären.

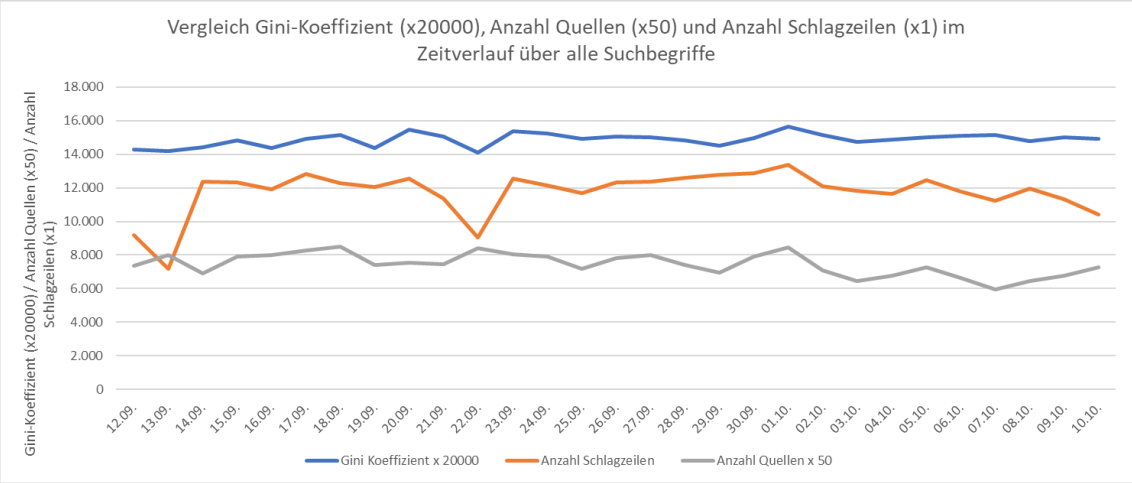


Abbildung 80 zeitlicher Verlauf des Gini Koeffizienten und der Anzahl der Schlagzeilen und der Quellen, tageweise Betrachtung

2.7.2 Einfluss des Betrachtungszeitraums

AUFGABENSTELLUNG DER FRAGE:

In diesem Kapitel wird der Frage nachgegangen, welchen Einfluss die Wahl des Betrachtungszeitraums auf die berechneten Gini Koeffizienten ausübt.

ABFRAGETECHNIK

Für jeden Tag bzw. jede Woche des Beobachtungszeitraum werden die zurückgelieferten Schlagzeilen in Teilmengen aufgeteilt, und für jede Teilmenge der zugehörige Gini Koeffizient hinsichtlich der Quellen berechnet und mit den Gini Koeffizienten des Gesamtzeitraums verglichen. Dabei erfolgt zusätzlich eine Unterscheidung in Bezug auf die Schlagzeilen aller Suchbegriffe und die Schlagzeilen, die sich für die Suchbegriffe „Annalena Baerbock", „Armin Laschet" und „Olaf Scholz" ergeben haben.

ERGEBNIS DER ANALYSE:

Aus dem zeitlichen Verlauf der Abbildung ist erkennbar, dass die Kurve der Gini Koeffizienten bei Betrachtung aller Suchbegriffe deutlich geringere Schwankungen aufweist als die Kurven der 3 Einzelsuchbegriffe. Zudem sind die Werte der Gini Koeffizienten für die Berechnung über alle Suchbegriffe höher als bei den Einzelsuchbegriffen. Der Grund für die beiden Effekte liegt in der Tatsache begründet, dass sich bei der detaillierten Betrachtung in Bezug auf die einzelnen Suchbegriffe der Umfang der jeweiligen Datenmenge stark reduziert und damit die Anzahl der Schlagzeilen, aber auch die Zahl der Quellen deutlich abnimmt (siehe Tabelle 22 mittlere Anzahlen Schlagzeilen und Quellen).

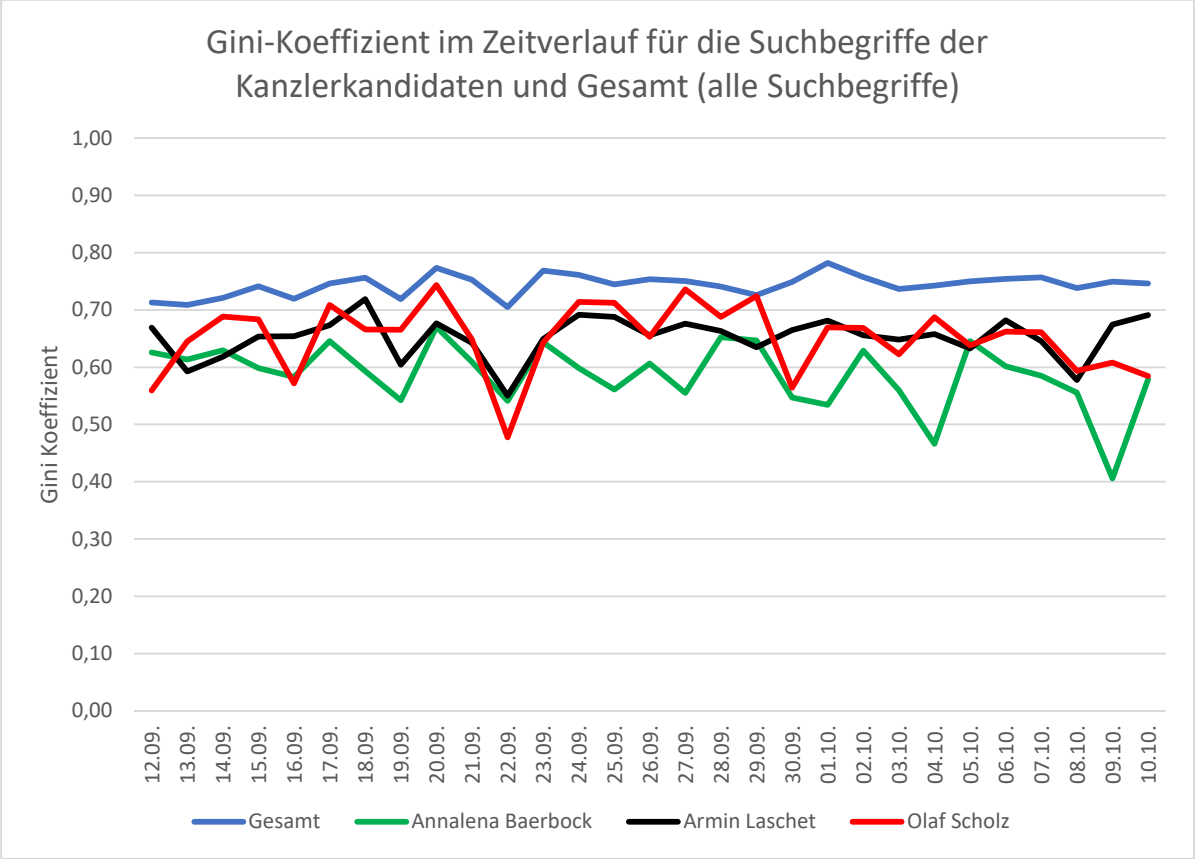


Abbildung 81 zeitlicher Verlauf Gini Koeffizient für alle und für ausgewählte Suchbegriffe (hier die 3 Kanzlerkandidaten), tageweise Betrachtung, Top10

Suchbegriff	Mittlere Anzahl Schlagzeilen (pro Tag)	Mittlere Anzahl Quellen (pro Tag)
Alle	11.740	149
Annalena Baerbock	909	33
Armin Laschet	953	38
Olaf Scholz	953	42

Tabelle 22 mittlere Anzahlen Schlagzeilen und Quellen

Alternativ zur täglichen Berechnung des Gini Koeffizienten zeigt die Abbildung 82 den Verlauf, wenn als Zeitraum jeweils eine Woche betrachtet wird und die Berechnung über alle Schlagzeilen der jeweiligen Woche erfolgt. Zusätzlich ist ganz rechts in der Abbildung der jeweilige Gini Koeffizient für den Gesamtbetrachtungszeitraum dargestellt. Erwartungsgemäß nimmt die Schwankungsbandbreite stark ab. Die Gini Koeffizienten für den Gesamtzeitraum sind am größten, da hier die größte Anzahl an Quellen in Berechnung mit eingeht (Tabelle 23).

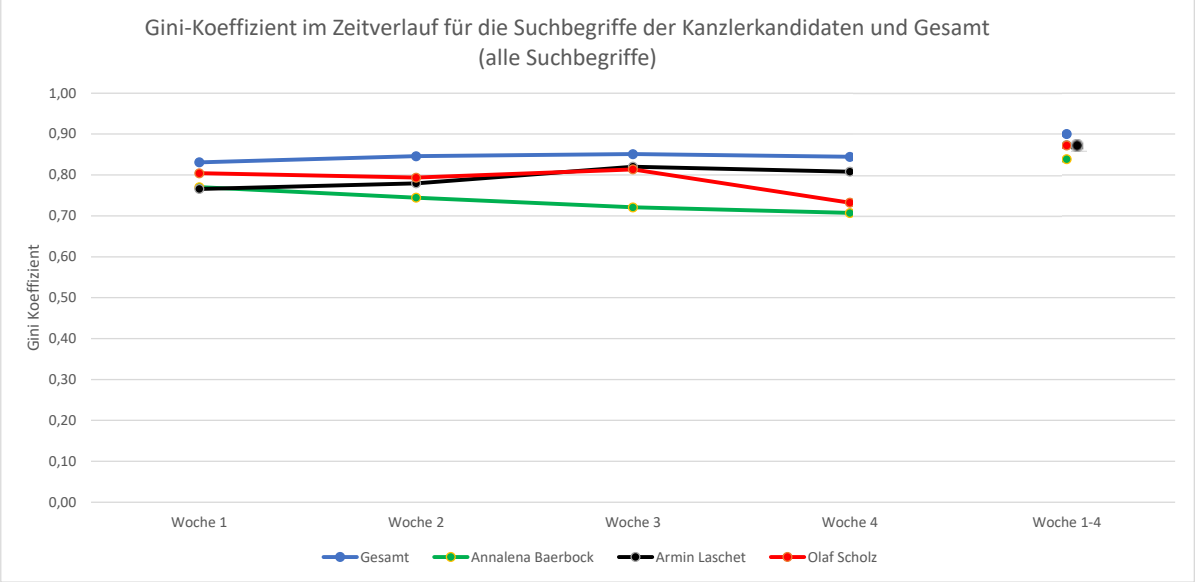


Abbildung 82 zeitlicher Verlauf Gini Koeffizient für alle und für ausgewählte Suchbegriffe (hier die 3 Kanzlerkandidaten), wochenweise und gesamte Betrachtung

Suchbegriff	Mittlere Anzahl Schlagzeilen (pro Woche)	Mittlere Anzahl Quellen (pro Woche)
alle	82.823	296
Annalena Baerbock	6.410	72
Armin Laschet	6.703	84
Olaf Scholz	6.711	97

Tabelle 23 mittlere Anzahlen Schlagzeilen und Quellen

2.7.3 Einfluss Top10 bzw. Top3 Platzierungen der Schlagzeilen

AUFGABENSTELLUNG DER FRAGE:

Abschließend soll geklärt werden, welche Auswirkungen die Betrachtung der Schlagzeilen-/ Quellenverteilung bei Berücksichtigung der Top10 bzw. Top 3 Schlagzeilen auf den Verlauf des Gini Koeffizienten ausübt.

ABFRAGETECHNIK

Für jeden Tag des Beobachtungszeitraum werden die zurückgelieferten Schlagzeilen in Teilmengen aufgeteilt und für jede Teilmenge der zugehörige Gini Koeffizient hinsichtlich der Quellen berechnet und zwar im Gegensatz zu Abschnitt nur für die Top3 Schlagzeilen im Schlagzeilenfenster.

ERGEBNIS DER ANALYSE:

Die Abbildung 84 zeigt analog zur Darstellung in Abbildung 80 den zeitlichen Verlauf von Gini-Koeffizient, Anzahl der Schlagzeilen und Quellen für die Top3 Schlagzeilen.

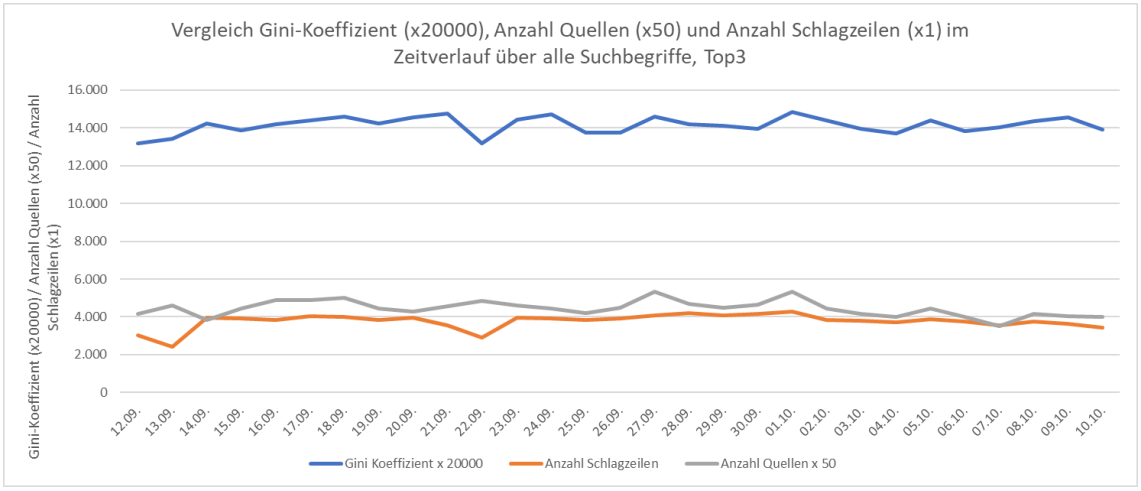
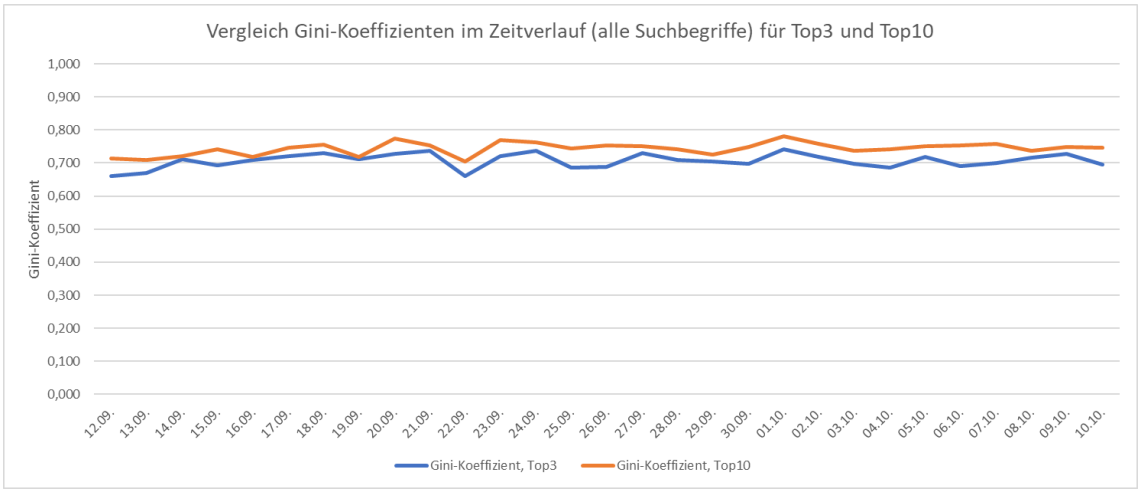
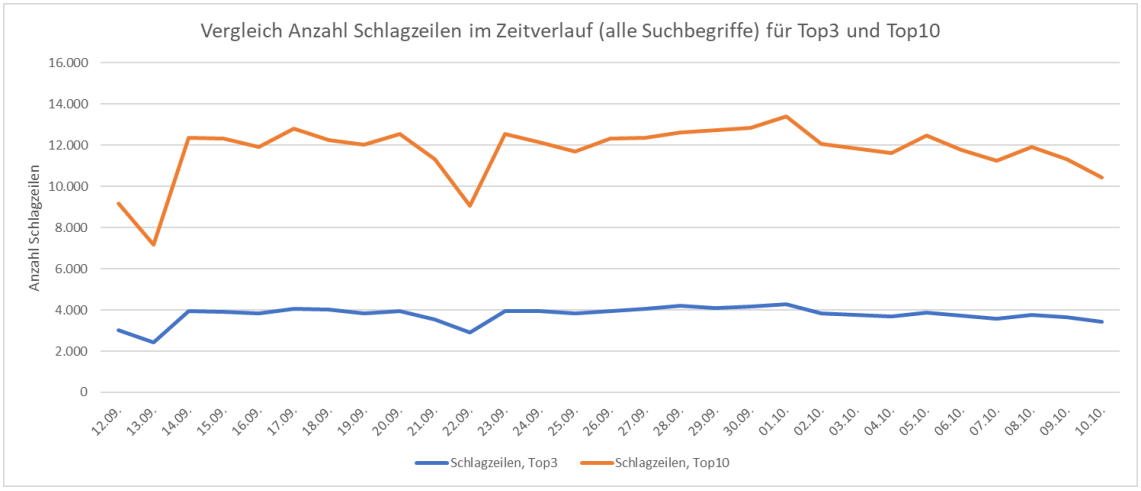


Abbildung 83 zeitlicher Verlauf Gini Koeffizient für alle und für ausgewählte Suchbegriffe (hier die 3 Kanzlerkandidaten), tageweise Betrachtung, Top3

Zur besseren Veranschaulichung sind in den folgenden Abbildungen die Verläufe jeweils für die Top3 und Top10 Schlagzeilen im direkten Vergleich dargestellt.





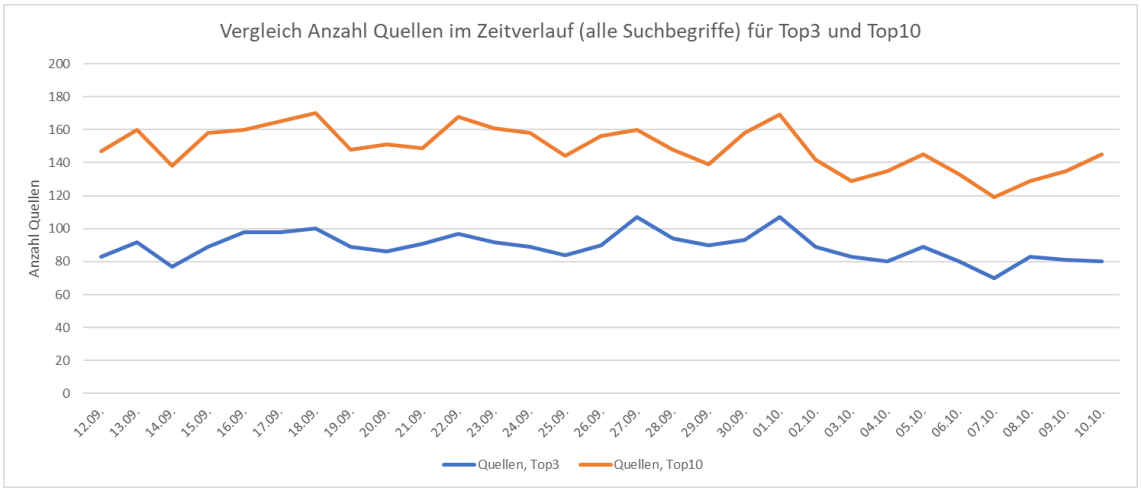


Abbildung 84 zeitlicher Verlauf Gini Koeffizient, Anzahl Schlagzeilen und Quellen für alle Suchbegriffe, tageweise Betrachtung, Top3 und Top10

Die Gini-Koeffizienten sind für die Top3 und Top10 Schlagzeilenauswahl sehr ähnlich, wobei erwartungsgemäß die Top10 Werte über den Top3 Gini-Koeffizienten liegen. Ebenfalls im Rahmen der Erwartungen liegen die Anzahlen der Schlagzeilen für Top10 bei ungefähr dem 3-fachen Wert von Top3. Auch die Anzahl der Quellen ist bei Top10 höher als bei Top3, wobei in diesem Fall der sehr ähnliche Verlauf beider Kurven auffällt.

Abschließend zeigen die Abbildung 85 und Abbildung 86 noch die Verläufe für ausgewählte Suchbegriffe in der täglichen und der wöchentlichen Betrachtung. Dabei fällt insbesondere beim täglichen Verlauf die noch höhere Schwankungsbreite der Gini-Koeffizienten auf, die durch die noch geringere Anzahl an Werten in den jeweiligen Teilmengen verursacht wird.

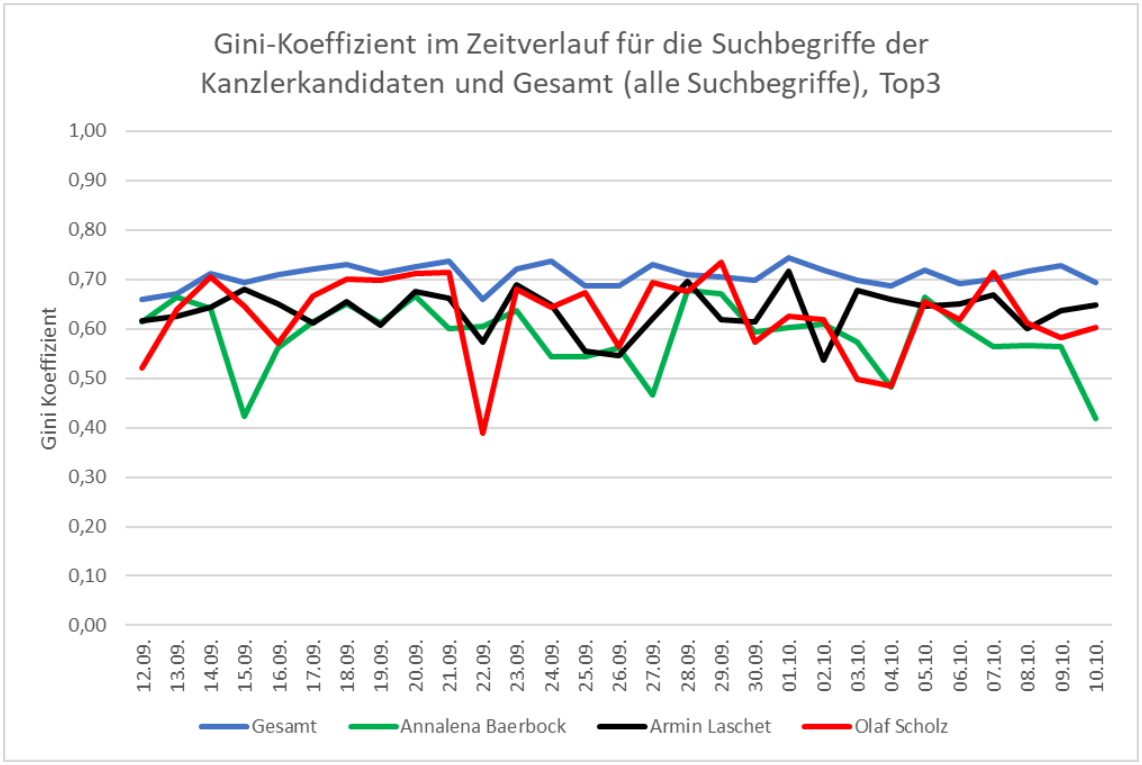


Abbildung 85 zeitlicher Verlauf Gini Koeffizient für alle und für ausgewählte Suchbegriffe (hier die 3 Kanzlerkandidaten), tageweise Betrachtung, Top3

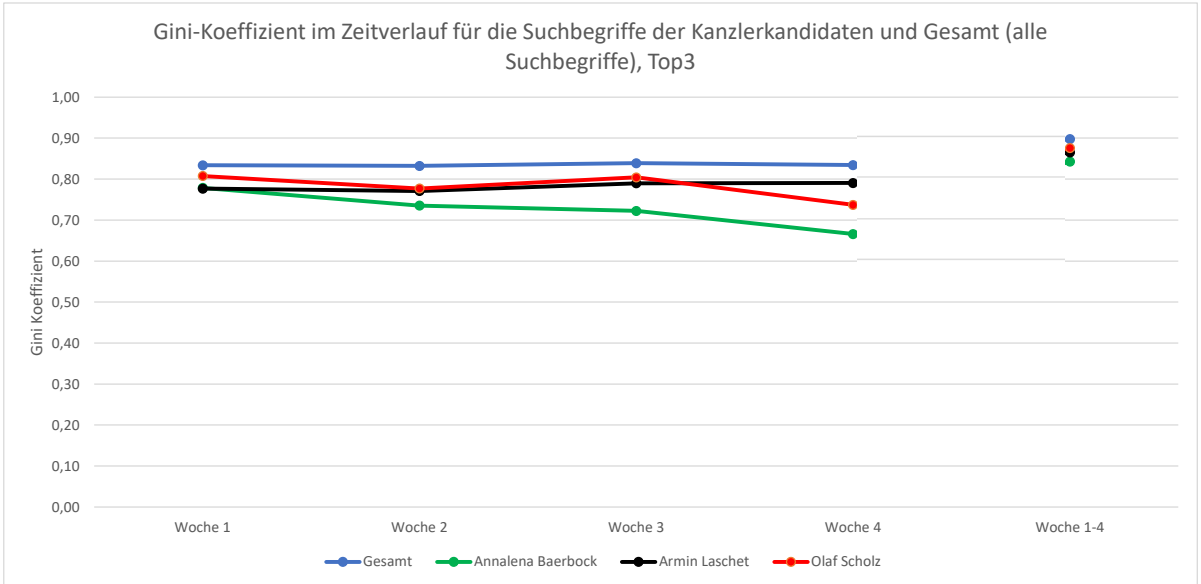


Abbildung 86 zeitlicher Verlauf Gini Koeffizient für alle und für ausgewählte Suchbegriffe (hier die 3 Kanzlerkandidaten), wochenweise und gesamte Betrachtung, Top3

3 Anhang

3.1 Listen/Datenquellen, die für die Studie genutzt wurden

In Folgenden eine Liste der der Dokumente, die signifikanten Einfluss auf die Analyse hatten:

- Gattungsliste, mit Definition der Medien aus Baden-Württemberg (Zuordnung_Gattung_LFK_21-10-27.xlsx)
- Medienliste Baden-Württemberg - diese Liste wurde in die Gattungsliste eingearbeitet durch die Kennzeichnung „BW Bezug" (Medienliste BW.xlsx)
- Liste der untersuchten Suchbegriffe (Finale Suchliste 09092021.xlsx)
- Liste der Orte/Gemeinden in Baden-Württemberg (Gemeinden_BaWü-2019.xls)

3.2 Weitere Anlagen

In Folgenden eine Liste der der Dokumente, die weitergehende Informationen bereitstellen:

- Auswertung der textuellen Ähnlichkeit im Detail (F20 Top 3 Similarities.xlsx)

3.3 Vollständige Darstellung der WordClouds

Im Folgenden werden die WordClouds dargestellt, die nicht exemplarisch unter Kapitel 2.5.1 aufgezeigt wurden.

3.3.1 WordCloud Suchbegriff Alice Weidel

Abbildung 87 WordCloud Suchbegriff Alice Weidel

3.3.2 WordCloud Suchbegriff Baden-Württemberg

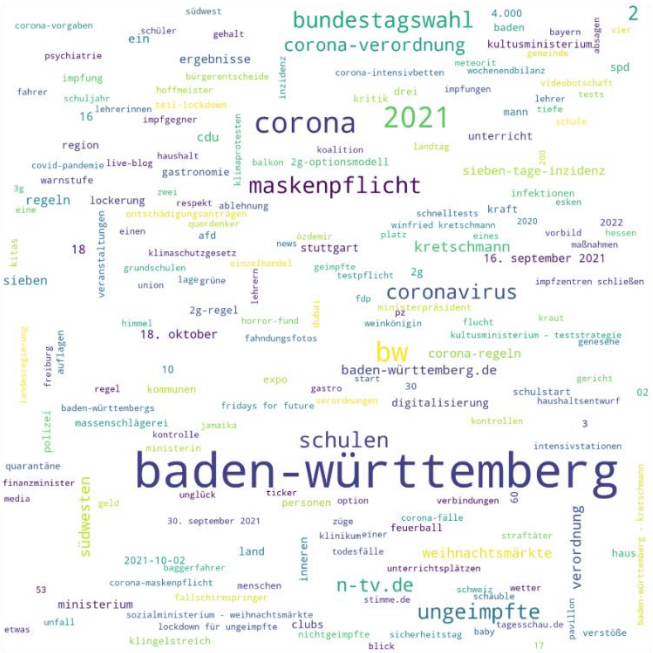


Abbildung 88 WordCloud Suchbegriff Baden-Württemberg

3.3.3 WordCloud Suchbegriff Bernd Riexinger

Abbildung 89 WordCloud Suchbegriff Bernd Riexinger

3.3.4 WordCloud Suchbegriff Bosch

Abbildung 90 WordCloud Suchbegriff Bosch

3.3.5 WordCloud Suchbegriff Daimler

Abbildung 91 WordCloud Suchbegriff Daimler

3.3.6 WordCloud Suchbegriff Franziska Brantner

Abbildung 92 WordCloud Suchbegriff Franziska Brantner

3.3.7 WordCloud Suchbegriff Karlsruhe

Abbildung 93 WordCloud Suchbegriff Karlsruhe

3.3.8 WordCloud Suchbegriff Klimawandel

Abbildung 94 WordCloud Suchbegriff Klimawandel

3.3.9 WordCloud Suchbegriff Michael Theurer

Abbildung 95 WordCloud Suchbegriff Michael Theurer

3.3.10 WordCloud Suchbegriff Saskia Esken

Abbildung 96 WordCloud Suchbegriff Saskia Esken

3.3.11 WordCloud Suchbegriff Stuttgart

Abbildung 97 WordCloud Suchbegriff Stuttgart

3.3.12 WordCloud Suchbegriff Wolfgang Schäuble

Abbildung 98 WordCloud Suchbegriff Wolfgang Schäuble