Projekttitel: Wandel der Worte - Langzeitdatenanalyse journalistischer Perspektiven

Teilnehmende: Levi Blumenwitz (17)

Erarbeitungsort: Privat/zu Hause

Projektbetreuende: Melanie Mestl

Thema des Projekts: Langzeitanalyse von Veränderungen der Zeitungen „The New York Times“ und „The Guardian“ bezüglich Sentiment und Artikelumfang

Fachgebiet: Arbeitswelt

Wettbewerbssparte: Jugend forscht

Bundesland: Bayern

Wettbewerbsjahr: 2025

Projektüberblick

Wie haben sich die Medien über die letzten Jahre verändert? In meinem Projekt befasse ich mich mit der Datenanalyse von zwei bedeutenden Zeitungen. Ausgewählt habe ich die "New York Times" (The New York Times Company, 03.01.25) (USA) und "The Guardian" (Guardian News, 03.01.25) (GB). Das Ziel ist, Artikel der beiden Zeitungen zu analysieren und auszuwerten, wobei ich mich bei beiden Zeitungen auf jeweils drei Rubriken - World, Politics und Opinion - beschränkt habe. Die drei Themen sind bei beiden Zeitungen vergleichbar, und ich werde alle Artikel dieser Rubriken analysieren. Kriterien bei der Analyse sind aktuell die Objektivität und Polarisation der Artikel, sowie deren Länge und Anzahl. Der geplante Zeitraum dieser Analyse ist 10 bis 20 Jahre.

Das Ziel dieser Forschungsarbeit ist die Überprüfung von Vorurteilen bezüglich der Veränderung in der Berichterstattung sowie, je nach Ergebnis, die Glaubwürdigkeit von Qualitätsmedien zu stärken beziehungsweise die Veränderung der Medien im Laufe der Zeit nachzuweisen.

Inhaltsverzeichnis

1. Fachliche Kurzfassung

Das Projekt untersucht die zeitliche Entwicklung von Sentiment (Polarisation und Subjektivität), Artikellänge und Artikelanzahl in den Rubriken „Politics“, „World“ und „Opinion“ in "The New York Times" und "The Guardian" zwischen 2010 und 2021. Mithilfe von Python wurden Artikel gesammelt und analysiert, um langfristige Trends zu identifizieren.
Da für die Auswertung der Analyse eine Vielzahl von Optionen möglich ist, wurden etwa 80 Graphen generiert, von denen zehn exemplarisch ausgewählt wurden, bei denen im Betrachtungszeitraum die deutlichsten Veränderungen sichtbar wurden.

Die Ergebnisse zeigen:
Die Artikelanzahl sinkt im „Guardian“ in der Rubrik "Opinion", während sie im Politikbereich der „New York Times“ deutlich ansteigt.
Das Sentiment bleibt in beiden Zeitungen konstant (durchschnittlich neutral). Jedoch zeigt die Kategorie "World" in beiden Zeitungen eine höhere Subjektivität als die Rubriken "Politics" und "Opinion".
Die durchschnittliche Artikellänge unterscheidet sich wie folgt: Artikel der „New York Times“ sind mit durchschnittlich 1100 Wörtern länger als der durchschnittliche „Guardian“-Artikel (800 Wörter), ohne signifikante Entwicklung über die Jahre.

2. Motivation und Fragestellung

Die Diskussion über Fake News und das schwindende Vertrauen in die Medien ist in allgegenwärtig. Im offiziellen „Edelman Trust Barometer 2024“ wird gezeigt, dass Menschen in 15 von 28 getesteten Ländern den Medien nicht vertrauen, dies gilt auch für Deutschland (Daniel J. Edelman Holdings Inc., S.49).

Obwohl Precht/Welzer Deutschland als „das Land der Qualitätspresse“ (Precht/Welzer, 2022, S. 7) bezeichnet, hat es „ein Problem mit dem Vertrauen in die Leitmedien. […] Nur 46% [der deutschen Befragten] gaben an, sie hätten Vertrauen in die Presse.“ (Precht/Welzer, 2022, S. 8).

Laut einer Journalismusumfrage aus dem Jahr 2024 geben 48% der Befragten an, der Journalismus in Deutschland sei schlechter geworden (TU Dortmund, 21.01.25). Auch in den USA haben laut einer Befragung „nur noch 31 Prozent der Befragten volles oder mehrheitliches Vertrauen in die Massenmedien“ (Marc Neumann, 21.01.25).
Laut Precht/Welzer in „Die vierte Gewalt“, können „Leitmedien der Versuchung zu polarisier[en] […] nicht widerstehen, […] [was die] Demokratie in eine schwierige Lage [bringt].“ (Precht/Welzer, 2022, S.65)

Doch kann man auch wissenschaftlich untersuchen, ob sich die Berichterstattung tatsächlich ins Negative verändert hat?

Als Nutzer von sozialen Medien war meine ursprüngliche Idee Artikel der Plattform „X“ (X Corp., 12.01.25) (ehemals Twitter) zu analysieren.

Nach ein wenig Durchsuchen der Webseite viel mir schnell auf, dass die meisten Meinungen auf „X“ in Form von Videos und Bildern dargestellt wurden. Da dies als Textanalyse nicht umsetzbar war, ging meine Suche weiter zu Facebook, wo das Auslesen der Artikel nicht unterstützt wurde, und zu Reddit, wo ein ähnliches Problem wie bei „X“ auftrat. Daraufhin habe ich meinen Blick auf die sogenannten Leitmedien gerichtet.

Bei meiner weiteren Recherche zur Analyse von Leitmedien stieß ich auf eine Studie von Michael Haller aus dem Jahr 2017. Haller hat die Berichterstattung zum Thema Flüchtlingsgeschehen von drei deutschen Zeitungen über einen Zeitraum von 20 Wochen untersucht. Er wertete 480 Zeitungsausgaben mit 2240 Seiten aus und analysierte 1687 Berichte und Kommentare. (Michael Haller, S.13)
Meine Idee war hingegen eine Langzeitanalyse durchzuführen, um die Entwicklung der Medien über viele Jahre hinweg zu analysieren. Mithilfe moderner Technologien wollte ich eine Analyse über einen viel größeren Zeitraum durchführen.

Hier habe ich zuerst mein Blick auf die „New York Times" geworfen. Diese ist bekannt für ihre objektive Berichterstattung und ist eine der größten Zeitungen in den USA.

Als zweite Zeitung habe ich „The Guardian“ aus dem Vereinigten Königreich als Vergleich ausgewählt, ebenfalls wertgeschätzt für ihren unabhängigen Journalismus, sowie meinen technischen Kriterien entsprechend.

Ziel meiner Arbeit ist die Überprüfung folgender Fragestellungen:

- Wie objektiv sind die beiden analysierten Zeitungen wirklich?
- Werden Zeitungsbeiträge tatsächlich immer negativer dargestellt?
- Sind Medien in den USA subjektiver als in Großbritannien?

- Wie stark hat sich die Medienberichterstattung bezogen auf Artikellänge und -anzahl in den letzten Jahren verändert?

3. Hintergrund und theoretische Grundlagen

Die „New York Times" ist eine der „renommiertesten Zeitungen der Welt" (Johann Oberauer GmbH, 12.01.2025). Aus diesem Grund habe nicht nur ich versucht, dieses Vertrauen zu überprüfen. Zum Beispiel „Ad Fontes Media" (Ad Fontes Media Inc., 12.01.25) hat viele verschiedene Zeitungen analysiert und ihnen eine Einstufung bezüglich „Reliability" und „Bias" gegeben, mit einem überdurchschnittlichen Ergebnis, sowohl für die „New York Times", als auch für „The Guardian". Bei diesem Test haben beide Zeitungen ein sehr ähnliches Ergebnis erzielt, doch die Größe der Analyse beschränkte sich nur auf rund ein Dutzend Artikel. Meine Idee war eine Analyse auf der Basis eines deutlich längeren Zeitraums und Umfangs.

4. Vorgehensweise, Materialien und Methoden

Der Vorgang, um die Daten zu sammeln und zu analysieren, ist sehr komplex und wird in mehreren Schritten durchgeführt. Es wird unterteilt in Sammeln, die Analyse und die Visualisierung der Daten.

4.a. Links sammeln

Der erste Hauptschritt ist, den Artikeltext zu bekommen. Dafür muss ich als erstes Zugriff auf die kompletten Links der beiden Zeitungen erhalten. Dafür benutze ich das „Application Programming Interface" (API) von "The New York Times" und "The Guardian". Da nur sehr wenige Zeitschriften so eine API haben, musste ich mich auf die beiden Zeitschriften beschränken. Die API gibt mir die Möglichkeit mithilfe von der Bibliothek „requests" (A Kenneth Reitz Project, 30.12.24), die Links der Artikel zu erhalten und in einer Datei zu speichern.
Mithilfe der Unterverzeichnisse des Links kann man das Datum, sowie Rubrik des Artikels auslesen und sortieren.

…/2020/01/02/us/politics/artikelname.html

Hier habe ich mich für drei Rubriken entschieden, welche bei beiden Zeitschriften vergleichbar sind. Diese Rubriken sind "World", "Politics" und "Opinion". Jetzt sortiere ich die Links nach Datum und Rubrik.

Mithilfe der Links kann ich jetzt auf die Webseiten zugreifen. Doch um die Artikel zu analysieren, brauche ich Zugriff auf den Artikeltext. Dies wird unterteilt in zwei wesentliche Schritte:

4.b. Quellcode herunterladen

Als nächstes benötige ich den Quellcode der Webseite. Das Beschaffen des Quellcodes war der vermutlich aufwendigste Prozess der ganzen Arbeit. Der Quellcode ist der HTML-Code der Webseite, der alle Informationen der Webseite enthält. Diesen Code kann ich durch verschiedene Methoden herunterladen. Der Prozess ist sehr unterschiedlich, je nachdem welche Webseite ich herunterlade.

Bei "The Guardian" war dieser Prozess relativ einfacher. Ich konnte mit einer einfachen Anfrage mit dem Python Modul „requests“ den Quellcode der Webseite herunterladen. Dieser wurde dann in einer Textdatei gespeichert, sortiert nach Datum und Rubrik.

Bei der New York Times war die Beschaffung des Quellcodes viel komplizierter. Die Methode „requests“ welche ich bei "The Guardian" genutzt habe, hat hier nicht funktioniert. Nach bereits 100 Artikeln wurde meine IP-Adresse blockiert (siehe Abbildung 1).

Abbildung 1: Fehlermeldung beim Abrufen der „New York Times“ Webseite Daten. (Levi Blumenwitz, E, 12.01.25)

Die zweite Methode ist die Python Bibliothek „selenium“ (Selenium Software Freedom Conservancy, 30.12.24), welche eine beliebte Methode ist um einen echten Browser wie Chrome zu simulieren. Doch auch hier gab es Probleme. Die „New York Times“ hat schnell meine ungewöhnliche Aktivität bemerkt, und nur den ersten Absatz des Artikels angezeigt. Außerdem wurde der Inhalt des Artikels hinter einer Paywall versteckt (siehe Abbildung 2).

Dies hat meine Analyse zunächst unmöglich gemacht und ich musste einen Weg finden, um die Paywall, sowie die vielen Captchas zu umgehen.

Mein erster Versuch, die Gegenmaßnahmen der New York zu umgehen bestand aus Rotation meiner Proxy. Leider haben hier Gratis-Proxys oft nicht funktioniert. Eine andere Methode war

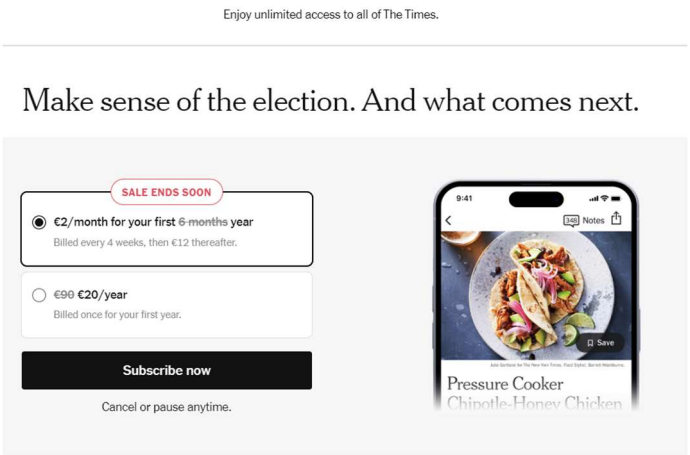

Abbildung 2: Aufruf zum Zahlen der „New York Times“ (Levi Blumenwitz, E, 12.01.25)

das Benutzen einer Externen API, zum Beispiel „Scraperapi“ (ScraperAPI, 12.01.25). Doch aufgrund von limitierten Tokens in der Freemium Version war dies auch keine Lösung. An diesem Punkt dachte ich, das wäre das Ende meiner Arbeit, doch nach ein wenig Herumprobieren, habe ich entdeckt, dass der komplette Artikeltext auch im "Backend" vorhanden ist. Aber leider in

einem komplizierten Geflecht aus JSON-ähnlichen Strukturen. Das bedeutet, ich konnte den Quellcode mit meiner vorher genannten Methode „selenium“ herunterladen und den Text nachträglich extrahieren.
Da dieser Schritt am zeitaufwendigsten war, habe ich nach Wegen gesucht um diesen Prozess zu optimieren.

Meine erste Idee war, das Projekt auf „Google Colab“ (Google Colaboratory, 03.01.25) als „Jupyter Notebook“ (Jupyter, 03.01.25) auszuführen. Hier kann man sowohl Python Code online ausführen, als auch die „selenium“ - Methode nutzen. Dies erlaubt jedoch weniger Zugriffsmöglichkeiten auf das Dateisystem, weshalb ich diese Methode verworfen habe. Meine zweite Idee war, das Projekt auf einem „Amazon Web Services“ (AWS) (Amazon Web Services Inc., 03.01.25) Server auszuführen. Hierbei habe ich einen AWS EC2-Server eingerichtet, um meine Anwendung zu hosten. Diese Methode hat mir ermöglicht den Quellcode im Hintergrund auf dem Server herunterzuladen. Mehr dazu kann in meinem erweitertem Blog gefunden werden. (Levi Blumenwitz, D, 10.01.25)

4.c. Text extrahieren

Für die Extraktion der „New York Times“ Texte habe ich eine Funktion erstellt, die den Text Stück für Stück aus dem Backend herausfiltert und dann wieder zusammenfügt. Mithilfe von „Regular Expressions“ (Python Software Foundation, 30.12.24) konnte der Text gefiltert und die relevanten Daten gespeichert werden. In Abbildung 3 ist die Funktion dargestellt, mit welcher ich den HTML-Text gefiltert habe. Als Parameter ist der HTML-Code erwartet und die Ausgabe ist der Text.

```
def get_text_from_html(html):
    matches = re.findall(r'"text":"(.*?)"', html)
    matches = list(dict.fromkeys(matches))
    text = "\n".join(matches)
    return text
```

Abbildung 3: Funktion zum Extrahieren des relevanten Textes aus dem Quellcode der New York Times

Das Extrahieren des Textes bei "The Guardian" war einfacher. Hier habe ich die herkömmliche Methode „BeautifulSoup“ (Leonard Richardson, 30.12.24) genutzt um anhand von HTML-Tags den Text zu lokalisieren. Dieser Text wurde dann in einer Textdatei gespeichert, sortiert nach Datum und Rubrik.

4.d. Text analysieren

Da ich nun den Artikeltext habe, kann ich diesen nach verschiedenen Kriterien analysieren.
Zum Speichern der ausgewerteten Daten werde ich mehrere SQL-Datenbanken erstellen. Hierfür benutze ich über das ganze Projekt hinweg Pythons Bibliothek „SQLite3“ (Python Software Foundation, 30.12.24). In dieser Datenbank speichere ich die Daten nach Datum und Rubrik, um sie später einfacher abrufen zu können.

4.d.i. Wörteranzahl

In meinem Code wird dies als "Wordcount" bezeichnet, und ist selbsterklärend. Zuerst habe ich den Text in einzelne Wörter aufgeteilt und diese zählen lassen. Diese Daten habe ich dann in einer weiteren SQL-Datenbank gespeichert.

4.d.ii. Sentiment Analyse

Die Sentiment Analyse ist ein wichtiger Bestandteil meiner Arbeit. Hierbei wird der Text auf Polarisation sowie Subjektivität hin analysiert. Dies wird mithilfe des Moduls „Textblob“ (Steven Loria, 30.12.24) durchgeführt.

4.d.ii.1. Polarisation

Dieses Modul gibt jedem Wort eine Wertung von -1 bis 1, wobei -1 negativ und 1 positiv ist. Daran kann man erkennen, ob ein Text tendenziell positiv oder negativ gefärbt ist.

4.d.ii.2. Subjektivität

„Textblob“ berechnet die Subjektivität, indem es die 'Intensität' betrachtet. Die Intensität bestimmt, ob ein Wort das nächste Wort modifiziert. (Parthvi Shah, 12.01.25)

Der Satz „The food was great” hat eine Subjektivität von 0.75 und eine Polarisation von 1.0, wie man an dem einfachen Code in Abbildung 4 erkennen kann.

```
import textblob
string = "The food was great!"
blob = textblob.TextBlob(string)
print(f"Polarity: {blob.sentiment.polarity}, Subjectivity: {
      blob.sentiment.subjectivity}")
```

Abbildung 4: Einfaches Beispiel zur Bestimmung der Polarisation und Subjektivität eines Strings. Mit dem Ergebnis: Polarity: 1.0, Subjectivity: 0.75

Diese beiden Werte werden anhand des Datums als Indikator in einer SQL-Datei gespeichert, um späteres Abrufen zu erleichtern.

4.d.iii. Artikelanzahl

Hierbei wird die Anzahl der Artikel pro Rubrik und Tag gezählt und in einer SQL-Datei gespeichert.
Dadurch kann man die Entwicklung der Artikelanzahl über die Jahre erkennen.
Hier wird unterteilt in die Anzahl der Artikel pro Tag sowie pro Monat.

4.e. Graphen erstellen

Die Daten, die ich in den SQL-Dateien gespeichert habe, werden graphisch dargestellt.
Hierfür habe ich als erstes eine globale Funktion programmiert, die mithilfe von verschiedenen Eingabeparametern die Graphen erstellt. (Levi Blumenwitz, C, 12.01.25)

```
def graph(rows, column_names, name, legend_title, title, drop_columns, color,
color_reg, regression, size, output):
```

Abbildung 5: Methodenkopf der Funktion graph(…) in der Datei „Plotting.py“ um einen Graphen zu erstellen (Levi Blumenwitz, C, 12.01.25)

Diese Funktion wird dann in den einzelnen Dateien aufgerufen. Zum Zeichnen des Graphen werden die Daten aus der Datenbank, alle Spalten, die entfernt werden sollen (z.B. ID), und andere Parameter wie z.B. Titel und Farben benötigt.

```
topics = ["Politics", "World", "Opinion"]
options = ['polarity', 'subjectivity']
colors = ['#1f77b4', '#ff7f0e', 'green']
colors_reg = ['blue', 'red', 'black']
news = ["NYT", "Guardian"]
for new in news:
    connection = sqlite3.connect(f"Database/Sentiment/{new}.db")
    cursor = connection.cursor()
    for o, option in enumerate(options):
        name = option.capitalize()
        drop_columns = ['id', 'idx', f'{options[1 - o]}']
        for i, topic in enumerate(topics):
            cursor.execute(f"SELECT * FROM {topic};")
            rows = cursor.fetchall()
            column_names = [col[0] for col in cursor.description]
            output = f"Output/Sentiment/{option}/{new}/{topic}.png"
            os.makedirs(os.path.dirname(output), exist_ok=True)
            graph(rows, column_names, option, f"{name} for {topic}", f"{name}
Analysis for {topic}", drop_columns, colors[i], colors_reg[i], True, 2,
output)
```

Abbildung 6: Beispiel Aufruf für die graph(…) Methode um mehrere Graphen über eine Wiederholungsanweisung zu erstellen (Levi Blumenwitz, F, 12.01.25)

Innerhalb der verschiedenen Analysetypen erstelle ich weitere Wiederholungsanweisungen um für jede Rubrik einen entsprechenden Graphen zu erstellen. Die Rubriken werden für jedes Jahr graphisch dargestellt, um die jeweilige Entwicklung innerhalb einzelner Jahre und für den Gesamtzeitraum zu visualisieren.

4.f. Interaktive Webseite erstellen

Um die Daten individualisiert und flexibel darzustellen habe ich zwei interaktive Webseiten mithilfe von Streamlit (Snowflake Inc., 30.12.24) erstellt. Eine Webseite zum Visualisieren der Wörteranzahl und die zweite für das Sentiment, also Polarisation und Subjektivität.

Die Webseiten wurden über den offiziellen Hosting-Service von Streamlit veröffentlicht und können im Quellenverzeichnis gefunden werden (Levi Blumenwitz, G, H, 12.01.25)

Beide Webseiten haben folgende Funktionen:

4.f.i. Datenauswahl

Zu Beginn kann man auswählen, wie viele verschiedene Graphen man übereinander angezeigt haben möchte. Man kann jedem Graphen die Daten zuordnen, die dieser anzeigen soll. Hier kann man auswählen zwischen einer oder beiden Zeitungen, sowie zwischen einer Rubrik, bzw. allen Rubriken vereint.

Nach Auswahl der Daten kann man auswählen für welchem Zeitraum die Daten angezeigt werden sollen. Wenn der Zeitraum kleiner als ein Jahr ist hat man die Möglichkeit auch auf genauere Monate zu begrenzen.

4.f.ii. Graph erstellen

Der Graph wird erst durch Drücken eines Buttons erstellt, um andauerndes Laden des Graphen im Hintergrund zu vermeiden. Hierfür wird die interaktive Bibliothek „plotly“ (2024 Plotly, 30.12.2024) benutzt, die es ermöglicht, den Graphen zu zoomen, zu verschieben und bestimmte Daten mithilfe der Legende auszublenden.

4.f.iii. Tabelle mit Top 10 Artikeln

Zusätzlich zu den Graphen habe ich eine Tabelle erstellt, welche den Datensatz nach den in 4.f.i. ausgewählten Kriterien sortiert und die Top 10 Artikel anzeigt.
Angezeigt wird dann der Titel des Artikels, das Datum und auch die Content ID des Artikels. Mit dieser ID kann man auf einer anderen von mir erstellten Webseite den Artikel direkt aufrufen und genauer lesen. Diese Webseite ist öffentlich nicht zugänglich, da die Daten nur lokal auf meinem Laptop gespeichert sind.

5. Ergebnisse

Ziel der Arbeit war eine Langzeitdatenanalyse aller Artikel in den ausgewählten Rubriken der beiden Zeitschriften. Aufgrund der zeitlichen Begrenzung in dieser Forschungsarbeit konnte ich nicht die kompletten Daten herunterladen. Deshalb habe ich mich darauf beschränkt die Entwicklung von 2010 - 2021 genauer zu analysieren, indem ich den Anfang des Zeitraums 2010 - 2011 und das Ende 2020 - 2021 heruntergeladen habe. Hierfür wurden von der „New York Times“ rund 52.000 Artikel und von „The Guardian“ rund 72.000 Artikel heruntergeladen und analysiert. Mit diesen Daten kann ich nun abschätzen, wie sich die verschiedenen Aspekte entwickelt haben.

Ich habe zum Vorstellen der Ergebnisse besonders aussagekräftige Diagramme ausgewählt, die die Entwicklung der beiden Zeitungen gut darstellen. Alle weiteren, bereits erstellten Graphen können auf meinem „GitHub-Repository“ (Levi Blumenwitz, B, 12.01.2025) eingesehen werden. Alternativ können eigene Graphen auf meinen Webseiten (Levi Blumenwitz, G, H, 12.01.2025) selber generiert werden.

5.a. Entwicklung der Artikelanzahl

Auf dem dargestellten Bild (Abbildung 7) kann man die Entwicklung der Artikelanzahl der Rubrik "Opinion" der Zeitung "The Guardian" sehen. Jeder Punkt stellt die Anzahl der Artikel pro Tag dar. Die Y-Achse zeigt die Anzahl der Artikel, die an diesem Tag veröffentlicht wurden und die X-Achse den Zeitraum. Man erkennt, dass in der Mitte des Graphen keine Punkte vorhanden sind.

Dies liegt an den nicht heruntergeladenen Daten in diesem Zeitraum. In schwarz ist die Regressionsgerade dargestellt, welche die Entwicklung der Artikelanzahl über die Jahre hinweg darstellt. Wie auch an dem Text in der rechten oberen Ecke zu erkennen ist, ist die Artikelanzahl in den letzten Jahren abgesunken. Durchschnittlich sind es ca. 11 Artikel pro Tag weniger im Jahr 2021 als im Jahr 2010.

Im Gegenzug dazu ist bei der „New York Times“ (Abbildung 8) im Bereich „Politik“ ein starker Anstieg zu erkennen. Im Durchschnitt wurden im Jahr 2021 8 Artikel pro Tag mehr veröffentlicht als im Jahr 2010.

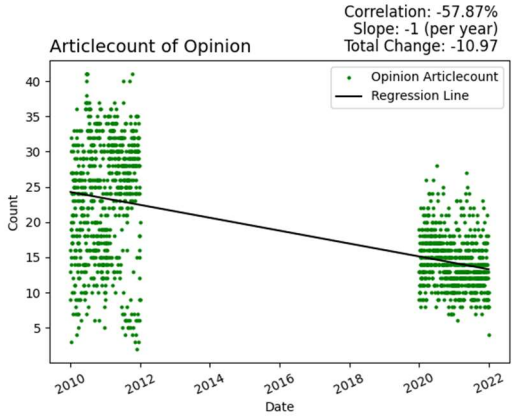

Abbildung 7: Artikelanzahl der Zeitung „The Guardian“, der Rubrik „Opinion“, jeder Punkt repräsentiert die Anzahl der Artikel pro Tag

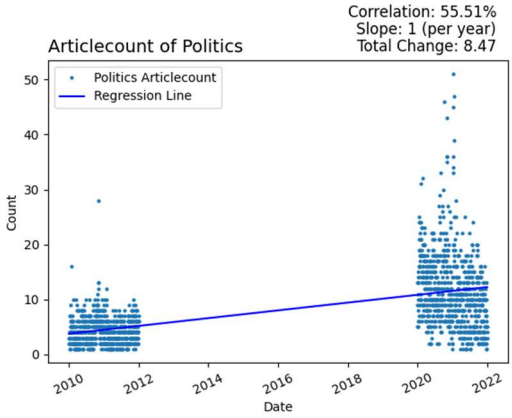

Abbildung 8: Artikelanzahl der Zeitung „The New York Times“, der Rubrik „Politics“, jeder Punkt repräsentiert die Anzahl der Artikel pro Tag

Bestätigt wird dies durch die Graphen der Artikelanzahl pro Monat. (Abbildung 9 und Abbildung 10) Hierbei ist ein starker Anstieg bei der New York Times (rechts) (Abbildung 10) zu erkennen, während bei "The Guardian" (links) (Abbildung 9) ein starker Abfall stattfindet.

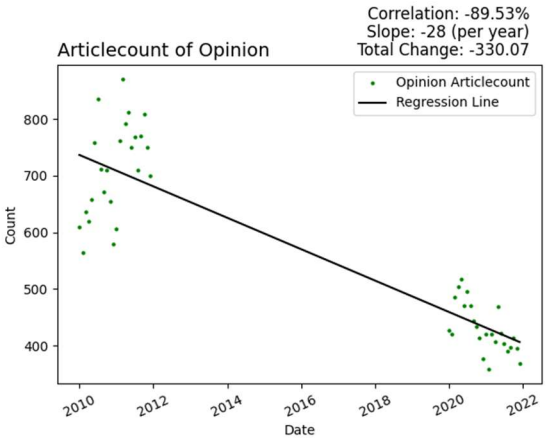

Abbildung 9.: Artikelanzahl der Zeitung „The Guardian“, der Rubrik „Opinion“, jeder Punkt repräsentiert die Anzahl der Artikel pro Monat

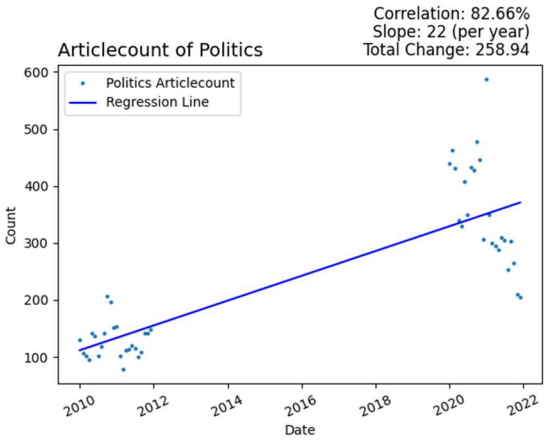

Abbildung 10.: Artikelanzahl der Zeitung „The New York Times“, der Rubrik „Politics“, jeder Punkt repräsentiert die Anzahl der Artikel pro Monat

5.b. Entwicklung des Sentiments

Die Sentiment Analyse wird unterteilt in Polarisation und Subjektivität.

5.b.i. Die Polarisation gibt an, ob ein Text positiv oder negativ ist. -1 bedeutet sehr negativ, 1 bedeutet positiv. Der durchschnittliche Wert beider Zeitungen ist 0.1.

In Abbildung 11 werden die 3 Kategorien der Zeitung "The Guardian" verglichen. Es ist klar ersichtlich, dass die Polarisation über die Jahre hinweg relativ konstant ist und auch zwischen den Kategorien vergleichbar ist. Es gibt keine großen Schwankungen und die Werte sind immer um ca. 0.1.

Auch bei der New York Times (Abbildung 12) ist ein ähnliches Bild zu erkennen. Die Werte sind relativ konstant und auch hier gibt es keine großen Schwankungen. Die Werte liegen immer bei ca. 0.1.

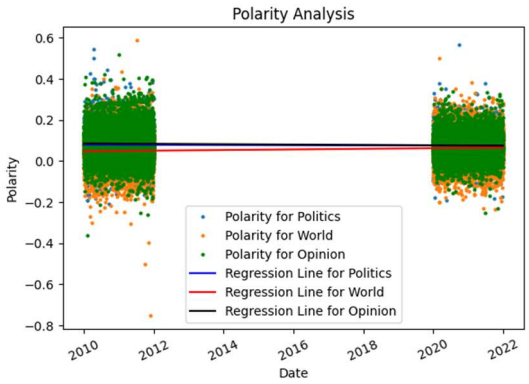

Abbildung 11.: Polarisation der Zeitung „The Guardian“, alle Rubriken in einer Darstellung, jeder Punkt repräsentiert einen Artikel.

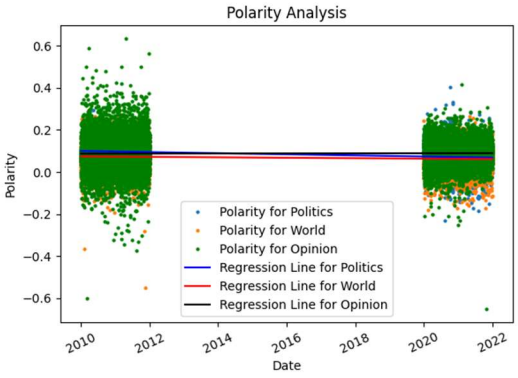

Abbildung 12.: Polarisation der Zeitung „The New York Times“, alle Rubriken in einer Darstellung, jeder Punkt repräsentiert einen Artikel

5.b.ii. Die Subjektivität gibt an, wie objektiv oder subjektiv ein Text ist. 0 bedeutet sehr objektiv, 1 bedeutet subjektiv. Der durchschnittliche Wert beider Zeitschriften liegt bei 0.4, wie man an der Grafik (Abbildung 13) erkennen kann. Hier wurden alle drei Kategorien zusammengefasst und in einem Graphen dargestellt.

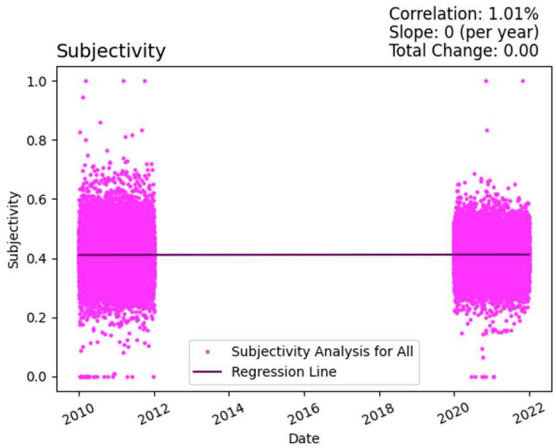

Abbildung 13: Subjektivität der Zeitung „The New York Times“, alle Rubriken vereint, jeder Punkt repräsentiert einen Artikel

Wenn man aber die Kategorien einzeln betrachtet (Bild unten), sieht man deutliche Unterschiede. Die Kategorie "World" ist deutlich objektiver als die beiden anderen Kategorien. Dies ist auch bei

"The Guardian" zu erkennen. Die Kategorie "World" ist zu Beginn des Zeitraumes deutlich objektiver als die anderen beiden Kategorien, und die Kategorie "Opinion" ist am subjektivsten. Die Kategorie „Politik“ wird über den Zeitraum gesehen eher objektiver.

Interessant ist auch, dass das Ergebnis (Abbildung 14) sehr vergleichbar mit dem Ergebnis von "The Guardian" (Abbildung 15) ist. Auch hier ist die Kategorie "World" am objektivsten und "Politics" nähert sich langsam vom eher subjektiven "Opinion" in Richtung der Kategorie "World".

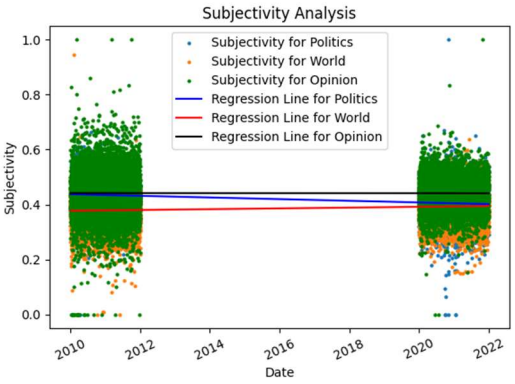

Abbildung 14: Subjektivität der Zeitung „The New York Times“, alle Rubriken in einer Darstellung, jeder Punkt repräsentiert einen Artikel. Ca. 52.000 Artikel

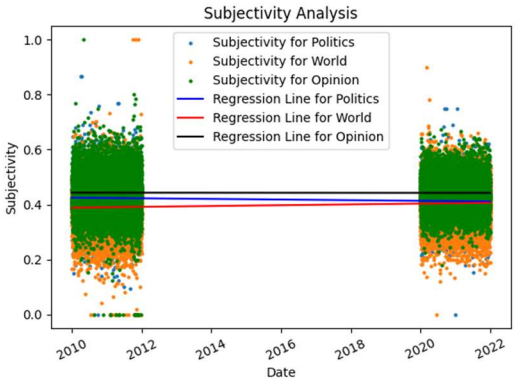

Abbildung 15: Subjektivität der Zeitung „The Guardian“, alle Rubriken in einer Darstellung, jeder Punkt repräsentiert einen Artikel. Ca. 73.000 Artikel

5.c. Wörteranzahl bzw. Artikellänge

Zuletzt noch die Wörteranzahl und die Artikellänge. "The New York Times" hat eine deutlich höhere durchschnittlichen Wörteranzahl mit ca. 1100 Wörtern pro Artikel. Dahingegen hat "The Guardian" eine durchschnittliche Wörteranzahl von 800 Wörtern.
Abgesehen von dem durchschnittlichen Unterschied der Wörteranzahl, gibt es auch Entwicklungen über die Jahre. Doch diese sind nicht sehr aussagekräftig, da die Korrelation sehr gering ist, was am Korrelationskoeffizient von nur 0.09 zu erkennen ist. (Abbildung 16)

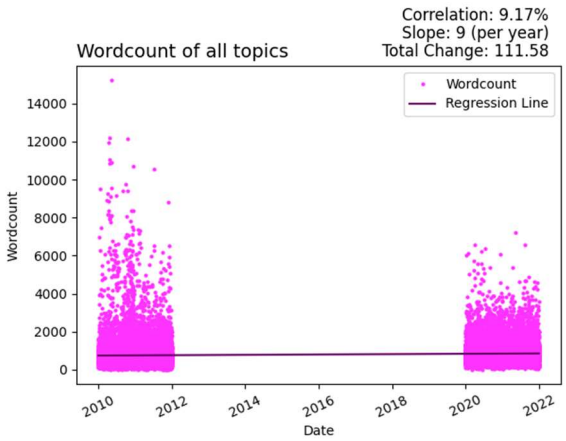

Abbildung 16: Wörteranzahl der Zeitung „The Guardian“, alle Rubriken vereint, jeder Punkt repräsentiert einen Artikel

6. Ergebnisdiskussion

Wenn man die Ergebnisse analysiert, können daraus viele Schlussfolgerungen gezogen werden.

Das Sinken der Artikelanzahl in der Rubrik „Opinion“ bei gleichbleibender Artikellänge in „The Guardian“ könnte bedeuten, dass diese weniger auf Meinungsbeiträge fokussiert ist, und beim Erstellen solcher Artikel nachgelassen hat. Bei der „New York Times“ erkennt man, dass diese den Fokus seit 2010 mehr auf politische Beiträge gelegt haben. Der Anstieg von fast 300 Artikeln pro Monat mehr, könnte allerdings auch daran liegen, dass es auf politischer Ebene mehr zu berichten gibt.

Überraschend war für mich die durchschnittliche Polarisation von 0.1. Der konstante Wert zeigt eine sehr geringe Änderung über 10 Jahre hinweg bei beiden Zeitungen. Durch diesen konstanten Wert wird gezeigt, dass die Medien nicht viel negativer geworden sind, wie ursprünglich vermutet. Der Wert, welcher nah an 0 liegt, zeigt, dass die Zeitung nicht negativ aber auch nicht positiv berichtet. Das ist ein positives Merkmal für Leser, da es das Vertrauen der Leser in die Zeitung stärkt.
Wenig überraschend ist, dass die Rubrik „Opinion“ subjektiver ausgerichtet ist, als die anderen Rubriken. Dies liegt am grundsätzlichen Charakter der Rubrik, welche ausschließlich aus Meinungsbeiträgen besteht.

Die Rubrik „World“ hingegen bleibt auf einem konstanten eher objektiv ausgerichtetem Level. Dies ist auch nachvollziehbar, da es sich hier um eine faktische Berichterstattung handelt. „Politics“ hingegen zeigt als einzige Rubrik eine Änderung, und zwar zunehmende Objektivität. Dies zeigt, dass das Thema „Politik“ bei beiden Zeitungen objektiver wird und weniger meinungsbasiert, was auf jeden Fall ein positiver Aspekt für den Leser ist. Womöglich folgt dies dem allgemeinen Trend der Qualitätsmedien, Meinungsbildung und objektive Berichterstattung deutlicher zu trennen.

Die komplett identische Entwicklung hinsichtlich der Subjektivität beider Zeitungen zeigt, dass die Berichterstattung zwischen Großbritannien und den USA sehr ähnlich ist und sich die journalistischen Standards ähneln. Außerdem wird durch meine Analyse gezeigt, dass seriöse Medien ähnlich arbeiten.

Um dies genauer zu analysieren könnte man zum Beispiel die Daten spezieller Jahre und Ereignisse, wie Wahljahre, oder dem Brexit herausfiltern, und genauer Analysieren.

Abschließend muss auch angemerkt werden, dass das automatische Auslesen von Sentiment und Polarisierung auch an Grenzen stößt, da es für ein Programm schwierig ist z.B. Sarkasmus und Ironie zu erkennen.

Ich habe erwartet, dass die Artikel kürzer werden, da die Aufmerksamkeitsspanne von Lesern allmählich durch soziale Medien verringert wird. Die konstante Wörteranzahl zeigt, dass die beiden Zeitungen trotz Annahme, dass Menschen „lieber kürzere Texte lesen“ (Konradin Medien GmbH, 12.01.25), ihrem Verständnis von qualitativem Journalismus treu bleiben.

7. Fazit und Ausblick

Das Ziel der Langzeitdatenanalyse wurde erreicht, jedoch konnte aufgrund der begrenzten Zeit nur ein Teil der Daten heruntergeladen werden, wobei trotzdem eine beachtliche Zahl von 125.000 Artikeln analysiert wurde.
Es konnte eine Veränderung der Medien über den Zeitraum von 10 Jahren festgestellt werden, jedoch konnte die Vermutung einer zunehmenden Polarisierung, also extremeren Meinungsbildung in den untersuchten Medien nicht bestätigt werden.
Eine konstante Polarisierung hat verschiedene Vorteile, denn durch eine konstante Polarisation kann das Vertrauen der Leser der Zeitungen gestärkt werden.
Auch beim Vergleich beider Zeitungen wird deutlich, dass die Berichterstattung in beiden Zeitungen ähnlich ist. Dies zeigt auf eine seriöse Berichterstattung beider Zeitungen, unabhängig vom Herkunftsland.
Ein deutlicher Unterschied zwischen der amerikanischen Zeitung "The New York Times" und der britischen Zeitung "The Guardian" konnte in dieser Forschungsarbeit nicht festgestellt werden.
Mein Projekt liefert ein konkretes Beispiel, wie Datenanalyse genutzt werden kann, um Vorurteile und Behauptungen über Medien zu überprüfen. Mit meiner Forschungsarbeit wurde für zukünftige Untersuchungen ein Tool erstellt, welches genutzt werden kann, um Artikel der beiden Zeitungen herunterzuladen und zu analysieren.
Diese Analyse kann auf weitere Zeitungen ausgeweitet werden, um so eine umfassende Analyse der Medienlandschaft zu ermöglichen.
Außerdem kann die Analyse auf weitere Jahre erweitert werden um Trends und Entwicklungen über einen längeren Zeitraum zu untersuchen.
Auch die verschiedenen Analysetools können weiter entwickelt werden. Weitere Ideen waren das Analysieren der Artikel auf Schlagwörter, um so die Themen der Artikel zu identifizieren, und dadurch die Entwicklung und Relevanz von Themen über die Jahre hinweg zu analysieren. Eine weitere Option wäre die Untersuchung der Komplexität der Artikel, um so die Veränderung der Verständlichkeit der Artikel zu analysieren.

Mein Projekt bietet eine Vielzahl an Möglichkeiten für Unternehmen, wie z.B. die Trendanalyse, welche Wettbewerbsvorteile bilden könnte. Außerdem bietet mein Projekt Vorteile für Bildungseinrichtungen, da die allgegenwertige Medienkritik, sowie sonstige Vorurteile genauer untersucht werden können.
Durch die Webseite kann die Auswahl der Daten individuell gefiltert werden, um spezielle Fragestellungen zu untersuchen.

Mit meiner Forschungsarbeit möchte ich einen Beitrag zur Aufklärung leisten, denn „in einer von Medien geprägten Welt, kann es gar nicht genug Medienkritik geben. […] Denn die Mediendebatte ist in letzter Instanz auch eine Debatte über den aktuellen Zustand und die Zukunft der Demokratie.“ (Ulrich Teusch, 2016, S.10)

8. Quellen- und Literaturverzeichnis

8.a. Python und Bibliotheken

https://www.python.org/: 31.12.24, Python Software Foundation, Python als Programmiersprache

https://www.crummy.com/software/BeautifulSoup/: 30.12.2024, Leonard Richardson, Beautiful Soup for HTML-Parsing

https://plotly.com/: 30.12.2024, © 2024 Plotly, Plotly für interaktive Graphen

https://requests.readthedocs.io/en/master/: 30.12.2024, © 2024. A Kenneth Reitz Project, Requests: HTTP for Humans

https://scikit-learn.org/: 30.12.2024, © 2007 - 2024 scikit-learn developers (BSD License), scikit-learn for Regression

https://www.selenium.dev/: 30.12.2024, © 2024 Selenium Software Freedom Conservancy, Selenium für Web-Scraping

https://docs.python.org/3/library/sqlite3.html: 30.12.2024, Python Software Foundation, SQLite3 für Datenbanken

https://docs.python.org/3/library/re.html: 30.12.2024, Python Software Foundation, Regular expression operations

https://streamlit.io/: 30.12.2024, © 2024 Snowflake Inc., Streamlit für Webseiten

https://textblob.readthedocs.io/en/dev/: 30.12.2024, © Steven Loria, TextBlob für NLP

8.b. Webseiten

https://colab.google/: 03.01.25, Google Colaboratory, Google Colab

https://jupyter.org/: 03.01.25, Jupyter, Jupyter Notebook

https://www.theguardian.com/: 03.01.25, © 2025 Guardian News, The Guardian

https://www.nytimes.com/: 03.01.25, © 2025 The New York Times Company, The New York Times

https://aws.amazon.com/de/: 03.01.25, 2024 Amazon Web Services Inc, Amazon Web Services

https://github.com/: 12.01.25, 2025 GitHub Inc., Github als Versionkontrollsystem

https://github.com/AdminL3/Jugend-Forscht: 12.01.25, Levi Blumenwitz, Mein Jugend-Forscht Projekt auf „Github“ (GitHub Inc, 12.01.25)

https://www.edelman.de/sites/g/files/aatuss401/files/2024-01/2024%20Edelman%20Trust%20Barometer_Germany%20Report_0.pdf: 12.01.2025, Daniel J. Edelman Holdings Inc., 2024 Edelman Trust Barometer - Germany Report

https://x.com/: 12.01.2025, © 2025 X Corp., Social Media Platform X (ehemals “Twitter”)

https://adfontesmedia.com/new-york-times-bias-and-reliability/: Ad Fontes Media Inc., The New York Times Bias and Reliability

https://www.scraperapi.com/: 12.01.25, ScraperAPI, Scale Data Collection with a Simple API

https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524: 21.01.2025, Parthvi Shah, Sentiment Analysis using TextBlob

https://www.otto-brenner-stiftung.de/fileadmin/user_data/stiftung/02_Wissenschaftsportal/03_Publikationen/AH93_Fluechtingskrise_Haller_2017_07_20.pdf: 12.01.25, Michael Haller, Die „Flüchtlingskrise“ in den Medien

https://www.newsroom.de/news/aktuelle-meldungen/vermischtes-3/die-10-renommiertesten-zeitungen-der-welt-957389/: 12.01.25, Johann Oberauer GmbH, Die 10 renommiertesten Zeitungen der Welt

https://www.pro-medienmagazin.de/glaubwuerdigkeit-des-journalismus-leidet-in-der-pandemie/: 12.01.25, Das christliche Medienmagazin, Glaubwürdigkeit des Journalismus leidet in der Pandemie

https://www.journalismusstudie.fb15.tu-dortmund.de/journalismus-und-demokratie/publikum-2024/: 12.01.25, TU Dortmund, Studie zu Journalismus & Demokratie 2024

https://www.nzz.ch/feuilleton/eine-grosse-mehrheit-vertraut-ihnen-ueberhaupt-nicht-oder-nicht-sonderlich-sind-die-massenmedien-in-den-usa-am-ende-ld.1854469: 21.01.25, Marc Neumann, Sind die Massenmedien in den USA am Ende?

https://www.wissenschaft.de/gesellschaft-psychologie/wie-die-digitalisierung-das-leseverhalten-veraendert/: 12.01.25, Konradin Medien GmbH, Wie die Digitalisierung das Leseverhalten verändert

8.c. Literatur

- Precht, Richard David und Welzer, Harald: Die vierte Gewalt – Wie Mehrheitsmeinung gemacht wird, auch wenn sie keine ist, Frankfurt am Main, 2022

- Teusch, Ulrich: Lückenpresse - Das Ende des Journalismus, wie wir ihn kannten, Frankfurt/Main 2016

- Pörksen, Bernhard: Die große Gereiztheit - Wege aus der kollektiven Erregung, München 2018

- Nocun, Katharina und Lamberty, Pia: Fake Facts - Wie Verschwörungstheorien unser Denken bestimmen, Köln 2020

8.d. Levi Blumenwitz

A: https://github.com/AdminL3/Jugend-Forscht/: 12.01.25, Levi Blumenwitz, Komplette Codebase für das Jugend Forscht Projekt

B: https://github.com/AdminL3/Jugend-Forscht/tree/main/Output: 12.01.25, Levi Blumenwitz, Generierte Graphen

C: https://github.com/AdminL3/Jugend-Forscht/blob/main/Plotting/Plotting.py: 12.01.25, Levi Blumenwitz, Code für die Globale Funktion „Plotting.py“

D: https://medium.com/@l-blu/running-python-in-the-cloud-3413fe59dfec: 12.01.25, Levi Blumenwitz, Running Python in the Cloud using AWS

E: https://github.com/AdminL3/Jugend-Forscht/tree/main/data-collection/Errors: 12.01.25, Levi Blumenwitz, Fehlermeldungen beim Sammeln des Quellcodes

F: https://github.com/AdminL3/Jugend-Forscht/blob/main/Plotting/Sentimental/Graph.py: 12.01.25, Levi Blumenwitz, Aufrufen der graph(…) Funktion um Graphen zu erstellen

G: https://jugend-forscht-wordcount.streamlit.app/: 21.01.25, Levi Blumenwitz, Webseite zum Visualisieren der Wörteranzahl

H: https://jugend-forscht-sentiment.streamlit.app/: 21.01.25, Levi Blumenwitz, Webseite zum Visualisieren der Polarisation und Subjektivität