Richtlinie zum „Software-Sprint“
PoliTrack Abschlussbericht

Zuwendungsempfänger:
Jonas Dippel, Michael Perk GbR

März 2022

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS21S51 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

1 Aufgabenstellung und Motivation

Einer der Grundpfeiler der Demokratie sind unabhängige Medien, die kritisch über die politische Lage im Land berichten. In Zeiten von Online-Journalismus und Live-Berichterstattung zu aktuellen Geschehnissen vergisst man leicht, dass sich die Entscheidung bei einer Wahl auch an bereits vergangenen Entscheidungen und Meinungen orientieren sollte. Möchte man über die zur Wahl stehenden Personen eines Landkreises mehr erfahren, kann man eine Online-Suche machen, sich Interviews ansehen und auf Wahlkampfveranstaltungen gehen. All diese Möglichkeiten haben einen großen Makel: Es fehlt die zeitliche Dimension. Mehr noch, ist es als Bürger:in unglaublich schwer, sich mit aktuellen Berichten auf dem Laufenden zu halten. Sobald einzelne Themen durch die Nachrichten in den Vordergrund geraten sind, bekommen sie für einige Tage oder Wochendie volle Aufmerksamkeit. Im Anschluss geraten sie dann wieder schnell in Vergessenheit. Für politische Partizipation und eine gute Wahlentscheidung ist es jedoch wichtig, möglichst gut informiert über die Positionen einzelner Politiker:innen zu sein. Bei der Internetrecherche zu einem/einer Politiker:in erscheinen nur aktuelle Themen. Um auch historische Daten einzubeziehen, muss ein enormer Aufwand betrieben werden, den nur die wenigsten in Kauf nehmen möchten. Gerade über Politiker:innen, die selten in Medien erwähnt werden, ist es sehr schwierig fundierte Informationen zu bekommen. Um die Meinungsbildung zu erleichtern, wollen wir mit unserem Projekt Nachrichtenartikel besser ordnen und zugänglich machen. Mithilfe von künstlicher Intelligenz möchten wir die Inhalte analysieren und so die Recherche so einfach wie möglich machen.

Meilenstein 1 *Entwicklung einer professionellen und skalierbaren Software* Es soll ein skalierbares Produktivsystem entwickelt werden, das sich durch einen professionellen Deployment-Prozess auszeichnet und nachhaltig genutzt werden kann.

Meilenstein 2 *Beschaffung der Nachrichtenartikel und Politiker:innen* Die Datenbeschaffung soll effizient durchgeführt werden können, um immer den aktuellen Stand abbilden zu können.

Meilenstein 3 *Aufbau einer Administrationsoberfläche* Um neue Features entwickeln zu können und den Datensatz zu pflegen, soll eine Oberfläche für die Administratoren entwickelt werden.

Meilenstein 4 *Politiker:innen-Profile* Auf Basis der extrahierten Daten sollen Profile über Politiker:innen automatisch erstellt werden.

Meilenstein 5 *Gruppierung der Nachrichten* Nachrichtenartikel über ein Thema sollen automatisch gruppiert werden, damit die Nutzenden die Quellen direkt vergleichen kann.

Meilenstein 6 *Zusammenarbeit mit FaceTheFacts* Das themenverwandte Projekt FaceTheFacts wird ebenfalls im Rahmen des Software-Sprints gefördert. Mögliche Ansatzpunkte für eine Zusammenarbeit sollen, wenn möglich, genutzt werden.

2 Ausführliche Darstellung der Ergebnisse

In diesem Abschnitt beschreiben wir, welche Meilensteine im Zuge des Projektes umgesetzt wurden. Zunächst gehen wir dabei auf die initial definierten Meilensteine ein. Im Anschluss erörtern wir weitere Meilensteine, die wir zusätzlich umsetzen konnten.

2.1 Umsetzung der initial definierten Meilensteine

Die gesetzten Meilensteine wurden erreicht. Meilenstein 5: *Gruppierung der Nachrichten* wurde nicht in vollem Maße umgesetzt, da die Gruppierung für personenbezogene Artikel nicht zuverlässig funktioniert hat (siehe Unterabschnitt 2.3). Stattdessen wurde eine Gruppierung der aktuellen Nachrichten eingebaut. Im Folgendenwerden die Ergebnisse jedes Meilensteins ausführlich vorgestellt.

Meilenstein 1: Entwicklung einer professionellen und skalierbaren Software

Es wurde ein Produktivsystem entwickelt, das aus zwei voneinander unabhängigen Servern besteht. Abbildung 1 bietet eine Übersicht des Systems. Der Frontend-Server ist für die Bearbeitung von Benutzeranfragen verantwortlich und ein zweiter Backend-Server für das Bereitstellen der Daten. Die Kommunikation zwischen diesen beiden Komponenten erfolgt über eine REST-API sowie vorberechneten JSON-Dateien, die auf einen Cloud Storage Server hochgeladen werden können. Eine Komponente im Backend wird außerdem für das regelmäßige Aktualisieren des Datenbestandes eingesetzt. Der Backend-Server wird in einem Docker Swarm Cluster[1] deployt. Je nach Auslastung können mehrere Instanzen der jeweiligen Applikation parallel ausgeführt werden, so dass das System

[1] https://docs.docker.com/engine/swarm/

mit der Nutzerzahl oder größeren Datenmengen skalieren kann. Das Cluster wird von Netcup[2] bereitgestellt. Das Frontend ist unter https://politrack.org erreichbar und wird statisch von Cloudflare Pages[3] gehostet. Aufgrund des statischen Frontendsunddes Verschiebens der Inhalte auf einen Cloud Storage Server, kann die Seite effizient an viele Nutzer gleichzeitig verteilt werden, ohne große Serverlasten oder Kosten zu generieren. Das Deployment beider Komponenten ist mithilfe von GitHub Workflows[4] vollständig automatisiert worden und kann in Zukunft von jedem berechtigten Benutzer über einen Git-Befehl gestartet werden.

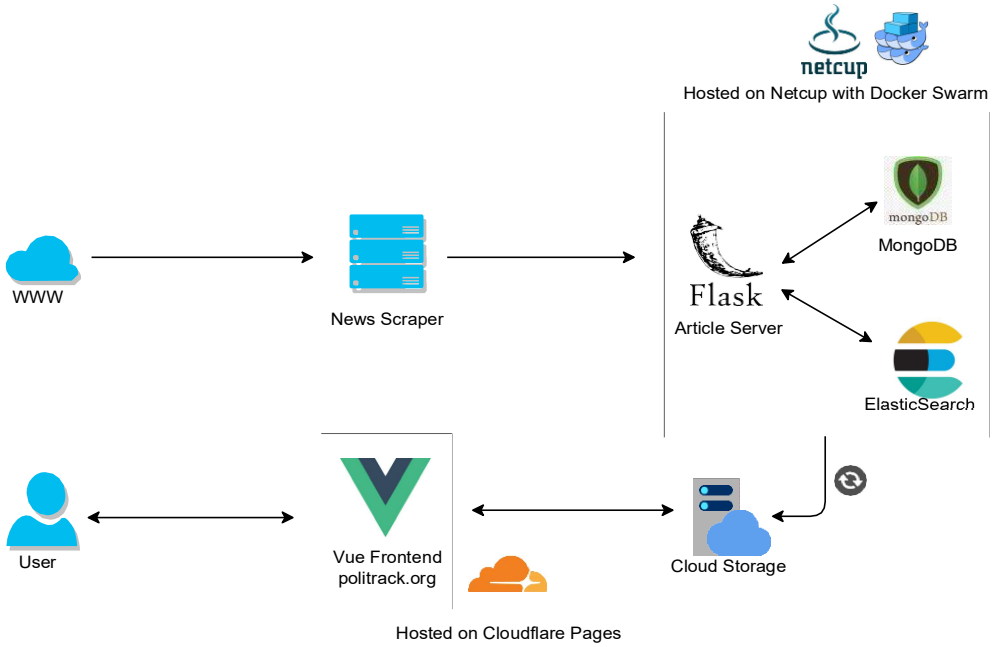


Abbildung 1: Komponenten und verwendete Technologien unseres PoliTrack-Services.

Meilenstein 2: Beschaffung der Nachrichtenartikel und Politiker:innen-Daten

Einer der ersten Meilensteine für PoliTrack war die Datenbeschaffung. Für unseren Service benötigen wir sowohl Nachrichtenartikel als auch grundlegende Informationen über Politiker:innen.

Nachrichtenartikel Zunächst haben wir die Webseiten der größten Zeitungen und Zeitschriften in Deutschland untersucht. Unsere Applikation ist in der Lage, die Seiten mithilfe des Scrapy[5] Frameworks auf neue Artikel zu durchsuchen. Die neuen Artikel werden dann

[2] https://www.netcup.de/
[3] https://pages.cloudflare.com/
[4] https://docs.github.com/en/actions/using-workflows
[5] https://scrapy.org/

über eine Rest-API auf dem Backend-Server gespeichert. Diese Architektur erlaubt es uns, das Extrahieren neuer Daten losgelöst von der restlichen Anwendung auf dedizierten Servern durchzuführen. Zu den heutigen Quellen gehören BILD, Frankfurter Allgemeine, Welt, Zeit, Spiegel, Süddeutsche, Tagesschau und taz. Der Aufbau unserer News Scraper erlaubt es uns in Zukunft, mit geringem Aufwand weitere Quellen hinzuzufügen. Ein gespeicherter Nachrichtenartikel enthält neben dem Titel, Text, Autor und Datum der Veröffentlichung auch Informationen über Schlüsselworte, die vom Herausgeber mit dem Artikel verknüpft wurden. Kostenpflichtige Artikel kennzeichnen wir ebenfalls. Hier ist es uns nur möglich, die Vorschau über den Inhalt abzuspeichern. Sobald neue Artikel in die Datenbank eingetragen werden, werden die rechenintensiven KI-Algorithmen auf dem Backend Server gestartet, um die Daten für die Hauptseite zu aktualisieren. Die aktualisierten Inhalte werden auf den Cloud Storage Server hochgeladen, von dem aus sie genutzt werden können. Für https://politrack.org haben wir mehr als eine halbe Million Artikel zwischen 2015 und 2022 untersucht (siehe Abbildung 2).



Abbildung 2: Anzahl der gespeicherten Artikel nach Monat und Jahr. Stand: März 2022.

Politiker:innen Die Informationen über die deutschen Poltiker:innen zu beschaffen, wurde uns durch die hervorragende Arbeit von Abgeordnetenwatch[6] enorm erleichtert. Abgeordnetenwatch liefert eine dedizierte Schnittstelle über die man grundlegende Informationen wie Name, Alter, Profession und auch jetzige und vergangene Mandate abfragen kann. Damit wir mit den Daten arbeiten können, klonen wir in regelmäßigen Abständen den Datensatz von Abgeordnetenwatch und speichern ihn in unserer Datenbank ab.

Meilenstein 3: Aufbau einer Administrationsoberfläche

Aufgrund der Größe des verwendeten Datensatzes ist es sehr schwierig diese Daten zu überblicken. Aus diesem Grund haben wir eine dedizierte Anwendung geschrieben, die uns das Administrieren der Seite und das Entwickeln neuer Funktionen erleichtert hat (siehe Abbildung 3). So haben wir beispielsweise verschiedene Analysen und Suchfunktionen eingebaut, mit denen wir fehlerhafte Artikel schnell identifizieren können. Außerdem haben wir jedes größere Feature, das heute auf https://politrack.org zu finden ist, zunächst dort getestet. Die Infrastruktur des Backend-Servers hat uns erlaubt, die Auswertungen anstatt über eine gecachte JSON-Datei, über eine Rest-API neu zu berechnen. Die Parameteroptimierung der verwendeten Algorithmen konnte dadurch bequem über die

[6] https://www.abgeordnetenwatch.de

Administrationsoberfläche geschehen. Ein Feature, das derzeit noch keine Anwendung auf der Hauptseite findet, ist diesemantische Suche, diewirschon bei dem Vorgängerprojekt Collabovid[7] erfolgreich eingesetzt haben. Diese Suche könnte es Journalistinnen und Journalisten erleichtern, Nachrichtenartikel zu einem bestimmten Thema zu finden und ist eine der möglichen Erweiterungen für die Zukunft (siehe Unterabschnitt 3.1).

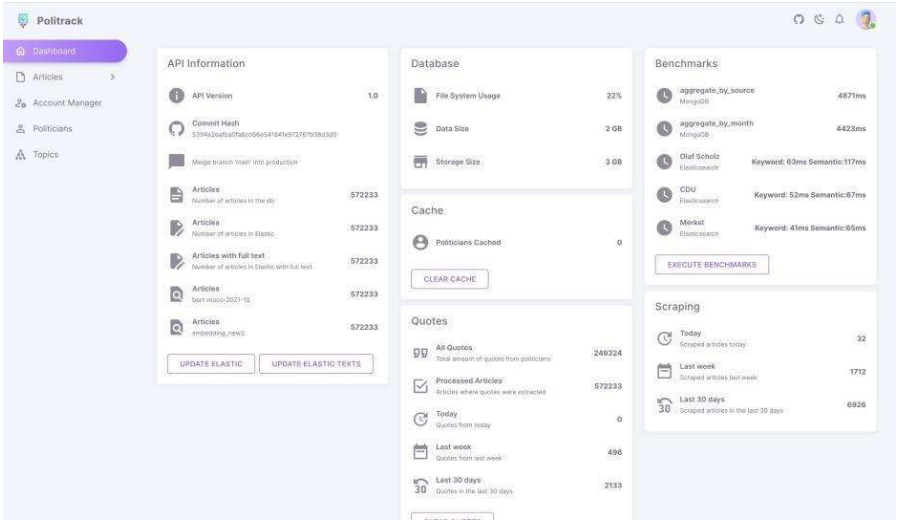


Abbildung 3: Die Administrationoberfläche von PoliTrack.

Meilenstein 4: Politiker:innen-Profile

Diese Menge an Artikeln erlaubt es uns, einige interessante Auswertungen zu machen. Wir speichern, welche Politiker:in zu welchen Nachrichtenartikeln gehört. Basierend auf dieser Information erstellen wir für jeden/jede Politiker:in ein Profil. Mithilfe der Namen und der Inhalte der Nachrichtenartikel kann eine Zuordnung von Artikeln zu Politiker:innen erfolgen. Sobald eine Person in einem Artikel erwähnt wird, wird dieser Inhalt mit der Person verknüpft. Auf dieser Datenbasis erstellen wir Politiker:innen-Profile, in denen die aktuellen Nachrichten über die jeweilige Person dargestellt werden. Mehr noch können andere Personen identifiziert werden, die oft mit einem Politiker oder einer Politikerin erwähnt werden. Auswertungen in dem Profil zeigen an, wann die Präsenz eines Politikers in den Medien besonders hoch war und aus welchen Medien die Artikel stammen. Ein hohes Aufkommen einer Person in den Medien korreliert meist mit einem besonderen politischen Ereignis, wie zum Beispiel einer Wahl. Die Politiker:in-Profile wurden um weitere Features ergänzt, die sich erst im Laufe des Projektes als nützlich und möglich erwiesen haben (siehe Unterabschnitt 2.2)

[7] https://collabovid.org

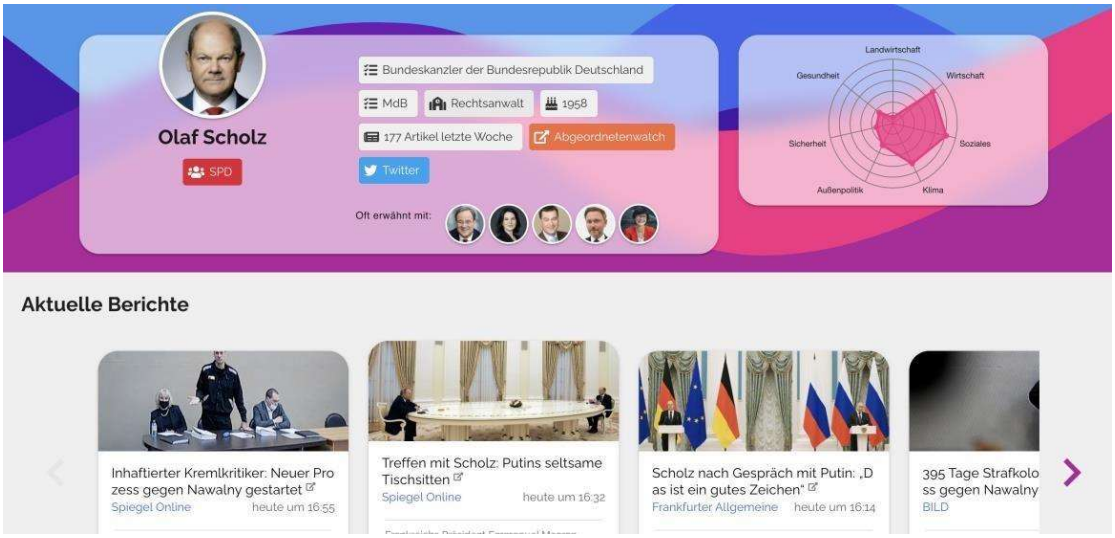


Abbildung 4: Exemplarisches Politiker-Profil von Olaf Scholz.

Meilenstein 5: Gruppierung der Nachrichten

Auf der Übersicht-Seite[8] zeigen wir die „Schlagzeilen des Tages" an (siehe Abbildung 5). Dabei gruppieren wir einzelne Schlagzeilen thematisch und verweisen auf die Artikel zu dem jeweiligen Thema. Dadurch kann man sich eine umfangreichere Meinung über ein Thema bilden und die Berichte von verschiedenen Herausgebern auf einen Blick vergleichen. Wir haben lange daran gearbeitet, die Gruppierung der Artikel zu verbessern. Es hat sich herausgestellt, dass die Feature-Extraktion mittels Tf-idf-Maß und das anschließende Clustering-Verfahren HDBSCAN [9] hier die besten Ergebnisse liefern. Mithilfe der zuvor extrahierten Assoziationen, können wir einer Themengruppe wie zum Beispiel den Koalitionsverhandlungen nach einer Wahl alle mit den Artikeln assoziierten Politiker:innen zuordnen und anzeigen. Dadurch kann der Nutzer auf einen Blick alle relevanten Personen zu diesem Thema sehen. Eine thematische Gruppierung von Artikelmengen, die mit einer bestimmten Person verknüpft sind, wurde ebenfalls getestet. Unsere Versuche haben gezeigt, dass die Zuordnung durch die dünnere Datenlage äußerst schwierig wird. In Unterabschnitt 2.3 gehen wir auf dieses Problem ein.

Meilenstein 6: Zusammenarbeit mit FaceTheFacts

Während der Entwicklung haben wir mit dem Projekt *FaceTheFacts* zusammengearbeitet. Die App macht alle wichtigen Informationen über Politiker:innen per Handykamera direkt am Wahlplakat abrufbar. Dazu gehören etwa politische Positionen, Nebentätigkeiten (bei MdB), vergangenes Abstimmungsverhalten oder der Lebenslauf. Wir haben den Entwicklern eine Schnittstelle geboten, mit dessen Hilfe sie die zu einem Politiker verknüpften

[8] https://politrack.org/#/dashboard
[9] https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

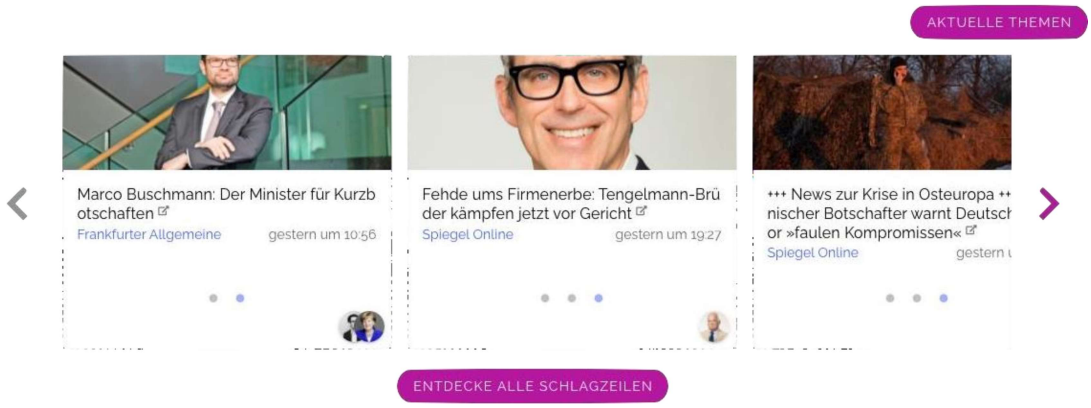


Abbildung 5: Schlagzeilen des Tages durch Gruppierung der Nachrichten.

Nachrichtenartikel abfragen können. Die Artikel ergänzen die Informationen über die Politiker um eine aktuelle Komponente.

2.2 Iterativ neu definierte Meilensteine und deren Umsetzung

Durch den iterativen Entwicklungsprozess wurden während des Förderzeitraumes mehrere zusätzliche Meilensteine definiert, die vollständig erreicht werden konnten und im Folgenden ausführlich beschrieben werden.

Meilenstein 7: Themenextraktion

Durch den Einsatz von Machine-Learning-Techniken, den sogenannten Transformer-Modellen, können wir weitere Informationen über die Themen extrahieren, mit denen sich bestimmte Politiker:innen beschäftigen. Transformer-Modelle sind in der KI-Forschung das State-of-the-Art-Werkzeug, um dem Computer zu ermöglichen, Sprache/Text zu verstehen. Dabei wird der zu analysierende Text zunächst in eine Repräsentation der im Text vorkommenden Themen und Inhalte überführt. Auf Basis dieser Repräsentation können im Anschluss Funktionen wie Themen-Extraktion oder eine Ähnlichkeitssuche stattfinden. Vergleichen wir nun die Repräsentationen der Artikeltexte mit denen bestimmter Themenbeschreibungen oder Schlagwörter, korrelieren Ähnlichkeiten im Inhalt mit Ähnlichkeiten in der Repräsentation. Durch diese grobe Kategorisierung der Artikel, können wir dann eine Übersicht geben, mit welchen politischen Themen sich eine Menge von Artikeln am meisten beschäftigt. Zu diesen Kategorien gehören zum Beispiel *Soziales*, *Außenpolitik*, *Gesundheit* oder *Klimaschutz*.

Meilenstein 8: Extraktion von Zitaten

Politische Nachrichtenartikel enthalten oft Zitate von Politiker:innen. Diese sind in verschiedenen Formen unstrukturiert im Text verteilt. Eine strukturierte Datenbank mit Zitaten und den zugehörigen Autoren bietet eine interessante Datenbasis für weitere

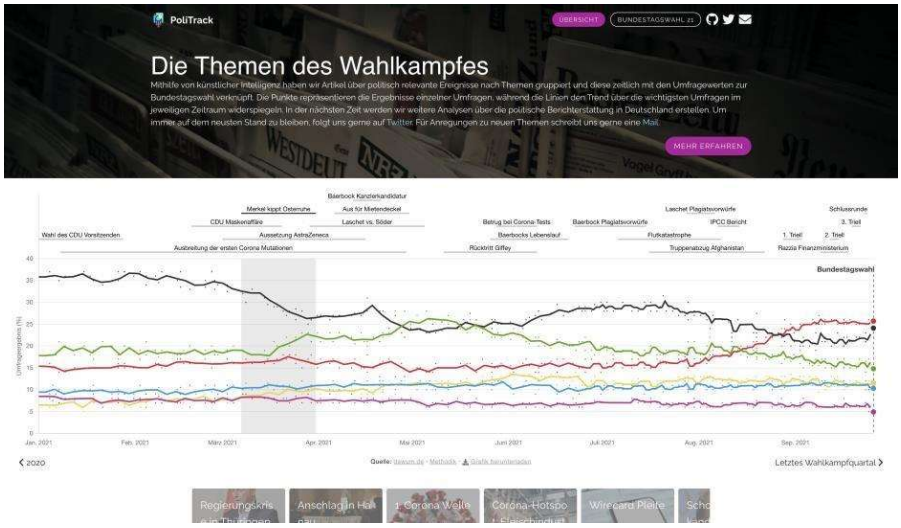


Abbildung 6: Übersicht zur Bundestagswahl 2021.



Abbildung 7: Beispiele für automatisch extrahierte Zitate.

Analysen. Daher extrahieren wir aus jedem Artikel die Zitate und korrespondierende Politiker:innen. Zur Extraktion der Zitate verwenden wir die Stanford Core NLP Bibliothek[10]. Jedoch extrahiert diese Bibliothek teilweise fehlerhafte Zitate oder ordnet Zitate dem/der falschen Politiker:in zu. Daher haben wir ein Tranformer-Modell trainiert, was gegeben einen Textausschnitt und ein Zitat, klassifiziert, ob das Zitat korrekt extrahiert wurde. Obwohl dieser Ansatz die Korrektheit der Zitate deutlich verbessert hat, kommen manchmal falsch extrahierte Zitate vor. Der Grund für die Fehler ist oft, dass die Zitationen in unterschiedlichsten sprachlichen Stilen vorkommen. Wegen der Fehleranfälligkeit zeigen wir überall eine Warnung an, wo die Zitate auf unserer Seite dargestellt sind und arbeiten weiter daran, das Modell zu verbessern. Auf dem Politiker:in-Profil stellen wir die neuesten Zitate eines/einer Politiker:in dar. Auf der Übersicht-Seite zeigen wir die Zitate der populärsten Politiker (siehe Abbildung 7) an.

[10]https://stanfordnlp.github.io/CoreNLP/quote.html

Meilenstein 9: Bundestagswahl 2021

Anlässlich der Bundestagswahl 2021 haben wir die Umfragewerte der größten Parteien in Zusammenhang mit den politischen Großereignissen der Jahre 2020 und 2021 gebracht (siehe Abbildung 6). Dafür wurden über 200.000 Nachrichtenartikel automatisch analysiert. Auf der Seite können Benutzer:innen in einer interaktiven Übersicht die Umfragewerte und Ereignisse gleichermaßen betrachten. Zu jedem Ereignis können direkt passende Nachrichtenartikel aus den zuvor genannten Quellen eingesehen werden. Die Zuordnung erfolgte halbautomatisch mithilfe eigens trainierter Transformer-Modelle. Die Übersicht zeigt, wie effektiv sich unsere Ansätze anwenden lassen und wie gut die Themenextraktion funktioniert. Es ist denkbar, vergleichbare Übersichten zu anderen Anlässen oder anderen Themen automatisch aus den Nachrichtenartikeln zu generieren (siehe Unterabschnitt 3.1).

Meilenstein 10: Gesamtübersicht der aktuellen Nachrichtenlage

Wir liefern nicht nur eine spezifische Übersicht über Politiker:innen, sondern auch über das aktuelle Tagesgeschehen in der Politik. Auf der Übersicht-Seite [11] zeigen wir die „Schlagzeilen des Tages". Dabei gruppieren wir einzelne Schlagzeilen thematisch und verweisen auf die Artikel zu dem jeweiligen Thema. Dadurch kann man sich eine umfassendere Meinung über ein Thema bilden und die Berichte von verschiedenen Herausgebern auf einen Blick vergleichen. Aktuelle Artikel werden außerdem ebenfalls in die Themenbereiche eingeteilt, die wir auf den Profilseiten verwenden. Daraus lässt sich auf einen Blick ableiten, auf welchen Bereich sich die Medien heute fokussieren. Basierend auf den Schlagworten eines Artikels klassifizieren wir, ob ein Artikel von einer bestimmten Partei handelt. Damit lässt sich berechnen, welche Parteien besonders im Fokus der Medien sind. Zudem zeigen wir auch, welche Politiker:innen einer Partei besonders viel in den Medien vorkommen und identifizieren Newcomer, die im Vergleich zu früheren Monaten nun besonders oft in Artikeln erscheinen.

2.3 Arbeiten, die zu keiner Lösung geführt haben

Wir haben sehr viel Zeit in eine gezielte Themenextraktion für Artikel über einzelne Personen ähnlich dem Clustering für die „Schlagzeilen des Tages" investiert. Da nur wenige Artikel über bestimmte Themen mit Personen verknüpft sind, hat sich die Extraktion als äußerst kompliziert herausgestellt. Uns war es bislang nicht möglich, die Algorithmen so anzupassen, dass sie die Artikel einer Person zuverlässig in Themen aufteilen können. Nach vielen Tests und Fehlversuchen haben wir diese Erweiterung zugunsten anderer Features (siehe Unterabschnitt 2.2) auf einen späteren Zeitpunkt verschoben.

2.4 Ergebnisse bei anderen Stellen

Die jetzige Plattform ist ein Informationsportal, auf der sich Bürger:innen über die derzeitige politische Lage sowie über einzelne Personen informieren können. Unsere Daten und Auswertungen sind jedoch auch für die Recherchearbeit von Journalist:innen von

[11] https://politrack.org/#/dashboard

großem Wert. Während der Umsetzungsphase haben wir an einem Coaching teilgenommen, um die Details unserer Visionen für das Projekt herauszuarbeiten. Im Zuge dessen haben wir mögliche Anwendungsmöglichkeiten und Features für eine Rechercheplattform entwickelt. Wir haben viele Interviews mit Journalist:innen geführt, um unsere Ideen zu validieren. Aus diesem Austausch und dem Feedback sind viele neue Ideen für zusätzliche Funktionen der Seite entstanden (siehe Unterabschnitt 2.2).

3 Schlussbetrachtung

3.1 Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Das Projekt ist an alle Bürger:innen gerichtet, zielt aber speziell auf diejenigen ab, die sich vor einer Wahlfür einen/eine Kandidat:in entscheiden müssen. Darüber hinaus hilft die Übersicht Journalist:innen, aktuelle Aussagen mit Bezug auf die Vergangenheit in Kontext zu setzen und trägt somit zur kritischen Berichterstattung bei. Die Plattform soll ein Bewusstsein über die Meinungen und Tatenunserer politischen Entscheidungsträger schaffen. Dies soll die Transparenz in politischen Entscheidungsprozessen stärken und voreilig getroffene populistische Aussagen entlarven. Es gibt viele Möglichkeiten, PoliTrack zu erweitern. Zunächst einmal planen wir, die Themenextraktion auch für Politiker:innen - Profile zu ermöglichen. Die Darstellung der relevanten Themen (anstatt der Artikel) würde die Übersicht der Profile verbessern. Dedizierte Themenseiten könnten den Nutzenden außerdem ermöglichen, zusammenhängende Artikel über einen langen Zeitraum zu verfolgen. So wäre beispielsweise eine Sammlung aller Inhalte zum Wirecard-Skandal denkbar, auch wenn dazugehörige Artikel über einen langen Zeitraum veröffentlicht wurden. Eine zur Bundestagswahlübersicht analoge Ansicht über Ereignisse, die kleinere (Unter-)Ereignisse darstellt, wäre ebenfalls denkbar. Das Hinzufügen weiterer Datenquellen ist ebenfalls geplant und durch den modularen Aufbau der Komponente sehr einfach. Eine kommerzielle Nutzung der Daten für Journalist:innen ist ebenfalls denkbar. Wir haben bereits mit der Möglichkeit einer semantischen Suche experimentiert. In Interviews mit Journalist:innen haben wir außerdem viele Ideen für nützliche Funktionen einer solchen Plattform sammeln können. Da das Projekt unter einer freien Lizenz veröffentlicht ist, können auch andere Interessierte die Codebasis nutzen, um Analysen auf textbasierten Datendurchzuführen.

3.2 Präsentationsmöglichkeiten für mögliche Nutzer:innen

Der Source Code ist auf der GitHub-Projektseite[12] von PoliTrack verfügbar. Informationen über den aktuellen Stand des Projektes, erreichte Meilensteine und weiteres mehr, veröffentlichen wir über den Twitter-Account @pltrck[13]. Die Finanzierung der Server-Infrastruktur ist mindestens bis 12/2022 gesichert, so dass der Service zunächst weiter unter https://politrack.org zur Verfügungsteht.

[12] https://github.com/politrack
[13] https://twitter.com/pltrck

3.3 Einhaltung der Arbeits- und Kostenplanung

Alle umgesetzten Meilensteine konnten im verfügbaren Zeitrahmen erreicht werden. Durch den Zuspruch und den Austausch mit Journalist:innen, Partnern und Freunden wurden wir motiviert und inspiriert, auch weitere, nachträglich definierte Meilensteine umzusetzen.