Richtlinie zum „Software-Sprint“
Collabovid Abschlussbericht

Zuwendungsempfänger:

Jonas Dippel, Yannic Lieder, Eike Niehs, Michael Perk, Moritz Pfister GbR

August 2020

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS20S07 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

1 Kurze Darstellung der Aufgabenstellung und Motivation

Seit dem globalen Ausbruch der Corona-Pandemie Anfang 2020 beschäftigen sich Virolog:innen, Mediziner:innen, Biolog:innen und andere Wissenschaftler:innen auf der ganzen Welt mit SARS-CoV-2 und der Infektionskrankheit COVID-19. Dadurch kommt es zu einer unverhältnismäßig hohen Anzahl neuer Publikationen, an manchen Tagen über 450 Dokumente (siehe Abbildung 1). Durch die mit dem Ausbruch verbundene Dringlichkeit, neue Forschungsergebnisse schnellstmöglich publik zu machen, haben viele dieser Artikel bei Veröffentlichung noch keinen offiziellen Peer-Review-Prozess durchlaufen, wie es in der Wissenschaft eigentlich üblich ist. Stattdessen werden die Ergebnisse ohne externe Überprüfung als Vorveröffentlichungen (Preprints) auf sogenannten Preprint-Servern hochgeladen.

Dieses Phänomen führt zu zwei verschiedenen Problemen: Erstens ist es für Wissenschaftler:innen kaum noch möglich, einen Überblick über die immense Menge an neuen Veröffentlichungen zu behalten und zweitens kann man sich auf Ergebnisse aus Preprints nicht ohne weiteres verlassen und muss sorgfältig abwägen, welche Studien professionell durchgeführt worden sind, und welche einen späteren Review-Prozess möglicherweise nicht bestehen werden.

Hier ist die Idee zu Collabovid entstanden, einer Plattform zum schnelleren Durchsuchen, Einordnen und Klassifizieren von Publikationen, die sich mit COVID-19, SARS-CoV-2 und der Corona-Krise 2020 beschäftigen. Zu Beginn der Förderung war bereits ein funktionsfähiger Prototyp in Form eines Webservices vorhanden, der unter anderem automatisch alle relevanten Preprints der Preprint-Server medRxiv und bioRxiv speichert und ein semantisches Durchsuchen sowohl des Inhalts als auch der Metadaten der Artikel

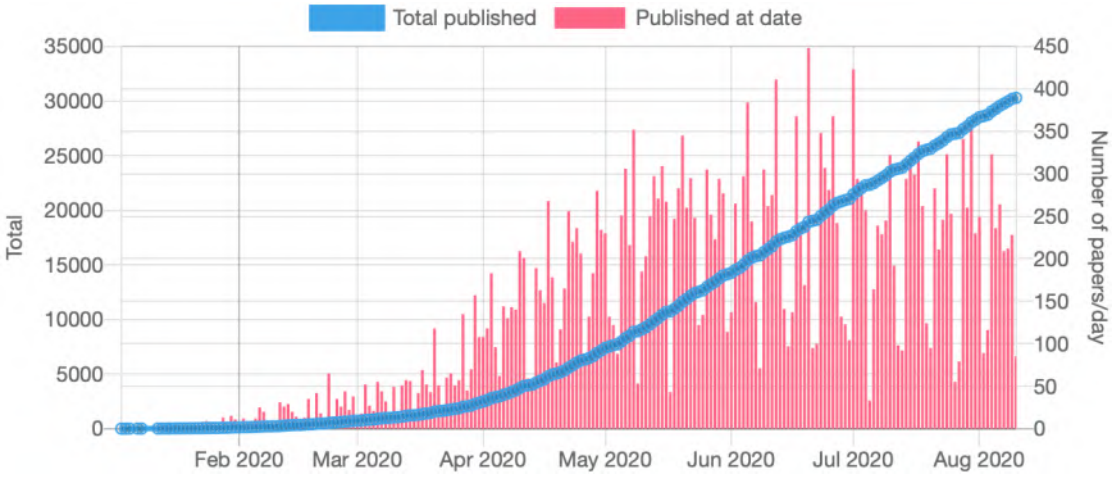


Abbildung 1: Anzahl neuer Publikationen zum Thema SARS-CoV-2, COVID-19 und der Corona-Pandemie

ermöglichte. Für den Förderzeitraum von drei Monaten haben wir uns die folgenden Meilensteine gesetzt:

Meilenstein 1 *Entwicklung einer professionellen und skalierbaren Software*

Auf Grundlage des existierenden Prototypen soll ein skalierbares Produktivsystem entwickelt werden, das sich durch einen professionellen Deployment-Prozess auszeichnet und nachhaltig genutzt werden kann.

Meilenstein 2 *Erweiterung der Datenquellen*

Die bereits im Prototypen vorhandenen Datenquellen medRxiv und bioRxiv werden um weitere Datenquellen erweitert. Als Beispiele sind hier arXiv, PubMed und Elsevier genannt.

Meilenstein 3 *Optimierung der semantischen Suche*

Die semantische Suche soll durch Evaluierung verschiedener Ansätze des maschinellen Lernens optimiert und in geeigneter Weise mit einer klassischen Stichwortsuche kombiniert werden.

Meilenstein 4 *Einordnung der Publikationen in verschiedene Oberthemen*

Es werden aussagekräftige Oberthemen gefunden, in die die Publikationen automatisiert eingeordnet werden, sodass ein gezielteres Durchsuchen der Daten möglich ist.

Meilenstein 5 *Qualitative Bewertung von Vorveröffentlichungen*

Es wird Forscher:innen die Möglichkeit geboten, Vorveröffentlichungen beispielsweise mittels kurzer Anmerkungen inhaltlich zu bewerten.

Meilenstein 6 *Ähnlichkeitsanalyse von Publikationen*

Es werden Methoden zur Bestimmung der Ähnlichkeit zweier Publikationen entwickelt. Dadurch können dem/der Nutzer:in gezielt relevante Publikationen angezeigt werden. Die Methoden können auch für die Klassifizierung (Meilenstein 4) genutzt werden.

Meilenstein 7 *Verbreitung der Publikationen in Sozialen Medien*

Es werden Daten aus sozialen Netzwerken wie Facebook und Twitter sowie weiteren relevanten Portalen ausgewertet, um viel diskutierte Publikationen anzuzeigen.

2 Ausführliche Darstellung der Ergebnisse

2.1 Umsetzung der initial definierten Meilensteine

Die gesetzten Meilensteine wurden weitestgehend erreicht. Meilenstein 7: *Verbreitung der Publikationen in Sozialen Medien* wurde mit Hilfe eines externen Anbieters realisiert und Meilenstein 5: *Qualitative Bewertung von Vorveröffentlichungen* wurde nicht umgesetzt, da wir nach vielen Experten-Interviews feststellen mussten, dass dieses Ziel für uns nicht realisierbar ist (siehe Abschnitt 4). Im Folgenden werden die Ergebnisse jedes Meilensteins ausführlich vorgestellt.

Meilenstein 1: Entwicklung einer professionellen und skalierbaren Software

Es wurde ein Produktivsystem entwickelt, das aus zwei voneinander unabhängigen Django[1]-Apps sowie drei eigenständigen Django-Projekten besteht. Dabei ist eine Komponente für die Bearbeitung von Benutzeranfragen verantwortlich und eine andere für das Berechnen von Suchergebnissen. Die Kommunikation zwischen diesen beiden Komponenten erfolgt über eine REST-API. Eine weitere Komponente wird im Backend für das regelmäßige Aktualisieren des Datenbestandes eingesetzt. Die Programme werden in einem Kubernetes[2]-Cluster deployed. Je nach Auslastung können mehrere Instanzen der jeweiligen Applikation parallel ausgeführt werden, so dass das System mit der Nutzerzahl skalieren kann. Das Cluster wird über Amazon Web Services (AWS)[3] bereitgestellt, AWS Elastic Load Balancing[4] verteilt die Anfragen automatisch an die Container (siehe Abbildung 2). Mit dem entwickelten Collabovid Command Line Tool ist ein nahezu vollautomatischer Deployment-Prozess möglich.

Meilenstein 2: Erweiterung der Datenquellen

Die Datenquellen wurden um den Preprint-Server arXiv und die Literatur-Metadatenbank PubMed erweitert. Der Elsevier-Verlag stellt über sein Novel Coronavirus Information

[1] `https://www.djangoproject.com`
[2] `https://kubernetes.io`
[3] `https://aws.amazon.com`
[4] `https://aws.amazon.com/de/elasticloadbalancing`

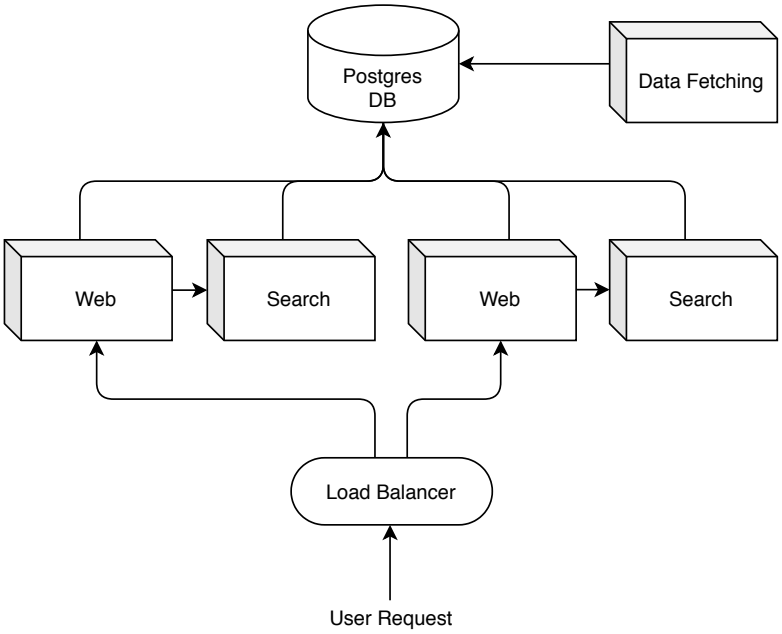


Abbildung 2: Collabovid Backend Infrastruktur

Center[5] sämtliche COVID-19 und SARS-CoV-2 relevante Literatur des Verlags zur Verfügung. Diese Datenquelle wird ebenfalls indexiert. Damit speichert Collabovid insgesamt Literatur von drei Preprint-Servern, eine Literatur-Metadatenbank und alle relevanten Artikel von Elsevier. Dadurch ist ein Großteil aller veröffentlichten und vorveröffentlichten Artikel zu dem Thema über Collabovid zu finden.

Meilenstein 3: Optimierung der semantischen Suche

Um den Sucherfolg für den/die Nutzer:in zu maximieren, haben wir eine klare Trennung zwischen einer semantischen Suche und einer klassischen Stichwortsuche vorgenommen. Dadurch bleibt es dem/der Nutzer:in überlassen, welche Art der Suche er/sie bevorzugt. Für die Stichwortsuche wird ein Elasticsearch[6] Backend benutzt. Die semantische Suche nutzt das bereits bei der Ähnlichkeitsanalyse eingesetzte Machine Learning-Modell (siehe Meilenstein 6: *Ähnlichkeitsanalyse von Publikationen*), um Zusammenhänge zwischen der Nutzereingabe und den Artikeln herzustellen. Die dabei gefundenen Suchergebnisse werden teilweise mittels Elasticsearch umsortiert.

Beide Suchen bieten umfangreiche Filter- und Sortiermöglichkeiten an. So kann zum Beispiel nach Autor:innen, Orten, Veröffentlichungsdatum, Thema oder Kategorie gefiltert werden und nach Relevanz, Veröffentlichungsdatum oder Beliebtheit sortiert werden. Ein Beispiel der Suche ist in Abbildung 3 dargestellt.

[5] `https://www.elsevier.com/connect/coronavirus-information-center`
[6] `https://www.elastic.co/`

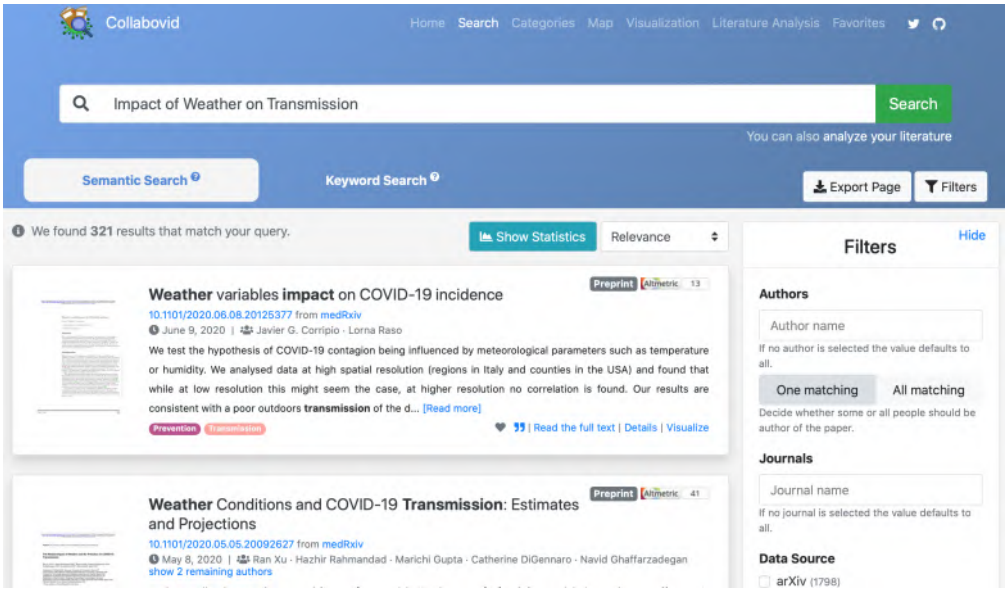


Abbildung 3: Semantische Suche auf Collabovid

Meilenstein 4: Einordnung der Publikationen in verschiedene Oberthemen

Die Einordnung der Publikationen in verschiedene Oberthemen geschieht mittels eines Machine Learning-Modells. Das Modell wurde auf dem LitCovid[7] Datensatz trainiert, der Publikationen enthält, die in acht Kategorien (General, Mechanism, Transmission, Diagnosis, Treatment, Prevention, Case Report, Forecasting) eingeteilt wurden. Das genutzte neuronale Netz basiert auf dem kürzlich publizierten Long-Document Transformer[8] (Longformer) Modell. Typischerweise arbeiten vergleichbare Modelle mit einem quadratisch skalierenden Attention Mechanismus. Dieser ermöglicht es, gute Transformationen auf einzelnen Sätzen durchzuführen. Bei langen Texten gibt es jedoch oft Probleme mit der großen Anzahl an Wörtern (Tokens), die verarbeitet werden müssen. Die neuartige Struktur von Longformer ist speziell für längere Dokumente ausgelegt. Wir übergeben dem Modell sowohl den Titel als auch den Abstract des jeweiligen Artikels als Eingabeparameter. Das Ergebnis wird dann von einem zweiten Teilmodell (Classification Head) in eine Zuordnung in die acht Klassen umgewandelt. Die Kombination aus dem zuvor trainierten Longformer Modell und unserem eigens erstellten Klassifizierungs-Zusatz wurde mit Beispielen aus dem LitCovid Datensatz trainiert. Sobald neue Artikel in unseren Datensatz aufgenommen werden, werden diese automatisch in eine (oder mehrere) Kategorien eingeteilt.

Meilenstein 6: Ähnlichkeitsanalyse von Publikationen

Für jede Publikation berechnen wir ähnliche Publikationen, die sich möglichst mit den gleichen Themen auseinandersetzen. Um diese Funktion zu realisieren, haben wir ein

[7] `https://www.ncbi.nlm.nih.gov/research/coronavirus/`
[8] `https://arxiv.org/pdf/2004.05150.pdf`



Abbildung 4: Von Altmetric bereitgestellter *Batch* zur Anzeige des Altmetric Scores

Siamese Netzwerk trainiert, dass auf einem BERT-Modell [9] aufbaut. Dieses Modell bekam während des Trainings die Aufgabe, basierend auf einem Titel und eines Abstracts zu bestimmen, ob beide zu der gleichen Publikation gehören. Das Netzwerk produziert für beide Textabschnitte einen Vektor. Innerhalb des Trainingsprozesses wird die Distanz zwischen den Vektoren von Titeln und Abstracts, die zu der gleichen Publikation gehören, minimiert und die Distanz zwischen denen, die zu unterschiedlichen Publikationen gehören, maximiert. Während des Trainings erhält das Netzwerk sowohl Beispiele, bei denen Titel und Abstract zu derselben Publikation gehören, als auch Beispiele, bei denen Titel und Abstract von zwei unterschiedlichen Publikationen miteinander vertauscht wurden. Damit das Modell eine gute Vorhersage treffen kann, muss es die unterliegenden Themen in den jeweiligen Textabschnitten in den Vektor codieren. Um die Ähnlichkeit einer Publikation A mit einer Publikation B zu bestimmen, berechnen wir die Distanz zwischen dem Titel-Vektor von A und dem Abstract-Vektor von B, sowie die Distanz zwischen dem Titel-Vektor von B und dem Abstract-Vektor von A. Je geringer die Summe der Distanzen ist, umso ähnlicher sind sich zwei Publikationen. Auf der Detail-Seite eines Artikels werden die zehn ähnlichsten Publikationen dargestellt. Unser Modell wird auch für die semantische Suche verwendet, wobei die Distanzen zwischen der Nutzereingabe und dem Titel und Abstract der einzelnen Publikationen berechnet werden.

Meilenstein 7: Verbreitung der Publikationen in den Sozialen Medien

Als Indikator für den Einfluss von Publikationen in den sozialen Medien nutzen wir den Altmetric Score, der von der Firma Altmetric[10] berechnet und uns kostenlos zur Verfügung gestellt wird. Der Altmetric Score einer Publikation ist eine gewichtete Summe der Anzahl an Artikeln, die diese referenzieren sowie der Anzahl von Erwähnungen in verschiedenen sozialen Medien und Fachzeitschriften. Dieser Wert wird für jede Publikation auf Collabovid angezeigt (siehe Abbildung 4). Außerdem können Suchergebnisse auch nach dem Altmetric Score bzw. nach dessen Entwicklung am letzten Tag/in der letzten Woche/dem letzten Monat sortiert angezeigt werden. So kann insbesondere nach Artikeln gesucht werden, die in einem bestimmten Zeitraum besonders stark verbreitet wurden.

[9] `https://arxiv.org/abs/1810.04805`
[10] `https://www.altmetric.com`

2.2 Iterativ neu definierte Meilensteine und deren Umsetzung

Darüber hinaus wurden durch den iterativen Entwicklungsprozess während des Förderzeitraumes mehrere weitere Meilensteine definiert, die vollständig erreicht werden konnten und im Folgenden ausführlich beschrieben werden.

Meilenstein 8: Visualisierung der Artikel-Ähnlichkeiten

Um näher verwandte Artikel zu einer Publikation zu finden und die Zusammenhänge zwischen Artikeln explorativ durchsuchen zu können, sollen die Ähnlichkeiten zwischen Artikeln visuell aufbereitet werden.

Für die Umsetzung dieses Meilensteins werden die der Ähnlichkeitsbestimmung zugrunde liegenden, multidimensionalen Vektoren mittels T-distributed Stochastic Neighbor Embedding (t-SNE) auf dreidimensionale Vektoren reduziert. Diese werden dann mit dem Grafik-Framework threejs[11] clientseitig visualisiert und so durchsuchbar gemacht. Dabei kann unter anderem nach Kategorien gefiltert werden (siehe Abbildung 5, links).

Meilenstein 9: Erkennung von ortsspezifischen Publikationen

Häufig geht es in medizinischen Publikationen in diesem Kontext um Beobachtungen oder Studien, die sich auf bestimmte Städte, Regionen oder Länder beziehen. Um diese ortsspezifischen Publikationen für den/die Nutzer:in aufbereiten zu können, sollen geografische Orte in den Titeln erkannt und auf einer Weltkarte visualisiert werden.

Mit Hilfe des Natural Language Processing Frameworks spaCy[12] werden geopolitische Entitäten aus den Titeln extrahiert. Diese werden mit der GeoNames[13]-Datenbank abgeglichen und darüber die Koordinaten des Ortes bestimmt. Die Koordinaten werden mit der JavaScript-Library jVectorMap[14] auf einer Weltkarte angezeigt. Durch Anklicken eines Landes oder einer Stadt/Region, werden alle Publikationen angezeigt, die sich auf diesen Ort beziehen (siehe Abbildung 5, rechts).

Meilenstein 10: Literaturanalyse

Wissenschaftler:innen arbeiten in der Regel mit Software, mit dessen Hilfe sie Literaturlisten verwalten können. Dazu gehören beispielsweise Programme wie Zotero[15] oder Mendeley[16]. Es soll eine Funktion implementiert werden, die eine eingegebene Literaturliste analysiert und dem/der Nutzer:in darauf basierend relevante Publikationen anzeigt.

In der Umsetzung erlaubt die Analyse die Eingabe einer solchen Liste im bibTeX[17]-Format, da alle gängigen Literaturverwaltungsprogramme einen Export in dieses Format

[11] `https://threejs.org/`
[12] `https://spacy.io`
[13] `https://www.geonames.org`
[14] `https://jvectormap.com`
[15] `https://www.zotero.org`
[16] `https://www.mendeley.com`
[17] `https://www.ctan.org/pkg/bibtex`

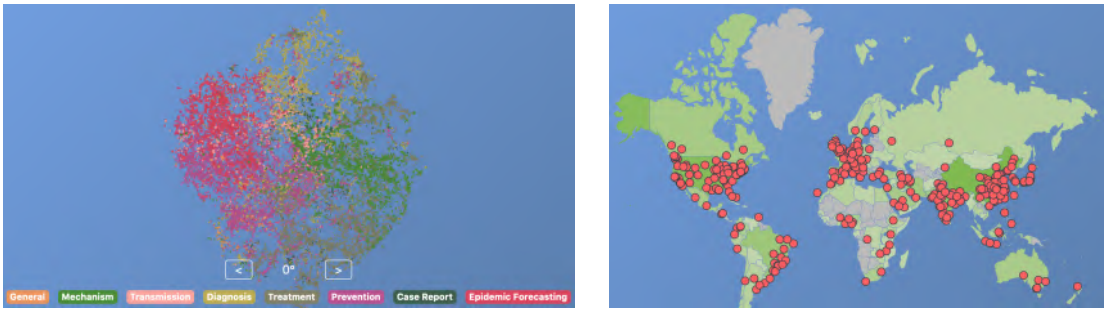


Abbildung 5: Die Ähnlichkeit von Publikationen wird im dreidimensionalen Raum dargestellt (links). Ortsspezifische Veröffentlichungen sind auf einer Weltkarte visualisiert (rechts).

unterstützen. Nach dem Einlesen der Liste, wird die bereits vorhandene Ähnlichkeitsanalyse genutzt, um dem/der Nutzer:in semantisch verwandte Artikel anzuzeigen. Alternativ können auch Publikationen aus einer auf Collabovid erstellten Favoritenliste als Eingabe genutzt werden.

3 Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Die Zielgruppe des Projekts sind Wissenschaftler:Innen, die aktuell an SARS-CoV-2/COVID-19 bezogenen Themen arbeiten. Durch unsere Plattform bieten wir diesen einen schnellen Zugang zu relevanten Publikationen. Vorrangig geht es nicht darum, eine klassische Literatursuchmaschine wie PubMed oder Google Scholar zur Verfügung zu stellen, sondern eine alternative Möglichkeit anzubieten, Publikationen vor allem *explorativ* zu durchsuchen, selbst wenn man zu Beginn der Suche noch nicht genau weiß, wonach man genau sucht. Wir haben von mehreren hundert Wissenschaftler:Innen, die sich aktuell mit dem Coronvirus beschäftigen, Feedback erhalten, das größtenteils sehr positiv ausgefallen ist. So berichteten viele, dass sie über Collabovid relevante Literatur gefunden haben, die sie über andere Plattformen nicht gefunden hatten.

Spezifische Ideen zur Weiterentwicklung beziehen sich hauptsächlich auf die engere Verknüpfung der Publikationen innerhalb des Datensatzes. So könnten Verknüpfungen zwischen zitierenden und zitierten Publikationen, zwischen Publikationen und klinischen Studien oder zwischen Publikationen und bestimmten Metadaten wie Wirkstoffen oder Methodiken hergestellt werden.

Die Plattform ist von Grund auf so generisch implementiert, dass sie mit angemessenem Aufwand auch auf andere Forschungsgebiete neben SARS-CoV-2/COVID-19 übertragen werden kann. Es müssten andere Datenquellen angebunden werden und die Klassifizierung müsste durch fachspezifische Themenbereiche ersetzt und neu trainiert werden.

Da das Projekt unter einer freien Lizenz veröffentlicht ist, können auch andere Interessierte die Codebasis nutzen, um vergleichbare Plattformen für andere Fachgebiete zur Verfügung zu stellen.

4 Arbeiten, die zu keiner Lösung geführt haben

Durch Interviews mit potentiellen Nutzer:innen hat sich herausgestellt, dass das von uns unter Meilenstein 5 geplante informelle Review-System keine sinnvolle Erweiterung der Plattform ist. Dafür müsste die Bereitschaft einer sehr großen Nutzerzahl bestehen, Artikel informell zu reviewen, zum Beispiel in Form von Kommentaren oder Likes. Diese Bereitschaft konnten wir in den geführten Interviews nicht erkennen. Durch die Anzeige des Altmetric Scores und dessen Sortierung haben wir eine Alternative gefunden, die ein Indikator der Qualität von Publikationen sein kann.

5 Präsentationsmöglichkeiten für mögliche Nutzer:innen

Der Source Code ist auf der GitHub-Projektseite[18] von Collabovid verfügbar. Informationen über den aktuellen Stand des Projektes, erreichte Meilensteine und weiteres mehr veröffentlichen wir über den Twitter-Account `@collabovid`[19]. Die Finanzierung der Server-Infrastruktur ist voraussichtlich bis 12/2021 gesichert, so dass der Service zunächst weiter unter `https://www.collabovid.org` zur Verfügung steht. Außerdem ist eine wissenschaftliche Veröffentlichung zu den eingesetzten Machine Learning Verfahren geplant.

6 Einhaltung der Arbeits- und Kostenplanung

Alle umgesetzten Meilensteine konnten im verfügbaren Zeitrahmen erreicht werden. Durch das Verwerfen von Meilenstein 5: *Qualitative Bewertung von Vorveröffentlichungen* konnten mehrere, nachträglich definierte Meilensteine erreicht werden.

7 Ergebnisse bei anderen Stellen

Während der Umsetzungsphase haben wir viele Interviews mit Wissenschaftler:Innen aus verschiedenen Ländern geführt, die sich aktuell mit der Forschung an SARS-CoV-2, COVID-19 oder der aktuellen Corona-Pandemie beschäftigen. Damit konnten wir gezielter auf die Bedarfe potentieller Nutzer:innen eingehen. Erst durch diese Gespräche konnten wir unsere festgelegten Meilensteine überprüfen und haben dadurch neue Anforderungen analysiert, die teilweise zu Änderungen der Meilensteine geführt haben. So haben wir uns dazu entschieden, Meilenstein 5: *Qualitative Bewertung von Vorveröffentlichungen* nicht umzusetzen und bei Meilenstein 3: *Optimierung der semantischen Suche* mehrmals Änderungen vorgenommen und damit auf häufige Nutzerwünsche reagiert.

[18] `https://github.com/collabovid/collabovid`
[19] `https://twitter.com/collabovid`