Optimierung und Implementierung eines Artikel-Empfehlungssystems für die SV-Gruppe auf der Google Cloud Platform

1. Projektarbeit

an der Fakultät für Wirtschaft
im Studiengang Data Science und künstliche Intelligenz

an der
DHBW Ravensburg

Autor/-in:	Finn Hertsch
Kurs:	RV-WDS124
DHBW-Betreuer/-in:	Dr. Andreas Heusler
Dualer Partner:	Schwäbischer Verlag GmbH & Co. KG Drexler, Gessler
Begleitperson Dualer Partner:	Mario Pfob
Abgabefrist:	30.09.2025

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

Formelverzeichnis

1 Einleitung

Die digitale Transformation hat die Art und Weise, wie Nachrichten konsumiert werden, grundlegend verändert. Die stark wachsende Informationsmenge im Internet erschwert es den Nutzern zunehmend, relevante Inhalte zu identifizieren. Dies erzeugt einen erheblichen Bedarf an effektiven Filtermechanismen. In diesem Kontext spielen Empfehlungssysteme (ES) eine entscheidende Rolle, indem sie die Informationsüberlastung von einem Hindernis für das Engagement in eine Chance für personalisierte Inhalte verwandeln und so die Leseerfahrung der Nutzer verbessern (vgl. Wu u. a. 2023). Diese Personalisierung birgt das Risiko, sogenannte Filterblasen zu erzeugen, also eine Verringerung der inhaltlichen Vielfalt in den individuellen Ergebnisräumen (vgl. Nguyen u. a. 2014). Eine Filterblase tritt auf, wenn einem Nutzer nur noch Artikel zu einem spezifischen Thema angezeigt werden. Das führt zu einer Verzerrung der Wahrnehmung von Informationen und begünstigt eine einseitige Meinungsbildung (vgl. Nguyen u. a. 2014)

Über ihre reine Filterfunktion hinaus dienen Empfehlungssysteme als Instrument zur Steigerung quantitativer Metriken wie Nutzerbindung, Verweildauer oder Artikel pro Sitzung. Empirische Arbeiten aus E-Commerce, Streaming und Nachrichten zeigen Verbesserungen von Engagement-Metriken (vgl. Linden, Smith und York 2003; Covington, Adams und Sargin 2016; Raza und Ding 2022).

Die Generierung qualitativ hochwertiger Artikel-Empfehlungen für die Nutzer der Schwäbischer Verlag Gruppe (SV-Gruppe) stellt eine mehrdimensionale Herausforderung dar. Ein zentrales Gütekriterium besteht in der Balance zwischen thematischer Relevanz und inhaltlicher Diversität. Nur so lassen sich die Interessen der Nutzer präzise abbilden und gleichzeitig die Entstehung von Filterblasen vermeiden.

Die technische Realisierung wird durch drei primäre Faktoren erschwert:

1. **Datenvolumen und Skalierbarkeit:** Das System muss über 440.000 Artikel sowie mehr als 6,5 TB an Nutzerinteraktionsdaten effizient verarbeiten können. Diese Kennzahlen basieren auf einer Analyse des im BigQuery-Data-Warehouse der SV-Gruppe vorgehaltenen Gesamtbestandes. Die performante Verknüpfung dieser beiden Datenquellen ist eine grundlegende Anforderung.
2. **Item-Cold-Start-Problem:** Neu publizierte Artikel verfügen definitionsgemäß über keine oder nur sehr wenige Nutzerinteraktionen. Modelle des Collaborative Filtering (CF), wie das hier eingesetzte Neural Collaborative Filtering (NCF)-Modell (He u. a. 2017), können für solche kalten Artikel keine Empfehlungen generieren. Es bedarf daher einer Strategie, um neue Inhalte unmittelbar nach der Veröffentlichung fair und effektiv in den Empfehlungsprozess zu integrieren. Der Ansatz des Content-Based Filtering (CBF), der auf textueller Ähnlichkeit basiert, wirkt diesem Problem entgegen (vgl. Lops, Gemmis und Semeraro 2011).

3. **Balance komplementärer Empfehlungslogiken:** Die beiden Ansätze weisen gegensätzliche Stärken und Schwächen auf. Während das Content-Based Filtering (CBF)-Modell eine hohe thematische Genauigkeit sicherstellt, birgt es die Tendenz, Nutzer in ihren bekannten Interessen zu isolieren. Das Collaborative Filtering (CF)-Modell hingegen kann durch das kollektive Nutzerverhalten neuartige Inhalte entdecken, neigt aber zu einem Popularity Bias (vgl. Abdollahpouri 2019). Die zentrale Problemstellung dieser Arbeit liegt in der Konzeption und Optimierung einer hybriden Architektur, die diese komplementären Eigenschaften gezielt kombiniert. Das Ziel ist es, durch die Fusion der beiden Ansätze eine Empfehlungsliste zu generieren, deren Gesamtrelevanz maximiert wird.

1.1 Zielsetzung und Aufbau der Arbeit

Das primäre Ziel dieser Arbeit ist die Konzeption, Entwicklung und Optimierung eines prototypischen, hybriden Empfehlungssystems für die SV-Gruppe, das auf der Google Cloud Platform [1] implementiert wird. Der Fokus liegt auf der Optimierung einer gewichteten Hybridisierungsstrategie, bei der die Empfehlungslisten eines CBF- und eines CF-Modells mittels einer gewichteten Summe kombiniert werden. Dieser Ansatz bildet eine validierte Grundlage, auf der künftig komplexere Hybridisierungsarchitekturen aufbauen können (vgl. Burke 2002).

Um die Übertragbarkeit der Resultate auf ein produktives Einsatzszenario zu gewährleisten, erfolgt die Evaluation des Systems auf Basis realer Nutzerdaten unter Anwendung einer chronologischen Aufteilung von Trainings- und Testdaten. Ein weiteres Ziel besteht darin zu zeigen, dass der hier gewählte hybride Ansatz eine deutlich höhere Empfehlungsqualität erzielt als etablierte Baseline-Strategien, wie die zufällige Auswahl oder die Empfehlung populärer Artikel (vgl. Wu u. a. 2023).

Die Arbeit ist wie folgt strukturiert:

Kapitel 2 legt die theoretischen Grundlagen für Empfehlungssysteme. Es werden sowohl die in dieser Arbeit angewandten fundamentalen Techniken (CF und CBF) erläutert als auch ein Überblick über fortgeschrittene State-of-the-Art-Konzepte gegeben.

Kapitel 3 beschreibt die technische Architektur und Implementierung des Systems auf der Google Cloud Platform (GCP). Der Fokus liegt auf den Anforderungen eines produktiven Systems und den gewählten Lösungsansätzen zur Erfüllung dieser Anforderungen.

Kapitel 4 präsentiert den Kern der Arbeit: die datengetriebene Optimierung der Gewichtungsfaktoren w_{cbf} und w_{cf} mittels „Optuna“ (Akiba u. a. 2019) sowie die detaillierte Darstellung und Analyse der Evaluationsergebnisse.

[1] Siehe `https://cloud.google.com/`

Kapitel 5 führt eine kritische Diskussion der erzielten Ergebnisse und beleuchtet die Limitationen der vorliegenden Arbeit. Dies schließt eine fundierte Einschätzung der Generalisierbarkeit und der praktischen Implikationen der Resultate ein.

Kapitel 6 fasst die zentralen Erkenntnisse zusammen und liefert einen Ausblick auf potenzielle Weiterentwicklungen, die auf dem hier entwickelten Prototypen aufbauen und einen inkrementellen Mehrwert generieren könnten.

2 Theoretische Grundlagen

Das Forschungsfeld der Empfehlungssysteme ist breit und umfasst zahlreiche methodische Ansätze. In der Praxis haben sich jedoch zwei grundlegende Paradigmen als besonders einflussreich erwiesen: Das CBF, das auf Inhaltsmerkmalen der zu empfehlenden Items basiert und das CF, das aus dem kollektiven Verhalten der Nutzer lernt. Aufgrund der komplementären Stärken und Schwächen dieser Systeme werden hybride Architekturen in der Nachrichtenbranche als praktikabler Ansatz beschrieben, um Relevanz, Diversität und Cold-Start auszubalancieren; sie bilden eine solide Grundlage für prototypische Systeme (vgl. Wu u. a. 2023; Raza und Ding 2022).

2.1 Content-based Filtering

Content-based Filtering (CBF) ist ein Empfehlungsansatz, der auf den Inhaltsmerkmalen von Items basiert, um Ähnlichkeiten zwischen ihnen zu bestimmen (vgl. Lops, Gemmis und Semeraro 2011). In der vorliegenden Arbeit wird eine spezifische Form des CBF, die Item-zu-Item-Ähnlichkeit, angewendet. Anstatt ein langfristiges Nutzerprofil aus der Lesehistorie zu erstellen, wird hierbei ausschließlich der aktuell vom Nutzer gelesene Artikel als Referenzpunkt genommen, um thematisch verwandte Inhalte zu finden. Das Nutzerverhalten dient in diesem Ansatz also nicht zur Profilbildung, sondern lediglich zur Auswahl des initialen Artikels.

Technisch wird dazu der Inhalt jedes Artikels – bestehend aus Titel und Text – in eine semantische, maschinenlesbare Form überführt. Dies geschieht durch den Einsatz von Transformer-basierten Sprachmodellen (vgl. Vaswani u. a. 2017), die den Text in hochdimensionale numerische Vektoren, sogenannte Embeddings, transformieren. Jeder Vektor repräsentiert dabei die semantische Position eines Artikels in einem vieldimensionalen Vektorraum. Die thematische Ähnlichkeit zwischen zwei Artikeln kann anschließend über gängige Maße wie Kosinus-Ähnlichkeit, Maximum Inner Product Search (MIPS) oder euklidische Distanz quantifiziert werden (vgl. Reimers und Gurevych 2019). Zunächst gilt

$$\cos(u, v) = \frac{u^T \cdot v}{\|u\| \, \|v\|} \tag{2.1}$$

Bei L2-normalisierten Vektoren sind Kosinus-Ähnlichkeit und Skalarprodukt identisch ($\|u\| = \|v\| = 1$); die euklidische Distanz ist eine monotone Transformation davon, sodass die Rangordnung erhalten bleibt.

$$\cos(u, v) = u^T \cdot v$$

In der Praxis wird daher häufig das Skalarprodukt auf normalisierten Vektoren verwendet.

Eine zentrale Herausforderung bei der Implementierung von CBF-Systemen ist die performante Durchführung dieser Ähnlichkeitssuche in Echtzeit, insbesondere bei einem großen Korpus von hunderttausenden Artikeln. Eine naive Brute-Force-Suche, bei der jeder Vektor mit jedem anderen verglichen wird, ist für produktive Anwendungen zu rechenintensiv und langsam. Daher kommen spezialisierte Vektorindizes für die Suche nach ungefähren nächsten Nachbarn (Approximate Nearest Neighbour (ANN)) zum Einsatz. Diese Indizes partitionieren und komprimieren den hochdimensionalen Vektorraum, um eine Suche in logarithmischer statt linearer Zeit zu ermöglichen. Moderne Techniken erreichen dies durch Ansätze wie die Vektor-Quantisierung, zum Beispiel mittels Anisotropic Vector Quantization (AVQ). Auch fortgeschrittene Methoden wie Spilling with Orthogonality-Amplified Residuals (SOAR), die auf redundanten Repräsentationen mit orthogonalisierten Residuen basieren, reduzieren den Speicherbedarf und die Suchlatenz drastisch. Dieser Geschwindigkeitsgewinn wird durch einen kontrollierten, geringen Recall-Verlust erreicht (vgl. Guo u. a. 2020; Sun u. a. 2023).

2.2 Collaborative Filtering

Im Gegensatz zum Content-Based Filtering, das auf den Inhaltsmerkmalen von Items basiert, nutzt das Collaborative Filtering (CF) das kollektive Verhalten und die Interaktionsmuster aller Nutzer, um Präferenzen zu modellieren. Der Grundgedanke ist, dass Nutzer, die in der Vergangenheit ähnliche Vorlieben zeigten, auch in Zukunft ähnliche Interessen haben werden. Eine besondere Herausforderung in Domänen wie Nachrichtenportalen besteht darin, dass die Interaktionen typischerweise als implizites Feedback vorliegen – beispielsweise als Klicks oder Lesezeit – und nicht als explizite Bewertungen wie eine Sterne-Vergabe. Dies erfordert spezielle mathematische Verfahren, um aus dem reinen Vorhandensein einer Interaktion eine Präferenz abzuleiten.

Ein etablierter und leistungsstarker Ansatz zur Modellierung impliziter Daten ist die gewichtete Matrixfaktorisierung (vgl. Hu, Koren und Volinsky 2008). Sie zielt darauf ab, die User-Item-Interaktionsmatrix in zwei niedrigdimensionale Matrizen zu zerlegen, die latente, also verborgene, Merkmale von Nutzern und Items repräsentieren. Das zugrundeliegende Optimierungsproblem wird durch die folgende Formel beschrieben:

$$\min_{\{x_u\},\{y_i\}} \sum_{u,i} c_{ui} \left(p_{ui} - x_u^\top y_i\right)^2 \; + \; \lambda\Big(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2\Big) \tag{2.2}$$

Diese Formel minimiert den Fehler zwischen den Vorhersagen und den tatsächlichen Interaktionen, gewichtet nach der Vertrauenswürdigkeit (Konfidenz) jeder Beobachtung.

- x_u und y_i: Dies sind die latenten Vektoren, die die verborgenen Präferenzen eines Nutzers u bzw. die Eigenschaften eines Items i repräsentieren. Das Ziel des Modells

ist es, diese Vektoren zu lernen.

- $x_u^\top y_i$: Das Skalarprodukt der beiden Vektoren. Es dient als Vorhersage, wie wahrscheinlich eine Interaktion zwischen Nutzer u und Item i ist.
- p_{ui}: Eine binäre Präferenzvariable, die angibt, ob eine Interaktion zwischen Nutzer u und Item i beobachtet wurde ($p_{ui} = 1$) oder nicht ($p_{ui} = 0$).
- c_{ui}: Der Konfidenz-Term, der angibt, wie viel Vertrauen in die Beobachtung p_{ui} gelegt wird. Er wird oft als $c_{ui} = 1 + \alpha r_{ui}$ berechnet, wobei r_{ui} die Häufigkeit der Interaktion (z.B. Anzahl der Klicks) und α ein Skalierungsparameter ist. Dies ermöglicht es dem Modell, wiederholten Interaktionen eine höhere Bedeutung beizumessen.
- λ: Ein Regularisierungsparameter, der verhindert, dass die Werte in den latenten Vektoren zu groß werden (Overfitting) und somit die Generalisierungsfähigkeit des Modells verbessert.

Ein modernerer Ansatz, das Neural Collaborative Filtering (NCF), erweitert dieses Prinzip, indem es das lineare Skalarprodukt ($x_u^\top y_i$) durch ein nichtlineares neuronales Netz, ein Multi-Layer Perceptron (MLP), ersetzt (vgl. He u. a. 2017). Siehe Abbildung 2.1. Dadurch kann das Modell komplexere und subtilere Zusammenhänge in den Nutzer-Item-Interaktionen erfassen. NCF-Modelle werden typischerweise mit den beobachteten positiven Interaktionen und einer Auswahl an zufällig gezogenen negativen Beispielen (Negative Sampling) oder den gesamten negativen Artikeln trainiert, was für die Modellqualität entscheidend ist.

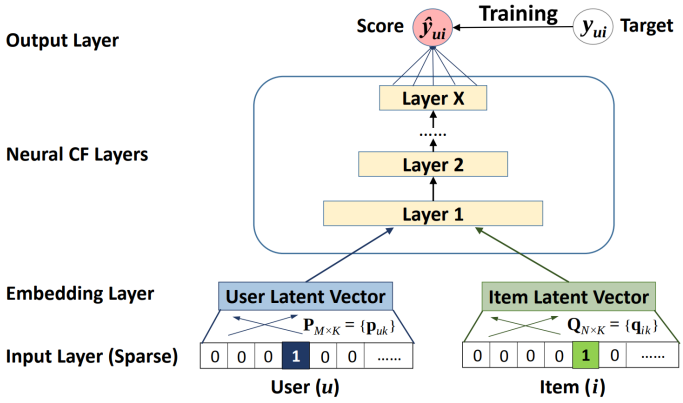

Abb. 2.1: Schematische Darstellung der Architektur eines Neural-Collaborative-Filtering-Modells Quelle: He u. a. 2017.

Trotz ihrer Leistungsfähigkeit weisen CF-Methoden zwei systematische Schwächen auf. Erstens leiden sie unter dem Item-Cold-Start-Problem, da sie für neue Artikel ohne Interaktionshistorie keine Empfehlungen generieren können. Zweitens neigen sie zu einem Popularity Bias, bei dem bereits sehr populäre Artikel überproportional oft empfohlen werden, was die Vielfalt und Personalisierung einschränkt (vgl. Abdollahpouri 2019). Um

diese Nachteile zu mitigieren, wird der CF-Ansatz in dieser Arbeit mit einem Content-Based-Ansatz kombiniert.

2.3 Hybrid Filtering

Da weder CBF noch CF in isolierter Form alle Anforderungen in der Nachrichtendomäne abdecken, gelten hybride Architekturen als praxiserprobter und in der Literatur breit beschriebener Entwurfsansatz. Das Ziel ist es, die komplementären Stärken der Ansätze zu kombinieren und deren typische Schwächen wie das Item-Cold-Start-Problem oder den Popularity Bias abzumildern (vgl. Burke 2002; Wu u. a. 2023; Raza und Ding 2022).

Die Fachliteratur unterscheidet eine Vielzahl von Hybridisierungsstrategien. Hierzu zählen unter anderem die gewichtete Kombination von Modell-Scores (weighted), eine fallweise Umschaltung zwischen Modellen (switching) sowie die Nutzung von Modellergebnissen als Eingangsmerkmale für nachfolgende Modelle (feature/meta-level) (vgl. Burke 2002). Die vorliegende Arbeit implementiert eine gewichtete Hybridisierung, bei der die Scores der CBF- und CF-Komponente zu einer finalen Empfehlungsliste zusammengeführt werden.

Dieser Ansatz grenzt sich von komplexeren, mehrstufigen Architekturen ab. Bei kaskadierenden Systemen (cascade) oder Re-Ranking-Verfahren wird beispielsweise zunächst eine breite Kandidatenliste durch ein schnelles Retrieval-Modell erzeugt und diese anschließend in einem zweiten Schritt durch ein präziseres, oft aufwendigeres Modell umsortiert. Solche Re-Ranking-Modelle können explizit auf weitere Kriterien wie Diversität optimieren, etwa durch die Anwendung von Algorithmen wie Maximal Marginal Relevance (MMR) (vgl. Carbonell und Goldstein 1998). Für diese Arbeit wurde bewusst eine einstufige, gewichtete Kombination als interpretierbare und robuste Baseline gewählt, deren Optimierung im Zentrum der Untersuchung steht.

2.4 Qualitätsdimensionen und Erfolgsmetriken

Um die Güte von Empfehlungssystem (ES) objektiv zu bewerten und verschiedene Ansätze miteinander vergleichen zu können, sind standardisierte Erfolgsmetriken erforderlich. Da ES typischerweise eine sortierte Liste von Items ausgeben, müssen diese Metriken nicht nur die Relevanz der Vorschläge, sondern auch deren Rangposition berücksichtigen.

Eine der etabliertesten Metriken für diese Anforderung ist der Discounted Cumulative Gain (DCG). Der Grundgedanke des DCG ist, die Relevanzwerte aller empfohlenen Items bis zu einer bestimmten Listenlänge K zu summieren. Um der Beobachtung Rechnung zu tragen, dass Nutzer den vorderen Plätzen mehr Beachtung schenken, wird der Relevanzwert (rel_i) eines Items an Position i durch einen logarithmischen Faktor abgewertet (diskontiert). In dieser Arbeit ist rel_i binär; 1 steht für einen angeklickten Artikel. Ein

relevantes Item an Position 1 trägt somit mehr zum Gesamtscore bei als dasselbe Item an Position 10.

$$DCG@K = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \tag{2.3}$$

Da der absolute DCG-Wert von der Anzahl der verfügbaren relevanten Items abhängt, ist er für einen fairen Vergleich zwischen verschiedenen Nutzern oder Anfragen ungeeignet. Aus diesem Grund wird der DCG normalisiert, indem er ins Verhältnis zum bestmöglichen DCG-Wert gesetzt wird. Dieser Idealwert, der Ideal Discounted Cumulative Gain (IDCG), repräsentiert den DCG einer perfekten, nach absteigender Relevanz sortierten Empfehlungsliste.

Das Ergebnis dieser Normalisierung ist der Normalized Discounted Cumulative Gain (nDCG). Sein Wert liegt immer zwischen 0.0 (kein Treffer) und 1.0 (perfekte Rangfolge) und ermöglicht so eine vergleichbare und aussagekräftige Bewertung der Performance eines ES (vgl. Järvelin und Kekäläinen 2002).

$$NDCG@K = \frac{DCG@K}{IDCG@K} \tag{2.4}$$

3 Systemarchitektur und Rahmenbedingungen

Die Konzeption eines produktiv einsetzbaren Empfehlungssystems für die SV-Gruppe erfordert eine Architektur, die klar definierten Rahmenbedingungen genügt. Aus dem Anwendungsfall lassen sich nicht-funktionale Anforderungen ableiten, die die technische Ausgestaltung leiten.

Für den initialen Proof-of-Concept wurden drei erfolgskritische Non Functional Requests (NFR)s identifiziert: Erstens muss das System niedrige Antwortzeiten unter Last gewährleisten. Als konkretes Service Level Objective (SLO) wird eine End-to-End-Latenz im 95. Perzentil von unter 2000 Millisekunden angestrebt, siehe Abbildung 3.1. Diese Anforderung orientiert sich an etablierten Erkenntnissen der Usability-Forschung, welche besagen, dass Antwortzeiten über einer Sekunde die Konzentration des Nutzers unterbrechen, während Latenzen von bis zu zehn Sekunden die Obergrenze darstellen, um die Aufmerksamkeit zu halten (vgl. Nielsen 1993). Zweitens ist eine horizontale Skalierbarkeit erforderlich, um auf wachsende Nutzerzahlen und Datenmengen im dynamischen Umfeld eines Nachrichtenportals reagieren zu können. Drittens ist eine einfache Integrierbarkeit sicherzustellen; hierzu wird eine standardisierte REST-API bereitgestellt, die die Einbindung in bestehende Redaktions- und IT-Workflows vereinfacht.

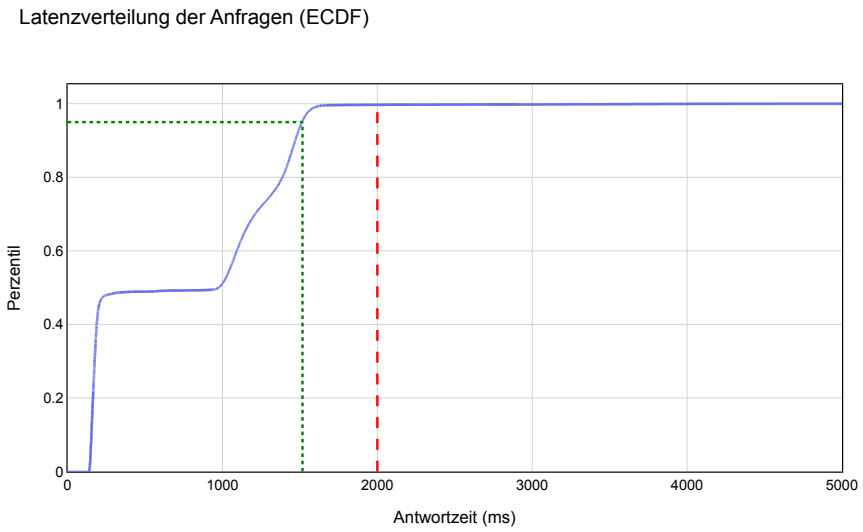

Abb. 3.1: ECDF-Plot der Latenzverteilung. Die rote, gestrichelte Linie markiert das SLO-Ziel von 2000 ms, während die grüne, gepunktete Linie das tatsächlich erreichte 95. Perzentil bei 1562.28 ms zeigt.

Obwohl es sich um einen Prototyp handelt, wurden weitere, für einen späteren Produktivbetrieb relevante Anforderungen berücksichtigt: Datenschutz und Sicherheit gemäß Datenschutz-Grundverordnung (DSGVO), hohe Verfügbarkeit und Ausfallsicherheit sowie Transparenz der Empfehlungslogik zur Stärkung des Nutzervertrauens. Diese Aspekte sind

im Entwurf konzeptionell verankert.

3.1 Rahmenbedingungen

Die bestehende, auf der GCP basierende IT-Infrastruktur der SV-Gruppe setzt die zentralen Rahmenbedingungen. Der Artikelkorpus sowie die aus Google Analytics 4 (GA4) stammenden Nutzerinteraktionsdaten liegen in BigQuery-Tabellen vor. Eine zentrale Ressource bilden hochdimensionale Artikel-Embeddings (3072 Dimensionen), die aus Titeln und Volltexten erzeugt wurden. Da sich diese Embeddings bereits in verwandten Anwendungsfällen innerhalb der SV-Gruppe (z.B. semantische Anzeigenausspielung) als qualitativ hochwertig erwiesen haben, werden sie in dieser Arbeit wiederverwendet. Dieser Ansatz reduziert den Entwicklungsaufwand und baut auf einer validierten Datenressource auf.

Die `user_pseudo_id` ist ein pseudonymer Geräte-/Browser-Identifier. Sie bleibt auf demselben Gerät/Browser meist über mehrere Sitzungen stabil, kann sich jedoch durch Cookie-/App-Resets oder Geräte-/Browserwechsel ändern. Eine typische `user_pseudo_id` hat beispielsweise das Format `18475638.1694782345`.

Für eingeloggte Nutzer wird zusätzlich eine eindeutige, geräteübergreifende `user_id` erfasst. Eine Analyse der Nutzerbasis im Untersuchungszeitraum (Januar bis März 2025) zeigt jedoch, dass mit 99,01 % der weitaus größte Teil der Nutzer anonym auf das Angebot zugreift und somit nur über eine `user_pseudo_id` verfügt.

Um ein Empfehlungssystem zu entwickeln, das der gesamten Leserschaft und nicht nur der kleinen Kohorte eingeloggter Nutzer dient, wurde die `user_pseudo_id` konsequent als primärer Schlüssel für die Personalisierung verwendet. Dieser Ansatz ist praxistauglich, da (i) die `user_pseudo_id` auf demselben Gerät über mehrere Sitzungen hinreichend stabil ist, um kurz- bis mittelfristige Präferenzen zu lernen, und (ii) im Nachrichtenkontext vor allem jüngste Interaktionen und der unmittelbare Artikelkontext prädiktiv sind. Letzteres wird durch die Hybridisierung genutzt: Der CBF-Anteil arbeitet kontextuell, während der NCF-Anteil gerätebezogene Wiederholungsmuster erfasst.

3.2 Systemarchitektur

Die technologische Architektur in Abbildung 3.2 folgt einem cloud-nativen Microservice-Ansatz auf der GCP. Diese Architektur wurde gewählt, da sie durch lose gekoppelte, unabhängige Dienste eine hohe Skalierbarkeit, Wartbarkeit und technologische Flexibilität ermöglicht (vgl. Newman 2015). Als zentraler Orchestrator dient ein in Python implementierter Service auf Basis von FastAPI. Der Service nimmt Anfragen entgegen, steuert die Modell-Endpunkte, aggregiert die Teilergebnisse und führt die Hybridisierung aus. Das Design ist modular, um künftig alternative Hybridisierungsstrategien mit geringem Integrationsaufwand aufnehmen zu können. Für die latenzkritische Online-Auslieferung

werden der Artikelkorpus und Embeddings in einer operativen Datenbank vorgehalten; die ML-Modelle sind auf Vertex AI[2] deployt.

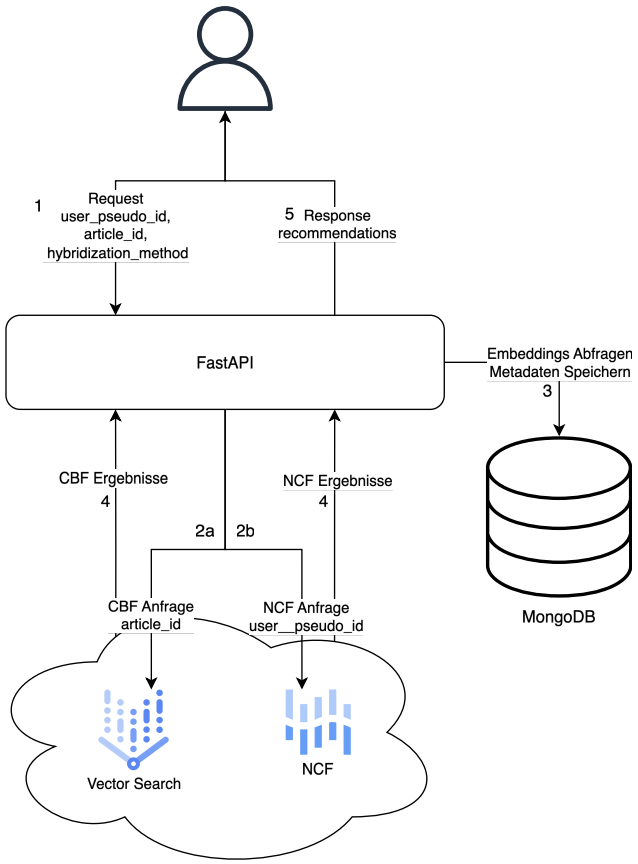

Abb. 3.2: Die technologische Architektur des hybriden Empfehlungssystems. Der nummerierte Datenfluss zeigt den Weg einer Anfrage vom Nutzer (1), über die parallelen Abfragen an die ML-Dienste (2a, 2b) und die Datenbank (3), die eintreffenden Ergebnisse (4) bis zur finalen Empfehlung (5).

Das Herzstück des Systems bildet der API-Service, der die externe Schnittstelle und die interne Orchestrierung bereitstellt. Die API exponiert einen REST-Endpunkt unter `/v1/recommendations` mit einem JSON-basierten Datenvertrag. Eine Anfrage umfasst die `user_pseudo_id`, die `article_id` als Kontext sowie die gewünschte Hybridisierungsstrategie. Die Implementierung nutzt die asynchrone Leistungsfähigkeit von FastAPI (Asynchronous Server Gateway Interface (ASGI)) und orchestriert pro Anfrage parallele Aufrufe der untergeordneten ML-Dienste und der Datenbank. Nach dem Zusammenführen der Teilergebnisse wird die Hybridisierungslogik angewandt und die finale Empfehlungsliste zurückgegeben. Eingebaute Schema-Validierung und die automatische Generierung einer interaktiven API-Dokumentation unterstützen eine robuste Integration.

Das CBF-Retrieval erfolgt über einen Vektorindex (Vertex AI Vector Search) auf den

[2]Siehe `https://cloud.google.com/vertex-ai`

3072-dimensionale Artikel-Embeddings. Abfragen nutzen L2-normalisierte Vektoren und MIPS; pro Kontextartikel werden die Top-K semantisch ähnlichsten Artikel mit niedriger Latenz geliefert. Die zugrunde liegenden ANN-Prinzipien und der Latenz–Recall-Trade-off sind in Abschnitt 2.1 erläutert.

Als kollaborative Komponente kommt ein NCF-Modell zum Einsatz, das implizites Feedback verarbeitet und nichtlineare Nutzer–Item-Interaktionen modelliert (methodische Details siehe Abschnitt 2.2). Das trainierte Modell wird als Echtzeit-Endpoint bereitgestellt und vom Orchestrator parallel zum CBF-Dienst abgefragt; die Ergebnisse werden mittels gewichteter Hybridisierung kombiniert. Effekte wie Popularity Bias werden in der Hybridisierung gezielt adressiert (vgl. 2.3).

3.3 Datenbasis und Einschränkungen

Der Datensatz *SM-News-Jan25* umfasst ausschließlich GA4-Interaktionsdaten des Nachrichtenportals schwaebische.de aus dem Januar 2025.

Um sicherzustellen, dass das Modell ausschließlich auf relevanten Nutzerinteraktionen trainiert wird, wurde der Rohdatensatz der `page_view`-Events aus GA4 einem entscheidenden Filterschritt unterzogen. Ein `page_view`-Event wird nicht nur für Artikel, sondern auch für Übersichts- oder Kategorieseiten ausgelöst. Daher wurden nur solche Events für die weitere Analyse berücksichtigt, deren zugehöriger Seitenpfad einem vordefinierten Muster für Artikelseiten entspricht. Der resultierende Datensatz repräsentiert somit ausschließlich explizite Artikelaufrufe.

Aus Kostengründen wird das NCF-Modell auf den ersten drei Wochen trainiert und in der vierten Woche getestet. Eine zentrale Eigenschaft des Trainingsdatensatzes ist die in Abbildung 3.3 gezeigte stark rechtsschiefe Verteilung der Artikelpopularität als auch der in Abbildung 3.4 gezeigten Nutzeraktivität (Long-Tail-Verteilung), ein für Mediendaten typisches Muster und eine Kernherausforderung für Empfehlungssysteme (vgl. Wu u. a. 2023; Raza und Ding 2022).

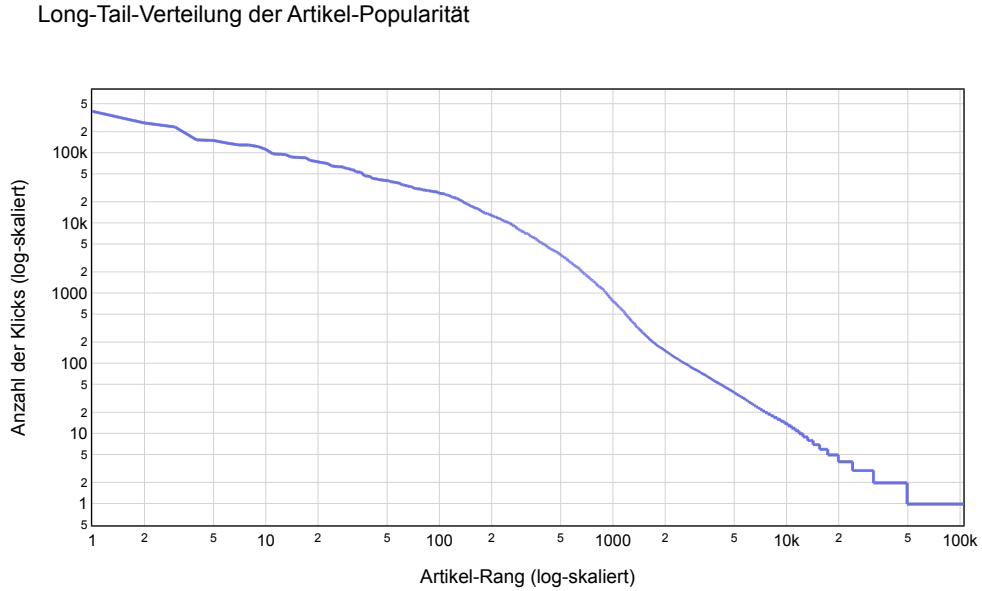

Abb. 3.3: Popularitätsverteilung der Artikel im Trainingsdatensatz. Die Grafik zeigt, dass eine geringe Anzahl von Artikeln einen Großteil der Klicks auf sich vereint, während die Mehrheit der Artikel nur wenige Interaktionen erhält (Long-Tail).

Die Long-Tail-Struktur manifestiert sich in zwei Dimensionen:

- **Artikelpopularität:** Wie in Abbildung 3.3 ersichtlich, konzentriert sich ein überproportional großer Anteil der Seitenaufrufe auf wenige virale „Hit"-Artikel (Kopf der Verteilung). Die kurze Lebensdauer von Nachrichtenartikeln verstärkt diesen Effekt zusätzlich.
- **Nutzeraktivität:** Analog zeigt sich beim Nutzerverhalten eine kleine Kohorte hochaktiver „Power-Nutzer", die einen Großteil der Artikelaufrufe generiert, während die Mehrheit nur sporadisch interagiert.

Diese ungleiche Verteilung induziert einen inhärenten Popularity Bias: Naive Modelle empfehlen tendenziell wiederholt dieselben Bestseller-Artikel, was Personalisierung untergräbt und das Risiko einer Filterblase erhöht. Gleichzeitig entsteht ein Kaltstart-Problem für Nischeninhalte im Long Tail. Eine zentrale Zielsetzung dieser Arbeit ist folglich die Konzeption eines Systems, das den Popularity Bias aktiv ausbalanciert und relevante Nischeninhalte an passende Nutzer ausspielt (vgl. Abdollahpouri 2019).

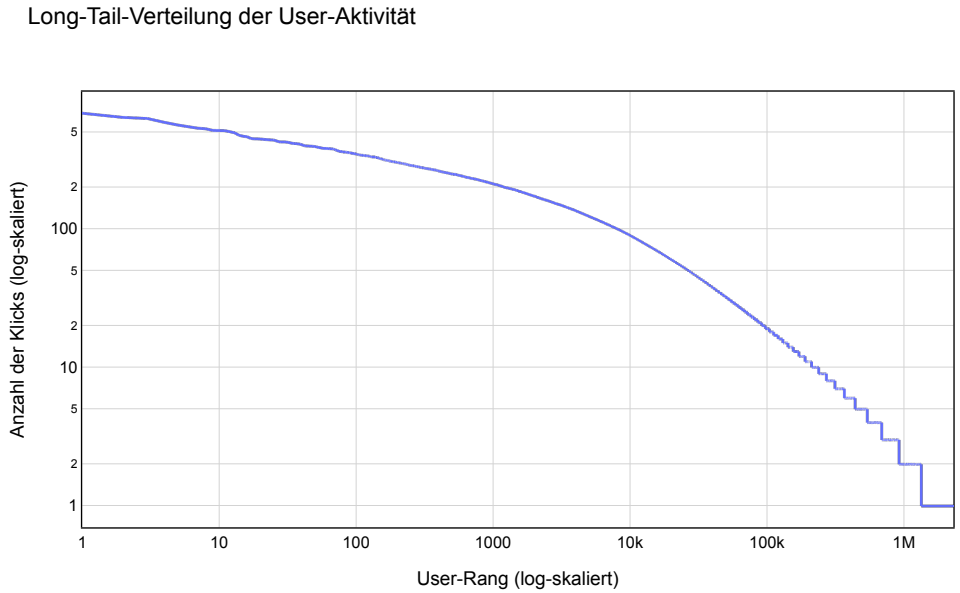

Abb. 3.4: Verteilung der Nutzeraktivität im Trainingsdatensatz. Die Darstellung verdeutlicht die typische Long-Tail-Verteilung: Eine große Anzahl von Nutzern interagiert nur selten mit Artikeln, während eine kleine Gruppe von "Power-Nutzern"für einen Großteil der Klicks verantwortlich ist.

Die statistischen Kennzahlen des Trainingsdatensatzes sind in Tabelle 3.1 zusammengefasst.

Tab. 3.1: Statistische Kennzahlen des Trainingsdatensatzes, basierend auf den Klick-Logs der ersten drei Januarwochen.

Metrik	**Wert**
Gesamte Artikelaufrufe	11.412.116
Einzigartige user_pseudo_ids	2.294.733
Einzigartige Artikel	104.462

Zur Nutzung der PageView-Events im NCF-Modell werden diese in eine User–Item-Interaktionsmatrix überführt. Zeilen repräsentieren eindeutige Nutzer, Spalten eindeutige Artikel des Trainingszeitraums. Eine Zelle (i, j) erhält den Wert 1, wenn Nutzer i mit Artikel j interagiert hat, andernfalls 0. Die resultierende binäre Matrix mit circa 2,3 Mio. Nutzern und 104,000 Artikeln weist eine Dichte von unter 0,005 % auf und ist damit extrem dünn besetzt. Diese Sparsity ist für CF-Modelle herausfordernd, da je Nutzer nur wenige Interaktionen im Verhältnis zur Gesamtheit der Artikel vorliegen und Generalisierung auf neue User–Item-Paare erschwert wird.

Bei der Interpretation der Ergebnisse sind folgende Limitationen zu berücksichtigen:

- **Zeitlicher Rahmen:** Der Datensatz deckt ausschließlich den Januar 2025 ab und bildet damit eine Momentaufnahme. Saisonale Effekte (z. B. Feiertage) oder längerfristige Trends können nicht erfasst werden.
- **Implizites Feedback:** Als Signal dienen ausschließlich Artikelaufrufe. Dieses implizite Feedback ist zwar reichhaltig, jedoch mehrdeutig: Ein Artikelaufruf ist kein direkter Indikator für Zufriedenheit (z. B. bei sofortigem Absprung). Weitere Signale wie Verweildauer werden im Prototyp nicht berücksichtigt.
- **Offline-Evaluation:** Die Modellgüte wird offline auf historischen Daten anhand etablierter Metriken (z. B. nDCG) evaluiert. Die tatsächliche Wirkung auf Nutzerverhalten im Live-Betrieb lässt sich damit nur approximieren. Ein Online-A/B-Test wäre der nächste Schritt, um den Business Impact auf KPIs wie Sitzungsdauer oder Nutzerbindung zu messen.

4 Optimierung der Empfehlungsqualität

4.1 Zielfunktion

Als Zielfunktion für die Optimierung des hybriden Empfehlungssystems dient eine gewichtete Linearkombination der Scores der beiden zugrundeliegenden Modelle. Dieser Ansatz wurde aufgrund seiner einfachen Interpretierbarkeit und seiner weiten Verbreitung als robuste Baseline für hybride Architekturen gewählt (vgl. Burke 2002). Die mathematische Formulierung der Funktion lautet:

$$s_{\text{hybrid}} = w_{\text{cbf}} \cdot s_{\text{cbf}} + w_{\text{cf}} \cdot s_{\text{cf}} \tag{4.1}$$

Die einzelnen Terme der Gleichung sind wie folgt definiert:

- s_{hybrid}: Der finale, kombinierte Score für ein potenziell zu empfehlendes Item. Die finale Empfehlungsliste wird durch das absteigende Sortieren der Items nach diesem Score erstellt.
- s_{cbf} und s_{cf}: Die von der Content-Based- bzw. der Collaborative-Filtering-Komponente generierten Roh-Scores. Vor der Kombination in der Zielfunktion ist eine Normalisierung dieser Scores erforderlich, um sicherzustellen, dass sie auf einer vergleichbaren Skala liegen. In dieser Arbeit werden die Scores mit der Min-Max-Skalierung auf $[0, 1]$ normalisiert.
- w_{cbf} und w_{cf}: Die nicht-negativen Gewichtungsparameter, welche den Einfluss der jeweiligen Modellkomponente auf das Endergebnis steuern. Für diese Gewichte gilt die Nebenbedingung $w_{\text{cbf}} + w_{\text{cf}} = 1$, wodurch die Optimierung auf die Bestimmung eines einzigen Parameters reduziert wird.

Das zentrale Ziel des in 4.4 beschriebenen Optimierungsprozesses ist es, die optimalen Werte für die Gewichtungsparameter w_{cbf} und w_{cf} zu finden, sodass die in Abschnitt 4.2 definierte Evaluationsmetrik (nDCG@10) maximiert wird.

4.2 Evaluationsmetriken

Die empirische Bewertung und Optimierung des hybriden Empfehlungssystems erfordert ein klar definiertes Evaluationsprotokoll sowie eine geeignete Zielfunktion. Als primäre Zielfunktion für die in Kapitel 4 beschriebene Hyperparameter-Optimierung dient die Maximierung der Empfehlungsqualität, gemessen durch den normalisierten diskontierten kumulativen Gewinn bei einer Listenlänge von 10 (nDCG@10). Diese Metrik wurde ausgewählt, da sie nicht nur die reine Trefferquote erfasst, sondern vor allem die exakte Position eines relevanten Treffers innerhalb der Empfehlungsliste gewichtet, was das reale Nutzerverhalten präzise abbildet.

Die besondere Eignung des nDCG als Metrik für Empfehlungssysteme liegt in seiner positions-sensitiven ("top-heavy") Natur. Mathematisch wird dies durch eine logarithmische Diskontierung erreicht, bei der der Wert eines Treffers am Rang r mit dem Faktor $1/\log_2(r+1)$ abgewertet wird. Dies führt dazu, dass ein Treffer an den vordersten Rängen überproportional mehr zur Gesamtbewertung beiträgt als ein Treffer an einer späteren Position. So ist beispielsweise ein relevanter Artikel an Rang 1 signifikant wertvoller als an Rang 2, während der Unterschied zwischen Rang 100 und 101 nur noch marginal ist. Diese Eigenschaft ist essenziell, da Nutzer Interaktionen auf den vordersten Plätzen der Ergebnisliste konzentrieren und weiter hinten platzierte Vorschläge selten beachten (vgl. Krichene und Rendle 2020).

Ein entscheidender Aspekt des Evaluationsdesigns ist die Handhabung negativer Instanzen. Anstatt auf Verfahren des Negative Samplings zurückzugreifen, bei denen eine kleine, zufällige Teilmenge nicht-interagierter Artikel als negative Beispiele dient, wird in dieser Arbeit eine methodisch rigorosere Strategie des *Full-Catalog Rankings* verfolgt. Bei diesem Vorgehen muss das Modell für jeden positiven Testfall den relevanten Artikel aus der Gesamtheit aller im Datensatz verfügbaren Artikel – abzüglich der bereits vom Nutzer gelesenen – identifizieren.

Diese Methode vermeidet systematische Verzerrungen (sampling bias), die durch ein unausgewogenes Sampling entstehen können, und simuliert ein anspruchsvolles, aber realistisches Anwendungsszenario. Die Evaluation auf dem gesamten Artikelkatalog stellt eine enorme Herausforderung für das Empfehlungssystem dar, was naturgemäß zu absolut gesehen niedrigen Metrikwerten führt. Wie in der Fachliteratur bestätigt wird, sind solche Ergebnisse jedoch nicht als Indikator für eine geringe Modellleistung zu interpretieren, sondern als Konsequenz eines besonders anspruchsvollen und unverzerrten Evaluationsprotokolls (vgl. Krichene und Rendle 2020). Die auf diese Weise gewonnenen Erkenntnisse bieten eine robuste und verlässliche Grundlage für die zukünftige Weiterentwicklung von Empfehlungssystemen bei der SV-Gruppe.

4.3 Experimenteller Aufbau und Validierungsdatensatz

Die Grundlage für die Optimierung und Evaluation bildet ein Validierungsdatensatz, der aus den Nutzerinteraktionen der letzten Januarwoche 2025 extrahiert wurde. Die statistischen Kennzahlen dieses Zeitraums sind in Tabelle 4.1 zusammengefasst.

Aus dem Gesamt-Pool an Nutzern wurde eine repräsentative Validierungsstichprobe von 1.000 einzigartigen Nutzern zufällig gezogen, um den für die Optimierung erforderlichen Rechenaufwand in einem praktikablen Rahmen zu halten. Für jeden Nutzer in der Stichprobe wurde nach dem *Leave-Last-Out*-Prinzip der letzte interagierte Artikel als Ground Truth für die Evaluation definiert (vgl. Rendle u. a. 2009).

Tab. 4.1: Statistische Kennzahlen des Test- und Validierungsdatensatzes, basierend auf den Klick-Logs der letzten Januarwoche.

Metrik	**Wert**
Gesamte Artikelaufrufe	3.627.024
Einzigartige user_pseudo_ids	1.099.289
Einzigartige Artikel	57.100

4.4 Optimierungsstrategie

Die Bestimmung der optimalen Gewichtungsparameter w_{cbf} und w_{cf} aus Gleichung 4.1 erfolgt durch einen automatisierten Hyperparameter-Optimierungsprozess.

Als Framework für diesen Prozess wurde *Optuna* (Akiba u. a. 2019) gewählt, das eine effiziente Suche im Parameterraum ermöglicht. Über insgesamt 515 Iterationen (Trials) wurden verschiedene Gewichtungskombinationen evaluiert, mit dem Ziel, die in Abschnitt 4.2 definierte nDCG@10-Metrik zu maximieren.

Der Ablauf eines einzelnen Optimierungs-Trials gestaltet sich wie folgt:

1. Optuna schlägt eine neue Wertekombination für w_{cbf} und w_{cf} vor.
2. Für jeden der 1.000 Nutzer in der Validierungsstichprobe werden die Empfehlungen der CBF- und CF-Modelle über deren jeweilige API-Endpunkte parallel und asynchron abgerufen.
3. Die Ergebnislisten der beiden Modelle werden mittels der vorgeschlagenen Gewichte zu einer finalen hybriden Empfehlungsliste fusioniert.
4. Der nDCG@10-Wert dieser finalen Liste wird berechnet, indem sie mit dem Ground-Truth-Artikel des Nutzers verglichen wird.
5. Der über alle 1.000 Nutzer gemittelte nDCG@10-Wert wird an Optuna zurückgegeben, um den nächsten, informierteren Suchschritt zu steuern.

Dieser Prozess entspricht einem Gesamtaufwand von 515.000 individuellen Evaluierungen des Hybridmodells, was eine gründliche und robuste Suche nach der optimalen Parameterkonfiguration gewährleistet.

Die Visualisierung der Optimierungslandschaft in Abbildung 4.1 verdeutlicht die komplexe, nicht-lineare Beziehung zwischen den Modellgewichten und der resultierenden Empfehlungsqualität. Angesichts dieser zerklüfteten Landschaft mit mehreren lokalen Optima ist ein naiver Ansatz wie eine Rastersuche (Grid Search) ineffizient. Der Einsatz eines fortschrittlichen Optimierungsframeworks wie Optuna, das mit intelligenten Suchalgorithmen arbeitet, ist daher notwendig, um den global optimalen Bereich im Parameterraum effizient und verlässlich zu identifizieren.

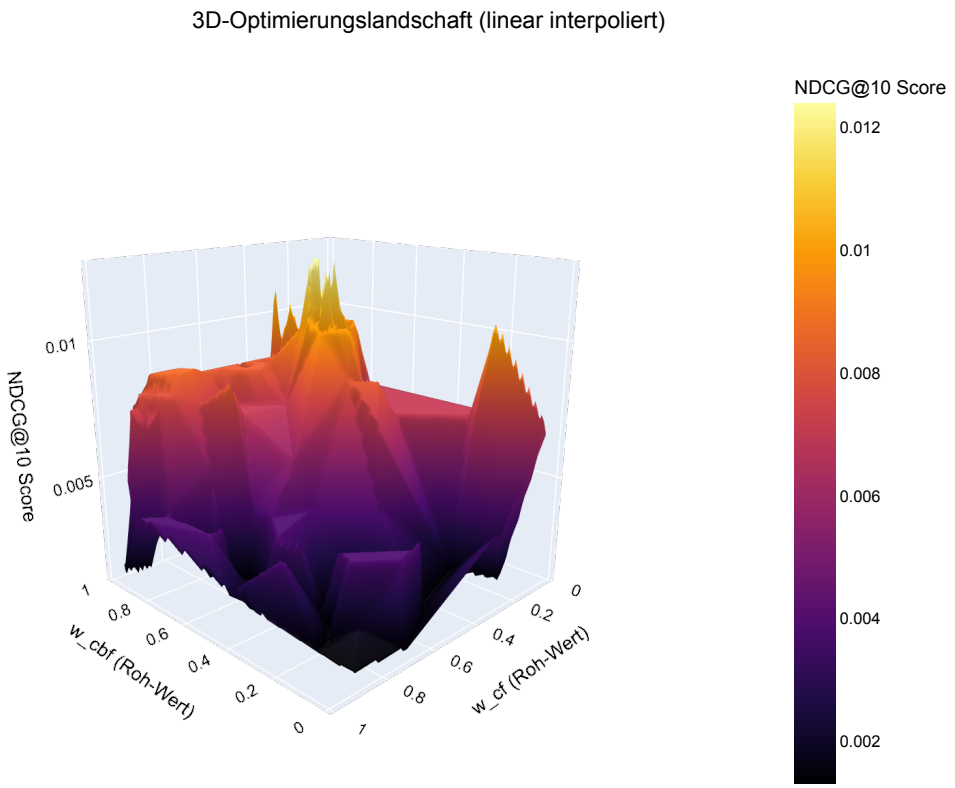

Abb. 4.1: Visualisierung des zweidimensionalen Hyperparameterraums der Modellgewichtungen. Die Achsen repräsentieren die Gewichte für das CBF-Modell (w_{cbf}) und das CF-Modell (w_{cf}). Die Einfärbung der Punkte visualisiert den resultierenden NDCG@10-Wert für jede Konfiguration aus dem Optuna-Suchlauf.

Konkret wurde für die Suche der in Optuna standardmäßig implementierte Tree-structured Parzen Estimator (TPE) als Suchalgorithmus verwendet (vgl. Akiba u. a. 2019). Der TPE-Algorithmus ist eine Variante der Bayes'schen Optimierung, die ein probabilistisches Modell der Zielfunktion erstellt, um intelligent zu entscheiden, welche Parameterkombination als Nächstes getestet werden soll. Dafür werden die bisherigen Evaluationsergebnisse in eine Gruppe mit vielversprechenden und eine mit weniger erfolgreichen Ergebnissen aufgeteilt.

Für beide Gruppen werden separate Wahrscheinlichkeitsverteilungen modelliert. Der Algorithmus maximiert anschließend die erwartete Verbesserung Expected Improvement, indem er gezielt nach Parametern sucht, die eine hohe Wahrscheinlichkeit unter der Verteilung der guten Ergebnisse und gleichzeitig eine niedrige Wahrscheinlichkeit unter der Verteilung der schlechten Ergebnisse aufweisen. Diese Strategie ermöglicht eine effiziente Balance zwischen der Exploitation bereits bekannter guter Regionen und der Exploration neuer, potenziell noch besserer Bereiche des Suchraums. Durch diesen informierten Suchprozess kann das Optimum in komplexen Landschaften, wie der in Abbildung 4.1 gezeigten, deutlich schneller und ressourcenschonender gefunden werden als durch uninformierte Methoden.

4.5 Ergebnisse

Der in Abschnitt 4.4 beschriebene Optimierungsprozess führte zur Identifikation einer robusten Gewichtungskonfiguration für das hybride Empfehlungssystem. Die finale Evaluation zeigt, dass das optimierte Hybrid-Modell die definierten Baseline-Modelle in allen Metriken signifikant übertrifft.

Tabelle 4.2 listet die detaillierten Performance-Werte der evaluierten Modellvarianten auf, während Abbildung 4.2 eine visuelle Gegenüberstellung der Ergebnisse bietet.

Tab. 4.2: Vergleich des optimierten Hybrid-Modells mit den Baseline-Modellen anhand der Metriken NDCG@10 und Hit Rate@10.

Modell	NDCG@10	Hit Rate@10
Hybrid-Modell	1.25%	1.8%
Popularity-Baseline	0.23%	0.6%
Recency-Baseline	0.14%	0.4%

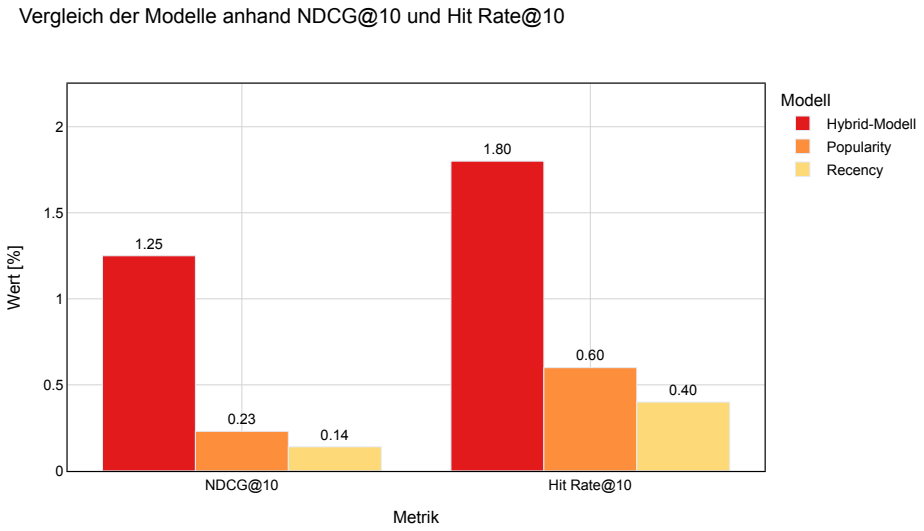

Abb. 4.2: Vergleich des Hybrid-Modells mit den Popularity- und Recency-Baselines anhand der Metriken NDCG@10 und Hit Rate@10.

Abbildung 4.3 zeigt die einzelnen Trials (Schwarze Punkte) auf der Hyperparameterlandschaft. Es ist gut zu erkennen, dass die größte Anzahl an optima in einem Bereich liegt, in dem w_{cbf} größer als 0.5 ist und w_{cf} kleiner als 0.5.

Diese Ergebnisse bestätigen nicht nur die Wirksamkeit des gwählten Hybridisierungsansatzes, sondern liefern auch den empirischen Beleg für dessen Überlegenheit gegenüber den reinen Einzelkomponenten. Der Optimierungsprozess durchsuchte den gesamten Gewichtungsraum, einschließlich der Extremfälle, die einem reinen CF-Modell ($w_{cf} = 1$) oder einem reinen CBF-Modell ($w_{cb} = 1$) entsprechen.

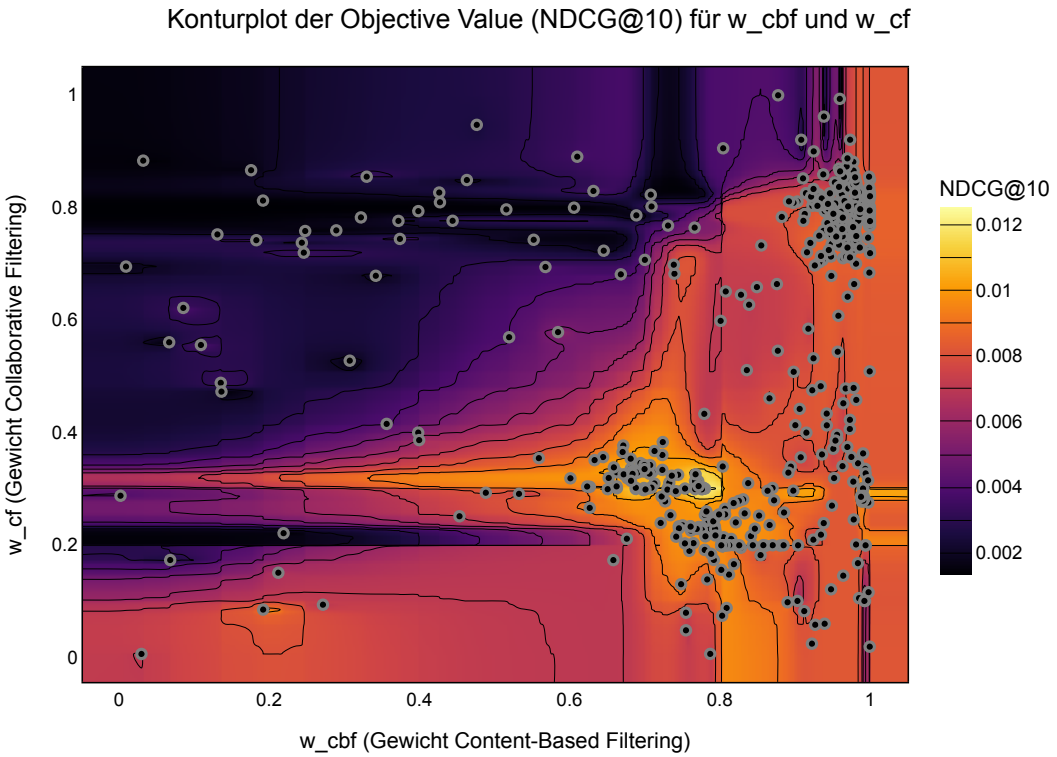

Abb. 4.3: Konturplot zur Darstellung der NDCG@10-Verteilung im zweidimensionalen Parameterraum der Modellgewichtungen. Die Isolinien verbinden Bereiche mit ähnlicher Performance.

Das identifizierte Optimum liegt jedoch klar bei einer Konfiguration von $w_{cbf} \approx 0.7226$ und $w_{cf} \approx 0.2774$. Diese Tatsache belegt, dass beide Komponenten einen positiven Beitrag zur Maximierung des nDCG@10 leisten und die gewichtete Kombination eine signifikant höhere Empfehlungsqualität erzielt, als es durch den alleinigen Einsatz des CBF- oder CF-Modells möglich gewesen wäre. Die konzeptionelle Notwendigkeit des hybriden Ansatzes zur Ausbalancierung von Relevanz und Diversität wird somit durch die datengetriebene Optimierung quantitativ untermauert.

Die dominante Rolle der CBF-Komponente wird in Abbildung 4.4 nochmals verdeutlicht. Die Visualisierung zeigt die nDCG@10 Performance aller 515 Trialsin Abhängigkeit vom normalisierten CBF-Gewicht.

Es ist ein deutlicher Leistungssprung (Phasensprung) bei einem Gewicht von $w_{cbf} \approx 0.5$ zu erkennen. Unterhalb dieser Schwelle werden durchweg nur niedrige nDCG@10-Werte erreicht. Oberhalb davon steigt die Performance signifikant an, und alle optimalen Ergebnisse (orange markiert) konzentrieren sich im Bereich $w_{cbf} > 0.6$

Diese Beobachtung untermauert die Hypothese, dass im Nachrichtenkontext der unmittelbare thematische Bezug zum gerade gelesenen Artikel der stärkste Prädiktor für die nächste Interaktion ist. Das CBF-Modell liefert hier das notwendige Fundament für eine relevante Empfehlung. Die CF-Komponente dient in diesem Zusammenspiel als entscheidender „Fein-Tuner", der auf dieser starken Basis aufsetzt, um die Performance durch Per-

sonalisierung auf das globale Maximum zu heben, wie es im Konturplot (Abbildung 4.3) ersichtlich wurde.

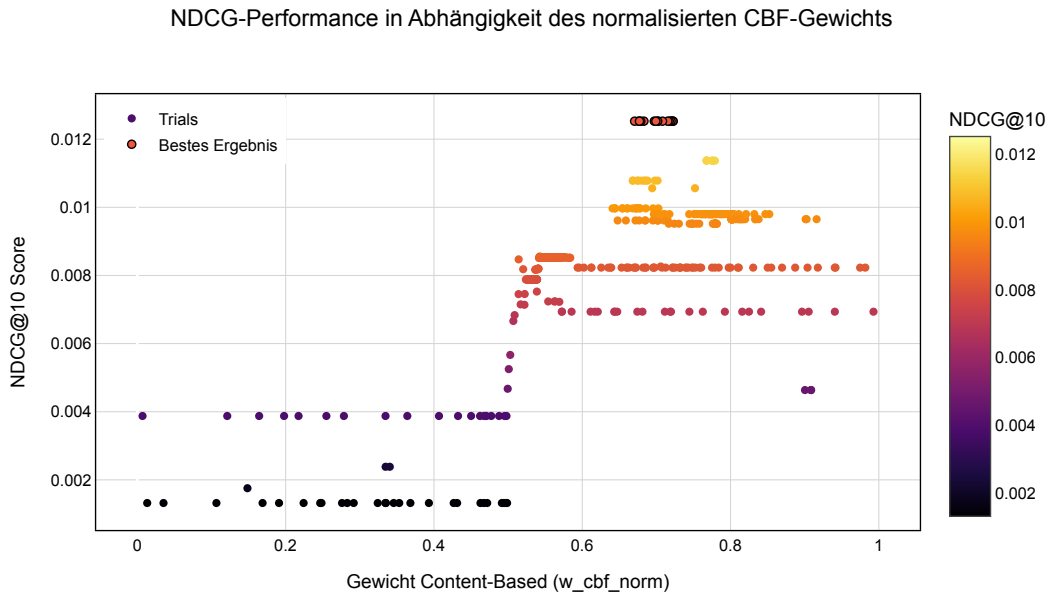

Abb. 4.4: 2D-Darstellung des Hyperparameterraums. Die Achsen zeigen die Gewichtungen für das CBF- (w_{cbf}) und CF-Modell (w_{cf}). Die Farbe der Punkte indiziert den erreichten NDCG@10-Score. Der optimale Punkt ist markiert.

5 Diskussion

In diesem Kapitel werden die im Rahmen der Arbeit erzielten Ergebnisse kritisch reflektiert, die Limitationen der Arbeit erörtert und die praktische Relevanz der Erkenntnisse für die SV-Gruppe beleuchtet. Abschließend wird die Übertragbarkeit des entwickelten Ansatzes auf andere Kontexte diskutiert.

5.1 Reflexion der Ergebnisse

Die in Kapitel 4 präsentierten Ergebnisse bestätigen nicht nur die grundsätzliche Funktionsfähigkeit des hybriden Ansatzes, sondern liefern auch tiefere Einblicke in die Dynamik des Nutzerverhaltens im Nachrichtenumfeld. Die datengetriebene Optimierung der Modellgewichtungen ermöglicht eine differenzierte Interpretation der Systemkomponenten.

Eine der zentralen Erkenntnisse ist die deutliche Dominanz der CBF-Komponente, deren optimales Gewicht bei über 0.7 liegt. Dies lässt sich auf die spezifischen Charakteristika des Nachrichtenkonsums zurückführen: Die unmittelbar nächste Interaktion eines Nutzers wird sehr stark vom thematischen Kontext des aktuell gelesenen Artikels beeinflusst. Das CBF-Modell bildet diese kontextuelle Relevanz präzise ab und liefert somit das notwendige Fundament für eine qualitativ hochwertige Empfehlung.

Gleichzeitig belegt das identifizierte Optimum, dass ein rein kontextuelles Modell nicht ausreicht. Der Beitrag des CF-Modells, wenngleich geringer gewichtet, ist entscheidend, um die Empfehlungsqualität auf das globale Maximum zu heben. Es fungiert als personalisierender "Fein-Tuner", der auf der starken CBF-Basis aufsetzt und Muster aus dem kollektiven Nutzerverhalten einbringt. Dadurch wird die Gefahr einer Überspezialisierung gemindert und dem Nutzer der Ausbruch aus einem engen thematischen Korridor ermöglicht.

Die Analyse der Optimierungslandschaft deutet zudem auf eine erfreuliche Robustheit des Systems hin. Das relativ breite Plateau um das Optimum signalisiert, dass das hybride ES nicht übermäßig empfindlich auf geringfügige Änderungen der Gewichtungsparameter reagiert, was für einen stabilen Betrieb in einer produktiven Umgebung von Vorteil ist.

Bei der Interpretation der absoluten nDCG@10-Werte muss das anspruchsvolle Evaluationsprotokoll des Full-Catalog Rankings berücksichtigt werden. Die erzielten Werte sind als Konsequenz dieser rigorosen und unverzerrten Methodik zu verstehen. Viel entscheidender als der absolute Wert ist daher die signifikante relative Steigerung gegenüber den Baseline-Modellen, welche die Wirksamkeit des optimierten hybriden ES klar belegt.

5.2 Limitationen

Trotz der vielversprechenden Ergebnisse unterliegt die vorliegende Arbeit bestimmten Limitationen, die für eine ausgewogene Einordnung essenziell sind und Ansatzpunkte für zukünftige Forschung bieten.

Eine wesentliche Einschränkung stellt die Datenbasis dar. Die Beschränkung auf GA4-Interaktionsdaten aus einem einzigen Monat verhindert die Modellierung saisonaler Effekte oder langfristiger Interessensentwicklungen. Ein Modell, das auf Daten aus dem Januar trainiert wurde, könnte im Sommer eine andere Performance aufweisen.

Darüber hinaus ist zu beachten, dass zum Zeitpunkt der Implementierung nur ein Teil des Artikelkorpus – rund 70.000 Artikel – über vorab berechnete und im Vektorindex gespeicherte Text-Embeddings verfügt. Um dennoch eine vollständige Abdeckung zu gewährleisten, wurde ein Ad-hoc-Mechanismus implementiert: Für einen Artikel ohne vorhandenes Embedding wird dieses bei der ersten Anfrage in Echtzeit über eine externe Schnittstelle generiert.

Dieser Prozess stellt zwar die grundsätzliche Funktionsfähigkeit des CBF-Ansatzes für alle Artikel sicher, führt jedoch zu einer potenziellen Inkonsistenz in der System-Performance. Der externe API-Aufruf und die anschließende Berechnung des Embeddings verursachen eine signifikant höhere Latenz für die betroffenen Anfragen. Diese kann das in Abschnitt 3.1 definierte SLO von 2000 Millisekunden verletzen und führt zudem zu variablen operationalen Kosten. Die Limitation liegt somit weniger in einer eingeschränkten Anwendbarkeit des ES, sondern vielmehr in der Gewährleistung einer durchgehend niedrigen Antwortzeit für den gesamten Artikelkorpus.

Die wohl wichtigste Limitation dieser Arbeit liegt in der Natur der Offline-Evaluation. Metriken wie nDCG@10 sind bewährte Indikatoren, können jedoch das tatsächliche Nutzerverhalten in einer Live-Umgebung nur approximieren. Um den kausalen Einfluss auf zentrale Geschäftsmetriken wie die Sitzungsdauer oder die Nutzerbindung zu messen, wäre ein Online-Experiment in Form eines A/B-Tests der unumgängliche nächste Schritt.

Zudem basiert die Modellierung ausschließlich auf Klick-Interaktionen. Dieses implizite Feedback ist zwar reichhaltig, aber auch ambivalent, da ein Klick nicht zwangsläufig Zufriedenheit oder tatsächliches Lesen signalisiert. Die Integration weiterer Signale wie der Verweildauer oder der Scrolltiefe könnte zu einer noch präziseren Abbildung der Nutzerpräferenzen führen.

5.3 Relevanz für die SV-Gruppe

Über den wissenschaftlichen Beitrag hinaus generiert diese Arbeit direkten und umsetzbaren Wert für die SV-Gruppe. Der entwickelte Prototyp ist nicht nur ein Proof-of-Concept, sondern dient als robuste und datenvalidierte Grundlage für die Produktivsetzung eines personalisierten Empfehlungsdienstes.

Aus geschäftlicher Sicht ist die nachgewiesene Steigerung der Empfehlungsrelevanz von großer strategischer Bedeutung. Es ist anzunehmen, dass sich die Verbesserung des nDCG@10 direkt in einer Erhöhung der Metrik „Artikel pro Session" niederschlägt. Dies würde nicht nur die Nutzerbindung stärken, sondern auch die Monetarisierungsmöglichkeiten

durch eine höhere Anzahl an Werbeeinblendungen verbessern.

Darüber hinaus adressiert das ES die Herausforderung der Content-Entdeckung. Indem es relevante Nischen-Inhalte aus dem "Long Tail"des Artikelarchivs an die passende Leserschaft ausspielt, erhöht es den Wert des gesamten Content-Portfolios und sorgt dafür, dass auch ältere, aber weiterhin relevante Beiträge sichtbar bleiben.

Neben der direkten Ausspielung an die Nutzer eröffnet die zugrundeliegende Technologie neue Möglichkeiten zur Effizienzsteigerung interner Redaktionsprozesse. Beispielsweise könnte die CBF-Komponente Redakteuren automatisiert thematisch passende Artikel für interne Verlinkungen vorschlagen und so den manuellen Rechercheaufwand reduzieren.

5.4 Übertragbarkeit

Die Erkenntnisse dieser Arbeit sind nicht auf den spezifischen Kontext der SV-Gruppe beschränkt, sondern lassen sich auf breitere Anwendungsfelder übertragen. Die identifizierten Prinzipien und methodischen Vorgehensweisen besitzen generischen Charakter.

Insbesondere für andere digitale Nachrichtenverlage sind die Resultate von hoher Relevanz, da die grundlegenden Datenstrukturen und Nutzerverhaltensmuster, wie die Dominanz des Lesekontexts, sehr ähnlich sind. Der hier entwickelte Ansatz kann als Vorlage für vergleichbare Medienhäuser dienen.

Der architektonische Grundgedanke, ein schnelles kontextuelles CBF-Modell mit einem personalisierten CF-Modell zu hybridisieren, ist auch in anderen Domänen ein bewährtes Muster. Im E-Commerce beispielsweise entspricht dies der Kombination von produktbasierter Ähnlichkeit für Neukunden mit personalisierten Empfehlungen für Bestandskunden.

Vor allem der hier demonstrierte methodische Prozess – von der Auswahl eines rigorosen Evaluationsprotokolls wie dem Full-Catalog Ranking bis zur systematischen Optimierung der Modellgewichte mittels Bayes'scher Verfahren – stellt eine generische und wiederverwendbare Blaupause für die Entwicklung und Validierung datengetriebener hybrider ES dar.

6 Fazit und Ausblick

6.1 Zusammenfassung zentraler Erkenntnisse

Die vorliegende Arbeit demonstrierte erfolgreich die Konzeption, Implementierung und Optimierung eines hybriden Empfehlungssystems zur Steigerung der Nutzerbindung für die SV-Gruppe. Es wurde der empirische Nachweis erbracht, dass ein datengetriebener Ansatz zur Gewichtung der Systemkomponenten zu einer signifikanten Verbesserung der Empfehlungsqualität führt.

Die zentrale Erkenntnis ist, dass eine hybride Architektur, die ein kontextuelles CBF-Modell mit einem personalisierten CF-Modell kombiniert, den reinen Baseline-Strategien signifikant überlegen ist. Die systematische Optimierung mittels Optuna identifizierte eine robuste Konfiguration, in der die thematische Relevanz des unmittelbaren Kontexts dominiert, aber durch kollaborative Signale entscheidend angereichert wird. Die Maximierung der nDCG@10-Metrik dient hierbei als validierter Indikator für eine verbesserte Nutzererfahrung.

Darüber hinaus wurde gezeigt, dass der Prototyp auf Basis der bestehenden GCP-Infrastruktur die definierten Anforderungen an Skalierbarkeit und Latenz erfüllt. Die Arbeit liefert somit nicht nur ein theoretisch fundiertes Modell, sondern auch eine praktisch umsetzbare und leistungsfähige technische Blaupause für den produktiven Einsatz.

6.2 Zukünftige Optimierungsmöglichkeiten

Aufbauend auf den gewonnenen Erkenntnissen ergeben sich mehrere vielversprechende Richtungen für die Weiterentwicklung des Systems.

Ein zentraler nächster Schritt ist die Durchführung von Online-A/B-Tests, um die in der Offline-Evaluation gemessenen Qualitätsgewinne unter realen Bedingungen zu validieren. Hierbei würde das entwickelte ES gegen das bestehende System ausgespielt, um den kausalen Einfluss auf zentrale Geschäftsmetriken wie die „Artikel pro Session", die Klickrate und die Verweildauer zu quantifizieren.

Darüber hinaus könnte die aktuell genutzte, einfache gewichtete Hybridisierung durch mehrstufige Architekturen ersetzt werden. Ein vielversprechender Ansatz wäre ein regelbasiertes Re-Ranking der Top-Empfehlungen. In diesem zweiten Schritt könnten die Kandidaten basierend auf weiteren Kriterien umsortiert werden, beispielsweise um Artikel hinter einer Paywall für abonnierte Nutzer zu priorisieren, eine höhere thematische Diversität sicherzustellen oder die Aktualität der Nachrichten zu berücksichtigen.

Ein weiterer vielversprechender Anwendungsfall liegt in der Entwicklung personalisierter Redaktionsassistenten. Die dem CBF-Modell zugrundeliegende Technologie zur semantischen Ähnlichkeit könnte Redakteuren in Echtzeit passende interne Verlinkungen vorschlagen, beim Tagging von Artikeln unterstützen oder aufzeigen, welche Inhalte aus

dem Archiv zu einem aktuellen Thema relevant sind. Dies würde nicht nur die Effizienz der Redaktionsprozesse steigern, sondern auch die Qualität und Vernetzung der Inhalte verbessern.

Literatur

Abdollahpouri, Himan (Jan. 2019). „Popularity Bias in Ranking and Recommendation“. In: DOI: 10.1145/3306618.3314309.

Akiba, Takuya u. a. (2019). „Optuna: A Next-generation Hyperparameter Optimization Framework“. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, S. 2623–2631. ISBN: 9781450362016. DOI: 10.1145/3292500.3330701. URL: https://doi.org/10.1145/3292500.3330701.

Burke, Robin (Nov. 2002). „Hybrid Recommender Systems: Survey and Experiments“. In: *User Modeling and User-Adapted Interaction* 12.4, S. 331–370. ISSN: 1573-1391. DOI: 10.1023/A:1021240730564. URL: https://doi.org/10.1023/A:1021240730564.

Carbonell, Jaime und Jade Goldstein (1998). „The use of MMR, diversity-based reranking for reordering documents and producing summaries“. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: Association for Computing Machinery, S. 335–336. ISBN: 1581130155. DOI: 10.1145/290941.291025. URL: https://doi.org/10.1145/290941.291025.

Covington, Paul, Jay Adams und Emre Sargin (Sep. 2016). „Deep Neural Networks for YouTube Recommendations“. In: S. 191–198. DOI: 10.1145/2959100.2959190.

Guo, Ruiqi u. a. (2020). „Accelerating large-scale inference with anisotropic vector quantization“. In: *Proceedings of the 37th International Conference on Machine Learning*. ICML'20. JMLR.org.

He, Xiangnan u. a. (2017). „Neural Collaborative Filtering“. In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, S. 173–182. ISBN: 9781450349130. DOI: 10.1145/3038912.3052569. URL: https://doi.org/10.1145/3038912.3052569.

Hu, Yifan, Yehuda Koren und Chris Volinsky (Dez. 2008). „Collaborative Filtering for Implicit Feedback Datasets“. In: *2008 Eighth IEEE International Conference on Data Mining*. Pisa, Italy: IEEE, S. 263–272. ISBN: 978-0-7695-3502-9. DOI: 10.1109/ICDM.2008.22. URL: http://ieeexplore.ieee.org/document/4781121/ (besucht am 04. 08. 2025).

Järvelin, Kalervo und Jaana Kekäläinen (Okt. 2002). „Cumulated gain-based evaluation of IR techniques“. In: *ACM Trans. Inf. Syst.* 20.4, S. 422–446. ISSN: 1046-8188. DOI: 10.1145/582415.582418. URL: https://doi.org/10.1145/582415.582418.

Krichene, Walid und Steffen Rendle (Aug. 2020). „On Sampled Metrics for Item Recommendation“. en. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM, S. 1748–1757.

ISBN: 978-1-4503-7998-4. DOI: 10.1145/3394486.3403226. URL: https://dl.acm.org/doi/10.1145/3394486.3403226 (besucht am 01. 09. 2025).

Linden, Greg, B. Smith und J. York (Feb. 2003). „Linden G, Smith B and York J: 'Amazon.com recommendations: item-to-item collaborative filtering', Internet Comput. IEEE, , 7". In: *Internet Computing, IEEE* 7, S. 76–80. DOI: 10.1109/MIC.2003.1167344.

Lops, Pasquale, Marco de Gemmis und Giovanni Semeraro (Jan. 2011). „Content-based Recommender Systems: State of the Art and Trends". In: S. 73–105. ISBN: 978-0-387-85819-7. DOI: 10.1007/978-0-387-85820-3_3.

Newman, Sam (2015). *Building Microservices: Designing Fine-Grained Systems*. O'Reilly Media. ISBN: 978-1491950357.

Nguyen, Tien T. u. a. (2014). „Exploring the filter bubble: the effect of using recommender systems on content diversity". In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. Seoul, Korea: Association for Computing Machinery, S. 677–686. ISBN: 9781450327442. DOI: 10.1145/2566486.2568012. URL: https://doi.org/10.1145/2566486.2568012.

Nielsen, Jakob (1993). „Response Times: The 3 Important Limits". In: *Nielsen Norman Group*. URL: https://www.nngroup.com/articles/response-times-3-important-limits/.

Raza, Shaina und Chen Ding (Jan. 2022). „News recommender system: a review of recent progress, challenges, and opportunities". In: *Artificial Intelligence Review* 55.1, S. 749–800. ISSN: 1573-7462. DOI: 10.1007/s10462-021-10043-x. URL: https://doi.org/10.1007/s10462-021-10043-x.

Reimers, Nils und Iryna Gurevych (2019). „Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Conference on Empirical Methods in Natural Language Processing*. URL: https://api.semanticscholar.org/CorpusID:201646309.

Rendle, Steffen u. a. (2009). „BPR: Bayesian personalized ranking from implicit feedback". In: *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09. Montreal, Quebec, Canada: AUAI Press, S. 452–461. ISBN: 9780974903958.

Sun, Philip u. a. (2023). „SOAR: improved indexing for approximate nearest neighbor search". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc.

Vaswani, Ashish u. a. (2017). „Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., S. 6000–6010. ISBN: 9781510860964.

Wu, Chuhan u. a. (Jan. 2023). „Personalized News Recommendation: Methods and Challenges". In: *ACM Trans. Inf. Syst.* 41.1. ISSN: 1046-8188. DOI: 10.1145/3530257. URL: https://doi.org/10.1145/3530257.

Erklärung zum Einsatz von KI-basierten Werkzeugen

Zur Verwendung KI-gestützter Werkzeuge erkläre ich in Kenntnis des Hinweisblatts „Hinweise zum Einsatz von KI-basierten Werkzeugen bei der Anfertigung wissenschaftlicher Arbeiten, u.a. im prüfungsrechtlichen Kontext“ Folgendes:

- Ich habe mich aktiv über die Leistungsfähigkeit und Beschränkungen der in meiner Arbeit eingesetzten KI-Werkzeuge informiert.
- Bei der Anfertigung der Arbeit habe ich durchgehend eigenständig und beim Einsatz KI-gestützter Werkzeuge maßgeblich steuernd gearbeitet.
- Insbesondere habe ich die Inhalte entweder aus wissenschaftlichen oder anderen zugelassenen Quellen entnommen und diese gekennzeichnet oder diese unter Anwendung wissenschaftlicher Methoden selbst entwickelt.
- Mir ist bewusst, dass ich als Autor/in der Arbeit die volle Verantwortung für die in ihr gemachten Angaben und Aussagen trage.
- Soweit ich KI-gestützte Werkzeuge zur Erstellung der Arbeit eingesetzt habe, sind diese jeweils mit dem Produktnamen, den formulierten Eingaben (Prompts), der Einsatzform sowie der entsprechenden Seiten-/Bereichsreferenzierung auf die Arbeit im KI-Verzeichnis am Ende der Arbeit vollständig ausgewiesen und im Text belegt (z.B. als Fußnote).

KI-Verzeichnis:

KI-Werkzeug	**Prompts (Beispiele)**	**Einsatzform**	**Bereich**
ChatGPT	"Hier ist Kapitel X meiner Arbeit. Bitte überprüfe es auf roten Faden, wissenschaftlichen Stil und technische Konsistenz."	Wissenschaftliches Lektorat und Feedback zur Argumentationsstruktur	Kapitel 1-6
Gemini	"Wie könnte ich diesen Satz weniger dramatisch formulieren?"	Gezielte Formulierungshilfe zur Verbesserung des wissenschaftlichen Ausdrucks	Kapitel 1
Gemini	"Wie kann ich die Überlegenheit des Hybrid-Modells argumentieren, ohne die Einzelkomponenten separat zu evaluieren?"	Unterstützung bei der Interpretation der Optimierungsergebnisse und Visualisierungen	Kapitel 4.5

Selbständigkeitserklärung

Ich versichere hiermit, dass ich meine
Seminararbeit mit dem Thema

Optimierung und Implementierung eines
Artikel-Empfehlungssystems für die SV-Gruppe auf der Google
Cloud Platform

selbständig verfasst und keine anderen als die angegebenen
Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass die
eingereichte elektronische Fassung mit der gedruckten Fassung
übereinstimmt.

Ravensburg, 28.09.2025

Ort, Datum	Unterschrift