Bachelorarbeit

im Studiengang Data Science Management

Unüberwachte Früherkennung von Trends durch statistische Analyse von Reddit Communities

---

vorgelegt von
Johannes Hald

Erstkorrektor:	Professor Dr. Stefan Faußer
Zweitkorrektor:	Professor Dr. Alexander Bartel
Matrikelnummer:	321761
Anschrift Verfasser:	Steubenstraße 2/3, 89231 Neu-Ulm
	johannes@hald.com
Themenvergabe:	1. November 2025
Arbeit eingereicht:	Neu-Ulm, 1. März 2026

Zusammenfassung

Die frühzeitige Identifikation aufkommender Trends verschafft Akteuren einen strategischen Vorteil durch proaktive Positionierung vor dem Mainstreamdurchbruch. Werkzeuge wie Google Trends messen jedoch die aggregierte Aufmerksamkeit breiter Bevölkerungsschichten und erfassen Phänomene daher erst nach ihrem Durchbruch. In der vorliegenden Arbeit wird ein System zur Früherkennung von Vorläufersignalen durch kontrastive Korpusanalyse spezialisierter Reddit Communities gegen einen Wikipedia-Referenzkorpus entwickelt. Die zentrale Hypothese lautet, dass emergente Trends zunächst in thematisch fokussierten Nischencommunities diskutiert werden, bevor sie den Mainstream erreichen. Das System setzt eine dreistufige Filterkaskade ein, um den Kandidatenraum schrittweise zu reduzieren. Die erste Phase extrahiert domänenspezifische Termkandidaten auf Basis des Unithood/Termhood Frameworks, das linguistische Stabilität durch Kohäsionsmetriken wie NPMI und Domänenspezifität durch kontrastive Frequenzvergleiche mittels Log-Likelihood Ratio quantifiziert. In der zweiten Phase wird die temporale Dynamik der Kandidaten mithilfe statistischer Zeitreihenverfahren analysiert, um transiente Spikes von nachhaltigen Wachstumsmustern zu unterscheiden. Die dritte Phase filtert semantisch irrelevante Kandidaten mittels Large Language Model gestützter Relevanzprüfung und transformiert heterogene Oberflächenformen in kanonische Termrepräsentationen. Die retrospektive Evaluation simuliert das System an 45 historischen Trends aus drei Domänen zu definierten Zeitpunkten vor dem Mainstreamdurchbruch. Das System erzielt eine Detection Rate von 40 % bei einem Mean Reciprocal Rank von 0,587. Drei Ablation Studies validieren zentrale Designentscheidungen: Der semantische LLM Filter steigert die Rangqualität, die inverse Gewichtung kleinerer Subreddits verbessert die Detection Rate um 13 Prozentpunkte, und ein Beobachtungszeitraum von 60 Tagen stellt den optimalen Kompromiss zwischen statistischer Stabilität und Frühzeitigkeit dar. Der domänenspezifische Gradient offenbart die Systemgrenzen: Organisch wachsende Trends werden zuverlässig detektiert, während releasegetriebene Trends ohne graduellen Diskursvorlauf konzeptionell unzugänglich bleiben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Quellcodeverzeichnis

Abkürzungsverzeichnis

AI Artificial Intelligence

API Application Programming Interface

ATR Automatic Term Recognition

CSV Comma-Separated Values

DR Detection Rate

GPT Generative Pre-trained Transformer

ICL In-Context Learning

IDF Inverse Document Frequency

JSON JavaScript Object Notation

LFS Large File Storage

LLM Large Language Model

LLR Log-Likelihood Ratio

LRU Least Recently Used

MRR Mean Reciprocal Rank

NLP Natural Language Processing

NPMI Normalized Pointwise Mutual Information

PMI Pointwise Mutual Information

POS Part-of-Speech

REST Representational State Transfer

TDT Topic Detection and Tracking

TF-IDF Term Frequency–Inverse Document Frequency

1 Einleitung

1.1 Problem und Relevanz

Die frühzeitige Identifikation aufkommender Trends verschafft Unternehmen, Content Creators und Marktanalysten einen erheblichen Wettbewerbsvorteil durch proaktive Strategieentwicklung in Bereichen wie Produktinnovation, Contentstrategie und Investmentallokation. Das Zeitfenster zwischen der ersten Erscheinung eines Phänomens in spezialisierten Diskursräumen und seinem Durchbruch in den Mainstream gibt Akteuren die Möglichkeit, strategische Positionen zu beziehen, bevor Wettbewerber auf bereits offensichtliche Signale reagieren. Diese zeitliche Vorauslaufphase bildet das zentrale Erkenntnisinteresse dieser Arbeit. Der ökonomische Wert dieser Früherkennung besteht nicht in der bloßen Quantifizierung bestehender Phänomene, sondern darin, Ressourcen auf Entwicklungen allokieren zu können, die noch keine breite Marktdurchdringung erreicht haben.

Bestehende Lösungen wie Google Trends erfassen Phänomene erst, nachdem sie sich in den aggregierten Suchanfragen der breiten Öffentlichkeit manifestiert haben, und eignen sich daher primär zur Quantifizierung offenkundiger Trends, nicht jedoch zur Identifizierung ihrer Vorläufersignale. Für Akteure, die von einer frühzeitigen Positionierung profitieren würden, ist dieser Detektionszeitpunkt systematisch zu spät. Die fundamentale Limitierung liegt in der Informationsquelle selbst: Aggregierte Suchvolumina repräsentieren per Definition bereits durchgebrochene Phänomene, da die schiere Masse an Suchanfragen das Überschreiten einer Mainstreamschwelle voraussetzt. Eine Lösung für das Früherkennungsproblem erfordert demnach keine algorithmische Innovation auf bestehenden Datenquellen, sondern die Erschließung alternativer Signalräume.

Die Hypothese dieser Arbeit besagt, dass spezialisierte Communities als ergänzender Datenraum für die Trendfrüherkennung genutzt werden können. Luhn [1] etablierte das Konzept der Technese, dem zufolge sich Fachsprache in spezialisierten Diskursräumen konzentriert und dort statistisch überrepräsentiert auftritt. Reddit macht sich diese strukturelle Eigenschaft durch seine communitybasierte Architektur zunutze, in der thematisch abgegrenzte Subreddits als spezialisierte Diskursräume fungieren, deren Mitglieder häufig als Early Adopters neuer Technologien, Konzepte oder kultureller Phänomene agieren. Die zentrale Forschungsfrage lautet: Eignen sich Aktivitätsmuster und sprachliche Anomalien in diesen Communities als quantifizierbare Frühindikatoren für spätere Mainstreamtrends, und wie kann ein System konzipiert werden, das diese Signale aus dem Informationsrauschen digitaler Diskussionen zuverlässig isoliert? Die methodische Herausforderung besteht darin, statistische Auffälligkeiten von semantischer Relevanz zu unterscheiden, da nicht jede terminologische Anomalie einen substantiellen Trend repräsentiert.

1.2 Stand der Forschung und Forschungslücke

Die Topic Detection and Tracking Initiative [2] etablierte First Story Detection als kanonisches Problem der automatisierten Eventerkennung. Der klassische Topic Detection and Tracking (TDT) Ansatz basiert auf Dokumentenclustering mittels Vektorraummodellen, wobei die Novelty eines Dokuments als dessen Dissimilarität zum nächstgelegenen bestehenden Cluster quantifiziert wird. Ein Dokument wird als First Story klassifiziert, wenn sein Novelty Score einen definierten Schwellwert überschreitet, der die Balance zwischen False Positives und False Negatives reguliert. Die Ähnlichkeitsmessung operiert dabei typischerweise auf Term Frequency–Inverse Document Frequency (TF-IDF) gewichteten Vektorrepräsentationen.

Die empirische Validierung des TDT Frameworks offenbarte jedoch fundamentale Limitationen der dokumentenzentrierten Clusteringstrategie für die Identifikation schwach strukturierter, emergenter Phänomene. Allan et al. konstatierten: „We believe that we have hit the limits of effectiveness that can be reached with simple IR-based approaches to story/topic comparison" und forderten explizit „substantially different approaches and ideas"[2]. Diese Einschränkung resultiert daraus, dass Dokumentenclustering auf globaler Ähnlichkeit basiert und somit anfällig für Rauschen durch stilistische Varianz, strukturelle Artefakte und thematische Drift ist. Dieser Ansatz erweist sich als unzureichend, um emergente Trends zu identifizieren, die sich zunächst durch subtile terminologische Verschiebungen manifestieren, bevor sie als kohärente thematische Cluster erkennbar werden.

Diese Limitation wird durch kontrastive Korpusanalyse adressiert, bei der die Analyseebene vom Dokument auf den Term verlagert wird. Dunning [3] demonstrierte, dass das Log-Likelihood Ratio domänenspezifische Terminologie durch den Vergleich der Frequenzverteilung zwischen Ziel- und Referenzkorpus identifiziert, wobei die Robustheit dieses Maßes bei niedrigen Frequenzen für die Früherkennung essenziell ist. Das von Kageura und Umino [4] formalisierte Unithood/Termhood Framework strukturiert die Extraktion durch zwei komplementäre Dimensionen: Unithood validiert die linguistische Stabilität eines Kandidaten durch Kohäsionsmetriken wie Normalized Pointwise Mutual Information (NPMI), während Termhood dessen Domänenspezifität durch kontrastive Frequenzvergleiche quantifiziert. Diese Perspektive ermöglicht die Identifikation terminologischer Anomalien in einem Stadium, in dem diese noch keine kohärenten thematischen Cluster bilden, sondern als statistisch auffällige Einzelsignale in heterogenen Diskussionskontexten auftreten. Die statistische Identifikation domänenspezifischer Terminologie ist jedoch nicht ausreichend für die Trenddetektion, da sich Domänenspezifität und zeitliche Emergenz nicht gleichsetzen lassen. Kleinberg [5] erweiterte die kontrastive Perspektive um die zeitliche Dimension, indem er Bursts als Zustandswechsel in stochastischen Automaten modellierte.

Die Kombination aus statistischer Termextraktion und Zeitreihenanalyse führt zwangsläufig zu False Positives, da beide Verfahren ausschließlich auf quantitativen Mustern basieren und somit strukturelles Rauschen nicht von semantisch relevanten Signalen unterscheiden können. Brown et al. [6] demonstrierten, dass Large Language Models (LLMs) durch In-Context Learning zur Generalisierung auf ungesehene Aufgaben befähigt sind, wobei die Modellparametrisierung implizites Weltwissen kodiert, das zur semantischen Inferenz genutzt werden kann. Diese Eigenschaft ermöglicht die semantische Validierung statistischer Kandidaten durch Relevanzprüfung gegenüber Nutzerinteressen sowie die Kanonisierung heterogener Oberflächenformen zu konsistenten Termrepräsentationen. Die systematische Integration statistischer Extraktion, zeitreihenbasierter Validierung und LLM gestützter semantischer Filterung zu einer mehrstufigen Filterkaskade wurde bislang nicht untersucht und bildet die methodische Lücke, die diese Arbeit adressiert.

1.3 Forschungsfrage und Zielsetzung

Die zentrale Forschungsfrage dieser Arbeit lautet: *Wie lässt sich ein System entwickeln, das Vorläufersignale von Mainstreamtrends aus dem Informationsrauschen auf Reddit zuverlässig isoliert?* Diese übergeordnete Frage wird durch drei spezifische Unterfragen operationalisiert, die zentrale Designentscheidungen des Systems empirisch validieren und deren Beitrag zur Gesamtperformance quantifizieren.

RQ1 (Semantische Filterung): Inwieweit steigert ein nachgelagerter semantischer LLM Filter die Präzision des Systems im Vergleich zu rein statistischen Baselineverfahren, und rechtfertigt der Qualitätsgewinn den zusätzlichen Ressourcenaufwand durch externe Application Programming Interface (API) Calls?

RQ2 (Inverse Gewichtung): Führt die inverse logarithmische Gewichtung von Subreddits nach Communitygröße zu einer messbaren Verbesserung der Detection Rate im Vergleich zu einer ungewichteten Aggregation, oder verstärkt sie primär irrelevante Signale aus marginalisierten Nischencommunities?

RQ3 (Beobachtungszeitraum): Welcher Beobachtungszeitraum für die Zeitreihenanalyse stellt den optimalen Kompromiss zwischen den konkurrierenden Evaluationsmetriken Detection Rate und Lead Time dar, wobei längere Zeiträume mehr historischen Kontext liefern, jedoch die Früherkennungsfähigkeit reduzieren?

1.4 Methodischer Beitrag und Evaluationsstrategie

Das entwickelte System implementiert eine dreistufige Filterkaskade, die statistische Robustheit mit semantischer Intelligenz kombiniert und eine systematische Reduktion des Kandidatenraums (ca. 25.000 Posts auf ca. 10 bis 15 finale Trends) erreicht. Phase 1 (Abschnitt 4.3) identifiziert domänenspezifische Termkandidaten durch kontrastive Korpusanalyse, indem Reddit-Diskussionen gegen einen Wikipedia-Referenzkorpus kontrastiert werden. Die Validierung erfolgt durch Unithoodfilter zur Prüfung linguistischer Stabilität sowie Termhoodscores zur Quantifizierung von Domänenspezifität. Phase 2 (Abschnitt 4.4) quantifiziert Trendsignale mittels Fisher's Exact Test (Burstereignisse), Mann-Kendall Test (Trendkonsistenz) und logarithmischer Growth Rate (Wachstumsdynamik). Phase 3 (Abschnitt 4.5) filtert die verbleibenden Kandidaten durch LLM basierte Relevanzprüfung und transformiert heterogene Oberflächenformen in kanonische Termrepräsentationen. Diese Architektur priorisiert Präzision über Vollständigkeit nach dem von Gale und Church [7] etablierten Prinzip, wonach für statistische Natural Language Processing (NLP) Verfahren Verlässlichkeit bedeutsamer ist als Abdeckung.

Die Evaluation (Kapitel 6) quantifiziert die Systemleistung durch retrospektive Simulation zu definierten Zeitpunkten vor dem historisch bekannten Mainstreamdurchbruch von 45 Trends über drei Domänen (Artificial Intelligence, Cryptocurrency/Web3, Gaming), wobei die Diversität der Domänen die Generalisierbarkeit der Methodik testet. Der Mainstreamdurchbruch wird algorithmisch durch Wikipedia Pageviews operationalisiert (Abschnitt 6.1.1): Ein Trend gilt als durchgebrochen, wenn der 7-Tage Rolling Mean das 1,5-fache der 30-Tage Baseline überschreitet und gleichzeitig einen absoluten Schwellwert von 300 Views oder 5% des globalen Peaks erfüllt. Simulationsfenster von 7 bis 90 Tagen vor diesem Ereignis testen die Früherkennungsfähigkeit bei variierender Lead Time. Die Forschungsfragen werden durch Ablation Studies beantwortet (Abschnitte 6.2.2, 6.2.3, 6.2.4), die systematisch Systemkomponenten isolieren und deren Beitrag zur Gesamtperformance durch Vergleich mit konfigurierten Baselineverfahren quantifizieren. Die primäre Evaluationsmetrik ist die Detection Rate, definiert als Anteil der korrekt detektierten Ground Truth Trends pro Simulationsfenster und Domäne (Abschnitt 6.1.2).

2 Grundlagen

2.1 Begriffsdefinitionen

Ein **Trend** bezeichnet ein Phänomen, das in spezialisierten Diskursgemeinschaften zunehmend thematisiert wird und das Potenzial für eine breitere gesellschaftliche Relevanz aufweist. Im Kontext dieser Arbeit manifestieren sich Trends als identifizierbare sprachliche Muster, die sich durch drei Dimensionen auszeichnen: Neuartigkeit gegenüber etablierter Fachsprache, zeitliches Wachstum der Diskussionsfrequenz sowie thematische Relevanz für ein definiertes Interessensgebiet. Diese Definition folgt der ereignisbasierten Konzeption des Topic Detection and Tracking Frameworks [2], das für die Identifikation konzeptioneller Entwicklungen anstelle konkreter Ereignisse adaptiert wurde.

Ein **aufkommender Trend** ist ein Trend in einem frühen Entwicklungsstadium, das sich durch eine messbare Aufwärtsdynamik auszeichnet. Die Abgrenzung zu etablierten Phänomenen erfolgt nicht durch die absolute Diskussionsfrequenz, sondern durch nachweisbares Wachstum im Verhältnis zu einer historischen Basislinie. Diese Definition orientiert sich an Kleinbergs Konzeption von Bursts als Zustandswechsel in zeitlich geordneten Datenströmen [5], die um die Anforderung einer nachhaltigen Entwicklung erweitert wurde.

Früherkennung bezeichnet die Identifikation von Trends in einer zeitlichen Phase, die ihrer Wahrnehmung in breiten Bevölkerungsschichten vorausgeht. Die zentrale Annahme lautet, dass spezialisierte Communities Phänomene diskutieren, bevor diese im Mainstream oder in allgemeinen Diskursräumen sichtbar werden. Diese Hypothese basiert auf Rogers' Diffusionstheorie [8], wonach Innovationen von einer kleinen Gruppe von Innovatoren initiiert werden, bevor sie sich über soziale Netzwerke verbreiten.

Eine **Nischencommunity** ist eine thematisch spezialisierte Online-Diskussionsgruppe, die sich durch zwei Merkmale auszeichnet: eine vergleichsweise geringe Mitgliederzahl und eine hohe thematische Fokussierung. Basierend auf Luhns Konzept der Technese [1] – der Konzentration von Fachsprache in spezialisierten Diskursräumen – werden kleine Communities als qualitativ wertvollere Frühindikatoren für emergente Phänomene verstanden. Die operative Definition erfordert nachweisbare kontinuierliche Aktivität, um sporadische von nachhaltigen Diskussionsgemeinschaften unterscheiden zu können [9].

2.2 Die Diskussionsplattform Reddit

Reddit ist mit über 73 Millionen täglichen Nutzern und mehr als 100.000 aktiven Communities (Stand: März 2024) [10] eine der größten nutzergenerierten Diskussionsplattformen im Internet. Die Plattform ist in thematisch spezialisierte Diskussionsforen, sogenannte Subreddits, organisiert. Diese bilden jeweils eine eigenständige Community mit spezifischen Regeln, Moderationspraktiken und kulturellen Normen.

In diesen Subreddits können Nutzer Beiträge veröffentlichen, kommentieren und durch ein Votingsystem (Upvotes/Downvotes) die Sichtbarkeit von Inhalten beeinflussen.

Reddit weist drei strukturelle Eigenschaften auf, die es als Datenquelle für die Früherkennung aufkommender Trends qualifizieren: Erstens ermöglicht die Unterteilung in Subreddits die gezielte Analyse hochspezialisierter Nischencommunities. Diese fungieren nach Rogers' Diffusionstheorie [8] als primäre Quelle für Innovationen. Die Bandbreite reicht von Communities mit mehreren Millionen Mitgliedern bis zu hochspezialisierten Foren mit wenigen Tausend aktiven Nutzern. Gerade letztere sind für die Identifikation von Vorläufersignalen von entscheidender Bedeutung. Zweitens operationalisiert das Abstimmungssystem (Upvotes/Downvotes) ein kollektives Relevanzranking. Dadurch werden diskussionswürdige Inhalte identifiziert und das System dient als Proxy für die wahrgenommene Bedeutung eines Themas innerhalb einer Community. Drittens gewährleistet die persistente und öffentlich zugängliche Archivierung aller Beiträge, dass historische Diskussionsdaten für retrospektive Zeitreihenanalysen verfügbar sind.

Reddit fungiert somit nicht als generisches soziales Netzwerk, sondern als stratifiziertes Ökosystem aus Diskursgemeinschaften, dessen Architektur eine systematische Analyse von Diffusionsprozessen von der Nische in den Mainstream ermöglicht.

2.3 Automatische Termextraktion

2.3.1 Unithood/Termhood Framework

Um Fachtermini automatisch identifizieren zu können, müssen allgemeinsprachliche und domänenspezifische Begriffe unterschieden werden. Ein Term kann häufig auftreten, ohne fachspezifisch zu sein (z. B. „System" in technischen Diskussionen), während hochspezifische Terme statistisch unauffällig bleiben können, wenn sie nur in Nischen diskutiert werden. Diese Ambiguität erfordert eine systematische Dekomposition der Termqualität.

Kageura und Umino [4] etablierten hierfür das Unithood/Termhood-Framework, das zwei komplementäre Qualitätskriterien definiert. Unithood quantifiziert die linguistische Stabilität einer Wortsequenz, also den Grad, zu dem Wörter als kohäsive Einheit auftreten. Termhood quantifiziert die domänenspezifische Relevanz, also den Grad, zu dem ein Begriff für eine Fachdomäne charakteristisch ist. Ein Term wie „Neural Network" weist eine hohe Unithood (die Komponenten treten nicht zufällig gemeinsam auf) und eine hohe Termhood im Kontext von Machine Learning auf (er kommt signifikant häufiger vor als in allgemeiner Sprache).

Mithilfe dieser Dekomposition ist ein zweistufiger Filteransatz möglich: Zunächst eliminieren Unithoodfilter linguistisch instabile Kandidaten, anschließend prüfen Termhood-Maße die statistische Domänenrelevanz. Diese Architektur reduziert die kombinatorische Komplexität und erhöht die Präzision, indem kategorial unplausible Terme eliminiert werden.

Für unüberwachte Systeme ist dieses Framework methodisch von entscheidender Bedeutung, da keine annotierten Trainingsdaten verfügbar sind. Die Anwendung auf Social Media-Diskurse erfordert eine domänenspezifische Anpassung: Unithoodfilter müssen plattformtypische Terminologie (Akronyme, Versionsnummern, Produktnamen) abbilden, während Termhood-Maße die statistische Auffälligkeit gegenüber einem allgemeinsprachlichen Referenzkorpus quantifizieren. Die konkrete Operationalisierung ist in Abschnitt 4.3 dokumentiert.

2.3.2 Statistische Assoziationsmaße

Zur Operationalisierung des Unithood-/Termhood-Frameworks sind quantitative Maße zur Bewertung der linguistischen Stabilität und der domänenspezifischen Relevanz erforderlich. Die Maßauswahl ist methodisch kritisch, da aufkommende Trends per Definition initial selten sind und somit statistische Verfahren benötigen, die bei niedrigen Frequenzen robuste Ergebnisse liefern. Die nachfolgend aufgeführten Maße wurden aufgrund dieser spezifischen Robustheitseigenschaften etabliert.

Log-Likelihood Ratio Das Log-Likelihood Ratio (LLR) [3] quantifiziert die statistische Signifikanz einer Frequenzdifferenz zwischen zwei Korpora durch einen Test, der auf dem Likelihood-Quotienten basiert. Das Maß liefert auch bei erwarteten Häufigkeiten von weniger als 1 exakte Ergebnisse und ist somit für die Identifizierung seltener Terme unerlässlich. Es operationalisiert die Intuition der „statistischen Überraschung“: Wie unwahrscheinlich ist die beobachtete Frequenzverteilung, wenn der Term in beiden Korpora gleich häufig vorkäme? Dunning zeigte, dass LLR bei niedrigen Frequenzen signifikant präzisere Signifikanzwerte liefert als traditionelle Verfahren. Dadurch etablierte es sich als Goldstandard für kontrastive Korpusanalyse, wie in [4] beschrieben.

Weirdness Ratio Der von Ahmad et al. [11] entwickelte Weirdness Index operationalisiert Termhood als Verhältnis relativer Frequenzen zwischen Fach- und Referenzkorpus. Ein Wert von 100 bedeutet, dass ein Term im Zielkorpus 100-mal häufiger auftritt als im Referenzkorpus – relativ zur jeweiligen Korpusgröße. Diese Ratio ist zwar intuitiv interpretierbar, jedoch instabil bei sehr seltenen Termen, da die Division durch eine Zahl, die nahe Null liegt, zu numerisch extremen Werten führt. Ahmad et al. adressierten diese Schwäche, indem sie Frequenzschwellen kombinierten, und etablierten einen empirischen Cutoff von 100.

Normalized Pointwise Mutual Information Pointwise Mutual Information (PMI) [12] quantifiziert die Assoziationsstärke zwischen zwei Wörtern durch Vergleich ihrer gemeinsamen Auftretenswahrscheinlichkeit mit dem Produkt ihrer individuellen Wahrscheinlichkeiten.

Bouma normalisierte den PMI auf das Intervall [-1, 1], wodurch Vergleiche zwischen Bigrammen unterschiedlicher Frequenz ermöglicht werden [13]. Normalized Pointwise Mutual Information (NPMI) operationalisiert Unithood durch die Quantifizierung der „Klebkraft" zwischen Wörtern. Ein Bigramm mit hohem NPMI verhält sich statistisch wie eine lexikalische Einheit, da seine Komponenten signifikant häufiger gemeinsam auftreten, als es anhand ihrer Einzelfrequenzen zu erwarten wäre. Kohäsive Kollokationen, wie sie etwa bei "machine learning" zu finden sind, unterscheiden sich so in ihrer Struktur von zufälligen syntaktischen Nachbarschaften.

Inverse Document Frequency Inverse Document Frequency (IDF) [14] quantifiziert die Diskriminationskraft eines Terms durch seine Streuung über Dokumente. Die Intuition: Ein Term, der nur in wenigen spezialisierten Dokumenten auftritt, ist informativer als ein allgegenwärtiger Term. Um die nicht-lineare Beziehung zwischen Dokumentfrequenz und Informationsgehalt abzubilden, werden die IDF-Werte logarithmisch skaliert. Dies ist methodisch relevant für Daten aus sozialen Medien, da Terme mit einer Dokumentfrequenz von über 30 Prozent wahrscheinlich allgemeinsprachlich oder plattformspezifisches Rauschen sind, unabhängig von ihrer domänenspezifischen Überrepräsentation.

2.4 Statistische Verfahren der Zeitreihenanalyse

Um Trendsignale statistisch zu validieren, muss man transiente Aktivitätsspitzen von nachhaltigen Wachstumsmustern unterscheiden. Ein Term kann durch ein einmaliges Ereignis eine kurzfristige Diskussion auslösen, ohne ein aufkommendes Phänomen zu repräsentieren. Diese Herausforderung wird durch eine Zeitreihenanalyse mit statistischen Tests adressiert. Diese quantifizieren sowohl die Signifikanz aktueller Aktivitäten als auch die Konsistenz der Entwicklung über den Beobachtungszeitraum. Die in diesem System eingesetzten Verfahren basieren auf dem im TDT-Framework etablierten Prinzip der statistischen Abweichung von einer Baseline.

Fisher's Exact Test Der Fisher's Exact Test [15] quantifiziert die statistische Signifikanz einer Frequenzdifferenz zwischen zwei Zeitfenstern mittels eines exakten Wahrscheinlichkeitstests für Kontingenztabellen. Im Gegensatz zu approximativen Tests wie dem Chi Quadrat Test berechnet er auch bei niedrigen Erwartungswerten exakte p-Werte. Das ist für die Analyse aufkommender Trends mit anfänglich geringen Frequenzen essenziell. Der Test operationalisiert das Konzept der „Violation of Stationarity"[2], indem er prüft, ob die aktuelle Diskussionsfrequenz eines Terms statistisch signifikant von seiner historischen Baseline abweicht. Im vorliegenden System dient Fisher's Test zur Identifikation von Bursts durch Kontrastierung eines aktuellen Zeitfensters gegen ein historisches Zeitfenster (siehe Abschnitt 4.4.1).

Mann-Kendall Test Der Mann-Kendall Test [16], [17] ist ein nichtparametrisches Verfahren zur Erkennung monotoner Trends in Zeitreihen. Er quantifiziert die Korrelation zwischen Zeitpunkten und beobachteten Werten mithilfe von Kendall's Tau. Ein positiver Tau-Wert deutet auf einen konsistenten Aufwärtstrend hin, ein negativer Wert auf einen Abwärtstrend. Die zentrale methodische Stärke liegt in der Robustheit gegenüber Ausreißern sowie dem Fehlen von Verteilungsannahmen. Dadurch ist der Test besonders gut für verrauschte Daten aus sozialen Medien geeignet. Der Test operationalisiert Kleinbergs Konzept der Unterscheidung zwischen Zustandswechseln und temporären Spikes [5], indem er die Konsistenz der Aufwärtsentwicklung über den gesamten Beobachtungszeitraum misst. Im vorliegenden System validiert der Mann-Kendall Test, ob ein statistisch signifikanter Burst (Fisher's Test) auch eine robuste, monotone Wachstumsdynamik aufweist (siehe Abschnitt 4.4.2).

TDT-Prinzip und Burst Detection Das TDT-Framework [2] etablierte das Prinzip, dass aufkommende Phänomene durch statistisch auffällige Abweichungen von erwarteten Frequenzmustern identifiziert werden können, wobei χ^2-basierte Tests sogenannte „Violations of Stationarity" durch Kontrastierung von Termfrequenzen über Zeitfenster operationalisieren. Die zentrale Einsicht lautet: Ein Trend manifestiert sich nicht durch absolute Häufigkeit, sondern durch einen signifikanten Anstieg im Verhältnis zu einer Baseline. Kleinberg erweiterte dieses Konzept, indem er Bursts als Zustandswechsel in einem probabilistischen Automaten formal modellierte [5], wodurch robuste Aktivitätsperioden von transienten Spikes differenziert werden.

2.5 Large Language Models und Few-Shot Learning

Brown et al. [6] demonstrierten, dass große vortrainierte Sprachmodelle die Fähigkeit entwickeln, neue Aufgaben durch kontextuelle Beispiele im Prompt zu erlernen, ohne dass eine Anpassung der Modellparameter erforderlich ist. Bei diesem als In-Context Learning (ICL) bezeichneten Paradigma werden Aufgaben und Demonstrationsbeispiele ausschließlich durch textuelle Interaktion mit dem Modell spezifiziert, ohne dass eine Gradientenoptimierung durchgeführt wird. Die Autoren interpretieren diesen Mechanismus als Meta-Learning, bei dem das eigentliche Lernen nicht durch Gewichtsänderungen, sondern durch Aktivierungsmuster innerhalb des Kontextfensters realisiert wird. Die zentrale empirische Erkenntnis lautet: Während Zero-Shot Performance linear mit der Modellgröße skaliert, steigt Few-Shot Performance überproportional an, was darauf hindeutet, dass größere Modelle signifikant bessere ICL-Fähigkeiten entwickeln.

ICL adressiert die Anforderungen des vorliegenden Systems durch einen fundamentalen Paradigmenwechsel von aufgabenspezifischen Architekturen hin zu aufgabenagnostischem Pretraining: Die Identifikation aufkommender Trends ist eine unüberwachte Entdeckungsaufgabe, für die per Definition keine annotierten Trainingsdaten existieren können, und das System muss domänenagnostisch operieren, da Nutzer beliebige Keywords angeben können. Die von Brown et al. dokumentierte Skalierungsgesetzmäßigkeit bildet die wissenschaftliche Grundlage für Forschungsfrage 1: Da die Leistungsdifferenz zwischen Zero-Shot und Few-Shot mit zunehmender Modellkapazität wächst, sollten größere Modelle (Generative Pre-trained Transformer (GPT)-4o) aufgrund ihrer überlegenen ICL-Fähigkeiten eine höhere semantische Validierungsqualität erreichen als kleinere Modelle (Llama 3 8B).

Die praktische Operationalisierung von ICL erfolgt durch Prompt Engineering. Liu et al. [18] definieren dies als systematischen Prozess zur Gestaltung von Prompts, die eine optimale Leistung auf der Zielaufgabe erzielen. Diese spezifizieren die Aufgabe, den Kontext und das erwartete Ausgabeformat. Liu et al. unterscheiden zwischen zwei Arten von Prompts: Prefix Prompts, die dem Input vorangestellt werden, um das Modellverhalten zu steuern, und die eigentliche Aufgabenspezifikation. Beim Trend Radar werden Prefix-Komponenten zur Rollensteuerung (System-Messages) mit aufgabenspezifischen Instruktionen (User-Messages) kombiniert. Um die maschinelle Verarbeitbarkeit zu gewährleisten, wird das Modell mittels Constrained Decoding dazu gezwungen, strukturierte JSON-Objekte auszugeben. Dabei werden nur Outputs zugelassen, die einer definierten Grammatik entsprechen. Die Reproduzierbarkeit wird durch eine deterministische Ausgabegenerierung sichergestellt. Liu et al. beschreiben diese Strategie als Argmax-Dekodierung, bei der der höchstbewertete Output ohne stochastisches Sampling ausgewählt wird. Diese Strategie wird durch das Setzen der Temperature auf 0,0 und die Spezifikation eines fixen Seeds operationalisiert. Dadurch wird das probabilistische Sprachmodell in ein deterministisches System transformiert. Die Kombination aus ICL und strukturiertem Prompt Engineering bildet das methodische Fundament für die semantische Validierung in Phase 3. Dabei wird das Modell mithilfe kontextueller Beispiele (Reddit-Posts) und einer expliziten Aufgabenspezifikation zur Filterung und Veredelung statistischer Trendsignale eingesetzt.

3 Stand der Forschung

3.1 Methoden zur Termextraktion

Die automatische Extraktion domänenspezifischer Terminologie ist eine fundamentale Herausforderung für die computerlinguistische Forschung. Manuelle Kuration durch Fachexperten gewährleistet zwar hohe Präzision, ist jedoch nicht skalierbar auf die Datenmengen moderner digitaler Diskurse.

Die methodische Entwicklung der Automatic Term Recognition (ATR)-Forschung lässt sich als Progression von statistischen Assoziationsmaßen über supervised Machine Learning bis hin zu Embedding-basierten Ansätzen charakterisieren. Dabei manifestiert jedes Paradigma spezifische Trade-offs zwischen Interpretierbarkeit, Performance und Generalisierbarkeit.

3.1.1 Statistische Fundamente

Luhn [1] etablierte das konzeptionelle Fundament durch die Beobachtung, dass sich Fachsprache in spezialisierten Diskursräumen konzentriert und dort statistisch überrepräsentiert auftritt. Dieses als „Technese" bezeichnete Phänomen bildet die Grundlage kontrastiver Korpusanalyse. Spärck Jones [14] formalisierte diese Intuition durch die IDF, die das inverse Verhältnis zwischen Dokumentfrequenz und Diskriminationskraft quantifiziert. Church und Hanks [12] erweiterten die statistische Terminologieforschung um die Dimension der Wortassoziationen durch PMI, basierend auf Firths distributionalistischem Prinzip „You shall know a word by the company it keeps". Die dokumentierte Einschränkung besteht in der Instabilität bei niedrigen Frequenzen.

Dunning [3] adressierte die Instabilität von PMI bei niedrigen Frequenzen durch den LLR als robustes Maß für die kontrastive Korpusanalyse. Die zentrale methodische Innovation bestand in der Verwendung exakter Likelihood-Quotienten anstelle von Normalverteilungsapproximationen, wodurch LLR zur Identifikation von „content-bearing words and nearly all the technical jargon" besonders geeignet ist. Für das vorliegende System ist LLR das primäre Termhood-Maß, da aufkommende Trends per Definition initial selten sind. Justeson und Katz [19] lieferten den empirischen Nachweis, dass über 97 Prozent der Fachterminologie aus einfachen Sequenzen von Nomen und Adjektiven bestehen. Diese strukturelle Regularität ermöglicht POS-basierte Filter zur Eliminierung grammatikalisch implausibler Kandidaten und operationalisiert das Unithood-Konzept. Ahmad et al. [11] operationalisierten die kontrastive Analyse durch die „Weirdness Ratio" und etablierten einen empirischen Schwellwert von 100 zur Differenzierung zwischen domänenspezifischer und allgemeiner Sprache.

Kageura und Umino [4] konsolidierten diese Entwicklungen im Unithood/Termhood-Framework (siehe Abschnitt 2.3.1), das sich als konzeptionelle Grundlage der ATR-Forschung etablierte und mehrstufige Filterarchitekturen ermöglicht. Dabei werden Kandidaten zunächst anhand von Unithood-Kriterien vorgefiltert, bevor rechenintensive Termhood-Bewertungen erfolgen.

3.1.2 Supervised Machine Learning

Lample et al. [20] etablierten neuronale Sequenzmodelle als State of the Art für die Named Entity Recognition durch eine Bi-LSTM-CRF-Architektur mit automatischem Feature-Learning.

Kucza et al. [21] demonstrierten die Übertragbarkeit auf ATR durch Formulierung als Sequence Labeling-Problem, wobei die Evaluation einen signifikanten Leistungsabfall bei Domänenwechsel offenbarte. Der TermEval 2020 Shared Task [22] lieferte systematische Evidenz für die Dominanz von Transformer-basierten Ansätzen. Hazem und Bouhandi [23] erzielten mit BERT-Fine-Tuning die höchste Performance, wobei die Evaluation über vier Domänen die starke Abhängigkeit von der Verfügbarkeit und Qualität domänenspezifischer Trainingsdaten demonstrierte.

Diese Befunde verdeutlichen die konzeptionelle Unvereinbarkeit von supervised Ansätzen mit den Anforderungen unüberwachter, domänenagnostischer Termidentifikation: Aufkommende Trends sind per Definition neuartig und existieren daher nicht in historischen Trainingskorpora. Die Generalisierung auf beliebige Keywords ohne aufgabenspezifische Annotation steht im fundamentalen Widerspruch zur Notwendigkeit, für jede neue Domäne annotierte Beispiele zu sammeln.

3.1.3 Embedding-basierte Ansätze

Amjadian et al. [24] quantifizierten die semantische Verschiebung zwischen allgemeinem und spezialisiertem Sprachgebrauch, indem sie globale (GloVe) und lokale (Word2Vec) Embeddings kombinierten. Die Distanz zwischen beiden Repräsentationen operationalisiert Termhood durch distributionelle Semantik. Die Limitation besteht in der Fokussierung auf Unigramme sowie dem Erfordernis ausreichender Datenmengen, um robuste, domänenspezifische Embeddings zu gewährleisten. Devlin et al. [25] setzten mit BERT neue Maßstäbe für kontextuelles Sprachverständnis. Die bidirektionale Transformer-Architektur generiert durch masked language modeling kontextsensitive Repräsentationen, die im Gegensatz zu statischen Word Embeddings kontextabhängige Bedeutungsnuancen erfassen.

Die Limitationen der unüberwachten Identifikation emergenter Phänomene zeigen sich auf drei Ebenen: Erstens erfordert die BERT-Inferenz für Millionen N-Gramm-Kandidaten erhebliche GPU-Ressourcen, während sich statistische Maße effizient auf CPUs skalieren lassen. Zweitens operieren Embeddings als „Black Box" ohne interpretierbare Entscheidungskriterien, was wissenschaftliche Reproduzierbarkeit beeinträchtigt. Drittens leiden sie unter dem Out-of-Vocabulary-Problem, da neue Trends wie „GPT-4" im Pretraining Korpus nicht existieren.

Die systematische Analyse zeigt, dass adaptierte statistische Verfahren die methodisch robusteste Wahl für die unüberwachte, domänenagnostische Identifikation emergenter Phänomene in verrauschten Social-Media-Daten darstellen. Klassische Assoziationsmaße sind wissenschaftlich etabliert, interpretierbar und effizient skalierbar. Supervised Methoden sind konzeptionell für neuartige Phänomene ungeeignet, Embeddings leiden unter Rechenintensität und dem Out-of-Vocabulary-Problem. Diese Erkenntnisse begründen die hybride Architektur des vorliegenden Systems, die statistische Robustheit für die initiale Kandidatenextraktion mit LLM-basierter semantischer Validierung kombiniert.

3.2 Methoden zur Trendanalyse

3.2.1 Statistische Zeitreihenanalyse für termbasierte Detektion

Um aufkommende Trends zu identifizieren, muss man transiente Aktivitätsspitzen von nachhaltigen Entwicklungsmustern in zeitlich geordneten Datenströmen unterscheiden. Hierfür etablierte das TDT-Framework das Prinzip der statistischen Abweichungserkennung als methodisches Fundament [2]. Allan et al. definierten Topics als „a seminal event or activity, along with all directly related events and activities" und operationalisierten deren Identifikation durch das Konzept der „violation of stationarity": Terme, deren Frequenzverteilung signifikant von der erwarteten Basislinie abweicht, deuten auf das Auftreten oder die Entwicklung eines diskussionswürdigen Phänomens hin. Die technische Umsetzung erfolgte durch χ^2-basierte Tests, die Frequenzen über Zeitfenster hinweg kontrastieren. Dieses Framework wurde primär für die retrospektive Clusterung von News-Artikeln zu kohärenten Event-Narrativen konzipiert. Die Limitationen für die Früherkennung manifestieren sich auf zwei Ebenen: Erstens operiert TDT auf diskreten, zeitlich abgrenzbaren Ereignissen wie Wahlen oder Naturkatastrophen, während emergente Trends graduell entstehen und keine klaren Ereignisgrenzen aufweisen. Zweitens erfolgt die Analyse auf Dokumentebene mithilfe von Ähnlichkeitsmaßen, während die Identifikation spezifischer Trendbegriffe eine termbasierte Granularität erfordert.

Kleinberg erweiterte die statistische Zeitreihenanalyse um ein probabilistisches Modell zur Burst Detection [5], das Aktivitätsströme als Sequenz von Intensitätszuständen formalisiert. Die zentrale Innovation besteht in der Unterscheidung zwischen Baseline-Aktivität und Burst-Phasen durch die Zuweisung von Kosten für Zustandswechsel. Dadurch werden robuste, langanhaltende Aktivitätsperioden von einzelnen Intensitätsspitzen differenziert. Das Modell basiert auf einem zustandsbasierten Automaten, der die Interarrivalzeiten zwischen Ereignissen Topic-Model-basierte Verfahrenanalysiert und signifikante Zustandswechsel als Bursts erkennt. Für aggregierte Zähldaten (batched arrivals) ist eine modifizierte Wahrscheinlichkeitsfunktion erforderlich, die statt kontinuierlicher Event-Streams tägliche oder stündliche Frequenzaggregationen verarbeitet. Kleinbergs Arbeit liefert die konzeptionelle Grundlage für die Unterscheidung zwischen „Zustandswechseln" (nachhaltigen Veränderungen) und „Spikes"Topic-Model-basierte Verfahren (transienten Fluktuationen), jedoch ohne semantische Validierung der identifizierten Terme.

Die Kombination aus Fisher's Exact Test zur Signifikanzprüfung und Mann-Kendall Test zur Konsistenzvalidierung (vgl. Abschnitt 2.4) operationalisiert Kleinbergs konzeptionelle Unterscheidung zwischen nachhaltigen Zustandswechseln und transienten Spikes und bildet das methodische Fundament für Phase 2 des vorliegenden Systems.

3.2.2 Event Detection in Social Media Streams

Für die Adaption von Trenderkennungsverfahren auf Social-Media-Daten müssen plattformspezifische Charakteristika berücksichtigt werden: eine hohe Veröffentlichungsfrequenz, eine heterogene Inhaltsqualität, ein nutzergeneriertes Vokabular und die Notwendigkeit der Echtzeitverarbeitung. Sakaki et al. [26] etablierten das Konzept der „Social Sensors", bei dem Twitter Nutzer als verteilte Sensoren für physische Ereignisse fungieren. Die Evaluation demonstrierte zeitliche Vorsprünge gegenüber offiziellen Messnetzwerken bei der Erdbeben Detektion. Die fundamentale Einschränkung besteht in der Domänenspezifität: Das System erfordert für jede Event Klasse annotierte Trainingsbeispiele, was die Anwendung auf die unüberwachte Identifikation neuartiger Trends konzeptionell ausschließt.

Velocity-basierte Ansätze wie TwitterMonitor [27] definieren Trends ausschließlich über Burstiness, ohne semantische Validierung. Diese rein statistische Definition von Relevanz führt zu einer nachgewiesenen Anfälligkeit gegenüber koordinierter Bot-Aktivität und strukturellem Rauschen. Das System delegiert die semantische Einordnung an Nutzer und kann nicht zwischen viral verbreiteten, aber inhaltlich irrelevanten Phänomenen und substanziellen Entwicklungen unterscheiden. Für die Früherkennung ist diese Einschränkung kritisch, da die anfängliche Phase emergenter Phänomene häufig durch moderate Aktivität in spezialisierten Communities gekennzeichnet ist, die durch reine Velocity nicht erfasst wird.

Topic-Model-basierte Verfahren adressieren die semantische Blindheit statistischer Ansätze, indem sie latente thematische Strukturen modellieren. Wang und Goutte [28] kombinierten Online Topic Modeling mit Bayes'scher Change Point Detection, um semantische Verschiebungen in Dokumentströmen zu identifizieren. Dieser Ansatz modelliert Texte kontinuierlich als Verteilungen über latente Themen und erkennt Zeitpunkte, zu denen sich die Wortzusammensetzung dieser Themen signifikant verändert. Die methodische Stärke liegt in der Fähigkeit, thematische Drifts jenseits reiner Frequenzänderungen zu erfassen. Die Limitation für die Identifikation spezifischer Trendbegriffe ist jedoch strukturell: In Latent Dirichlet Allocation werden Topics als Multinomialverteilungen über Vokabulare repräsentiert [29], wobei explizit festgestellt wird, dass keine epistemologischen Ansprüche bezüglich der Interpretierbarkeit dieser latenten Variablen erhoben werden. Die Bag-of-Words-Annahme fragmentiert zudem zusammengesetzte Begriffe wie „Stable Diffusion" in einzelne Tokens. Dadurch wird die kohärente Extraktion spezifischer Entitäten verhindert. Für ein System zur Früherkennung benennbarer Trends ist diese Repräsentationsform fundamental ungeeignet, da Nutzer konkrete Begriffe wie „GPT-4" erwarten und keine abstrakten Wahrscheinlichkeitsverteilungen über Token-Mengen.

Mredula et al. [30] identifizieren in ihrer Metaanalyse drei persistierende Lücken in der Forschung zur automatischen Event Detection.

Erstens konzentrieren sich bestehende Ansätze auf diskrete, abrupte Ereignisse wie Erdbeben oder Proteste und nicht auf graduell emergente Phänomene. Zweitens erfordern selbst flexible neuronale Architekturen domänenspezifische Trainingsdaten und generalisieren nicht auf beliebige Ereignistypen. Drittens fehlen Verfahren, die plattformübergreifend und mehrsprachig operieren. Die unüberwachte, domänenagnostische Identifikation emergenter Diskursphänomene bleibt eine ungelöste Herausforderung.

3.3 Synthese und Positionierung

Die systematische Analyse der Forschungslandschaft zeigt eine ungelöste Anforderungstrias: Statistische Ansätze liefern interpretierbare, reproduzierbare Ergebnisse, sind jedoch semantisch blind gegenüber dem Unterschied zwischen strukturellem Rauschen und inhaltlich relevanten Phänomenen. Ansätze der Topic Models [28] detektieren thematische Verschiebungen, repräsentieren Trends jedoch als abstrakte Wahrscheinlichkeitsverteilungen, nicht als benennbare Entitäten. Verfahren des Supervised Learning [23], [26] erreichen domänenspezifisch State-of-the-Art-Performance, sind jedoch konzeptionell unvereinbar mit der Identifikation neuartiger Phänomene, da für diese per Definition keine annotierten Trainingsdaten existieren können. Die unüberwachte Identifikation spezifischer, benennbarer Trendbegriffe mit semantischer Relevanzvalidierung in domänenagnostischen Systemen ist durch bestehende Einzelparadigmen nicht realisierbar. Diese Lücke wird durch eine dreistufige Filterkaskade adressiert, die statistische Robustheit und semantische Intelligenz vereint: Phase 1 extrahiert durch kontrastive Korpusanalyse [3], [11] und das Unithood/Termhood-Framework [4] linguistisch stabile und domänenspezifische Termkandidaten. Phase 2 validiert deren temporale Dimension durch eine an TDT angepasste Zeitreihenanalyse [2], [5]. Phase 3 bewertet durch ICL [6] die semantische Relevanz und transformiert Termkandidaten in ihre kanonische Form. Die statistische Vorqualifikation reduziert die LLM-Inferenz auf eine handhabbare Kandidatenmenge, während ICL domänenagnostische Semantik ohne aufgabenspezifisches Training bereitstellt. Die methodische Innovation zeigt sich in der funktionalen Dekomposition der Filterkaskade. Statistische Verfahren kommen dort zum Einsatz, wo sie am besten funktionieren: bei der Massenverarbeitung von Rohkorpora und robusten Signifikanztests. Semantische Modelle werden dagegen ausschließlich bei vorqualifizierten Kandidaten mit ausreichendem Kontext eingesetzt. Die inverse Gewichtung von Subreddits operationalisiert Rogers' Diffusionstheorie [8], indem kleine, spezialisierte Communities als Frühindikatoren priorisiert werden. Die Zeitreihenanalyse adaptiert das TDT-Prinzip der Stationaritätsverletzung [2], wobei die Kombination aus Signifikanz (Fisher's Test) und Konsistenz (Mann-Kendall) transiente Spikes von nachhaltigen Trends differenziert. Die operative Umsetzung dieser Prinzipien wird in Kapitel 4 ausführlich erläutert.

4 Methodik

4.1 Überblick: Dreistufige Filterkaskade

Um aufkommende Trends in sozialen Medien zu identifizieren, müssen massive Datenmengen gefiltert und systematisch von Rauschen bereinigt werden. Aus den Millionen täglicher Beiträge müssen diejenigen statistisch signifikanten und semantisch relevanten Diskussionsmuster extrahiert werden, die tatsächlich auf neue Phänomene hinweisen. Eine naive Frequenzanalyse würde vorwiegend generische Alltagssprache und plattformspezifisches Rauschen extrahieren. Eine ausschließlich auf LLM basierende Analyse wäre ressourcenintensiv und würde aufgrund fehlender Kontextualisierung zu inkonsistenten Ergebnissen führen.

Diese Herausforderung wird durch die vorliegende Implementierung mit einer dreistufigen Filterkaskade adressiert, die statistische Robustheit mit semantischer Intelligenz kombiniert. Unter Anwendung komplementärer Filterkriterien wird die Kandidatenmenge bei jeder Stufe progressiv reduziert (Abbildung 1).

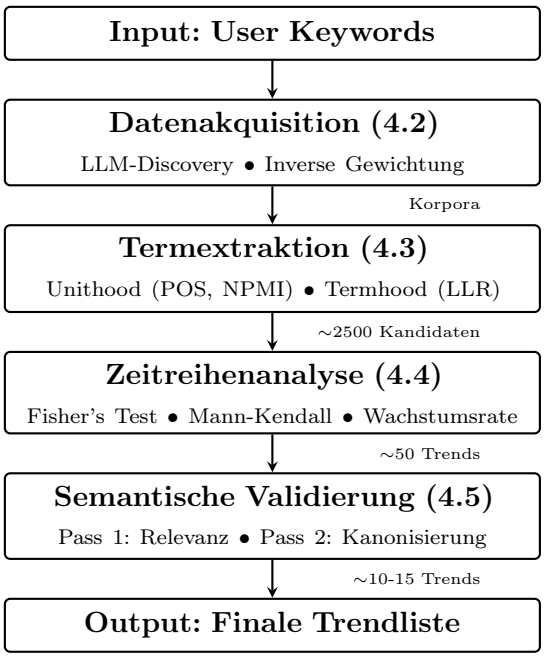

Abbildung 1: Architektur der Filterkaskade

4.2 Datenakquisition und Korpuskonstruktion

Die unüberwachte Termextraktion (Abschnitt 4.3) operationalisiert Luhns Konzept der Technese [1], dem zufolge Fachsprache in spezialisierten Diskursräumen konzentriert ist und dort statistisch überrepräsentiert auftritt.

Die Identifikation dieser domänenspezifischen Terminologie erfordert eine kontrastive Analyse zweier Korpora mit komplementären Funktionen [3], [11].

Der Zielkorpus (Abschnitt 4.2.1) repräsentiert die spezialisierte Diskursdomäne, in der aufkommende Trends als sprachliche Muster sichtbar werden. Der Referenzkorpus (Abschnitt 4.2.2) bildet mit seiner allgemeinsprachlichen Frequenzverteilung die Nullhypothese, gegen die statistisch auffällige Terme des Zielkorpus als domänenspezifische Technese detektiert werden.

4.2.1 Konstruktion des Zielkorpus (Reddit)

Die Konstruktion des Reddit Zielkorpus umfasst fünf sequenzielle Schritte zur Erstellung eines personalisierten, domänenspezifischen Korpus mit ausreichender Aktivität.

Schritt 1: Identifikation relevanter Communities Ein LLM (GPT-4o) identifiziert zunächst relevante Communities, indem es die abstrakten Keywords des Nutzers in konkrete Subreddits übersetzt. Dieser Ansatz basiert auf der von Brown et al. [6] gewonnenen Erkenntnis, dass große Sprachmodelle durch Few-Shot Learning implizites Weltwissen über semantische Zusammenhänge und plattformspezifische Strukturen erworben haben, ohne explizit für diese Aufgabe trainiert worden zu sein. Das Modell erhält einen strukturierten Prompt, der für jedes Keyword relevante Subreddits generiert. Die Anforderungen umfassen thematische Relevanz zum Keyword, einen Fokus auf wissensbasierte Diskussionen, nachweisbare Aktivität und ausschließlich englischsprachige Inhalte. Die Anzahl der identifizierten Communities variiert je nach Keyword und Breite der zugehörigen Diskussionslandschaft auf Reddit.

Die Wahl eines großen, proprietären Modells gegenüber kleineren, lokal ausführbaren Alternativen wird durch Forschungsfrage 1 motiviert und in der Evaluation (Abschnitt 6.2.2) systematisch untersucht. Um die Varianz der Ausgaben zu minimieren und somit die Reproduzierbarkeit der Korpuskonstruktion zu erhöhen, wird der Temperaturparameter des Modells auf null gesetzt. Das System erfordert weder manuelle Kuration noch domänenspezifisches Training und ist dadurch auf beliebige Interessensgebiete übertragbar. Das konzeptionelle Fundament dieser Methode lässt sich auf Luhns Konzept statistisch überrepräsentierter Fachsprache in spezialisierten Diskursräumen [1] zurückführen. Die identifizierten Communities sind spezialisierte Räume, in denen die für die Früherkennung von Trends entscheidende domänenspezifische Fachsprache gesprochen wird.

Schritt 2: Validierung und Gewichtung Die vom LLM identifizierten Subreddits durchlaufen mithilfe der Reddit API eine technische Validierung. Dabei wird sichergestellt, dass die Communities existieren und öffentlich zugänglich sind. Zudem wird für jeden Subreddit die Anzahl der Abonnenten ermittelt. Diese dient als Maß für die Reichweite der jeweiligen Community.

Auf Basis der Abonnentenzahlen wird eine inverse logarithmische Gewichtung berechnet, die kleineren Communities ein höheres Gewicht zuweist und damit die zentrale Hypothese von Forschungsfrage 2 operationalisiert. Demnach sind kleine, spezialisierte Communities stärkere Frühindikatoren für aufkommende Trends als große, heterogene Communities. Rogers [8] zeigt, dass eine kleine Gruppe von Innovatoren – 2,5 Prozent der Gesamtbevölkerung – Innovationen in spezialisierten Kontexten initiiert, bevor sie sich horizontal über Peer-Netzwerke verbreiten, und damit einen überproportionalen Einfluss auf die frühe Verbreitung neuer Ideen ausübt. Konzeptionell konvergiert dieser Befund mit Luhns Technese [1]: Fachsprache konzentriert sich in kleinen, spezialisierten Communities, die als primäre Quelle terminologischer Frühindikatoren fungieren.

Die logarithmische Transformation der Gewichte berücksichtigt die empirisch belegte Potenzgesetzverteilung der Mitgliederzahlen in sozialen Medien [31]: Die exponentiellen Größenunterschiede zwischen den kleinsten und größten Communities werden linearisiert, wodurch numerisch instabile Gewichte verhindert werden. Die anschließende Normalisierung auf das Intervall [1, 10] stellt sicher, dass einzelne extrem kleine Communities das Gesamtergebnis nicht dominieren. Die Effektivität dieser Gewichtung wird in Abschnitt 6.2.3 durch den Vergleich mit einer ungewichteten Baseline validiert.

Schritt 3: Datenabruf Die validierten und gewichteten Subreddits bilden die Grundlage für die Extraktion der Postdaten über den definierten Beobachtungszeitraum hinweg. Als Datenquelle dient die Arctic Shift API[32]. Im Gegensatz zur offiziellen Reddit API, die den Zugriff auf historische Daten auf die jüngsten 1000 Beiträge pro Subreddit beschränkt, ermöglicht sie den vollständigen strukturierten Zugriff auf beliebige historische Zeiträume. Das Zeitfenster ist dabei auf 60 Tage festgelegt. Diese Parametrisierung ist kein willkürlicher Wert, sondern das Ergebnis einer systematischen Evaluation. In Forschungsfrage 3 (Abschnitt 6.2.4) wird er im Hinblick auf die Abwägung zwischen Präzision, Sensitivität und Frühzeitigkeit untersucht.

Mithilfe sequenzieller API-Abfragen werden für jeden validierten Subreddit alle Beiträge aus dem definierten Zeitfenster extrahiert. Damit die Datenerhebung stabil bleibt und die Dateninfrastruktur nicht überlastet wird, ist zwischen aufeinanderfolgenden Anfragen eine konservative Verzögerung von 0,5 Sekunden eingebaut. Für jeden Beitrag werden die für die nachfolgende Analyse relevanten Metadaten erfasst. Hierzu zählen Titel, Engagementscore, Anzahl der Kommentare, Autor und Zeitstempel der Publikation.

Schritt 4: Event-basierte Aggregation Die extrahierten Posts durchlaufen eine Event-basierte Deduplikation, um redundante Diskussionen zu eliminieren. Identische Beiträge, die in mehreren Subreddits veröffentlicht wurden (Cross-Posts), werden als Instanzen desselben diskursiven Ereignisses behandelt und zu einem einzelnen Datensatz zusammengeführt.

Diese Vorgehensweise implementiert das TDT-Prinzip *same event, multiple reports* [2], das zwischen der semantischen Identität eines Ereignisses und seinen multiplen textuellen Repräsentationen unterscheidet.

Die Konsolidierung erfolgt durch Normalisierung der Posttitel (Lowercase-Transformation, Whitespace-Elimination) und anschließende Gruppierung. Posts mit identischem normalisiertem Titel werden aggregiert. Dabei werden ihre Engagementscores (Upvotes, Kommentare) summiert, ihre Subreddit-Zugehörigkeiten als Liste erfasst und der früheste Zeitstempel als kanonischer Publikationszeitpunkt festgelegt. Die Liste der Autoren wird dedupliziert, um die Diversität der an der Diskussion beteiligten Akteure zu erfassen. Durch diese Aggregation werden multiple textuelle Repräsentationen desselben Events in einen einzelnen, mit akkumulierten Engagementsignalen angereicherten Datensatz transformiert.

Die nachfolgende Analyse konzentriert sich ausschließlich auf die Posttitel, während Kommentarinhalte nicht extrahiert werden. Diese methodische Fokussierung ist durch drei Faktoren begründet: Erstens muss der semantische Kern einer Diskussion bereits durch den Titel kommuniziert werden, um auf Reddit Sichtbarkeit zu erlangen. Zweitens reduziert die Beschränkung auf Titel den Datenumfang, da jeder Post in der Regel mehrere Kommentare generiert, was wiederum die Verarbeitungseffizienz der statistischen Termextraktion erhöht. Drittens verhindert die Exklusion von Kommentaren das Problem der Diskursdrift, bei der die Diskussion vom ursprünglichen Thema abweicht und somit Noise in die Termextraktion einbringt.

Schritt 5: Aktivitätsbasierte Validierung Die aggregierten Beiträge werden einer aktivitätsbasierten Validierung unterzogen. Dies stellt sicher, dass ausschließlich Subreddits mit anhaltender Diskussionsaktivität in die Analyse einfließen. Mensah et al. [9] demonstrierten, dass sustained activity das entscheidende Qualitätskriterium für informative Communities darstellt. Diese Filterung erfolgt nach der Event-basierten Aggregation, um die Aktivität auf Basis der konsolidierten Diskussionen zu bewerten.

Zur Gewährleistung der statistischen Robustheit der nachfolgenden Zeitreihenanalyse muss jeder Subreddit an mindestens 40 Prozent der Tage im Beobachtungszeitraum Beiträge aufweisen, was bei einem Beobachtungszeitraum von 60 Tagen mindestens 24 aktiven Tagen entspricht. Der Schwellwert von 40 Prozent wird als pragmatische Heuristik ohne formale Optimierung festgelegt: Er vermittelt zwischen zwei konkurrierenden Anforderungen. Ein höherer Schwellwert würde hochspezialisierte Communities mit intermittierender, aber thematisch fokussierter Diskussion ausschließen, die für die Früherkennung von Trends besonders wertvoll sind. Ein niedrigerer Schwellwert würde Communities mit nur gelegentlichen Diskussionsschüben einschließen, deren sporadische Aktivität die statistische Robustheit der Zeitreihenanalyse beeinträchtigt.

Communities unterhalb dieses Schwellwerts werden vom Korpus ausgeschlossen. Die proportionale Skalierung mit dem Beobachtungszeitraum gewährleistet, dass die relative Aktivitätsanforderung unabhängig von der in Forschungsfrage 3 evaluierten Parametrisierung des Zeitfensters konstant bleibt.

4.2.2 Konstruktion des Referenzkorpus (Wikipedia)

Die kontrastive Termextraktion (Abschnitt 4.3.3) erfordert ein allgemeines Referenzkorpus zur Identifikation domänenspezifischer Sprache. Als Referenzkorpus dient eine Wikipedia-N-Gramm-Frequenzdatenbank, die durch ihre thematische Breite die allgemeinsprachliche Wortverteilung abbildet.

Korpusauswahl und Begründung Wikipedia wird als Referenzkorpus gewählt, da es drei zentrale Anforderungen erfüllt: Erstens bietet es eine umfassende thematische Abdeckung über diverse Wissensdomänen hinweg, was die Identifikation domänenspezifischer Sprache erst ermöglicht. Zweitens unterliegen die Inhalte redaktioneller Kuration. Dadurch wird eine höhere sprachliche Qualität im Vergleich zu nutzergenerierten Inhalten gewährleistet. Drittens ermöglicht die öffentliche Verfügbarkeit als standardisierte Ressource die vollständige Reproduzierbarkeit der Methodik. Die kontrastive Analyse eines spezialisierten Korpus gegen ein allgemeines Referenzkorpus zur Identifikation domänenspezifischer Terme ist methodisch etabliert [3], [11].

N-Gramm Extraktion und Preprocessing Die Wikipedia Daten werden aus der öffentlich verfügbaren DoltHub-Datenbank "wikipedia ngrams" [33] bezogen, die vorverarbeitete Unigramm- und Bigramm-Frequenzen über den gesamten englischsprachigen Wikipedia-Korpus bereitstellt. Diese Datenbank aggregiert die Häufigkeiten aller Unigramme und Bigramme aus Wikipedia Artikeln und eliminiert damit die Notwendigkeit einer eigenen, rechenintensiven Korpusverarbeitung. Die Rohdaten umfassen mehrere Millionen Unigramme und Bigramme mit ihren absoluten Frequenzen.

Skalierbare Lookup-Infrastruktur Für die effiziente Frequenzabfrage während der Termextraktion werden die N-Gramm-Daten in sortierte binäre Dateien transformiert. Jeder Term wird mittels xxHash (64 Bit) gehasht und gemeinsam mit seiner Frequenz in einem memory-mapped NumPy-Array gespeichert, wodurch Lookups in $O(\log n)$ mittels binärer Suche möglich sind und der RAM-Bedarf auf 43 MB reduziert wird. Ein Least Recently Used (LRU)-Cache mit 20.000 Einträgen beschleunigt wiederholte Abfragen häufig verwendeter Terme. Die technische Spezifikation ist in Anhang A.3 dokumentiert.

4.3 Phase 1: Unüberwachte Termextraktion

Die unüberwachte Termextraktion operationalisiert Luhns Konzept der Technese [1]. Demnach konzentriert sich Fachsprache in spezialisierten Diskursräumen und tritt dort statistisch überrepräsentiert auf. Die Identifikation dieser domänenspezifischen Terminologie erfolgt durch Anwendung des „Unithood/Termhood"-Frameworks [4]. Dieses definiert linguistische Stabilität (Unithood) und domänenspezifische Relevanz (Termhood) als komplementäre Qualitätskriterien für Termkandidaten. Die Implementierung gliedert sich in vier sequenzielle Stufen: Zunächst wird der Korpus durch Textbereinigung und frequenzbasierte Vorfilterung für die statistische Analyse vorbereitet (Abschnitt 4.3.1). Anschließend werden linguistisch stabile N-Gramme mithilfe syntaktischer und kohäsionsbasierter Filter identifiziert. (Abschnitt 4.3.2). Die verbleibenden Kandidaten werden mittels kontrastiver Frequenzanalyse gegen den Referenzkorpus auf domänenspezifische Relevanz geprüft (Abschnitt 4.3.3). Anschließend stellt eine finale Deduplikation redundanter Terme die sortierte Kandidatenliste für die Zeitreihenanalyse (Phase 2) bereit (Abschnitt 4.3.4). Bei diesem mehrstufigen Ansatz wird Präzision über Vollständigkeit priorisiert [7], da die verlässliche Identifikation weniger, aber qualitativ hochwertiger Trends für die Früherkennung kritischer ist als eine vollständige, aber verrauschte Kandidatenliste.

4.3.1 Preprocessing und N-Gramm-Extraktion

Die Vorbereitung des Rohkorpus für die statistische Analyse erfolgt in drei sequenziellen Schritten: Zunächst wird eine Textbereinigung durchgeführt, anschließend erfolgt eine frequenzbasierte Vorfilterung und schließlich eine POS-basierte N-Gramm-Generierung.

Textbereinigung Die Posttitel durchlaufen eine regelbasierte Normalisierung, um plattformspezifisches Rauschen zu eliminieren. Mithilfe regulärer Ausdrücke werden URLs, Subreddit- und Nutzerlinks (z.B. `r/MachineLearning`, `u/username`), Markdown sowie Reddit-spezifische Tags (`[OC]`, `[NSFW]`) entfernt. Um die nachfolgende linguistische Analyse zu standardisieren, werden die resultierenden Texte in Kleinschreibung transformiert und auf einzelne Leerzeichen zwischen Tokens normalisiert.

Frequenzbasierte Vorfilterung Frequenzbasierte Schwellwerte eliminieren statistisch unzuverlässige Einzelvorkommen, ohne potenzielle Frühindikatoren zu verlieren. Unigramme müssen mindestens 15-mal, Bigramme mindestens 5-mal im Gesamtkorpus auftreten. Zusätzlich muss jeder Termkandidat in mindestens 3 unterschiedlichen Beiträgen vorkommen, um eine ausreichende Streuung in den Diskussionen zu gewährleisten. Terme, die in mehr als 30 Prozent aller Dokumente auftreten, werden als allgemeinsprachlich klassifiziert und eliminiert. Die exakten Parameterwerte sind in Tabelle 1 dokumentiert.

Tabelle 1: Frequenzschwellen der Vorfilterung

Unigramme	**Bigramme**	**Min. Posts**	**Max. Anteil**
≥ 15	≥ 5	≥ 3	≤ 30 Prozent

Die absoluten Frequenzschwellen sind an die Korpusgröße angepasst: Etablierte Verfahren verwenden minimale Frequenzen von 2 bis 3 Vorkommen für kleine Fachkorpora [11], [19]; die höheren Werte von 15 für Unigramme und 5 für Bigramme spiegeln die größere Datenmenge wider [4]. Die Schwellen für die Dokumentfrequenz folgen der von Ahmad et al. [11] etablierten Heuristik, dass domänenspezifische Begriffe weder zu selten (weniger als 3 Beiträge) noch zu ubiquitär (mehr als 30 Prozent der Beiträge) auftreten.

POS-Tagging und N-Gramm-Extraktion Die Extraktion linguistisch plausibler N-Gramme erfolgt durch eine spaCy-Pipeline (`en_core_web_sm`), die Tokenisierung, POS-Tagging und Lemmatisierung umfasst. Da die syntaktische Stabilitätsprüfung ausschließlich auf POS-Sequenzen basiert, werden komplexere Annotationen wie Dependency Parsing und Named Entity Recognition deaktiviert. Für jedes Token werden neben der Oberflächenform auch die POS-Tags erfasst, die zur Validierung syntaktischer Muster dienen (Abschnitt 4.3.2). Für jeden Termkandidaten werden die absolute Termfrequenz über den Gesamtkorpus sowie die Dokumentfrequenz als Maß für die Streuung über die Beiträge erfasst. Für Bigramme werden zusätzlich die POS-Sequenzen gespeichert, um sie gegen linguistische Muster zu validieren (Abschnitt 4.3.2).

4.3.2 Unithood: Linguistische Stabilität

Kageura und Umino [4] verstehen Unithood als den Grad, zu dem eine Wortsequenz als eigenständige lexikalische Einheit wahrgenommen wird. Die Operationalisierung erfolgt durch zwei komplementäre Filter: Zunächst validiert eine syntaktische Prüfung die grammatikalische Plausibilität von Bigrammen anhand etablierter POS-Muster (Abschnitt 4.3.2), anschließend quantifiziert eine statistische Kohäsionsmessung die Assoziationsstärke zwischen den Wortkomponenten (Abschnitt 4.3.2). Nur Kandidaten, die beide Filter passieren, werden in Abschnitt 4.3.3 auf domänenspezifische Relevanz geprüft. Unigramme durchlaufen diese zweistufige Filterung nicht, da sie per Definition atomare lexikalische Einheiten darstellen.

Syntaktische Filterung Justeson und Katz [19] demonstrierten, dass 97 Prozent der Fachterme in technischen und wissenschaftlichen Korpora ausschließlich aus Nomen und Adjektiven bestehen. Diese strukturelle Regularität wird durch einen POS-basierten Filter operationalisiert, der Bigramme gegen eine Menge zulässiger Sequenzen validiert.

Tabelle 2: Zugelassene Abfolgen von Wortarten für Bigrammkandidaten

POS-Muster	**Beispiel**	**Begründung**
(JJ, NN/NNS)	neural network	J&K 1995
(NN, NN/NNS)	machine learning	J&K 1995
(NNP, NNP)	Claude Sonnet	Eigennamen-Terme
(VBG, NN/NNS)	training data	Gerundien als Modifier
(VBN, NN/NNS)	trained model	Partizipien als Modifier
(NN, VBG)	data mining	Gerundien als Head
(CD, NN/NNS)	gpt-4 model	Versionsnummern

Die Erweiterung um Gerundien (VBG), Partizipien (VBN), Kardinale (CD) und Eigennamen (NNP) spiegelt die terminologischen Charakteristika des Korpus wider, in dem versionierte Technologien (z. B. „GPT-4"), prozessbasierte Terme (z. B. „Machine Learning") und Produktnamen (z. B. „Claude Sonnet") als Trends auftreten. Kageura und Umino [4] dokumentieren, dass POS-basierte Termfilter aufgabenspezifisch kalibriert werden müssen, um domänencharakteristische Terminologie adäquat zu erfassen – eine Anforderung, der die vorliegende Erweiterung über Justeson und Katz [19] hinaus Rechnung trägt. Die Beschränkung auf diese syntaktischen Muster eliminiert linguistisch implausible Kandidaten a priori und operationalisiert die von Gale und Church [7] etablierte Priorisierung von Präzision über Vollständigkeit.

Statistische Kohäsionsmessung Die Kohäsionsbewertung von Bigrammkandidaten erfolgt durch NPMI [13], das klassische PMI [12] durch Normalisierung auf das Intervall $[-1, 1]$ erweiternd:

$$\mathrm{NPMI}(w_1, w_2) = \frac{\log \frac{p(w_1,w_2)}{p(w_1)\cdot p(w_2)}}{-\log p(w_1, w_2)}$$

Ein Bigramm gilt als kohäsiv, wenn $\mathrm{NPMI}(w_1, w_2) \geq 0{,}4$. Dieser Schwellwert balanciert die Identifikation stabiler Kollokationen gegen eine zu restriktive Filterung; die Sensitivität gegenüber diesem Parameter wird in Abschnitt 6.2.5 systematisch untersucht.

4.3.3 Termhood: Domänenrelevanz durch kontrastive Analyse

Kageura und Umino [4] definieren Termhood als den Grad, in dem eine lexikalische Einheit für eine bestimmte Domäne charakteristisch ist. Die Operationalisierung erfolgt durch zwei komplementäre statistische Größen, die jeweils unterschiedliche Aspekte der Domänenspezifität erfassen.

Log-Likelihood Ratio als Primärmaß Das Log-Likelihood Ratio (LLR) quantifiziert die statistische Signifikanz der Frequenzdifferenz zwischen Ziel- und Referenzkorpus durch einen χ^2-basierten Likelihood-Quotiententest [3].

Dunning [3] zeigt, dass LLR im Gegensatz zu klassischen χ^2-Tests auch bei niedrigen Erwartungswerten robuste Ergebnisse liefert, was es für die Identifikation selten auftretender Terme essenziell macht. Die Berechnung basiert auf einer 2×2-Kontingenztabelle, in der die Frequenzen des Terms und aller anderen Terme in beiden Korpora erfasst sind.

Tabelle 3: Kontingenztabelle zum Log-Likelihood Ratio

	Term t	**Andere Terme**
Reddit (Zielkorpus)	a	c
Wikipedia (Referenzkorpus)	b	d

Die G^2-Statistik wird als doppelte logarithmische Likelihood-Ratio berechnet:

$$G^2 = 2 \sum_i O_i \cdot \ln \left(\frac{O_i}{E_i} \right)$$

wobei O_i die beobachteten und E_i die unter Unabhängigkeitsannahme erwarteten Frequenzen der Tabellenzellen darstellen. Ein Term gilt als statistisch signifikant domänenspezifisch, wenn $G^2 \geq 10{,}83$ entsprechend dem 99,9. Perzentil der χ^2-Verteilung bei einem Freiheitsgrad ($p < 0{,}001$). Dieser konservative Schwellwert gewährleistet eine hohe statistische Konfidenz und reduziert Fehldetektionen durch zufällige Frequenzschwankungen.

Weirdness Ratio als Sekundärfilter Ahmad et al. [11] definieren den Weirdness Index als Verhältnis der relativen Frequenzen zwischen Fach- und Referenzkorpus, der ergänzend zum LLR als Sekundärfilter eingesetzt wird:

$$\text{Weirdness}(t) = \frac{f_{\text{Reddit}}(t)/N_{\text{Reddit}}}{f_{\text{Wiki}}(t)/N_{\text{Wiki}}}$$

Während LLR die statistische Signifikanz der Frequenzdifferenz quantifiziert, operationalisiert Weirdness das Ausmaß dieser Differenz als interpretierbare Ratio: Ein Wert von 100 bedeutet, dass der Term im Zielkorpus relativ 100-mal häufiger auftritt als im Referenzkorpus. Ahmad et al. [11] etablierten diesen Schwellwert empirisch für die kontrastive Korpusanalyse des TREC-8 Korpus gegenüber dem British National Corpus. Die Übertragbarkeit auf den vorliegenden Reddit-Wikipedia-Vergleich ist nicht formal validiert; der Schwellwert wird hier als konservative Orientierung übernommen.

Die Kombination beider Maße adressiert die dokumentierte Instabilität von Weirdness bei seltenen Termen [11]. Dank der vorgeschalteten LLR-Filterung werden nur Kandidaten mit bereits etablierter statistischer Signifikanz einer Weirdness-Prüfung unterzogen. Dies gewährleistet, dass nur Terme mit einer sowohl statistisch signifikanten als auch substanziellen Überrepräsentation in die finale Bewertung einfließen.

Multikriterien-Scoring und Normalisierung Die finale Bewertung erfolgt durch einen gewichteten Multikriterienscore, der die komplementären Stärken der verschiedenen statistischen Maße nutzt. Kageura und Umino [4] empfehlen mehrstufige Filterarchitekturen als konzeptionelle Grundlage; die konkrete Gewichtung ist eine eigenständige Designentscheidung [4]. Um eine vergleichbare Gewichtung zu ermöglichen, werden zunächst alle Maße auf das Intervall [0,1] normalisiert: Die LLR-Normalisierung erfolgt durch eine Sigmoidfunktion der Form $1 - \frac{1}{1+\text{LLR}/100}$, welche die unbeschränkte LLR-Verteilung progressiv auf [0,1) abbildet; Weirdness wird durch die logarithmische Transformation $1 - \frac{1}{1+\ln(\text{Weirdness})}$ gedämpft; IDF wird durch Division durch den theoretischen Maximalwert $\log(n_{\text{docs}})$ normalisiert.

Bei Unigrammen werden LLR (50 Prozent), Weirdness (30 Prozent) und IDF (20 Prozent) kombiniert. Die Gewichtung reflektiert die relative statistische Robustheit der Maße: LLR als χ^2-basiertes Verfahren erhält das höchste Gewicht, Weirdness wird aufgrund seiner Instabilität bei seltenen Termen konservativer gewichtet, und IDF ergänzt die Bewertung durch die Quantifizierung der Dokumentstreuung [14].

Die Gewichtung für Bigramme ist an die bereits erfolgte Unithoodfilterung angepasst: Da NPMI die interne Kohäsion bereits validiert hat, erfolgt die finale Bewertung durch LLR (40 Prozent), NPMI (40 Prozent) und IDF (20 Prozent). Die paritätische Gewichtung spiegelt wider, dass bei Mehrworttermen sowohl Domänenspezifität als auch linguistische Stabilität für die Termqualität gleichermaßen kritisch sind.

4.3.4 Deduplication und finale Selektion

Die Deduplikation eliminiert redundante Termkandidaten, die aufgrund unterschiedlicher N-Grammlängen entstehen: Durchlaufen sowohl ein Unigramm als auch ein Bigramm, das dieses Unigramm enthält, die Filterung, entstehen semantisch überlappende Kandidaten – so würden etwa „gpt", „gpt model" und „language model" als separate Terme extrahiert, obwohl sie dasselbe Konzept repräsentieren.

Die Deduplikationsstrategie identifiziert Überlappungen durch einen mengentheoretischen Vergleich der Tokenmenge jedes Kandidaten: Terme in einer Subset- oder Supersetbeziehung zueinander werden als redundant klassifiziert, wobei ausschließlich der höchstbewertete Term beibehalten wird. Die Deduplikation erfolgt in absteigender Reihenfolge der Scores, sodass qualitativ überlegene Terme Vorrang erhalten.

Nach der Deduplikation werden die verbleibenden Scores auf das Intervall [0, 100] normalisiert, wobei der höchste Score auf 100 skaliert wird. Kandidaten mit einem normalisierten Score unter 10 gelten als statistisch nicht ausreichend robust und werden eliminiert, sodass ausschließlich Terme mit substanzieller statistischer Evidenz in die Zeitreihenanalyse (Abschnitt 4.4) einfließen. Die nach Score absteigend sortierte Kandidatenliste bildet den Input für Phase 2.

4.4 Phase 2: Zeitreihenanalyse

Die unüberwachte Termextraktion (Abschnitt 4.3) liefert eine nach statistischer Domänenspezifität sortierte Kandidatenliste, die jedoch ausschließlich statische Frequenzmaße enthält und keine Rückschlüsse auf die zeitliche Entwicklung der Diskussion zulässt. Ein Term kann domänenspezifisch sein, ohne einen aufkommenden Trend zu repräsentieren – etablierte Fachbegriffe mit konstanter Diskussionsfrequenz sind das prototypische Gegenbeispiel. Die Zeitreihenanalyse adressiert diese Einschränkung, indem sie Terme identifiziert, die sich nicht nur durch Domänenspezifität, sondern auch durch signifikantes Wachstum und konsistente Aufwärtsentwicklung auszeichnen.

Die methodische Grundlage bildet das Prinzip der statistischen Abweichung von einer Baseline, wie es im TDT Framework etabliert wurde [2]. Bei der Implementierung werden drei komplementäre Metriken kombiniert: Eine Signifikanzprüfung identifiziert statistisch auffällige Aktivitätsanstiege, eine Konsistenzprüfung validiert die Robustheit der Aufwärtsentwicklung und eine Wachstumsrate quantifiziert das Ausmaß der Veränderung.

4.4.1 Statistische Signifikanzprüfung

Allan et al. [2] etablierten das Prinzip der „violation of stationarity" als Kernsignal für aufkommende Phänomene im TDT Framework und operationalisierten es durch χ^2-basierte Kontrastierung von Termfrequenzen über Zeitfenster. Die vorliegende Implementierung adaptiert dieses Konzept, ersetzt den χ^2-Test jedoch durch Fisher's Exact Test, der bei den für aufkommende Trends typischen niedrigen Erwartungswerten exakte p-Werte liefert und damit statistisch robuster ist.

Der Test basiert auf einer 2×2-Kontingenztabelle (Tabelle 4), wobei a und c die Frequenzen des Terms in den jeweiligen Zeitfenstern repräsentieren, während b und d die Frequenzen aller anderen Terme darstellen. Das recent window ($T_{\text{recent}} = 12$ Tage) wird gegen das past window ($T_{\text{past}} = 48$ Tage) kontrastiert, wobei beide Fenster gemeinsam den gesamten Beobachtungszeitraum von 60 Tagen abdecken. Die Parametrisierung dieser Zeitfenster wird im Rahmen der Evaluation (Abschnitt 6.2.4) untersucht. Ein p-Wert unterhalb des Signifikanzniveaus von $\alpha = 0{,}05$ indiziert, dass die Frequenzerhöhung im aktuellen Fenster statistisch signifikant ist und somit als Evidenz für einen aktuellen Burst gewertet wird. Fisher's Test operiert dabei auf den absoluten täglichen Engagementsummen, da der Test Häufigkeiten in einer Kontingenztabelle erfordert und normalisierte Werte die Zellenfrequenzen verfälschen würden. Die Anwendung von Fisher's Exact Test auf aggregierte Zeitfenster ist eine pragmatische Vereinfachung des im TDT Framework etablierten χ^2-basierten Ansatzes [2]. Die zeitliche Autokorrelation der Posts wird dabei nicht explizit modelliert, was bei der Interpretation der p-Werte berücksichtigt werden muss. Für die Identifikation statistisch auffälliger Aktivitätsanstiege ist diese Approximation jedoch ausreichend robust (vgl. Abschnitt 6.2.1).

Tabelle 4: Kontingenztabelle zum Fisher's Exact Test

	Term t	**Andere Terme**
Recent Window (12 Tage)	a	b
Past Window (48 Tage)	c	d

4.4.2 Konsistenzprüfung

Die Signifikanzprüfung (Abschnitt 4.4.1) identifiziert statistisch auffällige Aktivitätsanstiege, kann jedoch nicht zwischen transienten Spikes und nachhaltigen Aufwärtstrends unterscheiden: Ein Term kann einen signifikanten Burst aufweisen, ohne eine konsistente Wachstumsdynamik zu zeigen – wie etwa bei einmaligen Ereignissen, die kurzfristige Diskussionsspitzen auslösen, jedoch keine anhaltende Relevanz entwickeln. Kleinberg [5] etablierte die konzeptionelle Grundlage für diese Unterscheidung; die Konsistenzprüfung operationalisiert sie, indem sie die Robustheit der Aufwärtsentwicklung über den gesamten Beobachtungszeitraum ($T_{\text{lookback}} = 60$ Tage) quantifiziert. Der Mann-Kendall Test operiert hingegen auf den normalisierten Engagementwerten, die den täglichen Anteil des Terms am Gesamtkorpus abbilden. Diese Normalisierung isoliert genuines terminologisches Wachstum von allgemeinen Aktivitätsschwankungen des Korpus.

Die Implementierung verwendet den Mann-Kendall Test [16], [17], ein etabliertes nichtparametrisches Verfahren zur Erkennung monotoner Trends in Zeitreihen. Der Test berechnet Kendall's Tau (τ) als Korrelationsmaß zwischen den Zeitpunkten und den beobachteten Werten:

$$\tau = \frac{S}{\frac{n(n-1)}{2}}$$

wobei $S = \sum_{i<j} \operatorname{sgn}(x_j - x_i)$ die Summe der Vorzeichen aller Paardifferenzen ist und $\frac{n(n-1)}{2}$ die Gesamtanzahl möglicher Paare darstellt. Ein positiver Wert ($\tau > 0$) indiziert einen monoton steigenden Trend, während ein negativer Wert ($\tau < 0$) einen fallenden Trend anzeigt. Die Robustheit des Tests gegenüber Ausreißern sowie das Fehlen von Verteilungsannahmen machen ihn besonders geeignet für die Analyse von Diskussionsdaten aus sozialen Medien.

Für die Einstufung als aufkommender Trend wird $\tau \geq 0{,}0$ vorausgesetzt, wodurch ausschließlich Terme mit aufsteigender oder stabiler Diskussionsdynamik berücksichtigt werden. Der Consistencyscore wird als $\text{score}_{\text{consistency}} = \max(0, \tau) \times 100$ normalisiert, wobei die Maximumsfunktion negative Werte auf null abbildet.

4.4.3 Multifaktorielles Scoring

Da die drei statistischen Metriken unterschiedliche Aspekte der Trenddynamik auf inkompatiblen Skalen erfassen, erfolgt zunächst eine Normalisierung auf das Intervall $[0, 100]$, gefolgt von einer gewichteten Summierung zu einem einheitlichen Statisticalscore.

Die Wachstumsrate wird als logarithmisches Verhältnis zwischen dem recent window und dem past window berechnet:

$$r_{\text{growth}} = \log_{10}\left(\frac{\bar{f}_{T_{\text{recent}}} + 1}{\bar{f}_{T_{\text{past}}} + 1}\right)$$

wobei $\bar{f}_{T_{\text{recent}}}$ und $\bar{f}_{T_{\text{past}}}$ die mittlere tägliche Frequenz des Terms im jeweiligen Zeitfenster darstellen. Die Verwendung von Mittelwerten anstelle von Summen gewährleistet die Vergleichbarkeit bei variierenden Fensterlängen, insbesondere im Rahmen der Parameterevaluation in Abschnitt 6.2.4. Die logarithmische Transformation stabilisiert die Verteilung bei stark variierenden Frequenzen. Da das System per Definition aufkommende Trends identifiziert, werden nur Terme mit nicht negativem Wachstum berücksichtigt ($r_{\text{growth}} \geq 0$).

Die Normalisierung erfolgt metrikspezifisch: Die Wachstumsrate wird linear zum Maximum der Kandidaten skaliert ($s_{\text{growth}} = r_{\text{growth}}/r_{\max} \times 100$), der Consistencyscore direkt aus Kendall's Tau abgeleitet ($s_{\text{consistency}} = \max(0, \tau) \times 100$) und der Significancescore invers zum p-Wert transformiert ($s_{\text{significance}} = (1 - p) \times 100$). Die finale Bewertung erfolgt durch gewichtete Summation:

$$S_{\text{statistical}} = w_{\text{growth}} \cdot s_{\text{growth}} + w_{\text{consistency}} \cdot s_{\text{consistency}} + w_{\text{significance}} \cdot s_{\text{significance}}$$

wobei die Gewichte $w_{\text{growth}} = 0{,}45$, $w_{\text{consistency}} = 0{,}25$ und $w_{\text{significance}} = 0{,}30$ theoretisch motiviert und durch Vergleich mit Ablationsvarianten empirisch bestätigt werden (Abschnitt 6.2.6). Die höhere Gewichtung der Wachstumsrate spiegelt ihre zentrale Bedeutung für die Identifikation aufkommender Trends wider.

4.5 Phase 3: Semantische Validierung

Die statistische Zeitreihenanalyse (Abschnitt 4.4) identifiziert Terme mit signifikantem Wachstum, berücksichtigt jedoch nicht deren semantische Relevanz für die spezifischen Nutzerinteressen: Ein Term kann alle statistischen Kriterien erfüllen, ohne einen inhaltlichen Bezug zu den Keywords aufzuweisen. Die semantische Validierung adressiert diese Einschränkung durch LLM-basierte Relevanzprüfung und Termveredelung, wobei das von Brown et al. [6] etablierte Few-Shot-Lernparadigma operationalisiert wird.

Die Implementierung erfolgt als zweistufiger Prozess: Im ersten Schritt werden semantisch irrelevante Terme durch Relevanzscore Filterung eliminiert, im zweiten Schritt werden die validierten Terme in ihre kanonische Form transformiert. Diese funktionale Dekomposition dient der Effizienz – Kanonisierung erfolgt ausschließlich auf der validierten Kandidatenmenge – und verhindert Bias durch Akronymbewertung, da Terme wie „DAO" ohne Kanonisierung anders bewertet würden als ihre Langform.

4.5.1 Kontextaggregation

Für die LLM-basierte Bewertung ist kontextuelle Evidenz zur Interpretation des domänenspezifischen Vokabulars erforderlich. Für jeden Termkandidaten werden die zehn Beiträge mit dem höchsten Engagementscore aus dem Zielkorpus aggregiert und als In-Context-Beispiele verwendet. Diese Beiträge disambiguieren Akronyme, Plattformslang und durch die N-Grammextraktion abgeschnittene Phrasen, die ohne Kontext nicht interpretierbar sind.

4.5.2 LLM Pass 1: Relevanzfilterung

Pass 1 berechnet für jeden statistisch qualifizierten Termkandidaten einen Relevanzscore im Intervall [0,0, 1,0]. Das LLM (GPT-4o-mini) erhält einen strukturierten Prompt mit Keywords, Termkandidat und Reddit-Kontext. Zur Kalibrierung werden Richtlinien bereitgestellt, die Relevanzbereiche als Orientierung definieren (0,9–1,0 für Kernkonzepte, 0,7–0,89 für Werkzeuge, 0,5–0,69 für Methoden, 0,2–0,49 für periphere Themen, 0,0–0,19 für Rauschen). Diese Grenzen wurden auf Basis von Pilottests an einer Stichprobe von 20 Termkandidaten aus dem Entwicklungsdatensatz festgelegt und dienen als Kalibrierungsanker, erzwingen jedoch keine starre Kategorisierung.

Die finale Bewertung erfolgt durch Multiplikation: $S_{\text{final}} = S_{\text{statistical}} \times r_{\text{relevance}}$. Diese multiplikative Fusion operationalisiert den Relevanzscore als Discount-Faktor: Ein statistisch starkes Signal wird durch niedrige semantische Relevanz gedämpft, während ein moderates Signal mit hoher Relevanz aufgewertet wird. Durch die Multiplikation müssen beide Kriterien gleichzeitig erfüllt sein. Terme mit einem Relevanzscore unterhalb von 0,2 werden als semantisch irrelevant klassifiziert und unabhängig vom Statistical Score eliminiert. Die finale Selektion der verbleibenden Kandidaten erfolgt durch einen adaptiven Cutoff: Liegt das 75. Perzentil der Final Scores mehr als zehn Punkte über dem Median, wird das 75. Perzentil als Schwellwert verwendet; andernfalls wird der Median verwendet. Diese Logik stellt sicher, dass bei flachen Scoreverteilungen ohne klaren Qualitätsgradienten eine ausreichend große Kandidatenmenge für die Kanonisierung verbleibt, während bei stark differenzierten Verteilungen nur die oberen Schichten weitergeleitet werden.

4.5.3 LLM Pass 2: Kanonisierung

In Pass 2 werden die validierten Terme in ihre kanonische Form transformiert. Reddit-Diskussionen verwenden häufig Akronyme („DAO"), Plattformslang oder durch die N-Grammextraktion abgeschnittene Phrasen („supply" statt „ethereum supply"). Das LLM erhält für jeden Term einen strukturierten Prompt mit Keywords, Term und Reddit-Kontext und soll die spezifischste und vollständigste Darstellung des Trends liefern: Generische Begriffe werden kontextspezifisch vervollständigt, Akronyme unter Beibehaltung der Kurzform aufgelöst (z. B. „Decentralized Autonomous Organization (DAO)"), und bereits eindeutige Terme bleiben unverändert.

Die kanonisierten Terme durchlaufen abschließend eine Deduplikation: Verschiedene statistische Rohformen wie „dao" oder „decentralized autonomous" können auf denselben kanonischen Term abgebildet werden, wobei ausschließlich die Variante mit dem höchsten Finalscore beibehalten wird. Die resultierende, nach Score absteigend sortierte Liste stellt die finale Ausgabe des Systems dar.

5 Implementierung

5.1 Systemarchitektur und technologische Grundlagen

Die Implementierung realisiert die in Kapitel 4 beschriebene dreistufige Filterkaskade durch eine modulare Pipeline. Jede Phase ist als eigenständiges Modul mit definierten Ein- und Ausgabeschnittstellen umgesetzt. Diese Architektur ermöglicht die isolierte Modifikation und Evaluierung einzelner Systemkomponenten, sodass sich Änderungen nicht durch nachgelagerte Phasen ausbreiten.

`data.py` konstruiert den Zielkorpus (Abschnitt 4.2) mittels LLM-gestützter Identifikation relevanter Subreddits, deren Validierung sowie gewichtetem Datenabruf über die Arctic Shift API. Diese Schnittstelle wird genutzt, da die offizielle Reddit API den Zugriff auf historische Daten auf die jüngsten 1.000 Beiträge pro Subreddit beschränkt. Die Ausgabe ist ein ereignisbasiert dedupliziertes Korpus im Parquet-Format. `extraction.py` implementiert die kontrastive Termextraktion (Abschnitt 4.3). Das Modul vergleicht den Zielkorpus gegen einen memory-mapped Wikipedia-Referenzkorpus und produziert eine nach Unithood/Termhood-Scores sortierte Kandidatenliste. `analysis.py` führt Zeitreihenvalidierung und semantische Filterung durch (Abschnitte 4.4 und 4.5). Eingabe sind Termkandidaten mit täglichen Engagement-Zeitreihen, Ausgabe ist die finale Trendliste. `main.py` orchestriert die Pipeline und verwaltet persistente Zwischenergebnisse.

Caching persistiert Verarbeitungsergebnisse auf drei Granularitätsebenen. `posts.parquet` speichert den Rohkorpus im spaltenorientierten Parquet-Format. Dieses Format unterstützt partielle Lesezugriffe und verlustfreie Kompression.

Die Datei `timelines.pkl` speichert die extrahierten Termkandidaten mit den vorberechneten Zeitreihen als serialisiertes Dictionary. `results.json` persistiert finale Analyseergebnisse zur direkten Visualisierung. Diese Architektur entkoppelt die Datenakquisition, statistische Extraktion und semantische Validierung voneinander.

Die Pipeline unterstützt drei Ausführungsmodi. Der `full`-Modus durchläuft alle Phasen und aktualisiert sämtliche Caches. Der `analysis_only`-Modus operiert auf gecachten Zeitreihen und ermöglicht Parametervariationen zeitreihenanalytischer und semantischer Komponenten, ohne dass eine Neuextraktion erforderlich ist. Der `load_cached`-Modus lädt persistierte Endergebnisse zur Visualisierung. Diese Modularisierung ist für die systematische Parameterevaluation (Kapitel 6) methodisch erforderlich.

Die Implementierungssprache ist Python 3.12. Mithilfe von NumPy und SciPy können statistische Primitive wie das Log-Likelihood Ratio, der Fisher's Exact Test und der Mann-Kendall Test durchgeführt werden. spaCy (`en_core_web_sm`) realisiert Part-of-Speech Tagging für Unithoodfilter. Polars verarbeitet tabellarische Daten mithilfe spaltenorientierter Ausführung und nativer Parquet-Integration. Die LLM-Integration erfolgt über OpenAI REST-API mit deterministischer Konfiguration (seed=42, temperature=0.0). Für die Identifikation relevanter Subreddits (Abschnitt 4.2.1) wird GPT-4o eingesetzt, für die semantische Validierung (Abschnitt 4.5) GPT-4o-mini. Die Modellwahl reflektiert die Abwägung zwischen Aufgabenkomplexität und Inferenzkosten und wird in Abschnitt 6.2.2 evaluiert.

5.2 Optimierungen für Skalierbarkeit

5.2.1 Binäre Lookupstruktur für Referenzkorpus

Für die kontrastive Termextraktion (siehe Abschnitt 4.3.3) sind Millionen Frequenzabfragen gegen den Wikipedia-Referenzkorpus erforderlich. Die Rohdaten umfassen 95 MB komprimierte Unigramm Frequenzen und 1,9 GB komprimierte Bigramm Frequenzen aus dem Wikipedia-N-Gramm-Korpus [33]. Das naive Laden dieser Daten in den Arbeitsspeicher würde nach der Dekompression mehrere Gigabyte RAM beanspruchen und die Ausführung der Pipeline auf Systemen mit begrenztem Speicher unmöglich machen.

Die Implementierung realisiert die in Abschnitt 4.2.2 beschriebene Architektur: xxHash-64 Hashing, sortierte NumPy-Arrays mit Struktur `[('hash', '<u8'), ('count', '<u4')]` und binäre Suche via `np.searchsorted`. Nach Filterung auf Terme mit mindestens 10 Vorkommen resultieren 1,3 Millionen Unigramme (15 MB) und 10 Millionen Bigramme (114 MB) bei einem RAM-Verbrauch von 43 MB und einer durchschnittlichen Lookup-Latenz von 1,57 μs [1].

[1] Alle Benchmarks: Apple M2 (2022), 8 CPU-Kerne, 8 GB RAM, Python 3.12, macOS Tahoe 26.3.

5.2.2 Vektorisierte Statistikberechnung

Für die Termextraktion (Abschnitt 4.3) müssen statistische Assoziationsmaße für mehrere tausend N-Gramm-Kandidaten berechnet werden. Eine naive Implementierung würde für jeden Kandidaten nacheinander Frequenzen abfragen, LLR, NPMI und IDF berechnen und die Ergebnisse iterativ aggregieren. Diese Vorgehensweise weist eine Komplexität von $O(n \cdot m)$ auf, wobei n die Anzahl der Kandidaten und m die durchschnittliche Anzahl der Lookups pro Kandidat darstellt. Bei typischen Korpusgrößen (20.000–30.000 Posts) entstehen 5.000–10.000 Kandidaten mit jeweils 2–4 Lookups, was zu mehreren zehntausend sequenziellen Operationen führt.

Die Implementierung eliminiert diese Ineffizienz durch Batchprocessing und Vektorisierung. Mit spaCy werden alle Posttitel in einem einzigen Durchlauf mittels parallelem POS-Tagging verarbeitet, wodurch sich die Tokenisierung und linguistische Annotation auf $O(n)$-Komplexität reduzieren lässt. Die extrahierten N-Gramme werden als NumPy-Arrays strukturiert, deren Frequenzen mittels Batchlookups gegen den Referenzkorpus abgefragt werden (Abschnitt 5.2.1). Die statistischen Assoziationsmaße (LLR, NPMI, IDF) werden mithilfe vektorisierter NumPy-Operationen parallel für alle Kandidaten berechnet. Die Strategie ersetzt verschachtelte Iterationen durch Single Instruction Multiple Data Verarbeitung und senkt die algorithmische Komplexität auf $O(n)$.

Die gemessene Performance für einen repräsentativen Korpus von 23.664 Posts beträgt 29,8 Sekunden zur Extraktion von 1.238 Kandidaten, was einer Durchsatzrate von 794 Posts pro Sekunde entspricht. Der Arbeitsspeicherbedarf bleibt bei 43 MB konstant, da die vektorisierten Operationen auf adressierten Daten operieren (Abschnitt 5.2.1). Die Voraballokation von NumPy Arrays begrenzt die maximale Anzahl an Kandidaten auf etwa 5 Millionen N-Gramme bei 8 GB Arbeitsspeicher, was für die Analyse spezialisierter Subreddit Korpora ausreichend ist. Für eine Skalierung auf größere Korpora wäre entweder inkrementelles Processing oder eine erhöhte RAM-Kapazität erforderlich.

5.2.3 Strategisches Sampling für LLM-Integration

Die semantische Validierung (Abschnitt 4.5) erfordert LLM-basierte Relevanzprüfungen mit kontextuellem Input (Keywords, Termkandidat, Reddit-Posts). Die vollständige Evaluation aller statistisch qualifizierten Kandidaten (typischerweise 500–1.500 Terme) würde bei einer gemessenen durchschnittlichen Latenz von 1,73 Sekunden und Kosten von \$0,015 pro API-Call Gesamtlaufzeiten von 15–45 Minuten und Kosten von \$7,50–22,50 pro Durchlauf verursachen. Diese Ressourcenanforderungen sind bei iterativen Experimenten und Parameterevaluationen nicht tragbar.

Die Implementierung reduziert die Kandidatenmenge durch eine zweistufige Selektion. Zunächst werden die 15 höchstbewerteten Terme nach Statistical Score (Abschnitt 4.4.3) deterministisch ausgewählt.

Anschließend wird das 80. Perzentil der Score-Verteilung berechnet und alle Kandidaten oberhalb dieses Schwellwertes inkludiert, mit einer absoluten Obergrenze von 40 Termen. Mit dieser Strategie werden sowohl die statistisch dominantesten Signale als auch eine Stichprobe der oberen 20 Prozent evaluiert. Die Selektion spiegelt die in Abschnitt 4.3 etablierte Priorisierung von Präzision gegenüber Recall wider: Terme mit niedrigen Statistical Scores werden als statistisch weniger robuste Frühindikatoren klassifiziert und von der LLM-Evaluation ausgeschlossen. Die finale Trendliste umfasst in der Regel 10 bis 15 Terme, was den Erwartungen spezialisierter Nischen entspricht.

5.3 Reproduzierbarkeit

Die Reproduzierbarkeit wird durch deterministische Konfiguration aller stochastischen Systemkomponenten sichergestellt: `seed=42` und `temperature=0.0` erzwingen Argmax-Dekodierung bei der LLM Inferenz, alle Abhängigkeiten sind versionsfixiert und die Arctic Shift API liefert für identische Parameter deterministische Ergebnisse. Die vollständige technische Spezifikation ist in Anhang A.5 dokumentiert.

6 Evaluation

In diesem Kapitel wird die Leistungsfähigkeit des entwickelten Systems durch retrospektive Simulation zu definierten Zeitpunkten vor dem historisch bekannten Mainstream Durchbruch von 45 Trends über drei Domänen quantifiziert. Abschnitt 6.1 beschreibt das Evaluationsdesign bestehend aus Ground Truth Konstruktion, Metriken und experimentellem Protokoll. Abschnitt 6.2 präsentiert die quantitativen Ergebnisse des Gesamtsystems sowie die Befunde der Ablation Studies zur Beantwortung der drei Forschungsfragen. Abschnitt 6.3 ergänzt die quantitative Analyse durch qualitative Fallstudien und eine Bewertung der Termkanonisierung.

6.1 Evaluationsdesign

6.1.1 Ground Truth Konstruktion

Für die retrospektive Evaluation ist eine operationale Definition des Mainstreamdurchbruchs erforderlich, die ohne manuelle Annotation auskommt und deterministisch auf historische Daten anwendbar ist. Wikipedia Pageviews eignen sich als Proxy für diesen Zeitpunkt, da sie das Information Seeking Behavior der breiten Öffentlichkeit messen: Ein signifikanter Anstieg der Abrufzahlen zu einem Begriff deutet darauf hin, dass dieser die Grenzen spezialisierter Communities überschritten hat und in den allgemeinen öffentlichen Diskurs eingedrungen ist.

Im Gegensatz zu plattformspezifischen Engagement Metriken sind die Wikipedia Pageviews archiviert und plattformunabhängig frei zugänglich. Dadurch ist die Reproduzierbarkeit der Ground Truth Konstruktion gewährleistet.

Als Mainstream Breakthrough Datum wird der erste Tag definiert, an dem drei Kriterien simultan erfüllt sind: Der 7-Tage Rolling Mean muss mindestens das 1,5-fache der 30-Tage Baseline erreichen, dieser Wert muss entweder 300 Pageviews oder 5% des globalen Peaks überschreiten, und dieses Niveau muss in den drei Folgetagen zu mindestens 70% gehalten werden. Die Kombination aus relativem Schwellwert, absoluter Mindestrelevanz und Persistenzkriterium sorgt dafür, dass sowohl kurzfristige Ausreißer als auch Artefakte bei Artikeln mit geringem Grundvolumen ausgeschlossen werden.

Der Evaluationsdatensatz umfasst 45 historische Trends aus drei Domänen, jeweils bestehend aus 15 Trends: Artificial Intelligence, Cryptocurrency und Web3 sowie Gaming. Um die Generalisierbarkeit der Methodik unter verschiedenen Trenddynamiken zu testen, orientiert sich die Domänenauswahl an strukturell unterschiedlichen Emergenzmustern. Typischerweise wachsen Artificial Intelligence (AI) Trends wie ChatGPT oder Stable Diffusion organisch über mehrere Monate. Cryptocurrency Trends wie der FTX Collapse sind hingegen häufig ereignisinduziert und gehen mit abrupten Diskurssprüngen einher. Gaming Trends folgen dagegen releasegetriebenen Ankündigungszyklen. Die Trendauswahl erfolgte nach dem Prinzip des Convenience Sampling nach Bekanntheit und Verfügbarkeit hinreichender Wikipedia Pageview Daten, wodurch ausschließlich Trends mit nachgewiesenem Mainstreamdurchbruch in den Datensatz eingehen und ein Selection Bias entsteht, der in Abschnitt 7.2 diskutiert wird.

6.1.2 Quantitative Metriken

Als primäre Evaluationsmetrik dient die Detection Rate (DR). Sie ist definiert als Anteil der korrekt detektierten Ground Truth Trends an der Gesamtzahl der Trends im Datensatz. Ein Trend gilt als detektiert, wenn das System ihn in mindestens einem der sechs Simulationsfenster korrekt detektiert hat. Als sekundäre Metrik kommt der Mean Reciprocal Rank (MRR) zum Einsatz, der die Rangposition detektierter Trends positionsgewichtet erfasst: $\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$, wobei $|Q|$ die Anzahl der relevanten Trends und rank_i die Rangposition des i-ten detektierten Trends bezeichnet.

Die Kombination beider Metriken ist methodisch erforderlich, da sie komplementäre Aspekte der Systemleistung erfassen. Allein die DR würde das rein statistische Baselineverfahren als überlegenes System ausweisen, da es durch den Verzicht auf semantische Filterung einen deutlich höheren Recall erzielt. Für den praktischen Anwendungskontext des vorliegenden Systems, dessen Ausgabe eine priorisierte Liste von zehn bis fünfzehn Begriffen ist, ist die Rangqualität der Treffer jedoch entscheidend: Ein Trend, der korrekt detektiert wird, aber nur auf Rang zwölf erscheint, hat einen geringeren praktischen Wert als ein Trend, der konsistent auf den ersten Rängen platziert wird.

Diese Anforderung wird durch den MRR operationalisiert, der eine differenzierte Bewertung der Systemvarianten jenseits des reinen Recallvergleichs ermöglicht.

6.1.3 Experimentelles Protokoll

Die Evaluation simuliert das System zu sechs definierten Zeitpunkten, wobei w die Lead Time in Tagen vor dem Durchbruch bezeichnet, vor dem jeweiligen Breakthrough Datum: 7, 21, 30, 45, 60 und 90 Tage. Pro Simulationsfenster wird die identische Pipeline bestehend aus Termextraktion, Timeline Building und Trend Detection auf einem zeitlich beschnittenen Korpus ausgeführt, der am jeweiligen Reference Date endet. Insgesamt ergeben sich 270 Simulationsläufe aus allen 45 Trends und sechs Fenstern. Als Baselines dienen drei Verfahren: eine Random Baseline, die Kandidaten zufällig aus dem Pool selektiert, eine Volume Baseline, die Kandidaten ausschließlich nach Mention Count rankiert, sowie ein rein statistisches Systemverfahren ohne LLM Komponente, das den Mehrwert der semantischen Filterung isoliert.

Der Abgleich der Systemausgaben mit der Ground Truth erfolgt durch ein dreischichtiges Matching Verfahren: Substring Match als erster Pass, Token Overlap via Jaccard Koeffizient mit einem Schwellwert von 0,4 als zweiter Pass, sowie ein Synonym Lookup als dritter Pass. Für die drei primären Konfigurationen `full_system`, `local_llm` und `uniform_weights` wurden sämtliche Matches vollständig manuell durch den Autor identifiziert und validiert, um die Korrektheit der Zuordnung zwischen Systemausgabe und Ground Truth sicherzustellen. Für `statistical_only` entfiel die manuelle Validation aufgrund der überproportional hohen Trefferzahl, da diese Konfiguration ausschließlich als Baseline dient und nicht als primäre Systemvariante evaluiert wird.

6.2 Quantitative Ergebnisse

6.2.1 Performance des Gesamtsystems

Das Gesamtsystem (`full_system`) detektiert 18 von 45 Ground Truth Trends, was einer DR von 40,0 % bei einem MRR von 0,587 entspricht. Gegenüber beiden trivialen Baselines verdoppelt das System die DR: Die Random Baseline und die Volume Baseline erzielen jeweils 20,0 %, mit MRR-Werten von 0,355 beziehungsweise 0,298. Besonders aussagekräftig ist dabei der Abstand zur Volume Baseline, da es sich dabei um das nächstliegende naive Verfahren handelt, das ausschließlich auf dem rohen Erwähnungsvolumen basiert und somit weder eine kontrastive Korpusanalyse noch eine zeitliche Filterung nutzt. Das Gesamtsystem übertrifft dieses Verfahren in beiden Metriken deutlich und zeigt, dass die mehrstufige Filterkaskade einen wesentlichen Mehrwert gegenüber rein volumenbasierter Selektion bietet.

Tabelle 5: Performance der Systemkonfigurationen und Baselines

Konfiguration	Detektiert	DR	MRR
`full_system`	18/45	40,0 %	0,587
`statistical_only`	28/45	62,2 %	0,219
`local_llm`	15/45	33,3 %	0,683
`uniform_weights`	12/45	26,7 %	0,662
Random Baseline	9/45	20,0 %	0,355
Volume Baseline	9/45	20,0 %	0,298

Der domänenspezifische Breakdown zeigt einen klaren Gradienten: Die DR beträgt 60,0 % für AI, 46,7 % für Cryptocurrency und Web3 sowie 13,3 % für Gaming. Dieser Gradient korreliert mit den strukturellen Charakteristika der in Abschnitt 6.1 beschriebenen Emergenzprozesse: Organisch wachsende Trends mit gradueller Diskursentwicklung über mehrere Monate stellen den optimalen Systemauslöser dar, da sie einen stabilen terminologischen Fußabdruck in spezialisierten Communities hinterlassen, bevor sie den Mainstream erreichen. Ereignisinduzierte Trends im Kryptowährungsbereich sind teilweise detektierbar, sofern dem eigentlichen Ereignis eine Aufbauphase in der Community vorausgeht. Releasegetriebene Gaming Trends entstehen hingegen häufig durch externe Ankündigungsereignisse, denen kein gradueller terminologischer Vorlauf in spezialisierten Subreddits vorausgeht – ein strukturelles Mismatch zum Detektionsprinzip des Systems.

Die 18 unique detektierten Trends verteilen sich ungleichmäßig auf die sechs Simulationszeitpunkte, wobei ein Trend in mehreren Fenstern detektiert werden kann: Das Fenster mit 21 Tagen Vorlauf weist mit 8 Detektionsereignissen die höchste Einzelfensterpräsenz auf, gefolgt von $w = 45$ (6), $w = 30$ und $w = 60$ (je 5), $w = 7$ (4) sowie $w = 90$ (2). Dass das dem Mainstreamdurchbruch am nächsten liegende Fenster ($w = 7$) nicht die höchste Detektionsrate erzielt, ist ein bemerkenswerter Befund, dessen Ursachen in Abschnitt 7.2 diskutiert werden. Um die Systemleistung vollständig zu interpretieren, müssen DR und MRR gemeinsam betrachtet werden, da beide Metriken komplementäre Aspekte der Leistungsfähigkeit abbilden. Dieser Zusammenhang wird in den nachfolgenden Ablation Studies expliziert.

6.2.2 Semantischer Relevanzfilter (RQ1)

Der Vergleich dreier Konfigurationen mit unterschiedlicher semantischer Filtertiefe quantifiziert den Beitrag der LLM-basierten Validierung zur Systemleistung. Das rein statistische Baselineverfahren `statistical_only` erzielt eine DR von 62,2 % bei einem MRR von 0,219. Das Gesamtsystem `full_system` detektiert mit 40,0 % DR und MRR 0,587 deutlich weniger Trends, platziert die detektierten Trends jedoch signifikant höher. Die Konfiguration `local_llm`, die den semantischen Validierungsschritt durch Llama 3 8B ersetzt, erzielt 33,3 % DR bei einem MRR von 0,683.

Methodisch ist dabei eine Einschränkung des Vergleichs hervorzuheben: `local_llm` ersetzt ausschließlich beide Passes von Phase 3 der Pipeline in `analysis.py` durch das lokal ausgeführte Modell. Die Community Discovery in Phase 1 (Abschnitt 4.2.1) läuft in allen drei Konfigurationen identisch über GPT-4o. Der Vergleich isoliert somit den Einfluss des Validierungsmodells auf das Relevanzscoring und die Kanonisierung, nicht jedoch die Gesamtwirkung des LLM-Einsatzes im System.

Der Mechanismus hinter dem DR–MRR-Trade-off ist im Aufbau der Scoreberechnung begründet: Das LLM berechnet einen Relevanzscore im Intervall $[0{,}0, 1{,}0]$, der multiplikativ mit dem Statistical Score zu einem Final Score kombiniert wird (Abschnitt 4.5.2). Diese multiplikative Fusion bewirkt eine implizite Neuordnung der Kandidatenliste. Statistisch starke, aber semantisch irrelevante Terme – etwa generische Diskussionsartefakte oder plattformspezifisches Rauschen – werden durch niedrige Relevanzscores gedämpft und aus den oberen Rängen verdrängt. `statistical_only` verzichtet vollständig auf diesen Mechanismus und selektiert ausschließlich nach statistischer Auffälligkeit. Die daraus resultierende hohe DR von 62,2 % belegt, dass relevante Terme statistisch detektierbar sind. Der MRR von 0,219 zeigt jedoch, dass sie unter einer großen Menge statistisch gleichwertiger, aber semantisch irrelevanter Terme begraben werden und in einer priorisierten Liste von zehn bis fünfzehn Begriffen praktisch nicht auffindbar wären.

Das Verhalten von `local_llm` offenbart eine qualitative Asymmetrie zwischen großen und kleinen Sprachmodellen bei dieser Aufgabe. Mit einem MRR von 0,683 übertrifft Llama 3 8B das `full_system` (0,587) und deutet somit auf eine präzisere Priorisierung bei den korrekt detektierten Trends hin. Gleichzeitig sinkt die DR auf 33,3 %, was einem Verlust von 6,7 Prozentpunkten gegenüber `full_system` entspricht. Dieses Muster ist charakteristisch für ein überkonservatives Filterverhalten. Das kleinere Modell vergibt systematisch niedrigere Relevanzscores und eliminiert dabei nicht nur tatsächlichen Noise, sondern auch valide Trends mit peripherem oder implizitem Themenbezug. Brown et al. [6] zeigen, dass die Few-Shot-Fähigkeiten von LLMs überproportional mit der Modellgröße skalieren – eine Gesetzmäßigkeit, die sich im vorliegenden Befund empirisch manifestiert: So generalisiert das kleinere Modell schlechter auf die kontextuell anspruchsvolle Aufgabe, semantische Relevanz aus dem Vokabular von Reddit Diskursen zu inferieren.

Für Forschungsfrage 1 ist die entscheidende Frage, welche Konfiguration für den praktischen Anwendungskontext des Systems überlegen ist. Beide Varianten mit semantischer Filterung erzielen gegenüber `statistical_only` einen deutlichen MRR-Gewinn bei reduzierten DR-Werten. Da es sich bei der Systemausgabe um eine priorisierte Liste handelt, deren obere Ränge die primäre Nutzungsschnittstelle darstellen, ist die Rangqualität der Treffer wichtiger als die absolute Abdeckung. Ein Trend auf Rang 1 hat für den Nutzer einen deutlich höheren Wert als ein korrekt detektierter Trend auf Rang 12. Unter dieser Prämisse übertrifft `full_system` mit seiner höheren DR bei vergleichbarer Rangqualität die lokale Alternative.

Der durch GPT-4o-mini verursachte Ressourcenaufwand – im Rahmen des strategischen Samplings auf maximal 40 Kandidaten begrenzt (Abschnitt 5.2.3) – ist durch den kombinierten Gewinn aus DR und MRR gegenüber `statistical_only` gerechtfertigt. Forschungsfrage 1 ist somit positiv zu beantworten: Der semantische Relevanzfilter steigert die praktisch relevante Systemleistung, und der Ressourcenaufwand durch externe API-Calls ist durch die Qualitätsverbesserung begründbar.

6.2.3 Inverse Gewichtung (RQ2)

Der Vergleich zwischen `full_system` und `uniform_weights` isoliert den Beitrag der inversen logarithmischen Subreddit Gewichtung, da sich beide Konfigurationen lediglich in der Aktivierung der inversen Gewichtung unterscheiden, während alle übrigen Parameter identisch sind. `full_system` erzielt eine DR von 40,0 % bei einem MRR von 0,587, während `uniform_weights` auf 26,7 % DR bei einem MRR von 0,662 fällt. Der Gewinn von 13,3 Prozentpunkten in der DR belegt, dass die inverse Gewichtung einen substanziellen Beitrag zur Detektionsleistung leistet.

Dieser Befund ist auf das Zusammenspiel zwischen der Größenverteilung von Subreddits und der Signalqualität zurückzuführen. Ohne inverse Gewichtung dominieren große Communities mit einem hohen absoluten Postvolumen die aggregierten Termscores. Dadurch werden Nischensignale aus kleinen Subreddits algorithmisch unterdrückt. Da die Größenverteilung von Social Media Communities empirisch einer Power Law Verteilung folgt [31], sind ungewichtete Aggregationen strukturell gegenüber Mainstream Diskursen verzerrt, in denen emergente Terme bereits eine breitere Verbreitung gefunden haben. Die logarithmische Skalierung linearisiert diese Verteilung und hebt Termfrequenzen aus kleinen, thematisch fokussierten Communities proportional an. Dies ist konsistent mit Rogers' Befund, dass die Innovatoren Fraktion in kleinen Peer Netzwerken agiert, die in zentralisierten Aggregationen systematisch untergehen (vgl. Abschnitt 4.2.1).

Das gegenläufige Verhalten des MRR – `uniform_weights` erzielt 0,662 gegenüber 0,587 beim `full_system` – ist auf dasselbe statistische Artefakt zurückzuführen, das bereits in Abschnitt 6.2.2 für `statistical_only` beschrieben wurde: Bei einer geringeren absoluten Trefferzahl (12 gegenüber 18) entfällt der MRR Denominator auf eine kleinere Menge, was den Durchschnittswert gegenüber dem tatsächlichen Rangniveau aufwärts verzerrt. Der scheinbare MRR Vorteil von `uniform_weights` ist kein Indikator überlegener Rangqualität: Bei nur 12 detektierten Trends gegenüber 18 beim `full_system` wird der MRR Durchschnitt durch wenige, hoch platzierte Treffer stärker dominiert, was den Wert strukturell nach oben verzerrt.

Für die Interpretation der Magnitude ist eine methodische Einschränkung erforderlich: Bei einem Datensatz von 45 Trends entspricht eine Zunahme von sechs zusätzlichen Detektionen einem Anstieg von 13,3 Prozentpunkten in der DR, sodass die prozentuale Darstellung die absolute Effektgröße überzeichnet.

Die Richtung des Befunds ist jedoch eindeutig und durch Rogers' Diffusionstheorie theoretisch fundiert. Damit ist Forschungsfrage 2 positiv zu beantworten: Die inverse Gewichtung steigert den Recall substanziell, ohne die Rangqualität der verbleibenden Treffer praktisch relevant zu beeinträchtigen.

6.2.4 Beobachtungszeitraum (RQ3)

Die Variation des Beobachtungszeitraums von 30 auf 45, 60 oder 90 Tage führt zu einem nicht-monotonen Verlauf.: Die DR steigt von 31,1 % (30 Tage) auf 40,0 % beim optimalen Wert von 60 Tagen, fällt jedoch bei 45 Tagen auf 28,9 % und bei 90 Tagen weiter auf 24,4 %. Der MRR verhält sich gegenläufig zur DR, ohne ein eindeutiges Muster zu zeigen: 0,561 (30 Tage), 0,664 (45 Tage), 0,587 (60 Tage) und 0,615 (90 Tage). Bei gemeinsamer Betrachtung beider Metriken stellt `full_system` mit 60 Tagen den empirisch besten Kompromiss dar.

Für die Extremwerte lassen sich plausible Erklärungen finden: Ein Beobachtungszeitraum von 30 Tagen liefert eine zu schmale Datenbasis für eine stabile Schätzung des Mann-Kendall Tests (vgl. Abschnitt 4.4). Dies destabilisiert die Trendkonsistenzkomponente des Scores und reduziert die DR. Bei 90 Tagen tritt das entgegengesetzte Problem auf: Etablierte Terme zeigen über den gesamten Zeitraum kontinuierliche Aktivität. Dadurch detektiert Fisher's Exact Test keinen signifikanten Kontrast zwischen dem zurückliegenden und dem aktuellen Fensterabschnitt mehr und der Stationaritätsbruch als Detektionssignal verschwindet. Auf Basis der vorliegenden Daten lässt sich der nicht-monotone Verlauf zwischen 30 und 60 Tagen – insbesondere der DR-Einbruch bei 45 Tagen gegenüber 30 Tagen – nicht abschließend erklären und ist bei vier Messpunkten methodisch vorsichtig zu interpretieren.

Forschungsfrage 3 ist damit differenziert zu beantworten: Ein Beobachtungszeitraum von 60 Tagen stellt den empirisch optimalen Kompromiss zwischen statistischer Stabilität und Signalfrische dar. Die nicht-monotone Charakteristik des Verlaufs deutet jedoch darauf hin, dass der optimale Wert von der jeweiligen Domäne und dem Emergenztyp abhängen könnte. Eine domänenspezifische Kalibrierung dieses Parameters wäre daher ein naheliegender Ansatzpunkt für weiterführende Arbeiten (vgl. Abschnitt 8.2).

6.2.5 Kohäsionsschwelle

Die Variation des NPMI Schwellwerts von 0,0 bis 0,6 zeigt eine ausgeprägte Robustheit der Systemleistung: Die DR bewegt sich zwischen 28,9 % (`npmi_0.5`) und 31,1 % (`npmi_0.0`, `npmi_0.2`, `npmi_0.3`, `npmi_0.6`), der MRR zwischen 0,560 (`npmi_0.2`) und 0,654 (`npmi_0.3`). Der Ursprung dieser Robustheit liegt in der Struktur der Filterkaskade: Solange der Kandidatenpool grob sinnvolle Bigramme enthält, übernimmt der nachgelagerte LLM Filter die Feinfilterung und entkoppelt die Systemleistung vom vorgelagerten Kohäsionsschwellwert.

Dass selbst `npmi_0.0`, der vollständige Verzicht auf eine Kohäsionsprüfung, keine substanzielle Leistungsdegradation produziert, belegt diese Entkopplung empirisch. Der Standardwert von 0,4 ist damit als robuste Konfiguration bestätigt; die Kohäsionsschwelle erfüllt primär die Funktion, linguistisch implausible Kandidaten vor der kostenintensiven LLM Inferenz zu reduzieren.

6.2.6 Validierung der Metrikkombination

Der Vergleich zwischen `full_system` und den vier Einzelgewicht Varianten quantifiziert den Beitrag der kombinierten Scoring Funktion gegenüber einer Reduktion auf jeweils eine der drei Zeitreihenmetriken. `full_system` erzielt mit 40,0 % DR den höchsten Wert aller Varianten: `weights_consistency` erreicht 33,3 %, `weights_growth` und `weights_significance` je 31,1 % sowie `weights_equal` 28,9 %. Der MRR zeigt erneut das bekannte gegenläufige Muster – `weights_equal` erzielt mit 0,683 den höchsten Wert bei gleichzeitig niedrigster DR – und ist analog zu Abschnitt 6.2.2 als Artefakt der reduzierten Treffermenge zu interpretieren.

Der Leistungsvorsprung der kombinierten Scoring Funktion basiert auf der komplementären Natur der drei Metriken. Die Wachstumsrate reagiert sensibel auf kurzfristige Frequenzanstiege, ist jedoch anfällig für transiente Spikes, die nicht zu einer nachhaltigen Diskursentwicklung führen. Trendkonsistenz bevorzugt Terme mit stabiler Aufwärtsdynamik, läuft jedoch Gefahr, bereits etablierte Terme mit gleichmäßiger Basisaktivität zu stark zu gewichten. Der Fisher's Exact Test operiert als Schwellwertkriterium für statistische Signifikanz und liefert als kontinuierliches Rankingkriterium allein zu wenig Diskriminierungskraft. Erst die Kombination aller drei Metriken ermöglicht eine zuverlässige Unterscheidung zwischen echten Frühindikatoren und strukturellem Rauschen.

Die Gewichtung (Growth 0,45, Significance 0,30, Consistency 0,25) wurde a priori durch theoretische Überlegungen festgelegt, wobei die Wachstumsrate als zentrale Eigenschaft emergenter Trends das höchste Gewicht erhält, statistische Signifikanz als statistischer Anker das zweithöchste und Trendkonsistenz als nachgelagertes Qualitätskriterium das niedrigste. Die Evaluation bestätigt diese Priorisierung empirisch: `weights_equal` erzielt die schwächste DR aller Varianten, weist jedoch den höchsten MRR auf – ein Artefakt der reduzierten Treffermenge, das analog zum in Abschnitt 6.2.2 beschriebenen Mechanismus zu interpretieren ist, was belegt, dass die spezifische Gewichtung einen zusätzlichen Beitrag über die bloße Metrikkombination hinaus leistet.

6.3 Qualitative Ergebnisse

Während die quantitativen Metriken die Systemleistung auf Aggregatsebene beschreiben, ergänzen die nachfolgenden Fallstudien sowie die Analyse der Termkanonisierung diese Perspektive durch exemplarische Einblicke in konkrete Detektions- und Verarbeitungsmuster.

6.3.1 Fallstudien

Stable Diffusion (Best Case) Stable Diffusion wurde in vier der sechs simulierten Fenster detektiert, was den Idealfall des Systemdesigns darstellt. Im frühesten Fenster (w=7, Beobachtungszeitraum endet 7 Tage vor dem Mainstreamdurchbruch) erscheint der Term auf Rang 8 mit einem Score von 39 unter dem kanonischen Namen *text-to-image generation using stable diffusion* – das Signal ist vorhanden, aber noch schwach. Mit zunehmendem Vorlauf verdichtet sich das Signal: Bei w=45 (105 bis 45 Tage vor Durchbruch) und w=60 (120 bis 60 Tage vor Durchbruch) erreicht das System konsistent Rang 1 mit Scores von 76 bzw. 77. Aus mechanistischer Sicht ist dieser Verlauf auf organisches Wachstum zurückzuführen: Der Begriff akkumuliert über Monate hinweg eine stabile terminologische Präsenz, sodass Fisher's Exact Test einen klaren Stationaritätsbruch detektiert (xAbschnitt 4.4.1) und der Mann-Kendall Test eine monotone Aufwärtsdynamik bestätigt. Das mittlere Fenster liefert die höchste Signalqualität, weil der Diskurs dicht genug ist, um statistische Stabilität zu erreichen, der Begriff aber noch nicht im Mainstream angekommen ist und somit keinen flachen Konsistenzgradienten erzeugt.

FTX Collapse (Event-Trend) Der Kollaps der Kryptobörse FTX im November 2022 veranschaulicht das Detektionsprofil eines ereignisinduzierten Trends. Im kürzesten Fenster (w=7) findet keine Detektion statt. Bei w=21 erscheint der Term auf Rang 5 mit Score 44; bei w=30 steigt er auf Rang 1 mit Score 80, dem höchsten Einzelscore im gesamten Evaluationsdatensatz. Das System greift den Trend in der Phase auf, in der die Community die Implikationen des Ereignisses systematisch verarbeitet und dabei einen stabilen terminologischen Fußabdruck hinterlässt. Ereignisinduzierte Trends sind demnach detektierbar, sofern ihnen eine ausreichend lange Diskursaufbauphase vorausgeht; ob und unter welchen Bedingungen das System den initialen Schock selbst erfasst, lässt sich anhand dieses Einzelfalls nicht verallgemeinern.

Umbrella Brands (Failure Case) Die prominentesten AI Brands wie OpenAI, GPT-4, Claude und DALL-E erzielen über alle sechs Fenster hinweg null Detektionen. Anhand der Evaluation lassen sich zwei strukturell verschiedene Ursachen unterscheiden. Bei OpenAI und Claude liegt der Wikipedia Pageview Durchbruch weit nach dem jeweiligen Marktauftritt der Organisationen.

Dies deutet darauf hin, dass diese Begriffe bereits lange zuvor in einschlägigen Subreddits dauerhaft diskutiert wurden und somit keinen detektierbaren Stationaritätsbruch erzeugen (vgl. Abschnitt 4.4.1). Bei GPT-4 und DALL-E liegt der Durchbruch dagegen unmittelbar am Erscheinungsdatum: Die Simulationsfenster enden kurz vor dem Release, da zu diesem Zeitpunkt noch kein ausreichendes Diskursvolumen für einen statistisch stabilen Befund vorhanden ist. Beide Ursachen werden in Abschnitt 7.2 diskutiert.

6.3.2 Qualität der Termkanonisierung

Tabelle 6: Beispiele zur Qualität der Kanonisierung

Rohterm	**Kanonischer Term**	**Bewertung**
sora	openai sora text-to-video generation technology	✓
alphafold	deepmind's alphafold for protein structure prediction	✓
tornado cash	tornado cash sanctions and regulatory implications	✓
midjourney	midjourney invites	×
cyberpunk	cyberpunk role-playing games (rpgs)	△

Die erfolgreichen Fälle belegen, dass das LLM generische Einwortterme kontextspezifisch disambiguiert und Akronyme mit Hersteller- und Funktionsangaben anreichert. Der Fehlerfall *midjourney invites* illustriert eine strukturelle Grenze: In frühen Simulationsfenstern wird der Kontext durch temporäre, diskursdominante Features überlagert, sodass das LLM den zum Detektionszeitpunkt dominanten Diskursfokus kanonisiert, anstatt den zugrunde liegenden Trend. Der Grenzfall *cyberpunk role-playing games* überschreitet den eigentlichen Referenten (Cyberpunk 2077) in Richtung Gattungsbegriff, bleibt dabei jedoch thematisch kohärent. Beide Fälle deuten darauf hin, dass die Kanonisierungsqualität mit dem Reifegrad des Diskurses zum Detektionszeitpunkt korreliert, was in Abschnitt 7.2 weiter diskutiert wird.

7 Diskussion

7.1 Interpretation und Beantwortung der Forschungsfragen

Die Ergebnisse der Evaluation zeichnen ein konsistentes Bild: Das System funktioniert als Kaskade, wobei jede Komponente eine notwendige, aber nicht hinreichende Bedingung für die Gesamtleistung darstellt. Dies wird durch die Ablation Studies strukturell belegt, da jede Konfiguration, die eine Systemkomponente isoliert, zu einer Leistungsdegradation in mindestens einer der beiden Evaluationsmetriken führt.

Forschungsfrage 1 ist positiv zu beantworten, erfordert jedoch eine metrische Differenzierung. Der semantische Relevanzfilter steigert nicht primär die DR, sondern die Rangqualität der detektierten Trends. Die multiplikative Fusion wirkt als Discount Faktor, der semantisch irrelevante Terme aus den oberen Rängen verdrängt – bei einem System, dessen Ausgabe eine priorisierte Liste von zehn bis fünfzehn Begriffen ist, die entscheidend relevante Dimension. Der Vergleich zwischen `full_system` und `local_llm` ist unter dem Vorbehalt zu interpretieren, dass beide Konfigurationen durch praktische Randbedingungen limitiert sind: `local_llm` nutzt das größte auf der verfügbaren Hardware ausführbare Modell, `full_system` das kosteneffizienteste proprietäre Modell. Ob größere lokale Modelle das Gleichgewicht zwischen DR und MRR substanziell verschieben, lässt sich auf Basis der vorliegenden Evaluation nicht beurteilen.

Forschungsfrage 2 ist positiv zu beantworten. Die inverse Gewichtung steigert den Recall substanziell, wobei der scheinbare MRR-Vorteil der ungewichteten Variante als statistisches Artefakt der geringeren Trefferzahl zu interpretieren ist. Die Richtung des Befunds ist eindeutig und durch Rogers' Diffusionstheorie [8] theoretisch fundiert, wenngleich die absolute Effektgröße bei $n = 45$ vorsichtig zu interpretieren ist.

Forschungsfrage 3 ist differenziert zu beantworten. Ein Beobachtungszeitraum von 60 Tagen stellt den empirisch optimalen Kompromiss zwischen statistischer Stabilität und Signalfrische dar. Die nicht-monotone Charakteristik des Verlaufs deutet darauf hin, dass der optimale Wert von der jeweiligen Domäne und dem Emergenztyp abhängen könnte, was bei vier Messpunkten methodisch vorsichtig zu interpretieren ist.

7.2 Limitationen

Die zentrale methodische Einschränkung betrifft die Struktur des Evaluationsdesigns. Da die Ground Truth ausschließlich Trends mit nachgewiesenem Mainstreamdurchbruch enthält, ist die Präzision des Systems strukturell nicht messbar: Die DR misst den Recall unter der Bedingung aktiver Diskussion im Beobachtungszeitraum und nicht die Gesamtqualität der Systemausgabe. Ergänzend bevorzugt das Convenience Sampling Trends mit einem starken terminologischen Diskursmuster, sodass die berichteten DR-Werte gegenüber einer repräsentativen Trendpopulation tendenziell überschätzt sein dürften.

Der domänenspezifische Gradient der DR ist nicht als Messfehler, sondern als inhärentes Systemlimit zu betrachten. Das Detektionsprinzip setzt einen graduellen terminologischen Vorlauf voraus; releasegetriebene Trends entstehen durch externe Ankündigungsereignisse ohne diesen Vorlauf – ein konzeptionelles Mismatch, das durch Parameteranpassungen nicht zu beheben ist.

Ein verwandtes strukturelles Limit zeigen die nicht detektierten Markenbegriffe. Etablierte Begriffe erzeugen keinen nachweisbaren stationären Bruch mehr; Begriffe mit releaseinduziertem Durchbruch liefern vor dem Erscheinungsdatum kein ausreichendes Diskursvolumen. Beide Muster veranschaulichen, dass das System Trends in einem bestimmten Entwicklungsstadium erkennt – nach dem Aufbau einer Diskursbasis, aber vor dem Mainstreamdurchbruch – und für Signale außerhalb dieses Fensters strukturell blind ist.

Der Befund, dass $w = 7$ nicht die höchste Detektionsrate erzielt, lässt sich hypothetisch damit erklären, dass ein Term kurz vor dem Durchbruch zum Mainstream bereits so präsent ist, dass kein signifikanter Kontrast zwischen den Zeitfenstern mehr detektiert wird. Mit den vorliegenden Daten ist diese Hypothese nicht isoliert prüfbar.

Die Pageviews von Wikipedia sind ein valider, aber vereinfachender Proxy: Das Maß erfasst das Informationssuchverhalten, nicht die gesellschaftliche Relevanz im weiteren Sinne. Das algorithmisch bestimmte Breakthrough Datum repräsentiert zudem keinen objektiven Zeitpunkt der gesellschaftlichen Durchdringung.

Eine grundlegende Einschränkung der retrospektiven Evaluationsmethodik ist die Kontamination der Trainingsdaten der eingesetzten Sprachmodelle. GPT-4o-mini und Llama 3 8B wurden auf Korpora trainiert, die die evaluierten Trends bereits enthalten. Bei der Relevanzbewertung von Begriffen wie *stable diffusion* oder *ftx* stützen sich die Modelle möglicherweise nicht ausschließlich auf den bereitgestellten Reddit-Kontext, sondern auch auf im Training erworbenes Vorwissen. Dies dürfte zu einer systematischen Überschätzung der Relevanzscores für bekannte Trends führen. In der Produktionsumgebung bewertet das System hingegen emergente, dem Modell zum Zeitpunkt der Inferenz unbekannte Begriffe. Die retrospektive Evaluation misst damit nicht exakt die Leistung, die im realen Anwendungsfall relevant ist, und die berichteten Ergebnisse sind hinsichtlich ihrer externen Validität entsprechend einzuordnen.

Eine weitere methodische Einschränkung betrifft die Annahmen der statistischen Signifikanztests. Fisher's Exact Test setzt die Unabhängigkeit der Beobachtungen voraus, während Reddit-Posts zeitlich korreliert sind: Ein viraler Beitrag löst Folgediskussionen aus, die wie scheinbar unabhängige Datenpunkte in die Kontingenztabelle eingehen. Da diese zeitliche Autokorrelation nicht explizit modelliert wird, kann es zu einer systematischen Unterschätzung der p-Werte kommen. Für die Identifikation statistisch auffälliger Aktivitätsanstiege ist diese Approximation pragmatisch vertretbar. Die exakten Signifikanzniveaus sind jedoch mit Vorsicht zu interpretieren.

Das System ist außerdem anfällig für koordinierte Manipulationen der Eingabedaten. Reddit ist bekanntermaßen von automatisierter Bot-Aktivität und koordinierten Manipulationskampagnen betroffen, wobei gerade die evaluierten Domänen Cryptocurrency und AI aufgrund ökonomischer Anreizstrukturen besonders exponiert sind.

Da die Implementierung keine expliziten Mechanismen zur Erkennung manipulativer Aktivitätsmuster enthält, können koordinierte Kampagnen als genuine Trendsignale fehlinterpretiert werden.

Die Parametrisierung des Systems wurde a priori auf Basis theoretischer Überlegungen und domänenspezifischer Heuristiken festgelegt. Anschließend wurde sie auf demselben Datensatz evaluiert, auf dem die Schwellwerte kalibriert wurden. Da weder ein separates Holdout-Set noch ein Kreuzvalidierungsverfahren eingesetzt wurden, kann eine Überanpassung der berichteten Performance an die spezifische Trendauswahl und Domänenstruktur nicht ausgeschlossen werden. Die Generalisierbarkeit der konkreten Parameterwerte auf andere Domänen oder Zeiträume muss unabhängig validiert werden.

Schließlich steht die Qualität der Termkanonisierung in Korrelation zum Reifegrad des Diskurses zum Zeitpunkt der Detektion. In frühen Simulationsfenstern kanonisiert das LLM den temporär dominanten Diskursfokus anstelle des zugrunde liegenden Trends, da die Informationsdichte des bereitgestellten Kontexts in diesen Phasen zwangsläufig geringer ist [6].

8 Fazit

8.1 Zusammenfassung

Die vorliegende Arbeit adressierte die Frage, wie Vorläufersignale von Mainstreamtrends aus dem Informationsrauschen spezialisierter Reddit Communities isoliert werden können. Hierzu wurde eine dreistufige Filterkaskade entwickelt, die statistische Robustheit mit semantischer Intelligenz kombiniert. Die kontrastive Korpusanalyse identifiziert domänenspezifische Termkandidaten mithilfe des Unithood-/Termhood-Frameworks [4], die Zeitreihenanalyse validiert deren temporale Dynamik durch die Kombination aus Fisher's Exact Test und Mann-Kendall Test und die LLM-basierte Validierung filtert semantisch irrelevante Kandidaten durch ICL [6]. Die Architektur reduziert den Kandidatenraum von ca. 25.000 Beiträgen auf zehn bis 15 finale Trends und priorisiert dabei Präzision gegenüber Vollständigkeit [7].

Die retrospektive Evaluation von 45 historischen Trends aus drei Domänen quantifiziert die Systemleistung mithilfe von 270 Simulationsläufen. Das Gesamtsystem erzielt eine DR von 40,0 % bei einem MRR von 0,587 und übertrifft damit beide trivialen Baselines in beiden Metriken. Die drei Forschungsfragen konnten wie folgt beantwortet werden: Der semantische LLM-Filter steigert die Rangqualität detektierter Trends substanziell, wobei die multiplikative Fusion als Discount Faktor wirkt und den MRR gegenüber dem rein statistischen Baselineverfahren von 0,219 auf 0,587 erhöht (RQ1).

Die inverse logarithmische Gewichtung von Subreddits nach Communitygröße verbessert die DR um 13,3 Prozentpunkte gegenüber der ungewichteten Aggregation und operationalisiert Rogers' Diffusionstheorie [8] empirisch (RQ2). Ein Beobachtungszeitraum von 60 Tagen stellt den optimalen Kompromiss zwischen statistischer Stabilität und Signalfrische dar. Der nicht-monotone Verlauf der DR über verschiedene Zeitfenster deutet jedoch darauf hin, dass der optimale Wert vom Emergenztyp der jeweiligen Domäne abhängen könnte (RQ3).

Der domänenspezifische Gradient – 60,0 % DR für AI, 46,7 % für Cryptocurrency, 13,3 % für Gaming – offenbart ein inhärentes Systemlimit: Das Detektionsprinzip setzt einen graduellen terminologischen Vorlauf in spezialisierten Communities voraus und ist für releasegetriebene Trends ohne diesen Vorlauf konzeptionell ungeeignet. Die Arbeit zeigt damit sowohl das Potenzial als auch die strukturellen Grenzen kontrastiver Korpusanalyse für die Trendfrüherkennung auf.

8.2 Ausblick

Die in Abschnitt 7.2 identifizierten Limitationen eröffnen drei naheliegende Forschungsrichtungen. Erstens erfordert die Validierung der Generalisierbarkeit eine prospektive Evaluation. Dabei wird das System über einen definierten Zeitraum im Produktivbetrieb ausgeführt und die Ausgaben werden gegen nachträglich eintretende Mainstreamdurchbrüche abgeglichen. Eine solche Studie würde die Kontamination durch die Trainingsdaten der Sprachmodelle eliminieren und die externe Validität der berichteten Ergebnisse unabhängig überprüfen.

Zweitens könnte eine domänenadaptive Parametrisierung den nicht-monotonen Verlauf der DR über verschiedene Beobachtungszeiträume berücksichtigen. Die Hypothese, dass organisch wachsende Trends von längeren, ereignisinduzierten Trends aus kürzeren Zeitfenstern profitieren, ließe sich durch eine automatische Klassifizierung des Emergenztyps und eine entsprechende dynamische Anpassung der Zeitreihenparameter operationalisieren.

Drittens stellt die Integration komplementärer Signalquellen jenseits von Reddit einen vielversprechenden Erweiterungspfad dar. Die kontrastive Methodik ist prinzipiell plattformunabhängig und ließe sich auf weitere spezialisierte Diskursräume, wie beispielsweise Fachforen, GitHub-Repositories oder wissenschaftliche Preprint-Server, übertragen. Eine plattformübergreifende Triangulation von Trendsignalen könnte die Robustheit gegenüber plattformspezifischem Rauschen und koordinierter Manipulation erhöhen. Gleichzeitig würde sie die in Abschnitt 7.2 dokumentierte Anfälligkeit gegenüber Bot-Aktivität adressieren.

Literaturverzeichnis

[1] H. P. Luhn, „A Statistical Approach to Mechanized Encoding and Searching of Literary Information," *IBM Journal of Research and Development*, Jg. 1, Nr. 4, S. 309–317, 1957. DOI: 10.1147/rd.14.0309

[2] J. Allan, *Topic Detection and Tracking: Event-Based Information Organization*. Springer Science & Business Media, 2002, Bd. 12.

[3] T. Dunning, „Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, Jg. 19, Nr. 1, S. 61–74, 1993.

[4] K. Kageura und B. Umino, „Methods of Automatic Term Recognition: A Review," *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, Jg. 3, Nr. 2, S. 259–289, 1996. DOI: 10.1075/term.3.2.03kag

[5] J. Kleinberg, „Bursty and Hierarchical Structure in Streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Ser. KDD '02, New York, NY, USA: Association for Computing Machinery, 2002, S. 91–101. DOI: 10.1145/775047.775061

[6] T. Brown u. a., „Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, Bd. 33, Curran Associates, Inc., 2020, S. 1877–1901.

[7] W. A. Gale und K. W. Church, „Identifying Word Correspondences in Parallel Texts," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19–22, 1991*, 1991.

[8] E. M. Rogers, *Diffusion of Innovations*, Fifth. New York: The Free Press, 2003.

[9] H. Mensah, L. Xiao und S. Soundarajan, „Characterizing the Evolution of Communities on Reddit," in *Proceedings of the International Conference on Social Media and Society*, Ser. SMSociety '20, Toronto, ON, Canada: ACM, Juli 2020, S. 58–64. DOI: 10.1145/3400806.3400814

[10] Reddit, Inc., *Amendment No. 2 to Form S-1 Registration Statement*, U.S. Securities and Exchange Commission, File No. 333-277256, März 2024. Adresse: https://www.sec.gov/Archives/edgar/data/1713445/000162828024011448/reddit-sx1a2.htm

[11] K. Ahmad, L. Gillam und L. Tostevin, „Weirdness Indexing for Logical Document Extrapolation and Retrieval," in *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*, Bd. 82, 2000.

[12] K. W. Church und P. Hanks, „Word Association Norms, Mutual Information, and Lexicography," *Computational Linguistics*, Jg. 16, Nr. 1, S. 22–29, 1990.

[13] G. Bouma, „Normalized (Pointwise) Mutual Information in Collocation Extraction," in *Proceedings of the Biennial GSCL Conference 2009*, Potsdam, Germany, 2009, S. 31–40.

[14] K. Spärck Jones, „A Statistical Interpretation of Term Specificity and Its Application in Retrieval," *Journal of Documentation*, Jg. 28, Nr. 1, S. 11–21, 1972. DOI: 10.1108/eb026526

[15] R. A. Fisher, „On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P," *Journal of the Royal Statistical Society*, Jg. 85, Nr. 1, S. 87–94, 1922. DOI: 10.2307/2340521

[16] H. B. Mann, „Nonparametric Tests Against Trend," *Econometrica*, Jg. 13, Nr. 3, S. 245–259, 1945. DOI: 10.2307/1907187

[17] M. G. Kendall, *Rank Correlation Methods*, Fourth. London: Charles Griffin, 1975.

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi und G. Neubig, „Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, Jg. 55, Nr. 9, S. 1–35, 2023. DOI: 10.1145/3560815

[19] J. S. Justeson und S. M. Katz, „Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text," *Natural Language Engineering*, Jg. 1, Nr. 1, S. 9–27, 1995. DOI: 10.1017/S1351324900000048

[20] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami und C. Dyer, *Neural Architectures for Named Entity Recognition*, 2016. DOI: 10.48550/arXiv.1603.01360 arXiv: 1603.01360 [cs.CL].

[21] M. Kucza, J. Niehues, T. Zenkel, A. Waibel und S. Stüker, „Term Extraction via Neural Sequence Labeling: A Comparative Evaluation of Strategies Using Recurrent Neural Networks," in *Interspeech*, 2018, S. 2072–2076.

[22] A. Rigouts Terryn, V. Hoste, P. Drouin und E. Lefever, „TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset," in *Proceedings of the 6th International Workshop on Computational Terminology*, Marseille, France: European Language Resources Association, 2020, S. 85–94.

[23] A. Hazem, M. Bouhandi, F. Boudin und B. Daille, „TermEval 2020: TALN-LS2N System for Automatic Term Extraction," in *Proceedings of the 6th International Workshop on Computational Terminology*, Marseille, France, 2020, S. 95–100.

[24] E. Amjadian, D. Inkpen, T. S. Paribakht und F. Faez, „Local-Global Vectors to Improve Unigram Terminology Extraction," in *Proceedings of the 5th International Workshop on Computational Terminology*, 2016, S. 2–11.

[25] J. Devlin, M.-W. Chang, K. Lee und K. Toutanova, „BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding“ in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, S. 4171–4186. DOI: 10.18653/v1/N19-1423

[26] T. Sakaki, M. Okazaki und Y. Matsuo, „Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,“ in *Proceedings of the 19th International Conference on World Wide Web*, Ser. WWW '10, New York, NY, USA: Association for Computing Machinery, 2010, S. 851–860. DOI: 10.1145/1772690.1772777

[27] M. Mathioudakis und N. Koudas, „TwitterMonitor: Trend Detection over the Twitter Stream,“ in *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, Ser. SIGMOD '10, New York, NY, USA: Association for Computing Machinery, 2010, S. 1155–1158. DOI: 10.1145/1807167.1807306

[28] Y. Wang und C. Goutte, „Real-time Change Point Detection using On-line Topic Models,“ in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, S. 2505–2515.

[29] D. M. Blei, A. Y. Ng und M. I. Jordan, „Latent Dirichlet Allocation,“ *Journal of Machine Learning Research*, Jg. 3, S. 993–1022, März 2003.

[30] M. S. Mredula, N. Dey, M. S. Rahman, I. Mahmud und Y.-Z. Cho, „A Review on the Trends in Event Detection by Analyzing Social Media Platforms' Data,“ *Sensors*, Jg. 22, Nr. 12, S. 4531, 2022. DOI: 10.3390/s22124531

[31] M. P. Cameron, „Zipf's Law Across Social Media,“ University of Waikato, School of Accounting, Finance und Economics, Hamilton, New Zealand, Working Paper in Economics 7/22, März 2022.

[32] Arctic Shift Project, *Arctic Shift – Reddit Data API*, https://arctic-shift.photon-reddit.com, Accessed: 2025-12-01.

[33] DoltHub, *Wikipedia N-grams Database*, https://www.dolthub.com/repositories/Liquidata/wikipedia-ngrams, Accessed: 2025-02-04.

Übersicht verwendeter Hilfsmittel

Gemäß den Handlungsempfehlungen der Hochschule Neu-Ulm zum Einsatz generativer Künstlicher Intelligenz erkläre ich, dass bei der Erstellung dieser Bachelorarbeit folgende KI-gestützte Hilfsmittel eingesetzt wurden:

Tool	**Anbieter**	**Verwendungszweck**
Claude (Sonnet 4, 4.5, 4.6)	Anthropic	Strategische Beratung bei methodischen Entscheidungen, Strukturierung von Argumentationen sowie sprachliche Überarbeitung von Textpassagen
DeepL Write	DeepL SE	Stilistische und grammatikalische Überarbeitung einzelner Textpassagen

Sämtliche durch KI-gestützte Hilfsmittel generierten oder überarbeiteten Inhalte wurden von mir kritisch geprüft, inhaltlich verantwortet und in die Gesamtargumentation der Arbeit eigenständig eingebettet. Die inhaltliche Verantwortung für alle Aussagen, Schlussfolgerungen und methodischen Entscheidungen liegt ausschließlich bei mir. Die Eigenständigkeit der Arbeit wurde durch den Einsatz dieser Hilfsmittel nicht beeinträchtigt.

Neu-Ulm, 1. März 2026

Johannes Hald

A Technische Spezifikationen

A.1 Entwicklungsumgebung

Die Implementierung erfolgte in Python 3.12.10 auf einem MacBook Air mit Apple M2 Chip (8 CPU-Kerne) und 8 GB RAM unter macOS Tahoe Version 26.3. Die Entwicklungsumgebung wurde als isoliertes Virtual Environment konfiguriert. Um die Reproduzierbarkeit zu gewährleisten, sind sämtliche Package Dependencies mit exakten Versionsnummern in der Datei `requirements.txt` fixiert (Tabelle 7). Die zentrale Datenverarbeitung basiert auf Polars 1.32.3, einer spaltenorientierten DataFrame-Bibliothek. Durch parallele Ausführung und native Parquet-Integration ermöglicht sie eine effiziente Verarbeitung großer Datenmengen.

Tabelle 7: Dependencies mit exakten Versionsnummern

Package	**Version**
polars	1.32.3
numpy	2.3.2
scipy	1.16.1
spacy	3.8.7
en_core_web_sm	3.8.0
praw	7.8.1
prawcore	2.4.0
requests	2.32.4
python-dotenv	1.1.1
openai	1.100.0
xxhash	3.5.0
plotly	6.3.0
streamlit	1.48.1

A.2 Natural Language Processing

Die linguistische Vorverarbeitung für Part-of-Speech-Tagging und Lemmatisierung basiert auf dem spaCy-Framework in der Version 3.8.7. Das verwendete Sprachmodell `en_core_web_sm` Version 3.8.0 wurde auf dem Korpus OntoNotes 5 für englischsprachige Texte trainiert. Zur Effizienzsteigerung werden beim Laden nicht benötigte Komponenten deaktiviert. Die Initialisierung erfolgt durch `spacy.load('en_core_web_sm', disable=["parser", "ner"])`, wodurch ausschließlich Tokenisierung, POS-Tagging und Lemmatisierung aktiviert werden. Das Modell wird über das GitHub Release Repository[2] als Wheel-Datei bezogen und via `pip install` in die Entwicklungsumgebung integriert.

[2] `https://github.com/explosion/spacy-models/releases/download/en_core_web_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl`

A.3 Referenzkorpus

Als allgemeiner Referenzkorpus für die kontrastive Korpusanalyse dient ein Wikipedia-basierter N-Gramm-Korpus, der über das DoltHub Repository [33] bereitgestellt wird. Er umfasst Unigram- und Bigram-Frequenzen aus dem englischsprachigen Wikipedia-Korpus. Die Rohdaten liegen als gzip-komprimierte Comma-Separated Values (CSV)-Dateien im Dolt-Tabellenformat vor (`unigram_counts.csv.gz`: 95 MB, `bigram_counts.csv.gz`: 1.9 GB) und wurden in ein binäres, speichereffizientes Lookup-Format konvertiert.

Die Konvertierung erfolgt durch das Modul `background.py`, das Terme mit dem xxhash64-Algorithmus hasht und in sortierten Binärdateien ablegt. Dadurch wird Lookup über `numpy.memmap` ermöglicht. Das Verfahren unterstützt durch Memory-Mapping und binäre Suche (`np.searchsorted`) Millionen Frequenzabfragen pro Sekunde, ohne den gesamten Korpus in den Arbeitsspeicher laden zu müssen. Ein LRU-Cache mit 20.000 Einträgen beschleunigt wiederholte Abfragen häufig verwendeter Begriffe. Die Korpusstatistik nach Filterung mit `min_count=10` ist in Tabelle 8 dokumentiert.

Tabelle 8: Wikipedia-Referenzkorpus nach Filterung

Metrik	**Unigrams**	**Bigrams**
Unique Entries	1.336.552	9.986.570
Total Occurrences	2.111.968.086	2.218.437.729

A.4 Datenquellen und APIs

A.4.1 Arctic Shift API

Die historische Datenakquisition erfolgt über die Arctic Shift API[32], eine inoffizielle Schnittstelle zum Reddit Archiv. Im Gegensatz zur offiziellen Reddit API ermöglicht sie einen vollständigen, strukturierten Zugriff auf beliebige historische Zeiträume, ohne die Begrenzung auf 1000 Posts. Die Plattform basiert auf Pushshift-Dumps, deren Verfügbarkeit je nach Dataset (Posts, Kommentare, Subreddit-Statistiken) sowie Subreddit variiert. Für die vorliegende Arbeit wurde der Zeitraum 2020 bis 2024 verwendet, da dieser für alle evaluierten Subreddits vollständig verfügbar war. Die API liefert Metadaten im JavaScript Object Notation (JSON)-Format mit folgenden Feldern: `id`, `title`, `created_utc`, `score`, `num_comments`, `author`.

Die Anfragen erfolgen paginiert mit einer Stapelgröße von 100 Beiträgen pro Anfrage. Zwischen aufeinanderfolgenden Anfragen wird eine Ratenbegrenzungsverzögerung von 0.5 Sekunden eingehalten. Der Anfrage Timeout ist auf 30 Sekunden eingestellt.

Um Serverüberlastungen zu vermeiden, ist die maximale Anzahl der Seiten pro Subreddit auf 50 limitiert. Der Endpoint lautet `/api/posts/search` mit den Query Parametern `subreddit`, `after`, `before`, `limit`, `sort` und `fields`.

A.4.2 Reddit PRAW API

Zur Validierung der Subreddits wird die offizielle Reddit API über die PRAW-Bibliothek (Python Reddit API Wrapper) in der Version 7.8.1 genutzt. Die Authentifizierung erfolgt mithilfe von OAuth2-Credentials (Client-ID, Client-Secret, Refresh-Token), die über Umgebungsvariablen bereitgestellt werden. Der User Agent ist als `TrendRadar/1.0` spezifiziert. Bei der Validierung wird geprüft, ob die vom LLM vorgeschlagenen Subreddits existieren, öffentlich zugänglich sind und wie viele Abonnenten sie haben.

A.4.3 OpenAI API

Die semantische Validierung und Kanonisierung erfolgt über die OpenAI Chat Completions API (Version 1.100.0). Für die Community Discovery wird `gpt-4o` verwendet, für die zweistufige LLM-Validierung (Relevanzscoring und Kanonisierung) das kostengünstigere Modell `gpt-4o-mini`. Zur Sicherstellung der Reproduzierbarkeit werden die Konfigurationen `seed=42` und `temperature=0.0` verwendet, wodurch eine deterministische Argmax-Dekodierung erzwungen wird. Die strukturierte Ausgabe wird durch Constrained Decoding mit `response_format={"type: "json_object"}` garantiert, das nur Outputs zulässt, die valides JSON darstellen. Trotz dieser Konfiguration ist absolute Reproduzierbarkeit bei proprietären LLM-APIs nicht garantiert (siehe Abschnitt A.5).

A.4.4 Wikipedia Pageviews API

Die Ground Truth Konstruktion für die retrospektive Evaluation basiert auf der Wikimedia Representational State Transfer (REST)-API[3], die tägliche Pageview-Statistiken für Wikipedia Artikel bereitstellt. Die Abfragen erfolgen für den Zeitraum November 2020 bis Dezember 2024 mit einem Anfrage Timeout von 10 Sekunden und einer Ratenbegrenzungsverzögerung von 0.1 Sekunden. Der User Agent ist als `TrendRadar-Evaluation/1.0` spezifiziert. Die API liefert JSON-Objekte mit Zeitstempel und Pageview Anzahl. Diese werden durch einen Forward Scan Algorithmus mit kombinierten Momentum- und Relevanzprüfungen algorithmisch verarbeitet, um das Mainstream Breakthrough Datum zu bestimmen (siehe Abschnitt 6.1.1).

[3] `https://wikimedia.org/api/rest_v1/metrics/pageviews`

A.5 Reproduzierbarkeit

Die Reproduzierbarkeit wird durch die deterministische Konfiguration aller stochastischen Systemkomponenten gewährleistet. Die Spezifikation von `seed=42` und `temperature=0.0` erzwingt deterministische Argmax-Dekodierung bei der LLM-Inferenz. Alle Python Abhängigkeiten sind mit exakten Versionsnummern in der Datei `requirements.txt` fixiert. Über die Arctic Shift API werden für identische Zeitfenster und Subreddits deterministische Ergebnisse geliefert, da die Plattform keine Stichprobenziehung durchführt.

Die Determinismusgarantien proprietärer LLM-APIs sind eingeschränkt. OpenAI dokumentiert, dass identische Prompts mit deterministischer Konfiguration in seltenen Fällen abweichende Ergebnisse produzieren können, insbesondere bei Aktualisierungen des Modells oder Änderungen der Infrastruktur. Dies bedeutet, dass exakte Reproduzierbarkeit bei der Identifikation relevanter Subreddits (Abschnitt 4.2.1) und der semantischen Validierung (Abschnitt 4.5) nicht absolut garantiert werden kann.

Die vollständige Codebasis ist unter folgender URL verfügbar:

`https://github.com/johanneshald/trndrdr`

Zur Replikation der Ergebnisse müssen zunächst die Dependencies aus `requirements.txt` installiert und das spaCy-Modell `en_core_web_sm` Version 3.8.0 via `python -m spacy download en_core_web_sm` heruntergeladen werden. Die Authentifizierung erfolgt über Umgebungsvariablen für Reddit (`REDDIT_CLIENT_ID`, `REDDIT_CLIENT_SECRET`, `REDDIT_REFRESH_TOKEN`) und OpenAI (`OPENAI_API_KEY`). Die vorverarbeiteten N-Gramm-Frequenzen des Wikipedia-Referenzkorpus[33] liegen als binäre Lookup-Dateien via Git Large File Storage (LFS) im Repository bereit und werden automatisch beim Klonen heruntergeladen, sofern Git LFS installiert ist.[4] Die detaillierte Installationsanleitung ist in der Datei `README.md` des Repositories dokumentiert.

[4] `https://git-lfs.com`

B Parameter und Schwellwerte

Die nachfolgenden Tabellen konsolidieren sämtliche konfigurierbaren Parameter und Schwellwerte des Systems. Die Werte sind im Methodikteil (Kapitel 4) wissenschaftlich begründet und werden hier als Referenz für die Reproduzierbarkeit zusammengefasst. Die zugehörigen POS-Muster für die syntaktische Filterung sind in Tabelle 2 dokumentiert.

B.1 Datenakquisition und Korpuskonstruktion

Tabelle 9: Parameter für Datenakquisition und Korpuskonstruktion

Parameter	**Wert**	**Beschreibung / Referenz**
Beobachtungszeitraum	60 Tage	Optimaler Kompromiss zwischen Stabilität und Signalfrische (RQ3, Abschnitt 6.2.4)
Min. Aktivitätsanteil	40 %	Anteil aktiver Tage pro Subreddit im Beobachtungszeitraum [9]
Gewichtungsintervall	$[1, 10]$	Normalisierungsbereich der inversen logarithmischen Gewichte
LLM (Community Discovery)	GPT-4o	temperature=0.0, seed=42
API-Batchgröße	100	Posts pro Arctic Shift Anfrage
Rate Limit Delay	0,5 s	Verzögerung zwischen API-Anfragen
Request Timeout	30 s	Maximale Wartezeit pro Anfrage
Max. Seiten pro Subreddit	50	Obergrenze zur Vermeidung von Serverüberlastung

B.2 Phase 1: Termextraktion

Tabelle 10: Parameter für Vorfilterung und Unithood

Parameter	**Wert**	**Beschreibung / Referenz**
Min. Unigramm-Frequenz	15	Absolute Mindesthäufigkeit im Korpus [4]
Min. Bigramm-Frequenz	5	Absolute Mindesthäufigkeit im Korpus [4]
Min. Dokumentfrequenz	3	Mindestanzahl unterschiedlicher Posts [11]
Max. Dokumentfrequenzanteil	30 %	Obergrenze zur Elimination allgemeinsprachlicher Terme [11]
Min. NPMI	0,4	Kohäsionsschwelle für Bigramme [13]

Tabelle 11: Parameter für Termhood und Scoring

Parameter	**Wert**	**Beschreibung / Referenz**
Min. LLR (G^2)	10,83	99,9. Perzentil der χ^2-Verteilung ($p < 0{,}001$, $df = 1$) [3]
Min. Weirdness	100	Empirischer Schwellwert für Domänenspezifität [11]
Unigramm-Scoring-Gewichte:		
LLR	50 %	Primärmaß: statistische Robustheit
Weirdness	30 %	Sekundärmaß: Effektgröße
IDF	20 %	Dokumentstreuung [14]
Bigramm-Scoring-Gewichte:		
LLR	40 %	Domänenspezifität
NPMI	40 %	Linguistische Kohäsion
IDF	20 %	Dokumentstreuung
Min. normalisierter Score	10	Eliminationsschwelle nach Normalisierung auf $[0, 100]$

B.3 Phase 2 und 3: Zeitreihenanalyse und semantische Validierung

Tabelle 12: Parameter für die Zeitreihenanalyse

Parameter	**Wert**	**Beschreibung / Referenz**
Recent Window (T_{recent})	12 Tage	Zeitfenster für Burst-Detektion
Past Window (T_{past})	48 Tage	Historisches Vergleichsfenster
Max. p-Wert (Fisher)	0,05	Signifikanzniveau [2]
Min. Kendall's τ	0,0	Ausschluss fallender Trends [5]
Min. Autoren	2	Qualitätsfilter: Diskursdiversität
Min. Erwähnungen	5	Qualitätsfilter: Mindestaktivität
Scoring-Gewichte (Abschnitt 6.2.6):		
Wachstumsrate	45 %	Zentrale Eigenschaft emergenter Trends
Signifikanz	30 %	Statistischer Anker
Konsistenz	25 %	Nachgelagertes Qualitätskriterium

Tabelle 13: Parameter für die semantischen Validierung

Parameter	**Wert**	**Beschreibung / Referenz**
LLM (Pass 1 und 2)	GPT-4o-mini	temperature=0.0, seed=42
Ausgabeformat	JSON	Constrained Decoding via `response_format`
Top-k (deterministisch)	15	Höchstbewertete Terme für LLM Pass 1
Perzentilschwelle	P80	Kandidaten oberhalb des 80. Perzentils
Max. Kandidaten (Pass 1)	40	Absolute Obergrenze für LLM-Evaluation
Kontextposts pro Term	10	Höchstbewertete Posts als In-Context-Beispiele
Min. Relevanzscore	0,2	Eliminationsschwelle für Pass 1
Adaptiver Cutoff	Q75 / Median	Q75 wenn $(Q75 - \text{Median}) > 10$, sonst Median
Scorefusion	multiplikativ	$S_{\text{final}} = S_{\text{statistical}} \times r_{\text{relevance}}$

B.4 Evaluation

Tabelle 14: Parameter für Ground Truth und Evaluation

Parameter	**Wert**	**Beschreibung / Referenz**
Breakthrough-Detektion (Abschnitt 6.1.1):		
Rolling Mean (kurz)	7 Tage	Kurzfristiges Momentum-Fenster
Baseline (lang)	30 Tage	Langfristiges Referenzfenster
Rise Factor	1,5	Min. Verhältnis kurz/lang
Min. Baseline Views	300	Absoluter Relevanzschwellwert
Peak Fraction	5 %	Anteil des globalen Peaks
Persistenzdauer	3 Tage	Nachhaltigkeitsprüfung
Persistenzschwelle	70 %	Min. Anteil des erreichten Niveaus
Simulationsdesign:		
Simulationsfenster	7, 21, 30, 45, 60, 90	Tage vor Breakthrough-Datum
Ground Truth Trends	45	15 pro Domäne (3 Domänen)
Evaluationszeitraum	Nov. 2020 – Dez. 2024	Wikipedia Pageviews
Matching (Abschnitt 6.1.3):		
Jaccard-Schwellwert	0,4	Min. Token-Overlap für Matching

B.5 Ground Truth Datensatz

Tabelle 15 dokumentiert die 45 historischen Trends des Evaluationsdatensatzes mit dem algorithmisch bestimmten Mainstream-Breakthrough-Datum (vgl. Abschnitt 6.1.1) sowie dem Detektionsstatus des Gesamtsystems. Die Detektionsspalte gibt an, ob der Trend in mindestens einem der sechs Simulationsfenster korrekt identifiziert wurde.

Tabelle 15: Ground Truth: 45 historische Trends mit Breakthrough Datum

Trend	**Wikipedia-Artikel**	**Breakthrough**	**Det.**
Artificial Intelligence			
AlphaFold	AlphaFold	2020-12-11	✓
Bard	Bard_(chatbot)	2023-04-01	✓
ChatGPT	ChatGPT	2023-01-11	✓
Claude	Claude_(language_model)	2024-02-29	×
DALL-E	DALL-E	2022-04-08	×
Gemini	Gemini_(language_model)	2024-02-09	✓
Generative AI	Generative_artificial_intelligence	2023-05-05	✓
GPT-4	GPT-4	2023-03-12	×
Llama	Llama_(language_model)	2024-05-17	✓
Midjourney	Midjourney	2022-08-02	✓
Mistral AI	Mistral_AI	2024-01-24	×
OpenAI	OpenAI	2022-12-03	×
Runway	Runway_(company)	2024-02-21	×
Sora	Sora_(text-to-video_model)	2024-05-01	✓
Stable Diffusion	Stable_Diffusion	2022-10-07	✓
Cryptocurrency & Web3			
Bitcoin	Bitcoin	2021-01-03	✓
Celsius Collapse	Celsius_Network	2022-06-13	×
Coinbase	Coinbase	2020-12-20	✓
Dogecoin	Dogecoin	2021-01-29	×
Ethereum Merge	Ethereum	2021-01-03	×
FTX Collapse	FTX	2022-12-27	✓
Kraken	Kraken_(company)	2021-01-03	✓
OpenSea	OpenSea	2021-12-18	×
Polkadot	Polkadot_(cryptocurrency)	2021-02-10	✓
SafeMoon	SafeMoon	2021-08-29	✓
Shiba Inu	Shiba_Inu_(cryptocurrency)	2021-10-05	×
Solana	Solana	2021-09-03	×
Terra Luna Collapse	Terra_(blockchain)	2022-09-10	×
Tornado Cash	Tornado_Cash	2022-09-15	✓
Uniswap	Uniswap	2021-01-04	×
Gaming			
Baldur's Gate 3	Baldur's_Gate_3	2023-08-08	✓
Cyberpunk 2077	Cyberpunk_2077	2020-12-08	✓
Dead Space Remake	Dead_Space_(2023_video_game)	2022-10-04	×
Diablo IV	Diablo_IV	2021-02-20	×
Elden Ring	Elden_Ring	2021-06-10	×
Halo Infinite	Halo_Infinite	2021-06-14	×
Helldivers 2	Helldivers_2	2024-02-08	×
Hogwarts Legacy	Hogwarts_Legacy	2022-03-17	×
Lethal Company	Lethal_Company	2024-06-02	×
Overwatch 2	Overwatch_2	2022-05-01	×
Spider-Man 2	Spider-Man_2_(2023)	2023-05-25	×
Starfield	Starfield_(video_game)	2022-06-13	×
Steam Deck	Steam_Deck	2021-10-08	×
Street Fighter 6	Street_Fighter_6	2022-06-12	×
Tears of the Kingdom	TotK	2023-02-08	×

C Quellcode

Dieser Appendix dokumentiert ausgewählte Implementierungsdetails der methodisch kritischen Systemkomponenten. Die vollständige Codebasis ist unter `https://github.com/johanneshald/trndrdr` verfügbar. Die Listings konzentrieren sich auf Algorithmen, deren präzise Operationalisierung für die Reproduzierbarkeit der Ergebnisse essenziell ist und die zentrale methodische Entscheidungen aus Kapitel 4 direkt implementieren.

Die Docstrings und LLM-Prompts wurden zur Platzoptimierung auf das wissenschaftlich Wesentliche gekürzt, wobei alle methodischen Referenzen und Parameter vollständig erhalten bleiben. Weitere Komponenten sind vollständig im Repository dokumentiert.

C.1 LLM-gestützte Community Discovery

Listing 1 zeigt die Identifikation relevanter Subreddits (Abschnitt 4.2.1), die abstrakte Keywords in spezialisierte Communities übersetzt und Luhns Konzept des Inquirer's Document [1] operationalisiert.

Listing 1: LLM-gestützte Subreddit Discovery (`data.py`)

```python
def discover_subreddits(keywords: List[str]) -> List[str]:
    """Subreddit discovery via LLM (Luhn, 1957).

    Translates abstract keywords into specialized communities
    using LLM world knowledge. Operationalizes inquirer's
    document concept.

    Args:
        keywords: User's interest keywords

    Returns:
        List of relevant subreddit names
    """
    # Structured prompt for community discovery
    prompt = f"""For each keyword, list relevant Reddit subreddits.

Keywords: {", ".join(keywords)}

Requirements:
- Relevant to keyword
- Knowledge/discussion focused (not memes)
- Sustained activity (not dead)
- Primarily English

Output format:
# keyword1
- r/subreddit1
- r/subreddit2

# keyword2
..."""

    response = _query_llm(prompt, model="gpt-4o")

    # Extract subreddit mentions
    matches = re.findall(r"r/[a-zA-Z0-9_]+", response)

    # Deduplicate while preserving order
    return list(dict.fromkeys(matches))
```

C.2 Inverse Gewichtung von Subreddits (RQ2)

Listing 2 zeigt die Implementierung der inversen logarithmischen Gewichtung (Abschnitt 4.2.1), die Rogers' Diffusionstheorie [8] operationalisiert und in Forschungsfrage 2 (Abschnitt 6.2.3) evaluiert wird.

Listing 2: Inverse logarithmische Gewichtung von Subreddits (`data.py`)

```python
def _calculate_weights(
    sizes: Dict[str, int],
    inverse: bool = True
) -> Dict[str, float]:
    """Subreddit weighting via inverse log transform (Luhn, 1957).

    Smaller communities receive higher weight when inverse=True.
    Uniform weights when inverse=False (RQ2 baseline).

    Args:
        sizes: Subreddit -> subscriber count mapping
        inverse: Apply inverse weighting or uniform

    Returns:
        Subreddit -> weight [1.0, 10.0] mapping
    """
    if not sizes:
        return {}

    if not inverse:
        return {sub: 1.0 for sub in sizes}

    # Inverse logarithmic weighting
    max_size = max(sizes.values())
    weights = {
        sub: np.log(max_size / size + 1)
        for sub, size in sizes.items()
    }

    # Normalize to [1, 10] range
    min_w, max_w = min(weights.values()), max(weights.values())
    if max_w == min_w:
        return {k: 5.0 for k in weights}

    return {
        k: 1 + 9 * (v - min_w) / (max_w - min_w)
        for k, v in weights.items()
    }
```

C.3 Log-Likelihood Ratio Berechnung (Phase 1)

Listing 3 dokumentiert die kontrastive Korpusanalyse (Abschnitt 4.3.3), die als Primärmaß für Termhood dient und die statistische Signifikanz der Frequenzdifferenz zwischen Ziel- und Referenzkorpus quantifiziert [3].

Listing 3: Log-Likelihood Ratio für kontrastive Korpusanalyse (`extraction.py`)

```
def _calculate_llr(
    obs_freq: int,
    obs_total: int,
    bg_freq: int,
    bg_total: int
) -> float:
    """Log-Likelihood Ratio for Termhood (Dunning, 1993).

    Compares Reddit vs. Wikipedia frequency to measure
    domain-specificity. Robust for low-frequency terms.

    Args:
        obs_freq, obs_total: Observed corpus stats
        bg_freq, bg_total: Background corpus stats

    Returns:
        G2 LLR statistic (float)
    """
    # Build 2x2 contingency table
    a, b = obs_freq, bg_freq
    c, d = max(1, obs_total - a), max(1, bg_total - b)
    n = a + b + c + d
    if n == 0:
        return 0.0

    # Calculate expected frequencies
    e1 = (a + b) * (a + c) / n
    e2 = (a + b) * (b + d) / n
    e3 = (c + d) * (a + c) / n
    e4 = (c + d) * (b + d) / n

    # G2 = 2 * sum(observed * log(observed / expected))
    g2 = 0.0
    for obs, exp in [(a, e1), (b, e2), (c, e3), (d, e4)]:
        if obs > 0 and exp > 0:
            g2 += obs * log(obs / exp)
    return 2 * g2
```

C.4 NPMI für Unithood-Filterung (Phase 1)

Listing 4 zeigt die Kohäsionsmessung für Bigramme (Abschnitt 4.3.2), die linguistisch stabile Kollokationen von zufälligen Nachbarschaften unterscheidet [13].

Listing 4: Normalized Pointwise Mutual Information (`extraction.py`)

```python
def _calculate_npmi(
    bigram_freq: int,
    w1_freq: int,
    w2_freq: int,
    total: int
) -> float:
    """NPMI for Unithood (Church & Hanks, 1990; Bouma, 2009).

    Measures bigram cohesion. Normalizes PMI to [-1, 1],
    where +1 indicates perfect correlation.

    Args:
        bigram_freq: Frequency of (w1, w2)
        w1_freq, w2_freq: Individual word frequencies
        total: Total unigrams for probability calculation

    Returns:
        NPMI score (float)
    """
    if (bigram_freq == 0 or w1_freq == 0 or
        w2_freq == 0 or total == 0):
        return 0.0

    p_bigram = bigram_freq / total
    p_w1, p_w2 = w1_freq / total, w2_freq / total

    # PMI = log(P(w1,w2) / (P(w1) * P(w2)))
    pmi = log(p_bigram / (p_w1 * p_w2))

    # Normalize to [-1, 1] range
    return pmi / -log(p_bigram)
```

C.5 Syntaktische Termfilterung (Phase 1)

Listing 5 definiert die zulässigen POS-Muster (Abschnitt 4.3.2), die Justeson & Katz' Befund operationalisieren, dass 97 % der Fachterme aus Nomen und Adjektiven bestehen [19].

Listing 5: Zulässige POS-Muster für Bigramme (`extraction.py`)

```python
# Valid POS patterns for bigrams (Justeson & Katz, 1995)
VALID_BIGRAM_POS = {
    ("JJ", "NN"), ("JJ", "NNS"),       # Adjective + Noun
    ("NN", "NN"), ("NN", "NNS"),       # Noun + Noun
    ("NNS", "NN"),
    ("NNP", "NNP"),                    # Proper Noun + Proper
    ("VBG", "NN"), ("VBG", "NNS"),     # Gerund + Noun
    ("VBN", "NN"), ("VBN", "NNS"),     # Participle + Noun
    ("NN", "VBG"),                     # Noun + Gerund
    ("CD", "NN"), ("CD", "NNS"),       # Cardinal + Noun
}

def _is_valid_bigram_pos(pos_sequence: Tuple[str, str]) -> bool:
    """Validates bigram POS sequence (Justeson & Katz, 1995).

    Justeson & Katz found 97% of technical terms follow
    (ADJ/NOUN) + NOUN patterns. Extended for domain-specific
    patterns (gerunds, cardinals, proper nouns).

    Args:
        pos_sequence: Tuple of (POS1, POS2)

    Returns:
        True if pattern is valid
    """
    return pos_sequence in VALID_BIGRAM_POS
```

Listing 6 zeigt die Anwendung der POS-Filter während der N-Gramm-Extraktion.

Listing 6: N-Gramm-Extraktion mit POS-Filterung (`extraction.py`)

```
def _extract_ngrams_with_pos(texts: List[str]) -> Tuple:
    """Extracts POS-tagged n-grams with frequency counts.

    Tokenizes, tags, and filters based on syntactic patterns.

    Args:
        texts: List of cleaned post titles

    Returns:
        (unigram_freqs, bigram_freqs, bigram_pos_map)
    """
    unigram_freqs = defaultdict(int)
    bigram_freqs = defaultdict(int)
    bigram_pos_map = {}

    for doc_id, text in enumerate(texts):
        doc = NLP(text)
        tokens = [
            (t.text.lower(), t.tag_)
            for t in doc
            if t.is_alpha and len(t.text) > 2
            and t.text.lower() not in STOPWORDS
        ]

        # Extract unigrams
        for word, pos in tokens:
            unigram_freqs[word] += 1

        # Extract bigrams with POS validation
        for i in range(len(tokens) - 1):
            w1, pos1 = tokens[i]
            w2, pos2 = tokens[i + 1]
            pos_seq = (pos1, pos2)

            # Apply Justeson & Katz filter
            if pos_seq in VALID_BIGRAM_POS:
                bigram_freqs[(w1, w2)] += 1
                bigram_pos_map[(w1, w2)] = pos_seq

    return (unigram_freqs, bigram_freqs, bigram_pos_map)
```

C.6 Fisher's Exact Test für Burst Detection (Phase 2)

Listing 7 zeigt die Signifikanzprüfung (Abschnitt 4.4.1) zur Identifikation auffälliger Aktivitätsanstiege, die das TDT-Prinzip der Violation of Stationarity operationalisiert [2].

Listing 7: Fisher's Exact Test für Burst-Detektion (`analysis.py`)

```python
def _calculate_significance(
    term_counts: np.ndarray,
    total_counts: np.ndarray,
    recent_window: int = RECENT_WINDOW,
) -> Tuple[float, float]:
    """Fisher's Exact Test (TDT, Allan).

    Tests if recent frequency significantly exceeds past
    (violation of stationarity).

    Args:
        term_counts, total_counts: Daily engagement arrays
        recent_window: Recent vs. past days (default: 12)

    Returns:
        (p-value, odds-ratio)
    """
    # Construct 2x2 contingency table
    term_recent = int(np.sum(term_counts[-recent_window:]))
    term_past = int(np.sum(term_counts[:-recent_window]))
    other_recent = (
        int(np.sum(total_counts[-recent_window:])) - term_recent
    )
    other_past = (
        int(np.sum(total_counts[:-recent_window])) - term_past
    )

    if term_recent == 0:
        return 1.0, 0.0

    contingency_table = [
        [term_recent, term_past],
        [other_recent, other_past]
    ]

    try:
        result = fisher_exact(contingency_table, alternative="greater")
        return result.pvalue, result.statistic
    except ValueError:
        return 1.0, 0.0
```

C.7 Mann-Kendall Test für Trendkonsistenz (Phase 2)

Listing 8 dokumentiert die Konsistenzprüfung (Abschnitt 4.4.2), die robuste Aufwärtstrends von transienten Spikes unterscheidet [5].

Listing 8: Mann-Kendall Test für Trendkonsistenz (`analysis.py`)

```python
def _mann_kendall_trend(counts: np.ndarray) -> float:
    """Mann-Kendall test for monotonic trend (Kleinberg, 2002).

    Measures trend consistency via Kendall's Tau.
    Distinguishes steady rise from volatile burst.
    Robust to outliers, no distributional assumptions.

    Args:
        counts: Daily engagement array

    Returns:
        Kendall's Tau (float), 0.0 if insufficient data
    """
    # Require minimum data quality
    if (np.sum(counts) < 10 or
        len(np.where(counts > 0)[0]) < 3):
        return 0.0

    # Calculate Kendall's Tau correlation
    tau, _ = kendalltau(
        x=np.arange(len(counts)),
        y=counts
    )

    return tau if not np.isnan(tau) else 0.0
```

C.8 Multifaktorielles Scoring (Phase 2)

Listing 9 zeigt die gewichtete Kombination der drei Zeitreihenmetriken (Abschnitt 4.4.3), die in Forschungsfrage 3 evaluiert wird.

Listing 9: Multifaktorielles Scoring der Zeitreihenmetriken (`analysis.py`)

```
# Multi-Factor Scoring (from detect_trends function)

# Calculate individual metric scores
growth_rates = np.array([c["growth"] for c in qualified_candidates])
max_g = np.max(growth_rates) if growth_rates.size > 0 else 1

for c in qualified_candidates:
    # Normalize growth to [0, 100]
    c["growth_score"] = (c["growth"] / max(max_g, 1e-6)) * 100

    # Consistency score from Kendall's Tau
    c["consistency_score"] = max(0, c["trend"]) * 100

    # Significance score inverse to p-value
    c["significance_score"] = (1 - c["p_value"]) * 100

    # Weighted combination (Abschnitt 4.5.3)
    c["statistical_score"] = (
        scoring_weights["growth"] * c["growth_score"]
        + scoring_weights["consistency"] * c["consistency_score"]
        + scoring_weights["significance"] * c["significance_score"]
    )

# Default weights: theoretically motivated, validated via ablation
SCORING_WEIGHTS = {
    "growth": 0.45,         # Central property
    "consistency": 0.25,    # Secondary criterion
    "significance": 0.3     # Statistical anchor
}
```

C.9 LLM-basierte Relevanzvalidierung (Phase 3)

Listing 10 dokumentiert Pass 1 der semantischen Filterung (Abschnitt 4.5.2) mit strukturiertem Prompt und deterministischer Konfiguration, die ICL operationalisiert [6].

Listing 10: Semantische Relevanzvalidierung, Pass 1 (`analysis.py`)

```
def _get_llm_relevance(
    term: str,
    keywords: List[str],
    reddit_context: str,
    model: str = "gpt-4o-mini",
    client=None,
) -> float:
    """Semantic relevance via ICL (Brown et al., 2020).

    Filters statistically significant but semantically
    irrelevant noise. Addresses RQ1.

    Args:
        term, keywords, reddit_context: Input data
        model, client: LLM configuration

    Returns:
        Relevanzscore [0.0, 1.0]
    """
    if client is None:
        client = _LLM_CLIENT

    user_prompt = f"""INTERESTS: {", ".join(keywords)}
TERM: "{term}"
CONTEXT: {reddit_context}

Rate relevance [0.0-1.0]:
0.9-1.0: Core concept | 0.7-0.89: Domain tool/tech
0.5-0.69: Method/practice | 0.2-0.49: Peripheral
0.0-0.19: Noise/unrelated

JSON: {{"relevance_score": <float>}}"""

    try:
        response = client.chat.completions.create(
            model=model,
            seed=42,
            temperature=0.0,
            messages=[{"role": "user", "content": user_prompt}],
            response_format={"type": "json_object"}
        )
        result = json.loads(response.choices[0].message.content)
        return max(0.0, min(1.0, result.get("relevance_score", 0.0)))
    except (APIError, json.JSONDecodeError):
        return 0.0
```

C.10 LLM-basierte Kanonisierung (Phase 3)

Listing 11 zeigt Pass 2 der semantischen Filterung (Abschnitt 4.5.3), der generische Terme in kanonische Formen transformiert.

Listing 11: Semantische Kanonisierung, Pass 2 (`analysis.py`)

```
def _get_llm_canonical_term(
    term: str,
    keywords: List[str],
    reddit_context: str,
    model: str = "gpt-4o-mini",
    client=None,
) -> str:
    """Canonical term transformation via LLM.

    Transforms raw term into canonical form
    (e.g., dao -> Decentralized Autonomous Organization (DAO)).

    Args:
        term, keywords, reddit_context: Input data
        model, client: LLM configuration

    Returns:
        Canonical term string
    """
    if client is None:
        client = _LLM_CLIENT

    user_prompt = f"""INTERESTS: {", ".join(keywords)}
TERM: "{term}"
CONTEXT: {reddit_context}

Return most specific canonical form:
- Generic: Add context (supply -> Ethereum Supply)
- Acronyms: Expand + keep short (dao -> DAO (Decentralized...))
- Clear: Unchanged

JSON: {{"canonical_term": "..."}}"""

    try:
        response = client.chat.completions.create(
            model=model,
            seed=42,
            temperature=0.0,
            messages=[{"role": "user", "content": user_prompt}],
            response_format={"type": "json_object"}
        )
        result = json.loads(response.choices[0].message.content)
        canonical = result.get("canonical_term", term)
        return result.get("canonical_term", term)
    except (APIError, json.JSONDecodeError):
        return term
```