Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Mediensysteme

Retrieval-Modelle zum Filtern, Ranken und Zusammenfassen von Web-Kommentaren

Bachelorarbeit

Fabian Loose

Steffen Becker

1. Gutachter: Prof. Benno Stein

Betreuer: Dipl. Inf. Martin Potthast

Datum der Abgabe: 14. Oktober 2008

Erklärung der Selbstständigkeit

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen worden ist. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Weimar, 13. Oktober 2008 Steffen Becker

Weimar, 13. Oktober 2008 Fabian Loose

Kurzfassung

Benutzergenerierte Inhalte spielen im WWW eine immer größere Rolle. Eine Form davon stellt auf vielen Web-Portalen die Möglichkeit dar, Kommentare abzugeben. Dabei kommt das Problem auf, dass oft mehr Kommentare abgegeben wurden, als vom Benutzer in adäquater Zeit gelesen werden könnten. In dieser Arbeit werden Retrieval-Aufgaben zur Lösung des Problems der aufkommenden Flut von Kommentaren identifiziert: *Filterung*, *Zusammenfassung* und *Ranking*. Für die Bearbeitung dieser Aufgaben wurden drei Relevanzkriterien *thematische Relevanz*, *Polarität der Meinungsäußerung* und *Textqualität* herausgearbeitet. Diese Kriterien bilden die Grundlage für die Entwicklung von Retrieval-Modellen für Kommentare. Für die Berechnung thematischer Relevanz vergleichen wir bekannte Maße des textbasierten Information-Retrieval und stellen ein neues Maß vor, das entwickelt wurde, um die thematische Vervollständigung eines Textes durch einen anderen zu messen: das *Kontinuitätsmaß*. In Experimenten wird demonstriert, dass das Kontinuitätsmaß insbesondere eine gute Basis für das Ranking von Kommentaren sein kann. Weiterhin wird eine neue Anwendung vorgestellt, die eine Zusammenfassung aller Meinungsäußerungen der Kommentare zu einem Thema generiert und in Form einer dafür entwickelten Opinion-Cloud darstellt. Für das Relevanzkriterium Textqualität untersuchen wir die Anwendbarkeit einer Stilanalyse. Dabei hat sich herausgestellt, dass sich die Stilanalyse prinzipiell als Relevanzmaß eignet, jedoch schlecht auf kurzen Kommentaren skaliert und somit insbesondere als Ergänzung eines anderen Maßes in Frage kommt.

Inhaltsverzeichnis

Fabian Loose	**Steffen Becker**
1 Einleitung	1 Einleitung
	1.1 Retrieval Szenario
1.2 Verwandte Arbeiten	1.2 Verwandte Arbeiten
2 Modelle für das Retrieval	2 Modelle für das Retrieval
2.1 Modelle zur Berechnung thematischer Relevanz	
	2.2 Meinungsanalyse
2.3 Qualitative Analyse	
3 Ergebnisse	3 Ergebnisse
3.1 Korpora	3.1 Korpora
3.1.1 Slashdot-Korpus	
	3.1.2 YouTube
3.2 Evaluierung der Themen-Relevanzmodelle	
	3.3 Fallstudie zur Meinungsanalyse
3.4 Evaluierung der Stilanalyse	
4 Zusammenfassung und Ausblick	4 Zusammenfassung und Ausblick
	A.1 Firefox-Erweiterung: Opinion-Cloud
	A.2 YouTube-Crawler und Wörterbucherweiterung
A.3 Slashdot-Crawler und Modelle	

1 Einleitung

Mit voranschreitender Entwicklung des WWW, insbesondere seit der Prägung des Begriffes *Web 2.0*, spielen benutzergenerierte Inhalte eine immer größere Rolle. Bekannte Beispiele sind Foren, Blogs, Gästebücher, Video/Foto/Link-Communities etc. Eine weitere Form der inhaltlichen Gestaltung von Web-Portalen durch die Benutzer stellt die Möglichkeit dar, Kommentare auf Web-Seiten abzugeben. Hier wird es Benutzern ermöglicht, anders als bei klassischen Medien wie Rundfunk und Zeitungen, unmittelbares Feedback zu geben. Interessant können diese Kommentare sowohl für den Autor bzw. Betreiber der Webseite, als auch für die Besucher sein. Jedoch sind nicht alle Kommentare von Interesse, es kommt zu Wiederholungen und auch zu unerwünschten Kommentaren wie Spam. Des Weiteren macht es die oftmals sehr große Anzahl von Kommentaren zu zeitaufwändig, alle zu lesen. Die vorliegende Arbeit beschäftigt sich mit der Frage, wie es möglich ist, in der Menge von Kommentaren eine Übersicht für den Benutzer herzustellen und stellt konkrete Retrieval-Modelle zur Erfüllung dieser Aufgabe vor.

1.1 Retrieval Szenario

Die Hauptaufgabe eines Retrievals auf Kommentaren einer Webseite besteht darin dem Benutzer (Leser, Autor) in übersichtlicher Form, das zu geben, was ihn interessiert und was er lesen will. Zur Konkretisierung betrachten wir folgendes Szenario:

Der Autor einer Webseite stellt ein Thema in Form von Text, Bildern oder Videos vor. Die Leser werden eingeladen, das Vorgestellte zu kommentieren. Beispiele hierfür sind Slashdot[1], mit aktuellen Themen des IT-Bereichs oder auch YouTube[2], wo Videos vorgestellt und kommentiert werden. Gerade auf diesen Seiten ist die Flut von Kommentaren riesig. Es kommt vor, dass hunderte bis tausende Kommentare zu einem Thema abgegeben werden. In dieser Situation stellen sich jedem Besucher der Webseite folgende Probleme:

- Alle Kommentare können nicht mehr gelesen und erfasst werden.
- Nicht alle Kommentare interessieren den Leser.

[1] http://www.slashdot.com
[2] http://www.youtube.com

- Vandalismus in Form von *Spam* - als unerwünschte Werbung, und *Trolling* - gezielte Provokation ohne Beitrag zum Thema.
- Wiederholung der Themenbeiträge der Kommentare

Aufgabenstellung. Um Übersicht über große Kommentarmengen zu schaffen, werden Kommentare als erstes gefiltert, um so unerwünschte bzw. nicht lesenswerte Kommentare auszuschließen. Kommentare ohne Relevanz sowie Vandalismus werden dem Leser nicht präsentiert. Die Filterung von Kommentaren ist auch für den Autor bzw. Betreiber der Internetseite von großer Bedeutung. Anstößige und verbotene Inhalte in einem Kommentar können rechtliche Konsequenzen für den Seitenbetreiber haben, auch wenn er nicht Autor des entsprechenden Kommentars ist. Störende Inhalte, wie Vandalismus können der Glaubwürdigkeit der Seite schaden und somit die Leserschaft vertreiben, ohne dass der Betreiber darauf Einfluss hat. Vor allem Betreiber von Nachrichtenseiten filtern Kommentare mühsam von Hand.

Die verbleibenden Kommentare werden als nächstes nach Relevanz geordnet und dem Leser in absteigender Reihenfolge präsentiert. So wird jedem Leser das wichtigste zuerst präsentiert und er kann selbst entscheiden, bis wohin ihn Kommentare in der Liste interessieren. Bei den meisten Webseiten werden Kommentare chronologisch sortiert, was kein Anhaltspunkt für Relevanz ist. Für den Leser ist nicht ersichtlich, ob in der Liste zu einem anderen Zeitpunkt ein lesenswerter Kommentar steht ohne alle Kommentare zu lesen.

Des Weiteren werden Kommentare, die gleiches Aussagen zusammengefasst. Sehr kurze Kommentare, die beispielsweise Antworten der Leser zu einer gestellten Frage darstellen, gleichen sich oft sehr stark. Ein einzelner Kommentar stellt dabei oft keine neuen Information für den Leser dar. Alle Kommentare zusammen geben aber eine Information über die gegebenen Antworten und damit über die Allgemeinheit. Über eine automatisch generierte Zusammenfassung kann ein Gesamtüberblick über alle Kommentare geben werden.

Die drei Teilaufgaben für das Retrieval auf Kommentaren sind also:

1. *Filterung*: Entfernen von nicht relevanten und unerwünschten Kommentaren
2. *Ranking*: Sortierung und Bildung einer Reihenfolge bezüglich der Relevanz
3. *Zusammenfassung*: Generierung einer Zusammenfassen des Inhalts sich wiederholender Kommentare

Relevanz von Kommentaren. Bislang wurde außer Acht gelassen, was eigentlich die Relevanz eines Kommentars ausmacht. Die drei von Mishne [15] vorgeschlagen Aspekte zum Opinion-Retrieval für Weblogs können auch zur Bewertung von Kommentaren auf Webseiten sinnvoll verwendet werden: Thematische Relevanz, Meinungsausdruck und Beitragsqualität. Für Kommentare kann dies wie folgt aussehen:

- Ein Kommentar zum gegebenen Thema einer Webseite ist ein selbst verfasster Beitrag eines zweiten, unabhängigen Autors dazu. Dies kann eine Ergänzung bzw. Erweiterung zum Thema sein, eine Antwort auf eine eventuell gestellte Frage, aber auch mit dem gegebenen Thema nicht im Zusammenhang stehen. Einen Leser, den das gegebene Thema interessiert, interessieren sehr wahrscheinlich nur Kommentare, die mit dem Thema in engem Zusammenhang stehen. Ein guter, also lesenswerter Kommentar hält sich demnach inhaltlich nah am gegebenen Thema.

- Das vorgestellte Thema provoziert gewollt oder ungewollt Meinungen, die in Form von Kommentaren geäußert werden. Diese können für den Seitenautor und deren Leser von Interesse, also relevant sein. Beispielsweise sind persönliche Meinungen auf Produktbewertungsseiten für einen Leser als potentiellen Käufer relevant. Auch stellen die Kommentare bei YouTube in der Mehrzahl reine Meinungsäußerungen dar. Die Meinung der Allgemeinheit ist hier für einen Leser von Interesse. Andererseits stellen Kommentare, die sehr stark subjektiv sind oft auch eine vom Kommentator gewollte Provokation in Form von Trolling dar, die aber beim Leser oft unerwünscht also nicht relevant ist.

- Ein weiteres Relevanzkriterium ist die Kommentarqualität. Die Kommentarqualität ist in erster Linie der Schreibstil. Ein guter Schreibstil erhöht die Lesbarkeit und wird allgemein bevorzugt. Außerdem erhöht er die Glaubwürdigkeit des Kommentars. Hastig verfasste Kommentare, die meist auch wenig durchdacht sind, enthalten oft Rechtschreib- und Grammatikfehler. In die Bewertung der Glaubwürdigkeit eines Kommentars fließt auch die Reputation des Kommentators ein, wenn Informationen darüber existieren. Beispielsweise ist ein Kommentator, der in der Vergangenheit viele relevante Kommentare geschrieben hat, glaubwürdiger als einer mit vielen irrelevanten oder keinen Kommentaren.

1.2 Verwandte Arbeiten

Die Vorgestellten Probleme im Zusammenhang mit benutzergenerierten Inhalten sind Bestandteil der aktuellen Forschung. Dabei liegt der Fokus oft auf sehr speziellen Problemen. In [21] wird ein System zur automatischen Moderation auf dem Nachrichten-Portal Slashdot vorgestellt. Die Autoren analysieren dabei hauptsächlich die Struktur des sozialen Netzwerks der Mitglieder von Slashdot. Mit diesem Ansatz kann das Moderationssystem auf Slashdot gut automatisch nachempfunden werden. Allerdings ist dieses System nur bedingt verallgemeinerungsfähig, da nicht auf allen Portalen, die eine Kommentarfunktion anbieten entsprechende Informationen über die soziale Struktur des Benutzerkreises zur Verfügung stehen.

Eine weitere Form benutzergenerierten Inhalts stellen Weblogs dar. Die große Anzahl existierender Weblogs macht ein spezielles Retrieval erforderlich. Mishne [15] stellt Möglichkeiten vor, ein Ranking von Weblog-Artikeln bezüglich eines gegebenen Themas durchzuführen. Dabei werden ähnliche Kriterien zur Bewertung verwendet, wie sie auch für ein Kommentar-Retrieval sinnvoll sind.

Andere Arbeiten versuchen Produktbewertungen auf Portalen wie beispielsweise Amazon[3] automatisch zu sortieren [5, 12, 24]. Als Relevanzkriterium wird die Nützlichkeit bzw. *helpfulness*, die von Benutzern auf Amazon bewertet werden kann, automatisch nachempfunden. Die wichtigsten Merkmale zur Bewertung sind meist die Länge einer Produktbewertung und andere schwache Stilmerkmale.

Auch wird versucht, unerwünschte Kommentare wie Spam aus Produktbewertungen herauszufiltern [8, 9, 13]. Es werden Stilmerkmale, aber auch spezielle Merkmale für Produktbewertungen wie die Anzahl an Markennennungen und Produktbezeichnungen pro Artikel ausgewertet. Die Modelle sind sehr domänenspezifisch und deshalb nur bedingt auf Kommentare im Allgemeinen übertragbar.

Auf Produktbewertungen wird außerdem versucht, eine automatische Zusammenfassung der subjektiven und objektiven Aussagen zu generieren. Es werden beispielsweise positive und negative Meinungen zu Produktmerkmalen gesucht und zusammengefasst [7, 2].

[3] http://www.amazon.com

2 Modelle für das Retrieval

In diesem Abschnitt stellen wir Modelle vor, die für die Bearbeitung der oben genannten Aufgaben *Zusammenfassung*, *Filterung* und *Ranking* von uns untersucht wurden. Die Modelle sind nach den drei herausgearbeiteten Relevanzkriterien geordnet.

2.1 Modelle zur Berechnung thematischer Relevanz

Wir stellen zwei Modelle vor, die die inhaltliche Relevanz zwischen Kommentar und Webseite messen: das *Ähnlichkeitsmodell* und das *Kontinuitätsmodell*. In diesen Modellen werden Verfahren eingesetzt, die sich im Bereich des textbasierten Information-Retrieval (IR) als geeignet erwiesen haben.

IR ist das Bindeglied zwischen Informationsbedürfnissen einerseits und einer großen Menge an Informationen andererseits. IR-Systeme liefern dem Benutzer zu einer gegebenen Anfrage eine (überschaubare) Menge an relevanten Informationen, i.d.R in Form von geschrieben Texten.

In diesem Kontext können Webseiten mit Kommentarfunktion als Anfrage an die WWW-Gemeinschaft verstanden werden, die ihrerseits Informationen in Form von Kommentaren liefert.

Mit diesem Verständnis ist es möglich, bekannte Verfahren des IR einzusetzen und damit den Prozess des Kommentierens um die Berechnung der Relevanz zu erweitern. Die Fragestellung, die sich daraus ergibt lautet: *Wie relevant sind die Kommentare bezüglich der WWW-Seite, von der sie hervorgerufen wurden?* Die vorgestellten Verfahren unterscheiden sich insbesondere durch die Definition der Relevanz.

2.1.1 Begriffe

Die im Folgenden erklärten Begriffe sind die Grundlage für die vorgestellten Modelle.

Stoppwort Ein Stoppwort ist ein Füllwort, welches für den Inhalt eines Textes nicht von Bedeutung ist, aber mit großer Häufigkeit auftritt. Beispiele für Stoppwörter der deutschen Sprache sind: „zu“, „von“, „einer“, „der“, „oder“. Im IR werden Stoppwörter meist entfernt, da sie für die weitere Verarbeitung des Textes nicht relevant sind.

Stammformreduktion Bei der Stammformreduktion wird ein Wort auf seinen Wortstamm reduziert. Damit werden Wörter mit gleichem Wortstamm auf einen einzigen Term abgebildet, und somit die Anzahl verschiedener Terme verringert.

Retrieval-Modell[17] [16] Retrieval-Modelle ermöglichen die Berechnung der Relevanz eines Dokuments $d \in D$ bezüglich einer Anfrage $q \in Q$. D bezeichnet eine Menge von Dokumenten, Q eine Menge von Anfragen. Ein Retrieval-Modell $\mathcal{R}$ setzt sich zusammen aus dem Tupel $\langle \mathbf{Q}, \mathbf{D}, \rho_{\mathcal{R}} \rangle$. Dabei sind $\mathbf{d} \in \mathbf{D}$ und $\mathbf{q} \in \mathbf{Q}$ formalisierte Repräsentationen von $d \in D$ und $q \in Q$. Die Berechnung der Relevanz eines Dokumentes d bezüglich einer Anfrage q erfolgt durch die Retrieval-Funktion $\rho_{\mathcal{R}}(\mathbf{q}, \mathbf{d})$. Hier vorgestellte Retrieval-Modelle sind: *Vektorraummodell*, *Latent-Semantic-Indexing* und *Explicit-Semantic-Analysis*.

Dokumentvektor Der Dokumentvektor ist eine Repräsentation eines Textes als Vektor. Dabei stellt jeder Term t_k aus allen Texten $(D \cup Q)$ eine Dimension des Vektorraums dar. Die Vektoren $\mathbf{d}$ und $\mathbf{q}$ enthalten in der Dimension k das Gewicht des Terms t_k bezüglich d bzw. q. Das Gewicht wird mit einem Termgewichtsmaß bestimmt, das die Wichtigkeit eines Terms in Bezug auf den Inhalt des Dokuments oder der Dokumentkollektion quantifiziert.

TF-Gewicht Bei der TF-Gewichtung wird jeder Term mit seiner relativen Häufigkeit innerhalb des Dokumentes dargestellt. Dieses Maß ist also dokumentspezifisch.

TF-IDF-Gewicht Im TF-IDF-Schema wird die Termhäufigkeit TF multipliziert mit der inversen Dokumentfrequenz IDF. Das IDF-Maß gewichtet seltene Terme innerhalb der Kollektion höher und häufige Terme niedriger. Damit werden insbesondere kollektionsspezifische Stoppwörter geringer gewichtet oder eliminiert. Dieses Maß ist somit kollektions- und dokumentbezogen.

Term-Dokument-Matrix Die Term-Dokument-Matrix entsteht durch spaltenweise Aneinanderreihung aller Dokumentvektoren einer Kollektion. In den Zeilen enthält sie alle Terme der Kollektion.

Vektorraummodell (engl. Vector Space Model, VSM) Das Vektorraummodell ist ein Retrieval-Modell. Dokumente sind repräsentiert durch Dokumentvektoren. Als Retrieval-Funktion $\rho_{\mathcal{R}}$ dient der Kosinus zwischen den Dokumentvektoren $\mathbf{d}$ und $\mathbf{q}$. Der Kosinus bildet im VSM auf das Intervall $[0, 1]$ ab, da alle Dimensionen der Vektoren positiv sind. Dabei gilt:

$$\rho_{VSM}(\mathbf{d}, \mathbf{q}) = 1 \quad \Leftrightarrow \quad \mathbf{d} = \mathbf{q} \tag{1}$$

$$\rho_{VSM}(\mathbf{d}, \mathbf{q}) = 0 \quad \Leftrightarrow \quad \mathbf{d} \perp \mathbf{q} \tag{2}$$

Damit stellt der Kosinus ein Maß für die Ähnlichkeit zweier Dokumentvektoren

und der zugrunde liegenden Texte d und q dar.

Die Retrieval-Funktion lautet:

$$\rho_{VSM}(\mathbf{d}, \mathbf{q}) \quad = \quad \frac{\mathbf{d} \cdot \mathbf{q}}{|\mathbf{d}||\mathbf{q}|} \tag{3}$$

Latent Semantic Indexing (LSI)[1] LSI ist ebenfalls ein Retrieval-Modell. In diesem Modell werden die Dokumentvektoren $\mathbf{d}$ zu einer Term-Dokument-Matrix zusammengefasst, in der die Terme mit TF-IDF gewichtet sind. Die entstandene Matrix wird einer Singulärwertzerlegung unterzogen und damit in einen niedrigdimensionaleren Raum projiziert. Dieser Raum wird als Konzeptraum bezeichnet, in dem Terme zu abstrakten Konzepten zusammengefasst wurden. Die Anfragen $\mathbf{q}$ werden ebenfalls in diesen Raum projiziert. Als Ähnlichkeitsmaß dient der Kosinus zwischen den dimensionsreduzierten Vektoren des Dokumentes $\mathbf{d_r}$ und der Anfrage $\mathbf{q_r}$ im Konzeptraum.

Explizit Semantic Analysis (ESA)[4] ESA ist ein weiteres Retrieval-Modell. Im Gegensatz zu LSI, wo die Konzepte im Konzeptraum abstrakt sind, werden im ESA-Modell explizite Konzepte eingesetzt, um einen Text zu repräsentieren. Explizite Konzepte können z.B „Fahrrad“ oder „objektorientierte Programmierung“ sein. Dazu werden Artikel der Online-Enzyklopädie Wikipedia[4] verwendet. Populäre Wikipedia-Artikel zeichnen sich oft durch detaillierte Beschreibungen aus. Dadurch können Wikipedia-Artikel als Konzept angesehen werden. Die Beschreibungen der Artikel bilden die abstrakte Repräsentation $\mathbf{c} \in \mathbf{C}$ der Konzepte als Dokumentvektor. Die Dokumentvektoren $\mathbf{c}$ werden als Term-Dokument-Matrix $\mathbf{C_{TDM}}$ dargestellt und TF-IDF gewichtet. Der Konzeptraum wird durch die Menge $\mathbf{C}$ der Konzepte aufgespannt.

Dokumentvektoren $\mathbf{d}$ und Anfragevektoren $\mathbf{q}$ werden in den Konzeptraum projiziert. Die in den Konzeptraum projizierten Dokumentvektoren enthalten in Dimension i die Ähnlichkeit $\rho_{VSM}(\mathbf{d}, \mathbf{c_i})$ zwischen $\mathbf{d}$ und $\mathbf{c_i}$ nach dem Vektorraummodell. Liegt die Ähnlichkeit eines Vektors $\mathbf{d}$ bzw. $\mathbf{q}$ zu einem Konzept $\mathbf{c_i}$ unter einem festgelegten Insignifikanzschwellwert, so wird diese Dimension auf 0 gesetzt. Die Projektion der Dokumentvektoren $\mathbf{d}$ in den Konzeptraum erfolgt durch Matrixmultiplikation der transponierten Matrix $\mathbf{C_{TDM}^T}$ mit dem Vektor $\mathbf{d}$. Den

[4] http://en.wikipedia.org/

entstandenen Vektor bezeichnen wir als Interpretationsvektor[4] $\mathbf{d_{int}}$.

$$\mathbf{d_{int}} = \mathbf{C_{TDM}^{T}} \cdot \mathbf{d} \tag{4}$$

Die Ähnlichkeit $\rho_{ESA}(\mathbf{d}, \mathbf{q})$ wird als Kosinus der Interpretationsvektoren $\mathbf{d_{int}}$ und $\mathbf{q_{int}}$ berechnet.

2.1.2 Ähnlichkeitsmodell

Im hier vorgestellten Modell wird die Relevanz eines Kommentars zu einem Artikel durch die Ähnlichkeit der beiden Texte zueinander quantifiziert. Motiviert wird dieses Maß mit der Idee, dass ein Kommentar, der eine sehr geringe Ähnlichkeit zum Artikel besitzt, wenig inhaltliche Relevanz haben kann.

Zur Berechnung der Ähnlichkeit werden die vorgestellten Retrieval-Modelle *VSM*, *LSI* oder *ESA* verwendet. Die Menge D der Dokumente wird dabei durch die Menge der Kommentare repräsentiert, die zu einem Artikel a abgegeben wurden. Die Menge Q der Anfragen beinhaltet in diesem Modell nur das eine Element *Artikel* a.

Grundsätzlich ergibt sich bei der Ähnlichkeitsanalyse die Problematik, dass der Artikel selbst den besten Kommentar darstellen würde, weil er die Ähnlichkeit von 100% zu sich selbst hat. Anders formuliert hat ein Kommentar mit geringer Ähnlichkeit zum Artikel zwar eine geringe inhaltliche Relevanz, die Umkehrung gilt aber nicht in jedem Fall. Denkbar ist diese Situation, wenn ein Kommentar große Teile des Artikels zitiert, aber sehr wenig Neues hinzufügt.

Das Ähnlichkeitsmodell bildet auf den Wertebereich $[0, 1]$ ab, dabei deutet ein Wert nahe 1 auf eine große Ähnlichkeit des Kommentars zum Artikel hin und ein Wert nahe 0 auf eine geringe Ähnlichkeit. Mit dieser Eigenschaft kann das Ähnlichkeitsmodell direkt für ein Ranking benutzt werden, indem die Kommentare nach absteigender Ähnlichkeit sortiert werden.

Durch die Festlegung eines Schwellwertes τ kann dieses Maß ebenfalls für die Filterung eingesetzt werden. Kommentare mit einer Ähnlichkeit $\varphi < \tau$ zum Artikel werden entfernt.

2.1.3 Modell zur Berechnung der thematischen Kontinuität

Betrachtet man eine Webseite als Aufforderung zum Kommentieren, so ist eine der Motivationen, Kommentare abzugeben der Fakt, dass der Kommentator zusätzliches Wissen bezüglich des Themas des Artikels hat. Dieser Informationsvorsprung, der sich in

den Wörtern des Kommentars manifestiert, ist ein wichtiges Relevanzkriterium. Ein Kommentar ist im Kontinuitätsmodell dann relevanter, wenn er das Thema des Artikels in größerem Maße vervollständigt als ein anderer Kommentar. Anders formuliert ist ein Kommentar insbesondere dann inhaltlich relevant, wenn er zum Artikel zusätzliche Informationen in Form von Fakten zum Thema enthält.

Das hier vorgestellte Maß misst den Informationsvorsprung, den ein Text (Kommentar) bezüglich eines anderen Textes (Artikel) mit sich bringt. Grundlage dafür bildet das ESA-Modell. Es vergleicht Texte nicht auf Wortebene, sondern auf Konzeptebene.

Wir postulieren, dass ein Text dann einen Informationsvorsprung zu einem gegebenen Text darstellt, wenn sich die Texte konzeptuell sehr ähnlich sind, aber keine Wörter gemeinsam haben. Konkret kann sich eine thematische Ergänzung nur in den Wörtern des Kommentars manifestieren, die der Artikel noch nicht verwendet hat. Dementsprechend unterscheidet sich das Kontinuitätsmodell vom ESA-Ähnlichkeitsmodell in der formalen Repräsentation $\mathbf{d}$ eines Kommentars. Wir entfernen aus den Kommentaren alle Wörter, die auch im Artikel vorkommen, bevor die Dokumentvektoren $\mathbf{d}$ konstruiert werden.

Sei $\{a\}$ die einelementige Menge Q der Anfragen, $\mathbf{a}$ dessen Dokumentvektor und T_a die Menge der Terme in a. Sei weiterhin D eine Menge von Kommentaren zu a. Des weiteren ist T_d die Menge der Terme in Kommentar $d \in D$. Die abstrakte Darstellung $\mathbf{d}$ als Dokumentvektor wird dann aus der Menge der Terme $(T_d \backslash T_a)$ gebildet. Nach dem ESA-Modell werden die Vektoren $\mathbf{d}$ und $\mathbf{a}$ in den Konzeptraum projiziert und im Konzeptraum durch die Interpretationsvektoren $\mathbf{d_{int}}$ und $\mathbf{a_{int}}$ repräsentiert. Die Ähnlichkeit der Interpretationsvektoren von Artikel und Kommentardarstellung nach dem Kosinusmaß ist ein Maß für die thematische Vervollständigung von Artikel a durch Kommentar d.

Auch dieses Maß bildet auf das Intervall $[0, 1]$ ab und kann direkt für ein Ranking verwendet werden. Ebenso kann das Kontinuitätsmaß durch Festlegung eines Schwellwertes τ für die Filterung eingesetzt werden.

2.2 Meinungsanalyse

Meinungsanalyse, im englischen *Sentiment Analysis* oder *Opinion Mining* ist ein sehr aktuelles Forschungsgebiet, das die Bereiche Information-Retrieval und Computerlinguistik verbindet. Die meisten Arbeiten in diesem Bereich befassen sich mit der automatischen Extraktion und Klassifizierung von Empfindungs- und Meinungsäußerungen auf Dokument- oder Satzebene. Ausgesuchte Arbeiten werden im Folgenden vorgestellt.

Darin werden redaktionelle Beiträge aus Nachrichtentexten extrahiert und die Meinung von Produkt- und Filmbewertungen als beispielsweise positiv und negativ klassifiziert. Auf Dokumentebene sollen zum Beispiel Filmbesprechungen klassifiziert werden. Auf Satz- oder Phrasenebene kann in Produktbewertungen die Meinung über bestimmte Produktmerkmale identifiziert werden. Dies stellt eine meinungsorientierte Informationsextraktion dar [22].
Eine Meinungsäußerung ist nach Ding [2] wie folgt zusammengesetz:

- *Meinungsinhaber (Opinion holder)*: derjenige dem die Meinung zugeordnet wird. Das muss nicht der Autor sein. Beispielsweise ist im Satz „Peter fand den Film super“, „Peter“ der, dem die Meinung zugeordnet wird [11].
- *Meinung (Opinion)*: eine subjektive Äußerung, die ausgedrückt wird. Im vorhergehenden Beispielsatz war dies „super“. Eine Meinung besitzt eine semantische *Orientierung* oder *Polarität* [6] zwischen den Polen *positiv* und *negativ*.
- *Meinungsgegenstand (Object)*: das, worüber die Meinung geäußert wird. Im Beispielsatz ist das der „Film“.

2.2.1 Meinungserkennung

Bei Esuli [3] sind die drei wichtigsten Aufgabengebiete der Meinungsanalyse wie folgt eingeteilt:

1. Erkennung der Subjektivität bzw. Objektivität: Wird ein Fakt oder eine Meinung geäußert?
2. Erkennung der semantischen Orientierung (Polarität) der Meinung zwischen *positiv* und *negativ*
3. Erkennung des Grads der Orientierung bzw. der Stärke der Subjektivität

Daneben können auch die Identifizierung des Meinungsinhabers und die Identifizierung des Meinungsgegenstandes als weitere Aufgaben zugeordnet werden. Diese Aufgabenstellungen werden je nach Anforderung auf unterschiedlichen Ebenen bearbeitet: Wort-, Phrasen-, Satz- oder Dokumentebene.

Für die ersten drei Aufgabengebiete werden hauptsächlich zwei verschiedene Ansätze verfolgt. Im *lexikalischen Ansatz* werden Wörterbücher benutzt um Wörter und Phrasen zu finden, die eine Meinungsäußerung identifizieren. Über grammatikalische Regeln

können zu den identifizierten Wörtern weitere Informationen in die Klassifizierung einfließen. Verändert beispielsweise eine Konjunktion wie „nicht“ das gefundene Wort, wird das berücksichtigt. Die Erkennungsleistung der Meinungsanalyse hängt maßgeblich von der Qualität des Wörterbuchs ab.

In einem zweiten, *korpusbasierten Ansatz* werden für jedes Dokument aus einem Korpus verschiedene Merkmale (Features) berechnet. Die Auswahl geeigneter Merkmale ist für das Ergebnis von zentraler Bedeutung. Merkmale sind meist Unigramme (die einzelnen Wörter des Dokuments) oder n-Gramme (n fortlaufende Wörter). Auch verschiedene statistische Merkmale wie Satzlänge und Dokumentlänge kommen zum Einsatz. Die Anzahl an positiven und negativen Wörtern, wiederum mit Wörterbüchern ermittelt, werden ebenfalls als Merkmale benutzt. Verschiedene Merkmale werden zu Merkmalsvektoren zusammengefügt. Über einen manuell klassifizierten Trainingskorpus wird ein Klassifikator trainiert, der dann auf unbekannten Objekten automatisch klassifizieren kann. *Naive-Bayes-Klassifikatoren* und *Support-Vector-Mashines* werden dazu häufig verwendet. Da in diesem Ansatz die Informationen aus dem Korpus selbst kommen, gehen besondere Eigenschaften des Korpus und damit der Domäne in die Klassifikation ein. So wird die Erkennungsleistung im Vergleich zum lexikalischen Ansatz gesteigert. Dafür muss der Klassifikator für jeden Korpus anhand einer manuell vorklassifizierten Trainingsmenge neu trainiert werden, wenn sich die statistischen Verteilungen der Merkmale ändern. Beispielsweise unterscheiden sich Satzlänge und Auftrittswahrscheinlichkeiten von Wortklassen stark zwischen Korpora aus Nachrichtentexten und Korpora aus Produktbewertungen.

Die Korpusunabhängigkeit von lexikalischen Verfahren hat dagegen den Vorteil einer universellen Einsetzbarkeit. Da keine Informationen eines Korpus in die Klassifikation einfließen, kann das Verfahren ohne aufwendige Anpassungen auf verschiedenen Domänen eingesetzt werden. Ein Hauptnachteil ist, dass keine Wörter erkannt werden können, die nicht im Wörterbuch enthalten sind. Auch Wörter mit kontextabhängiger Bedeutung sind schwer handhabbar.

2.2.2 Wörterbucherstellung

Die Wörterbucherstellung spielt bei der Sentiment-Analyse eine zentrale Rolle. Wörterbücher, mit einer Einteilung in positive und negative Wörter, sind in *lexikalischen Ansätzen* notwendige Bedingung, stellen bei *korpus-basierten Ansätzen* aber auch wichtige Features dar, beispielsweise als Auftrittshäufigkeit. Bei den drei Aufgabengebieten der Meinungsanalyse ist eine Sammlung und Einteilung von Wörtern sehr nützlich:

1. Erkennung der Subjektivität bzw. Objektivität eines Terms:
   Subjektive Wörter sind beispielsweise „gut“, „schön“, „furchtbar“ und objektive Wörter z.B. „blau“, „rund“, „biegsam“.

2. Erkennung der semantischen Orientierung (Polarität):
   Positive Begriffe sind beispielsweise „exzellent“, „lustig“, „richtig“ und „schlecht“, „dumm“, „übel“ sind negative.

3. Erkennung des Grads der Orientierung bzw. Stärke der Subjektivität:
   „fantastisch“ ist stärker als „gut“ bzw. eine Steigerung davon.

Die meisten Wörter tragen allerdings mehrere Bedeutungen. Das englische Wort „cool“ kann umgangssprachlich als positiv gewertet werden, sowie als eher negative Temperaturempfindung. Auch im Kontext kann sich die Bedeutung eines Wortes stark verändern. Eine „schöne Bescherung“ kann eine positive oder auch negative Aussage seine. Besonders bei Umgangssprache bzw. *Slang* kann sich die Bedeutung auch mit der Zeit schnell ändern sowie in verschiedenen Bevölkerungsgruppen gegensätzlich verwendet werden.

In einem Wörterbuch für die Meinungsanalyse sollten Wörter mit möglichst eindeutiger Orientierung bzw. dem Grad der semantischen Orientierung gesammelt werden. In der Mehrzahl sind das Adjektive und Adverbien. Bei Substantiven sind es meist Schimpfwörter, die eine eindeutige Meinung über etwas ausdrücken.

Für die Erstellung von Wörterbüchern werden zwei Arten von Ansätzen verfolgt. In *korpus-basierten Ansätzen* werden die Wörter einer großen Kollektion bzw. eines Korpus ausgewertet. Des Weiteren werden *thesaurus-basierte Ansätze* verfolgt, bei denen aus vorhandenen Theasuren Wörter extrahiert werden.

Im Folgenden werden die Ansätze von Turney [19, 20] sowie von Hatzivassiloglou und McKeown [6] als die beiden wichtigsten *korpus-basierte Ansätze* vorgestellt:

- Im *Pointwise Mutual Information (PMI)* genannten Verfahren von Turney [19, 20] wird die Wahrscheinlichkeit des Zusammenauftretens, der Kookkurenz von zwei Wörtern, w_1, w_2 ermittelt und über die jeweiligen Wahrscheinlichkeiten des Einzeln-Auftretens die *Pointwise Mutual Information* berechnet. Die Wahrscheinlichkeiten werden über einen hinreichend großen Korpus durch die Häufigkeiten abgeschätzt.

$$PMI(w_1, w_2) = log_2 \left(\frac{p(w_1 \cap w_2)}{p(w_1)p(w_2)} \right) \tag{5}$$

PMI ist somit ein Maß für die statistische Unabhängigkeit zwischen w_1 und w_2. Die Annahme bei diesem Verfahren ist, dass Wörter mit gleicher semantischer Orientierung in einem Satz häufiger zusammen auftreten als Wörter mit ungleicher Orientierung. Berechnet man für ein gesuchtes Wort w PMI zu mehreren Wörter mit bekannter Orientierung, kann man damit ermitteln, mit welchen Wörtern es häufiger vorkommt und damit die gleiche semantische Orientierung besitzt.

$$O(w) = \sum_{w_{pos} \in Pos} PMI(w, w_{pos}) - \sum_{w_{neg} \in Neg} PMI(w, w_{neg}) \tag{6}$$

Ist $O(w) > 0$ kommt das Wort häufiger im positiven Zusammenhang vor, und es kann gefolgert werden, dass es ebenfalls eine positive Orientierung hat. Ist $O(w) < 0$, kann entsprechend gefolgert werden, das das Wort eine negative Orientierung besitzt. Der Betrag $|O(w)|$ gibt Auskunft über die Eindeutigkeit. Liegt $O(w)$ nahe bei 0, kommt das Wort mit der unbekannten semantischen Orientierung gleich häufig mit beiden Orientierungen vor. Einerseits kann das bedeuten, dass es selbst keine Orientierung hat, andererseits kann es auch zwei gegensätzlich orientierte Bedeutungen haben.

Turney nutzt für sein Verfahren den Korpus der Suchmaschine Altavista[5], auf dessen Grundlage die Wahrscheinlichkeiten errechnet werden. Die Anzahl der Treffer auf eine Suchanfragen, $hits(q)$ geteilt durch die Größe N des Korpus stellen hier die Häufigkeit dar.

$$PMI(w_1, w_2) = log_2 \left(\frac{\frac{1}{N} hits(w_1 \ NEAR \ w_2)}{\frac{1}{N} hits(w_1) \ \frac{1}{N} hits(w_2)} \right) \tag{7}$$

$NEAR$ steht hier für den NEAR-Operator der Suchmaschine. Zwischen beiden Wörtern dürfen maximalen zehn andere Wörter stehen. Der Operator wird benutzt, um die Kookkurenz der beiden Wörter zu ermitteln.

- Hatzivassiloglou und McKeown [6] benutzen Konjunktionen zwischen Adjektiven, um für unbekannte Adjektive die Orientierung zu ermitteln. Die Annahme ist, dass bei bestimmten Konjunktionswörtern wie beispielsweise „und“ und “weder noch“ zwei Adjektive, die damit verbunden werden die gleiche Orientierung besitzen. Bei Kombinationen wie z.B. „fair und ehrlich“ oder „korrupt und brutal“ besitzen jeweils beide Adjektive die gleiche Orientierung. Kombinationen wie

[5] http://altavista.com

„fair und korrupt“ oder „ehrlich und brutal“ kommen kaum vor. Bei anderen Konjunktionswörtern, wie „aber“ und „entweder oder“ ist es umgekehrt. Die beiden verknüpften Adjektive sind dabei meist gegensätzlich orientiert.

Es werden zuerst alle Adjektive mit Konjunktionen aus dem Korpus extrahiert. Im zweiten Schritt wird ein Klassifizierer trainiert, der die Adjektive in „gegensätzlich orientiert“ und „gleich orientiert“ klassifizieren kann. Damit wird auf den Adjektiven ein Graph erstellt, der im dritten Schritt geclustert wird. So entstehen zwei Cluster. Innerhalb eines Clusters sind alle Wörter gleich orientiert und zu den Wörtern des anderen Clusters gegensätzlich. Ein Cluster kann dann über bereits bekannte Wörter der positiven semantischen Orientierung zugeordnet werden, der andere der negativen Orientierung. Alle unbekannten Wörter können so einer der beiden Orientierungen zugeordnet werden.

Andere Ansätze basiert auf Thesauren, um Wörter für ein Meinungswörterbuch zu sammeln. Hauptvertreter für *thesaurus-basierte Ansätze* sind die Verfahren von Kim und Hovy [11] sowie von Kamps [10]:

- Kim und Hovy [11] suchen ausgehend von Startwörtern (seed words) mit bekannter semantischer Orientierung aus einem lexikalischen Thesaurus wie *WordNet*[14] Synonyme und Antonyme. Synonyme haben, so die Annahme, die gleiche semantische Orientierung, Antonyme dagegen eine gegensätzliche.

- Kamps [10] nutzt ein Distanzmaß zwischen synonymen Wörtern in *WordNet* ausgehend von bekannten Startwörtern. So kann für zwei Wörter mit minimaler Distanz eine Aussage über die Ähnlichkeit ihrer Bedeutung gemacht werden. Es wird ein Graph aufgebaut, mit Wörtern als Knoten und deren Synonym-Beziehungen als Kanten. Die semantische Orientierung wird über die Distanz eines Wortes zu den Wörtern eines gegensätzlichen Adjektivpaars z.B. „gut-schlech“ bestimmt. Das unbekannte Wort erhält die Orientierung, des Wortes mit dem minimalen Abstand zum unbekannten Wort.

2.2.3 Relevanzmodell zur Meinungsanalyse

Die drei Hauptaufgaben der Sentiment-Analyse eignen sich als Relevanzmaße für alle drei vorgeschlagenen Retrievaltasks. Zum Ranking von Kommentaren kann ein Modell aus dem Grad der Subjektivität für die Relevanz gebildet werden. Der Grad des Subjektivität legt dabei die Reihenfolge aller Kommentare C fest. Beispielsweise könnten einen Leser objektivere, emotional nicht eingefärbte Kommentare stärker interessieren

oder umgekehrt. Für diese Aufgabe ist es allerdings notwendig ein feingranulareres Maß als das beispielsweise von Wilson [23] vorgeschlagene (weak, medium,strong) zu benutzen. Mit diesem Maß wäre es eher möglich die Aufgabe der Filterung über diesem Relevanzkriterium zu erfüllen. Stark subjektiv geschriebene Kommentare könnten einen Anhaltspunkt zur Erkennung und Filterung nicht erwünschter Kommentare, wie Flamewars sein.

Am besten als Relevanzkriterium ist die Meinungsanalyse für den Retrievaltask der Zusammenfassung geeignet. Alle Meinungsäußerungen aus allen Kommentaren über den Artikel sollen in eine kompakte Form zusammengefasst werden, um so dem Leser einen Überblick über die allgemeine Meinung zum Thema zu geben. Ein Modell hierfür sieht folgendermaßen aus:

Für jeden Kommentar $c \in C$ wird eine abstrakte Darstellung $\mathbf{c} \in \mathbf{C}$ aus zwei Wortvektoren der beiden semantischen Orientierungen erzeugt $\mathbf{c} = (\mathbf{o_{c,pos}}, \mathbf{o_{c,neg}})$. Für alle $\mathbf{c} \in \mathbf{C}$ kann dann eine Zusammenfassung $S = (S_{pos}, S_{neg})$ aus $S_{pos} = \sum_{\mathbf{c} \in \mathbf{C}} \mathbf{o_{c,pos}}$ und $S_{neg} = \sum_{\mathbf{c} \in \mathbf{C}} \mathbf{o_{c,neg}}$ berechnet werden.

Die Erzeugung des Opinionvektors $\mathbf{o_c}$ stellt dabei eine Meinungsextraktion und -klassifizierung dar. Es werden nur Meinungsäußerungen des Kommentators als Meinungsinhaber und nur zum Thema der Webseite als Meinungsgegenstand extrahiert. Bei sehr kurzen Kommentaren mit ein bis drei Sätzen vereinfacht sich dies, da man davon ausgehen kann, dass der Kommentarautor auch Meinungsinhaber ist bzw. die Meinung die geäußert wird auch selbst vertritt. Ein lexikalischer Ansatz auf Wortebene kann dabei die meinungstragenden positiven und negativen Wörter identifizieren, die in den Wortvektoren $\mathbf{o_{c,pos}}$ und $\mathbf{o_{c,neg}}$ abgelegt werden.

In einem nächsten Schritt können dann die Wortvektoren aller Kommentare zusammengefasst werden. So entsteht eine Zusammenfassung der positiven und der negativen Meinungsäußerungen. Die Menge und Anzahl der Meinungsäußerungen der beiden Orientierungen stellt so die allgemeine Meinung der Kommentatoren dar. Als Präsentation bietet sich hier eine Wortwolke, ähnlich einer Tag-Cloud an. Hierbei werden die geäußerten Wörter in verschiedener Schriftgröße abhängig von der Anzahl der Äußerungen dargestellt. So kann Menge und Anzahl der Äußerungen gleichzeitig visualisiert werden.

2.3 Qualitative Analyse

In diesem Kapitel betrachten wir den Schreibstil der Kommentare als Relevanzkriterium. Neben dem Schreibstil ist für die Bewertung der Qualität auch die Beurteilung

des Kommentators möglich [15]. Diese Beurteilung wird im Rahmen dieser Arbeit aber nicht untersucht.

Die Stilanalyse wird mit der Annahme motiviert, dass ein Text, der stilistisch hochwertiger geschrieben wurde, auch eine größere Relevanz hat. Variiert werden kann dieses Modell durch die Verwendung von Vandalismusmerkmalen anstelle der Stilmerkmale. Offensichtlich sind Kommentare, die als Vandalismus erkannt werden, weniger relevant. Auch die Kombination von Stil- und Vandalismusmerkmalen in einem Modell ist möglich.

Sei d ein Kommentar aus der Menge der Kommentare D. Sei weiterhin F eine Menge an Stilmerkmalen, die Eigenschaften eines Textes als Zahl quantifizieren, dann ist die formalisierte Darstellung $\mathbf{d} \in \mathbf{D}$ eines Kommentars ein Vektor von Stilmerkmalen der Dimension $|F|$. Jedem Stilmerkmal ist eine Dimension zugeordnet. Zwei Kommentare $d, d' \in D$ können verglichen werden, indem der Kosinus von $\mathbf{d}$ und $\mathbf{d}'$ gebildet wird.

In in Abschnitt 2.3.2 wird besprochen, welche Möglichkeiten bestehen, aus diesem Vergleichsmaß zu erkennen, welcher Kommentar relevanter ist. Im folgenden Abschnitt werden Stil- und Vandalismusmerkmale vorgestellt.

2.3.1 Stil- und Vandalismusmerkmale

Aus der Linguistik sind Maße bekannt, die es ermöglichen auf den Bildungsgrad des Autors zu schließen bzw. (in Schuljahren gemessen) anzugeben, welchen Bildungsstand der Leser mindestens haben muss, um den Text zu verstehen. Ebenfalls sind weitere Merkmale bekannt, die es ermöglichen, verschiedene Autoren zu unterscheiden:

1. Dale-Chall Lesbarkeitsindex: Dieser Index benutzt eine Liste, die die 3000 einfachsten Wörter der englischen Sprache enthält, und verrechnet den Anteil der Wörter aus dieser Liste an einem Text mit dessen durchschnittlicher Satzlänge. Das Maß soll auf die Jahre der Schulbildung des Autors im amerikanischen Schulsystem abbilden.

2. Flesch-Kincaid-Index: es wird ein Maß aus der durchschnittlichen Satzlänge und der durchschnittlichen Anzahl an Silben pro Wort gebildet. Dieses Maß gibt die Jahre der Schulbildung an, die benötigt werden um einen Text zu verstehen.

3. Gunning-Fog-Index: verrechnet wird die durchschnittliche Satzlänge mit der Anzahl der Wörter, die aus mindestens drei Silben bestehen. Auch dieses Maß gibt die Jahre der Schulbildung an, die benötigt werden um einen Text zu verstehen.

4. Anzahl der Terme

5. Durchschnittliche Silbenzahl pro Wort
6. Relative Häufigkeit von Interjektionen
7. Relative Häufigkeit von Präpositionen

Um die Anzahl der Merkmale zu erhöhen, können in diesem Zusammenhang auch Vandalismusmerkmale eingesetzt werden:

1. Abweichung der Verteilung der Buchstaben von der durchschnittlichen Verteilung der Buchstaben der englischen Sprache
2. Grad der Komprimierbarkeit eines Textes mit einem Kompressionsalgorithmus
3. Verhältnis vulgärer Wörter zu allen Wörtern

2.3.2 Anwendung der Stilanalyse

Das Retrieval-Modell der Stilanalyse bietet zunächst nur die Möglichkeit, verschiedene Texte bezüglich stilistischer Ähnlichkeit miteinander zu vergleichen.

Die Aufgabe eines solchen Maßes ist es aber zu entscheiden, welcher Text stilistisch besser ist. Dazu sind verschiedene Ansätze möglich:

1. Heuristisch: Für jedes Merkmal $f \in F$ wird heuristisch die beste Ausprägung ermittelt. Daraus ergibt sich ein *idealer Stilvektor*. Die Ähnlichkeit zu diesem Vektor bildet das Qualitätsmaß.
2. Empirisch: Sei G eine Menge an Texten, die Beispiele für guten Stil darstellen. Für alle Texte $g \in G$ wird der Stilvektor berechnet. Aus der Menge $\mathbf{G}$ aller Stilvektoren ermittelt man den Zentroiden $\mathbf{g_z}$. Das Qualitätsmaß wird repräsentiert durch die Ähnlichkeit des Stilvektors eines Textes zum Zentroiden $\mathbf{g_z}$.

Das Qualitätsmaß kann verwendet werden, um ein Stilranking zu realisieren. Ebenfalls ist es möglich, einen Schwellwert τ zu definieren und damit eine Filterung durchzuführen. Weiterhin können für die Filterung auch Verfahren des maschinellen Lernens eingesetzt werden. Dazu notwendig ist ein vorklassifizierter Korpus. Damit kann die Filterung als Klassifikationsproblem verstanden werden.

3 Ergebnisse

Das Kapitel Ergebnisse beschreibt die Evaluierungen und Fallstudien, die wir zur Einschätzung der vorgestellten Maße durchgeführt haben.

3.1 Korpora

In diesem Abschnitt werden die Korpora vorgestellt, die speziell für die Evaluierung der vorgeschlagenen Maße erstellt wurden.

3.1.1 Slashdot-Korpus

Slashdot ist eine Nachrichtenplattform im Web, die es Benutzern ermöglicht besondere Ereignisse einer breiten Masse von Interessenten zur Verfügung zu stellen. Diese Nachrichten sind meist Zusammenfassungen von Nachrichten anderer Portale. Thematisch stammen die Nachrichten auf Slashdot aus dem Bereich Technik/IT. Ein zentrales Element von Slashdot ist die Kommentarfunktion. Um die Übersichtlichkeit in der großen Masse der Kommentare zu gewährleisten, existiert ein Moderationssystem, mit dessen Hilfe Punkte vergeben, und Kommentare in Kategorien eingeteilt werden können. Die Moderation wird von anderen Benutzern durchgeführt. Für jeden Kommentare können Punkte (Scores) zwischen -1 und 5 vergeben werden, ebenso kann jeder Kommentar einer der Kategorien *Informative*, *Insightfull*, *Interesting*, *Funny*, *Troll*, *Offtopic*, *Redundant* oder *Flamebait* zugeordnet werden. Durch die benutzergenerierte Vorklassifikation bietet Slashdot eine ideale Basis für einen Kommentar-Korpus, da für alle moderierten Kommentare ein Hinweis auf die Stärke der Relevanz gegeben ist.

Kommentare auf Slashdot können in der Form eines Diskussionsforums abgegeben werden. Aus diesem Grund unterscheiden wir Kommentare ersten und zweiten Grades.

- Erster Grad: Der Kommentar wurde direkt auf die Nachricht abgegeben.
- Zweiter Grad: Der Kommentar wurde auf einen anderen Kommentar abgegeben.

Wir haben die Nachrichtenartikel inklusive aller Kommentare aus dem Zeitraum von Januar 2006 bis Juni 2008 gecrawlt. Das sind 17.948 Nachrichten und $3.820.918$ Kommentare. 780.008 davon sind Kommentare ersten Grades, $3.040.910$ sind Kommentare zweiten Grades. Im Rahmen dieser Arbeit sind Kommentare ersten Grades von größerem Interesse, da vereinfachend die Relevanz bezüglich der Nachricht untersucht werden soll.

Tabelle 1: Gezeigt wird die Häufigkeitsverteilung der Kategorien im Slashdot-Korpus, sowie die durchschnittliche Punktezahl des Scorings innerhalb jeder Kategorie. Abkürzungen: Inform.=Informative, Flameb.=Flamebait, Offt.=Offtopic, Red.=Redundant

Klasse	**Inform.**	**Insightfull**	**Interesting**	**Funny**	**Flameb.**	**Offt.**	**Red.**	**Troll**
Häufigkeit	14%	32%	24%	16%	3%	4%	2%	5%
Score	3,1	3,1	2,9	3,2	-0,5	-0,6	-0,3	-0,6

In vielen Slashdot-Nachrichten wird auf Quellen im Internet verlinkt. Diese sind ebenfalls für den Korpus heruntergeladen worden.

Nicht alle Kommentare wurden von den Benutzern klassifiziert. Nicht klassifizierte Kommentare können keinen Hinweis auf die Relevanz für den Leser geben und sind demnach für die Evaluierung von geringerem Interesse. Die Anzahl der klassifizierten Kommentare ersten Grades liegt bei 311.167. Diese Menge bildet die Grundlage der nachfolgend beschriebenen Experimente. Tabelle 1 zeigt die Verteilung der Kategorien und den erwarteten Score für eine Nachricht aus der jeweilige Kategorie. Es lässt sich feststellen, dass 86% der Kommentare in positive Kategorien eingeordnet wurden und 14% in negative. Als positiv werden die Kategorien *Informative*, *Insightfull*, *Interesting* und *Funny* angesehen. Diese Einteilung spiegelt sich auch im Scoring wider. Die positiven Kategorien erreichen durchschnittlich 3 Punkte, die Negativen etwa $-0,5$.

Abbildung 1 zeigt die Verteilung der Anzahl der Kommentare pro Artikel. Die mittleren 50% der Artikel haben 16 bis 41 Kommentare ersten Grades, und 35 bis 160 Kommentare ersten und zweiten Grades.

In Abbildung 2 ist die Länge der Kommentare ersten Grades gegen die Häufigkeit aufgetragen. Die mittleren 50% der Kommentare sind zwischen 1 und 45 Wörter lang. 5% der Kommentare sind länger als 260 Wörter.

3.1.2 YouTube

YouTube[6] ist laut Hitwise[7] das populäres Video-Portal im Internet mit 73.18% Marktanteil. Es wurde 2005 gegründet und gehört seit 2006 zu Google Inc. Eine Wildcard-Suche mit „*“ auf YouTube bringt über 77 Millionen Einträge in Form von Videos. Jeder Nutzer kann Videos bei YouTube präsentieren, die von anderen Nutzern kommentiert werden können. Dabei sind mehrere Tausend Kommentare zu einem populären Video

[6] http://www.youtube.com

[7] http://www.hitwise.com/press-center/hitwiseHS2004/us-visits-to-youtube-14042008.php

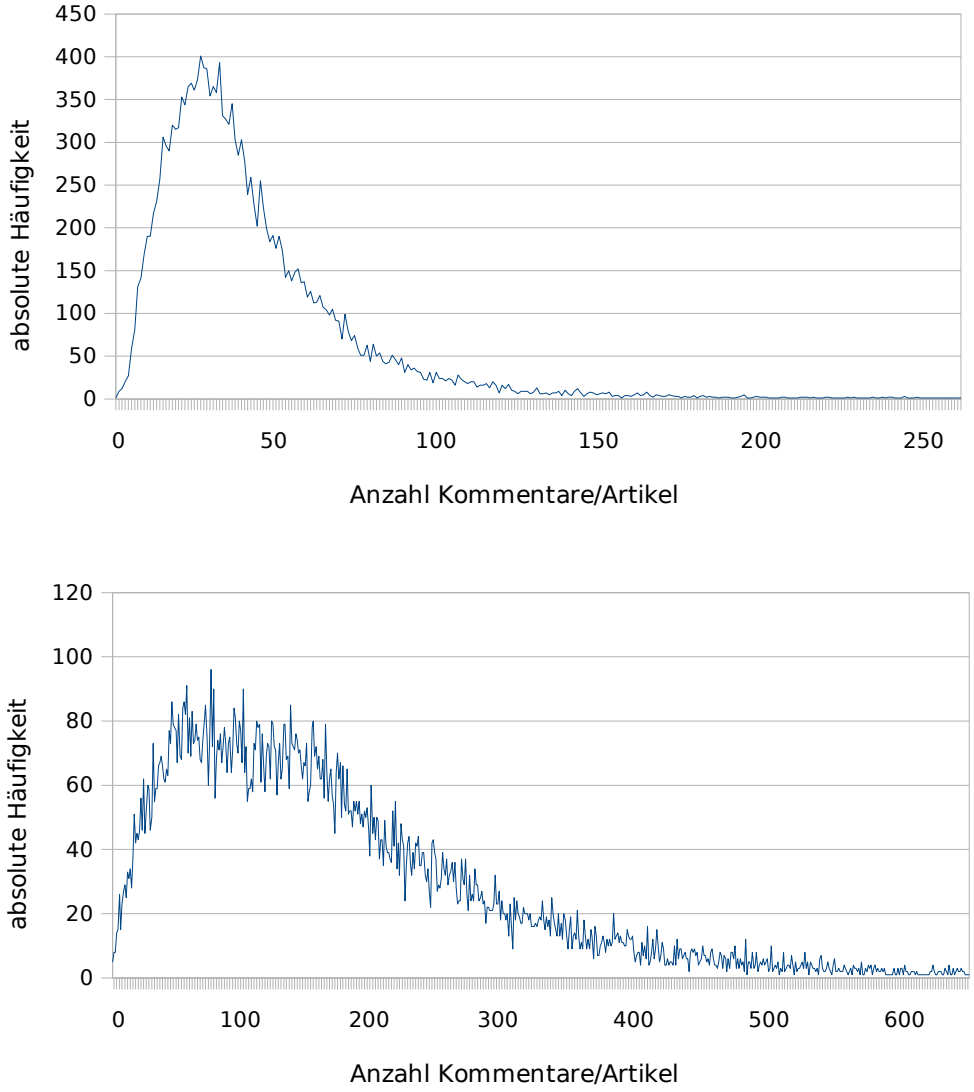

Abbildung 1: Verteilung der Kommentare im Slashdot-Korpus. Oben: Kommentare ersten Grades pro Artikel, unten: alle Kommentare pro Artikel.

keine Seltenheit. Ein Kommentar bei YouTube ist in der Regel sehr kurz und stellt die Meinung des Kommentators dar. Hier ist ein Ranking über thematische Relevanz nicht möglich, da die Kommentare zu kurz sind und eine Reaktion auf ein Video und nicht auf einen Text darstellen. Hier bietet sich an, eine Zusammenfassung der Kommentare zu generieren.

YouTube stellt eine gute Basis dar, um Retrieval-Modelle für meinungstragende Kommentare zu untersuchen. Dazu wurde eine Sammlung von YouTube-Kommentaren er-

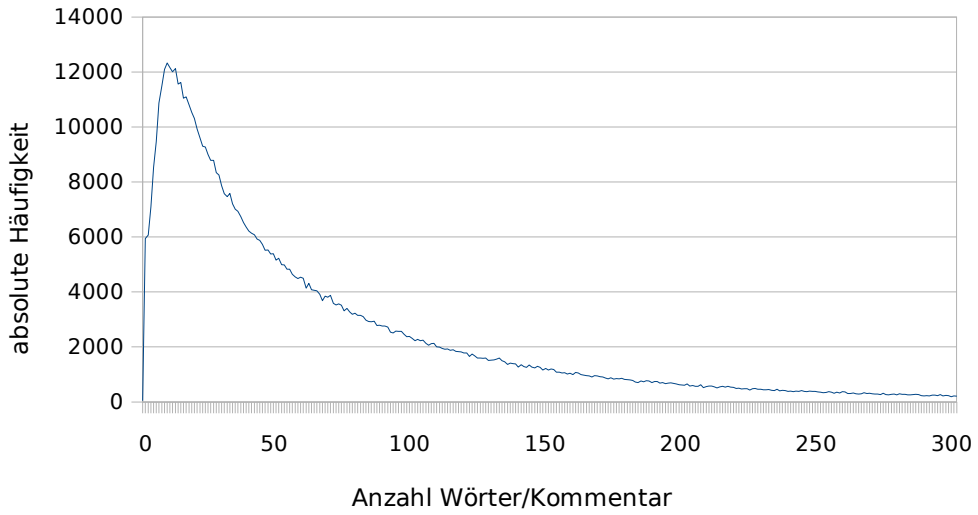

Abbildung 2: Verteilung der Länge der Kommentare ersten Grades im Slashdot-Korpus. Aus Gründen der Übersichtlichkeit ist das Diagramm gekürzt. Die nicht dargestellten Kommentare mit einer Länge von 300-1150 Wörtern machen 3,6 % aller Kommentare aus.

stellt. Über ein von YouTube bzw. Google zu Verfügung gestelltes Developer-API[8] wurden Kommentare zu 64.830 Videos gesammelt. Insgesamt konnten 9 Millionen Kommentare zusammengetragen werden. Es wurden alle sogenannte „Standard video feeds“ abgefragt. Dies sind jeweils die 100 besten Videos der Kategorien: „top rated“, „top favorites“, „most viewed“, „most popular“, „most discussed“, „most linked“, „most responded“, „most recent“, „recently featured“ und „watch on mobile“. Die Kategorien werden von YouTube zu Verfügung gestellt und ständig aktualisiert. Sie können jeweils für den aktuellen Tag, die aktuelle Woche, den aktuellen Monat und als Allzeitliste abgefragt werden. Zusätzlich wurden Suchanfragen mit einzelnen Buchstaben von A bis Z und den Zahlen 0 bis 9 gestellt. Über das Developer-API wurden für jede Anfrage maximal 1.000 Videos geliefert. Zu allen gesammelten Videos konnten jeweils maximal 1.000 Kommentare gecrawlt werden.

Vergleichbar mit dem in 3.1.1 vorgestellten Slashdot-Korpus ist die YouTube-Kommentarsammlung nicht. YouTube-Kommentare sind in unterschiedlichen Sprachen verfasst. Slashdot-Kommentare sind dagegen ausschließlich in englischer Sprache verfasst. Außerdem wird das Moderationssystem für Kommentare bei YouTube kaum genutzt. Dadurch stehen keine detaillierten Informationen über die Qualität der YouTube-Kommentare zu Verfügung.

[8] http://code.google.com/apis/youtube/overview.html

3.2 Evaluierung der Themen-Relevanzmodelle

Im Folgenden werden die in Kapitel 2.1 vorgestellten Modelle evaluiert.

3.2.1 Evaluierung der Ähnlichkeitsmodelle

Die Ähnlichkeitsmodelle messen die Ähnlichkeit der Kommentare zu dem Artikel, zu dem sie verfasst wurden. Die Ergebnisse zweier Experimente von zentraler Bedeutung werden hier vorgestellt. Zum einen werden die drei Modelle LSI, ESA und VSM zur Ähnlichkeitsberechnung verglichen, zum anderen wird die Ermittlung günstiger Parameter der zwei Modelle LSI und ESA demonstriert. Die Experimente wurden iterativ durchgeführt. Es wird in Experiment 1 gezeigt, dass die Slashdot-Kategorien in zwei Klassen eingeteilt werden können. Diese Klassifikation wird in Experiment 2 angewendet. Aus Gründen der Übersicht finden die günstigen Parameter aus Versuch 2 in Versuch 1 bereits Anwendung.

Die Basis der Experimente stellt der Slashdot-Korpus dar. Es wurden nur klassifizierte Kommentare ersten Grades ausgewählt, dabei mussten zu einem Artikel mindestens 40 solcher Kommentare existieren. Diese Bedingung wurde festgelegt, damit im LSI-Modell eine Projektion auf einen bis zu 40-dimensionalen Raum möglich wird. Scott Deerwester empfiehlt in seinem Papier [1] eine Reduktion auf 100 Dimensionen, jedoch existieren im Slashdot-Korpus zu wenige Artikel, die 100 oder mehr klassifizierte Kommentare ersten Grades aufweisen. Die Dimension 40 ergibt sich somit aus der Beschaffenheit des Slashdot-Korpus. Insgesamt enthält der Korpus 13.527 Kommentare, die diese Voraussetzungen erfüllen.

Der Korpus unterscheidet zwei verschiedene Klassifikationsschemata, einerseits die Einteilung in Kategorien und andererseits das Scoring. Da das Scoring nur die diskreten Werte -1 bis 5 annehmen kann, ist es möglich, auch diese Punkteverteilung als Klassifizierung zu betrachten.

Im ersten Experiment wurden die drei Verfahren VSM, LSI und ESA miteinander verglichen. Dabei dient das Vektorraummodell vorrangig als Baseline zum Vergleich von LSI mit ESA. Die Dimension des Konzeptraums im LSI-Modell wurde aus dem Intervall $[1, 40]$ mit 20 festgelegt. Die Insignifikanzschwelle des ESA-Modells liegt bei $0,01$. Diese Werte wurden in Experiment 2 als günstig ermittelt. In Abbildung 3 sieht man die Verteilung der Ähnlichkeiten innerhalb der Kategorien bzw. innerhalb der Scores für alle drei Modelle.

Es ist deutlich zu erkennen, dass Kommentare mit kleineren Score-Werten auch geringere Ähnlichkeiten zu dem Artikel aufweisen. Das betrifft insbesondere Kommentare mit

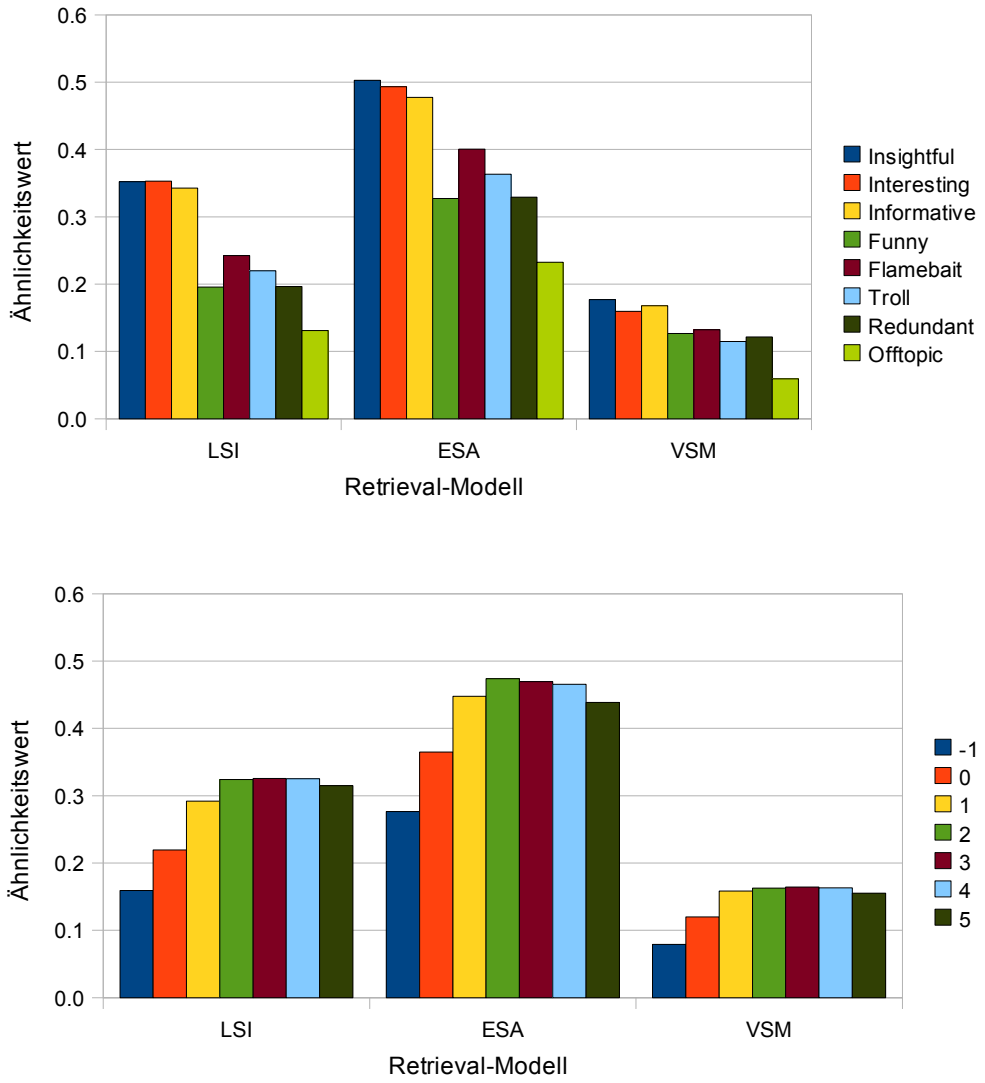

Abbildung 3: Ähnlichkeitsverteilung der Kommentare zum Artikel im Slashdot-Korpus innerhalb der Kategorien (oben) und innerhalb der Scores (unten). Darstellung jeweils für die Modelle LSI, ESA und VSM.

einem Score von -1 oder 0. Weiterhin ist zu erkennen, dass die verschiedenen Modelle den selben Verlauf beschreiben. LSI, ESA und VSM bilden zwar auf unterschiedliche Größenordnungen ab, die Tendenz ist bei allen Modellen aber die selbe.

Eine stärkere Aussage bezüglich der Qualität des Ähnlichkeitsmodells lässt sich anhand der Darstellung der Ähnlichkeiten innerhalb der Kategorien treffen (Abbildung 3 oben). Es ist zu sehen, dass die Kategorien *Insightfull*, *Interesting* und *Informative* eine deutlich größere Ähnlichkeit zum Artikel aufweisen, als die restlichen Kategorien. Untereinander unterscheiden sich die Ergebnisse dieser drei Kategorien jedoch kaum. Auch die

Werte der restlichen Kategorien unterscheiden sich nicht sehr stark. Daraus lesen wir, dass Kommentare der Kategorien *Insightfull*, *Interesting* und *Informative* (nachfolgend *I-Kategorien* genannt) eine größere inhaltliche Nähe zu dem Artikel haben als Kommentare der übrigen Kategorien.

Im zweiten Versuch wurde nach optimalen Parametern der Verfahren LSI und ESA gesucht. Dabei wird die Einteilung in *I-Kategorien* und *Andere Kategorien* aus Versuch 1 übernommen. Der wesentliche Parameter beim LSI-Modell ist die Dimension des Konzeptraums, auf die die Term-Dokument-Matrix reduziert wird. Es hat sich gezeigt, dass eine Dimension von 20 günstig ist (vgl. Abbildung 4 oben) [9] . Für das ESA-Modell wurden 10.000 Wikipedia-Artikel verwendet. Der Konzeptraum hatte somit 10.000 Dimensionen. Es hat sich in vorangegangenen Versuchen mit 100, 1.000 und 10.000 Dimensionen gezeigt, dass mit größeren Dimensionen bessere Ergebnisse erzielt werden. Die Größenordnung 10.000 stellt dabei einen Kompromiss zwischen Retrieval-Qualität und Rechen- bzw. Speicheraufwand dar. Der wichtigste Parameter im ESA-Modell ist der Insignifikanzschwellwert. Wird dieser bei der Projektion eines Dokumentvektors in den Konzeptraum in einer Dimension unterschritten, dann wird diese Dimension auf 0 gesetzt. Das ist wie folgt zu interpretieren: Ist die Ähnlichkeit eines Dokumentes zu einem Wikipedia-Artikel zu gering, so wird sie als nicht signifikant angesehen und nicht beachtet. Es hat sich ein Insignifikanzschwellwert von $0,01$ aus günstig erwiesen (vgl. Abbildung 4 unten).

3.2.2 Kontinuitätsmodell

Das Kontinuitätsmodell wurde unter anderem entwickelt, um die Probleme die das Ähnlichkeitsmodell mit sich bringt zu umgehen, denn würde der Artikel selbst als Kommentar abgegeben werden, hätte er die größte Ähnlichkeit zu sich selbst. Andererseits hat sich im Ähnlichkeitsmodell gezeigt, dass sehr unähnliche Kommentare im Slashdot-Korpus auch ein geringes Scoring erreichen und in Kategorien liegen, die keine oder wenig inhaltliche Relevanz bezüglich des Artikels haben. Diese zweite Eigenschaft bleibt im Kontinuitätsmodell erhalten, da im Extremfall Kommentare, die kein Wort mit dem Artikel gemeinsam haben, in diesem Modell auf den selben Wert abbilden wie im Ähnlichkeitsmodell mit dem ESA-Verfahren. Nun bleibt zu zeigen, dass die relevanten Kommentare auch einen größeren Wert im Kontinuitätsmodell erreichen.

[9] Die Boxplots zeigen den Bereich vom 25%-Quantil bis zum 75%-Quantil (Box), den Median (Strich in der Box), die kleinsten und größten Werte (Ausläufer) und die Ausreißer (graue Punkte, weiter als 1,5*Interquartilsabstand von den mittleren Quartilen entfernt). Die Parameter sind günstig, wenn die Boxplots weit gegeneinander verschoben sind und möglichst wenig streuen.

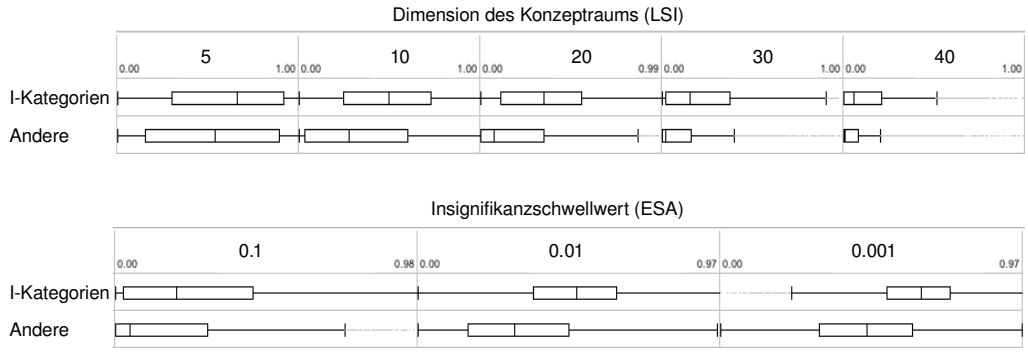

Abbildung 4: In dieser Abbildung sind die Parameter des LSI-Modells (oben) und des ESA-Modells (unten) variiert. Jeweils zeilenweise sind die Kategorien *Insightfull*, *Interesting* und *Informative* zusammengefasst zu einer Klasse (I-Kategorien) und die restlichen Kategorien ebenfalls zusammengefasst zu einer Klasse (Andere). Oben: Der wichtigste Parameter des LSI-Modells ist die Dimension des Konzeptraums. Anhand der Boxplots wurde die Dimension 20 als günstigste ausgewählt. Unten: Der wichtigste Parameter des ESA-Modells ist der Insignifikanzschwellwert. Für diesen wurde anhand der Boxplots 0,01 als günstig ausgewählt.

Dazu wurde für das erste Experiment der Versuchsaufbau des Ähnlichkeitsmodells übernommen. Abbildung 5 zeigt die Ergebnisse dieses Experiments. Deutlich erkennbar ist, dass die inhaltlich relevanten Kategorien *Insightfull*, *Interesting* und *Informative* größere Werte erreichen, als die restlichen, inhaltlich weniger relevanten Kategorien.

In der Definition des Kontinuitätsmaßes haben wir herausgearbeitet, dass es möglich ist, eine inhaltliche Vervollständigung zu messen. Um diese Aussage zu bekräftigen, haben wir ein weiteres Experiment durchgeführt. Es war dafür notwendig, eine andere Datenbasis zu schaffen. Zu 10 Slashdot-Artikeln wurden Kommentare so ausgewählt, dass bekannt war, welche Kommentare mit großer Wahrscheinlichkeit einen relevanten Informationsvorsprung haben und welche nicht. Die Konstruktion der Kommentare zu jedem Artikel geschah wie folgt:

1. Aus den zum jeweiligen Artikel abgegebenen Kommentaren wurden vier ausgewählt, die den *I-Kategorien* entstammen, und mindestens einen Score von 3 hatten. Bei diesen Kommentaren ist davon auszugehen, dass sie im Durchschnitt einen relevanten Informationsvorsprung haben.

2. Vier weitere Kommentare eines zufälligen anderen Slashdot-Artikels, die dort ebenfalls den *I-Kategorien* entstammen und ein Scoring von mindestens 3 hatten bildeten die zweite Gruppe. Diese entspringen immer noch der Domäne IT,

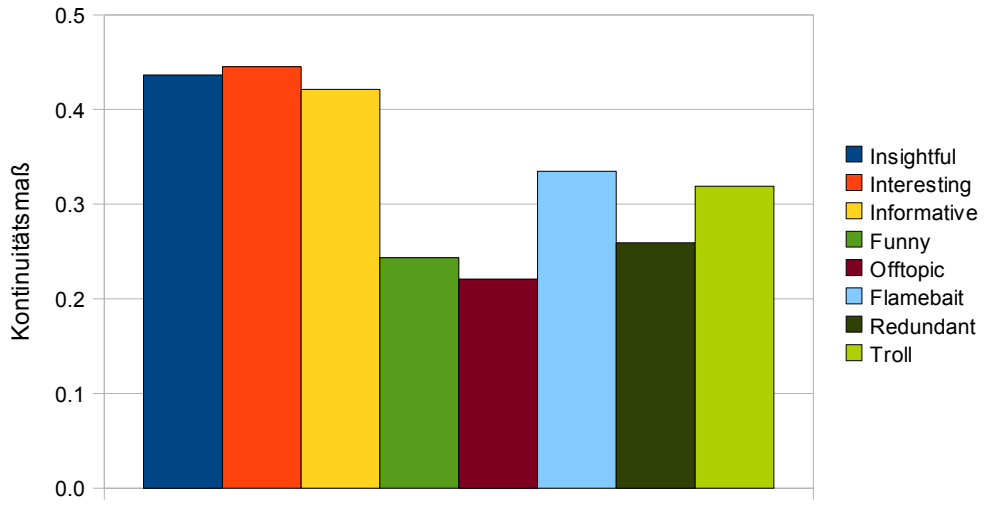

Abbildung 5: Das Kontinuitätsmaß im Verglich zwischen den Kategorien im Slashdot-Korpus bei einem Insignifikanzschwellwert von 0,01.

haben sonst aber höchstens zufällig thematisch etwas mit dem aktuellen Artikel gemeinsam.

3. Die letzte Gruppe bildet sich aus Abschnitten arbiträrer Nachrichtentexte, die verschiedenen Portalen wie *cnn.com*, *bbc.com*, *news.yahoo.com*, *reuters.com* und *news.google.com* entnommen wurden. Diese Texte haben tendenziell keinen Informationsvorsprung bezüglich des Themas des Artikels.

Es wurde darauf geachtet, dass die Texte alle etwa gleich groß sind. Die Länge lag bei 60 bis 120 Wörtern. Wir haben somit 10 Slashdot-Artikel mit jeweils drei Gruppen zu je vier Kommentaren, insgesamt 12 Kommentare pro Artikel. Abbildung 6 zeigt die Ergebnisse. Das Kontinuitätsmaß ergibt für Gruppe (1) der echten Kommentare einen deutlich größeren Wert. Ein entscheidender Parameter ist dabei die Insignifikanzschwelle des ESA-Modells. Bei einem Wert von $0,04$ ist der relative Unterschied von Gruppe (1) zu (2) und (3) am größten. Vergrößert man den Wert weiter, so ergibt das Maß häufiger 0, auch in Gruppe (1). Bei der manuellen Betrachtung der Einzelergebnisse der 10 Artikel haben wir einen günstigen Wert zwischen $0,03$ und $0,05$ ausmachen können.

Das Experiment kann als Indiz dafür betrachtet werden, dass das Kontinuitätsmaß in der Lage ist, die thematische Relevanz zu messen. Es ist zu beobachten, dass die Kommentare aus Gruppe (2), die anderen Slashdot-Artikeln entstammen, einen kleineren Kontinuitätswert ergeben als die echten Kommentare, dennoch wird eine größere Relevanz gemessen, als in Gruppe (3) der Nachrichtentexte. Tatsächlich sind die Kommentare

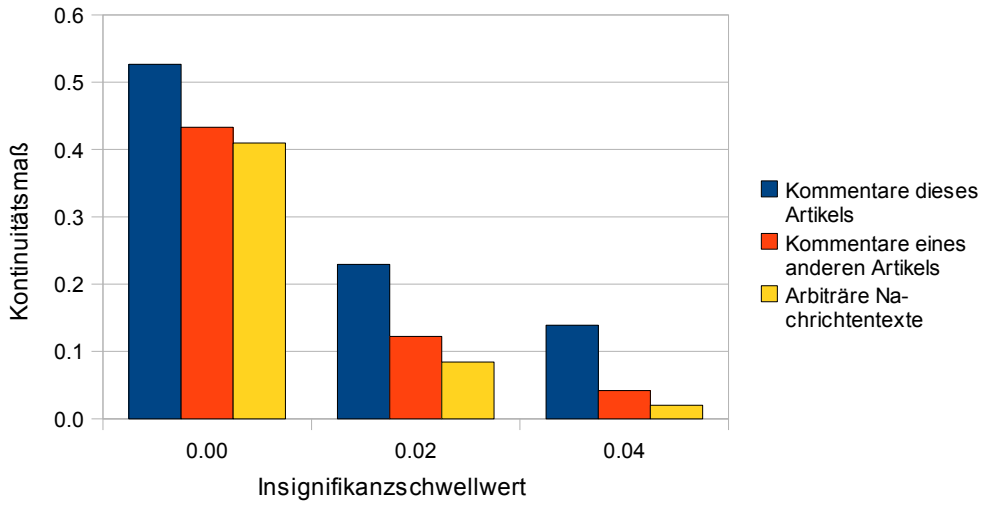

Abbildung 6: Aufgetragen ist das Kontinuitätsmaß, welches die thematische Vervollständigung verschieden konstruierter Kommentare misst. Auf der X-Achse wurde der Insignifikanzschwellwert des im Kontinuitätsmaß angewendeten ESA-Modells variiert.

der Gruppe (2) thematisch näher an dem Artikel, da sie dem selben Themenkomplex entstammen. Damit kann angenommen werden, dass auch Abstufungen richtig erkannt werden und das Maß die Kommentare in eine augenscheinlich sinnvolle Reihenfolge bringt.

3.3 Fallstudie zur Meinungsanalyse

Um die Qualität des vorgeschlagenen Modells für die Zusammenfassung von Meinungen aus Kommentaren objektiv zu bewerten, wird ein geeigneter Vergleichskorpus mit bereits identifizierten Meinungsäußerungen benötigt. Ein solcher Korpus steht uns für Kommentare nicht zu Verfügung und es bedarf eines erheblichen Aufwands einen solchen manuell zu erstellen. Aus diesem Grund entschieden wir uns zu einer Fallstudie. Es entstand eine Anwendung, die das vorgeschlagene Modell benutzt um die Anwendbarkeit zu zeigen und das Modell empirisch zu prüfen. Im Folgenden wird die Anwendung vorgestellt und die Arbeitsweise erläutert. Die Anwendung benutzt ein Meinungswörterbuch mit 1.812 positiven und 2.633 negativen Wörtern. Die Vorgehensweise bei der Erstellung des Wörterbuchs wird ebenfalls im Folgenden erläutert.

3.3.1 Opinion-Cloud für YouTube

Auf dem Internetportal YouTube werden wie in Abschnitt 3.1.2 beschrieben, Videos vorgestellt und kommentiert. Kommentare sind überwiegend sehr kurz. Kurze Kommentare stellen dabei keinen inhaltlichen Beitrag zum Thema, sondern reine persönliche Meinungsäußerung des Kommentators zum präsentierten Video dar. Ein Ranking der Kommentare nach thematischer Relevanz ist nicht möglich. Um trotzdem Übersicht bei sehr vielen Kommentaren zu einem Thema zu schaffen, bietet sich hier als neue Lösung die Generierung einer Zusammenfassung der Meinungen an. Dafür entwickelten wir eine Anwendung in Form einer plattformunabhängigen Browsererweiterung für den Webbrowser Firefox[10].

Die Zusammenfassung wird in Form einer Opinion-Cloud oder auch Meinungswolke präsentiert. Eine Opinion-Cloud ist, ähnlich einer Tag-Cloud oder Schlagwortwolke, eine flächige Darstellung von Wörtern mit unterschiedlicher Schriftgröße. Eine solche Wortwolke ermöglicht es, schnell die wichtigsten Wörter anhand ihrer Größe zu erfassen und auch einen Überblick über alle Wörter zu erhalten. Die Schriftgröße eines Wortes $s(w)$ in einer Wortwolke richtet sich nach der Auftrittshäufigkeit des Wortes $c(w)$. Je häufiger ein Wort in den Kommentaren vorkommt, desto größer wird es dargestellt. Die Schriftgröße wird dabei relativ zwischen der größten auftretenden Häufigkeit c_{max} und der geringsten auftretenden Häufigkeit c_{min} errechnet.

$$s(w) = \frac{s_{max} * (c(w) - c_{min})}{(c_{max} - c_{min}))} \; f\ddot{u}r \; c(w) > c_{min} \; sonst \; c(w) = 1 \tag{8}$$

Befinden sich sehr viele Wörter in einer Wortwolke, ist eine logarithmische Skalierung der Schriftgröße sinnvoll, da die Auftrittshäufigkeit der Wörter einem Potenzgesetz folgt: sehr wenige Wörter treten sehr häufig auf und sehr viele Wörter dagegen sehr selten. Die beste Übersichtlichkeit über die Wörter einer Wortwolke ergibt sich, wenn die größten Wörter zentral positioniert werden und die Größe der anderen Wörter nach außen hin abnimmt.

In unserer Opinion-Cloud werden im Gegensatz zu einer Tag-Cloud zwei Wortwolken nebeneinander dargestellt. Auf der einen Seite stehen dabei die positiven Äußerungen in grüner Farbe und auf der anderen die negativen in rot. Als alternative Möglichkeit der Darstellung wird eine Wortwolke erzeugt, die alle Wörter der Kommentare mit farbiger Hervorhebung der meinungstragenden Wörter enthält. Abbildung 7 stellt beide Varianten gegenüber. Die Wortwolke über alle Wörter der Kommentare dient hierbei auch

[10] http://www.mozilla-europe.org/de/firefox/

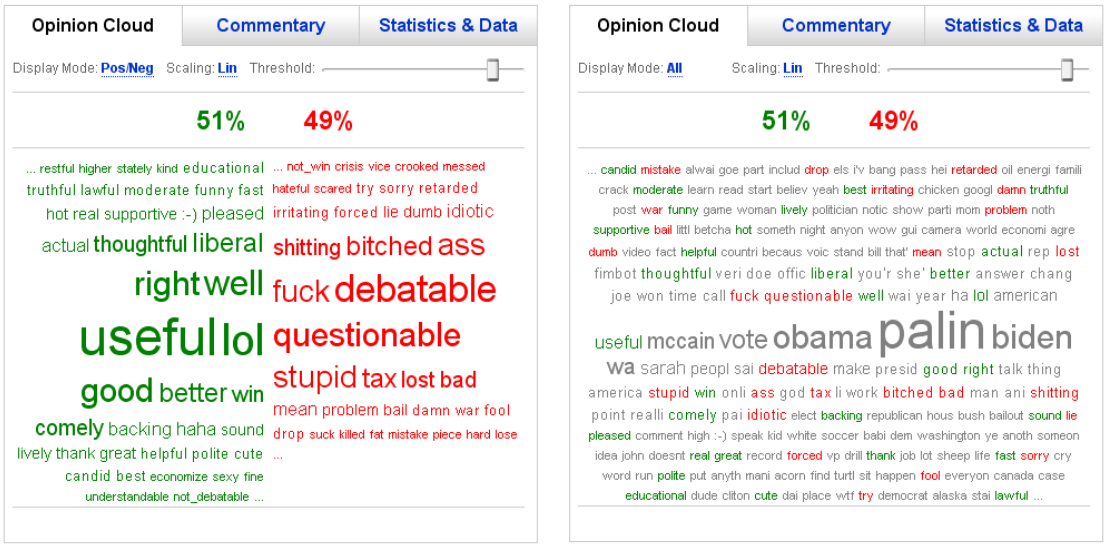

Abbildung 7: Opinion-Cloud: links über positive und negative Wörter; rechts über alle Wörter der Kommentare

der Evaluierung. Wörter die keine oder keine eindeutige Meinung ausdrücken werden in grau dargestellt. So kann intuitiv überprüft werden, ob in der Zusammenfassung alle relevanten Wörter stehen.

Zusätzlich zur Opinion-Cloud wird die prozentuale Verteilung von positiven und negativen Meinungsäußerungen als Zahlenwert in der Browsererweiterung über der Wolke dargestellt. Die Browsererweiterung stellt dem Nutzer außerdem verschiedene Einstellungsmöglichkeiten zu Verfügung, um die Darstellung zu beeinflussen. Es besteht die Möglichkeit zwischen der Meinungswolke und der Wortwolke aller Wörter per Link zu wechseln. Die Skalierung der Schriftgröße der Wörter ist logarithmisch und linear möglich und kann ebenfalls per Link gewechselt werden. Als dritte Interaktion existiert ein Schieberegler, über den ein Schwellwert für die Mindestanzahl des Auftretens eines Wortes einstellbar ist um in der Wolke dargestellt zu werden. Dies ist bei Wortwolken mit sehr vielen Wörtern sehr nützlich um die Übersicht über die häufigsten Wörter zu gewährleisten. Alle Bedienelemente sowie die Wortwolken sind in die graphische Gestaltung des YouTube-Portals integriert, um die Nutzung so intuitiv wie möglich zu machen. Abbildung 8 gibt einen Eindruck im Anwendungsfall.

Für die Erstellung der Meinungszusammenfassung nutzt die Browsererweiterung das in Kapitel 2.2.3 vorgeschlagene Modell. Dabei wird ein wörterbuchbasierter Ansatz verfolgt, der einzelne meinungstragende Wörter identifiziert. Die einzelnen Schritte bei der Analyse sehen wie folgt aus:

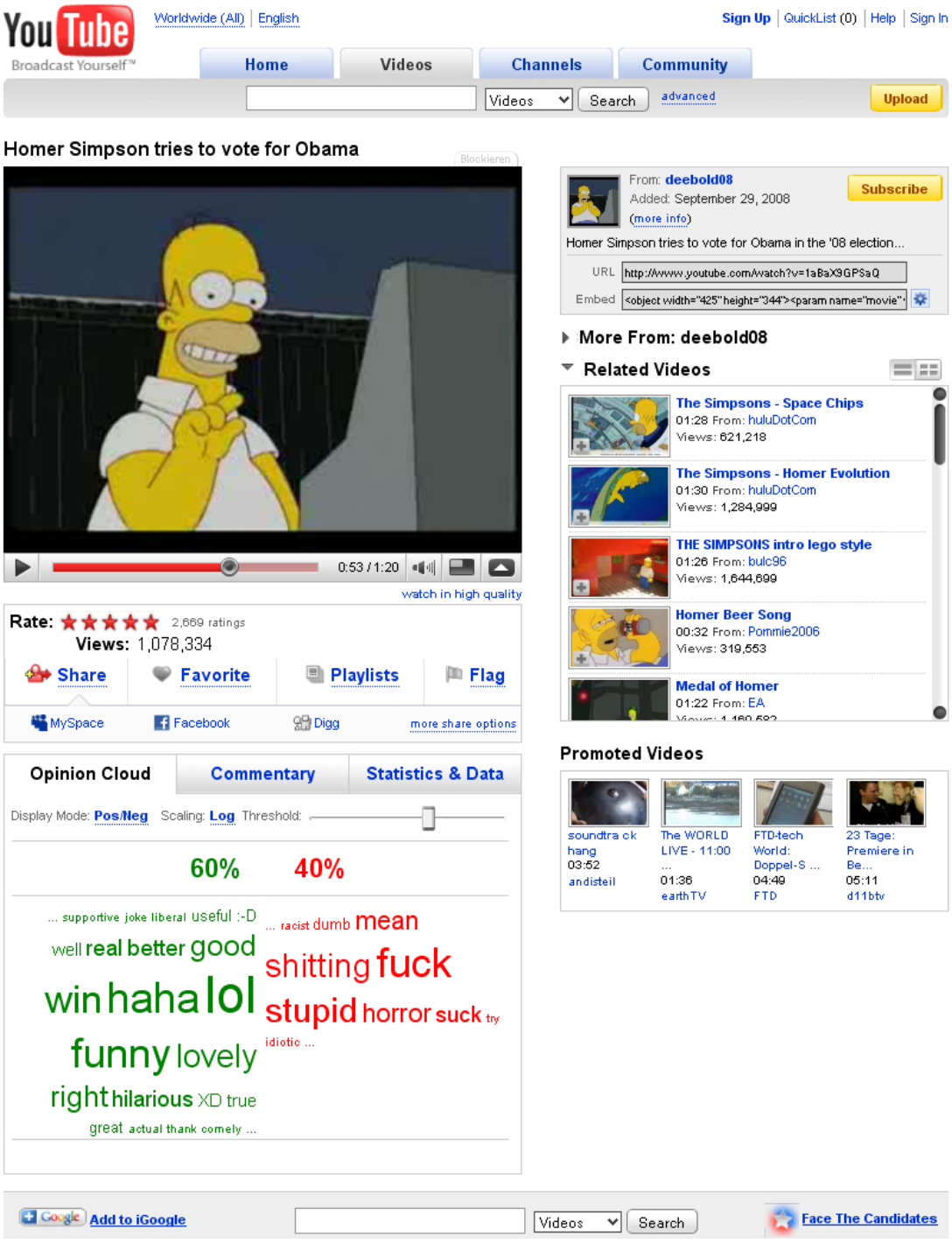

Abbildung 8: Gesamtübersicht einer YouTube-Seite mit Opinion-Cloud

1. Laden aller Kommentare zum angezeigten Video über ein von YouTube bzw. Google zu Verfügung gestelltes Developer-API[11].

[11] http://code.google.com/apis/youtube/overview.html

Die Einschränkung bei der Nutzung des YouTube-API ist, dass maximal 999 Kommentare zu einem Video abgefragt werden können. Die Kommentare werden aus angeforderten XML-Dokumente mit jeweils maximal 50 Kommentaren extrahiert.

2. Filterung von Kommentaren mit mehr als 2 Sätzen bzw. mehr als 30 Wörtern.
   Die intuitive Annahme ist, dass bei Kommentaren mit weniger als 3 Sätzen, der Autor auch der *Meinungsinhaber* ist und dass der *Meinungsgegenstand* (vgl. 2.2) auf den sich die Meinung bezieht nur das Video sein kann. Bei mehr als 3 Sätzen müssten subjektive von objektiven Aussagen unterschieden werden.

3. Identifizierung von Emoticons:
   Emoticons sind Kombinationen von Zeichen, die als Symbol verstanden werden. Meist werden Smileys nachgebildet, die Gefühle und Emotionen ausdrücken. Beispielsweise steht „:-)“ für Freude, Zustimmung und andere positive Emotionen. Die Erkennung erfolgt durch ein Wörterbuch, das eine Auswahl gebräuchlicher Emoticons verschiedener Quellen enthält [12] [13]. Die identifizierten Emoticons werden entsprechend der Orientierung in die Wortvektoren eingeführt.

4. Zusammenfassen von Buchstabenwiederholungen:
   Buchstaben mehrfach zu wiederholen ist eine beliebte Methode, das Geschriebene zu bekräftigen. Die sehr beliebte Abkürzung „lol“ für „laugh out loud“ kommt mit beliebiger Buchstabenwiederholung wie „loooool“ oder „lololol“ vor. In unserem Ansatz findet diese Ausdrucksform keine Berücksichtigung in der Bewertung der geäußerten Meinung. Sie sollte aber in Zukunft einfließen.

5. Entfernung aller Zeichen, die keine Buchstaben sind:
   In den folgenden Schritten werden nur Kleinbuchstaben verarbeitet. Hier fallen auch Kommentare in Sprachen weg, die kein ASCII-Zeichen verwenden. Dies dient der Vereinfachung der Bearbeitung und der Reduzierung der Wortanzahl. Das benutzte Wörterbuch beschränkt sich zur Zeit auf Wörter der englischen Sprache. Eine Erweiterung auf andere Sprachen soll in Zukunft folgen.

6. Entfernung von Stoppwörtern

7. Stammformreduktion:
   Wörter in verschiedener Beugung werden auf ihre Grundform reduziert. Beispielsweise wird „hating“ und „hated“ zu „hate“ reduziert.

[12] http://en.wikipedia.org/wiki/Emoticon
[13] http://www.greensmilies.com/smilie-lexikon/

8. Erzeugen der Wortvektoren für die Orientierungen *positiv* und *negativ* über ein Wörterbuch.
   Es wird für jedes Wort des Kommentars geprüft, ob es im Wörterbuch vorhanden ist und welche Orientierung es besitzt. Außerdem werden die vier vorausgehenden Wörter auf das Vorhandensein einer Negation geprüft. Diese Negationen sind „no“, „not“, „don't“, „neighter“ und „nor“. Das Wort wird in den entsprechenden Wortvektor aufgenommen.

9. Darstellung der Wortvektoren:
   Alle Wortvektoren der beiden Orientierungen werden zusammengefasst und dargestellt. Die Schriftgröße wird abhängig von der Auftrittshäufigkeit skaliert. Die Skalierung erfolgt relativ, das häufigste Wort erhält die größtmögliche Schriftgröße und umgekehrt. Außerdem ist eine lineare bzw. logarithmische Skalierung durch den Nutzer wählbar.

3.3.2 Erstellung eines Meinungswörterbuchs

Das Wörterbuch zur Meinungserkennung basiert auf ausgewählten Kategorien des *General Inquirer (GI)*[18]. Der General Inquirer wurde in den sechziger Jahren von Philip Stone und anderen am MIT (Massachusetts Institute of Technology) als Software zur Inhaltsanalyse entwickelt. Aus den 182 Kategorien entnahmen wir alle Wörter der beiden Kategorien „positive“ und „negative“, die Wörter positiver bzw. negativer semantischer Orientierung enthalten. Die beiden Listen wurden gefiltert, da Wörter mit mehrfacher Bedeutung in beiden Kategorien enthalten sind. Außerdem wurden nur Wörter gewählt, die im YouTube-Korpus vorkommen. So entstand ein Basiswörterbuch mit 1615 positiven und 1938 negativen Wörtern. Auszüge sind in Tabelle 2 zu sehen.

Um das Wörterbuch mit neuen meinungstragenden Wörtern zu erweitern, wurden auf den beiden Korpora von YouTube und Slashdot alle Adjektive mit einem Part-Of-Speech-Tagger extrahiert. Ein POS-Tagger ordnet zu Wörtern eines Textes die Wortart wie Substantiv, Adjektiv, Verb zu. Hierzu werden stochastische Modelle, wie das Hidden Markow Modell verwendet. Für unsere Untersuchung nutzten wir QTAG, ein Java POS-Tagger der Universität Birmingham[14]. In den über Neun Millionen Kommentaren der YouTube-Kommentarsammlung wurden ca. 168.000 verschiedene Adjektive durch den POS-Tagger identifiziert. Diese Sammlung schließen auch alle gefundenen Adjektive aus dem Slashdot-Korpus ein.

[14] http://www.english.bham.ac.uk/staff/omason/software/qtag.html

Tabelle 2: Auszüge der benutzten Kategorien des *General Inquirer*.

Wörter mit positiver semantischer Orientierung:
able, abound, absolve, absorbent, absorption, abundance, abundant, accede, accentuate, accept, acceptable, acceptance, accessible, accession, acclaim, acclamation, accolade, accommodate, accommodation, accompaniment, accomplish, accomplishment, accord, accordance, accountable, accrue, accuracy, accurate, accurateness, achieve, achievement, acknowledgement, ...

Wörter mit negativer semantischer Orientierung:
abandonment, abate, abdicate, abhor, abject, abnormal, abolish, abominable, abrasive, abrupt, abscond, absence, absent, absent-minded, absentee, absurd, absurdity, abuse, abyss, accident, accost, accursed, accusation, accuse, ache, acrimonious, acrimony, addict, addiction, admonish, admonition, adulteration, ...

Kommentare werden oft in Eile verfasst und enthalten daher viele Tipp- und Rechtschreibfehler. Es befinden sich sehr viele gleiche Adjektive in unterschiedlicher, falscher Schreibweise im Korpus. In Tabelle 3 ist eine Auswahl verschiedener Schreibweisen von *beautiful* zusehen. Insgesamt wurden 137 verschiedene Schreibweisen mit einer Suche nach Wörtern, die mit b anfangen und e,a,u,t,i,f und l enthalten gefunden. Bei einer automatischen Erstellung eines Wörterbuches und auch bei der Benutzung des Wörterbuchs zur Identifizierung von Wörtern stellt dies ein Problem dar. Hier sollten in Zukunft automatische Korrektursysteme verwendet werden. Die Korrektur sollte dabei über phonetische Ähnlichkeit erfolgen wie beispielsweise bei SmartSpell[15]. Eine falsche Schreibweise, die dem richtigen Wort phonetisch sehr ähnlich ist, wird bei *Slang* oft auch bewusst verwendet beispielsweise „kewl“ für „cool“. Die Wörter können dabei im Sinn auch verändert werden: ein “gansta“ ist laut *Urban Dictionary*[16] eine Übertreibung von „gangster“.

Um das Meinungswörterbuch mit weiteren Wörtern zu vergrößern, nutzten wir das von Turney[19] vorgeschlagene PMI-Verfahren, das in Kapitel 2.2.2 bereits vorgestellt wurde. Dabei verwendeten wir die Websuche von *Yahoo!*[17], da diese ein API zu Verfügung stellt[18] und den $NEAR$-Operator bei der Suche zulässt. PMI wurde für ein Wort zu je 7 Wörtern mit bekannter Orientierung berechnet. Dies waren die von Turney[20] vorge-

[15] http://www.artsystems.de/smartspell/index.htm
[16] http://www.urbandictionary.com/define.php?term=gansta
[17] http://yahoo.com
[18] http://developer.yahoo.com/search/

Tabelle 3: Auszug aus einer Liste gefundener Schreibweisen von „beautiful“.

beautiflu, baautiful, beauuttifull, beuataful, beaeuiful, buietiful, beaiutful, butafull, butilful, buttaful, beuutiful, butiiful, buetefull, beautiaful, beituful, beautiull, beautiieful, beaititiful, buteaful, butteful, buteaiful, beautifeel, beeaauutifull, beautitiful, beaauutiful, beutleful, bealtful, buatefful, beautifal, beaatuuiful, beutufull, ...

Tabelle 4: Ergebnisse der PMI-Analyse über den gewählten Kategorien des *General Inquirer*.

	Richtige Orientierung		**Falsche Orientierung**	
	Absolut	Prozent	Absolut	Prozent
negative	1647	84,98 %	291	15,02%
positive	1336	82,72 %	279	17,28%
alle	2983	83,96 %	570	16,04%

schlagenen und getesteten Startwörter:

1. Wörter mit positiver semantischer Orientierung:
   good, nice, excellent, positive, fortunate, correct, superior

2. Wörter mit negativer semantischer Orientierung:
   bad, nasty, poor, negative, unfortunate, wrong, inferior

Das bedeutet, dass für jedes unbekannte Wort w 14 Anfragen mit $w\ NEAR\ Startwort$ und einer Anfrage mit w allein an die Suchmaschine gestellt wurden und die Trefferanzahl, wie in Kapitel 2.2.2 beschrieben ausgewertet wurde.

Zunächst testeten wir das Verfahren PMI anhand des benutzten GI-Wörterbuches. Mit den 14 benutzten Wörtern der beiden Orientierungen konnten 83.96% der benutzten Wörter des General Inquirers der richtigen Orientierung zugeordnet werden. Tabelle 4 zeigt die Ergebnisse für beide Orientierungen. Turney [20] erreichte eine Genauigkeit von 82.8% bei der Klassifizierung der beiden gleichen Kategorien des GI-Wörterbuchs mit der Benutzung der Suchmaschine *Altavista*[19].

[19] http://www.altavista.com/

Bei der Benutzung des PMI-Verfahrens stellte sich heraus, dass die vielen falsch geschriebenen Wörter der Adjektivliste des Korpus zu Problemen führte. Auch für diese Wörter wird ein Wert für PMI berechnet, da diese Wörter auch unterschiedlich häufig im Korpus der Suchmaschine vorkommen. Für das Wörterbuch hieße das aber, das unterschiedliche Schreibweisen eines Wortes gesondert enthalten sind und bei einer Zusammenfassung auch gesondert behandelt werden. Aus diesem Grund war es notwendig, die Ergebnisse des Verfahrens manuell zu filtern. Falsche Schreibweisen wurden dabei erst einmal nicht berücksichtigt. Um den Aufwand der manuellen Filterung zu senken, wurden nur Wörter berücksichtigt, die einen PMI-Wert von größer 6 für positive semantische Orientierung bzw. kleiner -6 für negative berücksichtigt. Dieser Schwellwert erwies sich als guter Kompromiss. Wörter mit einem PMI-Wert außerhalb des Schwellwertes sind meist nicht mehr eindeutig einer Orientierung zuordbar. Zu den 20.000 häufigsten der 168.000 auftretenden Adjektive wurde mit dem Verfahren die semantische Orientierung bestimmt. Das Meinungswörterbuch konnte um 550 negative und 197 positive Wörter erweitert werden.

Durch die Benutzung einer automatischen Rechtschreibkorrektur, kann die Menge der zu prüfenden Wörter eingeschränkt werden um das Verfahren vollständig zu automatisieren. Die Qualität des POS-Taggers muss dafür aber auch verbessert werden, da in der Adjektivliste auch andere Wortklassen enthalten sind. Wörter anderer Wortklassen, wie Substantive können zwar einen aussagekräftigen PMI-Wert haben, trotzdem nicht unbedingt als eindeutige Meinung auftreten, da beim PMI-Verfahren nur die Kookkurenz betrachtet wird. Das gesuchte Substantiv kommt beispielsweise häufig mit den gegebenen positiven Wörtern vor, hat aber selbst nur in Verbindung mit einem positiven Adjektiv auch eine positive Bedeutung. Des Weiteren hatten auch offensichtlich objektive Adjektive einen aussagekräftigen PMI-Wert. Beispielsweise „slovenian“ und „californian“ hatten PMI-Werte größer zehn und „shiite“ mit -12 einen eindeutig negativen.

Die Qualität des Wörterbuchs kann nur empirisch über die Anwendung evaluiert werden, da ein Vergleichskorpus fehlt. Auf dem YouTube-Korpus kann man zeigen, dass 59,6% der auftretenden Adjektive durch das Wörterbuch erkannt werden, wenn man die Auftrittswahrscheinlichkeit berücksichtigt. Dies bedeutet, dass ein zufällig gezogenes Adjektiv aus diesem Korpus mit einer Wahrscheinlichkeit von ca. 60% in unserem Wörterbuch enthalten ist. Das Wörterbuch deckt also einen beträchtlichen Teil der häufig auftretenden Adjektive ab, wenn man berücksichtigt, dass nicht alle Adjektive meinungstragend bzw. nicht eindeutig meinungstragend sind. Tabelle 6 zeigt die 20 häufigsten Adjektive und deren semantische Orientierung.

Tabelle 5: Auszüge der gefundenen meinungstragenden Wörter aus dem YouTube-Korpus.

Wörter mit positiver semantischer Orientierung:
academic, affordable, alright, analytical, appreciative, atmospheric, breezy, brillant, calming, chill, colourful, comedian, comfortable, comprehensible, comprehensive, concise, conformable, cosy, crafted, creamy, delicious, delightful, desirable, dramatic, dreamy, durable, eclectic, emotional, enjoyable, enticing, evergreen, extraordinary, ...

Wörter mit negativer semantischer Orientierung:
abased, abhorrent, aborted, abusive, accident, accused, adulterated, aggresive, agoraphobic, alleged, allergic, amateurish, amputate, amputated, anal, antagonistic, antisocial, apathetic, appalled, appalling, asinine, ass, assed, asshole, atrocious, atrophied, attacking, authoritarian, badass, barbaric, bashing, beastly, ...

3.4 Evaluierung der Stilanalyse

Zu Beginn definieren wir wichtige Begrifflichkeiten:

Maschinelles Lernen Maschinelles Lernen ist die Fähigkeit eines Programms, auf der Grundlage eines Qualitätsmaßes durch die Einbeziehung vorheriger Ergebnisse (Erfahrung) die Qualität der Bearbeitung einer Aufgabe zu steigern.

Naive-Bayes-Klassifikator Der Naive-Bayes-Klassefikator ist ein Programm, dass ein Eingabedatum einer Klasse zuordnet. Die Menge der möglichen Klassen ist dem Programm vorher bekannt. Die Parameter des Klassifikators wurden mit einem statistischen Verfahren des maschinellen Lernens auf Grundlage von Beispielen ermittelt.

Precision Precision ist ein Maß, das insbesondere für die Bewertung der Qualität einer Klassifikation durch einen Klassifikator verwendet wird. Dabei wird das Verhältnis der *richtig einer Klasse zugeordneten Instanzen* zu *allen dieser Klasse zugeordneten Instanzen* gebildet.

Recall Recall ist ein weiteres wichtiges Maß zur Beurteilung einer Klassifikationsleistung. Hier wird das Verhältnis *der richtig einer Klasse zugeordneten Instanzen* zu *allen Instanzen* gebildet, *die dieser Klasse angehören* (aber nicht zwingend richtig zugeordnet wurden).

Tabelle 6: Dargestellt sind die 20 häufigsten, mit einem POS-Tagger ermittelten Adjektive des YouTube-Korpus mit Auftritshäufigkeit und semantsicher Orientierung. Die semantische Orientierung wurde mit dem erstellten Meinungswörterbuch identifiziert.

Auftritshäufigkeit in %	Wort in Stammform	semantische Orientierung
4.24607	good	positiv
3.11449	best	positiv
3.11228	great	positiv
2.12496	funni	positiv
2.09968	awesom	positiv
1.83663	amaz	positiv
1.59557	nice	positiv
1.17252	real	positiv
1.14183	other	–
1.05928	cool	positiv
1.05393	bad	negativ
1.02061	beauti	positiv
0.98401	just	–
0.95571	cute	positiv
0.95382	fuck	negativ
0.95229	old	–
0.93905	veri	–
0.88597	onli	–
0.86393	hot	positiv
0.80711	stupid	negativ

F-Measure Das F-Measure ist ein Maß, welches Precision und Recall miteinander Kombiniert. Es wird das gewichtete harmonische Mittel gebildet:

$$F = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \tag{9}$$

In Kapitel 2.3.2 haben wir zwei Möglichkeiten vorgestellt, die Stilanalyse als bewertendes Maß zu benutzen: heuristischer Ansatz und empirischer Ansatz. Wir haben den empirischen Ansatz für unsere Experimente benutzt.

Die Grundlage für unsere Tests bildete der Slashdot-Korpus. Um die Einheitlichkeit beizubehalten haben wir den selben Versuchsaufbau verwendet, der bei den Inhaltsmodellen zum Einsatz kam. In ersten Versuchen haben wir die Mittelwerte aller Stil- und Vandalismusmerkmale berechnet um die Möglichkeiten abzuschätzen, mittels maschinellen Lernens die Kategorien im Korpus voneinander zu trennen und es wurden verschiedene Versuche mit mehreren Lernverfahren durchgeführt. Dabei wurden verschiedene Varianten der Zusammenfassung von Kategorien getestet. Es hat sich herausgestellt, dass eine einzige sinnvolle Einteilung in zwei Klassen gute Ergebnisse liefert:

1. I-Kategorien: *Insightfull*, *Interesting* und *Informative*
2. Andere Kategorien: *Offtopic*, *Flamebait*, *Redundant*, *Troll* und *Funny*

Überraschenderweise hat sich erwiesen, dass die I-Kategorien in einem durchweg anderen Schreibstil verfasst sind als die übrigen Kategorien. Die Überprüfung mit Lesbarkeitsmaßen (z.B. Gunning-Fog-Index), zeigt dass der Leser der Kommentare der I-Kategorien durchschnittlich 11 Jahre Schulbildung hinter sich haben muss, zum Lesen der anderen Kommentare reichen $8,5$ Jahre. Dabei fällt auf, dass auch die lustigen Kommentare (Kategorie Funny) in schlechterem Stil verfasst sind. Eigentlich sollten diese auch in die Klasse der positiv orientierten Kommentare gefasst werden. Es ist zu vermuten, dass die Qualität lustiger Kommentare nicht in gutem Stil verankert ist. In Tabelle 7 sind die Ergebnisse der Klassifizierung mit den in Kapitel 2.3.1 vorgestellten Stil- und Vandalismusmerkmalen und einem trainierten Naive-Bayes-Klassifikator dargestellt.

Dabei ist zu erkennen, dass über 80% der Kommentare der *I-Kategorien* und über 55% der Kommentare der *Anderen Kategorien* richtig zugeordnet wurden. Insgesamt bessere Ergebnisse werden erzielt, wenn die Stil- und Vandalismusmerkmale kombiniert werden.

Auffällig ist, dass das wichtigste Merkmal (welches am schärfsten die Klassen trennt), das *TermCount*-Merkmal ist. Die Länge der Kommentare im Korpus hat entscheidenden Einfluss auf die Klassifizierung. Problematisch ist dieses Verhalten, da die Länge

Tabelle 7: Ergebnisse der Klassifikation der in Kapitel 3.2.1 definierten Kommentarmenge.

	Precision	**Recall**	**F-Measure**
Stilmerkmale			
I-Kategorien	0,801	0,836	0,818
Andere Kategorien	0,610	0,553	0,580
Vandalismusmerkmale			
I-Kategorien	0,815	0,824	0,819
Andere Kategorien	0,621	0,606	0,613
Stil- und Vandalismusmerkmale			
I-Kategorien	0,823	0,810	0,817
Andere Kategorien	0,614	0,633	0,623

eines Textes (insbesondere bei sehr kurzen Texten) implizit in vielen Stilmerkmalen mit verrechnet wird.

Um zu überprüfen, ob die Stilanalyse auch unabhängig von der Länge der Texte funktioniert, haben wir einen weiteren Versuch angeschlossen. Dabei haben wir nur Kommentare ausgewählt, die eine Länge von 90 bis 110 Termen haben. Wir haben diese Spanne gewählt, da die durchschnittliche Länge aller Kommentare im vorigen Versuch bei 98 Termen lag. Die größere der beiden zusammengefassten Klassen wurde anschließend verkleinert, so dass die beiden Klassen gleich groß waren. Damit ist die a-priori Wahrscheinlichkeit für beide Klassen gleich groß. Tabelle 8 zeigt die Ergebnisse der Klassifikation dieses Experiments.

Die Klassifikationsleistung bei fester Kommentarlänge ist schlechter. Bei der Kombination aus Stil- und Vandalismusmerkmalen wurden knapp 73% der Instanzen der I-Kategorien richtig zugeordnet, für die andere Klasse ergab sich ein Recall von 50%. Damit wird einerseits der Einfluss der Kommentarlänge deutlich, andererseits ist damit gezeigt, dass die Länge eines Kommentars nicht das einzige Merkmal ist, das bei der Klassifikation Einfluss nimmt.

Tabelle 8: Klassifikationsleistung über eine Menge von 4.188 Kommentare ersten Grades im Slashdot-Korpus mit einer Termanzahl von 90 bis 110.

	Precision	**Recall**	**F-Measure**
Stilmerkmale			
I-Kategorien	0,563	0,677	0,615
Andere Kategorien	0,595	0,475	0,528
Vandalismusmerkmale			
I-Kategorien	0,550	0,863	0,672
Andere Kategorien	0,682	0,293	0,410
Stil- und Vandalismusmerkmale			
I-Kategorien	0,594	0,727	0,654
Andere Kategorien	0,648	0,503	0,567

4 Zusammenfassung und Ausblick

Wir haben in dieser Arbeit eine Übersicht über das Problem der zunehmenden *Flut von Kommentaren im Web* hergestellt. Wir haben drei Retrieval-Aufgaben herausgearbeitet: Filterung, Zusammenfassung und Ranking von Kommentaren. Orthogonal dazu schlagen wir drei Relevanzkriterien vor, die einzeln oder in Kombination für die Bearbeitung jeder der drei Aufgaben in Frage kommen. Diese Kriterien sind: (1) die thematische Relevanz, (2) die Subjektivität des Kommentars bzw. die Polarität der Meinungsäußerung sowie (3) die Qualität des Textes.

Für jedes dieser Kriterien haben wir ein oder mehrere Modelle vorgestellt. Zur Berechnung der thematischen Relevanz wurde ein klassisches Ähnlichkeitsmodell vorgeschlagen, sowie ein neues Maß entwickelt: das Kontinuitätsmaß. Das Ähnlichkeitsmodell hat sich bei der Evaluierung als brauchbar herausgestellt, jedoch trägt es ein nicht zu vernachlässigendes inhaltliches Problem mit sich: Der Artikel würde zu sich selbst den besten Kommentar darstellen. Somit kann das Ähnlichkeitsmaß bei der Kombination mehrerer Modelle eine Rolle spielen, einzeln betrachtet ist es nur bedingt einsetzbar. Für das Kontinuitätsmaß konnte gezeigt werden, dass es innerhalb der Testmenge eine sinnvolle Ordnung herstellt. Hier besteht die weiterführende Aufgabe eine verallgemeinerungsfähige Evaluierung durchzuführen. Wenn diese weitere Evaluierung die Leistungsfähigkeit des Maßes bestätigt, dann kann das Kontinuitätsmaß das

Ähnlichkeitsmaß ablösen. Ebenfalls kann es dann allgemein für Aufgaben eingesetzt werden, die ein Maß für die thematische Vervollständigung eines Textes durch einen anderen Text benötigen. Die Betrachtung der Ergebnisse unserer Experimente mit dem Kontinuitätsmaß lassen optimistische Erwartungen zu.

Zum Umgang mit sehr kurzen Kommentaren haben wir ein Modell vorgestellt, das eine Zusammenfassung über Meinungsäußerungen ermöglicht. Dazu wurde eine Anwendung entwickelt, die den Nutzen eines solchen Modells zeigt. In der Anwendung wurde ebenfalls eine neue Darstellungsform, die Opinion Cloud verwendet, die eine sehr intuitive Erfassung der Gesamtheit aller Meinungsäußerungen ermöglicht. Um die Qualität bei der Identifizierung der Meinungsäußerungen weiter zu verbessern, muss eine Lösung im Umgang mit verschiedenen Schreibweisen gleicher Wörter gefunden werden. Auch kann das verwendete Wörterbuch weiter vergrößert werden. Dazu sollten weitere Verfahren zur automatischen Erweiterung getestet und weiterentwickelt werden. Ebenso ist eine Übertragung der Meinungsanalyse auf andere Sprachen denkbar.

Für die Realisierung der Relevanzberechnung auf Grundlage der Textqualität haben wir als Maß die Güte des Schreibstils untersucht. Dabei hat sich gezeigt, dass ein solches Maß sinnvoll für die Bestimmung der Relevanz von Kommentaren sein kann, allerding stellt die geringe Länge von Kommentaren ein Problem dar. Jedoch kann die Stilanalyse in Kombination mit anderen Maßen eine gute Ergänzung sein.

Unsere Arbeit stellt einen Einstieg in das Thema Kommentar-Retrieval dar. Weiterführend müssen die Modelle der thematischen Relevanz und der qualitativen Analyse in konkreten Anwendungen evaluiert werden.

Quellenverzeichnis

[1] DEERWESTER, S., S. DUMAIS, G. FURNAS, T. LANDAUER und R. HARSHMAN: *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[2] DING, X., B. LIU und P. YU: *A holistic lexicon-based approach to opinion mining*. In: *Proceedings of the international conference on Web search and web data mining*, S. 231–240. ACM New York, NY, USA, 2008.

[3] ESULI, A. und F. SEBASTIANI: *SentiWordNet: A publicly available lexical resource for opinion mining*. Proceedings of LREC, S. 417–422, 2006.

[4] GABRILOVICH, E. und S. MARKOVITCH: *Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis*. Proceedings of the 20th International Joint Conference on Artificial Intelligence, S. 6–12, 2007.

[5] GHOSE, A. und P. IPEIROTIS: *Designing ranking systems for consumer reviews: The impact of review subjectivity on product sales and review quality*. Proceedings of the 2007 9th International Conference on Decision Support Systems (ICDSS 2007), 2007.

[6] HATZIVASSILOGLOU, V. und K. MCKEOWN: *Predicting the semantic orientation of adjectives*. In: *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, S. 174–181. Association for Computational Linguistics Morristown, NJ, USA, 1997.

[7] HU, M. und B. LIU: *Mining and summarizing customer reviews*. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, S. 168–177. ACM New York, NY, USA, 2004.

[8] JINDAL, N. und B. LIU: *Analyzing and Detecting Review Spam*. In: *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, S. 547–552, 2007.

[9] JINDAL, N. und B. LIU: *Opinion spam and analysis*. In: *Proceedings of the international conference on Web search and web data mining*, S. 219–230. ACM New York, NY, USA, 2008.

[10] KAMPS, J., M. MARX, R. MOKKEN und M. DE RIJKE: *Using WordNet to measure semantic orientation of adjectives*. 2004.

[11] KIM, S. und E. HOVY: *Determining the sentiment of opinions*. In: *Proceedings of COLING*, Bd. 4, S. 1367–1373, 2004.

[12] KIM, S., P. PANTEL, T. CHKLOVSKI und M. PENNACCHIOTTI: *Automatically assessing review helpfulness*. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, S. 423–430, 2006.

[13] LIU, J., Y. CAO, C. LIN, Y. HUANG und M. ZHOU: *Low-Quality Product Review Detection in Opinion Summarization*. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, S. 334–342, 2007.

[14] MILLER, G., R. BECKWITH, C. FELLBAUM, D. GROSS und K. MILLER: *Introduction to WordNet: An On-line Lexical Database**. International Journal of Lexicography, 3(4):235–244, 1990.

[15] MISHNE, G.: *Multiple ranking strategies for opinion retrieval in blogs*. Online Proceedings of TREC, 2006.

[16] STEIN, B.: *Lectures in Web Technology (Advanced): Unit. Modelle und Prozesse im IR*. http://www.uni-weimar.de/medien/webis/teaching/lecturenotes/information-retrieval/unit-de-retrieval-models.pdf, 2008.
[online: 05.10.2008].

[17] STEIN, B. und M. POTTHAST: *Construction of Compact Retrieval Models*. In: DOMINICH, S. und F. KISS (Hrsg.): *Studies in Theory of Information Retrieval*, S. 85–93. Foundation for Information Society, Okt. 2007.

[18] STONE, P., J. KIRSH und C. C. ASSOCIATES: *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, 1966.

[19] TURNEY, P. et al.: *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, S. 417–424, 2002.

[20] TURNEY, P. und M. LITTMAN: *Measuring Praise and Criticism: Inference of Semantic Orientation from Association*. ACM TRANSACTIONS ON INFORMATION SYSTEMS, 21(4):315–346, 2003.

[21] VELOSO, A., W. MEIRA, T. MACAMBIRA, D. GUEDES und H. ALMEIDA: *Automatic Moderation of Comments in a Large On-line Journalistic Environment*. In: *Proc. of the Intl. Conf. on WebLogs and Social Media*, 2007.

[22] WILSON, T., J. WIEBE und P. HOFFMANN: *Recognizing contextual polarity in phrase-level sentiment analysis*. In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, S. 347–354. Association for Computational Linguistics Morristown, NJ, USA, 2005.

[23] WILSON, T., J. WIEBE und R. HWA: *Just How Mad Are You? Finding Strong and Weak Opinion Clauses*. In: *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, S. 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[24] ZHANG, Z. und B. VARADARAJAN: *Utility scoring of product reviews*. In: *Proceedings of the 15th ACM international conference on Information and knowledge management*, S. 51–57. ACM Press New York, NY, USA, 2006.

A Programmcode

A.1 Firefox-Erweiterung: Opinion-Cloud

Der Programmcode der Firefox-Erweiterung befindet sich auf der beiliegenden CD im Ordner *opinion_cloud*. Zur Installation kann die Datei *opinion_cloud.xpi* mit dem Browser Firefox[20] in Version 2 und 3 geöffnet werden.

A.2 YouTube-Crawler und Wörterbucherweiterung

Der Java-Programmcode des Crawlers für den YouTube-Korpus sowie für die Wörterbucherweiterung für das Meinungswörterbuch befinden sich auf der beiliegenden CD im Ordner *youTube* und kann als Eclipse-Projekt geöffnet werden.

A.3 Slashdot-Crawler und Modelle

Der Java-Programmcode des Crawlers für den Slashdot-Korpus und der darauf getesteten Modelle befinden sich auf der beiliegenden CD im Ordner *slashdot* und kann als Eclipse-Projekt geöffnet werden.

[20] http://www.mozilla-europe.org/de/firefox/