

# Automatische Suchbegriffsempfehlungen für die Kommunikationsanalyse

Jenny Felser, Michael Spranger

Kontakt: Jenny Felser, Hochschule Mittweida, felser@hs-mittweida.de

## Zusammenfassung

Die Kommunikation über mobile Endgeräte erfreut sich heutzutage großer Beliebtheit. Die zahlreichen Kurznachrichten, die auf mobilen Endgeräten gespeichert sind, dienen hierbei auch als wichtige Beweisquelle in strafrechtlichen Ermittlungen. Die hohe Anzahl an Nachrichten stellt jedoch die Ermittler:innen vor eine große Herausforderung. Eine Möglichkeit, die Ermittler:innen bei der Auswertung der enormen Menge an Nachrichten zu unterstützen, besteht darin, ihnen die relevantesten Suchbegriffe und -phrasen automatisiert vorzuschlagen. Das Ziel dieser Arbeit besteht deshalb darin, Methoden zur automatischen Ermittlung von Suchbegriffen aus dem Datensatz auf ihre Tauglichkeit für forensische Kurztexte zu überprüfen. Hierfür wurden insgesamt neun Methoden miteinander verglichen und auf realen Falldaten evaluiert. Als erfolgsversprechend erwies sich insbesondere die Themenmodellierung unter Berücksichtigung von globalen Wortkookkurrenzen sowie die Analyse von syntagmatischen Relationen unter Einbeziehung eines Referenzkorpus.

**Keywords:** Suchbegriffe, umgangssprachliche Texte, forensisches Text Mining, Text Retrieval.

## 1 Einführung

Im heutigen digitalen Zeitalter wird die Kommunikation über Short Message Service (SMS) und Messenger-Dienste wie WhatsApp immer beliebter. Demzufolge gewinnt die Analyse der Kurznachrichten, die auf mobilen Endgeräten gespeichert sind, ebenfalls an Bedeutung für die digitale Forensik (Evans & Gosalia, 2015). Die immense Menge an Nachrichten erweist sich jedoch als große Herausforderung für die Ermittler:innen. Oftmals befinden sich auf einem einzigen Mobilfunktelefon mehr als 15.000 SMS und 150.000 Nachrichten von Messenger-Diensten (Spranger, Heinke, Appelt, Puder, & Labudde, 2016; Spranger, Zuchantke, & Labudde, 2014). Hinzu kommt, dass es bei Bandenkriminalität und organisierter Kriminalität oft notwendig ist, die Kurznachrichten von mehreren Mobilfunktelefonen zu untersuchen (Spranger et al., 2016). Text-Retrieval-Systeme können zwar dabei unterstützen, beweisrelevante Informationen in den Chatnachrichten zu finden, jedoch hängt der Erfolg der Suche stark davon ab, ob geeignete Suchbegriffe eingegeben wurden. Den Ermittler:innen ist hierbei häufig nicht bekannt, welche Begriffe für ihren konkreten Fall relevant sind (abhängig von konkreten Tatumständen), welche von den Gesprächsteilnehmer:innen verwendet wurden (abhängig vom verwendeten Vokabular) und deshalb im Datensatz vorkommen. Deshalb bestand das Ziel dieser Arbeit darin, eine Methode zu entwickeln, die geeignete Suchbegriffe und -phrasen automatisiert aus dem Datensatz vorschlagen kann. Zu diesem Zweck wurden verschiedene Methoden mitei-

einander verglichen. Dies umfasste die Untersuchung von Ansätzen, die Wörter empfehlen, ohne dabei Vorwissen über den Datensatz vorauszusetzen sowie von Methoden, die Suchbegriffe basierend auf wenigen relevanten Wörtern vorschlagen.

Eine besondere Herausforderung ist hierbei, dass die relevantesten Wörter für einen Fall häufig spezielle Begriffe sind, bei denen man nicht erwarten würde, dass diese in den Chatnachrichten auftreten. Handelt beispielsweise ein Fall von der finanziellen Unterstützung einer terroristischen Vereinigung, würden die Ermittler:innen damit rechnen, dass Wörter wie „Geld“, „Überweisung“ oder „Konto“ in dem Datensatz enthalten sind. Wesentlich interessanter wäre es jedoch, wenn die Chatnachrichten beispielsweise Namen von Terrororganisationen oder Marken von Waffen enthielten, die für Terroranschläge gekauft wurden, und diese Wörter vorgeschlagen werden könnten. Der Artikel ist folgendermaßen aufgebaut: In Abschnitt 2 werden bisherige Arbeiten im Bereich von Suchbegriffsempfehlungen vorgestellt. Anschließend werden in Abschnitt 3 die untersuchten Methoden beschrieben. Die Ergebnisse werden in Abschnitt 4 präsentiert und diskutiert. Die Arbeit endet mit einer Zusammenfassung der wichtigsten Resultate sowie einem Ausblick auf die zukünftige Forschung in Abschnitt 5.

## 2 Verwandte Arbeiten

Bisher befassten sich nur wenige Arbeiten mit der Empfehlung von Suchbegriffen in forensischen Texten (Joseph & Viswanathan, 2023; Koven, Bertini, Dubois, & Memon, 2016), die sich zudem nicht auf die Analyse von umgangssprachlichen Kurznachrichten konzentrierten. Beispielsweise wurde von Koven et al. (2016) vorgeschlagen, mittels einer Variante des TF-IDF-Algorithmus (K. S. Jones, 1972) Suchbegriffe zum Finden fallrelevanter Inhalte in großen E-Mail-Datensätzen zu empfehlen. Als problematisch an dieser Methode ist jedoch zu sehen, dass, wie von Teng und Lv (2021) betont wird, mittels TF-IDF vor allem hochfrequente Wörter extrahiert werden, weshalb der Algorithmus nicht zum Vorschlagen von spezifischen und ungewöhnlichen Wörtern geeignet ist. Interessanter ist hingegen der von Joseph und Viswanathan (2023) vorgestellte Ansatz, der Suchbegriffe empfiehlt, um beweisrelevante Informationen in Dateien, die aus Festplattenspeichern extrahiert wurden, ausfindig zu machen. Die grundlegende Idee bestand darin, Topic Modeling mit einer adaptierten Latent Dirichlet Allocation (LDA) durchzuführen und aus den wahrscheinlichsten Wörtern der extrahierten Themen relevante Suchbegriffe manuell auszuwählen.

Dieser Ansatz kam ebenfalls bei dem von Keerthana (2017) entwickelten Recommender zum Einsatz, der mithilfe der Standard LDA Suchbegriffe aus Konversationen vorschlug, die mittels eines Spracherkennungssystems auf-

gezeichnet wurden. Da Topic Modeling in der Lage ist, latente semantische Strukturen zu extrahieren, bietet es sich zum Empfehlen von unerwarteten und speziellen Begriffen an (Chen, Li, Guo, & Guo, 2015), weshalb es ebenfalls in dieser Arbeit herangezogen wurde. Da jedoch vorherige Studien (z.B. Quan et al. 2015; Shi et al, 2018) zeigten, dass insbesondere die von Keerthana (2017) gewählte Standard LDA mangelhafte Ergebnisse auf kurzen Texten hervorbringt, wurde stattdessen, wie von Zuo (2015) vorgeschlagen, eine Variante der LDA unter Einbeziehung von globalen Kookkurrenzmustern verwendet. Darüber hinaus wurden zahlreiche Ansätze zum Vorschlagen von Queries im Kontext der Websuche entwickelt, die jedoch überwiegend auf Query Logs basieren und dementsprechend nicht geeignet sind, um relevante Wörter basierend auf einem bestimmten zugrundeliegenden Datensatz zu empfehlen (z.B. Baeza-Yates, Hurtado, & Mendoza, 2004; He et al., 2009; Mei, Zhou, & Church, 2008). Als relevantere Arbeit in diesem Bereich wäre der von Kubek und Unger (2014) entwickelte Doc Analyzer anzuführen. Dieser empfiehlt Benutzer:innen aus zuvor besuchten Webpages Suchbegriffe zum Finden von ähnlichen Webdokumenten, wozu ein modifizierter HITS-Algorithmus (Kleinberg, 1999) eingesetzt wurde. Jedoch eignet sich dieser Ansatz nach Kubek und Unger (2014) sowie nach Wang et al. (2017) insbesondere für einzelne, homogene Dokumente, jedoch weniger für immense Datensätze, die sich mit verschiedenen Themen befassen, da die meisten extrahierten Wörter in diesem Fall nicht in Verbindung mit geringer repräsentierten Themen stehen. Dies ist für die Kommunikationsdaten besonders problematisch, da in diesen Smalltalk im Gegensatz zu fallrelevanten Themen die Konversationen dominiert. Um dieses Problem zu adressieren, wurde von Wang et al. (2017) der graphenbasierte Algorithmus „Rapid automatic keyword extraction“ (RAKE) anstelle von HITS eingesetzt, um die wichtigsten Bigramme aus Online-Zeitungsartikeln vorzuschlagen. Obwohl die Autoren mit diesem Ansatz aussagekräftige Phrasen extrahieren konnten, ist darauf hinzuweisen, dass die Chats im Gegensatz zu Zeitungsartikeln zahlreiche umgangssprachliche Ausdrücke, Rechtschreibfehler und Akronyme enthalten (Spranger et al., 2014). Aus diesem Grund sind Ansätze, die gute Resultate auf sprachlich korrekten Texten lieferten, nicht zwangsläufig auf mobile Kommunikationsdaten übertragbar. Die besondere Herausforderung, die Suche in umgangssprachlichen Texten zu erleichtern, wurde beispielsweise von Feldman et al. (2023), Fu et al. (2015) sowie von Yamada et al. (2021) behandelt. Jedoch ist zu beachten, dass das Ziel der Arbeit von Fu et al. (2015) im Gegensatz zu dieser Arbeit nicht in der Suchbegriffsempfehlung, sondern in der Vervollständigung von bereits begonnenen Queries bestand. Der von Feldman et al. (2023) beschriebene Ansatz beruhte darauf, das autoregressive Sprachmodell Generative Pre-Trained Transformer 3 (GPT-3) (Brown et al., 2020) damit zu beauftragen, eine Liste von Suchbegriffen zu generieren, um mit diesen in Tweets Informationen zu einer bestimmten Bevölkerungssubgruppe zu finden. Allerdings ist, insbesondere wenn GPT-3 wie bei Feldman et al. (2023) ohne weiteres Fine-Tuning eingesetzt wird, nicht garantiert, dass sich in der Liste von

Suchbegriffen tatsächlich die spezifischen Wörter aus dem Kommunikationsdatensatz befinden, da GPT-3 primär auf Wikipedia-Artikeln, Websites und Büchern trainiert wurde (Brown et al. 2020). Die von Yamada et al. (2021) eingesetzte Methode kommt ebenfalls nicht für die Extraktion von relevanten Begriffen aus Nachrichten von Messenger Diensten wie WhatsApp infrage, da diese Begriffe basierend auf Twitter-spezifischen Metadaten wie der Anzahl an Likes empfiehlt.

Ansätze, die wie diese Arbeit den Fokus auf die Empfehlung von relevanten, aber nicht offensichtlichen Begriffen legten, wurden bisher beispielsweise durch Joshi und Motwani (2006) sowie durch Abhishek und Hosanagar (2007) für das Keyword Advertising im Bereich des Suchmaschinenmarketings entwickelt. Während Joshi und Motwani (2006) statistisch syntagmatische Relationen zu von Benutzer:innen vorgegebenen Begriffen untersuchten, schlugen Abhishek und Hosanagar (2007) Begriffe vor, die in paradigmatischer Relation zu relevanten Wörtern standen. Der zweitgenannte Ansatz wurde ebenfalls von Kilgarriff et al. (2015) zur Empfehlung von ähnlichen Begriffen für eine Plattform zum Erlernen der englischen Sprache sowie von Shimizu et al. (2020) für die Zuweisung von relevanten Begriffen zu Dokumenten über geologische Themen aus einem vordefinierten Vokabular angewendet. Da paradigmatische Relationen auf der Untersuchung von Kontextähnlichkeit beruhen (Otero, 2009), können diese ebenfalls erfolgsversprechend für forensische Texte sein, die Hidden Semantics enthalten und in denen Wörter in einer unüblichen Bedeutung und dementsprechend in einem ungewöhnlichen Kontext verwendet werden (Spranger et al., 2014). Aus diesem Grund wurden sie ebenfalls für die Suchbegriffsempfehlungen im Kontext von forensischen Kurznachrichten untersucht.

Während jedoch bisherige Arbeiten nur einen einzigen Ansatz für die Begriffsempfehlungen in Betracht zogen, wurde in dieser Arbeit eine umfassende Untersuchung mehrerer Methoden des Text Minings durchgeführt. Ein wesentlicher Beitrag dieser Arbeit besteht zudem darin, dass die Methoden sowohl in der Lage sein sollten, spezielle Begriffe zu empfehlen als auch mit umgangssprachlichen Texten umzugehen.

### 3 Untersuchte Methoden

Um die geeignetste Methode zur Suchbegriffsempfehlung zu finden, wurden insgesamt neun verschiedene Ansätze miteinander verglichen.

Die im Folgenden untersuchten Methoden können in zwei Kategorien unterteilt werden: Zum einen wurden Methoden untersucht, die darauf abzielten, Begriffe vorzuschlagen, wenn keine relevanten Wörter aus dem Datensatz bekannt sind. Zum anderen wurden mehrere Ansätze verglichen, die für Empfehlungen infrage kamen, vorausgesetzt, dass die Ermittler:innen bereits Kenntnis von einigen relevanten Begriffen hatten. Dieses Vorwissen könnten sie beispielsweise aus Vernehmungen oder der Fallakte entnehmen. Eine Übersicht über die grundlegende Idee zur Auswahl der Suchbegriffe und das benötigte Vorwissen bei den eingesetzten Methoden kann Tabelle 1 entnommen werden.

Tabelle 1: Übersicht über die empfohlenen Suchbegriffe und das erforderliche Vorwissen bei den angewendeten Methoden.

Methode	Grundlegende Idee zur Auswahl von Suchbegriffen	Vorwissen/ Input der Nutzer:innen
Differenzanalyse Chi-Quadrat-Test	signifikant hohes Auftreten im Vergleich zum Referenzkorpus	unüberwacht
Document Clustering	hohe Termfrequenz in Nachrichten eines fallrelevanten Clusters	Auswahl eines Clusters
Topic Modeling (LDA) Topic Modeling (WNTM)	hohe Wahrscheinlichkeit in fallrelevantem Thema	Auswahl eines Themas
Thematische Ähnlichkeit nach Rus et al. (2013)	hohe Wahrscheinlichkeit in gleichen Themen wie Referenzbegriff	Referenzbegriff
Thematische Ähnlichkeit nach Griffiths et al. (2007)	hohe thematische Ähnlichkeit zu mehreren Referenzbegriffen	Referenzbegriff(e)
Paradigmatische Relationen	hohe Kontextähnlichkeit zu Referenzbegriff	Referenzbegriff
Statistisch syntagmatische Relationen	unerwartet hohes gemeinsames Vorkommen mit Referenzbegriff	Referenzbegriff

Alle Methoden wurden auf realen Falldaten zu dem Delikt der finanziellen Unterstützung einer terroristischen Vereinigung evaluiert. Der verwendete Datensatz besteht aus ungefähr 180.000 überwiegend deutschsprachigen und wenigen türkischsprachigen Nachrichten des Messenger-Dienstes WhatsApp, die auf dem Mobilfunktelefon einer tatverdächtigen Person gespeichert waren. Zur Evaluierung der einzelnen Ansätze wurden die jeweils vorgeschlagenen Begriffe mit von der ermittelnden Person ausgewählten Wörtern verglichen. Diese waren bei der Auswertung der Falldaten mit dem Mobile Network Analyzer (MoNA) (Spranger, Xi, Jaeckel, Felser, & Labudde, 2022), einer forensischen Software zur Analyse von Chatnachrichten, von der ermittelnden Person als Suchbegriffe herangezogen worden. Die folgenden 25 Begriffe wurden hierbei als relevant erachtet: *Geld, Schwester, Nalan, Hawala, Niederlande, Buch, Bücher, scan, Seiten, Kapitel, Wörter, Bieber, Layth, Verleger, money, Druck, angekommen, Dolar, €, Euro, \$, Western, Union, PayPal* und *Dollar*. Je mehr dieser oder ähnlicher Begriffe wie zum Beispiel Synonyme oder semantisch ähnliche Wörter eine Methode vorschlug, als desto geeigneter wurde sie angesehen. Die Vorgehensweise von der Vorverarbeitung des Datensatzes bis zu der Evaluierung wird in Abbildung 1 veranschaulicht.

### 3.1 Vorverarbeitung

In einem ersten Schritt wurden eine Reihe von Vorverarbeitungsschritten durchgeführt. Dies umfasste die Umwandlung der Nachrichten in Kleinschreibung und die Tokenisierung der Nachrichten in einzelne Wörter. Darüber hinaus wurden Emoticons, Satzzeichen, Zahlen und überflüssiger Whitespace entfernt. Ebenfalls wurden Wörter beziehungsweise Tokens entfernt, die nicht als Suchbegriffsempfehlungen infrage kamen. Dies beinhaltete die Entfernung von deutschen und türkischen Stoppwörtern

unter Verwendung der Stoppwortlisten von Diaz (2023) sowie die Entfernung von Weblinks. Zudem wurden mithilfe eines POS-Taggings die berücksichtigten Wortarten auf Substantive, Namen, Verben und Adjektive begrenzt, da diese im Vergleich zu anderen Wortarten eine stärkere semantische Bedeutung aufweisen (de Arruda, Costa, & Amancio, 2016; Heyer, Quasthoff, & Wittig, 2016a). Für das POS-Tagging wurde der TreeTagger (Schmid, 1994, 1995) eingesetzt. Hinsichtlich des Clusterings und Topic Modelings wurden zudem Auxiliarverben und Modalverben entfernt. Bei Ansätzen, die einen Vergleich mit einem Referenzkorpus beinhalteten, d.h. bei Verfahren der Keyword Extraction und der Analyse von syntagmatischen Relationen, wurde sowohl auf den Begriffen in den Kommunikationsdaten als auch im Referenzkorpus Stemming durchgeführt, um zu gewährleisten, dass verschiedene morphologische Varianten desselben Wortes als identisch betrachtet wurden (Singh & Gupta, 2016). Für die Empfehlungen wurden die gestemmen Begriffe jedoch wieder vervollständigt, um sprachlich korrekte Wörter vorschlagen zu können.

### 3.2 Methoden ohne Vorwissen

Für das Szenario, dass keine relevanten Begriffe bekannt sind, wurden zwei Verfahren der Keyword Extraction, Dokument Clustering und zwei Methoden des Topic Modelings verglichen.

#### 3.2.1 Methoden der Keyword Extraction

Hinsichtlich der Keyword Extraction wurde die von Heyer et al. (2016b) vorgeschlagene Differenzanalyse sowie der Chi-Quadrat-Test angewendet, dessen Einsatz zur Termextraktion beispielsweise von McEnery und Wilson (2001) beschrieben wurde. Beide Verfahren identifizieren Wörter, die signifikant häufiger in den Nachrichten als in einem

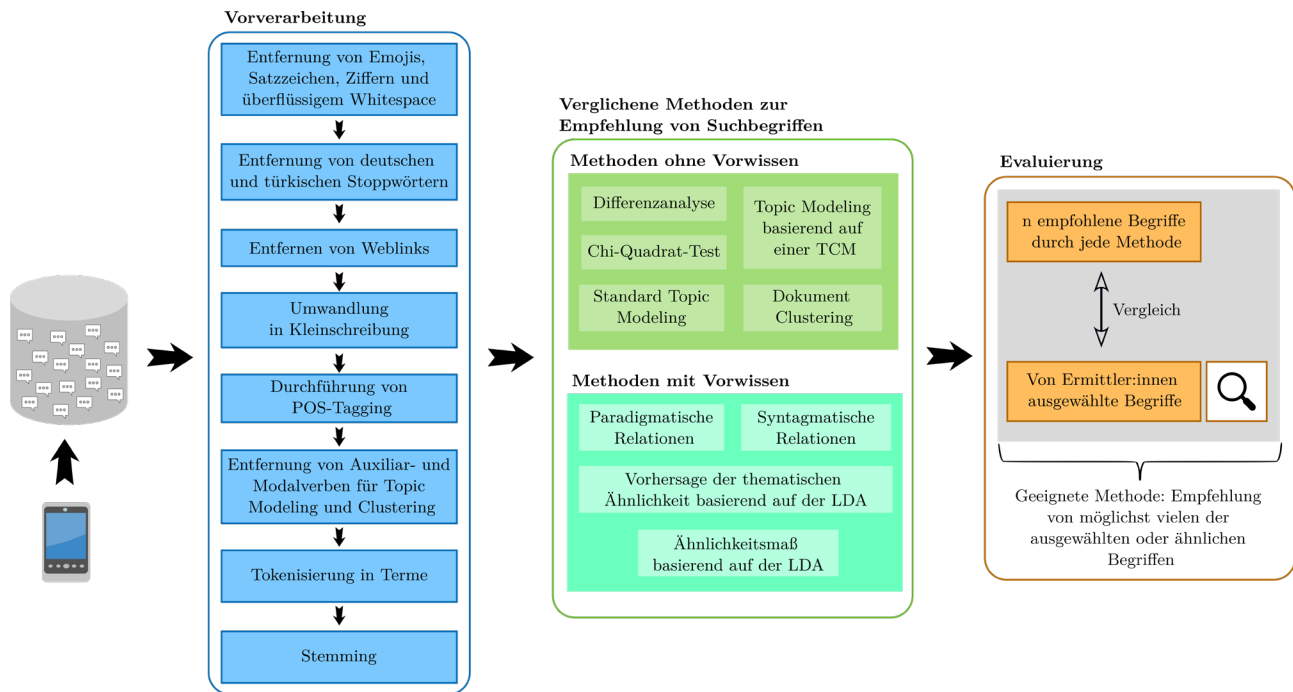


Abbildung 1: Verwendete Vorgehensweise zum Vergleich von Methoden für die Empfehlung von Suchbegriffen.

Referenzkorpus auftreten (Culpeper, 2009). Da die extrahierten Begriffe als charakteristisch für die Nachrichten angesehen werden (Mastropierro & Mahlberg, 2017), kommen sie ebenfalls als empfohlene Suchbegriffe infrage. Als Referenzkorpus diente der Datensatz des Projekts „Deutscher Wortschatz“ (Quasthoff, Richter, & Biemann, 2006) der Universität Leipzig, der primär aus Online-Zeitungsartikeln bestand. Die extrahierten Begriffe wurden bei der Differenzanalyse basierend auf der Differenz ihrer Frequenzen zum Referenzkorpus und bezüglich des Chi-Quadrat-Testes absteigend nach ihrem Chi-Quadrat-Wert sortiert.

### 3.2.2 Document Clustering

Ein weiterer Ansatz bestand darin, ein Document Clustering durchzuführen, wobei als Algorithmus eine agglomerative hierarchische Clusteranalyse gewählt wurde, die beispielsweise von Jaiswal und Janwe (2011) erklärt wird. Da jedoch die hohe Sparsität von Kurznachrichten für das hierarchische Clustering eine Herausforderung darstellt (Tsur, Littman, & Rappoport, 2013), wurden mehrere Nachrichten zu Pseudodokumenten zusammengefasst. Ein Pseudodokument beinhaltete alle Nachrichten, die an einem Tag in einem Chat geschrieben wurden. Die grundlegende Idee bestand darin, ein fallrelevantes Cluster auszuwählen und den Datensatz für die Empfehlung von Suchbegriffen auf die Nachrichten einzuschränken, die in diesem Cluster enthalten waren. Auf diese Weise sollte das Rauschen in den Daten reduziert werden, indem irrelevante Nachrichten ausgeschlossen wurden. Anschließend wurden, ähnlich wie bei dem von Jones (2021) beschriebenen Ansatz, die Wörter mit der höchsten Termfrequenz in dem eingeschränkten Datensatz vorgeschlagen.

### 3.2.3 Topic Modeling

Darüber hinaus kam Topic Modeling zum Einsatz, wobei als Algorithmus die LDA gewählt wurde. Diese wurde in einem ersten Schritt mit den zuvor gebildeten Pseudodokumenten als Eingabe durchgeführt. Darüber hinaus wurde die LDA unter Berücksichtigung von globalen Wortkookkurrenzen wiederholt, da dieser Ansatz sich nach Zuo et al. (2015) besonders für kurze Texte eignet. Die hierfür verwendete Vorgehensweise entsprach im Wesentlichen dem von Zuo et al. (2015) vorgeschlagenen Word Network Topic Model (WNTM). Als empfohlene Suchbegriffe dienten für beide Ansätze des Topic Modells die Wörter mit der höchsten Wahrscheinlichkeit in einem ausgewählten fallrelevanten Thema.

### 3.3 Methoden mit Vorwissen

Des Weiteren wurden mehrere Methoden untersucht, die darauf abzielten, Wörter zu empfehlen, die zu einem relevanten Begriff, der im Folgenden als Referenzbegriff bezeichnet wird, semantisch ähnlich sind beziehungsweise mit diesem in Verbindung stehen. Zudem wurde eine Methode mit einbezogen, die mehrere Referenzbegriffe als Eingabe entgegennahm.

#### 3.3.1 Einsatz der LDA als Ähnlichkeitsmaß

Eine Idee bestand darin, die LDA unter Einbeziehung von globalen Wortkookkurrenzen (Zuo et al., 2015) einzusetzen, um zu Referenzbegriffen thematisch ähnliche Wörter zu empfehlen. Hierbei wurden Wörter vorgeschlagen, die in den gleichen Themen wie der Referenzbegriff eine hohe Wahrscheinlichkeit aufwiesen (Griffiths, Steyvers, & Tenenbaum, 2007; Rus et al., 2013). Konkret wurde der von Rus et al. (2013) beschriebene Ansatz für Vorschläge basierend auf einem einzelnen Referenzbegriff angewendet

sowie die von Griffiths et al. (2007) entwickelte Vorgehensweise, die es erlaubte, die thematische Ähnlichkeit eines Begriffs zu mehreren relevanten, bekannten Wörtern vorherzusagen.

### 3.3.2 Paradigmatische Relationen

Zudem wurden paradigmatische Relationen zu dem Referenzbegriff analysiert, wozu der von Zhai und Massung (2016) beschriebene Ansatz „Expected Overlap of Words in Context“ (EOWC) verwendet wurde. Dieser beruhte darauf, die Skalarprodukt-Ähnlichkeit der Drei-Wort-Kontexte von Begriffen zu berechnen. Konkret wurde die von Zhai und Massung (2016) vorgeschlagene verbesserte Variante dieses Algorithmus unter Verwendung des BM25-Algorithmus (Robertson & Zaragoza, 2009) angewendet. Als empfohlene Suchbegriffe dienten die Wörter mit den ähnlichsten Kontexten zum Referenzbegriff.

### 3.3.3 Statistisch syntagmatische Relationen

Schließlich wurden, wie von Joshi und Motwani (2006) vorgeschlagen wurde, Wörter empfohlen, die mit einem Referenzbegriff in statistisch syntagmatischer Relation standen und somit mit diesem statistisch signifikant häufig gemeinsam in einer Nachricht auftraten (Rohrdantz et al., 2010). Da jedoch vor allem spezielle Begriffe empfohlen werden sollten, lag der Fokus darauf, Wörter zu extrahieren, bei denen man nicht damit rechnen würde, dass sie häufig mit einem Referenzbegriff in einer Nachricht vorkommen. Um solche Begriffe zu finden, wurde die Kookkurrenzhäufigkeit von jedem Wort des Datensatzes und einem Referenzbegriff in den Kommunikationsdaten mit der entsprechenden Kookkurrenzhäufigkeit in einem Referenzkorpus verglichen, wie dies beispielsweise von Pastor (2017) durchgeführt wurde. Genauer gesagt wurden die Log-Likelihood-Werte (Dunning, 1993) im Datensatz von Kurznachrichten mit Log-Likelihood-Werten von vorberechneten Wortkookkurrenzen des Projekts „Deutscher Wortschatz“ (Quasthoff et al., 2006) verglichen. Anschließend wurden die Wörter mit der höchsten Differenz bezüglich der Log-Likelihood Werte empfohlen.

## 4 Ergebnisse

Die folgenden Unterabschnitte präsentieren und diskutieren die Ergebnisse, die die verglichenen Algorithmen für Empfehlungen ohne und mit Einbeziehung von Referenzwörtern hervorbrachten.

### 4.1 Empfehlungen ohne Vorwissen

Hinsichtlich der Empfehlungen ohne Berücksichtigung von Vorwissen erwiesen sich insbesondere die Ansätze basierend auf Topic Modeling als vielversprechend.

#### 4.1.1 Methoden der Keyword Extraction

Beide Methoden der Keyword Extraction, sowohl die Differenzanalyse als auch der Chi-Quadrat-Test, führten nicht zu der Empfehlung von relevanten Begriffen. Die ersten sieben Wörter, die mit den beiden Ansätzen empfohlen wurden, sind in Tabelle 2 mit ihrem Rang in der Liste von Suchbegriffsempfehlungen angegeben. Wie die-

ser entnommen werden kann, befanden sich unter den vorgeschlagenen Wörtern keiner der 25 als relevant betrachteten Begriffe oder ähnliche Terme. Dies konnte darauf zurückgeführt werden, dass Verfahren der Keyword Extraction nicht nur die inhaltlichen Unterschiede zwischen den Nachrichten und dem Referenzkorpus hervorheben, sondern auch die stilistischen Unterschiede (Quasthoff, 2009). Da der Referenzkorpus im Gegensatz zu den Nachrichten in Schriftsprache verfasst war, handelte es sich bei Wörtern, die ausschließlich in den Nachrichten vorkamen und somit vorgeschlagen wurden, insbesondere um umgangssprachliche Begriffe, Anglizismen und türkische Füllwörter, wie Tabelle 2 entnommen werden kann.

### 4.1.2 Dokument Clustering

Tabelle 2: Empfohlene Suchbegriffe durch die Differenzanalyse und den Chi-Quadrat-Test.

Rang	Differenzanalyse	Chi-Quadrat-Test
1	Allah	hahaha
2	Alhamdulillah	vllt
3	Shaa	gehts
4	Inshallah	Bescheid
5	Subhanallah	khalid
6	haha	hayirli
7	hahaha	Song

Hinsichtlich des Dokument Clusterings konnte kein fallrelevantes Cluster identifiziert werden, da keiner der häufigsten Begriffe in den entsprechenden Pseudodokumenten zu den von dem Ermittler als relevant betrachteten Wörtern zählte. Hingegen handelte es sich bei den häufigsten Wörtern in allen Clustern überwiegend um umgangssprachliche Begriffe und Slang-Terme wie beispielsweise „diggi“, „moin“, „gucks“ und „ersma“. Eine mögliche Ursache ist darin zu sehen, dass hierarchische Clusteralgorithmen vor allem bei sprachlich korrekten Dokumenten gute Ergebnisse erzielen (Gaikwad & Patwardhan, 2014), wovon die Chatnachrichten stark abweichen.

### 4.1.3 Topic Modeling

Fallrelevante Begriffe konnten hingegen mit beiden Ansätzen des Topic Modelings identifiziert werden. Sowohl bei dem gewöhnlichen Topic Modeling als auch bei dem Topic Modeling basierend auf globalen Wortkookkurrenzen konnte eindeutig das Thema erkannt werden, das von Bedeutung für den Fall war. Aus diesem Grund wurde er als geeignetste Methode zur Empfehlung von Suchbegriffen ohne Berücksichtigung von Referenzwörtern betrachtet. Tabelle 3 sind die relevantesten Wörter der beiden Themen dargestellt, wobei die vorgeschlagenen Wörter absteigend nach ihrer Wahrscheinlichkeit in den Themen sortiert wurden. Die Tabellen zeigen aus Lesbarkeitsgründen nur die repräsentativsten Begriffe aus den Top 100 jedes Ansatzes. In Klammern ist dabei die exakte Position der Wörter innerhalb der Liste von Suchbegriffen dargestellt.

Wie in Tabelle 3 zu sehen ist, wurden mithilfe des Standard Topic Modelings zwei Begriffe von den zuvor genannten relevanten Termen vorgeschlagen, nämlich „Druck“ und „PayPal“, sowie weitere thematisch ähnliche Begriffe wie „Rechnung“ und „Preise“. Allerdings fehlten bei diesem Ansatz speziellere Wörter.

Hingegen befanden sich bei dem Topic Modeling unter Berücksichtigung von globalen Wortkookkurrenzen unter den wahrscheinlichsten Wörtern des fallrelevanten Themas ebenfalls Begriffe wie die Crowdfunding Plattform „Tipeee“, der Verein „Streaming with Heart (SWH)“ und das Live-Streaming-Portal „Twitch“. Diese Wörter waren zwar nicht von der ermittelnden Person als relevante Suchbegriffe ausgewählt worden, jedoch war unter Betrachtung ihres Kontextes in den Chatnachrichten ersichtlich, dass sie tatsächlich für den Fall relevant waren. Dementsprechend ermöglicht es dieser Ansatz, Suchbegriffe vorzuschlagen, die Ermittler:innen ansonsten nicht in Erwägung gezogen hätten. Aus diesem Grund wurde er als geeignetste Methode zur Empfehlung von Suchbegriffen ohne Berücksichtigung von Referenzwörtern betrachtet.

Tabelle 3: Relevante Suchbegriffsempfehlungen durch Topic Modeling.

Standard Topic Modeling	Topic Modeling mit globalen Wortkookkurrenzen
spenden (7)	Twitter (1)
Rechnung (11)	Team (2)
Druck (23)	Event (4)
fees (43)	Statement (5)
PayPal (54)	Tipeee (8)
Preise (58)	Twitch (9)
Belege (64)	Discord (15)
Gebühren (71)	SWH (18)
Sponsoren (75)	fees (19)

#### 4.2 Empfehlungen mit Vorwissen

Hinsichtlich der Empfehlungen unter Berücksichtigung von Vorwissen muss darauf hingewiesen werden, dass die Qualität der Vorschläge je nach verwendetem Referenzbegriff variierte. Hierbei waren die Wörter, die basierend auf einem speziellen Referenzwort wie „PayPal“ empfohlen

wurden, grundsätzlich relevanter und vor allem unerwarteter als Empfehlungen zu einem allgemeineren Referenzbegriff wie „Geld“. Zur besseren Vergleichbarkeit wurde für alle nachfolgenden Beispiele „PayPal“ als Referenzwort verwendet.

##### 4.2.1 Thematische Ähnlichkeit

Die Ermittlung von thematisch ähnlichen Wörtern zu einem einzelnen Referenzbegriff mithilfe des von Rus et al. (2013) beschriebenen Ansatzes führte grundsätzlich zu relevanten, empfohlenen Suchbegriffen, wovon einige exemplarisch in der ersten Spalte von Tabelle 4 dargestellt sind. Wie dieser entnommen werden kann, waren die vorgeschlagenen Wörter jedoch überwiegend identisch zu den in Tabelle 3 aufgeführten Termen, die mit dem Topic Modeling basierend auf globalen Wortkookkurrenzen empfohlen wurden. Die Einbeziehung des Referenzbegriffes konnte allerdings die Rangfolge verbessern, sodass relevante Wörter wie „fees“ und „SWH“ weiter oben in der Liste von Suchbegriffsempfehlungen erschienen als bei dem vollkommen unüberwachten Topic Modeling. Wurden mithilfe der von Griffiths et al. (2007) vorgeschlagenen Vorgehensweise Wörter zu mehreren Referenzbegriffen, konkret zu „Geld“, „Euro“, „PayPal“ und „Dollar“ vorgeschlagen, entsprachen diese, abgesehen von geringen Abweichungen bezüglich der Rangfolge, den thematisch ähnlichen Wörtern zu dem Referenzbegriff „PayPal“.

##### 4.2.2 Paradigmatische Relationen

Sinnvolle Empfehlungen wurden ebenfalls mithilfe von paradigmatischen Relationen erreicht. Jedoch war es aufgrund der hohen Laufzeit zur Berechnung der paradigmatischen Relationen notwendig, den Datensatz auf einige Chats einzugrenzen. Dies konnte durch eine Kombination mit den Ergebnissen des Topic Modelings realisiert werden, indem nur Chats berücksichtigt wurden, in denen das fallrelevante Thema mit einer hohen Wahrscheinlichkeit vorkam. Einige der als relevant angesehenen Wörter, die auf diese Weise zu dem Referenzbegriff „PayPal“ empfohlen wurden, sind in der zweiten Spalte von Tabelle 4 aufgeführt. Unter diesen befanden sich ebenfalls speziellere Wörter wie „Spendenbetrag“ und „Krebshilfe“, die ebenfalls in Zusammenhang mit Spendenaktionen standen, wie durch Betrachtung des Kontextes des Begriffs ersichtlich war. Allerdings bedarf es weiteren Untersuchungen, um beurteilen zu können, ob die paradigmatischen Relationen

Tabelle 4: Empfohlene Suchbegriffe mithilfe von syntagmatischen und paradigmatischen Relationen und thematischer Ähnlichkeit zu dem Referenzbegriff „PayPal“.

Thematische Ähnlichkeit	Paradigmatische Relationen	Syntagmatische Relationen
Twitter (1)	Berechnungen (2)	€ (5)
Team (2)	Spendemöglichkeiten (5)	Tipeestream (6)
Tipeee (5)	Konto (6)	Anzahlung (9)
Twitch (7)	Spendenbetrag (7)	überweisen (12)
fees (11)	Krebshilfe (9)	Konto (14)
SWH (13)	Quittungen (23)	Kohle (16)

ebenfalls auf dem gesamten Datensatz zu sinnvollen Empfehlungen führen oder ob die guten Resultate vor allem auf die Einschränkung des Datensatzes auf das fallrelevante Thema zurückzuführen sind.

#### 4.2.3 Statistisch syntagmatische Relationen

Als erfolgsversprechend erwiesen sich die Empfehlungen mithilfe von syntagmatischen Relationen und dem Abgleich mit einem Referenzkorpus. Die relevanten Begriffe unter den ersten 20 vorgeschlagenen Wörtern zu dem Referenzbegriff „PayPal“ können der letzten Spalte von Tabelle 4 entnommen werden. Neben den aufgeführten Begriffen befanden sich auf niedrigeren Positionen in der Liste von Suchbegriffsempfehlungen weitere relevante und spezielle Wörter wie „Transaktion“ auf Rang 24, „Spendendose“ auf Rang 35 und insbesondere „Terror“ auf Rang 46. Dass diese Begriffe auf unteren Rängen der Suchbegriffsliste erschienen, lag daran, dass sich unter den ersten empfohlenen Wörtern neben den in Tabelle 4 aufgeführten Termen vor allem auch Begriffe mit Rechtschreibfehlern wie „vlg“ anstelle von „vgl“ und umgangssprachliche Wörter wie „durchblättern“ befanden. Dies konnte damit begründet werden, dass diese Wörter im Referenzkorpus nicht oder selten mit „PayPal“ zusammen auftraten, was ähnlich wie bei der Differenzanalyse daraus resultierte, dass die Dokumente des Referenzkorpus im Gegensatz zu den Kurznachrichten in Schriftsprache verfasst waren.

## 5 Fazit

Das Finden von beweisheblichen Informationen in zahlreichen Chatnachrichten, die im Rahmen von Ermittlungen ausgewertet werden müssen, erweist sich oftmals als herausfordernd. Daher wurden in dieser Arbeit verschiedene Ansätze untersucht, um Suchbegriffe aus umgangssprachlichen Texten zu empfehlen. Insgesamt waren Topic Modeling, unter Berücksichtigung von globalen Wortkookkurrenzen, sowie die Analyse von paradigmatischen und syntagmatischen Relationen die Methoden, die besonders vielversprechende Ergebnisse erzielten. Diese zeichneten sich dadurch aus, dass sie auch Terme empfehlen konnten, die die Ermittler:innen nicht erwartet hätten.

Um endgültig eine Entscheidung treffen zu können, welche Methode sich am besten für die Empfehlung von Suchbegriffen eignet, ist eine ausführlichere quantitative Evaluierung der vorgestellten Ansätze erforderlich. Eine Möglichkeit, die von Joshi und Motwani (2006) vorgeschlagen wurde, besteht darin, die empfohlenen Begriffe von mehreren Annotatoren danach bewerten zu lassen, wie relevant und unerwartet diese sind, wobei als Annotatoren Ermittler:innen ausgewählt werden sollten, die über umfassende Kenntnisse über den Fall verfügen. Anschließend können Evaluierungsmaße des Information Retrievals berechnet werden. Die quantitative Evaluierung sollte zudem auf mehreren Datensätzen zu forensischen Fällen aus verschiedenen Deliktbereichen erfolgen.

## Literaturverzeichnis

- Abhishek, V., & Hosanagar, K. (2007). Keyword Generation for Search Engine Advertising using Semantic Similarity between Terms. Proceedings of the Ninth International Conference on Electronic Commerce, 89–94. Minneapolis, Minnesota, USA: Association for Computing Machinery (ACM).  
<https://doi.org/10.1145/1282100.1282119>
- Baeza-Yates, R., Hurtado, C., & Mendoza, M. (2004). Query Recommendation Using Query Logs in Search Engines. In W. Lindner, M. Mesiti, C. Türker, Y. Tzitzikas, & A. I. Vakali (Hrsg.), Proceedings of the International Conference on Current Trends in Database Technology (EDBT'04) (S. 588–596). Heraklion, Griechenland: Springer Science+Business Media.  
[https://doi.org/10.1007/978-3-540-30192-9\\_58](https://doi.org/10.1007/978-3-540-30192-9_58)
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 159, 1877–1901. Vancouver, Kanada: Association for Computing Machinery (ACM). Abgerufen von  
<https://dl.acm.org/doi/abs/10.5555/3495724.3495883>
- Chen, Y., Li, W., Guo, W., & Guo, K. (2015). Popular Topic Detection in Chinese Micro-Blog Based on the Modified LDA Model. Proceedings of the 12th Web Information System and Application Conference (WISA), 37–42. Jinan, China: IEEE Computer Society. <https://doi.org/10.1109/WISA.2015.58>
- Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. International Journal of Corpus Linguistics, 14(1), 29–59.  
<https://doi.org/10.1075/ijcl.14.1.03cul>
- de Arruda, H. F., Costa, L. da F., & Amancio, D. R. (2016). Topic segmentation via community detection in complex networks. Chaos: An Interdisciplinary Journal of Nonlinear Science, 26(6), 063120.  
<https://doi.org/10.1063/1.4954215>
- Diaz, G. (2023). Stopwords ISO. Abgerufen 10. Juni 2023, von <https://github.com/stopwords-iso/stopwords-de>
- Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), 61–74.
- Evans, G., & Gosalia, V. (2015). The Coming Storm: Companies Must Be Prepared to Deal with Text Messages on Employee Mobile Devices. Digital Discovery & E-Evidence, 1–4.
- Feldman, P. G., Pan, S., & Foulds, J. (2023). The Keyword Explorer Suite: A Toolkit for Understanding Online Populations. Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion), 21–24. Sydney, Australien: Association for Computing Machinery (ACM).  
<https://doi.org/10.1145/3581754.3584122>

- Fu, K., Lu, C.-T., Nune, R., & Tao, J. X. (2015). Steds: Social Media Based Transportation Event Detection with Text Summarization. Proceedings of the 18th International Conference on Intelligent Transportation Systems, 1952–1957. Gran Canaria, Spanien: IEEE Computer Society. <https://doi.org/10.1109/ITSC.2015.316>
- Gaikwad, K. S., & Patwardhan, M. S. (2014). Tweets clustering: Adaptive PSO. Proceedings of the Annual IEEE India Conference (INDICON), 1–6. Pune, Indien: IEEE Xplore. <https://doi.org/10.1109/INDICON.2014.7030584>
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211–244. <https://doi.org/10.1037/0033-295X.114.2.211>
- He, Q., Jiang, D., Liao, Z., Hoi, S. C. H., Chang, K., Lim, E.-P., & Li, H. (2009). Web Query Recommendation via Sequential Query Prediction. Proceedings of the 25th International Conference on Data Engineering, 1443–1454. Shanghai, China: IEEE Computer Society. <https://doi.org/10.1109/ICDE.2009.71>
- Heyer, G., Quasthoff, U., & Wittig, T. (2016a). Clustern von Wortformen. In *Text Mining: Wissensrohstoff Text* (4. Aufl., S. 209–219). Bochum: W3L-Verlag.
- Heyer, G., Quasthoff, U., & Wittig, T. (2016b). Differenzanalyse. In *Text Mining: Wissensrohstoff Text* (4. Aufl., S. 95–99). Bochum, Deutschland: W3L-Verlag.
- Jaiswal, A., & Janwe, N. (2011). Hierarchical Document Clustering: A Review. *International Journal of Computer Applications*, 3(1), 37–41.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21. <https://doi.org/10.1108/eb026526>
- Jones, T. W. (2021, Juni 27). Document clustering. Abgerufen 11. Juni 2023, von [https://cran.r-project.org/web/packages/textmineR/vignettes/b\\_document\\_clustering.html](https://cran.r-project.org/web/packages/textmineR/vignettes/b_document_clustering.html)
- Joseph, P. D., & Viswanathan, P. (2023). SDOT: Secure Hash, Semantic Keyword Extraction, and Dynamic Operator Pattern-Based Three-Tier Forensic Classification Framework. *IEEE Access*, 11, 3291–3306. <https://doi.org/10.1109/ACCESS.2023.3234434>
- Joshi, A., & Motwani, R. (2006). Keyword Generation for Search Engine Advertising. Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), 490–496. Hong Kong, China: IEEE Computer Society. <https://doi.org/10.1109/ICDMW.2006.104>
- Keerthana, S. (2017). Recommended Search of Documents from Conversation with Relevant Keywords Using Text Similarity. *Journal of Network Communications and Emerging Technologies*, 7(2), 1–6.
- Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and Language Learning with the Sketch Engine and SKELL. *Revue Française de Linguistique Appliquée*, 20(1), 61–80. <https://doi.org/10.3917/rfla.201.0061>
- Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5), 604–632. <https://doi.org/10.1145/324133.324140>
- Koven, J., Bertini, E., Dubois, L., & Memon, N. (2016). InVEST: Intelligent visual email search and triage. *Digital Investigation*, 18(16), 138–148. <https://doi.org/10.1016/j.diin.2016.04.008>
- Kubek, M., & Unger, H. (2014). On the Interactive Search with Web Documents. *International Journal of Computer, Information, Systems and Control Engineering*, 8(12), 1808–1813. <https://doi.org/10.5281/zenodo.1097152>
- Mastropierro, L., & Mahlberg, M. (2017). Key words and translated cohesion in Lovecraft's *At the Mountains of Madness* and one of its Italian translations. *English Text Construction*, 10(1), 78–105. <https://doi.org/10.1075/etc.10.1.05mas>
- McEnery, T., & Wilson, A. (2001). Approaching quantitative data. In *Edinburgh Textbooks in Empirical Linguistics. Corpus Linguistics* (S. 448–463). Edinburgh, Großbritannien: Edinburgh University Press.
- Mei, Q., Zhou, D., & Church, K. (2008). Query suggestion using hitting time. Proceedings of the 17th ACM Conference on Information and Knowledge Management, 469–478. Napa Valley, Kalifornien, USA: Association for Computing Machinery (ACM). <https://doi.org/10.1145/1458082.1458145>
- Otero, P. G. (2009). Comparing Different Properties Involved in Word Similarity Extraction. In L. S. Lopes, N. Lau, P. Mariano, & L. M. Rocha (Hrsg.), Proceedings of the 14th Portuguese Conference on Artificial Intelligence (EPIA) (S. 634–645). Aveiro, Portugal: Association for Computing Machinery (ACM). [https://doi.org/10.1007/978-3-642-04686-5\\_52](https://doi.org/10.1007/978-3-642-04686-5_52)
- Pastor, G. C. (2017). Collocational Constructions in Translated Spanish: What Corpora Reveal. In R. Mitkov (Hrsg.), Proceedings of the Second International Conference on Computational and Corpus-Based Phraseology (EUROPHRAS) (S. 29–40). London, Großbritannien: Springer International Publishing. [https://doi.org/10.1007/978-3-319-69805-2\\_3](https://doi.org/10.1007/978-3-319-69805-2_3)
- Quasthoff, U. (2009). Korpusbasierte Wörterbucharbeit mit den Daten des Projekts *Deutscher Wortschatz*. *Linguistik Online*, 39(3), 151–162. <https://doi.org/10.13092/lo.39.484>
- Quasthoff, U., Richter, M., & Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), 1799–1802. Genua, Italien: European Language Resources Association (ELRA). Abgerufen von <https://aclanthology.org/L06-1396/>



- Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- Rohrdantz, C., Koch, S., Jochim, C., Heyer, G., Scheuermann, G., Ertl, T., ... Keim, D. A. (2010). Visuelle Textanalyse: Interaktive Exploration von semantischen Inhalten. *Informatik-Spektrum*, 33(6), 601–611. <https://doi.org/10.1007/s00287-010-0483-x>
- Rus, V., Niraula, N. B., & Banjade, R. (2013). A Study of Probabilistic and Algebraic Methods for Semantic Similarity. *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 232–237. Florida, USA: Association for the Advancement of Artificial Intelligence (AAAI). Abgerufen von <https://digitalcommons.memphis.edu/facpubs/2420>
- Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *New Methods in Language Processing* (1. Aufl., S. 154–164). Manchester, Großbritannien: Routledge.
- Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application To German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzouker-mann, & D. Yarowsky (Hrsg.), *Natural Language Processing Using Very Large Corpora* (S. 13–25). Dublin, Ireland: Springer Science+Business Media. [https://doi.org/10.1007/978-94-017-2390-9\\_2](https://doi.org/10.1007/978-94-017-2390-9_2)
- Shimizu, T., Kido, J., & Yoshikawa, M. (2020). Keyword Recommendation Methods for Earth Science Data Considering Hierarchical Structure of Vocabularies. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL '20)*, 353–356. Virtual Event China: Association for Computing Machinery (ACM). <https://doi.org/10.1145/3383583.3398622>
- Singh, J., & Gupta, V. (2016). Text Stemming: Approaches, Applications, and Challenges. *ACM Computing Surveys*, 49(3), 1–46. <https://doi.org/10.1145/2975608>
- Spranger, M., Heinke, F., Appelt, L., Puder, M., & Labudde, D. (2016). MoNA: Automated Identification of Evidence in Forensic Short Messages. *International Journal On Advances in Security*, 9(1 & 2), 14–23.
- Spranger, M., Xi, J., Jaeckel, L., Felser, J., & Labudde, D. (2022). MoNA: A Forensic Analysis Platform for Mobile Communication. *Künstliche Intelligenz*, 36, 163–169. <https://doi.org/10.1007/s13218-022-00762-w>
- Spranger, M., Zuchantke, E., & Labudde, D. (2014). Semantic Tools for Forensics: Towards Finding Evidence in Short Messages. *Proceedings of the Fourth International Conference on Advances in Information Mining and Management (IMMM 2014)*, 1–4. Paris, Frankreich: IARIA. <https://doi.org/10.13140/RG.2.1.3391.3440>
- Teng, M. (2021). Using the Ship-Gram Model for Japanese Keyword Extraction Based on News Reports. *Complexity*, 21, 1–9. <https://doi.org/10.1155/2021/9965843>
- Tsur, O., Littman, A., & Rappoport, A. (2013). Efficient Clustering of Short Messages into General Domains. *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 621–630. Cambridge, Massachusetts, USA: Association for the Advancement of Artificial Intelligence (AAAI). <https://doi.org/10.1609/icwsm.v7i1.14420>
- Wang, Z., Hahn, K., Kim, Y., Song, S., & Seo, J.-M. (2017). A news-topic recommender system based on keywords extraction. *Multimedia Tools and Applications*, 77(4), 4339–4353. <https://doi.org/10.1007/s11042-017-5513-0>
- Yamada, M., Wang, X., & Yamasaki, T. (2021). Preference Analysis of Shopping Malls' Followers and Keyword Recommendation on Twitter. *Proceedings of the 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, 293–298. Tokyo, Japan: IEEE. <https://doi.org/10.1109/MIPR51284.2021.00055>
- Zhai, C., & Massung, S. (2016). Word Association Mining. In *ACM Books. Bd. 12. Text Data Management and Analysis—A Practical Introduction to Information Retrieval and Text Mining* (1. Aufl., S. 251–274). San Rafael, Kalifornien, USA: Morgan & Claypool. Abgerufen von <https://doi.org/10.1145/2915031>
- Zuo, Y., Zhao, J., & Xu, K. (2015). Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2), 379–398. <https://doi.org/10.1007/s10115-015-0882-z>