Extraktion und Analyse von Schlüsselwörtern in einer Literaturrecherche zu Quantum Computing

Mazlum Copurkuyu [1], Thomas Barton[2]

Abstract: Durch die große Menge an wissenschaftlichen Publikationen, die meist als unstrukturierte Daten vorliegt, nehmen Komplexität und Arbeitsaufwand eines Literature-Review Prozesses stetig zu. Auch im Forschungsgebiet Quantum Computing hat sich die Anzahl wissenschaftlicher Veröffentlichungen in den letzten Jahren stark erhöht. Dieser Beitrag gibt einen Überblick, wie man Text-Mining-Methoden zur Informationsextraktion bei der Literaturrecherche zu Quantum Computing einsetzen kann. Das zentrale Forschungsziel besteht in der Anwendung von Text-Mining zur automatischen Extraktion und Visualisierung von Schlüsselwörtern auf Basis der Abstracts von wissenschaftlichen Publikationen. Dieser Ansatz verwendet zum einen die TF-IDF-Methode und auf der anderen Seite den Word2Vec-Algorithmus, um die automatische Erfassung sowie die Verarbeitung relevanter Literatur zu ermöglichen. Anschließend wird eine visuelle Darstellung der Ergebnisse wie z.B. dynamische Word-Clouds durchgeführt. Aus der Analyse werden Erkenntnisse für den Forschungsbereich Quantum Computing abgeleitet.

Keywords: Literaturrecherche, Text-Mining, Keyword Extraction, TF-IDF, Word2Vec, Quantum Computing

1 Einführung

Eine grundlegende Aufgabe in der Forschung besteht darin, die vorhandene Literatur durch ein Literatur-Review zu identifizieren und zu verstehen, um den Kontext zu ermitteln, weiterführende Forschung zu betreiben sowie die Forschungsgemeinschaft auf dem neuesten Stand zu halten [Ta20]. Zu diesem Zweck ist es zwingend notwendig, die vorhandene Literatur zu einem Forschungsthema zu berücksichtigen. Diese Aufgabe ist jedoch bei der ständig steigenden Anzahl von Veröffentlichungen kaum noch zu bewältigen, und ihre Auswertung gestaltet sich schwierig. Um das Problem der Informationsflut zu lösen, sollte der Prozess der Literaturrecherche strukturiert sein und etablierten Rahmenwerken folgen.

Einer der bekanntesten und weit verbreiteten Vorgehensweise zur Durchführung eines (manuellen) Literatur-Reviews ist das Framework vom Brocke et al. [vo09], das in Abb. 1 dargestellt ist.

1 Hochschule Worms, Fachbereich Informatik, Erenburgerstr.19, 67549 Worms
2 Hochschule Worms, Fachbereich Informatik, Erenburgerstr.19, 67549 Worms, barton@hs-worms.de, https://orcid.org/0000-0001-6736-7040

https://doi.org/10.30844/AKWI_2022_16

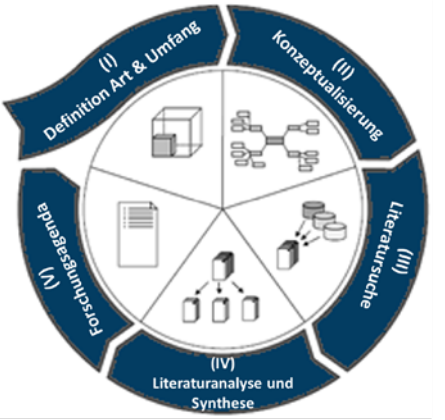

Abb. 1: Prozess zur Durchführung einer Literaturanalyse (vom Brocke et al. [vo09])

Die Durchführung eines Literaturreviews ist jedoch aufgrund der vielen manuellen Tätigkeiten wie Suchen und Herunterladen, Dokumentation des Prozesses, Textscreening usw. sehr zeitaufwändig und mühsam. Das Verfahren zur Durchführung von Literatur-Reviews zu unterstützen und zu automatisieren wäre sehr hilfreich, um mit der immer schnelleren Veröffentlichung von Artikeln Schritt halten zu können [Ta20]. Dabei gilt es zu prüfen, ob die zeitaufwendige Arbeit für die Auswertung sowie die Analyse von wissenschaftlichen Artikeln mit Hilfe von Text-Mining unterstützt werden kann.

In dieser Arbeit werden Methoden des Text-Mining eingesetzt, um Informationen aus wissenschaftlichen Publikationen zu generieren. Hierbei werden Verfahren der Keyword-Extraction und Methoden für Word-Embeddings zur Extraktion von Informationen aus den wissenschaftlichen Abstracts verwendet. Das Ziel ist es, Muster und Verbindungen zu verwandten Themen in einem Forschungsbereich zu ermitteln.

2 Grundlagen

Dieses Kapitel führt in das Quantum Computing und Text-Mining ein. Zuerst wird der Begriff Quantum Computing erläutert. Anschließend wird ein Überblick über Text-Mining gegeben und damit die für diesen Beitrag essenziellen Begriffe sowie Zusammenhänge vermittelt.

2.1 Quantum Computing

Die global vernetzte mobile Welt macht eine nachhaltige und effiziente Verarbeitung und Analyse von Informationen zu einer immer komplexeren Herausforderung. An dem Punkt, an dem klassische Prozessoren heute an ihre physikalischen Grenzen stoßen, könnten Quantencomputer in Zukunft Lösungen liefern. Diese basieren auf der Wechselwirkung quantenmechanischer Zustände [Ba20].

Ein Quantencomputer arbeitet nicht mit klassischen Bits, sondern mit Quantenbits - kurz Qubits. Sie können z.B. durch den Spin eines Elektrons gebildet werden. Die Funktion von Qubits in Quantencomputern basiert auf zwei Schlüsselprinzipien der Quantenphysik: Superposition und Verschränkung, die im Folgenden kurz erläutert werden [Ba20].

Superposition bedeutet, dass ein Qubit nicht nur den Zustand 1 oder 0 einnehmen kann, sondern beide „gleichzeitig“ sowie alle Zustände dazwischen, ähnlich wie eine sich drehende Münze [Ho18]. Solange sie sich bewegt, kann das Ergebnis sowohl Kopf oder Zahl werden. Erst wenn man sie stoppt und den Ausgang misst, wird das Ergebnis bestimmt [Ho18]. Durch die Messung wird dieser Quantenzustand, die Superposition, zerstört. Solange man den Zustand eines Qubits nicht misst, beschreibt man diesen als Superposition, also als Überlagerung aller möglichen Zustände [Ho18]. Der Zustand der Superposition wird durch die Messung und die damit verbundene Interaktion entweder zu 0 oder 1 kollabieren [Ba20].

Die Verschränkung ist der Schlüssel zur Effizienz von Quantencomputern. Verschränkung bedeutet, dass Qubits miteinander unter Wechselwirkungen stehen [Ba20]. Die Verschränkung funktioniert nur, wenn beide Qubits sich in Superposition befinden. Beeinflusst man eines, werden zeitgleich auch alle Partner-Qubits beeinflusst. Es besteht also eine Korrelation zwischen den Zuständen beider Qubits. Sobald der Zustand des einen gemessen wird, ist auch der Zustand des zweiten Qubits bekannt und so kollabiert die Superposition beider Qubits [Ba20]. Dieser Umstand gilt auch bei verschränkten Qubits, die voneinander räumlich getrennt sind. Es sei erwähnt, dass die Verschränkung nach der Messung der Qubits aufgehoben wird [Ho18].

Dadurch ist es möglich, in einem Quantencomputer parallele Berechnungen durchzuführen, aber der Zugriff auf die Ergebnisse der Berechnungen sind eingeschränkt. Der Zugriff auf die Ergebnisse ist gleichbedeutend mit einer Messung, wodurch der Quantenzustand gestört wird. Da die Messung probabilistisch ist, erhält man ein zufälliges Ergebnis [Ba20]. Zudem muss auch die Superposition der Qubits möglichst lange erhalten bleiben, also die Kohärenzzeit möglichst lang sein, denn selbst kleinste Zustandsänderungen verändern den Rechenvorgang. Weiterhin ist die Übertragung der Informationen so empfindlich, dass sie bereits durch das Rauschen der Elektronik gestört werden kann [Ba20]. Solche Fehlerquellen zu minimieren, ist eine spannende Herausforderung.

Die potenziellen Vorteile und Anwendungen des Quantum Computing für die Gesellschaft sind vielfältig. Richtig eingesetzt, sind Quantencomputer unglaublich schnell und effektiv.

Sie können in wenigen Sekunden Berechnungen durchführen, für die die heutigen klassischen Supercomputer viel länger brauchen würden. Diese Tatsache wird von Experten auch als Quantenüberlegenheit bezeichnet [Ba20]. Quantencomputer stellen einen Paradigmenwechsel in der Datenverarbeitung dar. Umso faszinierender ist es, dieses Gebiet zu verfolgen. Daher wird in dieser Arbeit dieses Forschungsgebiet mithilfe von Text-Mining-Verfahren genauer untersucht, um die Themenschwerpunkte, die bisherige Entwicklung sowie die zukünftigen Trends des Forschungsgebietes – sofern möglich – zu erfassen.

2.2 Überblick über Text-Mining

Text-Mining ist ein weit gefasster Oberbegriff, die eine Reihe von unterschiedlichen Verfahren zur Analyse und Verarbeitung von semistrukturierten und unstrukturierten Textdaten beschreiben. Das gemeinsame Ziel hinter jeder dieser Verfahren ist „Texte in Zahlen zu verwandeln“, damit leistungsstarke Algorithmen auf große Dokumenten-datenbanken angewendet werden können [Mi12]. Die Konvertierung von Text in ein strukturiertes numerisches Format und die Anwendung analytischer Algorithmen erfordert das Verständnis, wie man diese Verfahren für die Analyse von Texten sowohl nutzt als auch kombiniert [Mi12].

Text-Mining ist ein weitgehend automatisierter Prozess für die Gewinnung von neuen Erkenntnissen aus Textdokumenten [Mi12]. Die Anwendungen des Text-Mining sind so breit gefächert und ihre Ziele so vielfältig, dass es schwierig ist, seine Leistung in allgemeinen Worten zu beschreiben. Im Gegensatz zu anderen etablierten statistischen Methoden ist Text-Mining eine relativ neue und nicht standardisierte analytische Methode zur Wissensentdeckung. Folglich ist es eine Herausforderung, eine allgemeine Vorgehensweise für Text-Mining zu erstellen. Anwendungen des Text-Mining werden in erster Linie durch “trial and error“ auf der Grundlage persönlicher Erfahrungen und Vorlieben bestimmt [Mi12]. Während Data-Mining-Methoden relativ ausgereift sind, gibt es in der Literatur keine allgemein akzeptierten Methoden, die die besten Praktiken im Text-Mining in einem beliebigen Bereich widerspiegeln.

Feldmann [FS07] geht auf funktionaler Ebene davon aus, dass Text-Mining dem allgemeinen Modell einiger klassischer Data-Mining-Anwendungen folgt. Auch Hippner und Rentzmann [HR06] betonen, dass Text-Mining einen ähnlichen Aufbau wie ein klassischer Data-Mining-Prozess aufweist, sich aber in der Datenaufbereitung unterscheidet, weil beim Text Mining eine zusätzliche linguistische Datenaufbereitung erforderlich ist, damit die fehlende Datenstruktur rekonstruiert werden kann.

Die Vorgehensmodelle in der Literatur sind zu allgemein, um sie für einen konkreten Anwendungsfall einzusetzen. Die Prozessmodelle aus der Literatur können aber als Grundgerüst dienen und auf den Anwendungsfall dieser Arbeit übertragen werden (siehe folgenden Abschnitt); wobei die große Bedeutung der Datenvorverarbeitungsphase hervorzuheben ist, da sie eine wichtige Rolle in Text-Mining spielt.

3 Einsatz von Text-Mining

Die Abb. 2 zeigt das Vorgehen für die durchgeführte Studie. Es basiert auf dem Vorgehensmodell von Tauchert et al. [Ta20].

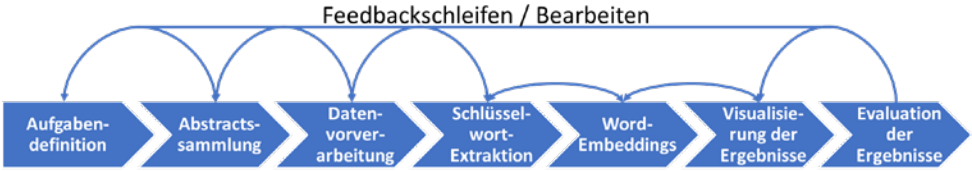

Abb. 2: Grafische Darstellung zur Vorgehensweise bei der Durchführung der Literaturrecherche und -auswertung, die mit Hilfe von Text Mining unterstützt wird [Ta20]

Zunächst werden die Phasen nachstehend kurz erläutert, anschließend werden diese im darauffolgenden Kapitel angewendet. Eine erst kürzlich erschienene Publikation hat die Entwicklung einen digitalen Plattform zum Inhalt, die Forschende dabei unterstützt, eine systematische Literaturanalyse zu unterstützen [Am22].

3.1 Phase 1: Aufgabendefinition

Ähnlich zu Schieber et al. [SH14] und Hippner et al. [HR06] erfolgt eine Festlegung des Zwecks der Recherche und ein Festlegung der Suchbegriffe und der Datenquellen. Um ein gründliches Verständnis zu erlangen und die Ziele genau zu definieren, ist das notwendige Wissen zur Durchführung eines Literatur-Reviews zu ermitteln. Die Definition der Problem- sowie Zielstellung dieser Arbeit ergeben sich aus der Einleitung (siehe Kapitel 1).

3.2 Phase 2: Abstracts-Sammlung

Die zweite Phase besteht daraus, die wissenschaftlichen Publikationen zu selektieren und die ermittelten Abstracts herunterzuladen, die in den nachfolgenden Prozessschritten analysiert werden sollen. Dazu wird eine Recherche in der englischsprachigen ACM Digital Library [AC22] durchgeführt, und Abstracts zu wissenschaftlichen Publikationen heruntergeladen, die dem vom Benutzer definierten Suchbegriff und dem Datumsbereich entsprechen.

3.3 Phase 3: Aufbereitung der Abstracts mittels NLP-Methoden

In Text-Mining liegen die Textdaten in natürlicher Sprache, d.h. in halbstrukturierter Form, vor. Es ist schwierig, Muster aus halbstrukturierten Daten zu extrahieren, da diese Daten nicht für Text-Mining-Verfahren einsetzbar sind. Um Text-Mining durchzuführen, ist es daher notwendig, die Rohtexte einem Prozess zu unterziehen, bei dem verschiedene Methoden auf sie angewandt werden. Als Ergebnis werden bereinigte Token zurückgegeben. Token sind einzelne Wörter oder Wortgruppen, die als Merkmale für weitere Analysen sowie als Input für Text-Mining-Verfahren dienen [FS07]. Diese verfeinerten Daten sind für den Anwender besser geeignet, um sachkundige Daten zu extrahieren. Dieser Prozess wird als „Vorverarbeitung" von Daten bezeichnet. Die Auswahl der Textvorverarbeitungsverfahren beeinflussen die Darstellung des Textdokuments und damit auch die darauffolgenden Ergebnisse erheblich [FS07]. Viele der Methoden zur Textvorbereitung und -aufbereitung haben ihre Wurzeln in Natural-Language-Processing.

Es wurden drei Vorverarbeitungsmethoden gewählt, wobei die erste die Filterung von Stoppwörtern ist. Dies ist eine einfache und schnelle Methode, um Wörter herauszufiltern, die keine Bedeutung haben und nur in Verbindung mit anderen Wörtern verwendet werden [AHN19]. Das zweite gewählte Verfahren war die Lemmatisierung. Da die eingesetzten Text-Mining-Methoden zur Schlagwortextraktion auf der Häufigkeit des Auftretens eines Wortes in einem Dokument beruhen, ist es notwendig, das Lemma zu identifizieren. Dadurch wird sichergestellt, dass verschiedene morphologische Varianten eines Wortes auf ihr gemeinsames Lemma zurückgeführt und somit als Teil seiner Häufigkeit gezählt werden. In diesem Fall wurde die Lemmatisierung dem Stemming vorgezogen, denn beim Stemming wird einfach ein Wort aufgrund seines Aussehens ausgeschnitten, während die Lemmatisierung ein kalkulierter Prozess ist [AHN19]. Bei der Lemmatisierung ist außerdem zu beachten, dass es oft schwieriger ist, diese für eine neue Sprache durchzuführen als für einen Stemming-Algorithmus. Da eine Lemmatisierung viel mehr Wissen über die Struktur einer Sprache erfordert. Glücklicherweise sind alle Abstracts auf Englisch, so können vorgefertigte Pakete in Python bei der Lemmatisierung verwendet werden. Um gute Ergebnisse zu erzielen, sollten vorher jedoch Tags bezüglich der Wortart mithilfe von POS-Tagging an die Lemmatisierung übergeben werden, da sonst möglicherweise nicht alle Wörter auf die gewünschten Lemmata reduziert werden können [AHN19].

3.4 Phase 4: Keyword-Extraktion

Nachdem alle relevanten Abstracts in einem einheitlichen Format vorliegen, werden im nächsten Schritt die wichtigsten Stichwörter daraus extrahiert. Keywords sind Schlüsselbegriffe, die dabei helfen, die Dokumente zusammenzufassen und einen Überblick über den Inhalt zu vermitteln [FPW99]. Das manuelle Auffinden der relevanten Schlüsselbegriffe, die den Inhalt des Dokuments beschreiben, kann je nach Länge und Anzahl der Dokumente sehr anspruchsvoll sein. Daher wäre die Anwendung eines automatisierten Prozesses, der Schlüsselwörter aus Dokumenten extrahiert, sinnvoll. Im Bereich des Text-Mining gibt es

zahlreiche Methoden der Keyword-Extraction, die z.B. in statistische, überwachte, teilüberwachte und unüberwachte Ansätze kategorisiert werden können [SS15].

Bereits Barton und Kokoev [BK21] nutzten die unüberwachte Lernmethode RAKE, um Schlüsselwörter aus den wissenschaftlichen Arbeiten zu extrahieren. Für diesen Beitrag wird die statistisch-basierte TF-IDF-Methode ausgewählt, da sie Schlüsselwörter für ein bestimmtes Dokument in einer Dokumentensammlung gewichtet. Hierbei werden Wörter als einzigartig (Schlüsselwort) oder unwichtig (Nicht-Schlüsselwort) für das Dokument und somit für den gesamten Korpus eingestuft. Die zu Grunde liegende Idee ist, eine Dokumentensammlung als Beurteilung für die Worthäufigkeit zu verwenden, um die Wichtigkeit eines einzelnen Wortes für ein Dokument festzustellen [MRS08]. Wenn ein Term in vielen Dokumenten vorkommt, geht man davon aus, dass dessen Bedeutung gering ist. Falls ein Begriff in wenigen Artikeln sehr häufig vorkommt, scheint die Relevanz des relevanten Wortes für diese bestimmten Artikel hoch zu sein.

Zunächst berechnet man die Komponente „Termfrequenz“ (TF). Diese bewertet (gewichtet) jedes Wort entsprechend der Vorkommenshäufigkeit. Nun wird ein Blick auf die Formel geworfen. Davor werden zunächst folgende Notationen definiert:

- ist die Gesamtanzahl der Dokumente
- ist ein gegebenes Dokument aus dem Datenbestand
- ist Sammlung aller Dokumente
- ist ein bestimmtes Word in einem Dokument

TF wird nach der folgenden Formel berechnet

$$tf(w,d)=\log(1+f(w,d)),$$

wobei die reine (Vorkommens) Häufigkeit des Wortes in Dokument ist. Die Logarithmierung wurde verwendet, um die Termfrequenz der Dokumente und des Korpus zu normalisieren.

Mit der obigen Formel werden alle Terme als gleich wichtig angesehen. Daher führt man die zweite Komponente „inverse Dokumentenhäufigkeit“ ein, um das Gewicht von Begriffen zu verringern. Die inverse Dokumentenhäufigkeit beschreibt, wie viel Information ein Wort liefert, d. h. ob es in allen Dokumenten häufig oder selten vorkommt. Man definiert die inverse Dokumenthäufigkeit () eines Terms wie folgt [MRS08]:

$$\mathrm{idf}(w, D) = \log\left(\frac{N}{f(w, D)}\right)$$

Man kombiniert nun die Definitionen der Termhäufigkeit und der inversen Dokumenthäufigkeit. Dadurch erhält man das statistische Maß TF-IDF, welches dem Term eine Gewichtung im Dokument zuweist, das durch

$$tfidf(w,d,D)=tf(w,d)*idf(w,D)$$

gegeben ist [MRS08].

Diese Arbeit untersucht, wie sich die Themen im Bereich Quantum Computing innerhalb jedes Jahres sowie im Zeitverlauf entwickelt haben. Die Themen werden hierbei als Schlüsselwörter betrachtet und die wichtigsten Begriffe werden auf der Grundlage der Ergebnisse des TF-IDF-Algorithmus pro Jahr identifiziert. Dazu berechnet man den durchschnittlichen TF-IDF-Score aller Wörter über alle Abstracts aus einem bestimmten Jahr.

3.5 Phase 5: Wortrepräsentation mit Word2Vec

Die Methode Worteinbettung (englisch “word embedding“) wird in der Computerlinguistik verwendet, um die Bedeutung eines Wortes aus seinem Kontext zu extrahieren bzw. darzustellen. Dieses Verfahren stellt Wörter als numerische Wortvektoren dar, bei denen semantisch ähnliche Wörter auf nahe gelegene Punkte im geometrischen Raum abgebildet werden. Wörter, die in einem ähnlichen Kontext verwendet werden, werden in einem unmittelbaren Vektorraum abgebildet. Das Praktische an der Darstellung von Wörtern als Vektoren ist, dass sie sich dadurch für mathematische Operatoren eignen. Zum Beispiel kann man Vektoren miteinander addieren und subtrahieren.

Es gibt mehrere mögliche Algorithmen, die die Aufgabe der Berechnung von Worteinbettungen lösen können. Angesichts der Beliebtheit und Nützlichkeit wurde für dieses Projekt der Word2Vec-Algorithmus gewählt. Word2Vec ist ein neuronales Netz und kann in Kombination mit der Kosinus-Ähnlichkeitsformel dazu verwendet werden, ähnliche Wörter für ein vorgegebenes Wort zu finden. Das Verfahren wurde im Jahr 2013 von Mikolov et al. [Mi13] bei Google entwickelt. Es sei erwähnt, dass in diesen Beitrag die Funktionsweise sowie die mathematischen Details des word2vec-Modells nicht weiter eingegangen wird. Weiterführende Information sind bei Goldberg et al. [GL14] und Weng [We21] zu finden.

Die Vorgehensweise hier ist es, die Schlüsselwortextraktion zu erweitern. Die Basis-Schlüsselwörter können aus der vorherigen Phase 4 (siehe Abschnitt 3.4) identifiziert werden. Anschließend können weitere ähnliche Schlüsselwörter, die im gleichen Kontext wie die Basis-Schlüsselwörter vorkommen, mithilfe des Word2Vec-Verfahrens sowie dem Kosinus-Ähnlichkeitsmaß ermittelt werden. In diesem Beitrag wird Word2Vec mit demselben Datensatz trainiert, aus dem die vorherigen Schlüsselwörter erstellt wurden. Hierbei dienen die eindeutig identifizierten Schlüsselwörter aus Abschnitt 3.4 als Input, um eine neue Liste von weiteren Schlüsselwörtern zu erhalten.

3.6 Visualisierung der Ergebnisse

Der letzte fehlende Baustein für dieses Projekt ist die Visualisierung der Ergebnisse. Daher werden im Folgenden kurz Informationen über die beiden verwendeten Visualisierungsmethoden bereitgestellt.

Ein Bündel visuell dargestellter Wörter wird als Tag-Cloud, auch Word-Cloud (Wortwolke), bezeichnet [Ku07]. Dieses Bündel von Wörtern wird nach bestimmten Kriterien ausgewählt und hilft bei der explorativen Textanalyse, wichtige Wörter und kontextbezogene Themen aus einem großen Textkorpus hervorzuheben [Ku07]. Im Allgemeinen wird bei der Erstellung von Tag-Clouds die Darstellungsgröße der Wörter herangezogen. Diese ist abhängig von der Häufigkeit eines Wortes. D. h. je häufiger ein bestimmtes Wort in einem Text vorkommt, desto größer erfolgt seine Darstellung in der Tag-Cloud. In dieser Arbeit wird die Häufigkeit durch den berechneten TF-IDF-Wert der Wörter aus der Phase 4 ersetzt und dadurch wird für die Abbildung von Schlüsselwörtern eine relativ aussagekräftige Darstellung ermöglicht.

Für eine Visualisierung der mehrdimensionalen Wort-Vektoren (siehe Abschnitt 3.5) müssen diese in zweidimensionale Wort-Vektoren umgewandelt werden. Dies wird mittels t-Distributed Stochastic Neighbor Embedding (t-SNE) umgesetzt. T-SNE ist ein unüberwachtes und nicht-lineares Verfahren, das von Laurens van der Maatens und Geoffrey Hinton [vH08] im Jahr 2008 entwickelt wurde.

Die Grundidee von t-SNE besteht darin, den mehrdimensionalen Raum unter Beibehaltung des relativen paarweisen Abstands zwischen den Punkten zu reduzieren, wodurch die Abstände zwischen den Punkten identisch bleiben. D. h. der Algorithmus bildet mehrdimensionale Daten auf zwei oder mehr Dimensionen ab, wobei die Punkte, die ursprünglich weit voneinander entfernt waren, auch weit entfernt liegen, und nahe Punkte auch in nahe Punkte umgewandelt werden. Man kann sagen, dass t-SNE nach einer neuen Datendarstellung sucht, bei der die Nachbarschaftsbeziehungen erhalten bleiben.

4 Ergebnisse der Literaturanalyse mittels Text-Mining-Methoden

In diesem Kapitel werden die Ergebnisse der Text-Mining-Anwendung nach der zuvor beschriebenen Methode vorgestellt. Alle Wörter, die mit dem TF-IDF- oder Word2Vec-Verfahren extrahiert wurden, werden bewusst kursiv geschrieben.

4.1 Extraktion von wissenschaftlichen Abstracts

Für diesen Beitrag wird der Suchbegriff „Quantum Computing“ ausgewählt. Um die Suche zu erweitern, wurden die Begriffe „quantum computation“, „quantum computer“ und

„quantum computers“ hinzugefügt. Die daraus resultierende Treffermenge wurde durch die Auswahl „Research Article“ in der Kategorie „Content Type“ weiter eingegrenzt. Für eine bessere Übersicht wurden ausschließlich Artikel von 2008 bis 2021 in Betracht gezogen. Vor dem Jahr 2008 sowie – bedingt durch den Zeitpunkt der Recherche – nach dem Jahr 2021 standen wenige wissenschaftlichen Arbeiten zur Verfügung.

Das folgende Balkendiagramm in Abb. 3 zeigt die Gesamtzahl der Veröffentlichungen für jedes einzelne Jahr. Es ist zu erkennen, dass die Anzahl der Veröffentlichungen von 2008 bis 2021 im Bereich Quantum Computing deutlich gestiegen ist. Im Jahr 2008 wurden in diesem Bereich nur 29 Research-Papers veröffentlicht, während die entsprechende Zahl der Arbeiten im Jahr 2021 196 betrug. Das Wachstum nimmt über die Jahre stetig zu, vor allem ab dem Jahr 2015 ist eine stärkere Zunahme der Veröffentlichungen zu verzeichnen.

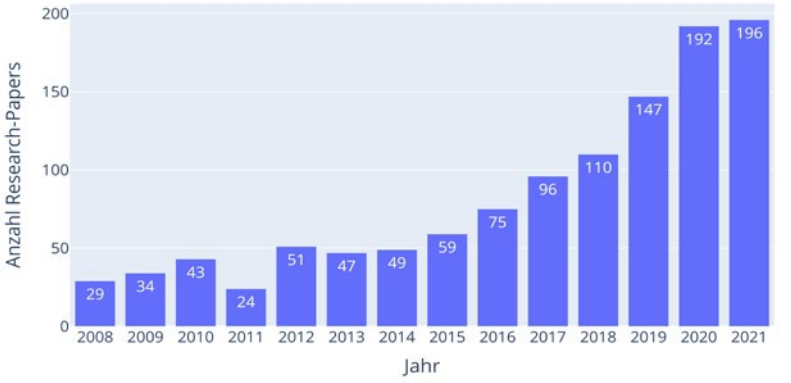

Abb. 3: Entwicklung der Anzahl von wissenschaftlichen Publikationen zu Quantum Computing zwischen 2008 und 2021

4.2 Extraktion der Schlüsselwörter sowie deren Visualisierung mithilfe von Wortwolken

Im Folgenden werden diese Schlüsselwörter ermittelt und visuell analysiert. Hierbei wird die Methode von Weiwei et al. [We10] übernommen. Die vorgeschlagene Methode verwendet die dynamische Tag-Clouds, um die zeitliche Entwicklung von Inhalten zu veranschaulichen. Diese zeitbasierten Word-Clouds stellen verschiedene statische Wortwolken zu konkreten Zeitpunkten dar. Dadurch entsteht ein aussagekräftiges Trenddiagramm, das die Veränderung der Word-Clouds im Laufe der Zeit darstellt. So erhält man einen visuellen Überblick über die unterschiedlichen thematischen Schwerpunkte.

Bei der Erstellung von Word-Clouds werden die durchschnittlichen TF-IDF-Werte herangezogen. Die Größe der Wörter spiegelt den durchschnittlichen TF-IDF-Wert für das Wort in diesem entsprechenden Jahr wider und gibt dabei die Wichtigkeit für das jeweilige Schlüsselwort – in diesem Fall sind es Bigramme – Abb. 4 zeigt die Wortwolken zu dem beschriebenen Korpus aus dem obigen Abschnitt. Es ist anzumerken, dass Begriffe wie *quantum computing*, *quantum computation*, *quantum computer* und *quantum computers* aus dem Korpus entfernt wurden, da ihre Verwendungshäufigkeit beträchtlich wäre und die anderen Begriffe überdecken würde.

Für die Visualisierung werden zunächst die Abstracts nach Jahr gruppiert und für jedes Jahr wird eine Wortwolke mit jeweils 15 Bigrammen erstellt. Die folgende Abb. 4 zeigt die Ergebnisse für die Jahre von 2019 bis 2021.

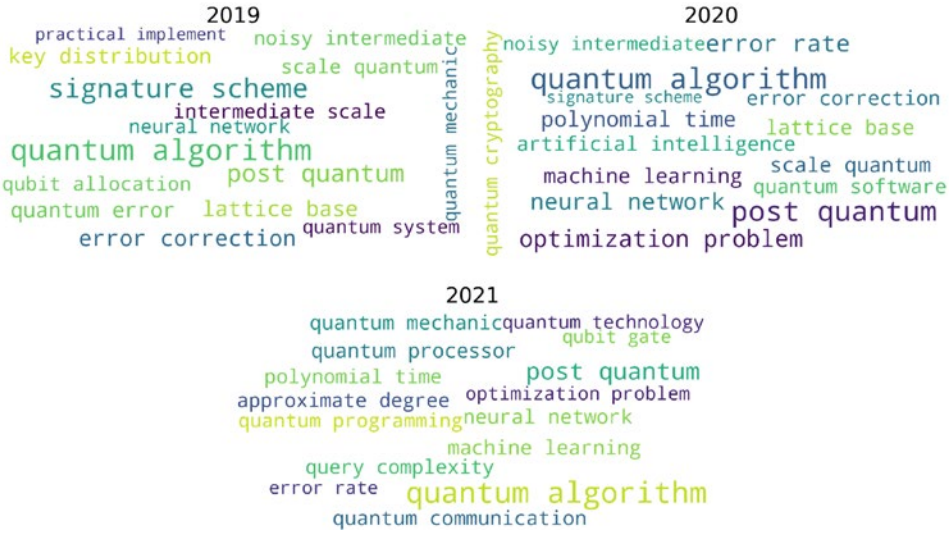

Abb. 4: Visualisierung von extrahierten Schlüsselwörtern mit Hilfe von Word-Clouds für die Jahre 2019 bis 2021

4.3 Zukunft des Quanten Computing

Auf Basis der Wortwolken konnten wichtige Themen für die Zukunft des Quanten Computing identifiziert werden, z. B. *noisy intermediate*, *intermediate scale*, *machine learning*, *artifical intellegence*, *neural network* und *error correction*. Die beiden Bigramme *noisy intermediate* sowie *intermediate scale* deuten auf das offene Kompositum „Noisy Intermediate Scale Quantum Computing“, kurz NISQ.

Im Folgenden wird die zeitliche Relevanz der Wörter untersucht. Für die zeitliche Betrach-

tung der extrahierten Schlüsselwörter wird eine Kennzahl eingeführt, um eine objektivere Interpretation zu ermöglichen sowie die Bedeutung von Forschungsthemen im zeitlichen Verlauf abzubilden. Die Kennzahl spiegelt die durchschnittliche Anzahl der Schlüsselwörter pro Artikel für jedes Jahr. Anschließend wird diese Zahl wird für die Erstellung der Liniendiagramme verwendet, wie die folgende Abb. 5 zeigt:

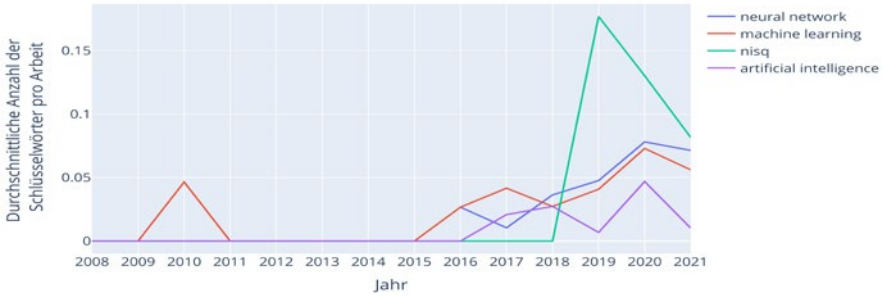

Abb. 5: Zeitliche Entwicklung von Publikationen zu ausgewählten Schlüsselwörtern

Ab dem Jahr 2015 wächst die Bedeutung von Themen wie *nisq*, *machine learning* und a*rtificial intelligence*. Die hohe Anzahl an wissenschaftlichen Publikationen in den letzten drei Jahren (siehe Abb. 3) sowie diese genannten Trends deuten auf eine rasche Ausweitung des Forschungsgebiets Quanten Computing hin und zeigen, dass sich mehr und mehr Wissenschaftler für das Quantum Computing interessierten.

Anhand der identifizierten Stichwörter lässt sich die Zukunft von Quanten Computing in drei Phasen einteilen:

Die erste Phase lautet NISQ. Dies ist die Ära der kleinen, fehleranfälligen oder „noisy intermediate scale quantum“ (NISQ) Maschinen, wie Preskill [Pr18] es zum ersten Mal im Jahr 2018 formuliert hat. Verrauscht deshalb, weil man nicht genügend Qubits für die Quantenfehlerkorrektur zur Verfügung hat. Und „*Intermediate Scale*“ wegen ihrer kleinen Qubit-Anzahl.

Nach der NISQ-Ära würde das Rauschen verhindern, dass Quantencomputer mit größerer Anzahl an Qubits eingesetzt werden. Quantencomputer benötigen daher eine Fehlerkorrektur. Ein Ziel wird daher sein, ein verlässliches Verfahren für die Quanten-Fehlerkorrektur in einem Quantencomputer zu implementieren, welches das Rauschen schneller korrigiert als es erzeugt wird. Daher wird solch ein Fortschritt in der Forschung als ein technologischer Meilenstein für die Entwicklung eines fehlertoleranten Quantencomputers gesehen.

Die letzte Phase betrifft vor allem das Wort *machine learning*. Für die nähere Untersuchung der letzten Phase wurde mit dem t-SNE-Algorithmus und dem Word2-Vec-Modell eine zweidimensionale Darstellung der semantisch ähnlichen Wörter im Umfeld des Bi-

gramms *machine learning* durchgeführt. Es werden zunächst die ersten 50 Datenpunkte manuell ausgewertet und anschließend diejenigen mit einer Beschriftung angezeigt, die von Interesse sein könnten. Es sei erwähnt, dass eine Messung der Distanz zwischen zwei Wortvektoren zu keiner gültigen Aussage führt. In der Abb. 6 kann der Abstand zwischen zwei Wortvektoren lediglich als nah oder weit entfernt bezeichnet werden.

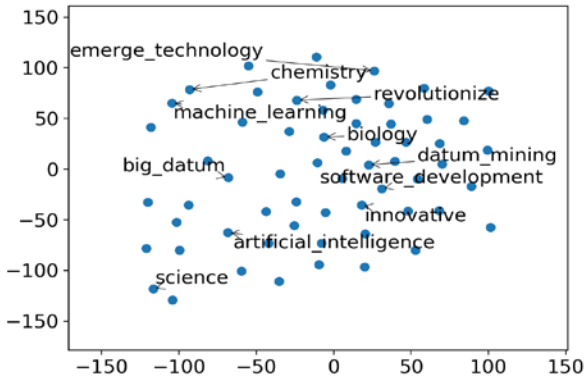

Abb. 6: Zweidimensionale Darstellung semantisch ähnlicher Wörter zu „*machine learning*“

Unter Quanten-Machine-Learning versteht man den Einsatz von Quantencomputern für das maschinelle Lernen. Die in Abb. 6 dargestellten Wörter *innovative*, *emerge technology* sowie *revolutionize* weisen darauf hin, dass Quantum-Machine Learning ein aufstrebendes Gebiet in Wissenschaft und Technologie ist. Es ist der Schnittpunkt von Quantenphysik und maschinellem Lernen. Das maschinelle Lernen könnte mithilfe von Quantencomputern zukünftig in den Bereichen *biology*, *chemistry* und *design automation* eingesetzt werden.

5 Fazit

Im Rahmen einer Literaturrecherche und -analyse zu Quanten Computing lassen sich mit Hilfe von Text-Mining-Methoden Schlüsselwörter identifizieren und auswerten, welche die zeitliche Entwicklung relevanter Themenfelder für das Forschungsgebiet Quanten Computing gut beschreiben. Auch aktuelle Themenfelder, die in einzelnen Zeiträumen vorherrschend waren, lassen sich so ableiten. Die Auswertung einer Literaturanalyse kann mit Hilfe von Text-Mining unterstützt und beschleunigt werden.

Die Anwendung der dynamischen Word-Clouds und der TF-IDF-Methode führten zu dem Ergebnis, dass sich die Forscher/innen in den letzten Jahren verstärkt für die Themen maschinelles Lernen, künstliche Intelligenz, NISQ und Fehlerkorrektur konzentrierten. Es ist zu erwarten, dass diese auch die zukünftigen Trends im Bereich des Quanten Computing

bilden werden, da viele Forschungsfragen noch offen sind. Die Einführung eines zusätzlichen Worteinbettungsverfahren erwies sich als nützlich für das weitere Verständnis der extrahierten Schlüsselwörter und für die Identifikation weiterer Themen und möglicher Trends.

Nach wie vor sind Forschende gefordert, die Literatur zu verstehen, zu verarbeiten und relevantes Wissen zu extrahieren. Insbesondere müssen die Schlüsselwörter, die mithilfe von TF-IDF oder Word2Vec extrahiert werden, genau überprüft werden, da sie nicht immer die eigenen Erwartungen widerspiegeln. So kann beispielsweise eine Word-Cloud ein Schlüsselwort beinhalten, welches zur weiteren Untersuchung weniger geeignet ist. Daher empfiehlt es sich, auf Basis der Word-Clouds oder t-SNE-Visualisierung Stichwörter manuell auszuwählen und diese in einen Wörterkatalog zu übertragen. Darüber hinaus müssen auch Forschungslücken weiterhin eigenständig von Forscher/innen identifizieret werden.

Zusammenfassend lässt sich sagen, dass die Extraktion von Schlüsselwörtern einen unterstützenden Mechanismus zur Beschleunigung des Literature-Review-Prozesses bietet und als ein Hilfsinstrument für die Teilautomatisierung eines manuellen Prozesses betrachtet werden kann.

Literaturverzeichnis

[AC22] ACM Digital Library: ACM Digital Library. https://dl.acm.org/, Abruf am 01.04.2022.

[AHN19] Anandarajan, M.; Hill, C.; Nolan, T.: Practical Text Analytics. Maximizing the Value of Text Data. Springer International Publishing, Cham, 2019.

[Am22] Ammirato, S et al. Digitalising the Systematic Literature Review process: the MySLR platform. Knowledge Management Research & Practice, 2022, S. 1-18.

[Ba20] Bauckhage, C. et al.: Quantum Machine Learning. Eine Analyse zu Kompetenz, Forschung und Anwendung, 2020. https://www.bigdata-ai.fraunhofer.de/de/publikationen/quantum-ml.html. Abruf am 04.04.2022.

[BK21] Barton, T.; Kokoev, A.: Text Mining bei einer wissenschaftlichen Literaturauswertung: Extraktion von Schlüsselwörtern zur Beschreibung von Inhalten. In Barton, T.; Müller, C. (Hrsg.): Data Science anwenden (Angewandte Wirtschaftsinformatik). Springer Vieweg, S. 193–200, 2021.

[FPW99] Frank, E.; Paynter Gordan W.; Witten, I. H.: Domain-Specific Keyphrase Extraction. International Journal of Computer Applications, S. 668–673, 1999.

[FS07] Feldman, R.; Sanger, J.: The text mining handbook. Advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge, 2007.

[GL14] Goldberg, Y.; Levy, O.: word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method, 2014.

[Ho18] Homeister, M.: Quantum Computing verstehen. Springer Fachmedien Wiesbaden, Wiesbaden, 2018.

[HR06] Hippner, H.; Rentzmann, R.: Text Mining. Informatik-Spektrum 4/29, S. 287–290, 2006.

[Ku07] Kuo, B. Y.-L. et al.: Tag clouds for summarizing web search results. In (Williamson, C. et al. Hrsg.): Proceedings of the 16th international conference on World Wide Web - WWW '07. ACM Press, New York, New York, USA, S. 1203, 2007.

[Mi12] Miner, G. et al.: Practical text mining and statistical analysis for non-structured text data applications. Elsevier Academic Press, Waltham, Amsterdam, Boston, Heidelberg, London, New York, Oxford, Paris, San Diego, San Francisco, Singapore, Sydney, Tokyo, 2012.

[Mi13] Mikolov, T. et al.: Efficient Estimation of Word Representations in Vector Space, 2013.

[MRS08] Manning, C. D.; Raghavan, P.; Schütze, H.: Introduction to information retrieval. Cambridge University Press, New York, 2008.

[Pr18] Preskill, J.: Quantum Computing in the NISQ era and beyond. Quantum 2, S. 79, 2018.

[SH14] Schieber, A.; Hilbert, A.: Entwicklung eines generischen Vorgehensmodells für Text Mining. Technische Universität Dresden, Fakultät Wirtschaftswissenschaften, 2014.

[SS15] Siddiqi, S.; Sharan, A.: Keyword and Keyphrase Extraction Techniques: A Literature Review. International Journal of Computer Applications 2/109, S. 18–23, 2015.

[Ta20] Tauchert, C. et al.: Towards an Integrative Approach for Automated Literature Reviews Using Machine Learning. In (Bui, T. Hrsg.): Proceedings of the 53rd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences, 2020.

[vH08] van der Maaten, L.; Hinton, G. E.: Visualizing Data using t-SNE. Journal of Machine Learning Research 9, S. 2579–2605, 2008.

[vo09] vom Brocke, J. et al.: Reconstructing the giant: On the importance of rigour in documenting the literature search process: ECIS, 2009.

[We10] Cui, W. et al.: Context-preserving, dynamic word cloud visualization. In: 2010 IEEE Pacific Visualization Symposium (PacificVis). IEEE, 2010. S. 121-128

[We21] Weng, L.: Learning Word Embedding. https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html, Stand: 19.09.2021.