Bauhaus-Universität Weimar
Fakultät Medien
Studiengang Medieninformatik

Erfassung thematisch relevanter Argumentationseinheiten in Wikipedia

Bachelorarbeit

Aaron Solbach

1. Gutachter: Prof. Dr. Benno Stein
2. Gutachter: PD Dr. Günther Schatter

Datum der Abgabe: 7. September 2016

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Weimar, 7. September 2016

.....
Aaron Solbach

Zusammenfassung

Argumente sind wichtiger Bestandteil bei der Meinungsbildung und Entscheidungsfindung. Das sammeln von Argumenten ist aber zeitintensiv, sodass es lohnenswert wäre, diese Aufgabe an einen Computer abzugeben. Levy et al. [9] haben ein Programm zur Erfassung thematisch relevanter Argumentationseinheiten entwickelt, das darauf basiert, für einzelne Sätze Merkmale zu bestimmen, die auf Relevanz und Argumentativität schließen lassen. In dieser Arbeit soll darauf aufbauend das Erfassen relevanter Sätze durch die Verwendung von Textrepräsentationsmodellen aus dem Bereich Topic Modeling verbessert werden. Zusätzlich sollen die einzelnen Bestandteile des Programms evaluiert werden, um ihre Bedeutung für die Erfassung relevanter Themen zu bestimmen. Die Evaluation macht deutlich, dass der Ansatz, Sätze losgelöst vom Kontext zu betrachten, zur Erfassung thematisch relevanter Argumentationseinheiten ungeeignet ist.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Der Mensch nutzt Argumente zur Meinungsbildung, zur Entscheidungsfindung und zur Überzeugung anderer, doch woher nimmt er sie? Argumente werden einerseits selbst gebildet: Es wird eine Behauptung aufgestellt, begründet und bestenfalls durch Beispiele belegt. Andererseits werden sie in verbalen oder schriftlichen Argumentationen ausgetauscht. Der Austausch von Argumenten ist für die vorliegende Arbeit von besonderem Interesse. Er findet meist im jeweiligen sozialen Umfeld[1] statt, in welchem bereits bestimmte Anschauungen vorherrschen. Dies hat zur Folge, dass solche Argumente, die anderen Anschauungen entsprechen, hier keinen Raum finden. Bei der Meinungsbildung, der Entscheidungsfindung und auch zur Überzeugung anderer wäre die Verfügbarkeit von Argumenten aus unterschiedlichen Blickwinkeln wünschenswert. Jedoch erweisen sich sowohl das Ausmachen relevanter Quellen als auch die Erfassung der eigentlichen Argumente als zeitintensive Aufgaben.

Diese Arbeit befasst sich mit einem computergestützten Ansatz zur Erfassung thematisch relevanter Argumentationseinheiten. Um Argumente unterschiedlichster Anschauungen zu erfassen, kann das Internet genutzt werden. Hier steht eine riesige Menge an Dokumenten zur Verfügung, die häufig Argumente zu kontrovers diskutierten Themen enthalten. Solche Dokumente stammen unter anderem aus Enzyklopädien, Weblogs, Nachrichtenportalen oder Diskussionsforen. Das Ziel dieser Arbeit ist die Entwicklung eines Programmes, das aus den Dokumenten heraus Argumente extrahiert und den Nutzern bereitstellt. Die Nutzer würden davon auf zweierlei Weise profitieren: Zum einen sparen sie sich die Zeit, Argumente zu sammeln, zum anderen erhalten sie auch Informationen aus ihnen unbekannten Quellen und haben dadurch Zugriff auf ein repräsentativeres Meinungsbild. Das Programm würde es erlauben,

[1] virtuelle Räume wie Internetplattformen sind hierbei nicht ausgeschlossen

eine kontroverse Behauptung einzugeben, und in vier Schritten Argumente für und wider die Behauptung zurückliefern:

1. Zur Behauptung werden thematisch relevante Dokumente gesammelt
2. Argumentationseinheiten mit Bezug zur Behauptung werden erfasst
3. Die erfassten Argumentationseinheiten werden entsprechend ihrer Relevanz gewichtet
4. Die gesammelten Daten werden dem Nutzer übersichtlich präsentiert

Ähnlich zu diesem Ansatz ist eine Arbeit von Braunstain et al. [5]: Dort sollen auf Frageportalen wie Yahoo Answers[2] subjektive Antworten durch relevante Aussagen aus Wikipedia-Artikeln unterstützt werden. Im Gegensatz zum oben beschrieben Programm werden jedoch nur unterstützende und nicht zwingend argumentative Aussagen gesucht. Ausgangspunkt für die vorliegende Arbeit ist die Arbeit von Levy et al. [9] (siehe Kapitel 2). Diese beschäftigt sich mit dem zweiten Punkt des oben genannten Ablaufs: der Erfassung thematisch relevanter Argumentationseinheiten. Dieser Punkt ist für das oben beschriebene Programm von grundlegender Bedeutung, wird aber noch nicht zufriedenstellend gelöst. Der grundlegende Gedanke bei Levy et al. ist es, Merkmale sowohl für die Relevanz als auch für die Argumentativität eines Satzes zu definieren und zu bestimmen. Mittels dieser Merkmale soll für den Satz geschlussfolgert werden, ob er ein thematisch relevantes Argument enthält. In dieser Arbeit soll einerseits das Verfahren zur Erfassung thematisch relevanter Argumentationseinheiten optimiert werden. Andererseits sollen die einzelnen Bestandteile auf die Bedeutung für diese Aufgabe hin evaluiert werden.

Bei der Optimierung des Verfahrens stehen die Relevanzmerkmale im Fokus. Sie nutzen die Bedeutung von Satzteilen, um aus deren Worten die Relevanz für ein gegebenes Thema zu bestimmen. Verschiedene Detektoren zur Bestimmung solcher Relevanzmerkmale werden im Programm implementiert. Die verschiedenen Relevanzmerkmale können sowohl separat als auch kombiniert genutzt werden, um die Relevanz eines Satzes für ein Thema zu bestimmen. Zusätzlich kommen innerhalb der Relevanzdetektoren verschiedene Textrepräsentationsmodelle (siehe Kapitel 3) zum Einsatz. Die Art und Weise, wie Text repräsentiert wird, hat maßgeblich Einfluss darauf, ob ein Zusammenhang zwischen Textpassagen erkannt wird oder nicht. Durch die Verwendung von Repräsentationsmodellen aus dem Bereich Topic Modeling werden nicht nur Wörter verglichen, um Zusammenhänge zu erkennen, sondern die im Text

[2] http://answers.yahoo.com, letzter Zugriff 05.09.2016, 11:43 Uhr

enthaltenen Konzepte. Auf diese Weise wird ein Kontext zwischen Texten gefunden, auch wenn ein unterschiedliches Vokabular verwendet wird.

Das Programm, das für diese Arbeit implementiert wurde, besteht aus vier Modulen (siehe Kapitel 4). Zwei Module dienen der Erfassung von Relevanz und Argumentativität. Ein weiteres Modul kombiniert verschiedene Merkmale, um die enthalten Informationen zeitgleich zur Klassifizierung der Sätze zu nutzen. Das letzte Modul dient der Auswertung der Evaluierung des Systems.

Zur Evaluierung der einzelnen Bestandteile des Programms (siehe Kapitel 5) werden drei verschiedene Versuchsabläufe durchlaufen. Die Relevanzmerkmale werden sowohl unabhängig voneinander als auch in Kombination evaluiert. In beiden Fällen werden zuerst alle relevanten Sätze zu einem Thema erfasst und diese anschließend auf Argumentativität geprüft. Im ersten Fall sollen Eigenschaften der Relevanzmerkmale untersucht werdem, im zweiten Fall Eigenschaften der Repräsentationsmodelle. Bei dem dritten Versuchsablauf werden Relevanz- und Argumentmerkmale kombiniert, sodass in einem Schritt geprüft werden kann, ob ein Satz thematisch relevante Argumente enthält. Dabei soll evaluiert werden, ob die parallele Verarbeitung der Merkmale einen Vorteil gegenüber der sequenziellen Verarbeitung hat.

In der Evaluierung wird sichtbar, dass es im Programm noch einen deutlichen Verbesserungsbedarf gibt. Ein wesentliches Problem liegt in der aus dem Kontext gelösten Betrachtung der Sätze, die das Erfassen thematischer Relevanz stark erschwert. Ein anderes Problem betrifft den Umstand, dass Relevanz und Argumentativität im Programm unabhängig voneinander erfasst werden. Zur Evaluierung der jeweiligen Detektoren liegt aber nur ein Datensatz vor, in welchem Sätze annotiert wurden, die relevant *und* argumentativ sind. Dieser Datensatz erweist sich für die Evaluation und für das maschinelle Lernen als ungeeignet. Eine qualitative Analyse der Fehlerquellen soll zum Abschluss dieser Arbeit mögliche Schritte zur Beseitigung dieser Mängel aufzeigen.

Kapitel 2

Der Ansatz von Levy et al.

In diesem Kapitel wird ein Überblick über den Ansatz von Levy et al. [9] gegeben, der als Grundlage für die vorliegende Arbeit dient. Es werden die Parallelen und Unterschiede dieser Arbeit aufgezeigt und begründet.

Forscher von IBM Haifa veröffentlichten im Jahr 2014 zwei Arbeiten zur Erfassung thematisch relevanter Argumentationseinheiten. Einerseits präsentierten Levy et al. ihr eigenes Programm zu dieser Problematik, andererseits stellten Aharoni et al. [2] einen Korpus zur Verfügung, der dazu dient, dieses und gleichartige Programme zu evaluieren.

Das von Levy et al. präsentierte Programm nimmt als Eingabe ein Thema und ein dazu relevantes Dokument entgegen. Als Thema dient ein Satz, der eine kontroverse Aussage enthält. Aus dem Dokument sollen Argumente für und wider diese Aussage extrahiert werden. Im Programm gibt es dafür drei Komponenten, die nacheinander durchlaufen werden:

1. Die Satzkomponente separiert die Sätze im Dokument und prüft jeweils, ob thematisch relevante Argumentationseinheiten enthalten sind. Jeder Satz erhält dazu eine numerische Bewertung und nur die 200 besten Sätze werden an die nächste Komponente weitergereicht.

2. Die Grenzkomponente bestimmt die genauen Grenzen der Argumentationseinheiten innerhalb der 200 Sätze. Dazu wird ein Satz in mehrere sich überlappende Abschnitte aufgeteilt und für jeden Abschnitt bewertet, mit welcher Wahrscheinlichkeit er den Grenzen des Arguments entspricht. Der Abschnitt mit der besten Bewertung wird an die nächste Komponente weitergereicht.

3. Die Rangkomponente erhält für jeden Abschnitt aus der Grenzkomponente die Bewertungen der beiden vorherigen Komponenten. Aus ihnen wird eine Rangfolge unter den erkannten Argumentationseinheiten gebildet. Die besten 50 werden als Resultat ausgegeben.

Das Lösungsansatz von Levy et al. dient als Grundlage für die vorliegende Arbeit. Dabei steht die Satzkomponente im Fokus, weshalb hier darauf verzichtet wird, die genauen Grenzen der Argumente zu bestimmen. Die Satzkomponente soll in dieser Arbeit einerseits optimiert und andererseits ausführlicher evaluiert werden.

Levy et al. verwenden die Satzkomponente um nach Merkmalen von Sätzen zu suchen, die auf Relevanz oder Argumentativität schließen lassen. Einige der Relevanzmerkmale sollen auch in dieser Arbeit reimplementiert werden. Dagegen werden komplett unterschiedliche Merkmale zur Bestimmung der Argumentativität verwendet, was darin begründet ist, dass diese nicht ausreichend genau beschrieben oder zu komplex sind. Bei Levy et al. werden die Relevanzmerkmale eines Satzes in Form von repräsentativen Worten mit dem Themensatz verglichen. Je mehr übereinstimmende Worte gefunden werden, desto höher ist die Relevanz des Satzes. Hierbei kommt es zu Problemen, wenn unterschiedliche Wörter mit der gleichen Bedeutung verwendet werden oder das gleiche Wort unterschiedliche Bedeutungen hat. Als Maßnahme gegen diese Probleme werden in dieser Arbeit nicht die Worte selbst, sondern damit verbundene Konzepte verglichen. Dann gehören die Wörter ‚Rechner' und ‚Computer' dem gleichen Konzept an und zählen deshalb als übereinstimmend.

Zur Evaluierung nutzen Levy et al. den Korpus von Aharoni et al. Dieser beinhaltet sowohl Themen, als auch für jedes Thema relevante Dokumente. Die Themen wurden nach Zufallsprinzip dem Debattierportal „iDebate"[1] entnommen. Bei den Dokumenten handelt es sich um manuell zugeordnete und annotierte Wikipedia-Artikel. Die Annotationen zeigen auf, an welcher Stelle thematisch relevante Argumentationseinheiten enthalten sind. Die Evaluation des Programms fällt bei Levy et al. rudimentär aus. Es werden lediglich anhand der Annotationen Genauigkeit und Trefferquote der einzelnen Komponenten berechnet. Diese Werte werden mit Erwartungswerten bei Zufallsentscheidungen verglichen. Eine genauere Untersuchung der Merkmale selbst findet nicht statt. Auch wird nicht begründet, warum die verschiedenen Merkmale geeignet sein sollen, Aussagen über die Relevanz oder Argumentativität eines Satzes zu treffen.

[1] http://idebate.org, letzter Zugriff: 29.08.2016, 13:04 Uhr

Um Eigenschaften der einzelnen Merkmale untersuchen zu können, werden in dieser Arbeit Argumentativität und Relevanz der Sätze unabhängig voneinander bestimmt. Durch diese Trennung kann der Einfluss von Relevanz- und Argumentmerkmalen, wie auch der Einfluss des verwendeten Repräsentationsmodells[2] auf das gesamte System geprüft werden.

Im folgenden Kapitel wird im Detail darauf eingegangen, wie Text für Zwecke des Information Retrievals repräsentiert wird und warum die Verwendung spezieller Repräsentationsmodelle für das Programm von Vorteil sein sollte.

[2]Das Repräsentationsmodell bestimmt, wie die Ähnlichkeit von zwei Texten berechnet wird; etwa durch übereinstimmende Wörter oder Konzepte (siehe oben und Kapitel 3)

Kapitel 3

Repräsentation von Text für Information Retrieval

Aufgaben aus dem Bereich des Information Retrievals (im Deutschen auch Informationsrückgewinnung) erfordern eine Repräsentation von Textdokumenten, die einen einfachen Zugriff auf enthaltene Informationen möglich macht. Gespeichert werden Texte meist als Zeichenketten. In dieser Form ist es für ein Programm weder einfach auf enthaltene Informationen zuzugreifen, noch Gemeinsamkeiten oder Unterschiede zwischen Texten auszumachen.

Informationen in Texten liegen auf unterschiedlichen Ebenen. Einerseits trägt jedes Wort für sich zur Bedeutung eines Textes bei, anderseits kann der Kontext, in dem ein Wort steht, seine Bedeutung beeinflussen. Bei dem verbreitesten Repräsentationsmodell im Information Retrieval wird aus der Häufigkeit von Wörtern im Text geschlussfolgert, welche Informationen im Text enthalten sind. Dabei wird darauf verzichtet, einen Kontext zwischen den Wörtern zu berücksichtigen. Dagegen beschreibt Topic Modeling eine Klasse von Repräsentationsmodellen, die auch den Kontext eines Wortes innerhalb des Textes berücksichtigt. Topic Modeling Modelle sollen es ermöglichen, die Konzepte, die in einem Text enthalten sind, zugänglich zu machen.

In diesem Kapitel wird zunächst das traditionelle Repräsentationsmodell tf-idf erläutert. Trotz der weiten Verbreitung und der guten Leistung liegen hier auch Probleme vor, die im Anschluss besprochen werden. Danach wird die Idee des Topic Modelings, das diesen Problemen entgegenwirken soll, zusammen mit zwei Modellen aus diesem Bereich, der Latenten Dirichlet Allokation und der Expliziten Semantischen Analyse, beschrieben. Kapitel 3.3 untersucht dann den Einfluss dieser drei Repräsentationsmodelle auf die Erfassung thematisch relevanter Argumentationseinheiten.

3.1 Gewichtete Worthäufigkeiten im Vektorraum (das tf-idf Modell)

Bei der Verwendung gewichteter Worthäufigkeiten als Repräsentationsmodell wird angenommen, dass die einzelnen Wörter eines Textes unabhängig voneinander Informationen enthalten. Die Reihenfolge der Wörter ist dabei irrelevant: Wenn in zwei Texten die gleichen Wörter verwendet werden, wird davon ausgegangen, dass auch die gleichen Informationen enthalten sind. Die gewichteten Worthäufigkeiten, auf denen das Modell basiert, werden als tf-idf Maß (siehe unten) bezeichnet. Es ist in dieser Arbeit deshalb in Folge von dem tf-idf Modell die Rede.

3.1.1 Die Bedeutung eines Wortes für ein Dokument (das tf-idf Maß)

Die Bedeutung eines Wortes für ein Dokument ist nach dem tf-idf Maß von zwei Faktoren abhängig:

1. Dem Informationsgehalt des Wortes. Er wird daran gemessen, wie gut sich das Wort eignet, verschiedene Dokumente voneinander zu unterscheiden. Je seltener das Wort in Dokumenten enthalten ist, desto höher der Informationsgehalt. Als Maß dafür dient die Inverse Dokumenthäufigkeit (engl. ‚inverse document frequency').
2. Der Häufigkeit des Wortes im Dokument. Als Maß dafür dient die Suchwortdichte (engl. term frequency), die den relativen Anteil des Wortes am Text angibt.

Das Produkt aus beiden Faktoren wird aufgrund der englischen Bezeichnungen (term frequency und inverse document frequency) als tf-idf Maß bezeichnet. Es eignet sich unter anderem hervorragend dafür, Schlüsselwortkandidaten auszumachen. Das tf-idf Maß gibt direkt an, wie aussagekräftig ein Wort für das Dokument ist. Ein hoher Wert signalisiert, dass ein Wort für dieses Dokument typisch ist und nur in wenigen anderen Dokumenten auftritt. Ein entsprechendes Wort ist ein guter Kandidat für ein Schlüsselwort (vergleiche Kapitel 1 in Blei et al. [4]).

3.1.2 Repräsentation eines Textes im Vektorraum

Um für ein beliebiges Wort in einem Text schnell auf das tf-idf Maß zugreifen zu können, werden die Werte in Vektoren gespeichert. Jedes verschiedene Wort im Korpus bildet dabei eine Dimension des Vektors[1]. Jedes Dokument wird als ein solcher Vektor repräsentiert. Wie bedeutsam ein Wort für das Dokument ist, kann aus dem Vektor ausgelesen werden, indem der Wert für die dem Wort entsprechende Dimension abgefragt wird.

Ein großer Vorteil der Repräsentation des Textes im Vektorraum liegt darin, dass verschiedene Texte sehr einfach verglichen werden können (vergleiche Kapitel 1 in Blei et al. [4]). Zwei Dokumente sind einander ähnlich, wenn der Winkel zwischen den entsprechenden Vektoren klein ist. Als Ähnlichkeitsmaß kann die Kosinus-Ähnlichkeit berechnet werden: Der Kosinus des Winkels zwischen zwei Vektoren ergibt dann 1, wenn die Gewichtung aller Worte zweier Vektoren gleich ist. Wenn es kein Wort gibt, das sich zwei Dokumente teilen, ergibt der Kosinus den Wert 0. Die Kosinus-Ähnlichkeit kann unabhängig davon berechnet werden, was die Dimensionen repräsentieren (etwa verschiedene Wörter im Korpus oder verschiedene Konzepte, siehe Kapitel 3.2) oder wie lang die zu vergleichenden Dokumente sind. Die Anzahl der Dimensionen muss gleich sein, was dadurch gewährleistet ist, dass sie durch den Korpus bestimmt wird und nicht durch die einzelnen Dokumente. Die Dokumentenlänge ist irrelevant, da die repräsentativen Vektoren bei der Ähnlichkeitsberechnung normalisiert werden.

3.1.3 Synonymie und Polysemie als Fehlerquellen

Die tf-idf Modell ist ein einfaches, aber effektives Repräsentationsmodell. Dadurch, das alle Worte isoliert betrachtet werden, entsteht ein wesentliches Problem: Das Modell ist fehleranfällig bei Synonymie und Polysemie (vergleiche zweiten Abschnitt in Deerwester et al. [6]).

Synonymie beschreibt den Umstand, dass für die selbe Sache verschiedene Worte verwendet werden können. Relevante Informationen werden nicht gefunden, da in manchen Textabschnitten ein anderes Vokabular vorherrscht, als in der Suchanfrage. Beispielsweise kann in einem Text von ‚Rechnern' die Rede sein, im anderen aber von ‚Computern'. Die Ähnlichkeit der Texte wird nicht erkannt, wodurch die Trefferquote sinkt.

[1] Um die Anzahl von Dimensionen geringer zu halten, können Maßnahmen wie das Filtern von Stopwords (Wörter, die in jedem Dokument enthalten sind) oder das Stemmen (Rückführung von Wörtern auf den Wortstamm) ergriffen werden. Darauf wird in dieser Arbeit aber nicht weiter eingegangen.

Polysemie dagegen bezeichnet das Phänomen, dass Worte in verschiedenen Kontexten unterschiedliche Bedeutungen haben können. Zum Beispiel kann mit dem Wort ‚Bank' ein Sitzmöbel oder aber ein Geldinstitut gemeint sein. Dies hat zur Folge, dass bei Suchanfragen die Genauigkeit abnimmt, wenn polyseme Wörter auf einen Zusammenhang schließen lassen, der nicht besteht.

In beiden Fällen, beim Auftreten von Synonymie und Polysemie, kann der Kontext, in dem die Worte stehen, zur Klärung des Problems beitragen.

3.2 Topic Modeling

Repräsentationsmodelle aus dem Bereich des Topic Modeling versuchen die oben genannten Probleme der Synonymie und Polysemie zu lösen, indem sie Dokumente nicht auf Wortebene betrachten, sondern auf der Ebene von Konzepten[2]. Die Charakteristik eines Dokumentes wird nicht mehr auf gewichtete Worthäufigkeiten, sondern auf den Anteil von Konzepten am Text zurückgeführt.

Konzepte repräsentieren das Wissen über eine Sache (vergleiche Kapitel 1 in Gabrilovich et al. [7]). Wenn der Mensch einen Text ließt, verbindet er die Wörter des Textes mit verinnerlichten Konzepten. Auch bei Wörtern, die mit mehreren, unterschiedlichen Konzepten assoziiert werden (Polysemie), kann mit Hilfe des Kontextes die passende Verbindung hergestellt werden. Zum Beispiel könnte das Wort „Bank" zusammen mit dem Wort „Geld" einem Konzept „Finanzwesen" zugeordnet werden, im Kontext mit dem Wort „Baum" dagegen einem Konzept „Stadtpark". Auch die Verwendung verschiedener Wörter für die gleiche Sache (Synonymie) stellt beim Abgleich von Konzepten kein Problem dar, da die Wörter mit dem gleichen Konzept verbunden werden. Konzepte werden vereinfacht als Ansammlung von Worten zu einer Sache repräsentiert. Die Worte eines Konzeptes werden nach Häufigkeit gewichtet: Worte, die im Rahmen eines Konzeptes häufiger auftreten, sind bedeutsamer für das Konzept.

Topic Modeling Verfahren setzen Konzepte an die Stelle der gewichteten Worthäufigkeiten im tf-idf Modell. Es wird angenommen, dass Dokumente einander ähnlich sind, wenn sie die gleichen Konzepte enthalten. Dokumente werden als Vektoren mit den Konzepten als Dimensionen repräsentiert. Dadurch

[2]In der Literatur wird für gewöhnlich von Themen (engl. topic) gesprochen. In dieser Arbeit wird der Begriff ‚Konzept' verwendet, da die Begriffe ‚Thema' und ‚thematische Relevanz' sich hier auf den Inhalt einer Behauptung (siehe Kapitel 4.1) beziehen.

bleibt praktischerweise auch die Vergleichbarkeit von Texten mittels der Kosinusähnlichkeit erhalten. Beim Topic Modeling gibt es im Wesentlichen zwei Problemstellungen, die es für die unterschiedlichen Verfahren zu bewältigen gilt: (1) Sinnvolle Konzepte müssen modelliert werden. (2) Den Dokumenten müssen passende Konzepte zugewiesen werden. Diese Arbeit nutzt die Latente Dirichlet Allokation und die Explizite Semantische Analyse als Repräsentationsmodelle aus dem Bereich Topic Modeling. Im Folgenden sollen für beide Verfahren die Lösungsstrategien zu den aufgeführten Problemstellungen dargelegt werden.

3.2.1 Latente Dirichlet Allokation

Die Latente Dirichlet Allokation wurde im Jahr 2003 von Blei et al. [4] vorgestellt. Der Grundgedanke hier ist es, dass Dokumente aus einer Zusammensetzung verschiedener Konzepte bestehen. Die einzelnen Wörter im Korpus sind für verschiedenen Konzepte unterschiedlich relevant. Es wird nun angenommen, dass die Wörter einem Dokument entsprechend der Zusammensetzung der Konzepte zugewiesen wurden. Umgekehrt soll nun aus den Wörtern eines Dokuments bestimmt werden, welche Zusammensetzung von Konzepten diesem zugrunde liegt.

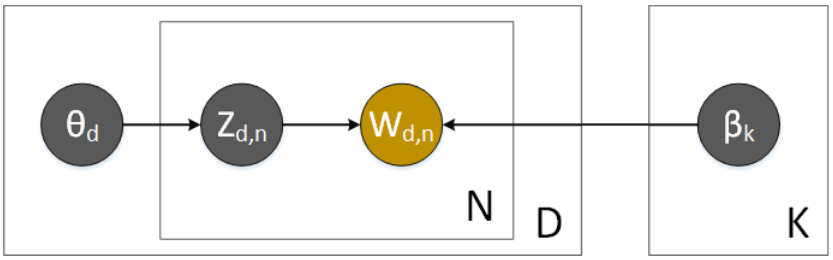

Abbildung 3.1: Der generative Prozess, der bei der Latente Dirichlet Allokation für die Entstehung der Dokumente angenommen wird. Die Kreise stellen Zufallsvariablen dar. Die Kästen geben an, dass es von den enthaltenen Variablen eine feste Menge (N, D und K) von Instanzen gibt (bspw. $\beta_1, \beta_2, ..., \beta_K$). Nur die Wörter $W_{d,n}$ können beobachtet werden; alle anderen Variablen sind verborgen.

Ein mathematisches Modell beschreibt den generativen Prozess, der für die Entstehung der Dokumente angenommen wird (vergleiche dazu Abbildung 3.1): Gegeben sind K Konzepte (rechter Kasten). Ein Konzept β_k wird modelliert als eine Wahrscheinlichkeitsverteilung über die verschiedenen Worte im Korpus. Es soll eine feste Anzahl an Dokumenten D generiert werden (linker Kasten). Jedem Dokument wird gemäß einer Poisson-Verteilung eine Länge

in Wörtern N zugewiesen (innerer Kasten). Einem Dokument d wird einer Dirichlet-Verteilung entsprechend eine Wahrscheinlichkeitsverteilung über die Konzepte θ_d zugewiesen. Über θ_d wird jedem Wort n im Dokument d ein Konzept zugewiesen. Die Zuweisung wird mit $Z_{d,n}$ notiert. $Z_{d,n}$ enthält für Wort $W_{d,n}$ in Dokument d eine Referenz auf ein Konzept β_k. Über dieses β_k wird nun aus der Menge der verfügbaren Wörter das Wort $W_{d,n}$ gewählt. Wenn alle Worte gewählt wurden, ist das Dokument vollständig generiert. Gleichzeitig ist bekannt, welche Konzepte zu welchem Anteil im Dokument enthalten sind.

An dieser Stelle sei noch einmal betont, dass der oben beschriebene Prozess so in der Realität nicht abläuft. Die Dokumente, mit denen gearbeitet wird, wurden nicht über Wahrscheinlichkeitsverteilungen generiert, sondern von Menschen verfasst. Die Konzepte sind bei der Latenten Dirichlet Allokation nicht vorgegeben, sondern sollen aus einem Datensatz heraus errechnet werden. Die Konzepte werden jedoch berechnet, indem die Latente Dirichlet Allokation die Startbedingungen für einen Prozess ermittelt, der die gegebenen Dokumente hervorbringen könnte. Dafür kann die Variable $W_{d,n}$ beobachtet werden, während alle anderen Variablen unbekannt sind.

Dieses mathematische Problem ist zu komplex, um eine exakte Lösung zu berechnen. Stattdessen wird versucht, eine möglichst gute Näherung zu erreichen. Um die Konzepte aus den vorliegenden Dokumenten zu extrahieren, wird zunächst allen Wörtern in den unterschiedlichen Dokumenten der Dirichlet-Verteilung entsprechend ein Konzept zugeordnet. Durch diese Zuordnung wird bereits die erste Annäherung an die Konzepte bestimmt, die ja Wahrscheinlichkeitsverteilungen über alle Wörter sind. Die erste Annäherung wird nun schrittweise verbessert, indem die Wörter in den Dokumenten wiederholt durchlaufen und neu einem Konzept zugeordnet werden. Für die Neuzuweisung wird angenommen, dass das aktuelle Wort die einzige Variable im Korpus ist. Die Neuzuordnung eines Wortes zu einem Konzept ist von zwei Faktoren abhängig: der Häufigkeit des Wortes im Konzept und der Häufigkeit von Worten des Konzeptes im Dokument. Das Produkt dieser beiden relativen Werte ergibt die Wahrscheinlichkeit dafür, dass das Wort zum Konzept gehört. Gemäß der Wahrscheinlichkeiten für alle Konzepte kann das Wort abschließend zugeordnet werden. Das wiederholte Neuzuordnen aller Wörter sorgt dafür,

- dass Wörter in Konzepten, in denen sie vermehrt vertreten sind, noch stärker vertreten sein werden
- dass Konzepte in Dokumenten, in denen sie vermehrt vertreten sind, noch stärker vertreten sein werden

Andersherum verschwinden Worte aus Konzepten, in denen sie vermindert

vertreten sind (das heißt: die Häufigkeit der Worte im Konzept geht gegen null). Analog verhält es sich für die Konzepte in den Dokumenten. In Folge dessen nimmt die interne Validität des Modells zu. Für das Verfahren kann ein prozentualer Schwellwert s als Endkriterium angegeben werden: Wenn in einem Durchlauf weniger als $s\%$ der Worte neu zugeordnet wurden, endet der Prozess.

Die auf diese Art errechneten Konzepte werden als latent bezeichnet. Es ist oftmals zwar möglich, jedoch nicht vorgesehen, aufgrund der vorrangig vertretenen Wörter einen Titel für das jeweilige Konzept zu finden. Stattdessen sollen Konzepte die hintergründige Struktur abbilden, nämlich Gruppen von Wörtern, die häufig gemeinsam auftreten. Es wird impliziert, dass diese Gruppen eine semantische Einheit bilden.

Beliebige Texte, von ganzen Dokumenten über einzelne Paragraphen bis zu kurzen Suchabfragen, können mittels Latenter Dirichlet Allokation repräsentiert werden, indem die Wörter des Textes wie oben beschrieben Konzepten zugeordnet werden. Wenn die Annäherung an die Konzepte bereits abgeschlossen ist, verändern sich die Konzepte bei der Zuordnung zu weiteren Texten nicht. Der Anteil eines Konzeptes an dem Text entspricht dem Anteil der Worte an dem Text, die dem Konzept zugewiesen werden.

3.2.2 Explizite Semantische Analyse

Bei der Expliziten Semantischen Analyse von Gabrilovich et al. [7] werden die Konzepte nicht errechnet, sondern explizit vorgegeben. Es wird nicht versucht, Zusammenhänge zwischen Worten im Dokument zu entdecken, sondern zwischen den Dokumenten und den vorliegenden Konzepten. Dieses Vorgehen soll stärker der menschlichen Denkweise entsprechen als die Latente Dirichlet Allokation: Der Mensch versucht nicht, Konzepte aus Texten heraus abzuleiten, sondern Inhalte mit vorhandenem Wissen zu verknüpfen.

Die explizite Vorgabe von Konzepten erfolgt durch einen Indizierungsdatensatz. Jedes Dokument im Korpus stellt ein Konzept dar. Als Indizierungsdaten kann prinzipiell eine beliebige Sammlung von Dokumenten dienen. Als besonders wertvoll haben sich aber Enzyklopädie-Artikel herausgestellt. In Enzyklopädien wird versucht, Wissen in einer ähnlichen Art zu speichern, wie es im Gehirn passiert: als vernetzte Konzepte. Ein Artikel behandelt in der Regel einen Gegenstand ausführlich und kann somit als ein Konzept für die Explizite Semantische Analyse dienen.

Konzepte sind auch bei der Expliziten Semantischen Analyse Wahrschein-

lichkeitsverteilungen über alle verschiedenen Worte. Es genügt deshalb die Dokumente im Indizierungsdatensatz mittels tf-idf Modell zu repräsentieren (siehe Kapitel 3.1), um aus ihnen Konzepte zu erzeugen: Der Anteil eines Wortes am Konzept entspricht der Wahrscheinlichkeit, dass ein zufällig aus dem Konzept gewähltes Wort eben dieses Wort ist. Zusätzlich zu den Konzepten wird ein inverser Index gebildet, der für jedes Wort im Korpus angibt, in welchen Konzepten es enthalten ist. Der inverse Index ermöglicht es, dass nur der Anteil jener Konzepte am Dokument berechnet wird, die auch tatsächlich im Dokument enthalten sind.

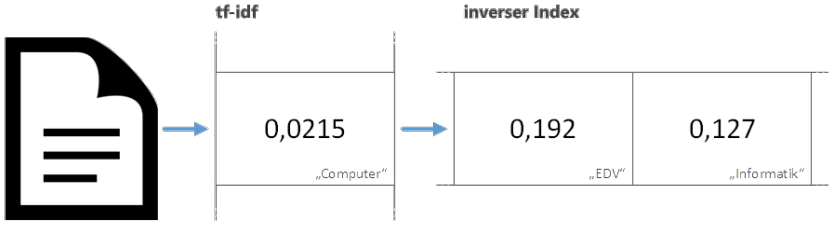

Abbildung 3.2: Mittels tf-idf wird die Relevanz eines Wortes für ein Dokument bestimmt. Durch den inversen Index ist die Relevanz dieses Wortes für die Konzepte bekannt. Die Relevanz eines Konzeptes für das Dokument steigt um das Produkt der Relevanz des Wortes für das Dokument und der Relevanz des Wortes das Konzept.

Um zu bestimmen, welche Konzepte in einem Dokument enthalten sind, wird zunächst das Dokument mittels tf-idf Modell repräsentiert. Dadurch wird für jedes Wort im Dokument ein Wert berechnet, der die Relevanz des Wortes für das Dokument wiedergibt. In dem Beispiel in Abbildung 3.2 liegt das tf-idf Maß für das Wort ‚Computer' bei $0,0215$. Für jedes Wort wird über den inversen Index geprüft, in welchen Konzepten es vorkommt. Im Beispiel kommt das Wort ‚Computer' unter anderem in den Konzepten ‚EDV' und ‚Informatik' vor. Die Relevanz des Wortes für die Konzepte wurde ebenfalls mittels tf-idf berechnet. Sie kann aus dem inversen Index ausgelesen werden und beträgt $0,192$ für das Konzept ‚EDV' sowie $0,127$ für das Konzept ‚Informatik'. Der Anteil der Konzepte an dem Dokument erhöht sich um das Produkt aus der Relevanz des Wortes für den Text und der Relevanz des Wortes für das Konzept. Für jedes Wort erhöht sich also der Anteil der Konzepte, die das Wort enthalten. Im Beispiel erhöht sich durch das Wort ‚Computer' der Anteil der Konzepte ‚EDV' und ‚Informatik' um $0.0215 \cdot 0.192 = 0.0041$ und $0.0215 \cdot 0.127 = 0.0027$. Natürlich kann der Anteil dieser Konzepte am ganzen Dokument größer sein, wenn das Dokument auch andere Worte des Konzepts enthält.

3.3 Bedeutung für die Arbeit

Die Probleme, welche durch Synonymie und Polysemie verursacht werden, beeinträchtigen sowohl die Trefferquote als auch die Genauigkeit bei Suchanfragen. Durch die Verwendung von Repräsentationsmodellen aus dem Bereich Topic Modeling soll diesen Problemen entgegengewirkt werden. Es wird angenommen, dass insbesondere die Trefferquote verbessert werden kann, da die Synonymie das deutlich häufiger auftretende Phänomen darstellt.

Die beiden vorgestellten Topic Modeling Modelle haben ihrerseits Vor- und Nachteile. Die aufwändige Extraktion von Konzepten in der Latenten Dirichlet Allokation mindert die Performanz des Verfahrens. Dafür ist es weniger stark abhängig von der Qualität der Indizierungsdaten[3], weil nicht explizite Konzepte, sondern unterliegende Strukturen genutzt werden.

Die Explizite Semantische Analyse ist gegenüber der Latenten Dirichlet Allokation deutlich schneller, wenn ein geeigneter Korpus aus Enzyklopädie-Artikeln vorliegt. Zudem kommt das Verfahren der Idee näher, Texte auf die Art zu verarbeiten, wie es der Mensch macht: nicht durch statistische Zusammenhänge, sondern durch menschlich organisiertes Wissen. Wenn der Trainingsdatensatz nicht sorgfältig gewählt wird, kann es andererseits schnell dazu kommen, dass die Trennschärfe der Konzepte stark abnimmt. Dies passiert, wenn viele Worte in einem Konzept auftreten, die eigentlich nicht mit diesem verbunden werden. Auch können Konzepte fehlen, wenn die Indizierungsdaten nicht alle Themen aufweisen, die in den Testdokumenten enthalten sind.

In der vorliegenden Arbeit werden beide Topic Modeling Modelle in Kombination mit verschiedenen Indizierungsdatensätzen verwendet. Auch das tf-idf Modell ist in die Arbeit eingebunden und dient als Referenz für die Topic Modeling Modelle. Alle drei Repräsentationsmodelle bilden die Dokumente im Vektorraum ab, wodurch sie im Programm beliebig ausgetauscht werden können. Die Einbettung der Repräsentationsmodelle in das System zur Erfassung von Argumentationseinheiten wird im folgenden Kapitel dargelegt. In diesem Rahmen werden auch die Indizierungsdaten beschrieben, die im Kontext dieser Arbeit ausgewählt worden sind.

[3]Genau genommen ist kein zusätzlicher Datensatz notwendig; er kann aber von Nutzen sein, wenn die Qualität oder der Umfang der relevanten Dokumente nicht ausreichend sind.

Kapitel 4

Technische Herangehensweise

In dieser Arbeit wird ein System zur Erfassung thematisch relevanter Argumentationseinheiten entwickelt, dass an die Arbeit von Levy et al. [9] (siehe Kapitel 2) fortführt. Das System soll die folgende Aufgabe lösen. Gegeben ist ein Thema und eine Menge von zu diesem Thema relevanten Dokumenten. Aus den Dokumenten sollen diejenigen Sätze zurückgegeben werden, die Argumentationseinheiten zu dem Thema enthalten. Um das System zu evaluieren wird als Datengrundlage der gleiche Korpus wie bei Levy et al. genutzt, jedoch in einer erweiterten Version [2] (siehe Kapitel 4.1).

Das Prinzip hinter dem System ist es, für jeden Satz Merkmale zu finden, aus denen sich rückschließen lässt, ob er relevant und argumentativ ist. Einige der in dieser Arbeit verwendeten Relevanzmerkmale wurden von Levy et al. übernommen. Da die Argumentmerkmale nicht reproduzierbar beschrieben wurden, werden sie in dieser Arbeit durch ein anderes Verfahren ersetzt.

Das System zur Erfassung thematisch relevanter Argumentationseinheiten besteht aus vier Modulen, die unterschiedliche Teilaufgaben ausführen:

- Ein Relevanzdetektor (siehe Kapitel 4.2) extrahiert Relevanz-Merkmale aus den Sätzen, indem der Satz auf unterschiedliche Arten mit dem Themensatz verglichen wird
- Ein Argumentdetektor (siehe Kapitel 4.3) extrahiert mit Hilfe von Wörterbüchern Argument-Merkmale aus den Sätzen
- Ein Merkmalsverknüpfer (siehe Kapitel 4.4) gruppiert Merkmale und lernt, welche Kombinationen von Merkmalsausprägungen relevante oder argumentative Sätze kennzeichnen

- Ein Evaluator (siehe Kapitel 5) bewertet die von den anderen Modulen weitergereichten Sätze anhand der Annotation des zugrundeliegenden Datensatzes

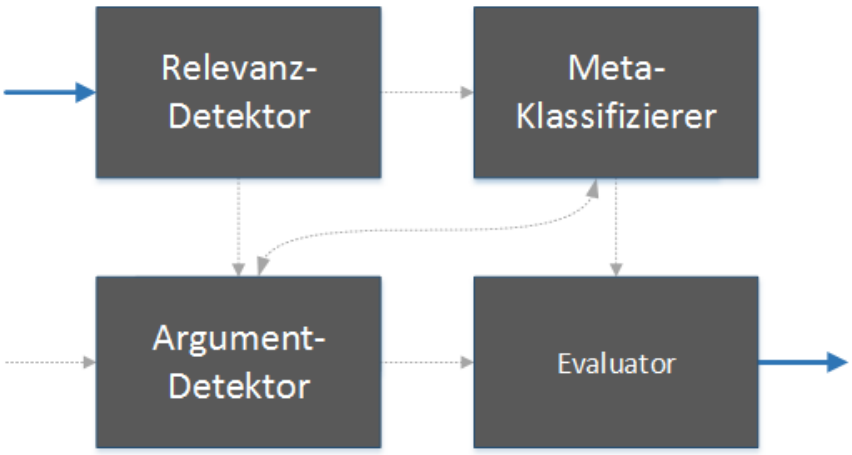

Abbildung 4.1: Das System besteht aus vier Modulen, die in unterschiedlicher Reihenfolge durchlaufen werden können.

Die einzelnen Module können in unterschiedlichen Programmabläufen durchlaufen werden (vergleiche die grauen Pfeile in Abbildung 4.1). Je nach Reihenfolge werden etwa erst irrelevante Sätze ausgefiltert, um anschließend zu prüfen, ob die relevanten Sätze auch argumentativ sind. Oder es werden Relevanz- und Argument-Merkmale zugleich gesammelt, sodass der Metaklassifizierer bestimmen muss, welche Sätze sowohl argumentativ als auch relevant sind. In jedem Fall wertet der Evaluator am Ende Trefferquote und Genauigkeit des Verfahrens aus. Welcher Programmablauf zu welchem Zweck durchlaufen wird, wird in Kapitel 4.5 beschrieben.

4.1 Korpus

Um das System zur Erfassung thematisch relevanter Argumentationseinheiten zu evaluieren, wird ein Datensatz von Aharoni et al. [2] herangezogen, der zu eben diesem Zweck bereitgestellt wurde. Eine ältere Version dieses Datensatzes mit geringerem Umfang wird auch bei Levy et al [9] verwendet.

Der Korpus enthält in der jetzigen Version 58 verschiedene Themen. Ein Thema wird in Form einer Forderung („This house would ban gambling“[1])

[1]frei übersetzt: „Diese Gruppierung würde Glücksspiel verbieten“

oder Überzeugung („This house believes atheism is the only way“[2]) in einem Satz formuliert. Die Sätze beginnen alle mit ähnlichen Phrasen (This house believes/thinks/...), die zum Ausdruck bringen, dass hier die Meinung einer Gruppe vertreten wird. Die Aussagen wurden zufällig der Debattierplattform http://idebate.org entnommen. Zu jedem Thema wurden manuell mindestens ein und maximal 25 relevante Wikipedia-Artikel gefunden. Insgesamt wurden 533 Artikel den Themen zugewiesen und 2778 thematisch relevante Argumentationseinheiten annotiert. Etwa jeder dreißigste Satz eines Artikels enthält Argumentationseinheiten.

Um die Qualität der Annotationen zu sichern haben Aharoni et al. etwa 20 Beurteiler sorgfältig instruiert und Ergebnisse kontrolliert. Die Beurteiler annotierten überlappende Textabschnitte, sodass mittels Cohens Kappa die Urteilsübereinstimmung evaluiert werden konnte. Der resultierende Wert von 0,4 wird von Aharoni et al. als hinreichend gut eingestuft, da die Klassifizierung der Argumentationseinheiten von Feinheiten abhängt und deshalb zu einem gewissen Grad der Subjektivität unterliegt. Durch die sich überlappenden Textabschnitte konnte auch die Trefferquote der einzelnen Beurteiler und darüber die Trefferquote des gesamten Korpus abgeschätzt werden: Es wird angenommen, dass etwa 90% der thematisch relevanten Argumentationseinheiten erfasst werden konnten.

4.2 Erfassung thematischer Relevanz

Für jedes Relevanzmerkmal gibt es einen Relevanzdetektor, der dazu dient, die Merkmalsausprägung für einen Satz zu ermitteln. Die Detektoren arbeiten alle nach dem gleichen Prinzip (vergleiche Abbildung 4.2): Für jedes Thema werden einem Detektor der Themensatz und eine Menge von möglicherweise relevanten Sätzen übergeben. Sowohl der Themensatz, als auch die anderen Sätze durchlaufen eine Vorverarbeitung. Dem Themensatz wird dabei die einleitende Phrase (siehe Kapitel 4.1) entfernt. Die anderen Sätze werden je nach Detektor auf sprachliche Merkmale reduziert (zum Beispiel auf die vorhandenen Nomen)[3]. Der Themensatz und auch die anderen Sätze werden nach dieser Vorverarbeitung im Vektorraummodell repräsentiert. Dann wird die Kosinus-Ähnlichkeit der Sätze zum Themensatz berechnet. Der resultierende Wert stellt die Merkmalsausprägung des jeweiligen Relevanz-Merkmales dar.

[2] frei übersetzt: „Diese Gruppierung glaubt, dass Atheismus der einzige Weg ist“

[3] In der Regel sind weder der Themensatz noch die anderen Sätze nach der Vorverarbeitung grammatisch korrekte Sätze. Das stellt aber kein Problem dar, weil es nur von Interesse ist, welche Wörter enthalten sind.

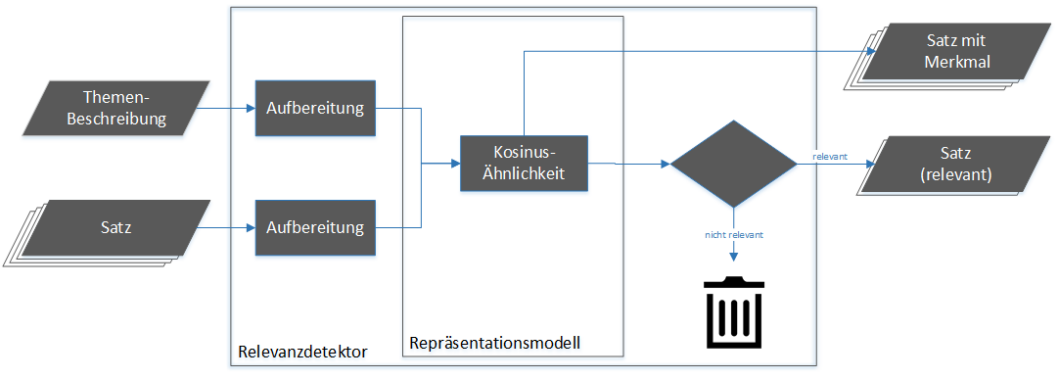

Abbildung 4.2: Im Relevanzdetektor werden Sätze mit Themensätzen verglichen, um festzustellen, ob sie thematisch relevant sind. Wenn die Relevanz direkt durch den Detektor bestimmt werden soll, werden nur relevante Sätze zurückgegeben. Dient das Merkmal als Information im Metaklassifizierer, so müssen alle Sätze zurückgegeben werden.

Die schwarzen Rahmen in Abbildung 4.2 stellen Schablonen für jeden Relevanzdetektor und analog für jedes Repräsentationsmodell dar. Ein Pfeil durch eine solche Schablone bedeutet, dass alle Instanzen (also alle Detektoren oder Repräsentationsmodelle) durchlaufen werden. Durch die Verschachtelung von Schablonen entstehen Kombinationen: In jedem Relevanzdetektor wird jedes Repräsentationsmodell durchlaufen. Folglich erhält jeder Satz für jede Kombination aus Relevanzmerkmal und Repräsentationsmodell eine Merkmalsausprägung.

Für jedes Relevanzmerkmal können (je nach Verwendungszweck im Programm) entweder alle Sätze mit ihrer jeweiligen Merkmalsausprägung zurückgegeben werden[4] oder nur die Sätze, die als relevant eingestuft werden. Im letzteren Fall muss definiert werden, wann ein Satz relevant ist. Ein Schwellwert stellt die Grenze zwischen relevanten und irrelevanten Sätzen dar. Für jedes Thema kann er unterschiedlich sein. Der optimale Schwellwert kann nur berechnet werden, wenn für jeden Satz bekannt ist, ob er relevant ist oder nicht. Dies ist in der Praxis nicht der Fall, weshalb der Schwellwert eines Themas unabhängig von den vorliegenden Annotationen für das Thema ermittelt werden muss. Dies geschieht durch eine sogenannte Kreuzvalidierung: Es wird für alle anderen Themen ein idealer Schwellwert ermittelt und dieser für das aktuelle Thema verwendet.

Im Folgenden werden die vier Relevanzmerkmale und die Idee für deren

[4]Alle Sätze werden benötigt, wenn der Metaklassifizierer in einem Schritt bestimmen soll, ob ein Satz relevant und argumentativ ist (vergleiche Kapitel 4.5).

Arbeitsweisen erläutert, sowie die Repräsentationsmodelle mit den jeweiligen Indizierungsdaten vorgestellt.

4.2.1 Relevanzmerkmale

Die Relevanz-Merkmale entsprechen Funktionen der Satzkomponente von Levy et al. [9] (vergleiche Kapitel 2). Die Funktionen der Satzkomponente werden dort in kontextbezogen und kontextfrei untergliedert. Die kontextfreien Funktionen der Satzkomponente dienen für Levy et al. zur Unterscheidung der argumentativen Sätze von nicht argumentativen. Diese Funktionen können in dieser Arbeit nicht reproduziert werden, da sie nicht hinreichend beschrieben wurden. Über die kontextbezogenen Funktionen soll bestimmt werden, ob ein Zusammenhang mit dem jeweiligen Thema vorliegt und damit der Satz für dieses Thema inhaltlich relevant ist oder nicht. Von den kontextbezogenen Funktionen werden drei als Relevanzmerkmale übernommen.

Der Zusammenhang zwischen einem Satz und einem Thema (das als Aussage formuliert wird, vergleiche Kapitel 4.1) wird über linguistische Merkmale bestimmt. Die grundlegende Idee dahinter ist, dass verschiedene Satzteile unterschiedlich bedeutsam für den Kontext eines Satzes sind. Beispielsweise ist das Subjekt von herausragender Bedeutung: Es ist die Sache, um die es in einem Satz geht. Der Satz beschreibt eine Eigenschaft oder Handlung des Subjektes. Objekte, insbesondere Nomen, bilden dagegen den Kontext zum Subjekt: Sie beschreiben, wozu das Subjekt in Relation steht oder womit es interagiert. Verben können zwar auch kontextuelle Informationen enthalten, beschreiben aber meist die Relation und nicht den Kontext. Sie eignen sich deshalb weniger, um den Kontext eines Satzes zu bestimmen.

Ein kurzes Beispiel soll die hier getroffenen Aussagen verdeutlichen. Vereinfacht werden nicht Objekte, sondern stellvertretend die Nomen betrachtet:

„Der Angeklagte verneint, zur Tatzeit im Kasino gewesen zu sein.“

Aus dem Subjekt (‚der Angeklagte‘) und dem ersten Objekt (‚Tatzeit‘) geht sehr deutlich der Kontext (Gericht, Rechtsprechung) dieses Satzes hervor. Das zweite Nomen (‚Kasino‘) dagegen ist fehlleitend. Das Verb liefert in diesem Beispiel keine Informationen zum Kontext, weil es in jedem Kontext verwendet werden könnte.

Im Folgenden werden die einzelnen Relevanzmerkmale beschrieben. Dabei handelt es sich einerseits um die Subjektähnlichkeit und zwei Varianten der

Nomenähnlichkeit, die in ähnlicher Form auch bei Levy et al. verwendet werden, und anderseits um die Satzähnlichkeit, die als Referenz für die anderen Merkmale hinzugefügt wurde.

Die **Subjektähnlichkeit** nutzt die Bedeutung des Subjektes in Bezug auf den Kontext des Satzes: Wenn das Subjekt relevant für ein Thema ist, kann davon ausgegangen werden, dass der ganze Satz relevant ist. Bei der Subjektähnlichkeit wird deshalb jeder Satz mittels des Computerlinguistik-Frameworks Stanford CoreNLP [10] auf seine Subjekte reduziert.

Es gibt zwei Varianten der **Nomenähnlichkeit**. Bei beiden wird ein Satz mittels Stanford CoreNLP [10] auf seine Nomen reduziert. Diese Nomen sind auf Satzebene immer auch Teil von Objekten und können auch das Subjekt enthalten. In der ersten Variante (erweiterte Nomenähnlichkeit), werden für die Nomen mittels der lexikalischen Datenbank Wordnet [1] zusätzlich Synonyme gesucht, um dem Problem der Synonymie (vergleiche Kapitel 3.1 und letzten Absatz) zu begegnen. Weil auch die Topic-Modeling Modelle dazu dienen, den Einfluss von Synonymie zu schmälern, wird in der zweiten Variante (einfache Nomenähnlichkeit) auf das Hinzufügen von Synonymen verzichtet.

Bei der **Satzähnlichkeit** werden beide Sätze ohne weiterer Vorverarbeitung im Vektorraummodell repräsentiert und verglichen. Es gibt keine Gewichtung von Wortarten oder Satzteilen, wodurch vor allem das jeweilige Repräsentationsmodell Einfluss auf das Ähnlichkeitsmaß hat. Die Satzähnlichkeit entspricht einer ‚naiven' Herangehensweise und dient deshalb als Referenz für die anderen Relevanzmerkmale.

4.2.2 Repräsentationsmodelle

In dieser Arbeit werden alle in Kapitel 3 beschriebenen Repräsentationsmodelle verwendet: Das traditionelle Modell mit TF-IDF (siehe Kapitel 3.1.1), die Latente Dirichlet Allokation (LDA, siehe Kapitel 3.2.1) und die Explizite Semantische Analyse (ESA, siehe Kapitel 3.2.2). Alle Modelle bilden Texte im Vektorraum ab, wodurch sie im System ohne weitere Anpassung ausgetauscht werden können.

Die Verwendung von Repräsentationsmodellen aus dem Bereich Topic Modeling soll die Qualität der Relevanzmerkmale steigern. Topic Modeling Modelle wirken dem Synonymieproblem entgegen, bei dem ein Zusammenhang zwischen zwei Texten nicht erfasst wird, weil unterschiedliche Worte für die gleiche Sache verwendet werden. Die Worte verweisen auf die gleichen Konzepte, die nun beim Topic Modelling verglichen werden, wodurch der Anteil

der relevanten Sätze, die auch als solche klassifiziert werden, steigen sollte. Im Programm sollen Sätze erfasst werden, die gleichzeitig relevant und argumentativ sind, weshalb die Steigerung der Trefferquote im Relevanzdetektor nicht auf das gesamte Programm übertragen werden kann. Es ist aber davon auszugehen, dass ein Teil der zusätzlich als relevant klassifizierten Sätze auch argumentativ ist. Es kann angenommen werden, dass dieser relative Anteil ebenso groß ist, wie in den Sätzen, deren Relevanz auch im tf-idf Modell erfasst wurde.

Die Repräsentationsmodelle aus dem Bereich Topic Modeling müssen vor der Anwendung mittels Indizierungsdaten trainiert werden. Zu diesem Zweck wurden drei verschiedene Datensätze zusammengestellt: Die Wikipedia-Artikel aus dem Korpus von Aharoni et al. [2], Einführungstexte zu kontroverse Themen von der Debattierplattform http://idebate.org und Wikipedia-Artikel zu Ereignissen aus den Jahren 2014/15. Aus den Indizierungsdaten werden die Konzepte generiert, die zum Vergleich von Texten herangezogen werden (vergleiche Kapitel 3.2). Eine sorgfältige Auswahl ist deshalb wichtig. Für LDA bedeutet dies, dass die Indizierungsdaten sich sprachlich nicht stark[5] von den Testdokumenten unterscheiden sollten, in denen nach thematisch relevanten Argumentationseinheiten gesucht wird. Die Themen in den Indizierungsdaten sollten zudem breit gefächert sein, damit für die Testdokumenten passende Konzepte zur Verfügung stehen. Noch empfindlicher für die Qualität der Indizierungsdaten ist ESA: Da jedes Dokument hier ein Konzept abbildet, sollte ein Dokument möglichst nur eine Sache thematisieren. Je schwächer hier die Trennschärfe zwischen den Konzepten ist, desto geringer wird später die Aussagekraft eines Vergleichs zweier Texte. Für die Aufgabenstellung ist es wichtig, dass die erzeugten Konzepte Gegenstände kontroverser Themen aufweisen. Im Folgenden werden die einzelnen Indizierungsdatensätze beschrieben und dargelegt, warum sie für die Erfassung thematisch relevanter Argumentationseinheiten geeignet sind.

Die **Wikipedia-Artikel** aus dem Korpus von Aharoni et al. erfüllen auf triviale Weise alle oben genannten Kriterien. Sie unterscheiden sich sprachlich nicht von den Testdokumenten und decken alle relevanten Themen ab (letzteres ist wegen der manuellen Zusammenfügung der Dokumente zu den Themen durch Aharoni et al. gegeben). Es handelt sich um Einträge aus einer Enzyklopädie, die für gewöhnlich einen Sache thematisieren. Dennoch scheint es insbesondere für ESA ungünstig, die Artikel, in denen nach thematisch rele-

[5]Ab wann Texte sprachlich zu stark auseinandergehen, ist schwer auszumachen. Extreme sprachliche Differenzen zur Alltagssprache finden sich beispielsweise in Gesetzestexten oder Jahrhunderte alten, literarischen Werken. Die Alltagssprache selbst wiederum unterscheidet sich stark von der Sprache, die in Wikipedia-Artikeln verwendet wird.

vanten Argumentationseinheiten gesucht werden soll, als Indizierungsdaten zu verwenden. Bei der Suche nach relevanten Sätzen werden dann die Sätze eines Artikels, der auch ein Konzept darstellt, auf enthaltene Konzepte geprüft. Mit anderen Worten: für die einzelnen Sätze wird geprüft, ob zwischen ihnen und dem Text, dem sie entstammen, ein Zusammenhang besteht. Intuitiv wird man annehmen, dass ein starker Zusammenhang zwischen dem Text und dem Satz bestehen muss. Dies ist wegen der Gewichtung der Worte im Konzept nicht zwingend der Fall. Aus dem gleichen Grund können auch andere Konzepte in dem einzelnen Satz stärker vertreten sein, als das Konzept, dem er entspringt.

Die **iDebate Einführungstexte** stammen aus dem Webis-Debate-16 Korpus von Al-Khatib et al. [3]. Für diesen Korpus wurden insgesamt 445 Einführungen zu kontroversen Themen aus der Debattier-Website ‚iDebate' ausgelesen. Die Themen wurden in 14 Kategorien untergliedert und decken damit ein breites Spektrum ab. Auch die Themen des Korpus von Aharoni et al. [2] stammen von iDebate und bilden eine Teilmenge der Themen, zu denen die Einführungstexte vorliegen. Sie werden also durch diesen Indizierungsdatensatz abgedeckt. Auch behandelt jeder Text genau ein Thema, wodurch die einzelnen Texte als ESA-Konzepte dienen können.

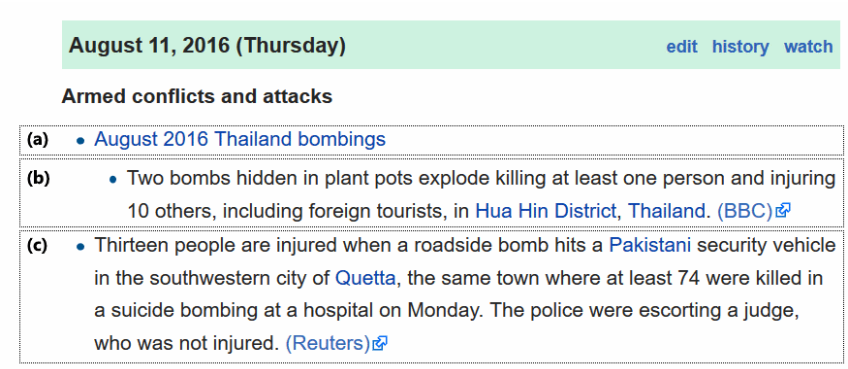

Abbildung 4.3: An jedem Tag werden besondere Ereignisse auf Wikipedia festgehalten. Es werden hier drei Ereignistypen unterschieden: (a) Andauernde Ereignisse, (b) Unterereignisse und (c) Einmalige Ereignisse.

Als weiterer Indizierungsdatensatz dienen **Wikipedia-Ereignisse**. Der Gedanke dahinter ist die Annahme, dass Ereignisse und Nachrichten Grundlage für kontroverse Diskussionen sind. Auf dem Wikipedia-Portal ‚Current Events'[6] sind für jeden Tag bedeutende Ereignisse aufgelistet (siehe Abbildung 4.3). Die aufgeführten Ereignisse können unabhängig von ihren Katego-

[6]http://en.wikipedia.org/wiki/portal:current_events, letzter Zugriff: 04.09.2016, 17:06 Uhr

rien in drei Typen unterteilt werden: (a) Andauernde Ereignisse, (b) Unterereignisse und (c) einmalige Ereignisse. Einmalige Ereignisse sind an einem Tag abgeschlossen. Andauernde Ereignisse dagegen sind dadurch gekennzeichnet, dass sie über mehrere Tage hinweg in Form von Unterereignissen aktualisiert werden. Für sie gibt es einen Wikipedia-Artikel, der das komplette Geschehen zusammenfasst, während einmalige und auch Unterereignisse oft auf keinen Wikipedia-Artikel mit direktem Bezug zum Ereignis verlinken. Aus diesem Grund werden nur die Wikipedia-Artikel zu andauernden Ereignissen ausgelesen und als Indizierungsdaten verwendet. Erneut handelt es sich um Einträge aus einer Enzyklopädie, die sich auch gut als ESA-Konzepte verwenden lassen.

Die Wikipedia-Ereignisse werden auch bei Mishra et al. [11] im wissenschaftlichen Kontext verwendet. Dort dienen die Wikipedia-Artikel über andauernde Ereignisse als eine Übersicht über diese Ereignisse, die in detaillierten Zeitungsberichten verloren geht. Auch hier wird deutlich, dass die Wikipedia-Ereignisse einen guten Überblick über die Nachrichten bieten und damit die Themen abdecken sollten, über die kontrovers diskutiert wird.

Tabelle 4.1: Charakteristik der verschiedenen Indizierungsdatensätze

Indizierungsdatensatz	Anzahl Dokumente	Wörter pro Dokument	Wörter gesamt
Wikipedia-Artikel	543	3.631	1.971.670
iDebate Einführungstexte	561	4.006	2.407.942
Wikipedia-Ereignisse	601	1.634	917.134

In Tabelle 4.1 werden Maße aufgelistet, die den Umfang der verschiedenen Indizierungsdatensätze charakterisieren. Alle Datensätze haben eine ähnliche Anzahl an Dokumenten, wodurch für die Explizite Semantische Analyse auch eine ähnliche Anzahl an Konzepten zur Verfügung steht. Auffällig ist, dass die Dokumente und dadurch auch die Gesamtanzahl der Wörter unter den Wikipedia-Ereignissen deutlich kleiner sind. Dies könnte sich für die Latente Dirichlet Allokation als Problem herausstellen, da weniger Text genutzt werden kann, um aus statistischen Zusammenhängen von Wörtern Konzepte zu modellieren (vergleiche Kapitel 3.2.1).

4.3 Erfassung von Argumentationseinheiten

Zur Erfassung von Argumentationseinheiten in Sätzen werden Wörterbücher verwendet, die aus der Arbeit von Somasundaran et al. [13] stammen und dieser Arbeit dankenswerter Weise zur Verfügung gestellt wurden. Die 20 Wörterbücher enthalten insgesamt 293 für Argumente typische Ausdrücke und Phrasen.

Es gibt verschiedene Konstrukte wie Bedingungen, Begründungen oder Verweise, die oft in Argumenten auftreten. Jedes Wörterbuch entspricht einem solchen Konstrukt und enthält entsprechende Ausdrücke. Manche Ausdrücke sind einfach gehalten wie zum Beispiel im Wörterbuch für Verweise ‚according to' (dt. ‚gemäß', ‚laut'). Andere sind dagegen komplexere reguläre Ausdrücke: `,(cannot|will not|won\'t|can\'t) ([\w]+[ ]+){1,7}(if|unless)`'.

Mit Hilfe der Wörterbücher wird die Merkmalsbestimmung durchgeführt. Anders als bei den Relevanzmerkmalen liegen die Merkmalsausprägungen nicht zwischen null und eins, sondern können gemäß der Boolschen Algebra nur zwei Werte annehmen: null oder eins (analog auch ‚falsch' oder ‚wahr'). Für jede Kombination aus Satz und Wörterbuch wird geprüft, ob mindestens ein regulärer Ausdruck enthalten ist. Wenn dies der Fall ist, ist die Ausprägung eins, ansonsten null.

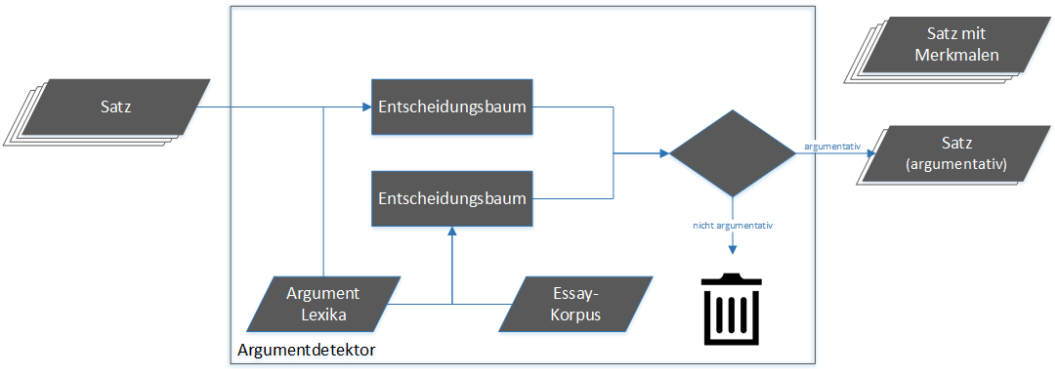

Abbildung 4.4: Im Argumentdetektor werden Sätze mittels Wörterbücher und einem Entscheidungsbaum auf argumentative Strukturen geprüft. Wenn die Argumentativität direkt durch den Detektor bestimmt werden soll, werden nur argumentative Sätze zurückgegeben. Dient das Merkmal als Information im Metaklassifizierer, so müssen alle Sätze zurückgegeben werden.

Das WEKA-Framework für maschinelles Lernen [8] wird verwendet, um aus den verschiedenen Argumentmerkmalen die Argumentativität eines Satzes zu bestimmen. Für das Lernverfahren werden Beispielsätze benötigt, für die annotiert ist, ob sie Argumente enthalten oder nicht. Die Annotationen von Aharoni et al. [2] sind dafür nicht geeignet, da hier nur die Sätze gekennzeichnet wurden die Argumente enthalten *und* relevant für ein Thema sind. Stattdessen wird ein Korpus mit kurzen Essays von Stab et al. [14] verwendet. Mithilfe der Beispielsätze wird ein C4.5 Entscheidungsbaum [12] aufgestellt[7], welcher

[7]In WEKA wurde der C4.5 Entscheidungsbaum als Klassifizierer mit der Bezeichnung „J48" implementiert.

anhand der Merkmalsausprägungen entscheidet, ob ein Satz argumentativ ist oder nicht.

4.4 Verknüpfung von Merkmalen

Sowohl der Argumendetektor, als auch die unterschiedlichen Relevanzdetektoren können eine direkte Beurteilung geben, ob ein Satz argumentativ beziehungsweise relevant ist. Prinzipiell würde es reichen beide Kriterien für die Sätze nacheinander zu überprüfen: Zuerst werden alle Sätze bestimmt, die thematisch relevant sind, und anschließend werden aus diesen Sätzen jene ausgewählt, die zusätzlich Argumente enthalten.

Dass aber die Verwendung der einzelnen Relevanz- oder Argumentmerkmale unabhängig voneinander nicht sinnvoll ist, wird schnell deutlich. Bei dem Argumentdetektor ist es gar nicht erst vorgesehen, die Merkmale separiert zu betrachten. Die einzelnen Konstrukte wie Bedingungen oder Verweise (vergleiche Kapitel 4.3) können nicht für sich als Indiz für Argumente verwendet werden. Wenn ein Satz aber sowohl eine Bedingung als auch einen Verweis enthält, scheint es wahrscheinlicher, dass er ein Argument enthält. Auch die unterschiedlichen Relevanzmerkmale könnten kombiniert bessere Ergebnisse erzielen. Beispielsweise kann es bei der Subjektähnlichkeit dazu kommen, dass bei relevanten Sätze Pronomen vorliegen und deshalb das Thema des Satzes gar nicht erkannt wird. Als Folge nimmt die Trefferquote ab. Bei der Nomenähnlichkeit dagegen (und noch stärker bei der Erweiterten Nomenähnlichkeit) können Wörter enthalten sein, die mit dem Thema nichts zu tun haben und deshalb irreführend sind. Dadurch nimmt die Genauigkeit ab. Eine Kombination aus beiden Merkmalen könnte die einzelnen Nachteile überwinden.

Um unterschiedliche Merkmale zu verknüpfen, müssen diese dem Metaklassifizierer hinzugefügt werden. Für jedes Merkmal werden alle Paare aus Satz und Merkmalsausprägung übergeben. Der Metaklassifizierer sammelt alle Merkmale zu einem Satz und erhält aus dem Korpus von Aharoni et al. [2] die Information, ob dieser Satz thematisch relevante Argumentationseinheiten enthält. Es soll wie im Argumentdetektor (vergleiche Kapitel 4.3) ein C4.5 Entscheidungsbaum [12] aufgebaut werden[8]. Dieser benötigt Trainingsdaten, die wie im Relevanzdetektor (vergleiche Kapitel 4.2) über eine Kreuzvalidierung gewonnen werden: für jedes Thema dienen die Sätze der anderen Themen als Trainingsdaten. Über den Entscheidungsbaum wird je nachdem,

[8]Auch hier wird das WEKA Framework [8] genutzt, dass den Entscheidungsbaum im Klassifizierer „J48“ implementiert.

welche Merkmale übergeben wurden, entschieden, ob ein Satz relevant ist oder ob er thematisch relevante Argumente enthält. Welche Merkmale übergeben werden ist abhängig vom jeweiligen Programmablauf. Die verschiedenen Programmabläufe werden im folgenden Abschnitt erläutert.

4.5 Programmabläufe

Wie in der Einleitung dieses Kapitels und auch in Kapitel 4.4 beschrieben, gibt es verschiedene Möglichkeiten, wie die einzelnen Module des Systems durchlaufen werden können. Die verschiedenen Programmabläufe, die in dieser Arbeit verwendet werden, und der Zweck ihrer Verwendung sollen im Folgenden erläutert werden.

In der Theorie können Relevanzmerkmale verwendet werden, um alle relevanten Sätze zu erfassen. Es würden dann alle Sätze gefunden werden, die thematisch relevante Argumentationseinheiten enthalten, und zusätzlich weiter Sätze, die thematisch relevant sind, aber nicht argumentativ. Das Resultat wäre eine sehr hohe Trefferquote, aber eine geringere Genauigkeit. Um die nicht argumentativen Sätze herauszufiltern würden Argumentmerkmale verwendet werden. Dadurch würde im Idealfall die Genauigkeit ansteigen und die Trefferquote nicht abnehmen.

In der Praxis sind die Merkmale und die Methoden, sie zu verknüpfen und auszuwerten, nicht perfekt. Es werden durch die Relevanzmerkmale weder alle relevanten Sätze gefunden, noch sind alle gefundenen Sätze relevant. Analog verhält es sich für die Argumentmerkmale: nicht alle argumentativen Sätze werden gefunden, noch sind alle gefundenen Sätze argumentativ. Ein Ziel dieser Arbeit ist es, die Genauigkeit und die Trefferquote des Relevanzdetektors zu steigern. Um die einzelnen Relevanzmerkmale zu untersuchen und das System zu evaluieren, werden die verschiedenen Module des Systems in unterschiedlicher Reihenfolge verknüpft (siehe Abbildung 4.5).

Der Programmablauf in Abbildung 4.5 (a) verzichtet darauf, die einzelnen Relevanzmerkmale im Metaklassifizierer zu verknüpfen. Für jedes Merkmal werden die Sätze, die als relevant bestimmt wurden, an den Argumentdetektor weitergereicht. Es wird angenommen, dass letzterer keinen wesentlichen Einfluss auf die Trefferquote hat[9], aber die Genauigkeit verbessert. Dadurch,

[9]Da durch den Argumentdetektor keine neuen Sätze hinzugefügt werden können, ändert sich die Trefferquote nur dann, wenn wahre Treffer des Relevanzdetektors fälschlicherweise aussortiert werden. Die Annahme ist also, dass der Argumentdetektor eine geringe Falsch-Negativ-Rate aufweist.

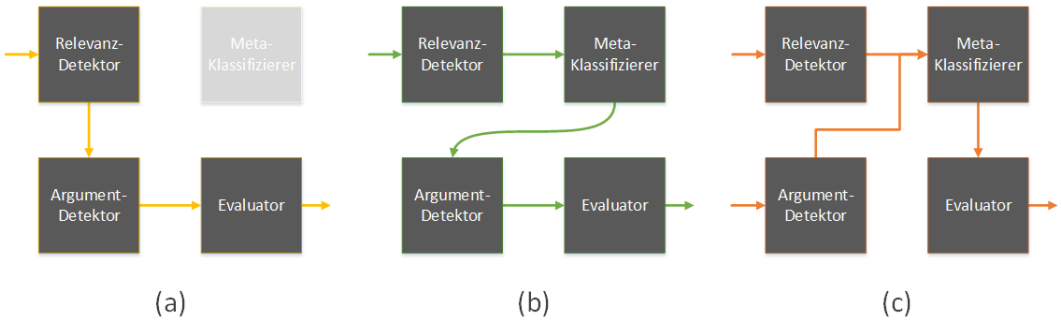

Abbildung 4.5: Es gibt drei unterschiedliche Programmschleifen, die dazu dienen, verschiedene Aspekte des Systems zu evaluieren.

dass die Relevanzmerkmale separiert bleiben, können ihre Eigenschaften in Bezug auf die Trefferquote und die Genauigkeit[10] besser verglichen werden. Insbesondere soll so der Einfluss der Repräsentationsmodelle auf die einzelnen Relevanzmerkmale sichtbar werden.

In Abbildung 4.5 (b) wird ein Programmablauf dargestellt, bei dem für jedes Repräsentationsmodell die einzelnen Relevanzmerkmale im Metaklassifizierer kombiniert werden. Für jedes Modell gibt es dann ein Ergebnis, dass an den Argumentdetektor weitergereicht und letztlich evaluiert wird. Auf diese Weise soll überprüft werden, wie gut die Relevanzmerkmale zusammenarbeiten. Es können die Ergebnisse der kombinierten Merkmale einerseits untereinander, andererseits mit den Ergebnissen der separaten Merkmale verglichen werden. Es zeigt sich also, welchen Vorteil die Metaklassifizierung bringt und welches Repräsentationsmodell die besten Ergebnisse liefert.

Der in Abbildung 4.5 (c) visualisierte Programmablauf sorgt für eine Kombination von Relevanz- und Argumentmerkmalen. Es werden die Ergebnisse aller Detektoren gesammelt und im Metaklassifizierer ausgewertet. Es wird also nicht mehr wie zuvor Relevanz und Argumentativität unabhängig voneinander bewertet. Dies entspricht auch dem Vorgehen von Levy et al. [9]. Es soll dabei sichtbar werden, ob die Kombination der Merkmale im Metaklassifizierer einen Vorteil gegenüber der separaten Klassifizierung von Relevanz und Argumentativität bringt.

[10]Für Relevanzmerkmale wird eine geringe Genauigkeit angenommen, da relevante *und* argumentative Sätze gesucht werden. In Bezug auf die Genauigkeit wird hier deshalb nur untersucht, ob es bedeutende Unterschiede in der Falsch-Positiv-Rate gibt.

Kapitel 5

Evaluierung

Die Evaluierung des Systems zur Erfassung thematisch relevanter Argumentationseinheiten erfolgt aus unterschiedlichen Perspektiven. Es wird versucht Unterschiede auszumachen zwischen:

1. Den einzelnen Relevanzmerkmalen
2. Den verschiedenen Repräsentationsmodellen
3. Der sequentiellen und gleichzeitigen Berücksichtigung von Relevanz- und Argumentmerkmalen

Aus der Aufzählung wird deutlich, dass die Relevanzmerkmale und die Repräsentationsmodelle (die nur innerhalb der Relevanzdetektoren Anwendung finden) im Fokus dieser Arbeit stehen. Es soll gezeigt werden, ob mithilfe geeigneter Modelle die Trefferquote und/oder die Genauigkeit des Systems gesteigert werden kann (vergleiche Kapitel 4.2.2). Der Argumentdetektor dient hier vorrangig dazu, die Ergebnisse zu verfeinern, indem relevante, aber nicht argumentative Sätze aussortiert werden.

Als Maße zur Evaluierung des Systems dienen vorrangig Genauigkeit und Trefferquote. Für einen Satz, den das System als relevant und argumentativ klassifiziert, wird das Ergebnis ‚positiv' genannt. Die Genauigkeit bezeichnet den Anteil der positiven Ergebnisse, die auch bei der Annotation des Korpus' als positiv eingestuft wurden (auch ‚richtig-positive' Ergebnisse genannt). Sie gibt an, wie viele der gefunden Sätze richtig sind. Die Trefferquote bezeichnet den Anteil der richtig-positiven Ergebnisse an den bei der Annotation als positiv eingestuften Sätze. Sie gibt an, wie viele der gesuchten Sätze gefunden wurden. Ein Problem mit den beiden Maßen im Sinne der Evaluierung

ist, dass eine höhere Trefferquote oftmals recht leicht, jedoch auf Kosten einer niedrigeren Genauigkeit erzielt werden kann. Beide Maße werden deshalb kombiniert, um in solchen Situationen zu bestimmen, welche Ansätze besser sind. Als Kombination der Maße wird das F-Maß[1] verwendet. Es bildet das harmonische Mittel aus Genauigkeit und Trefferquote.

Für die Untersuchung der oben aufgezählten Aspekte des Systems dient jeweils einer der Programmabläufe aus Kapitel 4.5. Im Folgenden werden zuerst für jeden der Punkte Hypothesen aufgestellt, Ergebnisse und Beobachtungen beschrieben und abschließend deren Bedeutung interpretiert.

5.1 Die einzelnen Relevanzmerkmale

Zur Untersuchung der einzelnen Relevanzmerkmale dient der erste Programmablauf aus Kapitel 4.5 (vergleiche Abbildung 4.5 (a) auf Seite 28). Im Programm werden für jedes Relevanzmerkmal nur die Sätze ausgegeben, die gemäß des jeweiligen Detektors relevant sind. Die Sätze werden anschließend durch den Argumentdetektor gefiltert. Nur argumentative Sätze werden an den Evaluator weitergeleitet.

5.1.1 Hypothesen

$H_{1.1}$: Die Subjektähnlichkeit hat eine höhere Genauigkeit als die anderen Merkmale. Durch die herausragende Bedeutung des Subjektes für den Kontext des Satzes (vergleiche Kapitel 4.2.1) wird erwartet, dass die Anzahl der falsch-positiven Ergebnisse bezüglich der Relevanz minimal ausfällt. In Hinblick auf das gesamte Verfahren bedeutet dies, dass sowohl weniger Sätze ausgegeben werden können, die nicht relevant, aber argumentativ sind (richtig-positive Ergebnisse des Argumentdetektors), als auch weniger Sätze die weder relevant noch argumentativ sind (falsch-positive Ergebnisse des Argumentdetektors). Infolge dessen sollte die Genauigkeit im Vergleich zu anderen Relevanzmerkmalen steigen.

$H_{1.2}$: Beide Varianten der Nomenähnlichkeit haben eine höhere Trefferquote als andere Merkmale. Für Nomen wird angenommen, dass sie gut dafür geeignet sind, den Kontext eines Satzes zu bestimmen (vergleiche Kapitel 4.2.1). Es kann zwar wegen fehlleitender Nomen zu falsch-positiven Ergebnissen bezüg-

[1] In dieser Arbeit wird nur das ausgeglichene F-Maß, auch F_1-Maß genannt, verwendet, bei dem Genauigkeit und Trefferquote gleichgewichtet sind.

lich der Relevanz kommen, dafür wird die Relevanz im Gegensatz zur Subjektähnlichkeit auch dann festgestellt, wenn das Subjekt etwa ein Pronomen oder ein Name ist, der nicht mit dem Thema in Verbindung gebracht wird.

$H_{1.3}$: Die Trefferquote der erweiterten Nomenähnlichkeit ist bei der Verwendung des tf-idf Modells besser als die der einfachen. Bei Verwendung von Topic Modeling Modellen weichen die Quoten nur geringfügig voneinander ab. Sowohl die Suche von Synonymen als auch die Verwendung von Topic Modeling Modellen dienen dazu, dem Synonymie-Problem (vergleiche Kapitel 3.1.3) entgegenzuwirken. Beide Maßnahmen sollen also die Trefferquote verbessern. Bei dem tf-idf Modell (also auch im Verfahren von Levy et al. [9]) ist es deshalb sinnvoll, die Nomen durch Synonyme zu ergänzen. Bei der Verwendung von Topic Modeling werden diese nicht gebraucht, da sie in den Konzepten enthalten sind. Es wird deshalb angenommen, dass die Synonyme dann keinen signifikanten Unterschied ausmachen.

5.1.2 Beobachtungen

Die Eigenschaften der Relevanzmerkmale stehen im Fokus dieser Arbeit. Um auszuschließen, dass die Filterung des Argumentdetektors diese Eigenschaften verschleiert, wurden zusätzlich die Ergebnisse der Relevanzdetektoren ohne Klassifizierung nach Argumentativität evaluiert. Im Vergleich mit den Rohdaten aus den Relevanzdetektoren fällt auf, dass der Argumentdetektor die Ergebnisse des Systems verschlechtert. In diesem Kapitel werden deshalb die Ergebnisse der Relevanzdetektoren ohne Filterung auf Argumentativität präsentiert[2]. Eine Fehleranalyse des Argumentdetektors wird in Kapitel 5.4 durchgeführt.

Bei einem Blick auf Tabelle 5.1 fällt direkt auf, dass die Werte für das F-Maß überall niedrig sind. Überraschender Weise ist der Mittelwert für die Verfahren unter tf-idf der einzige, der im zweistelligen Bereich liegt. Ohne Kombination der Relevanzmerkmale werden also unter diesem Modell die besten Ergebnisse erzeugt. Der generell niedrige Wert des F-Maßes wird insbesondere durch eine sehr geringe Genauigkeit verursacht. Der Durchschnitt liegt für alle Zusammensetzungen von Relevanzmerkmalen und Repräsentationsmodellen bei $5,4\%$, während der Wert für die Trefferquote analog immerhin $20,9\%$ beträgt. Die Satzähnlichkeit, die dem naiven Vergleich des Satzes mit dem Themensatz entspricht, erweist sich gemessen am durchschnittlichen F-Maß von $7,9\%$ als zweitbestes Relevanzmerkmal.

[2]Die Werte mit Filterung auf Argumentativität finden sich im Anhang A.

Tabelle 5.1: Evaluierung der Relevanzmerkmale in Abhängigkeit vom jeweiligen Repräsentationsmodell (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Model [Indizierungsdaten]	Relevanzmerkmal	G	T	F
tf-idf	- Satzähnlichkeit	7,9	33,8	12,8
	- Subjektähnlichkeit	12,7	18,0	14,9
	- Nomenähnlichkeit	8,7	18,5	11,8
	- Erweiterte Nomenähnlichkeit	9,8	19,0	13,0
	Mittelwert	9,8	22,3	13,1
LDA [WikiArtikel]	- Satzähnlichkeit	3,5	42,5	6,4
	- Subjektähnlichkeit	5,6	16,2	8,4
	- Nomenähnlichkeit	3,7	37,7	6,8
	- Erweiterte Nomenähnlichkeit	4,0	17,1	6,5
	Mittelwert	4,2	28,4	7,0
LDA [iDebate]	- Satzähnlichkeit	5,3	17,5	8,2
	- Subjektähnlichkeit	5,7	30,2	9,6
	- Nomenähnlichkeit	4,3	19,7	7,0
	- Erweiterte Nomenähnlichkeit	4,3	27,6	7,4
	Mittelwert	4,9	23,7	8,0
LDA [WikiEvents]	- Satzähnlichkeit	3,8	12,8	5,8
	- Subjektähnlichkeit	4,8	14,7	7,3
	- Nomenähnlichkeit	3,0	47,9	5,6
	- Erweiterte Nomenähnlichkeit	3,7	17,6	6,1
	Mittelwert	3,8	23,2	6,2
ESA [WikiArtikel]	- Satzähnlichkeit	5,6	26,9	9,2
	- Subjektähnlichkeit	9,0	14,5	11,1
	- Nomenähnlichkeit	4,6	13,0	6,8
	- Erweiterte Nomenähnlichkeit	4,2	6,9	5,2
	Mittelwert	5,8	15,3	8,1
ESA [iDebate]	- Satzähnlichkeit	4,9	16,7	7,5
	- Subjektähnlichkeit	11,0	16,3	13,1
	- Nomenähnlichkeit	4,6	20,4	7,5
	- Erweiterte Nomenähnlichkeit	4,5	8,2	5,8
	Mittelwert	6,2	15,4	8,5
ESA [WikiEvents]	- Satzähnlichkeit	3,7	9,8	5,4
	- Subjektähnlichkeit	3,3	15,1	5,4
	- Nomenähnlichkeit	2,6	16,2	4,5
	- Erweiterte Nomenähnlichkeit	3,2	30,4	5,8
	Mittelwert	3,2	17,9	5,3

In Bezug auf Hypothese $H_{1.1}$ lässt sich feststellen, dass die Subjektähnlichkeit mit einer Ausnahme überall die höchste Genauigkeit aufweist. Sie wird nur bei ESA [WikiEvents] von der erweiterten Nomenähnlichkeit übertroffen. Die Genauigkeit für die Subjektähnlichkeit beträgt im Mittel $7,4\%$, während sie bei den anderen Merkmalen zwischen $4,5\%$ und $4,9\%$ liegt (siehe Tabelle 5.2).

Tabelle 5.2: Mittelwerte der einzelnen Relevanzmerkmale über alle Repräsentationsmodelle (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Relevanzmerkmal	G	T	F
Satzähnlichkeit	4,9	22,8	7,9
Subjektähnlichkeit	7,4	17,9	10,0
Nomenähnlichkeit	4,5	24,7	7,1
Erweiterte Nomenähnlichkeit	4,8	18,1	7,1
alle Merkmale	5,4	20,9	8,0

Weniger deutlich fällt das Ergebnis für Hypothese $H_{1.2}$ aus: Zwar liegt die Trefferquote für die einfache Variante der Nomenähnlichkeit im Mittel vor den anderen Relevanzmerkmalen, jedoch wird die erweiterte Variante von der Satzähnlichkeit übertroffen und liegt auf einem Niveau mit der Subjektähnlichkeit (vergleiche Tabelle 5.2). Es gibt keine klare Tendenz, dass ein Relevanzmerkmal unabhängig von den Repräsentationsmodellen eine bessere Trefferquote erreicht. Bei LDA [iDebate] liegt sogar jene der Subjektähnlichkeit vorne (vergleiche Tabelle 5.1), bei der zugunsten einer hohen Genauigkeit eine geringere Trefferquote erwartet wurde.

Die Hypothese $H_{1.3}$ kann durch die Daten nicht bestätigt werden: Die Trefferquote für die erweiterte Nomenähnlichkeit ist nur um $0,5\%$ besser als jene für die einfache Nomenähnlichkeit. Während die Trefferquote im tf-idf Modell für beide Varianten sehr ähnlich ist, unterscheidet sie sich für alle anderen Modelle: In nur zwei Fällen (LDA [iDebate] und ESA [WikiEvents]) übertrifft die erweiterte Nomenähnlichkeit die einfache deutlich, in allen anderen Fällen ist es umgekehrt.

5.1.3 Analyse

Nur eine der drei Hypothesen wird durch die Daten bestätigt: Hypothese $H_{1.1}$. Die Genauigkeit der Subjektähnlichkeit ist den anderen Relevanzmerkmalen überlegen. Hypothese $H_{1.2}$ kann nicht belegt werden. Viel mehr muss angenommen werden, dass die Trefferquote weniger stark von dem Relevanzmerkmal abhängig ist, als von dem Zusammenspiel mit dem Repräsentationsmo-

dell. Selbst für das gleiche Topic Modeling Modell mit unterschiedlichen Indizierungsdaten unterscheiden sich die Trefferquoten eines Relevanzmerkmales stark. Weil die Trefferquote stärker von dem Repräsentationsmodell abhängig ist, als vom Relevanzmerkmal, unterscheiden sich die Merkmale überwiegend durch die Genauigkeit. Die Subjektähnlichkeit hat deshalb trotz niedrigster Trefferquote das höchste F-Maß vorzuweisen und damit die beste Performanz.

Erstaunlich ist, dass Hypothese $H_{1.3}$ durch die Daten nicht klar bestärkt wird. Zwar ist die Trefferquote besser als die der anderen Verfahren, doch handelt es sich dabei um einen geringfügigen Unterschied. Auch wurde durch Hypothese $H_{1.2}$ implizit angenommen, dass der Abstand der Trefferquote zwischen Nomen- und Subjektähnlichkeit größer ist. Ein Grund für die unerwartet kleinen Unterschiede könnte die generell geringe Performanz der Relevanzmerkmale sein. Es gibt zwei mögliche Ursachen für geringe Performanz: erstens der Umstand, dass im Korpus von Aharoni et al. [2] Relevanz und Argumentativität nicht unabhängig voneinander annotierten wurden, und zweitens die aus dem Kontext gelöste Betrachtung einzelner Sätze. Diese Ursachen werden detaillierter in Kapitel 5.4 besprochen.

5.2 Die verschiedenen Repräsentationsmodelle

Zur Untersuchung der verschiedenen Repräsentationsmodelle dient der zweite Programmablauf aus Kapitel 4.5 (vergleiche Abbildung 4.5 (b) auf Seite 28). Die Ergebnisse der einzelnen Relevanzdetektoren werden für jedes Repräsentationsmodell im Metaklassifizierer zusammengefügt, um durch maschinelles Lernen die relevanten Sätze auszugeben. Die relevanten Sätze wiederum werden wie in Programmablauf (a) im Argumentdetektor gefiltert, sodass möglichst die Sätze erhalten bleiben und evaluiert werden, die relevant und zusätzlich argumentativ sind.

5.2.1 Hypothesen

$H_{2.1}$*: Die Trefferquoten der einzelnen Relevanzmerkmale sind für das tf-idf Modell am niedrigsten.* Andersherum formuliert bedeutet dies, dass durch die Verwendung von Topic Modeling Modellen die Trefferquote bei allen Relevanzmerkmalen steigt. Diese Hypothese stützt sich wie bereits $H_{1.3}$ auf die Eigenschaft von Topic Modeling Modellen, dem Synonymieproblem entgegenzuwirken: Wenn alle Synonyme für ein Wort im zugehörigen Konzept enthalten sind, gibt es keine aufgrund unterschiedlichen Vokabulars falsch-negativen

Ergebnisse mehr (vergleiche Kapitel 3.2). Dementsprechend steigt die Trefferquote, wenn geeignete Konzepte vorliegen.

$H_{2.2}$: Das F-Maß der kombinierten Relevanzmerkmale ist höher als das Maximum der separaten Relevanzmerkmale. Es wird davon ausgegangen, dass der Metaklassifizierer besser funktioniert, je mehr Merkmale des Satzes ihm bekannt sind. Zum Beispiel können sich Subjektähnlichkeit und Nomenähnlichkeit ergänzen, wenn etwa im Subjekt ein nichtssagendes Pronomen steht, in den Nomen aber kontextrelevante Informationen enthalten sind. Andersherum kann das Subjekt Klarheit schaffen, wenn verschiedene Nomen den Kontext nicht eindeutig bestimmen. Wenn Subjekt und Nomen zusammen keine klare Einschätzung ermöglichen, ist die Satzähnlichkeit hilfreich, in die auch Verben und Adjektive einfließen. Die Relevanzmerkmale sollten sich also eignen, um kombiniert sowohl die Genauigkeit, als auch die Trefferquote zu erhöhen.

5.2.2 Beobachtungen

Das bereits aus Kapitel 5.1.2 bekannte Problem des Argumentdetektor bleibt auch für die kombinierten Relevanzmerkmale bestehen: Auch in diesem Kapitel werden deshalb die Ergebnisse der Relevanzdetektoren ohne Filterung nach Argumentativität präsentiert[3].

Zum Prüfen von Hypothese $H_{2.1}$ wird ein weiteres Mal Tabelle 5.1 auf Seite 32 herangezogen. Es zeigt sich, dass das Mittel der Trefferquote für die unterschiedlichen LDA-Varianten knapp höher ist als für tf-idf (bei LDA liegt das Mittel zwischen $23,2\%$ und $28,4\%$; bei tf-idf beträgt es $22,3\%$). Umgekehrt verhält es sich für die ESA-Modelle: hier liegt die mittlere Trefferquote zwischen $15,3\%$ und $17,9\%$ und damit deutlich unter jener von tf-idf.

In Tabelle 5.3 werden Genauigkeit, Trefferquote und F-Maß für die einzelnen Repräsentationsmodelle abgebildet. Für jedes Maß wird das Mittel der separaten Relevanzmerkmale dem Ergebnis des Metaklassifizierers gegenübergestellt. Es lässt sich schnell erkennen, dass die Kombination der Relevanzmerkmale im Metaklassifizierer ausnahmslos schlechtere Ergebnisse hervorbringt, als das Maximum der separaten Merkmale. Nicht einmal in einem Einzelmaß (Genauigkeit oder Trefferquote) kann das Maximum durch den kombinierten Relevanzmerkmale überboten werden.

[3]Die Werte unter Verwendung des Argumentdetektors finden sich ebenfalls in Anhang A

Tabelle 5.3: Evaluierung der kombinierten Relevanzmerkmale im Vergleich mit dem Maximum der separaten Relevanzmerkmale (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Relevanzmerkmal	Genauigkeit		Trefferquote		F-Maß	
	Maximum	Kombiniert	Maximum	Kombiniert	Maximum	Kombiniert
tf-idf	12,7	7,8	33,8	14,3	14,9	10,1
LDA [WikiArtikel]	5,6	3,9	42,5	9,5	8,4	5,5
LDA [iDebate]	5,7	4,3	30,2	9,5	9,6	5,9
LDA [WikiEvents]	4,8	2,9	47,9	7,1	7,3	4,1
ESA [WikiArtikel]	9,0	3,8	26,9	8,8	11,1	5,3
ESA [iDebate]	11,0	4,5	20,4	14,0	13,1	6,8
ESA [WikiEvents]	3,7	4,0	30,4	8,2	5,8	5,4

5.2.3 Analyse

Die Hypothese $H_{2.1}$ wird durch die Daten nicht bestätigt. Zwar weisen die LDA Modelle eine bessere Trefferquote auf, als das TF-IDF Modell, die ESA Modelle jedoch nicht. Das Gegenteil ist dort der Fall: Die Trefferquote ist bei den ESA Modellen einheitlich niedriger. Der Grund dafür, dass die LDA Modelle bezüglich der Trefferquote besser funktionieren als die ESA Modelle liegt vermutlich darin, dass die Konzepte bei ESA weniger trennscharf und unter Umständen unvollständig sind. Es liegen nur die Konzepte vor, die explizit bereitgestellt werden, und diese können auch Wörter enthalten, die für andere Konzepte wichtiger sind. Dagegen werden in den Konzepten bei LDA statistische Zusammenhänge festgehalten. Es entsteht dadurch eine höhere Trennschärfe und ein größerer Anteil inhaltlicher Zusammenhänge wird abgedeckt. Die Schlussfolgerungen aus den Daten zu Hypothese $H_{2.1}$ unterstützen aus einem anderen Blickwinkel die Annahme aus Kapitel 5.1.3, dass das Repräsentationsmodelle einen stärkeren Einfluss auf die Trefferquote haben, als die Relevanzmerkmale.

Auf Grundlage der Daten muss Hypothese $H_{2.2}$ klar verworfen werden. Das F-Maß der kombinierten Relevanzmerkmale ist nicht höher als das Maximum der separaten Merkmale. Darüber hinaus ist es sogar bis auf eine Ausnahme[4] niedriger als das Mittel der Relevanzmerkmale. Dass die Resultate des Metaklassifizierers schlechter sind als die Ergebnisse der separaten Relevanzmerkmale, wird einerseits erneut darauf zurückgeführt, dass in den Trainingsdaten Relevanz und Argumentativität nur in Kombination annotiert wurden (vergleiche Kapitel 5.4). Andererseits ist die Performanz der Relevanzmerkmale auf einem so niedrigen Niveau, dass diese für das Lernverfahren unter Umständen keine brauchbaren Informationen liefern.

[4]Die Ausnahme ist ESA [WikiEvents]. Die Genauigkeit war hier bei den separaten Merkmalen bereits so niedrig, dass der Steigerung durch Kombination der Merkmale keine Bedeutung zugemessen werden kann.

5.3 Die gleichzeitige Verarbeitung von Relevanz- und Argumentmerkmalen

Zur Untersuchung der gleichzeitigen Verarbeitung von Relevanz- und Argumentmerkmalen dient der dritte Programmablauf aus Kapitel 4.5 (vergleiche Abbildung 4.5 (c) auf Seite 28). Die Relevanzdetektoren und der Argumentdetektor bewerten die Sätze unabhängig voneinander. Alle Merkmale werden anschließend dem Metaklassifizierer zur Verfügung gestellt. Dieser kann mit den gegebenen Informationen die Sätze in einem Schritt auf Relevanz und Argumentativität prüfen. Die Ergebnisse werden mit den Beobachtungen aus Kapitel 5.2.2 verglichen, bei denen die Relevanzmerkmale kombiniert und anschließend bezüglich der Argumentativität gefiltert wurden.

5.3.1 Hypothesen

$H_{3.1}$: Das F-Maß ist bei der gleichzeitigen Berücksichtigung von Relevanz- und Argumentmerkmalen höher als bei der sequenziellen Filterung. Ein Problem für die Bestimmung der Relevanzmerkmale (und analog der Argumentmerkmale) liegt in der Annotation des Datensatzes. Es werden dort nur Sätze annotiert, die sowohl argumentativ als auch relevant sind. Bei der separaten Betrachtung der Merkmale ist der Lernprozess deshalb unsauber, weil etwa tatsächlich relevante Sätze nicht als solche erkannt werden sollen, wenn sie nicht argumentativ sind. Ein Verfahren, dass beide Merkmale gleichzeitig testet, sollte deshalb besser Ergebnisse erzielen.

5.3.2 Beobachtungen

Die Ergebnisse der kombinierten Relevanzmerkmale aus Tabelle 5.3 werden den Ergebnissen der gleichzeitigen Berücksichtigung von Relevanz- und Argumentmerkmalen in Tabelle 5.4 gegenübergestellt. Bis auf zwei Ausnahmen ist das F-Maß tatsächlich bei allen Repräsentationsmodellen höher. Die Ausnahmen sind tf-idf und erneut ESA [WikiEvents]. Für die anderen Modelle ist die Genauigkeit besser, wenn auch nur um durchschnittlich 0.6%. Bei der Trefferquote lässt sich keine Tendenz feststellen.

Tabelle 5.4: Die Ergebnisse der gleichzeitigen Berücksichtigung von Relevanz- und Argumentmerkmalen im Vergleich mit den Ergebnissen der kombinierten Relevanzmerkmale

Repräsentationsmodell [Indizierungsdaten]	Genauigkeit ohne Arg.	Genauigkeit mit Arg.	Trefferquote ohne Arg.	Trefferquote mit Arg.	F-Maß ohne Arg.	F-Maß mit Arg.
tf-idf	7,8	5,7	14,3	16,1	10,1	8,4
LDA [WikiArtikel]	3,9	4,5	9,5	9,4	5,5	6,1
LDA [iDebate]	4,3	4,6	9,5	9,7	5,9	6,3
LDA [WikiEvents]	2,9	3,7	7,1	8,4	4,1	5,1
ESA [WikiArtikel]	3,8	4,6	8,8	9,7	5,3	6,2
ESA [iDebate]	4,5	5,1	14,0	11,4	6,8	7,0
ESA [WikiEvents]	4,0	3,7	8,2	8,4	5,4	5,1

5.3.3 Analyse

Die Hypothese $H_{3.1}$ kann nicht bestätigt werden. Zwar gibt es eine leichte Tendenz, dass sich die Genauigkeit und dadurch auch das F-Maß verbessert, doch ist diese Verbesserung marginal. Für eine brauchbare Untersuchung der Hypothese mangelt es bereits bei den einfließenden Relevanz- und Argumentmerkmalen zu sehr an Performanz. Dennoch können die resultierenden Daten zumindest als Indiz dafür betrachtet werden, dass die gleichzeitige Berücksichtigung der Merkmale zur Steigerung der Performanz beiträgt.

5.4 Qualitative Analyse von Fehlerquellen

Die Evaluation brachte mehrere Fehlerquellen zum Vorschein, welche die Performanz des Programms oder einzelner Module stark beeinträchtigen. In diesem Abschnitt sollen diese Fehlerquellen qualitativ analysiert werden. Die Analyse des kompletten Korpus' von Aharoni et al. [2] wäre zu aufwändig, weshalb sich auf ein Thema konzentriert wird. Das Thema, das zur qualitativen Analyse herangezogen wurde, ist als Forderung formuliert: „This house believes that the sale of violent video games to minors should be banned.“[5] In den folgenden Kapiteln wird auf dieses Thema Bezug genommen.

Eine Ursache für die schlechten Ergebnisse der Detektoren wird darin vermutet, dass die Annotationen von Aharoni et al. für die maschinellen Lernverfahren ungeeignet sind. In den Annotationen werden Relevanz und Argumentativität zugleich berücksichtigt, während die Detektoren jeweils nur eins der Attribute erfassen sollen. Als Ursache für die geringe Wirksamkeit der Topic

[5]frei übersetzt: Diese Partei ist der Meinung, dass der Verkauf von Videospielen mit Gewaltdarstellung an Minderjährige verboten sein sollte.

Modeling Modelle wird der Mangel an Kontextinformationen in den einzelnen Sätzen untersucht.

5.4.1 Kombination aus Relevanz und Argumentativität

Durch die Evaluation des Programms wurde ein Problem sichtbar, das sowohl die Relevanzdetektoren als auch den Argumentdetektor betrifft: Die Detektoren sollen Sätze danach klassifizieren, ob sie relevant *oder* argumentativ sind, wohingegen die im Korpus vorhandenen Annotationen nur aufzeigen, ob ein Satz relevant *und* argumentativ ist.

Das Problem wird gut am Beispiel des Relevanzdetektors deutlich. Dieser soll alle relevanten Sätze erfassen, darunter argumentative und nicht argumentative[6]. Es kann angenommen werden, dass sich für beide Gruppen die Merkmalsausprägungen bezüglich der Relevanz nur geringfügig unterscheiden. Der Erfolg des Detektors wird nur an den relevanten und argumentativen Sätzen gemessen. Nach einer visuellen Analyse des Korpus' von Aharoni et al. [2] kann jedoch vermutet werden, dass die meisten der relevanten Sätze nicht argumentativ sind. Wenn also alle relevanten Sätze gefunden würden, wäre die Trefferquote zwar maximal, die Genauigkeit jedoch sehr niedrig, da die meisten gefundenen Sätze nicht argumentativ wären. Die schlechte Trefferquote in den Experimenten kann folglich zum Teil dadurch erklärt werden, dass sich das Programm nun einen leichten Anstieg der Genauigkeit mit einer starken Abnahme der Trefferquote erkauft, um das F-Maß zu maximieren. Analog verhält es sich für den Argumentdetektor, für den die Mehrheit der argumentativen Sätze nicht thematisch relevant ist.

Als Lösung für dieses Problem liegt es nahe, zur Beurteilung der Ergebnisse im Lernprozess Trainingskorpora zu verwenden, die jeweils nur Relevanz *oder* Argumentativität beurteilen. Für den Argumentdetektor ist ein solcher Korpus durch die Essays von Stab et al. [14] gegeben. Innerhalb dieser Essays, in denen die argumentativen Sätze annotiert sind, erreicht der Argumentdetektor eine Genauigkeit von $85,4\%$ und eine Trefferquote von $46,9\%$[7]. Auf den Wikipedia Daten konnte jedoch durch die Verwendung des Argumentdetektors keine Verbesserung des F-Maßes erzielt werden. Möglicherweise unterscheidet

[6] Relevante, aber nicht argumentative Sätze finden sich häufig in Form genereller Aussagen zum Thema wie „Video games have been studied for links to addiction and aggression." (frei übersetzt: Videospiele wurden auf Verknüpfungen hin zu Suchtverhalten und Aggressionen untersucht.)

[7] Im Versuchsaufbau wurde ein Zehntel der Essaysätze als Trainingsdaten und neun Zehntel als Testdaten verwendet.

sich die Struktur der Argumente in den Essays oder es liegen bei Aharoni et al. und Stab et al. unterschiedliche Definitionen dafür vor, was ein Argument ist.

Für die Beurteilung der Relevanz unabhängig von der Argumentativität liegt im Rahmen dieser Arbeit kein Trainingskorpus vor. Stattdessen kann die Abhängigkeit von der Argumentativität dadurch beseitigt werden, dass nur argumentative Sätze auf Relevanz geprüft werden. Als Eingabe in das Programm dienen dann ein Thema und die thematisch relevanten, argumentativen Sätze aller Themen[8]. Aus den Sätzen müssen dann nur noch die für das aktuelle Thema relevanten erfasst werden. Dieses Vorgehen entspricht der Annahme, man besäße einen perfekten Argumentdetektor, der im Programm vor dem Relevanzdetektor eingesetzt wird. Die Genauigkeitswerte bei diesem Vorgehen müssen mit Vorsicht betrachtet werden: Wäre tatsächlich der Argumentdetektor vorgeschaltet, so würden nur die Sätze relevanter Artikel untersucht werden, nicht die Argumente anderer Themen. Die argumentativen, nicht relevanten Sätze würden deshalb dem Thema deutlich näher stehen, als die für andere Themen relevanten Sätze. Die Wahrscheinlichkeit, dass ein Satz dann als relevant (also falsch-positiv) eingestuft wird, ist vermutlich deutlich größer als in diesem Versuchsaufbau.

Die Ergebnisse für diesen Versuchsaufbau werden in Tabelle 5.5 abgebildet. Die Performanz der Relevanzmerkmale ist erwartungsgemäß besser: Das F-Maß hat sich für jede Kombination aus Relevanzmerkmal und Repräsentationsmodell verbessert (vergleiche dazu Tabelle 5.1 auf Seite 32). Dennoch ist die Performanz nicht zufriedenstellend. Das beste Ergebnis liefert die Satzähnlichkeit unter LDA [WikiArtikel]. Die Genauigkeit erreicht hier mit $59,4\%$ ein brauchbares Niveau. Zusammen mit einer Trefferquote von $65,4\%$ ergibt sich ein F-Maß von $62,3\%$. Insbesondere die Genauigkeit muss hier aber mit Skepsis betrachtet werden. Die meisten anderen Kombinationen bleiben zudem weit hinter diesem Ergebnis zurück. Das drittbeste Ergebnis liefert erneut die Satzhäufigkeit, diesmal unter tf-idf. Dieses einfache Verfahren zur Berechnung der Relevanz sollte also nicht unterschätzt werden.

Auch über die Repräsentationsmodelle können durch diesen Versuch weitere Aussagen getroffen werden. Das einzige Modell, dass die Erwartungen erfüllt, ist LDA [WikiArtikel]: Hier bleibt die Genauigkeit konstant, während die Trefferquote im Mittel von $38,6\%$ auf $52,8\%$ steigt. Die ESA Modelle bleiben erneut hinter den LDA Modellen zurück. Es lässt sich jedoch nicht sagen, ob nur die explizit bereitgestellten Konzepte ungeeignet waren oder LDA im Kontext

[8] Zuvor bestand die Eingabe aus dem Thema und allen Sätzen eines Artikels, der für das Thema relevant ist.

Tabelle 5.5: Evaluierung der Relevanzmerkmale unter der Annahme eines perfekten Argumentdetektors (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Model [Indizierungsdaten]	Relevanzmerkmal	G	T	F
tf-idf	- Satzähnlichkeit	59,6	47,4	52,8
	- Subjektähnlichkeit	45,0	19,2	26,9
	- Nomenähnlichkeit	55,3	44,9	49,6
	- Erweiterte Nomenähnlichkeit	43,3	43,0	43,2
	Mittelwert	50,8	38,6	43,1
LDA [WikiArtikel]	- Satzähnlichkeit	59,4	**65,6**	**62,3**
	- Subjektähnlichkeit	43,8	43,9	43,8
	- Nomenähnlichkeit	55,2	59,0	57,1
	- Erweiterte Nomenähnlichkeit	41,9	42,7	42,3
	Mittelwert	50,1	52,8	51,4
LDA [iDebate]	- Satzähnlichkeit	33,3	48,9	39,6
	- Subjektähnlichkeit	21,1	45,4	28,8
	- Nomenähnlichkeit	27,5	49,9	35,5
	- Erweiterte Nomenähnlichkeit	22,1	34,9	27,0
	Mittelwert	26,0	44,8	32,7
LDA [WikiEvents]	- Satzähnlichkeit	20,0	19,9	19,9
	- Subjektähnlichkeit	13,5	14,9	14,2
	- Nomenähnlichkeit	12,6	26,8	17,1
	- Erweiterte Nomenähnlichkeit	8,5	23,3	12,4
	Mittelwert	13,6	21,2	15,9
ESA [WikiArtikel]	- Satzähnlichkeit	23,4	24,9	24,1
	- Subjektähnlichkeit	34,4	26,3	29,8
	- Nomenähnlichkeit	19,3	27,2	22,6
	- Erweiterte Nomenähnlichkeit	13,3	13,8	13,6
	Mittelwert	22,6	23,1	22,5
ESA [iDebate]	- Satzähnlichkeit	14,7	12,3	13,4
	- Subjektähnlichkeit	**75,0**	17,0	27,7
	- Nomenähnlichkeit	60,0	10,1	17,3
	- Erweiterte Nomenähnlichkeit	65,4	6,3	11,5
	Mittelwert	53,8	11,4	17,5
ESA [WikiEvents]	- Satzähnlichkeit	6,5	4,9	5,6
	- Subjektähnlichkeit	33,0	5,0	8,7
	- Nomenähnlichkeit	42,9	4,2	7,6
	- Erweiterte Nomenähnlichkeit	56,9	5,3	9,7
	Mittelwert	34,8	4,8	7,9

dieser Arbeit generell besser funktioniert. Die Wikipedia Ereignisse erweisen sich als Indizierungsdaten für LDA und ESA ungeeignet. Die Wikipedia Artikel, die zugleich als Quelle der thematisch relevanten Argumentationseinheiten dienen, eignen sich offenbar besser als Indizierungsdatensatz als die iDebate Einführungstexte. Dies liegt auch nahe, da sich die Indizierungsdaten im Falle von LDA sprachlich nicht von den Wikipedia-Artikeln unterscheiden (sie sind identisch) und im Falle von ESA alle relevanten Themen enthalten sollten (die Artikel wurden manuell und auf Grundlage der Themen gesammelt). Dass der Indizierungsdatensatz und der Korpus, in welchem nach relevanten Argumenten gesucht wird, bei der Verwendung von ESA identisch sind, wird allerdings für eine reale Anwendung ausgeschlossen.

Die Kombination der Relevanzmerkmale im Metaklassifizierer liefert überraschenderweise erneut schlechtere Ergebnisse als die separaten Relevanzmerkmale (vergleiche Tabelle 5.6). Um die Ursachen dafür auszumachen, bedarf es einer ausführlichen Analyse, die im Rahmen dieser Arbeit nicht durchgeführt werden konnte.

Tabelle 5.6: Evaluierung der kombinierten Relevanzmerkmale unter der Annahme eines perfekten Argumentdetektors (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Relevanzmerkmal	Genauigkeit [%]		Trefferquote [%]		F-Maß	
	Maximum	Kombiniert	Maximum	Kombiniert	Maximum	Kombiniert
tf-idf	59,6	37,4	47,4	53,4	52,8	44,0
LDA WikiArtikel	59,4	40,7	65,6	65,0	62,3	50,1
LDA iDebate	33,3	23,3	49,9	38,9	39,6	29,2
LDA WikiEvents	20,0	4,7	26,8	8,3	19,9	6,9
ESA WikiArtikel	34,4	13,5	27,2	29,9	29,8	18,6
ESA iDebate	75,0	5,6	17,0	11,2	27,7	7,4
ESA WikiEvents	56,9	1,4	5,3	45,7	9,7	2,8

5.4.2 Fehlender Kontext bei der Klassifizierung

Bei der Auswertung der Repräsentationsmodelle zeigt sich, dass das tf-idf Modell von den anderen Modellen nicht oder nur leicht überboten wird. Bei den Ergebnissen in Tabelle 5.1 auf Seite 32 übersteigt zwar die Trefferquote der LDA Modelle die Quote des tf-idf Modells, doch wegen der geringeren Genauigkeit erweist sich das F-Maß als niedriger. Bei dem Versuch mit der Simulation eines perfekten Argumentdetektors aus Kapitel 5.4.1 kann nur LDA [WikiArtikel] das tf-idf Modell übertreffen. Hier wird die gleiche Genauigkeit erreicht und die Trefferquote deutlich überboten, sodass das F-Maß von $43,1\%$ auf $51,4\%$ steigt (siehe Tabelle 5.5). Alle anderen Modelle sind deutlich weniger performant: das höchste F-Maß unter ihnen beträgt nur noch $32,7\%$, das niedrigste

sogar nur $7,9\%$. Es stellt sich die Frage, warum die Topic Modeling Modelle oft schlechter abschneiden, als das tf-idf Modell.

Aus einer Untersuchung der falsch-negativen Ergebnissen wird geschlussfolgert, dass nicht genügend Kontextinformationen in den Sätzen enthalten sind. Bei diesen Sätzen sollte der Kontext zwar mit dem aktuellen Thema verbunden sein, doch er bleibt entweder unklar oder verweist sogar auf andere Themen. Dafür gibt es drei unterschiedliche Ursachen:

1. Der Kontext erschließt sich oft erst über Satzgrenzen hinweg
2. Argumentative Sätze enthalten einen hohen Anteil themenfremder Worte
3. Argumente gegen eine Behauptung benutzen oft gegenteiliges Vokabular

Sätze stehen innerhalb eines Textes nicht unabhängig voneinander. Sie enthalten viele Referenzen auf andere Sätze. Ein Beispiel bietet folgender Satz, der von keinem der Relevanzdetektoren positiv klassifiziert wurde: „He counters the premise of these studies with the concept that not all depictions of violence are even bad to witness.“[9] Das Wort ‚he‘ (er) verweist als Pronomen auf eine Person, die bereits zuvor im Text eingeführt wurde. Auch das Demonstrativpronomen in ‚of these studies‘ (dieser Studien) verdeutlicht, dass schon vorher die Rede von besagten Studien war. Aus den Sätzen, auf die hier verwiesen wird, könnte hervorgehen, was diese Person mit dem Thema verbindet (sie belegt, dass Gewaltdarstellung in Medien nicht negativ sein muss) und worum es in den Studien geht (diese zeigen an Versuchen mit Kindern, dass mediale Gewaltdarstellung aggressive Handlungen fördern kann). Der Kontext eines Satzes geht also über den Satz hinaus. Eine Anaphernresolution, das ist die Rückführung der Pronomen auf dasjenige, worauf sie sich beziehen, kann hilfreich sein, um diesen Kontext nachzuverfolgen.

Argumentative Sätze bewegen sich oft auf einer Metaebene. Es wird dabei nicht direkt das Argument genannt, sondern beschrieben, dass gewisse Autoritäten ein Argument bekräftigen oder ihm widersprechen. Beispiele dafür sind etwa „This study found no evidence [sic!] violent games are psychologically harmful to minors.“[10] oder „The surveys also found correlations between violent gameplay and some common childhood problems.“[11] Es stehen hier

[9] frei übersetzt: Er entgegnete der Prämisse dieser Studien, dass nicht alle Darstellung von Gewalt als schlecht ist.

[10] frei übersetzt: Diese Studie hat keinen Beleg dafür gefunden, dass Gewaltspiele für Minderjährige psychologisch schädlich sind.

[11] frei übersetzt: Die Erhebung konnte zusätzlich einen Zusammenhang zwischen dem Spielen von Gewaltspielen und verbreiteten Kindheitsproblemen ausmachen.

die Studien (‚study', ‚surveys') und deren Ergebnisse (‚evidence', ‚correlation') im Vordergrund des Satzes. Der eigentliche Inhalt, also das, was die Studien bekräftigen oder entkräften, rückt dadurch in ein Objekt. Insbesondere die Subjektähnlichkeit ist in solchen Situationen ungeeignet, die Relevanz des Satzes zu bestimmen; aber auch die Nomenähnlichkeit erweist sich wegen der Metaebene in solchen Sätzen als ungenau.

Eine bemerkenswerte Beobachtung konnte bei Argumenten gegen die These, die das Thema definiert, festgestellt werden. Thesen werden durch Gegenargumente oftmals nicht verneint, sondern das Gegenteil wird gezeigt. Dabei entsteht das Problem, dass der Inhalt der Gegenargumente keine Schlagwörter zur These mehr enthält und deshalb nicht mehr mit dem Thema assoziiert wird. Im folgenden Beispiel werden also nicht Aggressionen oder Gewaltbereitschaft verneint, sondern die Förderung von Teamgeist hervorgehoben: „Recent research has suggested that some violent video games may actually have a prosocial effect in some contexts, for example, team play."[12] Andere, gleichartige Gegenargumente bezeugen einen positiven Einfluss auf die Entwicklung sozialer Fähigkeiten[13] und das psychologische Wohlbefinden[14]. Eine Relevanz kann hier aufgrund des unterschiedlichen Vokabulars weder mit Topic Modeling Modellen, noch mittels tf-idf festgestellt werden. Eine Unterrepräsentation der Gegenargumente wäre fatal für die Absicht, Zugriff auf ein repräsentativeres Meinungsbild zu liefern. Dieses Problem sollte deswegen dringend behoben werden.

[12] frei übersetzt: Die aktuelle Forschung legt die Vermutung nahe, dass manche Videospiele mit Gewaltdarstellung in ihrem Kontext einen sozial positiven Effekt haben, zum Beispiel auf Zusammenarbeit.

[13] „Video games also develop the individual's intelligence, and in social games develop the social capabilities of the individual." (frei übersetzt: Videospiele fördern die Aufassungsgabe und in sozialen Spielen die soziale Kompetenz des Spielers.)

[14] „Furthermore, numerous researchers have proposed potential positive effects of video games on aspects of social and cognitive development and psychological well-being." (frei übersetzt: Weiterhin haben zahlreiche Forscher einen möglichen positiven Effekt von Videospielen auf Aspekte der sozialen und kognitiven Entwicklung sowie das psychologische Wohlbefinden erklärt.)

Kapitel 6

Fazit

Die vorliegende Arbeit befasste sich mit der Erfassung thematisch relevanter Argumentationseinheiten. Es wurden vier Relevanzdetektoren und ein Argumentdetektor implementiert, sieben Repräsentationsmodelle basierend auf drei Indizierungsdatensätzen kamen zur Anwendung und drei verschiedene Experimente wurden durchlaufen, um das entstandene System zu evaluieren. Es wurden damit zwei Ziele verfolgt: Erstens wurde versucht aufbauend auf Levy et al.[9] die Erfassung thematisch relevanter Argumentationseinheiten zu optimieren; zweitens sollte die Bedeutung einzelner Bestandteile des Programms für die Suche relevanter Argumente evaluiert werden.

Bei Levy et al. wurde keine Motivation für die Einführung der Relevanzmerkmale gegeben. Eine Analyse der Subjekt- und der Nomenähnlichkeit in der vorliegenden Arbeit macht deutlich, dass sie sich bei einem wesentlichen Teil der argumentativen Sätze nicht eignen, die Relevanz zu bestimmen (siehe Kapitel 5.4.2). Die hier als Referenz eingeführte, naive Satzähnlichkeit erzielte ähnliche Ergebnisse wie die Subjekt- und die Nomenähnlichkeit, wodurch der jener Relevanzmerkmale weiter infrage gestellt wird (siehe Kapitel 5.1). Zur Optimierung der Relevanzmerkmale wurden Repräsentationsmodelle aus dem Bereich des Topic Modelings verwendet. Dies war nur bei der Latenten Dirichlet Allokation erfolgreich. Bei den anderen Modellen verschlechterte sich das Resultat sogar deutlich (siehe Kapitel 5.4.2). Es wird angenommen, dass die einzelnen Sätze zu kurz sind, um die enthaltenen Konzepte präzise bestimmen zu können.

Die Evaluation der einzelnen Bestandteile des Programms wurde durch die schlechten Ergebnisse der Relevanzmerkmale stark eingeschränkt, sodass die Ergebnisse keine Aussagekraft haben. Eine wichtigere Rolle spielte deshalb

die Analyse der Fehlerquellen, die wichtige Erkenntnisse für weitere Forschung hervorbrachte. Zwei wesentliche Probleme wurden deutlich: Erstens mangelt es an passenden Trainingskorpora um Relevanz und Argumentativität unabhängig voneinander evaluieren und die Detektoren trainieren zu können (siehe Kapitel 5.4.1); zweitens geht bei der Trennung der Sätze der Zusammenhang verloren (siehe Kapitel 5.4.2), weil Verweise wie zum Beispiel Pronomen nicht mehr aufgelöst werden können.

Weitere Forschung sollte dementsprechend einerseits geeignete Trainingskorpora erarbeiten, die es ermöglichen Relevanz- und Argumentmerkmale separat voneinander zu trainieren und zu evaluieren. Andererseits sollte dem Kontext von Sätzen und Satzteilen eine größere Bedeutung zukommen. Der Kontext von Sätzen wird für das Topic Modeling (siehe oben) und für die Erfassung von Relevanz (siehe Kapitel 5.4.2) benötigt. Der Kontext von Satzteilen könnte genutzt werden um zwischen argumentativen Strukturen und der inhaltlichen Aussage eines Satzes zu unterscheiden, damit nur letztere zur Bewertung der Relevanz herangezogen wird. Eine besondere Herausforderung bei der Erfassung thematisch relevanter Argumentationseinheiten stellen Argumente dar, die der Aussage des Themas widersprechen. Hier ist es möglich, dass über das Vokabular keine Ähnlichkeit zum Thema festgestellt werden kann (siehe Kapitel 5.4.2). Auch hier ist weitere Forschung notwendig.

In Zukunft könnte ein Programm entstehen, das zuerst größere Textabschnitte auf Relevanz prüft. In relevanten Abschnitten würde dann mit anspruchsvolleren Algorithmen nach argumentativen Sätzen gesucht werden. Nur die inhaltliche Aussage des Satzes würde erneut auf Relevanz geprüft werden, um sicherzugehen, dass der argumentative Satz tatsächlich einen Bezug zum Thema hat. Innerhalb der inhaltlichen Aussage könnten dann auch Subjekt- und Nomenähnlichkeit (siehe Kapitel 4.2.1) wirksam eingesetzt werden.

Das Ziel, das Verfahren von Levy et al. zu optimieren, wurde in dieser Arbeit nur bedingt erreicht. Abschließend lässt sich jedoch sagen, dass die Evaluation des Programms einen vielversprechenden Weg aufzeigt, thematisch relevante Argumente zuverlässig zu erfassen.

Anhang A

Weitere Ergebnistabellen

Die folgenden Tabellen enthalten die Ergebnisse des ersten Programmablaufs (siehe Kapitel 4.5). Bei der Evaluation in der Arbeit wurde der Argumentdetektor nicht berücksichtigt, weil die Ergebnisse zu einer geringeren Performanz führten. Zum Vergleich werden die Daten in Tabellen A.1 (Seite 48) und A.2 (Seite 49) präsentiert.

Tabelle A.1: Evaluierung der Relevanzmerkmale in Abhängigkeit vom jeweiligen Repräsentationsmodell. Die Ergebnisse wurden mittels Argumentdetektor auf Argumentativität geprüft. (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Model [Indizierungsdaten]	Relevanzmerkmal	Genauigkeit [%]	Trefferquote [%]	F-Maß
tf-idf	- Satzähnlichkeit	8,8	12,1	10,2
	- Subjektähnlichkeit	14,7	7,1	9,6
	- Nomenähnlichkeit	9,3	6,5	7,6
	- Erweiterte Nomenähnlichkeit	12,4	6,6	8,7
	Mittelwert	11,3	8,1	9,0
LDA [WikiArtikel]	- Satzähnlichkeit	4,2	17,0	6,7
	- Subjektähnlichkeit	5,8	6,5	6,1
	- Nomenähnlichkeit	4,7	15,6	7,2
	- Erweiterte Nomenähnlichkeit	5,6	7,1	6,2
	Mittelwert	5,1	11,5	6,6
LDA [iDebate]	- Satzähnlichkeit	6,4	5,8	6,1
	- Subjektähnlichkeit	5,8	11,3	7,6
	- Nomenähnlichkeit	5,0	7,0	5,8
	- Erweiterte Nomenähnlichkeit	5,2	10,1	6,8
	Mittelwert	5,6	8,6	6,6
LDA [WikiEvents]	- Satzähnlichkeit	7,4	6,0	6,6
	- Subjektähnlichkeit	3,9	6,2	4,8
	- Nomenähnlichkeit	3,8	17,4	6,2
	- Erweiterte Nomenähnlichkeit	4,0	7,7	5,3
	Mittelwert	4,8	9,3	5,7
ESA [WikiArtikel]	- Satzähnlichkeit	6,2	8,7	7,2
	- Subjektähnlichkeit	11,2	5,4	7,3
	- Nomenähnlichkeit	5,7	4,5	5,0
	- Erweiterte Nomenähnlichkeit	4,7	2,0	2,8
	Mittelwert	6,9	5,1	5,6
ESA [iDebate]	- Satzähnlichkeit	4,9	5,4	5,1
	- Subjektähnlichkeit	13,3	6,9	9,1
	- Nomenähnlichkeit	4,3	6,0	5,0
	- Erweiterte Nomenähnlichkeit	4,3	2,0	2,7
	Mittelwert	6,7	5,1	5,5
ESA [WikiEvents]	- Satzähnlichkeit	4,3	3,2	3,6
	- Subjektähnlichkeit	4,4	6,0	5,1
	- Nomenähnlichkeit	3,3	6,2	4,3
	- Erweiterte Nomenähnlichkeit	3,2	9,8	4,9
	Mittelwert	3,8	6,3	4,5

Tabelle A.2: Evaluierung der kombinierten Relevanzmerkmale im Vergleich mit dem Maximum der separaten Relevanzmerkmale. Die Ergebnisse wurden mittels Argumentdetektor auf Argumentativität geprüft. (G = Genauigkeit, T = Trefferquote, F = F-Maß)

Relevanzmerkmal	Genauigkeit [%]		Trefferquote [%]		F-Maß	
	Mittel	Kombiniert	Mittel	Kombiniert	Mittel	Kombiniert
tf-idf	11,3	8,8	8,1	5,1	9,0	6,5
LDA WikiArtikel	5,1	5,5	11,5	3,7	6,6	4,4
LDA iDebate	5,6	4,4	8,6	3,3	6,6	3,7
LDA WikiEvents	4,8	3,6	9,3	2,1	5,7	2,6
ESA WikiArtikel	6,9	4,9	5,1	3,4	5,6	4,0
ESA iDebate	6,7	4,8	5,1	5,1	5,5	4,9
ESA WikiEvents	3,8	5,6	6,3	2,9	4,5	3,8

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

[1] Wordnet an electronic lexical database. Cambridge, MA ; London, May 1998. The MIT Press. 4.2.1

[2] Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation and Computation, ACL 2014*, 2014. 2, 4, 4.1, 4.2.2, 4.3, 4.4, 5.1.3, 5.4, 5.4.1

[3] Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. Cross-Domain Mining of Argumentative Text through Distant Supervision. In *15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 16)*, pages 1395–1404. Association for Computational Linguistics, 2016. 4.2.2

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. volume 3, pages 993–1022, 2003. 3.1.1, 3.1.2, 3.2.1

[5] Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. Supporting human answers for advice-seeking questions in cqa sites. In *Proceedings of European Conference on Information Retrieval*, 2016. 1

[6] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. volume 41, pages 391–407, 1990. 3.1.3

[7] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *In Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, 2007. 3.2, 3.2.2

[8] G. Holmes, A. Donkin, and I.H. Witten. Weka: A machine learning workbench. In *Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994. 4.3, 8

[9] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. Context dependent claim detection. 2014. (document), 1, 2, 4, 4.1, 4.2.1, 4.5, 5.1.1, 6

[10] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. 4.2.1

[11] Arunav Mishra and Klaus Berberich. Leveraging semantic annotations to link wikipedia and news archives. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, pages 30–42, Cham, 2016. Springer International Publishing. 4.2.2

[12] Ross Quinlan. C4.5: Programs for machine learning. San Mateo, CA, 1993. Morgan Kaufmann Publishers. 4.3, 4.4

[13] S. Somasundaran, J. Ruppenhofer, and J. Wiebe. Detecting arguing and sentiment in meetings. In *SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007 (SIGdial Workshop 2007)*, 2007. 4.3

[14] Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In Junichi Tsujii and Jan Hajic, editors, *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. 4.3, 5.4.1