Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bachelorarbeit

Quantifizierung von Unsicherheiten in der Duplikaterkennung

vorgelegt von

Shawn Ray Söker

Matrikelnummer 7322384

Studiengang Informatik

MIN-Fakultät

Fachbereich Informatik

eingereicht am 13. September 2025

Betreuer und Erstgutachter: Prof. Dr. Fabian Panse

Zweitgutachter: Dr. Jannek Squar

Zusammenfassung

Diese Arbeit untersucht die Unsicherheit in der Duplikaterkennung, einer wichtigen Problemstellung der Datenqualität im Kontext von Big Data. Moderne Verfahren wie Ditto, HierGAT und Unicorn wurden bereits mit klassischen Metriken wie Precision, Recall und F1-Score umfangreich evaluiert, aber bislang blieb unklar, wie konsistent ihre Vorhersagen tatsächlich sind, also wie häufig sie sich selbst widersprechen. Hier setzt die Arbeit an. Es wurde ein Unsicherheitsindex definiert, der auf der Übereinstimmung verschiedener Modellinstanzen mit unabhängigem Training basiert. In verschiedenen Experimenten wurden die Effekte von mehrfacher Ausführung, Variation der Hardware und Parametern und dem Vergleich mehrerer Verfahren analysiert. Des Weiteren wurde untersucht, ob Unsicherheit als Größe genutzt werden kann, um durch gezielte manuelle Klassifikation die Performance effizient zu steigern, und ob die Konfidenz, die Ditto ausgibt, genutzt werden kann, um die Unsicherheit vorherzusagen. Außerdem wurden die Schnittmengen verschiedener Verfahren betrachtet und besonders unsichere Datenpaare manuell und durch ChatGPT charakterisiert.
Die Ergebnisse zeigen, dass die untersuchten modernen Verfahren in ihren Vorhersagen recht konsistent, aber nicht vollständig zuverlässig sind. Die experimentell bestimmte Unsicherheit erwies sich als nützliche Größe, um besonders Datenpunkte zu identifizieren, die für eine manuelle Klassifikation besonders geeignet sind. So kann die Effizienz in einem praktischen Einsatz erhöht werden. Damit liefert die Arbeit eine methodische Grundlage zur Quantifizierung von Unsicherheit und praktische Hinweise zur Nutzung unsicherer Daten in der Duplikaterkennung.

Inhaltsverzeichnis

1 Einführung

Die vorliegende Arbeit untersucht die Frage, wie stabil und konsistent moderne Verfahren der Duplikaterkennung in ihren Vorhersagen sind. Mit diesem besonderen Fokus auf Unsicherheit statt auf Performance-Metriken wird analysiert, inwieweit aktuelle Modelle in ihren Vorhersagen schwanken, welche Faktoren diese Schwankungen beeinflussen und wie Unsicherheit praktisch genutzt werden kann.

1.1 Motivation

Im Zeitalter von Big Data ist die Qualität von Daten entscheidend, um in Wirtschaft, Wissenschaft und öffentlicher Verwaltung fundierte Entscheidungen treffen zu können. Ein zentrales Problem der Datenqualität ist die Duplikaterkennung, auch genannt Entity Matching (EM): Mehrfache Einträge, die derselben realen Entität entsprechen, müssen erkannt und zusammengeführt werden [NH10], um Redundanz zu verhindern.
Die Aufgabe ist jedoch schwierig, weil reale Entitäten in Datenbanken sehr unterschiedlich repräsentiert sein können. Zwei Datensätze können sich stark ähneln, ohne dass es sich um dieselbe Entität handelt. Oder sie können die gleiche Entität beschreiben, obwohl die Darstellung, Attribute und Formate deutlich voneinander abweichen.

KI-Systeme, die auf Large Language Models (LLMs) basieren, wie Ditto [Li+20], haben in den letzten Jahren große Fortschritte erzielt und sind in der Lage, semantische Zusammenhänge zu erfassen und somit über oberflächliche Ähnlichkeiten hinauszugehen. Sie sind damit durchaus leistungsstark. Aufgrund ihrer Komplexität sind ihre Entscheidungen und die Stabilität der Entscheidungen jedoch schwer nachvollziehbar. Es lassen sich zwar mit etablierten Metriken wie Precision, Recall und F1-Score verlässlich Aussagen über die Güte des Verfahrens treffen, doch geben diese Metriken keinen Aufschluss darüber, wie sicher oder unsicher die Entscheidungen in einzelnen Fällen sind.

Vor diesem Hintergrund ist es für Forschung und Praxis relevant, Unsicherheit in der Duplikaterkennung systematisch zu untersuchen. Zum einen eröffnet dies die Möglichkeit, die Stabilität aktueller Verfahren einzuordnen und Unterschiede zwischen Modellinstanzen, Verfahren oder Hardware-Umgebungen sichtbar zu machen. Zum anderen kann die Unsicherheit auch als praktische Größe verwendet werden. Wenn sich zeigt, dass unsichere Datenpunkte besonders geeignet sind, manuell klassifiziert zu werden, um die Performance des Modells zu erhöhen, ergibt sich daraus ein klarer praktischer Nutzen. Damit wird Unsicherheit von einem bislang kaum betrachteten Aspekt zu einer verwendbaren Größe, die unmittelbare Bedeutung für den Einsatz moderner Duplikatserkennungsverfahren hat.

1.2 Fragestellung und Aufbau der Arbeit

In der Arbeit soll die Unsicherheit in der Duplikaterkennung untersucht und quantifiziert sowie ihre praktischen Konsequenzen analysiert werden. Der Ausgangspunkt ist, die Beobachtung, dass die Performance moderner Verfahren der Duplikaterkennung, wie Ditto [Li+20], HierGAT [Yao+22] und Unicorn [Tu+23], gut mit klassischen Metriken, wie Precision, Recall und F1-Score, untersucht ist, die Modelle jedoch nur unzureichend Auskunft über die Stabilität und Zuverlässigkeit ihrer Vorhersagen geben.

Im ersten Schritt wird die Unsicherheit der Modelle quantifiziert, indem ein Unsicherheitsindex berechnet wird, welcher auf der Übereinstimmung der Vorhersagen verschiedener Modellinstanzen basiert. Dabei wird insbesondere untersucht, wie die Unsicherheit durch veränderte Hardware, unterschiedliche Hyperparameter und die Wahl des Verfahrens beeinflusst wird.

In einem zweiten Schritt wird untersucht, ob die Unsicherheit in der Duplikaterkennung ein geeigneter Maßstab ist, um besonders schwierige Datenpunkte auszuwählen und diese manuell zu klassifizieren, um die Modellleistung zu erhöhen. Hierzu wird ein Orakel als Ersatz für die manuelle Klassifikation verwendet.

Darüber hinaus wird analysiert, ob die Unsicherheit bereits aus der internen Konfidenzangabe eines Modells wie Ditto entnommen werden kann und welche gemeinsamen Eigenschaften besonders unsichere Daten aufweisen. Schließlich werden Überschneidungen in der Vorhersage verschiedener Verfahren, einschließlich der klassischen ML-Verfahren Random Forest und XGBoost, untersucht, um Unterschiede und Gemeinsamkeiten in der Duplikatklassifikation verschiedener Ansätze zu erkennen.

Der Aufbau der Arbeit ist wie folgt: In Kapitel 2 wird der Stand der Forschung in modernen Verfahren der Duplikaterkennung dargestellt und auf Konzepte der Unsicherheit in neuronalen Netzen eingegangen. In Kapitel 3 wird das theoretische Konzept entwickelt. Dieses umfasst die Definition eines Unsicherheitsindexes, ein Orakelmodell und die Definition eines Konfidenzwertes. Kapitel 5 zeigt das methodische Vorgehen der Experimente und Analysen. In Kapitel 5 werden die Ergebnisse dargestellt, welche in Kapitel 6 diskutiert werden. Schließlich gibt Kapitel 7 eine Zusammenfassung und einen Ausblick.

2 Stand der Forschung

Die Forschung in der Duplikaterkennung hat sich in den letzten Jahren weiterentwickelt. In diesem Abschnitt werden die theoretischen Grundlagen von neuronalen Netzen und Large Language Models erklärt und es wird Ditto als aktuelles repräsentierendes Verfahren der Duplikaterkennung vorgestellt sowie HierGAT und Unicorn als aufbauende Entwicklungen. Anschließend wird der aktuelle Stand der Forschung bei Unsicherheitsbestimmung in der Duplikaterkennung und in neuronalen Netzen erklärt.

2.1 Neuronale Netze und Large Language Models (LLMs) für Entity Matching

Neuronale Netze sind Modelle, die durch die Verknüpfung vieler simpler Verarbeitungseinheiten (Neuronen) in der Lage sind, komplexe Muster in Daten zu lernen und so eine Vielzahl von Aufgaben lösen können. [GBC16] Im Kontext von Duplikaterkennung kommen vor allem tiefe neuronale Netze (Deep Neural Networks), also Netze mit vielen Schichten von Neuronen, zum Einsatz.
Eine zentrale Architektur in diesem Bereich sind Transformer. [Vas+23] Sie verarbeiten die Eingabe parallel, statt sequenziell, und gewichten die Relation zwischen Token durch den Mechanismus der Self-Attention. So sind sie in der Lage, lange Abhängigkeiten im Text zu identifizieren. Transformer sind die Grundlage für Large Language Models (LLMs).
LLMs wie BERT [Dev+19] sind Deep Neural Networks, die auf großen Textmengen vortrainiert wurden und so ein Verständnis für natürliche Sprache entwickelt haben. Sie sind besonders geeignet für das Entity-Matching, weil sie semantische Zusammenhänge im Text erfassen können. LLMs sind bekannt dafür, intransparent in ihren Entscheidungen zu sein [Lip17] und geben ihre Ausgaben zwar mit einer Wahrscheinlichkeit aus, aber es ist fraglich, ob diese auch beobachtbarer Unsicherheit entspricht. Intransparenz meint hierbei die fehlende Nachvollziehbarkeit der Entscheidungen. Gerade diese Intransparenz macht eine zusätzliche Prüfung der Unsicherheit hilfreich. Des Weiteren gibt es im Bereich des Entity Matching aktuelle Forschung, um die Entscheidungen der Modelle erklärbar zu machen. [Bar+21]

2.2 Ditto für Entity-Matching

Vor diesem Hintergrund hat sich in den vergangenen Jahren bei der Duplikaterkennung der Fokus von klassischen regelbasierten [EIV07] und einfachen ML-Verfahren [EIV07] hin zu tieferen und komplexeren ML-Verfahren verlagert. Ein prominentes Beispiel für diesen Fortschritt ist Ditto [Li+20]. Ditto ist ein Verfahren für Entity-Matching (EM), das auf vortrainierten Sprachmodellen wie BERT basiert und deren Fähigkeit nutzt, semantische Beziehungen im Text zu erfassen. Zusätzlich führt Ditto drei weitere Optimierungen ein. Diese umfassen (1) das Zusammenfassen von langen Einträgen mittels TF-IDF-basierter Verfahren, um Informationen bei sehr umfangreichen Textfeldern zu extrahieren, (2) Injektion domainspezifischen

Wissens durch Markierung und Normalisierung wichtiger Informationseinheiten innerhalb der Datensätze und (3) Datenaugmentation, um die Robustheit des Modells gegenüber Rauschen und unvollständigen Informationen zu erhöhen. Ditto war im Jahr 2020 SOTA und erzielte in umfassenden Benchmark-Vergleichen durchweg bessere Ergebnisse als klassische Modelle und damals führende Deep-Learning-Ansätze. Bemerkenswert ist, dass Ditto besonders unter Bedingungen mit begrenztem Trainingsmaterial oder starker Datenverschmutzung vergleichsweise hohe Matching-Qualität beibehält.

2.3 Entwicklungen nach Ditto

Nach Ditto wurden weitere spezialisierte Verfahren entwickelt, von denen zwei in dieser Arbeit aufgesetzt wurden.

2.3.1 HierGAT

HierGAT [Yao+22] versteht Entity Matching als hierarchisches Problem und berücksichtigt Abhängigkeiten auf verschiedenen Ebenen. Auf der inneren Ebene verarbeitet es die Attribute eines Eintrags in Attention-Schichten, um innerhalb und zwischen Attributen relevante Informationen zu erfassen. Auf der äußeren Ebene verknüpft ein Graph Attention Network die Kandidatenpaare untereinander und trifft gemeinsame Entscheidungen im Matching. Somit kann das Verfahren Abhängigkeiten explizit ausnutzen und Widersprüche reduzieren, wenn zum Beispiel mehrere Paare dieselbe Entität referenzieren. HierGAT zeigte in Experimenten robuste Leistungen, insbesondere bei komplexen oder verrauschten Daten.

2.3.2 Unicorn

Ein anderer Ansatz ist Unicorn [Tu+23], wo das Ziel verfolgt wird, unterschiedliche Matching-Aufgaben wie Entity Matching, Entity Linking oder Schema Matching in einem einheitlichen Modell zusammenzuführen. Das Verfahren basiert auf der Idee, verschiedene Eingaben zuerst in Textrepräsentationen zu überführen, diese durch ein vortrainiertes Sprachmodell zu kodieren und mit einer „Mixture-of-Experts“-Schicht an die jeweilige Aufgabe anzupassen. Dadurch kann Unicorn Wissen zwischen unterschiedlichen Matching-Aufgaben teilen. So wird die Effizienz gesteigert und Zero-Shot- und Few-Shot-Anwendungen ermöglicht. In Benchmarks zeigte Unicorn, dass es mit aufgabenspezifischen Modellen mithalten oder sie sogar übertreffen kann.

HierGAT und Unicorn werden als zwei Modelle mit vergleichbarer Leistung aufgesetzt und zum Vergleich mit Ditto herangezogen.

2.3.3 ChatGPT

Neuere Untersuchungen wie „Using ChatGPT for Entity Matching“ [PB23] zeigen, dass Zero-Shot-Anwendungen mit LLMs wie ChatGPT kompetitive Ergebnisse zu Modellen erreichen können, die Finetuning benötigen. Der Vorteil hierbei ist, dass keine Trainingsdaten mit Labels benötigt werden, was sicherlich für die praktische Anwendung von großem Vorteil ist.

2.4 Untersuchung von Unsicherheit im Entity-Matching und in Neuronalen Netzen

In der aktuellen Literatur gibt es kaum Arbeiten, die die Unsicherheit von EM-Anwendungen systematisch untersuchen. Klassische, nicht auf ML basierende Verfahren und moderne Verfahren geben häufig einen Matching Score oder eine Wahrscheinlichkeit für ein Duplikat aus. Zusätzlich gibt es einen Schwellenwert, ab wann ein Paar als Match angesehen wird. Ein naiver Ansatz zur Bestimmung der Unsicherheit ist es folglich, die Unsicherheit als Abstand des Matching Scores beziehungsweise der Wahrscheinlichkeit zum Schwellenwert zu definieren. Es ist jedoch nicht gegeben, dass dieser Abstand auch automatisch der Unsicherheit entspricht, die man experimentell feststellen würde.
In der allgemeinen Forschung zu neuronalen Netzen gibt es hingegen umfangreiche Forschung zur Quantifizierung von Unsicherheit. In der Statistik und insbesondere im Machine Learning wird zwischen epistemischer und aleatorischer Unsicherheit unterschieden. Epistemische Unsicherheit resultiert aus Unkenntnis oder Unvollständigkeit des Modells. Diese könnte theoretisch durch mehr Trainingsdaten, ein besseres Modell oder bessere Wahl der Parameter reduziert werden. Aleatorische Unsicherheit resultiert aus Variabilität oder Rauschen in den Daten und ist nicht reduzierbar. [Wim+23]
Einige Methoden zur Abschätzung der Unsicherheit in Neuronalen Netzen sind: (1) Monte Carlo Dropout, wo die Vorhersage mehrfach simuliert wird, wobei Neuronen stochastisch deaktiviert werden [GG16], (2) Bayesian Neural Networks, die statt eines bestimmten Wertes eine Wahrscheinlichkeitsverteilung für jedes Modellgewicht lernen [Blu+15] und (3) Deep Ensembles, bei denen die Ausgabe von mehreren unabhängig trainierten Modellen betrachtet wird. [LPB17] Diese drei Methoden simulieren eine Variation der Modellparameter und schätzen somit in erster Linie epistemische Unsicherheit ab.
Diese Arbeit soll sich jedoch nicht auf neuronale Netze beschränken, sondern eher die Form einer Studie haben, wo die Unsicherheit in der Duplikaterkennung auf allgemeinerer Ebene experimentell quantifiziert wird. Hierzu ließ sich keine einschlägige Literatur finden.

3 Theoretisches Konzept

Um die Unsicherheit in der Duplikaterkennung angemessen untersuchen und bewerten zu können, bedarf es eines theoretischen Verständnisses der zugrunde liegenden Konzepte. Im Folgenden werden die zentralen Grundlagen definiert, auf denen diese Arbeit basiert: Der konzeptionelle Rahmen zur Quantifizierung von Unsicherheit, der Orakelklassifikation und der Konfidenzanalyse.

3.1 Quantifizierung der Unsicherheit

Die Unsicherheit der Duplikaterkennung wird in dieser Arbeit als Übereinstimmung beziehungsweise Nichtübereinstimmung der Vorhersagen der verschiedenen Modellinstanzen definiert.
Sei ein Datensatz von Paaren von Einträgen und eine Menge von Modellinstanzen gegeben. Jede Modellinstanz macht ihre Vorhersage auf dem gesamten Datensatz. Anschließend wird für jedes Paar im Datensatz bestimmt, wie viele Modellinstanzen dieses als Match identifiziert haben. Wird jedes Paar von 100% oder 0% der Modellinstanzen als Duplikat erkannt, beschreibt dies die größtmögliche Sicherheit. Wird jedes Paar von einem Anteil der Modellinstanzen nahe 50% als Duplikat erkannt, beschreibt dies die größtmögliche Unsicherheit. Die Anzahl der Datenpunkte wird also auf die Anteile der Modellinstanzen verteilt, die sie als Duplikat erkannt haben.

Formal sei c die Anzahl der Modellinstanzen, die ein Paar als Duplikat klassifizieren, und n die Gesamtzahl der Modellinstanzen. Der relative Anteil der positiven Klassifikationen ist damit $p = \frac{c}{n}$. Der Unsicherheitsindex U wird definiert als:

$$U = 1 - 2 \cdot |p - 0.5| \tag{3.1}$$

Dieser Index nimmt Werte zwischen 0 und 1 an. $U = 0$ beschreibt die maximale Sicherheit, wenn alle Modellinstanzen einstimmig entscheiden ($p = 0$ oder $p = 1$). $U = 1$ beschreibt die maximale Unsicherheit, wenn genau die Hälfte der Modellinstanzen ein Paar als Duplikat klassifiziert ($p = 0.5$).

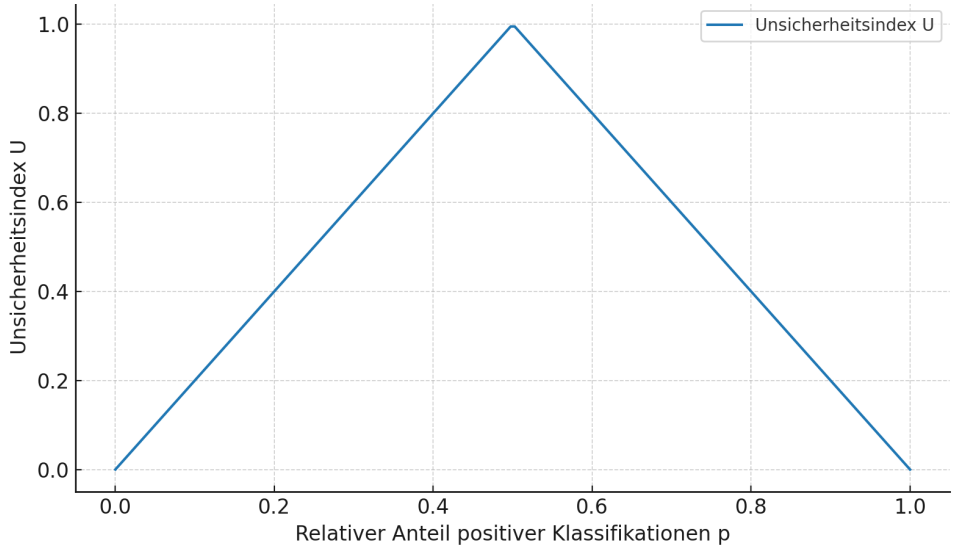

Abbildung 3.1: Unsicherheitsindex U als Funktion von p

3.2 Orakelklassifikation und Parameter t

Um die praktische Bedeutung von Unsicherheit zu untersuchen, wird untersucht, wie stark sich die Performance des Modells (gemessen in F1-Score) erhöht, wenn besonders unsichere Datenpunkte manuell klassifiziert werden. Statt einer manuellen Klassifikation wird ein Orakel verwendet, das die Labels der Daten kennt.
Es wird ein Parameter t definiert, der die Stärke der Abweichung zwischen den Modellinstanzen quantifiziert und somit steuert, welche Datensätze vom Orakel klassifiziert werden. Bei $t = x$ und n Durchläufen klassifiziert das Orakel alle Datenpunkte, die einen Unsicherheitsindex von mindestens $2x/n$ nach Gleichung 3.1 haben. Das entspricht dann allen Datenpunkten, bei denen die Minderheit in den Vorhersagen über alle Durchläufe mindestens x ist.
Somit gilt: Desto höher t, desto höher die durchschnittliche Unsicherheit der Daten, die vom Orakel klassifiziert werden.

3.3 Konfidenzanalyse

Ditto gibt für einen Datenpunkt einen Wert zwischen 0 und 1 aus. Je nachdem, ob dieser Wert über oder unter einem Threshold liegt, wird entschieden, ob es sich um ein Duplikat handelt. Der Threshold wird von Durchlauf zu Durchlauf unterschiedlich festgelegt.
Die Konfidenz von Ditto wird als Abstand von diesem Threshold definiert. Auf diese Weise erhalten wir ein Maß dafür, wie konfident Ditto in seiner Vorhersage ist, unabhängig davon, ob der Datenpunkt als Match klassifiziert wird oder nicht. Sei $prob$ der ausgegebene Wert von Ditto und $thresh$ der Threshold, dann ist die Konfidenz $conf$ definiert als:

$$conf = |prob - thresh| \tag{3.2}$$

4 Methodisches Vorgehen

Die Unsicherheit in der Duplikaterkennung wird in einem experimentellen Vergleich mit verschiedenen SOTA-Verfahren quantifiziert. Das methodische Vorgehen lässt sich in drei zentrale Phasen gliedern: (1) Auswahl und Aufsetzen geeigneter SOTA-Modelle, (2) Durchführung verschiedener Testszenarien zur Messung der Unsicherheit, (3) Analysen im Zusammenhang mit der gemessenen Unsicherheit und Interpretation der erzeugten Ergebnisse.

4.1 Aufsetzen der Verfahren

Die Verfahren Ditto[1], HierGAT[2] und Unicorn[3] werden aus den öffentlich verfügbaren Repositories aus den entsprechenden Publikationen entnommen und in Python aufgesetzt. Für die Experimente wurden die Verfahren so modifiziert, dass alle Ergebnisse der Vorhersagen auf den Testdaten in CSV-Dateien gespeichert werden. Außerdem wurde in Unicorn die Möglichkeit eingeführt, das Verfahren mit einem zufälligen Random-Seed auszuführen. Zusätzlich werden Random Forest[4] und XGBoost[5] mithilfe der scikit-learn-Bibliothek aufgesetzt. Alle Verfahren werden in einer einheitlichen Testumgebung auf Benchmark-Datensätzen ausgeführt.

4.2 Unsicherheitsmessung in verschiedenen Szenarien

Die Unsicherheit wird als Stabilität der Vorhersagen in verschiedenen Testszenarien untersucht. In jedem Testszenario werden 9 unabhängige Durchläufe verglichen. Mit 9 Durchläufen ist auch eine Analyse eines Mixes der 3 Hauptverfahren möglich, indem je 3 Durchläufe gewählt, alle Durchläufe zusammengenommen werden und somit wieder 9 Durchläufe zur Analyse zur Verfügung stehen. Für jedes Szenario wird der über alle Datenpunkte durchschnittliche Unsicherheitsindex nach Gleichung 3.1 berechnet. Die untersuchten Szenarien sind Folgende:

(a) **Ausführung auf derselben GPU**
Jedes Verfahren wird mehrfach auf derselben Hardwareumgebung mit derselben GPU und identischen Parametern trainiert und ausgeführt. Die verschiedenen Vorhersagen werden hinsichtlich ihrer Übereinstimmung und Nichtübereinstimmung analysiert. In diesem Szenario lässt sich die Unsicherheit der verschiedenen Verfahren vergleichen. Als Parameter wurden die laut Verfahren vorgeschlagenen Standard-Parameter verwendet. Diese sind:

- Ditto: Batch Size: 32, Learning Rate: 3e-05, Epochs: 30, Data Augmentation: swap, Domain Knowledge: general, Language Model: RoBERTa

1. https://github.com/megagonlabs/ditto
2. https://github.com/CGCL-codes/HierGAT
3. https://github.com/ruc-datalab/Unicorn
4. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
5. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

- HierGAT: Batch Size: 16, Learning Rate: 1e-05, Length: 512, Language Model: RoBERTa
- Unicorn: Batch Size: 32, Learning Rate: 3e-06, Length: 128, Model: DeBERTa-base

(b) **Ausführung auf unterschiedlichen GPUs des gleichen Mddells**
Ditto wird auf verschiedenen GPUs des gleichen Modells trainiert und ausgeführt. Ziel ist es, potenzielle Abweichungen in den Vorhersagen zu identifizieren, die sich aus Hardwareeffekten ergeben. Es wird auf verschiedenen Knoten des Clusters und somit auf unterschiedlichen GPUs trainiert. Bei Ditto fungiert die Run-Id als Random-Seed. Um sicherzustellen, dass nur der Einfluss der Hardware-Variation beobachtet wird, werden dieselben Run-Ids wie in (a) verwendet.

(c) **Variation von Hyperparametern**
Ditto wird mit veränderten Parametern trainiert und evaluiert. Hier kann die Sensitivität gegenüber der Parameterwahl als Quelle der auftretenden Unsicherheit bestimmt werden.
Die getesteten Parameterkonfigurationen sind (nicht genannte Parameter entsprechen der Standard-Konfiguration):

- 1.Konfiguration: Batch Size 16 statt 32
- 2. Konfiguration: Learning Rate 1e-05 statt 3e-05
- 3. Konfiguration: no Data Augmentation statt Data Augmentation swap
- 4. Konfiguration: no Domain Knowledge statt Domain Knowledge general
- 5. Konfiguration: No Data Augmentation, No Domain Knowledge statt Data Augmentation swap, Domain Knowledge general
- 6. Konfiguration: Language Model BERT statt RoBERTa
- 7. Konfiguration: Language Model DistilBERT statt RoBERTa
- 8. Konfiguration: 30 Epochs statt 15

Außerdem wird zusätzlich pro Datensatz der erste Durchlauf jeder Parameter-Konfiguration, einschließlich der Standard-Konfiguration, entnommen und zusammengeführt. So erhalten wir wieder 9 Durchläufe, anhand welcher der Unsicherheitsindex berechnet wird.

(d) **Vergleich verschiedener Verfahren**
Unterschiedliche EM-Verfahren werden auf denselben Datensätzen angewendet. Die Übereinstimmung und Abweichung ihrer Vorhersagen wird analysiert, somit können modellbedingte Unterschiede quantifiziert werden. Hierzu werden die ersten drei Durchläufe jedes Verfahrens in Standard-Konfigurationen aus (a) entnommen und kombiniert. So erhalten wir wieder 9 Durchläufe, sodass die Analyse des Unsicherheitsindexes vergleichbar ist.

4.3 Analysen

Auf Grundlage des bestimmten Unsicherheitsindexes werden weitere Analysen der Daten durchgeführt, die Aufschluss über die Aussagekraft und praktische Relevanz des Unsicherheitsindexes geben sollen.

- **Orakelanalyse:** Für einige ausgewählte Szenarien wird analysiert, wie gut die Performance nach F1-Score ist, wenn man die 9 Durchläufe als Ensemble betrachtet und die Klassifikationen nach Mehrheitsvotum entschieden werden. Dann wird untersucht, wie stark sich die Performance gegenüber dem Ensemble-Fall verbessert, wenn das Orakel unsichere Datenpunkte für verschiedene t-Werte klassifiziert. Schließlich wird analysiert, ob es einen Zusammenhang zwischen der Unsicherheit der vom Orakel klassifizierten Datenpunkte und dem Performance-Gewinn pro Klassifikation durch das Orakel gibt.
- **Konfidenzanalyse:** Für einige Szenarien wird analysiert, wie die Konfidenz nach Gleichung 3.2 für alle Datenpunkte in einem Durchlauf ist, also wie sicher sich Ditto bei der Klassifikation ist. Da der Schwellenwert bei der Klassifikation bei Ditto von Durchlauf zu Durchlauf unterschiedlich gewählt wird, ist zu erwarten, dass dieser je nach Durchlauf deutlich verschieden sein kann. Deshalb werden die Konfidenzwerte nur aus dem ersten Durchlauf für die Analyse genommen. Anschließend wird untersucht, ob dieser Konfidenzwert mit dem über alle 9 Durchläufe bestimmten Unsicherheitsindex zusammenhängt. So soll die Frage beantwortet werden, ob die Konfidenz ein geeigneter Prädiktor für die experimentell ermittelbare Unsicherheit ist.
- **Schnittmengenanalyse:** Die Schnittmengen der positiven Duplikat-Klassifikationen der 5 Verfahren werden in Venn-Diagrammen aufgeführt und auf Auffälligkeiten analysiert.
- **Untersuchung besonders unsicherer Daten:** Es werden Daten extrahiert, bei denen es besonders häufig Uneinigkeiten zwischen den Durchläufen gab. Diese Daten werden auf gemeinsame Charakteristiken untersucht. Zuerst werden die Daten manuell betrachtet und anschließend an ChatGPT zur Analyse übergeben.

4.4 Datensätze

Für die Experimente werden öffentlich zugängliche Entity-Matching-Datensätze verwendet, die vorzugsweise in den Veröffentlichungen der jeweiligen EM-Verfahren genutzt wurden. Die Datensätze wurden so ausgewählt, dass sie eine große Bandbreite an erreichter Performance abdecken und somit als verschieden schwierig angesehen werden können. Es wurden fünf Datensätze ausgewählt:

- Amazon-Google aus dem ER-Magellan-Datensatz[6]
- DBLP-ACM aus dem ER-Magellan-Datensatz
- Walmart-Amazon aus dem ER-Magellan-Datensatz
- Abt-Buy aus dem ER-Magellan-Datensatz
- Organizations[7], welcher aus Daten von Wikidata erstellt wurde

6. https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md
7. https://github.com/dekelcohen/ditto/tree/master/data/wiki/ORG_new_comp_org

4.5 Systemumgebung

Die Implementierung erfolgt auf dem Cluster „Levante“[8] des Deutschen Klimarechenzentrums (DKRZ) in Python unter Verwendung von gängigen Bibliotheken wie PyTorch, Transformers (Hugging Face), scikit-learn und Pandas. Die Modelle werden auf den GPU-Knoten trainiert, die mit NVIDIA A100 GPUs ausgestattet sind.

8. https://docs.dkrz.de/doc/levante/configuration.html

5 Ergebnisse

5.1 Messung des Unsicherheitsindexes

In diesem Abschnitt werden die Ergebnisse der im Abschnitt 5.2 beschriebenen Analysen vorgestellt.

5.1.1 Ausführung auf derselben GPU

Die vollständigen Performance-Metriken der einzelnen Durchläufe können im Appendix in der Tabelle A.1 eingesehen werden.

In der Tabelle 5.1 sind die minimalen, maximalen und mittleren F1-Scores der Verfahren in den 9 Durchläufen dargestellt.

Tabelle 5.1: F1-Score: Minimum, Maximum und Mittelwert je Verfahren und Datensatz

Verfahren	**Datensatz**	**min. F1**	**max. F1**	**mittl. F1**
Ditto	Abt-Buy	0,815	0,905	0,876
Ditto	Amazon-Google	0,720	0,765	0,750
Ditto	DBLP-ACM	0,983	0,991	0,986
Ditto	Organizations	0,831	0,843	0,835
Ditto	Walmart-Amazon	0,839	0,892	0,862
HierGAT	Abt-Buy	0,799	0,901	0,880
HierGAT	Amazon-Google	0,726	0,764	0,743
HierGAT	DBLP-ACM	0,984	0,990	0,987
HierGAT	Organizations	0,806	0,836	0,823
HierGAT	Walmart-Amazon	0,806	0,841	0,829
Unicorn	Abt-Buy	0,906	0,924	0,916
Unicorn	Amazon-Google	0,718	0,768	0,742
Unicorn	DBLP-ACM	0,982	0,990	0,987
Unicorn	Organizations	0,855	0,868	0,860
Unicorn	Walmart-Amazon	0,839	0,882	0,862

Im Allgemeinen erzielen die drei Verfahren auf den Test-Datensätzen eine ähnliche Performance. Je nach Datensatz unterscheidet sich das beste Verfahren und wie groß die Unterschiede sind. Im Ganzen ist die Performance von Unicorn sehr stark, während HierGAT etwas schwächer ist.

Die Übereinstimmung der einzelnen Durchläufe sowie den berechneten durchschnittlichen Unsicherheitsindex nach Gleichung 3.1 können in den Plots im Appendix im Abschnitt B

eingesehen werden. Zur Lesbarkeit wurde in allen Plots der Balken zu den Tupel-Paaren, die nie als Duplikat klassifiziert wurden, entfernt.

In folgender Tabelle sind die experimentell bestimmten Unsicherheitsindizes zusammengefasst.

Verfahren	Abt-Buy	Amazon-Google	DBLP-ACM	Organizations	Walmart-Amazon
Ditto	0,025	0,038	0,004	0,048	0,016
HierGAT	0,031	0,044	0,004	0,047	0,023
Unicorn	0,013	0,032	0,012	0,027	0,011

Tabelle 5.2: Durchschnittlicher Unsicherheitsindex nach jeweils 9 Durchläufen

In vier von fünf Fällen ist Unicorn das Verfahren mit der geringsten Unsicherheit. Nur auf dem DBLP-ACM-Datensatz hat es die höchste Unsicherheit. HierGAT ist das Verfahren mit der stärksten Unsicherheit. Auf drei von fünf Datensätzen erzielt es den höchsten Unsicherheitsindex und ist auf dem Organizations-Datensatz nur knapp hinter Ditto. Die Unsicherheit über alle Datensätze gemittelt ist bei Ditto 0,026, bei HierGAT 0,030 und bei Unicorn 0,019.
Wie hoch die Unsicherheit der Verfahren ist, hängt maßgeblich vom Datensatz ab. Während die mittlere Unsicherheit über alle Verfahren auf dem Amazon-Google-Datensatz 0,038 ist, ist sie auf dem DBLP-ACM-Datensatz nur etwa 0,007. Es gibt deutliche Unterschiede in den Spannweiten des erreichten Unsicherheitsindizes der verschiedenen Datensätze. Sie ist bei DBLP-ACM mit 0,008 am geringsten und bei Organizations mit 0,021 am höchsten.

5.1.2 Ausführung auf unterschiedlichen GPUs des gleichen Modells

Datensatz	Durchschn. Unsicherheitsindex	Differenz zu gleicher HW
Abt-Buy	0,025	0,000
Amazon-Google	0,038	0,000
DBLP-ACM	0,004	0,000
Organizations	0,048	0,000
Walmart-Amazon	0,016	0,000

Tabelle 5.3: Unsicherheitsindex nach Ausführung auf verschiedenen GPUs

Die Durchläufe auf verschiedenen GPUs ergaben exakt die gleichen Klassifikationen, wie die Durchläufe auf der gleichen GPU. Somit sind Precision, Recall und F1-Score sowie der durchschnittliche Unsicherheitsindex exakt die gleichen wie bei der Ausführung aller Durchläufe auf demselben Knoten. Die Variation der Hardware hat in diesem Setting also keinen Einfluss auf den Unsicherheitsindex.

5.1.3 Variation von Hyperparametern

Die vollständigen Ergebnisse der Durchläufe mit den einzelnen Konfigurationen können in der Tabelle A.2 eingesehen werden. Hier wird auch die Differenz zum entsprechenden mittleren F1-Score aus Tabelle 5.1 gegeben.

Das Ändern der Batch Size und die Verwendung von BERT oder DistilBERT hat auf jedem Datensatz über die Durchläufe gemittelt zu einem schlechteren F1-Score als mit der Standard-Konfiguration geführt. Die Änderungen der anderen Parameter führten nicht einheitlich zu einem besseren oder schlechteren F1-Score.

	Abt–Buy			Amazon–Google		
Konfiguration	U-Index	Δ vs. Standard	Δ in %	U-Index	Δ vs. Standard	Δ in %
batch size	0,033	0,008	32,00	0,043	0,005	13,16
bert	0,030	0,005	20,00	0,037	−0,001	−2,63
distilbert	0,036	0,011	44,00	0,040	0,002	5,26
epochs	0,021	−0,004	−16,00	0,040	0,002	5,26
learning rate	0,021	−0,004	−16,00	0,041	0,003	7,89
no da	0,028	0,003	12,00	0,043	0,005	13,16
no da + dk	0,018	−0,007	−28,00	0,033	−0,005	−13,16
no dk	0,023	−0,002	−8,00	0,035	−0,003	−7,89
Konfig.-Mix	0,030	0,005	20,00	0,048	0,010	26,32

Tabelle 5.4: Unsicherheitsindex nach Parameter-Varianten (Teil 1): Abt-Buy und Amazon-Google

	DBLP–ACM			Organizations		
Konfiguration	U-Index	Δ vs. Standard	Δ in %	U-Index	Δ vs. Standard	Δ in %
batch size	0,005	0,001	25,00	0,049	0,001	2,08
bert	0,004	0,000	0,00	0,044	−0,004	−8,33
distilbert	0,004	0,000	0,00	0,046	−0,002	−4,17
epochs	0,004	0,000	0,00	0,048	0,000	0,00
learning rate	0,002	−0,002	−50,00	0,041	−0,007	−14,58
no da	0,003	−0,001	−25,00	0,044	−0,004	−8,33
no da + dk	0,005	0,001	25,00	0,050	0,002	4,17
no dk	0,005	0,001	25,00	0,045	−0,003	−6,25
Konfig.-Mix	0,004	0,000	0,00	0,064	0,016	33,33

Tabelle 5.5: Unsicherheitsindex nach Parameter-Varianten (Teil 2): DBLP–ACM und Organizations

Die größten absoluten Effekte der Parameter-Änderungen (Mix der Konfigurationen ausgenommen), in Bezug auf den Unsicherheitsindex, treten bei Abt-Buy (Spannweite 0,018) und Walmart-Amazon (Spannweite 0,013) auf, während DBLP-ACM mit 0,03 in absoluten Werten am unempfindlichsten ist. Die stärkste relative Erhöhung der Unsicherheit zeigt die Verwendung von Distiblert auf Walmart-Amazon (+81,25%), die stärkste relative Senkung ergibt sich auf DBLP-ACM durch die geänderte Lernrate (-50%). Je Datensatz variiert der durchschnittliche Unsicherheitsindex zwischen 0,018 und 0,036 (Abt-Buy), 0,033 und 0,043 (Amazon-Google), 0,002 und 0,005 (DBLP-ACM), 0,041 und 0,050 (Organizations) sowie 0,016 und 0,029 (Walmart-Amazon).
Das Ändern der Batch Size führt zu einem höheren durchschnittlichen Unsicherheitsindex auf allen Datensätzen. Bei allen anderen Änderungen der Parameter gibt es keine einheitliche

	Walmart–Amazon		
Konfiguration	U-Index	Δ vs. Std.	Δ in %
batch size	0,025	0,009	56,25
bert	0,025	0,009	56,25
distilbert	0,029	0,013	81,25
epochs	0,022	0,006	37,50
learning rate	0,022	0,006	37,50
no da	0,018	0,002	12,50
no da + dk	0,021	0,005	31,25
no dk	0,016	0,000	0,00
Konfig.-Mix	0,028	0,012	75,00

Tabelle 5.6: Unsicherheitsindex nach Parameter-Varianten (Teil 3): Walmart–Amazon

Erhöhung oder Verringerung des Unsicherheitsindexes. Bei der Verwendung von DistilBERT jedoch tritt in vier von fünf Fällen ein höherer oder gleichbleibender Unsicherheitsindex auf. Der Mix der Konfigurationen führt immer zu einer gesteigerten oder gleichbleibenden Unsicherheit, verglichen mit der Standard-Konfiguration. Besonders deutlich wird dies auf dem Amazon-Google- und Organizations-Datensatz, wo der Mix der Konfigurationen eine höhere Unsicherheit erreicht als jede Konfiguration für sich. Der Unsicherheitsindex von Organizations ist mit 0,064 sogar der höchste gemessene Wert in allen Experimenten dieser Arbeit.

5.1.4 Vergleich verschiedener Verfahren

In Tabelle 5.7 ist der berechnete Unsicherheitsindex aus den 9 Durchläufen mit der Kombination der 3 Verfahren dargestellt. Zum Vergleich sind in den ersten 3 Zeilen nochmal die Ergebnisse aus Abschnitt 5.1.1 gegeben.

Verfahren	Abt-Buy	Amazon-Google	DBLP-ACM	Organizations	Walmart-Amazon
Ditto	0,025	0,038	0,004	0,048	0,016
HierGAT	0,031	0,044	0,004	0,047	0,023
Unicorn	0,013	0,032	0,012	0,027	0,011
Mix	0,031	0,052	0,006	0,062	0,028

Tabelle 5.7: Durchschnittlicher Unsicherheitsindex nach jeweils 9 Durchläufen und dem Mix der 3 Verfahren

Für vier von fünf Datensätzen ergibt sich aus dem Mix der Verfahren eine höhere oder gleichbleibende Unsicherheit verglichen mit den Unsicherheiten der einzelnen Verfahren. Nur auf dem DBLP-ACM-Datensatz ist die Unsicherheit aus dem Mix der Verfahren mit 0,006 geringer als die von Unicorn mit 0,012.

5.2 Verbesserung der Ergebnisse durch Orakelbefragung

Für die 9 Durchläufe mit Ditto mit Standard-Konfiguration wurde berechnet, wie Precision, Recall und F1-Score ausfallen, wenn man den Durchschnitt der Metriken aus den Durchläufen nimmt, wenn man die Daten per Mehrheitsvotum klassifiziert und wenn man einen Teil der Daten für verschiedene Parameter t vom Orakel klassifizieren lässt. Im Orakel-Fall werden die Daten, die nicht vom Orakel klassifiziert wurden, ebenfalls per Mehrheitsvotum entschieden. Zum Vergleich wurde zusätzlich berechnet, wie der F1-Score ausfällt, wenn ein anderes Orakel die gleiche Anzahl an Daten klassifiziert, wie das erste Orakel, die Auswahl der Daten aber zufällig ist. Hierzu wurde die Berechnung 50-mal für verschiedene Zufallsorakel durchgeführt und der Mittelwert der Metriken genommen.
Die Ergebnisse sind in den Tabellen 5.8 bis 5.12 zusammengefasst. Die Spalte „Anteil Orakel in %“ zeigt den Anteil der Testdaten, die durch das Orakel klassifiziert wurden, sowie die absolute Anzahl in Klammern. Die Spalte „$\Delta F1$ zu Mehrh.“ zeigt die Differenz vom F1-Score des Orakel-Modus zum F1-Score des Mehrheitsvotum-Modus. Die Spalte „$\Delta F1$ zu Zufallsor.“ zeigt die Differenz vom F1-Score des Orakel-Modus zum Mittelwert des F1-Scores des Zufallsorakels.

Modus	Precision	Recall	$F1$	Anteil Orakel in % (absolut)	$\Delta F1$ zu Mehrh.	$\Delta F1$ zu Zufallsor.
Mittelwert	0,883	0,869	0,876	—	—	—
Mehrheitsvot.	0,914	0,879	0,896	—	—	—
Orakel $t = 1$	0,980	0,961	0,971	6,0 (115)	0,075	0,068
Orakel $t = 2$	0,960	0,927	0,943	2,8 (53)	0,047	0,045
Orakel $t = 3$	0,945	0,913	0,928	1,7 (32)	0,032	0,031
Orakel $t = 4$	0,929	0,883	0,905	0,6 (12)	0,009	0,009

Tabelle 5.8: Orakelanalyse auf Abt-Buy

Modus	Precision	Recall	$F1$	Anteil Orakel in % (absolut)	$\Delta F1$ zu Mehrh.	$\Delta F1$ zu Zufallsor.
Mittelwert	0,708	0,801	0,751	—	—	—
Mehrheitsvotum	0,724	0,829	0,773	—	—	—
Orakel $t = 1$	0,868	0,927	0,897	8,3 (191)	0,124	0,105
Orakel $t = 2$	0,813	0,893	0,851	4,8 (110)	0,078	0,068
Orakel $t = 3$	0,793	0,868	0,829	2,8 (65)	0,056	0,050
Orakel $t = 4$	0,755	0,855	0,802	1,2 (27)	0,029	0,026

Tabelle 5.9: Orakelanalyse auf Amazon-Google

Modus	Precision	Recall	$F1$	Anteil Orakel in % (absolut)	$\Delta F1$ zu Mehrh.	$\Delta F1$ zu Zufallsor.
Mittelwert	0,980	0,991	0,986	—	—	—
Mehrheitsvotum	0,982	0,991	0,987	—	—	—
Orakel $t = 1$	0,998	0,998	0,998	1,1 (26)	0,011	0,011
Orakel $t = 2$	0,995	0,995	0,995	0,6 (16)	0,009	0,009
Orakel $t = 3$	0,982	0,995	0,989	0,2 (6)	0,002	0,002
Orakel $t = 4$	0,982	0,993	0,988	0,0 (1)	0,001	0,001

Tabelle 5.10: Orakelanalyse auf DBLP-ACM

Modus	Precision	Recall	$F1$	Anteil Orakel in % (absolut)	$\Delta F1$ zu Mehrh.	$\Delta F1$ zu Zufallsor.
Mittelwert	0,810	0,863	0,835	—	—	—
Mehrheitsvotum	0,836	0,874	0,855	—	—	—
Orakel $t = 1$	0,948	0,943	0,945	10,7 (4966)	0,091	0,075
Orakel $t = 2$	0,915	0,921	0,918	5,9 (2732)	0,064	0,055
Orakel $t = 3$	0,888	0,906	0,897	3,4 (1569)	0,043	0,038
Orakel $t = 4$	0,862	0,890	0,876	1,6 (722)	0,021	0,019

Tabelle 5.11: Orakelanalyse auf Organizations

Modus	Precision	Recall	$F1$	Anteil Orakel in % (absolut)	$\Delta F1$ zu Mehrh.	$\Delta F1$ zu Zufallsor.
Mittelwert	0,893	0,837	0,864	—	—	—
Mehrheitsvotum	0,922	0,860	0,890	—	—	—
Orakel $t = 1$	0,966	0,896	0,930	3,9 (79)	0,040	0,035
Orakel $t = 2$	0,949	0,876	0,911	1,9 (39)	0,021	0,019
Orakel $t = 3$	0,938	0,865	0,900	0,9 (18)	0,010	0,009
Orakel $t = 4$	0,922	0,860	0,890	0,4 (8)	0,000	0,000

Tabelle 5.12: Orakelanalyse auf Walmart-Amazon

Das Orakel erreicht für alle Werte t einen höheren oder gleichen F1-Score als im Mehrheitsvotum und wie beim Zufallsorakel.

In den Abbildungen 5.2 bis 5.6 wird der Parameter t gegen den Performance-Gewinn pro Orakel-Klassifikation nach F1-Score im Vergleich zum Mehrheitsvotum-Modus abgebildet und der Pearson-Korrelationskoeffizient zwischen den beiden Größen berechnet. Zu erwarten ist eine positive Korrelation, also dass mit der steigenden Unsicherheit der Daten, die durch das Orakel klassifiziert werden, der Performancegewinn, den eine einzelne Klassifikation durch das Orakel bringt, steigt. Dies ist zu erwarten, da die Daten, wo sich Ditto öfter selbst widerspricht, schwierig durch Verfahren zu klassifizieren sind und so eine manuelle Klassifikation besonders hilfreich ist. Ein generisches zu erwartendes Ergebnis ist in Abbildung 5.1 dargestellt.

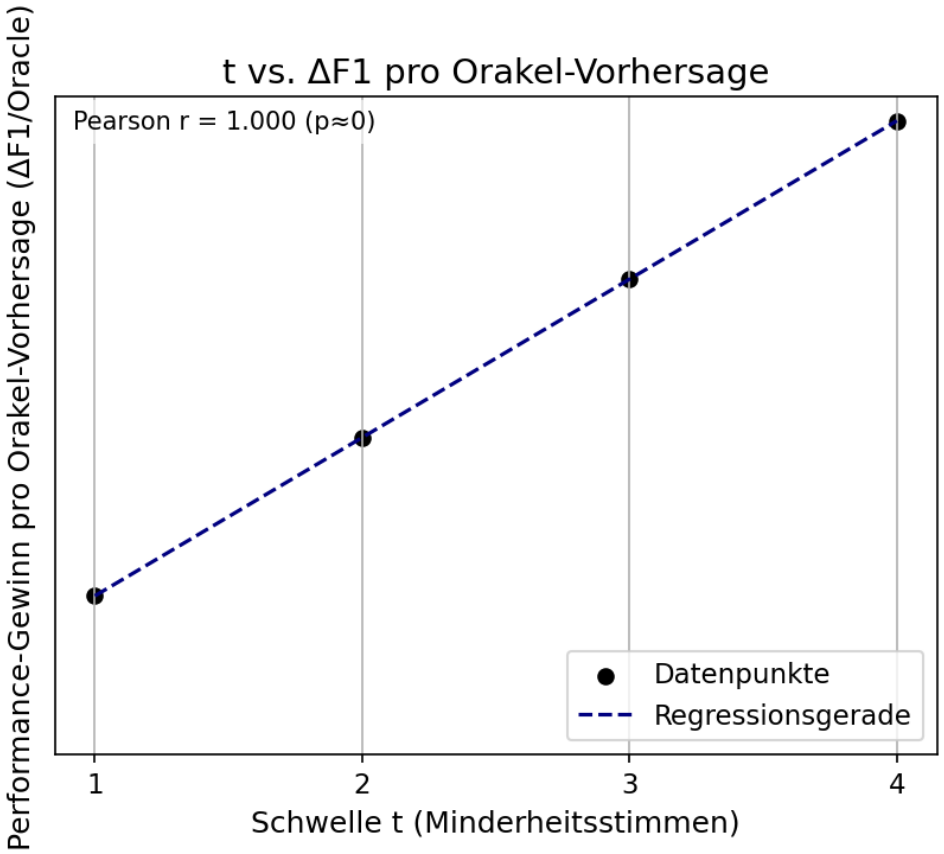

Abbildung 5.1: t gegen den Performance-Gewinn pro Orakel-Klassifikation (erwatetes Ergebnis)

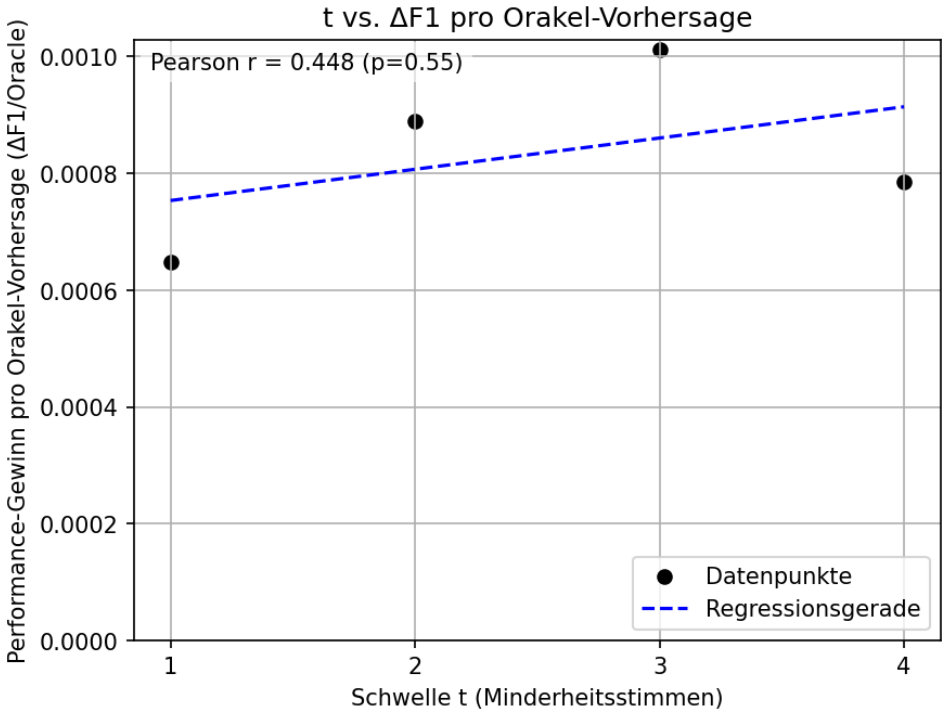

Abbildung 5.2: *t* gegen den Performance-Gewinn pro Orakel-Klassifikation bei Abt-Buy

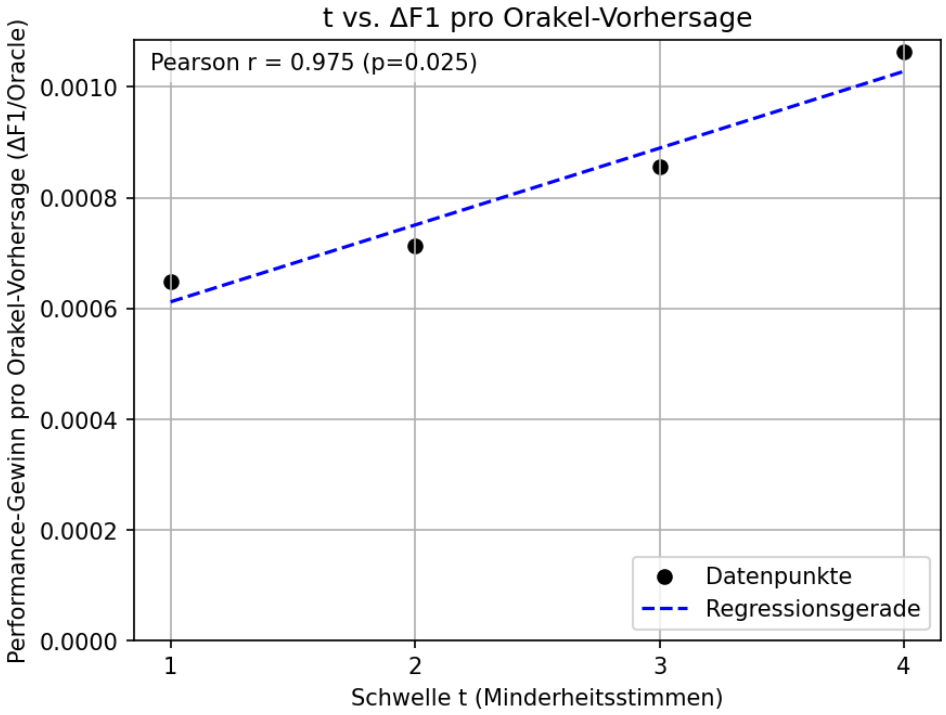

Abbildung 5.3: *t* gegen den Performance-Gewinn pro Orakel-Klassifikation bei Amazon-Google

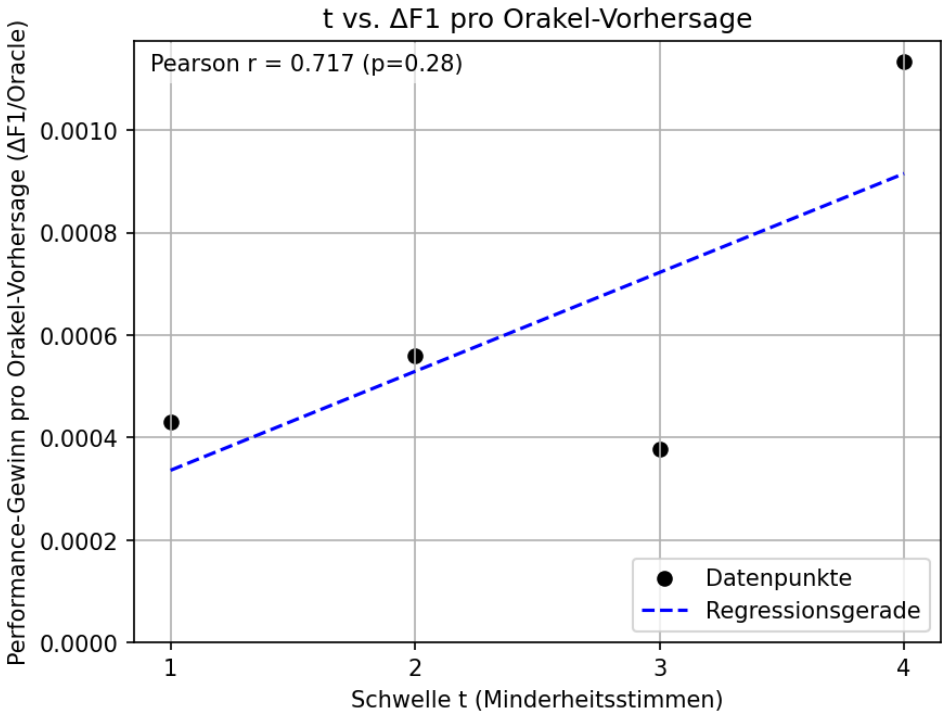

Abbildung 5.4: *t* gegen den Performance-Gewinn pro Orakel-Klassifikation bei DBLP-ACM

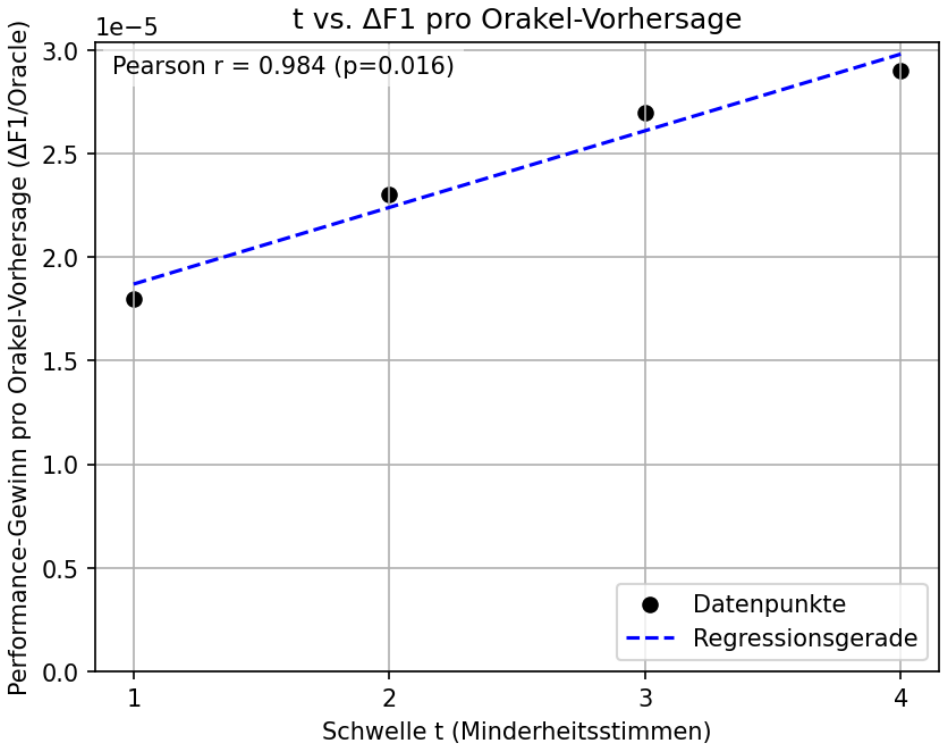

Abbildung 5.5: *t* gegen den Performance-Gewinn pro Orakel-Klassifikation bei Organizations

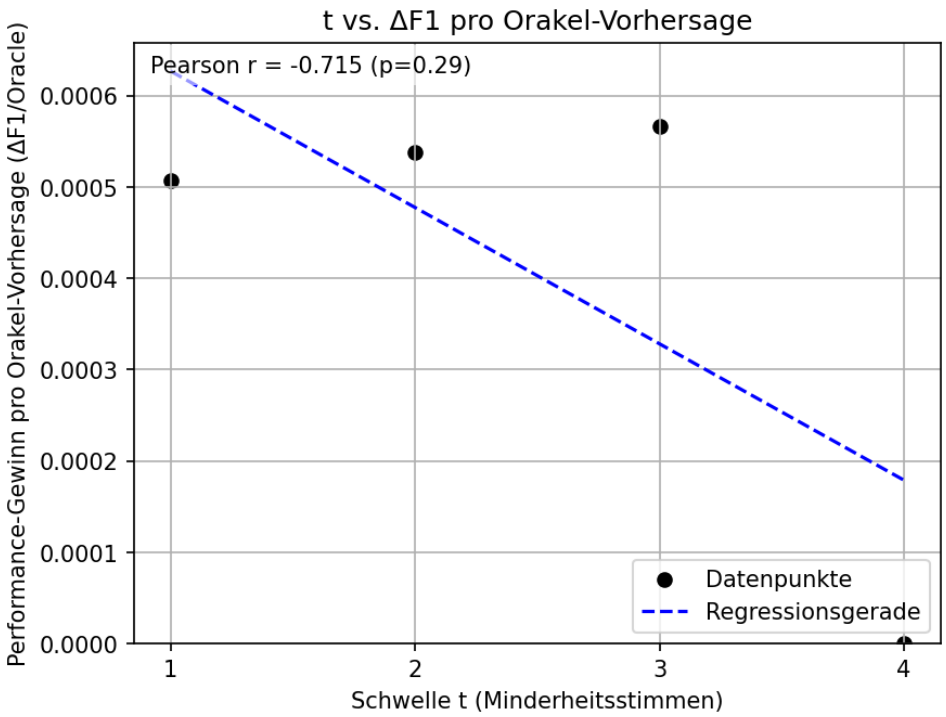

Abbildung 5.6: t gegen den Performance-Gewinn pro Orakel-Klassifikation bei Walmart-Amazon

Bei vier von fünf Datensätzen zeigt die Analyse eine positive Korrelation zwischen t und dem Performance-Gewinn pro Orakel-Klassifikation. Nur auf Walmart-Amazon ergibt sich ein negativer Korrelationskoeffizient.

5.3 Schnittmengen in den Vorhersagen der Verfahren

Die Venn-Diagramme zu den Schnittmengen der Duplikatklassifikationen der drei diskutierten Verfahren und Random Forest und XGBoost sind den Abbildungen 5.7 bis 5.11 zu entnehmen.

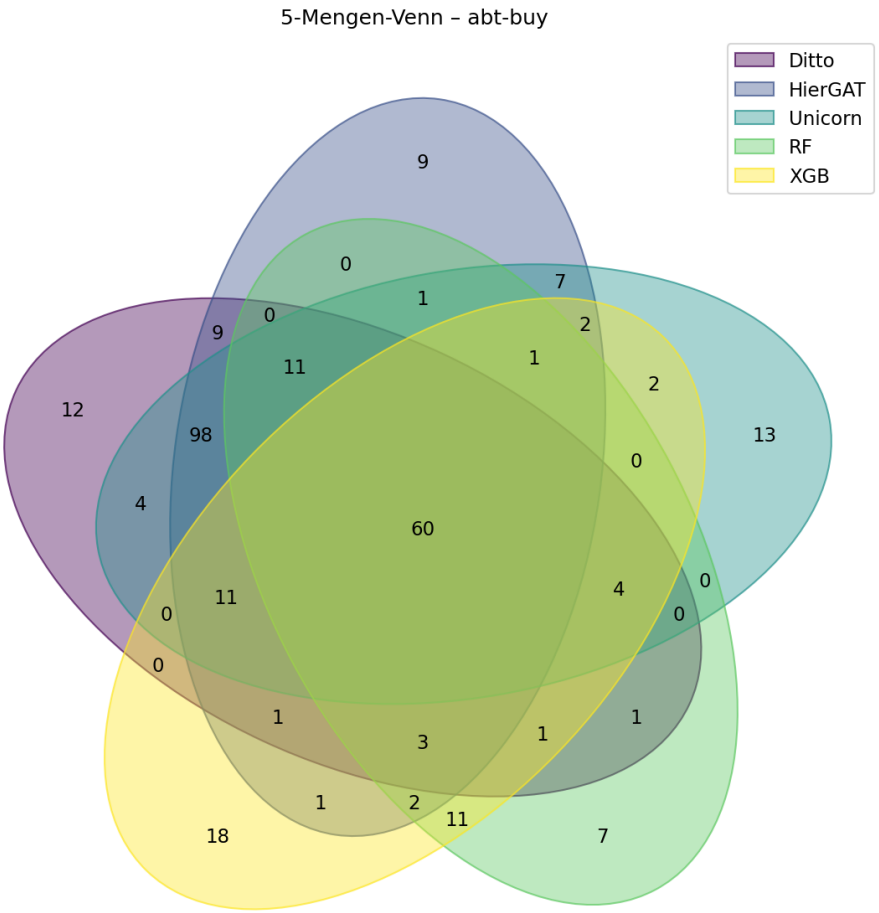

Abbildung 5.7: Venn Digramm der Duplikatklassifikationen auf Abt-Buy

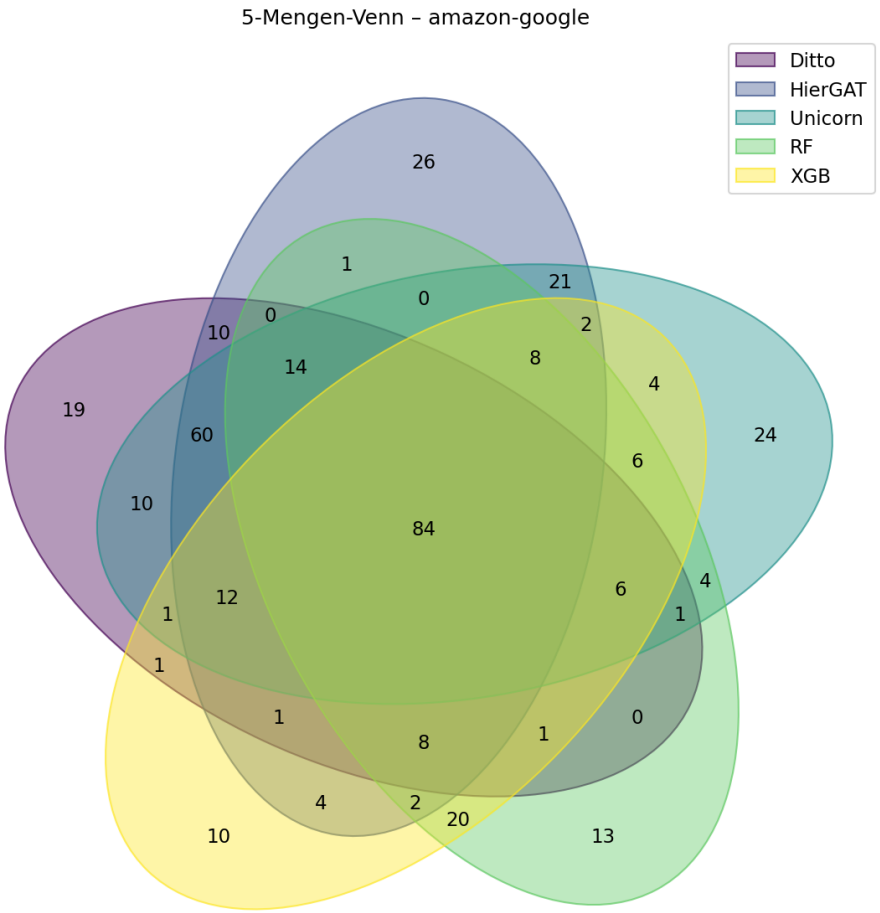

Abbildung 5.8: Venn Digramm der Duplikatklassifikationen auf Amazon-Google

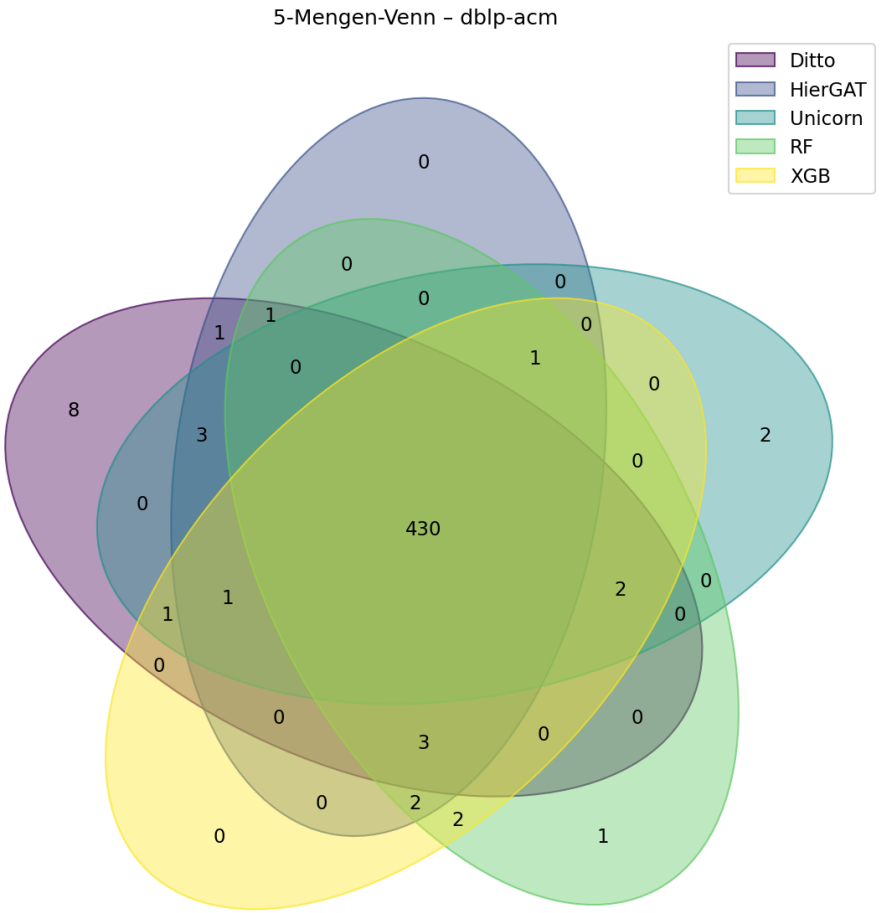

Abbildung 5.9: Venn Digramm der Duplikatklassifikationen auf DBLP-ACM

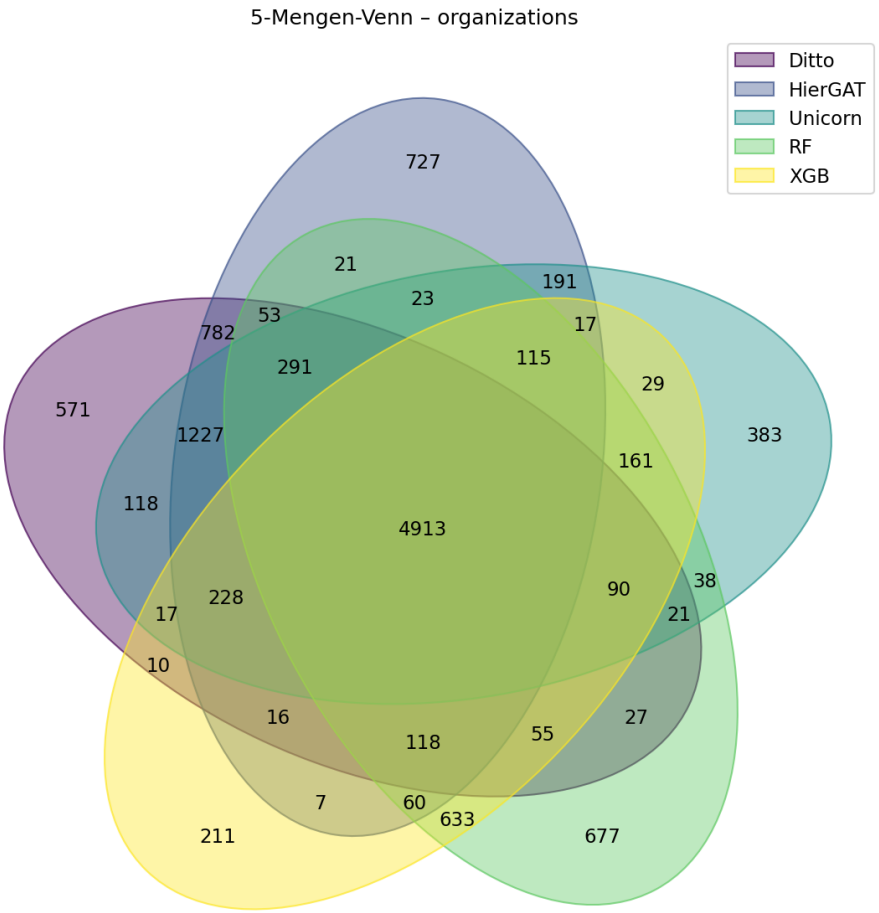

Abbildung 5.10: Venn Digramm der Duplikatklassifikationen auf Organizations

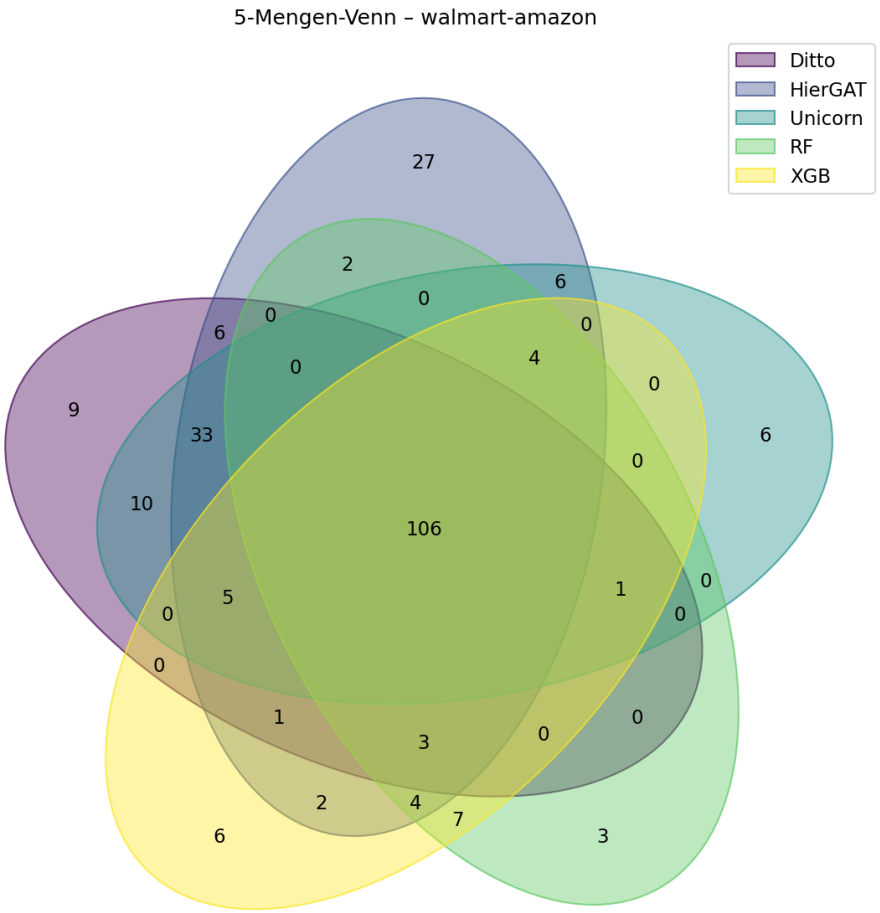

Abbildung 5.11: Venn Digramm der Duplikatklassifikationen auf Walmart-Amazon

Die Schnittmengenanalyse liefert je nach Datensatz ein anderes Bild. Auf DBLP-ACM zeigt sich zwischen allen Verfahren eine starke Kohärenz. 430 Datenpunkte wurden von allen Verfahren als Duplikat klassifiziert. Daneben gibt es kaum weitere nennenswerte Schnittmengen. Lediglich Ditto klassifiziert 8 Datenpunkte als Duplikat, die von keinem anderen Verfahren als Duplikat klassifiziert werden. Diese Ergebnisse passen zur hohen Performance aller Verfahren und dem niedrigen Unsicherheitsindex auf DBLP-ACM. Es ist ein einfacher Datensatz, der von allen Verfahren gut klassifiziert wird, somit gibt es wenig Uneinigkeiten.

Auf Abt-Buy zum Beispiel sehen die Ergebnisse deutlich anders aus. Die größte Schnittmenge mit 98 Datenpunkten ist die Menge, die nur Ditto, HierGAT und Unicorn, also den Verfahren, die auf LLMs basieren, als Duplikat klassifiziert wurden. Von allen Verfahren wurden 60 Datenpunkte als Duplikat klassifiziert. Diese Beobachtung ist damit zu erklären, dass die klassischen ML-Verfahren sehr schlecht auf diesem Datensatz performen. Sie erreichen einen F1-Score von etwa 0,5 mit einem Recall von nur etwa 0,36-0,39. Die LLM-basierten Verfahren erreichen einen F1-Score von über 0,8.

Auf den übrigen Datensätzen zeigt sich ein ähnliches Bild, wie auf Abt-Buy, wobei die beschriebenen Effekte teilweise oder deutlich schwächer sind. Es lässt sich ein „Kern“ von Daten beobachten, der von allen Verfahren als Duplikat klassifiziert wird, eine weitere große Menge von Daten, die nur von den LLM-basierten Verfahren als Duplikat klassifiziert wird, und einige Daten, die pro Verfahren spezifisch klassifiziert werden.

5.4 Zusammenhang zwischen Modellkonfidenz und Unsicherheit

In den Abbildungen 5.13 bis 5.17 wurde die Konfidenz nach Gleichung 3.2 des ersten Durchlaufs von Ditto in Standard-Konfiguration gegen die über alle 9 Durchläufe bestimmte Unsicherheit nach Gleichung 3.1 aufgetragen und der Pearson-Korrelationskoeffizient der beiden Größen berechnet.
Es ist zu erwarten, dass sich eine negative Korrelation zeigt, also dass mit steigender Konfidenz in der Vorhersage die empirisch berechnete Unsicherheit sinkt, wie in Abbildung 5.12 dargestellt.

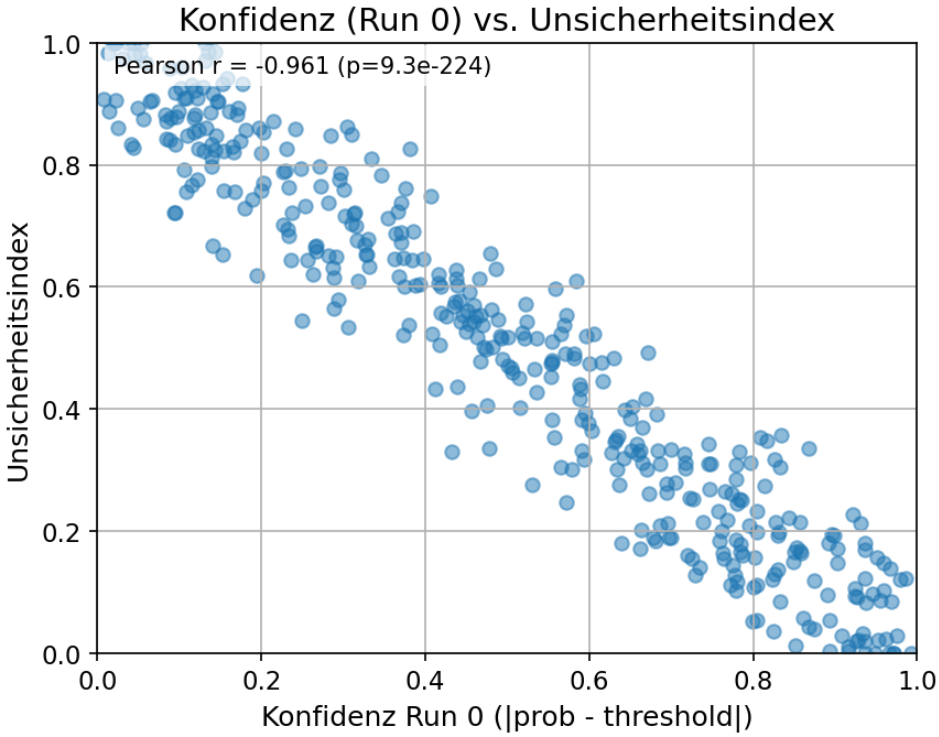

Abbildung 5.12: Konfidenz gegen Unsicherheit (erwartetes Ergebnis)

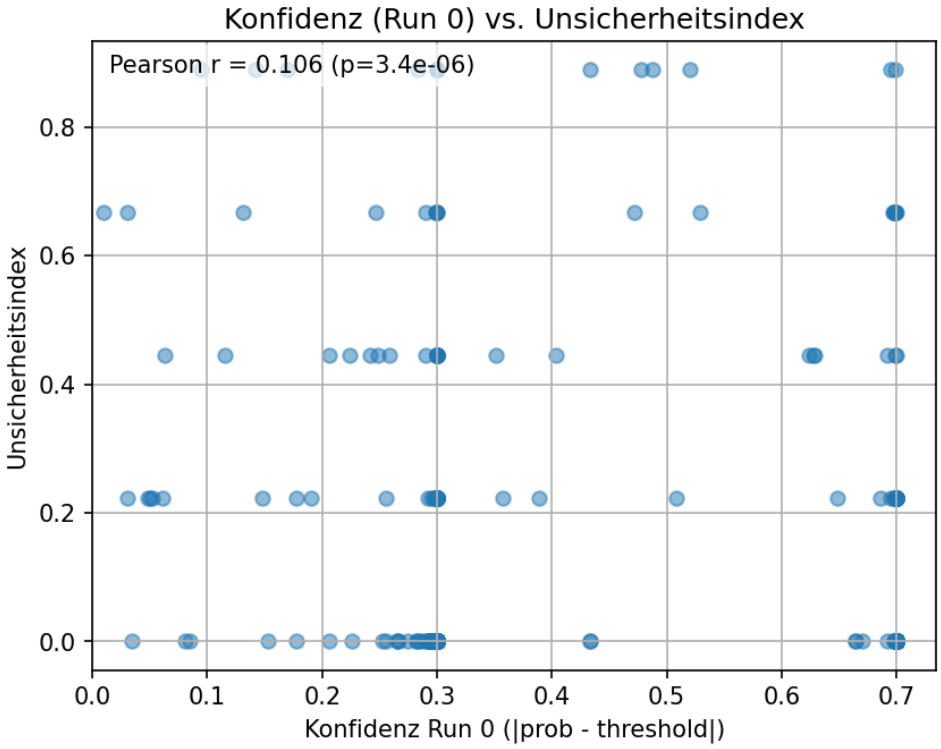

Abbildung 5.13: Konfidenz gegen Unsicherheit auf Abt-Buy

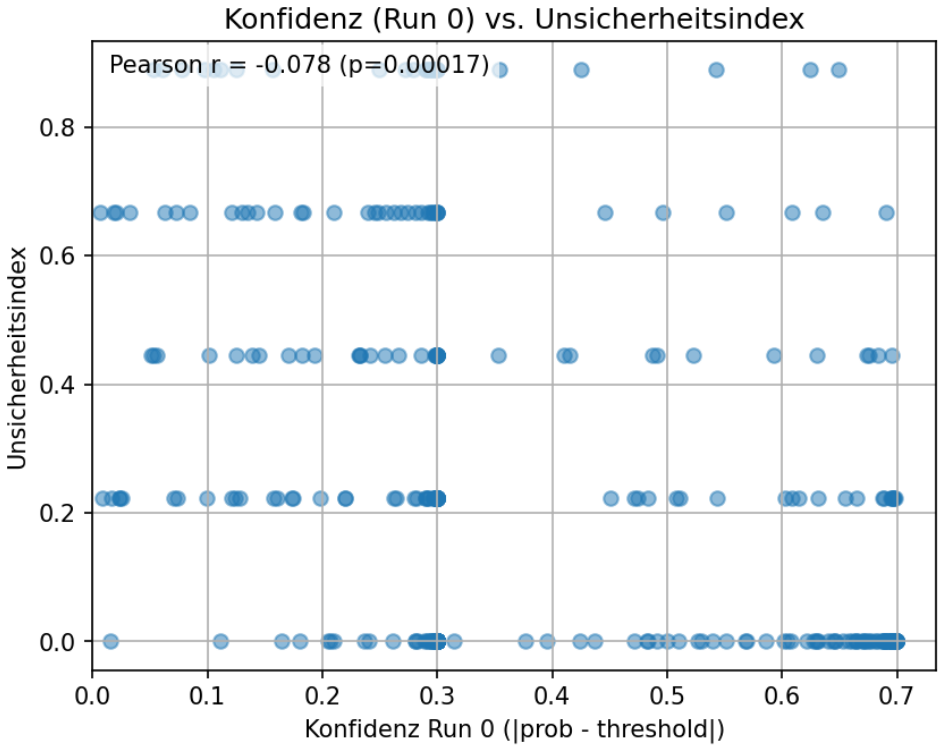

Abbildung 5.14: Konfidenz gegen Unsicherheit auf Amazon-Google

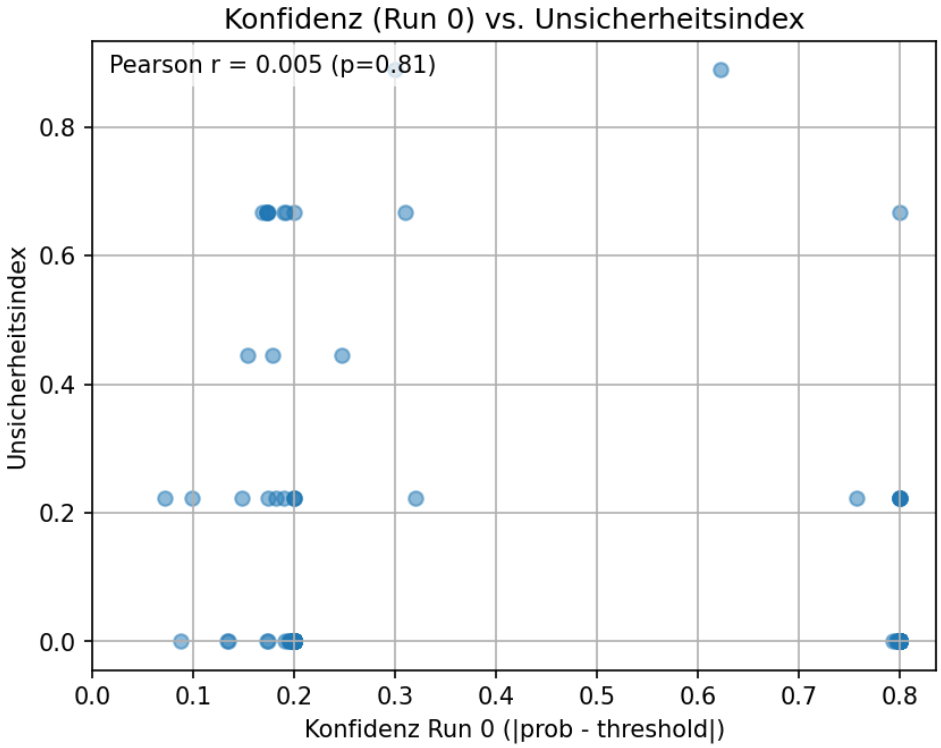

Abbildung 5.15: Konfidenz gegen Unsicherheit auf DBLP-ACM

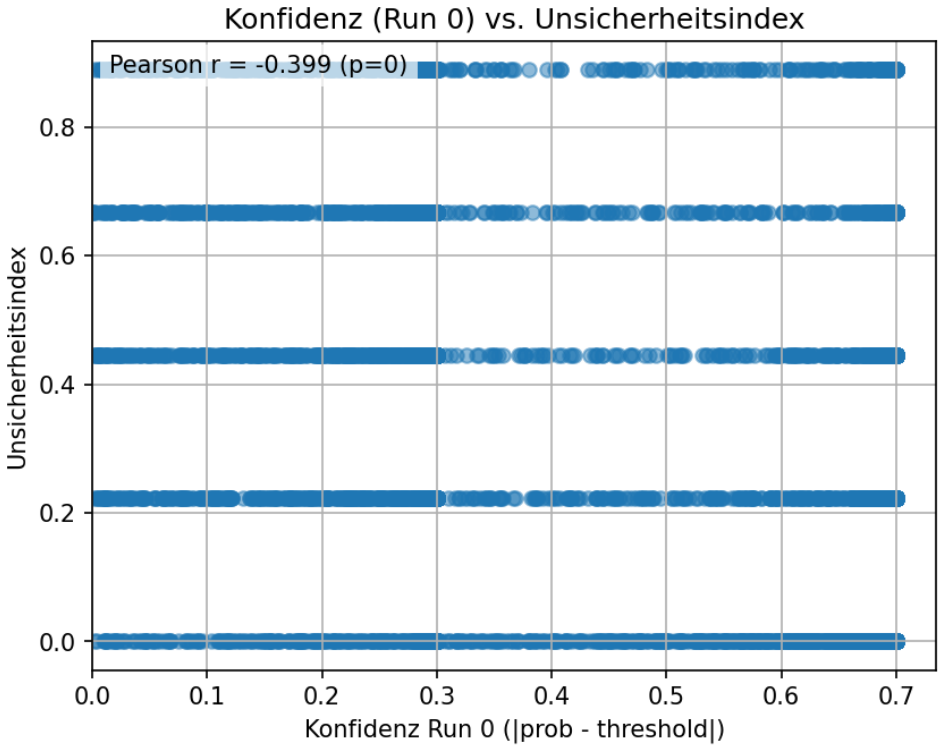

Abbildung 5.16: Konfidenz gegen Unsicherheit auf Organizations

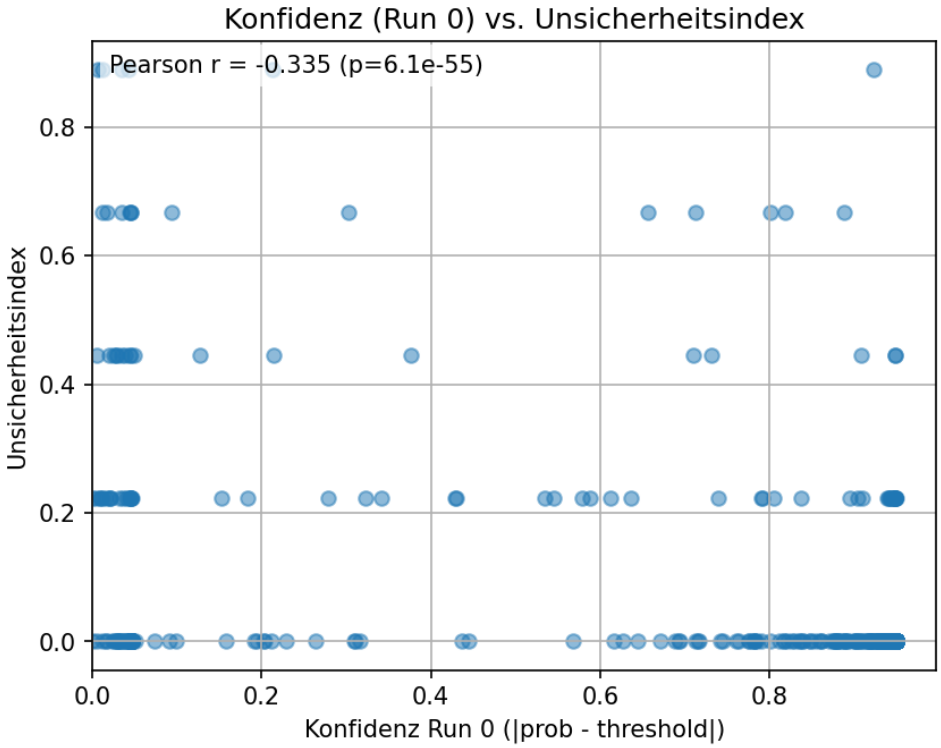

Abbildung 5.17: Konfidenz gegen Unsicherheit auf Walmart-Amazon

Auf Abt-Buy (r = 0,106, p = 3,4e-06) und DBLP-ACM (r = 0,005, p = 0,81) ergibt sich eine schwach positive Korrelation zwischen Konfidenz und Unsicherheit.
Auf Amazon-Google (r = -0,078, p = 0,00017), Organizations (r = -0,399, p = 0) und Walmart-Amazon (r = -0,335, p = 6,1e-55) ergibt sich eine negative Korrelation zwischen Konfidenz und Unsicherheit, welche auf Amazon-Google sehr schwach und auf Organizations und Walmart-Amazon deutlich stärker ist.
Eine Betrachtung der Scatter-Plots zeigt, dass die Ergebnisse keineswegs aussehen, wie das zu erwartende Ergebnis, es also keinen sauberen Zusammenhang gibt.

5.5 Charakterisierung besonders unsicher klassifizierter Datenpaare

Aus den 9 Durchläufen der 3 Modelle in Standard-Konfigurationen wurden jeweils die Daten extrahiert, die nach Abschnitt 4.4 mit $t = 3$ als unsicher gelten. Für jeden Datensatz wurde dann die Menge an unsicheren Daten der 3 Verfahren vereinigt.

So ergibt sich eine Liste von Datenpaaren für jeden Datensatz. Für Organizations wurden 3408 Datenpaare (7,3 % der gesamten Testdaten), für Walmart-Amazon 62 Datenpaare (3,0 % der gesamten Testdaten), für Amazon-Google 157 Datenpaare (6,8 % der gesamten Testdaten), für Abt-Buy 70 Datenpaare (3,7 % der gesamten Testdaten) und für DBLP-ACM 13 (0,5 % der gesamten Testdaten) extrahiert.

Zuerst wurden diese Daten manuell betrachtet und mit den Validierungsdaten verglichen, um spezifische Gemeinsamkeiten der unsicheren Testdaten herauszuarbeiten. Die extrahierten Daten wurden nicht mit den Testdaten verglichen, da die extrahierten Daten eine Teilmenge der Testdaten sind und so teilweise die gleichen Daten untereinander verglichen worden wären. Es konnten folgende Beobachtungen getroffen werden:

In Abt-Buy konnte bei Matches häufig beobachtet werden, dass eine Eigenschaft der Produkte auf unterschiedliche Weise beschrieben ist. Außerdem kam es deutlich häufiger als in den Validierungsdaten vor, dass nur ein oder zwei Attribute bei einem Eintrag des Paares gefüllt sind, während beim anderen Eintrag alle gefüllt sind.

Bei Amazon-Google fällt auf, dass das Attribut „price“ deutlich häufiger bei einem Eintrag im Paar nicht gefüllt ist, verglichen mit den Validierungsdaten. Auch das Attribut „manufacturer“ bleibt bei einem Eintrag häufiger leer, wobei die Information dann oft im Attribut „title“ zu finden ist. Wenn in beiden Einträgen der Preis vorhanden ist, liegen die beiden Werte im Durchschnitt deutlich näher beieinander. Bei Datenpaaren, die kein Match sind, ist häufig zu sehen, dass die beiden Produkte sich nur in einem Detail unterscheiden. Zum Beispiel:

[
"rainbow fish and the whale (win/mac),global-software-publishing,9.99,",
"rainbow fish and the whale,nan,6.95,",
"0"
]

Bei allen unsicheren Datenpaaren aus DBLP-ACM ist das gleiche Jahr beim Attribut „year“ bei beiden Einträgen gegeben. Die Autorenlisten sind sehr unterschiedlich. Oft ist bei einem Eintrag nur ein Autor oder gar keiner eingetragen, während beim anderen Eintrag mehrere gegeben sind. Die Titel der beiden Einträge sind jedoch immer sehr ähnlich.

Bei den unsicheren Datenpaaren aus Walmart-Amazon, die Matches sind, ist recht häufig zu beobachten, dass in einem Eintrag der Preis oder die Seriennummer fehlt. Außerdem ist der Preis bei Matches relativ weit auseinander, verglichen mit den Validierungsdaten. Bei Datenpaaren, die keine Matches sind, sieht man häufig, dass ein Eintrag viel Rauschen, wie die Wiederholung von Eigenschaften, beinhalten. Des Weiteren ist hier oft zu beobachten, dass die Seriennummern der Produkte sehr ähnlich zueinander sind, wie hier:
[
"tripp lite smartpro 1500va ups,computers,tripp lite,smart1500rm2u,499.0,",
"tripp lite smart1500 smart 1500va line-interactive ups 6 outlets,uninterrupted power supply ups,tripp lite,smart1500,320.61,",
"0"
]

Bei Organizations ist zu erkennen, dass die unsicheren Daten häufig sehr wenig Inhalt haben. Das bedeutet, dass die Namen der Organisationen oft sehr kurz mit unspezifischen Informationen sind. Datenpaare, die keine Matches sind, sind häufig trotzdem sehr ähnlich, wie hier zu sehen:
[
"COL name VAL virginia league of women voters",
"COL name VAL west virginia league of women voters",
"0"
]

Anschließend wurden die unsicheren Daten und die Validierungsdaten als JSON-Datei an ChatGPT zur Analyse gegeben. Die Analysen fanden am 02.09.2025 statt und es wurde ChatGPT 5 verwendet. Bevor die Daten gesendet wurden, wurde dieser Prompt übergeben:

„Ich gebe dir nun nacheinander für je einen Datensatz Testdaten, die sich bei einer Analyse durch mehrere Duplikatserkennungsverfahren als besonders unsicher herausgestellt haben, also wo die Verfahren sich in mehreren Durchläufen häufig widersprochen haben. Zum Vergleich erhältst du auch die gesamten Validierungsdaten. Finde charakteristische gemeinsame Eigenschaften der unsicheren Testdaten, die entsprechend weniger stark in den Validierungsdaten zu finden sind."

Chat-GPT gab diese Eigenschaften aus:

Bei Amazon-Google haben die unsicheren Daten durchschnittlich höhere Werte bei „price“ eingetragen. Die Produkte enthalten im Allgemeinen öfter Zahlenangaben. Software bezogene Begriffe, wie „software“, „home“ oder „draft“ kommen überproportional häufig vor. Abkürzungen und kleine Wörter, wie „to“, „s“, „only“, „card“ und „final“ sind überrepräsentiert.

Bei Abt-Buy enthalten viele unsichere Paare sehr ähnliche Produktnamen mit feinen Abweichungen, wie zum Beispiel Farben („silver“, „black“, „red“) oder Modellnummern. In unsicheren Paaren ist oft nur auf einer Seite ein Preis vorhanden. Oft kommt es vor, dass ein Paar aus einem Produkt und einem Zubehör zu diesem Produkt besteht, es werden also Produktvarianten und Zubehör mit einem Hauptgerät verglichen.

Bei Walmart-Amazon enthalten 63 % der Datenpaare die gleiche Marke, während es auf Validierungsdaten nur 16 % sind. Die unsicheren Daten unterscheiden sich im Preis viel weniger

als die Validierungsdaten (durchschnittliche Differenz von 36 vs. 354). Unsichere Titelpaare haben im Durchschnitt eine größere Differenz in der Länge.

Bei Organizations treten in den unsicheren Daten vermehrt kleine orthographische Unterschiede auf (zum Beispiel „benesov“ vs. „benešov“). Es treten häufiger Fälle auf, wo bei einem Eintrag der Name abgekürzt ist, während er beim anderen Eintrag ausgeschrieben wird. Viele unsichere Paare betreffen Organisationen mit verschiedenen rechtlichen Endungen oder Standortzusätzen. Auch verschiedene internationale Schreibweisen und Übersetzungen sind vermehrt zu beobachten.

6 Diskussion

In diesem Kapitel werden die Ergebnisse der Experimente interpretiert, diskutiert und Empfehlungen für die Praxis und weitere Forschung entworfen.

6.1 Interpretation des Unsicherheitsindexes

Zunächst wird der Unsicherheitsindex selbst betrachtet, bevor auf Unterschiede zwischen Parametern, Hardware und Verfahren eingegangen wird. Die Ergebnisse zeigen, dass die absolute Höhe des gemessenen Unsicherheitsindexes in allen Experimenten sehr niedrig ist. Der Index kann bei 9 Durchläufen einen Maximalwert von 8/9 erreichen, wenn also für alle Daten 5 Durchläufe das eine Ergebnis vorhersagen und 4 Durchläufe genau das andere. In den betrachteten Szenarien überschritt der durchschnittliche Unsicherheitsindex jedoch nie den Wert 0,064. Typische Mittelwerte waren je nach Daten und Konfiguration eher im Bereich von 0,02 bis 0,05.

Diese Beobachtung für sich genommen legt zunächst nahe, dass die untersuchten Verfahren in sich recht konsistent sind. Das heißt, sie widersprechen sich selten zwischen verschiedenen Durchläufen auf den gleichen Daten, auch wenn verschiedene Parameter getestet werden oder wenn die Hardware variiert wird. Grenzfälle, die das Potenzial für maximale Unsicherheit beinhalten, scheinen in den verwendeten Benchmarks selten vorzukommen. Bei genauerer Betrachtung wird jedoch klar, dass der niedrige Unsicherheitsindex vor allem daran liegt, dass ein Großteil der Daten in den Datensätzen nie als Duplikat klassifiziert wird und so eine große Einigkeit entsteht. Um eine bessere Aussage über die Konsistenz der Verfahren zu erhalten, müsste man die Experimente auf weiteren Datensätzen wiederholen, die ausgewogener im Verhältnis von Matches zu Nichtmatches sind.

Gleichzeitig bedarf der niedrige Unsicherheitsindex auch weiterer Interpretation. Er bedeutet nicht, dass das Verfahren sicher im Sinne korrekter Klassifikation ist, sondern lediglich, dass es konsistent in seinen eigenen Entscheidungen ist. Ein Modell kann sich auch auf stabile Weise irren. Daher wird die Metrik erst aufschlussreich, wenn man Performance-Metriken, wie den F1-Score hinzunimmt. Hier ist dennoch auffällig, dass bei manchen Parametervarianten, beispielsweise DistilBERT auf Walmart-Amazon, die Unsicherheit steigt und die F1-Werte sinken.
Des Weiteren wird die Interpretation des Unsicherheitsindexes dadurch erschwert, dass ein Vergleich zu modernen Verfahren für andere Problemstellungen mit ähnlicher Performance fehlt. Es ist möglich, dass solche Verfahren deutlich sicherer oder unsicherer im Sinne des Unsicherheitsindexes sind.

Ein wichtiger Befund betrifft den Vergleich zwischen mehrfacher Ausführung auf identischer Hardware und unterschiedlicher Hardware (Tabelle 5.3). Es zeigt sich, dass alle Durchläufe exakt das gleiche Ergebnis in den Vorhersagen und somit auch in den Metriken (Precision, Recall, F1-Score und Unsicherheitsindex) liefern. Das bedeutet, dass im gewählten Setting unabhängig

vom Cluster-Knoten oder der GPU die Ergebnisse stabil reproduzierbar sind, vorausgesetzt, es werden die gleichen Parameter verwendet, einschließlich der Run-ID als Random Seed.
Im Bezug auf den Unsicherheitsindex bedeutet dies, dass das Variieren zwischen GPUs des gleichen Modells keine zusätzliche Unsicherheit in die Vorhersagen von Ditto gebracht hat. Dieses Ergebnis sollte jedoch vorsichtig interpretiert werden. Es zeigt Reproduzierbarkeit in der konkreten Konfiguration und Umgebung, garantiert aber nicht, dass man beim Wechsel zu einer anderen Hardware-Architektur, Software-Version, Verfahren oder bei Variation zwischen GPUs unterschiedlicher Modelle zum gleichen Schluss kommt.
Um ein besseres Bild zu diesen Unklarheiten zu erhalten, wären weitere Experimente notwendig.

Deutlich interessantere Effekte zeigen sich bei der Variation der Hyperparameter (Tabelle 5.4 - 5.6). Hier sticht besonders die verringerte Batch Size hervor. In allen Datensätzen führte eine verringerte Batch Size zu einem höheren Unsicherheitsindex, etwa +32% bei Abt-Buy oder +56,25% bei Walmart-Amazon. Dieser Befund ist plausibel, weil kleinere Batches bei gleicher Lernrate zu einer höheren Varianz in den Gradienten beim SGD führen. Somit wird die Variabilität zwischen verschiedenen Durchläufen des Trainings erhöht und die Stabilität der Vorhersagen vermindert. In der Optimierungsforschung wird der Zusammenhang von Batchgröße, Lernrate und Varianz beschrieben, so kann eine Abweichung vom optimalen Verhältnis die Generalisierung und Konsistenz verschlechtern. [SL18]

Auch die Wahl des Language Models wirkt sich stark auf die Unsicherheit aus. Die Verwendung von Distilbert auf Walmart-Amazon führte zu einem stark erhöhten Unsicherheitsindex (+81,25%) und niedrigeren F1-Score (-0,075 im Mittel), verglichen mit der Ditto-Standard-Konfiguration, welche RoBERTa verwendet. Ein Grund für diese beiden Beobachtungen könnte die verminderte Modellkapazität von DistilBERT gegenüber RoBERTa sein, welche Performance und Stabilität der Vorhersagen beeinträchtigen könnte. Auf den anderen Datensätzen ist das gleiche Muster, wenn auch schwächer, zu beobachten. BERT produzierte in den meisten Fällen nur schwächere Abweichungen. Diese Ergebnisse legen den Schluss nahe, dass vor allem Modelle mit reduzierter Kapazität instabiler in ihrer Klassifikation sind.

Des Weiteren zeigt sich, dass durch die Kombination verschiedener Durchläufe mit unterschiedlicher Parameter-Konfiguration teilweise eine sehr hohe Unsicherheit erreicht werden kann. Tatsächlich wird in dieser Situation der höchste Unsicherheitsindex der Arbeit mit 0,064 auf dem Organizations-Datensatz erreicht. Dieser ist sogar höher als die erreichte Unsicherheit des Mixes der Verfahren mit 0,062. Auch wenn hier damit ein Mix von 3 Verfahren mit einem Mix aus 9 Konfigurationen verglichen wird, zeigt dies, dass Änderungen in der Parameter-Konfiguration zu erheblich unterschiedlichen Vorhersagen und somit Uneinigkeit zwischen Modellinstanzen führen können.

Die Datensätze selbst unterscheiden sich stark in der erreichten Unsicherheit. DBLP-ACM weist durchweg extrem niedrige Unsicherheit auf, mit einem Index von meistens zwischen 0,002 und 0,005, bei gleichzeitig nahezu perfekten F1-Scores. Organizations und Amazon-Google zeigen deutlich höhere Werte, bis zu 0,048. Dies deutet darauf hin, dass die Daten heterogener sind und mehr Grenzfälle enthalten.
Überraschenderweise hängt die Unsicherheit des Datensatzes nicht so stark von der Performance der Verfahren auf dem Datensatz ab, wie man denken könnte. Unicorn erzielt auf DBLP-ACM, Abt-Buy und Walmart-Amazon nahezu die gleiche Unsicherheit, obwohl die Performance sich stark unterscheidet. Dies wirft die Frage auf, welche Eigenschaften eines Datensatzes

zu hoher oder niedriger Unsicherheit der Vorhersagen führen, abgesehen von der erreichten Performance.

Schließlich zeigt der Vergleich der Verfahren (Tabelle 5.7), dass Unicorn in den meisten Fällen den niedrigsten Unsicherheitsindex erzielt. Der Mix der Verfahren (Ditto, HierGAT und Unicorn) erzielt in vier von fünf Datensätzen eine höhere oder mindestens gleichbleibende Unsicherheit. Am deutlichsten wird dieser Effekt, wenn man das Ergebnis auf dem Organization-Datensatz betrachtet. Hier erzielt der Mix der Verfahren eine Unsicherheit von 0,062. Diese Beobachtungen lassen sich mit Erkenntnissen aus der Ensemble-Forschung erklären. Verschiedene Modelle mit einem unterschiedlichen Bias erhöhen normalerweise die Diversität der Vorhersagen im Ensemble. Das kann für die Leistung nützlich sein, wenn zum Beispiel ein Mehrheitsvotum verwendet wird, führt aber zwangsläufig auch zu mehr Uneinigkeit und somit Unsicherheit zwischen den Modellen. [KW03] Dies spiegelt sich im Unsicherheitsindex wider.

6.2 Unischerheitsindex als Möglichkeit der Effizienzsteigerung bei menschlicher Einschätzung

Die Ergebnisse der Orakel-Analyse (Tabellen 5.8 - 5.12) zeigen, dass das Orakel durchgängig eine bessere oder gleichbleibende Performance erzielt als die Mittelwerte der Durchläufe und dem Verfahren mit Mehrheitsvotum. Natürlich kann das Orakel per Definition im Vergleich mit dem reinen Mehrheitsvotum nur zu einer besseren oder gleichbleibenden Performance führen. Dennoch ist bemerkenswert, wie hoch der Performance-Gewinn im Vergleich zur Menge der Klassifikationen durch das Orakel ist. Bei Organizations und $t = 1$ steigt der F1-Score um 0,091, wobei 10,7 % der Daten durch das Orakel klassifiziert werden. Bei Amazon-Google und $t = 1$ steigt der F1-Score sogar um 0,124, während nur 8,3 % der Daten durch das Orakel klassifiziert werden. Selbst auf Datensätzen mit hohem Ausgang-F1-Score, wie Abt-Buy, konnte noch eine deutliche Verbesserung (+0,011 im F1-Score) erreicht werden, bei nur 1,1 % Klassifikationen durch das Orakel.
Des Weiteren fällt auf, dass das Orakel durchweg bessere oder gleiche Ergebnisse liefert, als das Zufallsorakel mit gleichem Budget an Klassifikationen. Daraus lässt sich schließen, dass Daten, die teilweise unsicher sind, also bei denen es mindestens eine Uneinigkeit zwischen den Durchläufen gab, tatsächlich sinnvoller für eine manuelle Klassifikation sind, als zufällig ausgewählte Daten.

Das bedeutet, dass nicht alle Datenpunkte gleich informativ für eine manuelle Nachklassifikation sind. Wählt man also gezielt unischere Instanzen, steigt der Nutzen einer manuellen Klassifikation erheblich. Diese Beobachtung ist konsistent mit Erkenntnissen aus der Literatur zu Active Learning und insbesondere dem Ansatz des Query-by-Comittee [SOS92] Hier wird argumentiert, dass Daten mit hoher Ungenauigkeit beziehungsweise Uneinigkeit zwischen Modellen besonders wertvoll für das Labeln zum Training sind. In den Experimenten dieser Arbeit wurden keine Daten zum Training ausgewählt, aber die Ergebnisse sind vermutlich trotzdem übertragbar.

Die Korrelationsanalyse (Abbildungen 5.2 - 5.6) bestätigt diesen Befund teilweise. Auf Amazon-Google ($r = 0{,}975$, $p = 0{,}025$) und Organizations ($r = 0{,}984$, $p = 0{,}016$) zeigt sich eine sehr starke und signifikante positive Korrelation zwischen t und dem Performance-Gewinn pro Orakelklassifikation (bei einem gängigen Signifikanzniveau von 0,05). Das bedeutet: Wenn man die Auswahl der Datenpunkte restriktiver gestaltet, also je unsicherer ein Datenpunkt sein

muss, damit er durch das Orakel klassifiziert wird, desto höher ist der Nutzen, den das Orakel pro klassifizierten Fall erzielt. Auch auf DBLP-ACM (r = 0,717, p = 0,28) und Abt-Buy (r = 0,448, p = 0,55) zeigen sich positive Korrelationskoeffizienten, welche jedoch nicht statistisch signifikant sind. Dies kann eventuell mit der geringen Zahl an Messpunkten (nur vier Werte für t) erklärt werden. Auf Walmart-Amazon ist sogar ein negativer Korrelationskoeffizient zu beobachten (r = -0,715, p = 0,29). Dies ist durch den Ausreißer zu erklären: Bei $t = 4$ gab es keinen Performance-Gewinn pro Orakelklassifikation, da bereits alle relevanten Daten korrekt durch das Mehrheitsvotum klassifiziert wurden.

Aus diesen Ergebnissen lassen sich folgende Schlüsse ziehen: Erstens zeigt sich klar, dass der Unsicherheitsindex mindestens für manche Datensätze geeignet ist, um Daten für eine Orakelklassifikation oder in der Praxis für eine manuelle Klassifikation auszuwählen. Denn unsichere Datenpunkte tragen wesentlich mehr zum Performancegewinn bei als zufällig ausgewählte, und der Nutzen steigt sogar mit steigender Unsicherheit.
Zweitens ist diese Beobachtung in der Tendenz zwar konsistent, aber es ist fragwürdig, ob sich der Effekt auch in allen Datensätzen experimentell nachweisen ließe. Da pro Szenario nur 9 Trainingsdurchläufe durchgeführt wurden, gibt es nur vier t-Werte. Um zu überprüfen, ob sich der festgestellte Effekt weiter verallgemeinern lässt, sollte das Experiment mit deutlich mehr Durchläufen, mehr Datensätzen und vorher definiertem Signifikanzniveau wiederholt beziehungsweise weitergeführt werden.
Wenn sich hier zeigt, dass der Unsicherheitsindex für bestimmte Datensätze praktikabler ist als für andere, deckt sich das mit bekannten Befunden aus der Literatur, nach denen Unsicherheitsmetriken zwar wertvolle Indikatoren sind, aber ihre Zuverlässigkeit stark vom Datensatz abhängt. [Ova+19]

Zusammenfassend lässt sich festhalten, dass Unsicherheit wahrscheinlich eine geeignete Metrik ist, um menschliche Ressourcen effizient auf schwierige Fälle in der Duplikaterkennung zu konzentrieren und somit eine Performance-Steigerung zu erzielen.
Ein passender Use-Case könnte zum Beispiel eine kritische Anwendung sein, die zwingend eine hohe Performance benötigt. Je nachdem, wie viel Rechenkapazität die Benutzer zur Verfügung haben, führen sie mehrmals ein oder mehrere SOTA-Verfahren auf ihrem Datensatz aus. Sie berechnen den Unsicherheitsindex für jeden Datenpunkt und sortieren die Daten absteigend nach Unsicherheitsindex. Anschließend klassifizieren sie so viele Daten absteigend in der Liste, wie sie manuelle Ressourcen haben. Auf diese Weise kann die Stärke moderner Verfahren in der Duplikaterkennung genutzt werden und mit nur einem Bruchteil des manuellen Aufwands, verglichen mit vollständiger manueller Klassifikation, eine gesteigerte Performance erreicht werden.

6.3 Implikationen der Schnittmengenanalyse

Bei der Betrachtung der Schnittmengen wird die Stärke der LLM-basierten Verfahren deutlich. Sie können die Semantik der Datenpaare verstehen und so mehr Duplikate identifizieren, als es mit klassischem ML und Textähnlichkeit möglich ist. Des Weiteren zeigt sich auf allen Datensätzen auch, dass es für jedes Verfahren einige Datenpunkte gibt, die nur von diesem Verfahren als Duplikat klassifiziert werden. Hier zeigt sich, dass jedes Verfahren eine einzigartige Funktionsweite hat, auch wenn die Verfahren teilweise auf ähnlichen oder gleichen Architekturen basieren. Dieser Befund spiegelte sich zuvor in Abschnitt 7.1.4 wider, wo die Unsicherheit durch

den Mix der drei betrachteten Verfahren gesteigert wurde, verglichen mit jedem Verfahren für sich allein.

Darüber hinaus kann man die Schnittmengen auch im Kontext von Ensembles deuten. Große Schnittmengen bilden einen stabilen Konsenskern, der durch ein Mehrheitsvotum bestätigt wird. Kleinere exklusivere Mengen zeigen die Diversität der Modelle, wodurch zwar die Varianz in den Vorhersagen erhöht wird, aber auch Potenzial eröffnet wird, durch kombinierte Verfahren zusätzliche Duplikate zu entdecken.

Eventuell sind aus praktischer Perspektive gerade jene Datenpunkte interessant, die nicht im Konsenskern liegen. Sie beinhalten entweder besonders schwer zu entscheidenden Fälle oder spiegeln spezifische Stärken einzelner Verfahren wider. Das kann bedeuten, dass diese unsicheren Treffer priorisiert in einer manuellen Betrachtung überprüft werden sollten, weil aus ihnen die größte Erkenntnis über die Daten oder die Modelle gewonnen werden kann.

6.4 Modellkonfidenz als Prädiktor von Unsicherheit

Die Untersuchung des Zusammenhangs zwischen Modellkonfidenz und Unsicherheit zeigt gemischte Ergebnisse über die verschiedenen Datensätze hinweg. Dies zeigt, dass höhere Modellkonfidenz nicht notwendigerweise mit geringerer Unsicherheit im Sinne der Konsistenz über verschiedene Durchläufe zusammenhängt. Es tritt sogar teilweise das Gegenteil auf.

Eine mögliche Erklärung liegt darin, wie beide Größen definiert wurden. Die Konfidenz bezieht sich auf genau einen Durchlauf und ist der Abstand des vorhergesagten Wertes von Ditto zum Schwellenwert, der in diesem Durchlauf verwendet wurde. Hingegen wird die Unsicherheit über 9 Durchläufe hinweg berechnet und misst die Varianz der Vorhersagen für je einen Datenpunkt. Aus diesem methodischen Unterschied folgt, dass es sich um zwei verschiedene Konzepte handelt. Die Konfidenz ist eine lokale Eigenschaft einer Vorhersage, aber die Unsicherheit ist in dieser Arbeit eine globale Konsistenz über mehrere Instanzen desselben Modells.

Hinzu kommt auch, dass Ditto den Threshold für die Klassifikation von Lauf zu Lauf unterschiedlich wählt. Dies beeinflusst die Klassifikationsergebnisse kaum, da die meisten Datenpunkte nahe bei 0 oder 1 liegen und deshalb stabil klassifiziert werden. Dennoch führt dies in den Plots zu erkennbaren Clustern: Wenn die Konfidenz den Wert des Thresholds und den Wert $1 - \text{Threshold}$ annimmt. Somit ist erklärt, warum die Datenpunkte nicht gleichmäßig über das gesamte Intervall verteilt sind, sondern bestimmte Bereiche stärker besetzt sind.

Im Kontext neuronaler Netze gliedern sich diese Befunde in eine breite Diskussion zur Frage ein, ob Wahrscheinlichkeitswerte neuronaler Netze verlässlich als Indikator für Unsicherheit gelten können. Mehrere Studien zeigten bereits, dass dies nur eingeschränkt der Fall ist.
[Guo+17] untersuchten die Kalibrierung moderner neuronaler Netz-Architekturen, also die Übereinstimmung von vorhergesagter Wahrscheinlichkeit und tatsächlich empirisch bestimmter Genauigkeit. Sie führten ihre Analysen auf Bildklassifikationsdaten durch und fanden, dass moderne Architekturen im Vergleich zu älteren Modellen deutlich bessere Performance erzielen, aber gleichzeitig stark überkonfident in ihrer Vorhersage sind. Das meint, dass die Modelle auch bei falscher Klassifikation oft Wahrscheinlichkeiten nahe 1 ausgaben. Die Autoren ziehen das Fazit, dass Modellkonfidenz nicht direkt als Unsicherheitsmaß interpretiert werden kann, ohne zusätzliche Kalibrierung.

[LPB17] stellten mit Deep Ensembles eine einfache, aber dennoch leistungsstarke Methode vor. Mehrere Modelle werden unabhängig mit verschiedenen Initialisierungen trainiert und ihre Vorhersagen gemittelt. Dies führt oft zu einer verbesserten Performance und die Abweichungen zwischen den Modellen können als robustes Maß für Unsicherheit genutzt werden, ähnlich wie es in dieser Arbeit gemacht wird.

Vor dem Hintergrund der vorhandenen Literatur ist es nicht überraschend, dass in den durchgeführten Experimenten keine einheitliche Korrelation zwischen Konfidenz und Unsicherheitsindex gefunden werden konnte. Dass die Korrelation in einigen Fällen sogar negativ ist, kann bedeuten, dass hohe Konfidenzwerte durchaus mit instabilen Entscheidungen über mehrere Wiederholungen einhergehen können. Dies ist ein Hinweis auf die Grenzen einfacher Wahrscheinlichkeiten als Grundlage für ein Unsicherheitsmaß.

6.5 Interpretation der Analyse unsicherer Daten

Die Analyse der unsicheren Daten zeigt, dass Unsicherheit nicht zufällig entsteht, sondern zu finden ist, wenn sich starke Ähnlichkeiten mit feinen Unterschieden überlagern. Es ist charakteristisch, dass betroffene Datenpaare entweder sehr ähnlich wirken, aber im Detail voneinander abweichen, oder dass durch unvollständige Angaben oder unterschiedliche Schreibkonventionen eine eindeutige Entscheidung erschwert wird.

Es zeigt sich Heterogenität in der Attributdarstellung als zentrales Merkmal der unsicheren Daten. Häufig sind in einem Eintrag eines Paares alle Attribute mit Werten gefüllt, während sie beim anderen Eintrag fehlen. Oder die Informationen sind vorhanden, stehen aber im falschen Attribut. Dieses Ergebnis bestätigt eine Beobachtung, die in der Literatur wiederholt betont wird. Fehlende oder inkonsistente Daten gehören zu einem zentralen Grund für Unsicherheit in der Duplikaterkennung.

Außerdem zeigen sich Variantenbildungen durch orthografische Abweichungen, Abkürzungen gegenüber ausgeschriebenen Formen und sprachliche Übersetzungen oder Zusätze, wie Farben und Modellnummern. Solche Unterschiede sind für Menschen einfach und intuitiv zu verstehen, aber für Modelle problematisch, weil sie entscheiden müssen, ob es sich um einen wesentlichen Unterschied oder nur um eine alternative Darstellung handelt. Diese semantische Komplexität kann von klassischen ML-Ansätzen nicht abgebildet werden und auch LLM-basierte Verfahren stoßen aktuell hier noch an ihre Grenzen.

Ein weiteres Muster ist eine Asymmetrie in der Textdarstellung. In unsicheren Paaren gibt es oft einen Eintrag mit einem langen und detaillierten Text und einen anderen Eintrag, der auf Kurzformen reduziert ist. Diese Unterschiede bewirken eine hohe Differenz in der Textlänge, obwohl sie inhaltlich gleich sind. Auf diese Weise kann Mehrdeutigkeit und somit Unsicherheit entstehen.

Schließlich ist festzustellen, dass unsichere Daten überproportional Paare betreffen, die semantisch nahe, aber nicht identisch sind. Dazu gehören Produktvarianten, Publikationen mit gleichen Titeln, aber unterschiedlichen Autorenlisten oder Organisationen mit ähnlichen Namen oder unterschiedlichen rechtlichen Endungen. Diese Daten liegen semantisch an der Grenze zwischen „gleicher Entität“ und „verwandten, aber verschiedenen Entitäten“. Diese Ambiguität wird durch die Verfahren als Unsicherheit widergespiegelt. Hier wird noch einmal deutlich, dass gerade die

Daten, bei denen hohe Uneinigkeit auftritt, besonders informativ für zum Beispiel das Training sind, wie es in der Active-Learning-Literatur bekannt ist. [SOS92]

Gleichzeitig ist hervorzuheben, dass die hier dargestellten Ergebnisse mit Vorsicht zu interpretieren sind. Sie basieren nicht auf objektiven Messungen oder quantitativen Auswertungen, sondern zum einen auf einer manuellen Analyse und zum anderen auf einer Befragung von ChatGPT. Damit sind die Ergebnisse eher Hypothesen und qualitative Einsichten, aber sie erheben keinen Anspruch auf statistische Allgemeingültigkeit.

7 Schlussfolgerung und Ausblick

Die Untersuchungen zeigen, dass Unsicherheit in der Duplikaterkennung ein eigenständiger Aspekt ist, der als wichtige Ergänzung zur Betrachtung von reinen Performance-Metriken gesehen werden kann.
Der definierte Unsicherheitsindex macht durch mehrere Experimente deutlich, dass die Verfahren Ditto, HierGAT und Unicorn im Durchschnitt konsistente Entscheidungen treffen, aber Unterschiede durch die Wahl der Parameter, die Architektur der Modelle oder Charakteristika der Datensätze entstehen können. Teilweise ist die Konsistenz der Entscheidungen mit der Wahl der Datensätze zu erklären, die einen disproportional hohen Anteil an Matches enthalten.
Besonders hervorzuheben ist, dass Unsicherheit gezielt genutzt werden kann, um menschliche Ressourcen gezielt auf bestimmte Daten zu konzentrieren, um eine Effizienzsteigerung zu bewirken: Datenpunkte mit relativ hoher Unsicherheit liefern den größten Mehrwert bei einer manuellen Klassifikation, und dieser Mehrwert steigt sogar tendenziell mit steigender Unsicherheit.
Des Weiteren wurde deutlich, dass Konfidenzwerte eines Modells wie Ditto für sich nicht als Indikator für Unsicherheit dienen können. Eventuell wäre es von praktischem Wert, wenn bei der Entwicklung von zukünftiger Modellarchitektur darauf geachtet wird, dass die Konfidenz des Modells tatsächlich der empirisch beobachtbaren Wahrscheinlichkeit der Klassifikation entspricht, ohne dass die Konfidenz nachträglich kalibriert werden muss.
Die Analyse besonders unsicherer Daten zeigte auf, wo aktuelle Verfahren an semantische Grenzen stoßen. Hier zeigte sich auch, dass Unsicherheit nicht ausschließlich ein methodisches Problem ist, sondern auch eine gewisse inhärente Mehrdeutigkeit in den Daten vorhanden ist.

Für zukünftige Forschung könnte eine stärkere empirische Basis mit mehr Durchläufen und weiteren Datensätzen die Robustheit der Befunde stärken.
Des Weiteren erscheint es vielversprechend, Unsicherheit zu verwenden, um herausfordernde Trainingsdaten für zukünftige Modellarchitekturen zu identifizieren und so die Qualität der in der Duplikaterkennung weiter zu erhöhen. Eine weitere Möglichkeit, die schon jetzt umgesetzt werden kann, ist es, moderne Verfahren auszuwählen, die sich möglichst stark widersprechen und sie in einer Ensemble-Strategie zu verknüpfen und so maximale Performance zu erhalten, indem die Stärken der einzelnen Verfahren genutzt werden.
Schließlich könnten weitere qualitative Analysen unsicherer Daten helfen, die semantischen Gründe für Unsicherheit noch besser zu verstehen und die Verfahren dahingehend zu optimieren.

Literatur

[Bar+21] Andrea Baraldi u. a. *Landmark Explanation: An Explainer for Entity Matching Models*. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. CIKM '21. Virtual Event, Queensland, Australia: Association for Computing Machinery, 2021, S. 4680–4684. ISBN: 9781450384469. DOI: 10.1145/3459637.3481981. URL: https://doi.org/10.1145/3459637.3481981.

[Blu+15] Charles Blundell u. a. *Weight Uncertainty in Neural Networks*. 2015. arXiv: 1505.05424 [stat.ML]. URL: https://arxiv.org/abs/1505.05424.

[Dev+19] Jacob Devlin u. a. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis und Vassilios S. Verykios. *Duplicate Record Detection: A Survey*. In: *IEEE Transactions on Knowledge and Data Engineering* 19.1 (2007), S. 1–16. DOI: 10.1109/TKDE.2007.250581.

[GBC16] Ian Goodfellow, Yoshua Bengio und Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GG16] Yarin Gal und Zoubin Ghahramani. *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*. 2016. arXiv: 1506.02142 [stat.ML]. URL: https://arxiv.org/abs/1506.02142.

[Guo+17] Chuan Guo u. a. *On Calibration of Modern Neural Networks*. 2017. arXiv: 1706.04599 [cs.LG]. URL: https://arxiv.org/abs/1706.04599.

[Hua+22] Jiacheng Huang u. a. *Deep entity matching with adversarial active learning*. In: *The VLDB Journal* 32 (Apr. 2022), S. 1–27. DOI: 10.1007/s00778-022-00745-1.

[KW03] Ludmila Kuncheva und Chris Whitaker. *Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy*. In: *Machine Learning* 51 (Mai 2003), S. 181–207. DOI: 10.1023/A:1022859003006.

[Li+20] Yuliang Li u. a. *Deep entity matching with pre-trained language models*. In: *Proceedings of the VLDB Endowment* 14.1 (Sep. 2020), S. 50–60. ISSN: 2150-8097. DOI: 10.14778/3421424.3421431. URL: http://dx.doi.org/10.14778/3421424.3421431.

[Lip17] Zachary C. Lipton. *The Mythos of Model Interpretability*. 2017. arXiv: 1606.03490 [cs.LG]. URL: https://arxiv.org/abs/1606.03490.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel und Charles Blundell. *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles*. 2017. arXiv: 1612.01474 [stat.ML]. URL: https://arxiv.org/abs/1612.01474.

[NH10] Felix Naumann und Melanie Herschel. *An Introduction to Duplicate Detection*. Morgan und Claypool Publishers, 2010. ISBN: 1608452204.

[Ova+19] Yaniv Ovadia u. a. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift*. 2019. arXiv: 1906.02530 [stat.ML]. URL: https://arxiv.org/abs/1906.02530.

[PB23] Ralph Peeters und Christian Bizer. *Using ChatGPT for Entity Matching*. 2023. arXiv: 2305.03423 [cs.CL]. URL: https://arxiv.org/abs/2305.03423.

[PSB24] Ralph Peeters, Aaron Steiner und Christian Bizer. *Entity Matching using Large Language Models*. 2024. arXiv: 2310.11244 [cs.CL]. URL: https://arxiv.org/abs/2310.11244.

[SL18] Samuel L. Smith und Quoc V. Le. *A Bayesian Perspective on Generalization and Stochastic Gradient Descent*. 2018. arXiv: 1710.06451 [cs.LG]. URL: https://arxiv.org/abs/1710.06451.

[SOS92] H. S. Seung, M. Opper und H. Sompolinsky. *Query by committee*. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1992, S. 287–294. ISBN: 089791497X. DOI: 10.1145/130385.130417. URL: https://doi.org/10.1145/130385.130417.

[Tu+23] Jianhong Tu u. a. *Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration*. In: *Proc. ACM Manag. Data* 1.1 (Mai 2023). DOI: 10.1145/3588938. URL: https://doi.org/10.1145/3588938.

[Vas+23] Ashish Vaswani u. a. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[Wim+23] Lisa Wimmer u. a. *Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures?* 2023. arXiv: 2209.03302 [cs.LG]. URL: https://arxiv.org/abs/2209.03302.

[Yao+22] Dezhong Yao u. a. *Entity Resolution with Hierarchical Graph Attention Networks*. In: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD '22. Philadelphia, PA, USA: Association for Computing Machinery, 2022, S. 429–442. ISBN: 9781450392495. DOI: 10.1145/3514221.3517872. URL: https://doi.org/10.1145/3514221.3517872.

A Performance Metriken der verschiedenen Experimente

Tabelle A.1: Ergebnisse je Datensatz (Precision/Recall/F1) über Läufe und Mittelwert

		Abt-Buy			Amazon-Google			DBLP-ACM			Organizations			Walmart-Amazon		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Verfahren	Durchlauf															
Ditto	1	0,851	0,888	0,869	0,772	0,752	0,762	0,978	0,991	0,984	0,790	0,881	0,833	0,885	0,798	0,839
	2	0,884	0,888	0,886	0,745	0,786	0,765	0,980	0,991	0,985	0,803	0,863	0,832	0,907	0,855	0,880
	3	0,897	0,883	0,890	0,709	0,812	0,757	0,976	0,993	0,984	0,802	0,863	0,831	0,884	0,865	0,874
	4	0,901	0,883	0,892	0,689	0,786	0,735	0,993	0,989	0,991	0,808	0,861	0,834	0,880	0,834	0,856
	5	0,938	0,874	0,905	0,740	0,765	0,752	0,982	0,986	0,984	0,844	0,842	0,843	0,828	0,876	0,851
	6	0,896	0,874	0,885	0,731	0,709	0,720	0,989	0,989	0,989	0,806	0,864	0,834	0,883	0,819	0,849
	7	0,819	0,811	0,815	0,722	0,812	0,765	0,976	0,993	0,984	0,799	0,870	0,833	0,906	0,798	0,848
	8	0,893	0,854	0,873	0,737	0,765	0,751	0,989	0,977	0,983	0,787	0,888	0,834	0,904	0,881	0,892
	9	0,865	0,874	0,870	0,716	0,765	0,740	0,982	0,991	0,987	0,847	0,838	0,842	0,910	0,834	0,870
	Mittelwert	0,883	0,870	0,876	0,729	0,772	0,750	0,983	0,989	0,986	0,810	0,863	0,835	0,887	0,840	0,862
HierGAT	1	0,870	0,913	0,891	0,700	0,756	0,727	0,989	0,984	0,986	0,776	0,889	0,829	0,794	0,819	0,806
	2	0,919	0,883	0,901	0,734	0,765	0,749	0,991	0,984	0,988	0,736	0,890	0,806	0,825	0,808	0,817
	3	0,762	0,840	0,799	0,760	0,756	0,758	0,984	0,989	0,987	0,772	0,894	0,828	0,851	0,829	0,840
	4	0,868	0,927	0,897	0,744	0,709	0,726	0,991	0,989	0,990	0,764	0,890	0,822	0,847	0,834	0,841
	5	0,919	0,883	0,901	0,726	0,769	0,747	0,987	0,989	0,988	0,764	0,886	0,820	0,832	0,845	0,838
	6	0,899	0,903	0,901	0,754	0,774	0,764	0,989	0,980	0,984	0,763	0,897	0,824	0,847	0,829	0,838
	7	0,859	0,859	0,859	0,732	0,735	0,733	0,995	0,984	0,990	0,773	0,890	0,827	0,849	0,819	0,834
	8	0,883	0,917	0,900	0,738	0,756	0,747	0,986	0,982	0,984	0,786	0,892	0,836	0,862	0,808	0,834
	9	0,866	0,879	0,872	0,763	0,714	0,737	0,989	0,986	0,988	0,749	0,891	0,814	0,857	0,777	0,815
	Mittelwert	0,872	0,889	0,880	0,739	0,748	0,743	0,989	0,985	0,987	0,765	0,891	0,823	0,840	0,819	0,829
Unicorn	1	0,907	0,942	0,924	0,704	0,774	0,737	0,993	0,984	0,989	0,848	0,870	0,859	0,918	0,813	0,863
	2	0,905	0,922	0,913	0,758	0,778	0,768	0,991	0,989	0,990	0,849	0,873	0,861	0,880	0,839	0,859
	3	0,894	0,947	0,920	0,713	0,808	0,758	0,989	0,989	0,989	0,834	0,876	0,855	0,838	0,860	0,849
	4	0,913	0,922	0,918	0,726	0,735	0,730	0,987	0,989	0,988	0,835	0,884	0,859	0,909	0,829	0,867
	5	0,906	0,932	0,919	0,695	0,808	0,747	0,980	0,984	0,982	0,843	0,886	0,864	0,916	0,850	0,882
	6	0,889	0,937	0,913	0,679	0,778	0,725	0,973	0,991	0,982	0,864	0,854	0,859	0,914	0,829	0,870
	7	0,896	0,917	0,906	0,683	0,756	0,718	0,993	0,984	0,989	0,825	0,895	0,859	0,891	0,845	0,867
	8	0,930	0,898	0,914	0,710	0,786	0,746	0,984	0,989	0,987	0,833	0,886	0,859	0,919	0,819	0,866
	9	0,948	0,888	0,917	0,691	0,821	0,750	0,982	0,991	0,987	0,851	0,886	0,868	0,925	0,767	0,839
	Mittelwert	0,910	0,923	0,916	0,707	0,783	0,742	0,986	0,988	0,987	0,842	0,879	0,860	0,901	0,828	0,862

Tabelle A.2: Ergebnisse je Datensatz und Parameter-Variante mit ΔF1 zum Standard-Mittelwert

Konfiguration	Durchlauf	Abt–Buy				Amazon–Google				DBLP–ACM				Organizations				Walmart–Amazon			
		P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1
batch size	1	0,845	0,874	0,859	–0,017	0,703	0,799	0,748	–0,002	0,978	0,989	0,983	–0,003	0,760	0,878	0,815	–0,020	0,820	0,876	0,847	–0,015
	2	0,777	0,864	0,818	–0,058	0,729	0,803	0,764	0,014	0,984	0,989	0,987	0,001	0,809	0,863	0,835	0,000	0,872	0,808	0,839	–0,023
	3	0,858	0,879	0,868	–0,008	0,697	0,786	0,739	–0,011	0,980	0,993	0,987	0,001	0,762	0,883	0,818	–0,017	0,911	0,793	0,848	–0,014
	4	0,931	0,850	0,888	0,012	0,695	0,808	0,747	–0,003	0,980	0,986	0,983	–0,003	0,789	0,877	0,831	–0,004	0,886	0,850	0,868	0,006
	5	0,803	0,830	0,816	–0,060	0,698	0,791	0,741	–0,009	0,989	0,986	0,988	0,002	0,811	0,855	0,833	–0,002	0,761	0,725	0,743	–0,119
	6	0,890	0,903	0,896	0,020	0,656	0,756	0,702	–0,048	0,982	0,993	0,988	0,002	0,809	0,852	0,830	–0,005	0,877	0,850	0,863	0,001
	7	0,840	0,816	0,828	–0,048	0,738	0,748	0,743	–0,007	0,976	0,995	0,986	0,000	0,798	0,872	0,834	–0,001	0,847	0,834	0,841	–0,021
	8	0,902	0,850	0,875	–0,001	0,705	0,829	0,762	0,012	0,967	0,993	0,980	–0,006	0,799	0,870	0,833	–0,002	0,803	0,865	0,833	–0,029
	9	0,876	0,859	0,868	–0,008	0,661	0,833	0,737	–0,013	0,989	0,984	0,986	0,000	0,788	0,868	0,826	–0,009	0,853	0,870	0,862	0,000
	Mittelwert	0,858	0,858	0,857	–0,019	0,698	0,795	0,743	–0,007	0,981	0,990	0,985	–0,001	0,792	0,869	0,828	–0,007	0,848	0,830	0,838	–0,024
bert	1	0,793	0,854	0,822	–0,054	0,738	0,722	0,730	–0,020	0,971	0,995	0,983	–0,003	0,783	0,835	0,808	–0,027	0,859	0,788	0,822	–0,040
	2	0,861	0,840	0,850	–0,026	0,714	0,756	0,734	–0,016	0,969	0,995	0,982	–0,004	0,780	0,847	0,812	–0,023	0,802	0,819	0,810	–0,052
	3	0,793	0,854	0,822	–0,054	0,746	0,679	0,711	–0,039	0,969	0,995	0,982	–0,004	0,787	0,841	0,813	–0,022	0,834	0,808	0,821	–0,041
	4	0,853	0,816	0,834	–0,042	0,707	0,782	0,742	–0,008	0,978	0,993	0,985	–0,001	0,761	0,851	0,803	–0,032	0,824	0,824	0,824	–0,038
	5	0,782	0,835	0,808	–0,068	0,717	0,833	0,771	0,021	0,986	0,984	0,985	–0,001	0,805	0,838	0,821	–0,014	0,808	0,829	0,818	–0,044
	6	0,850	0,850	0,850	–0,026	0,708	0,735	0,721	–0,029	0,976	0,995	0,986	0,000	0,778	0,850	0,812	–0,023	0,905	0,793	0,845	–0,017
	7	0,763	0,845	0,802	–0,074	0,760	0,744	0,752	0,002	0,976	0,995	0,986	0,000	0,771	0,857	0,812	–0,023	0,856	0,803	0,829	–0,033
	8	0,759	0,874	0,813	–0,063	0,720	0,735	0,727	–0,023	0,971	0,993	0,982	–0,004	0,756	0,860	0,805	–0,030	0,851	0,798	0,824	–0,038
	9	0,831	0,835	0,833	–0,043	0,713	0,765	0,738	–0,012	0,976	0,993	0,984	–0,002	0,779	0,856	0,816	–0,019	0,891	0,762	0,821	–0,041
	Mittelwert	0,809	0,845	0,826	–0,050	0,725	0,750	0,736	–0,014	0,975	0,993	0,984	–0,002	0,778	0,848	0,811	–0,024	0,848	0,803	0,824	–0,038
distilbert	1	0,748	0,893	0,814	–0,062	0,658	0,765	0,708	–0,042	0,969	0,993	0,981	–0,005	0,757	0,849	0,801	–0,034	0,710	0,813	0,758	–0,104
	2	0,838	0,801	0,819	–0,057	0,688	0,765	0,725	–0,025	0,969	0,991	0,980	–0,006	0,764	0,861	0,810	–0,025	0,823	0,772	0,797	–0,065
	3	0,771	0,850	0,808	–0,068	0,681	0,765	0,720	–0,030	0,971	0,995	0,983	–0,003	0,784	0,851	0,816	–0,019	0,830	0,808	0,819	–0,043
	4	0,759	0,840	0,797	–0,079	0,701	0,791	0,743	–0,007	0,974	0,995	0,984	–0,002	0,764	0,854	0,807	–0,028	0,827	0,767	0,796	–0,066
	5	0,796	0,835	0,815	–0,061	0,689	0,778	0,731	–0,019	0,989	0,991	0,990	0,004	0,766	0,856	0,808	–0,027	0,770	0,813	0,791	–0,071
	6	0,808	0,816	0,812	–0,064	0,707	0,774	0,739	–0,011	0,971	0,998	0,984	–0,002	0,791	0,845	0,817	–0,018	0,803	0,741	0,771	–0,091
	7	0,761	0,850	0,803	–0,073	0,687	0,808	0,743	–0,007	0,967	0,995	0,981	–0,005	0,775	0,846	0,809	–0,026	0,790	0,762	0,776	–0,086
	8	0,794	0,786	0,790	–0,086	0,670	0,782	0,722	–0,028	0,969	0,998	0,983	–0,003	0,762	0,843	0,801	–0,034	0,822	0,720	0,768	–0,094
	9	0,733	0,864	0,793	–0,083	0,682	0,761	0,719	–0,031	0,969	0,993	0,981	–0,005	0,752	0,852	0,799	–0,036	0,803	0,803	0,803	–0,059
	Mittelwert	0,779	0,837	0,806	–0,070	0,685	0,777	0,728	–0,022	0,972	0,994	0,983	–0,003	0,768	0,851	0,808	–0,027	0,798	0,778	0,787	–0,075
epochs	1	0,900	0,874	0,887	0,011	0,731	0,756	0,744	–0,006	0,974	0,993	0,983	–0,003	0,794	0,864	0,828	–0,007	0,916	0,845	0,879	0,017
	2	0,913	0,864	0,888	0,012	0,756	0,688	0,720	–0,030	0,978	0,995	0,987	0,001	0,769	0,879	0,821	–0,014	0,900	0,839	0,869	0,007
	3	0,909	0,825	0,865	–0,011	0,740	0,765	0,752	0,002	0,978	0,989	0,983	–0,003	0,800	0,866	0,832	–0,003	0,764	0,756	0,760	–0,102
	4	0,896	0,879	0,887	0,011	0,685	0,782	0,731	–0,019	0,984	0,986	0,985	–0,001	0,808	0,856	0,831	–0,004	0,899	0,829	0,863	0,001
	5	0,901	0,888	0,895	0,019	0,679	0,769	0,721	–0,029	0,982	0,989	0,985	–0,001	0,799	0,866	0,831	–0,004	0,864	0,855	0,859	–0,003
	6	0,902	0,850	0,875	–0,001	0,714	0,748	0,731	–0,019	0,984	0,991	0,988	0,002	0,776	0,866	0,818	–0,017	0,795	0,824	0,809	–0,053
	7	0,898	0,859	0,878	0,002	0,723	0,735	0,729	–0,021	0,984	0,989	0,987	0,001	0,811	0,855	0,832	–0,003	0,871	0,839	0,855	–0,007
	8	0,853	0,874	0,863	–0,013	0,739	0,761	0,749	–0,001	0,982	0,993	0,988	0,002	0,762	0,878	0,816	–0,019	0,894	0,829	0,860	–0,002
	9	0,920	0,888	0,904	0,028	0,772	0,765	0,768	0,018	0,978	0,993	0,985	–0,001	0,775	0,882	0,825	–0,010	0,872	0,850	0,861	–0,001
	Mittelwert	0,899	0,867	0,882	0,006	0,727	0,752	0,738	–0,012	0,980	0,991	0,986	0,000	0,788	0,868	0,826	–0,009	0,864	0,830	0,846	–0,016

Auf der nächsten Seite weitergeführt

Tabelle A.2: Ergebnisse je Datensatz und Parameter-Variante mit ΔF1 zum Standard-Mittelwert (Fortsetzung)

Konfiguration	Durchlauf	Abt–Buy				Amazon–Google				DBLP–ACM				Organizations				Walmart–Amazon			
		P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1	P	R	F1	ΔF1
learning rate	1	0,882	0,869	0,875	–0,001	0,620	0,859	0,720	–0,030	0,982	0,989	0,985	–0,001	0,815	0,865	0,839	0,004	0,895	0,798	0,844	–0,018
	2	0,906	0,888	0,897	0,021	0,712	0,791	0,749	–0,001	0,980	0,993	0,987	0,001	0,759	0,897	0,822	–0,013	0,825	0,855	0,840	–0,022
	3	0,906	0,893	0,900	0,024	0,707	0,825	0,761	0,011	0,982	0,989	0,985	–0,001	0,810	0,873	0,840	0,005	0,893	0,824	0,857	–0,005
	4	0,901	0,883	0,892	0,016	0,716	0,829	0,768	0,018	0,982	0,991	0,987	0,001	0,804	0,871	0,836	0,001	0,891	0,845	0,867	0,005
	5	0,878	0,908	0,893	0,017	0,727	0,829	0,774	0,024	0,989	0,991	0,990	0,004	0,795	0,880	0,835	0,000	0,867	0,876	0,871	0,009
	6	0,921	0,854	0,887	0,011	0,688	0,812	0,745	–0,005	0,982	0,991	0,987	0,001	0,821	0,871	0,845	0,010	0,865	0,829	0,847	–0,015
	7	0,923	0,869	0,895	0,019	0,672	0,859	0,754	0,004	0,976	0,995	0,986	0,000	0,781	0,893	0,834	–0,001	0,883	0,824	0,853	–0,009
	8	0,861	0,874	0,867	–0,009	0,725	0,778	0,751	0,001	0,982	0,991	0,987	0,001	0,786	0,883	0,832	–0,003	0,906	0,845	0,874	0,012
	9	0,822	0,917	0,867	–0,009	0,711	0,842	0,771	0,021	0,978	0,993	0,985	–0,001	0,814	0,876	0,844	0,009	0,773	0,829	0,800	–0,062
	Mittelwert	0,889	0,884	0,886	0,010	0,698	0,825	0,755	0,005	0,981	0,991	0,987	0,001	0,798	0,879	0,836	0,001	0,866	0,836	0,850	–0,012
no da	1	0,889	0,898	0,894	0,018	0,699	0,774	0,734	–0,016	0,965	0,993	0,979	–0,007	0,763	0,879	0,817	–0,018	0,886	0,850	0,868	0,006
	2	0,899	0,903	0,901	0,025	0,695	0,778	0,734	–0,016	0,982	0,991	0,987	0,001	0,763	0,879	0,817	–0,018	0,880	0,876	0,878	0,016
	3	0,843	0,888	0,865	–0,011	0,710	0,786	0,746	–0,004	0,980	0,998	0,989	0,003	0,792	0,871	0,830	–0,005	0,901	0,850	0,875	0,013
	4	0,924	0,883	0,903	0,027	0,693	0,821	0,751	0,001	0,982	0,993	0,988	0,002	0,792	0,871	0,830	–0,005	0,904	0,829	0,865	0,003
	5	0,844	0,922	0,882	0,006	0,752	0,765	0,758	0,008	0,978	0,991	0,984	–0,002	0,758	0,884	0,816	–0,019	0,902	0,860	0,881	0,019
	6	0,948	0,879	0,912	0,036	0,729	0,769	0,748	–0,002	0,987	0,989	0,988	0,002	0,758	0,884	0,816	–0,019	0,880	0,834	0,856	–0,006
	7	0,861	0,874	0,867	–0,009	0,736	0,812	0,772	0,022	0,980	0,986	0,983	–0,003	0,770	0,889	0,825	–0,010	0,886	0,850	0,868	0,006
	8	0,809	0,966	0,881	0,005	0,693	0,752	0,721	–0,029	0,984	0,991	0,988	0,002	0,770	0,889	0,825	–0,010	0,838	0,829	0,833	–0,029
	9	0,837	0,850	0,843	–0,033	0,745	0,761	0,753	0,003	0,984	0,991	0,988	0,002	0,788	0,879	0,831	–0,004	0,862	0,839	0,850	–0,012
	Mittelwert	0,873	0,896	0,883	0,007	0,717	0,780	0,746	–0,004	0,980	0,991	0,986	0,000	0,773	0,881	0,823	–0,012	0,882	0,846	0,864	0,002
no da, dk	1	0,914	0,874	0,893	0,017	0,708	0,799	0,751	0,001	0,982	0,995	0,989	0,003	0,749	0,881	0,810	–0,025	0,868	0,855	0,862	0,000
	2	0,913	0,869	0,891	0,015	0,722	0,744	0,733	–0,017	0,976	0,995	0,986	0,000	0,772	0,869	0,818	–0,017	0,885	0,834	0,859	–0,003
	3	0,919	0,883	0,901	0,025	0,715	0,803	0,757	0,007	0,980	0,993	0,987	0,001	0,772	0,885	0,825	–0,010	0,868	0,850	0,859	–0,003
	4	0,859	0,859	0,859	–0,017	0,691	0,821	0,750	0,000	0,986	0,984	0,985	–0,001	0,763	0,883	0,819	–0,016	0,806	0,839	0,822	–0,040
	5	0,880	0,893	0,887	0,011	0,733	0,786	0,759	0,009	0,980	0,989	0,984	–0,002	0,755	0,883	0,814	–0,021	0,902	0,860	0,881	0,019
	6	0,906	0,893	0,900	0,024	0,715	0,795	0,753	0,003	0,973	0,991	0,982	–0,004	0,772	0,886	0,825	–0,010	0,856	0,860	0,858	–0,004
	7	0,925	0,898	0,911	0,035	0,700	0,756	0,727	–0,023	0,980	0,998	0,989	0,003	0,766	0,871	0,815	–0,020	0,909	0,876	0,892	0,030
	8	0,919	0,879	0,898	0,022	0,720	0,791	0,754	0,004	0,978	0,991	0,984	–0,002	0,744	0,888	0,809	–0,026	0,822	0,813	0,818	–0,044
	9	0,882	0,908	0,895	0,019	0,695	0,769	0,730	–0,020	0,980	0,991	0,985	–0,001	0,775	0,875	0,822	–0,013	0,823	0,865	0,843	–0,019
	Mittelwert	0,902	0,884	0,893	0,017	0,711	0,785	0,746	–0,004	0,979	0,992	0,986	0,000	0,763	0,880	0,817	–0,018	0,860	0,850	0,855	–0,007
no dk	1	0,923	0,879	0,900	0,024	0,736	0,812	0,772	0,022	0,982	0,995	0,989	0,003	0,812	0,863	0,837	0,002	0,877	0,850	0,863	0,001
	2	0,901	0,888	0,895	0,019	0,721	0,774	0,746	–0,004	0,982	0,986	0,984	–0,002	0,811	0,861	0,835	0,000	0,896	0,845	0,869	0,007
	3	0,766	0,888	0,822	–0,054	0,730	0,808	0,767	0,017	0,976	0,993	0,984	–0,002	0,814	0,857	0,835	0,000	0,906	0,850	0,877	0,015
	4	0,915	0,888	0,901	0,025	0,710	0,816	0,759	0,009	0,978	0,991	0,984	–0,002	0,813	0,873	0,842	0,007	0,901	0,850	0,875	0,013
	5	0,905	0,879	0,892	0,016	0,716	0,829	0,768	0,018	0,974	0,993	0,983	–0,003	0,765	0,883	0,820	–0,015	0,902	0,813	0,856	–0,006
	6	0,849	0,898	0,873	–0,003	0,723	0,782	0,752	0,002	0,980	0,993	0,987	0,001	0,809	0,873	0,840	0,005	0,854	0,850	0,852	–0,010
	7	0,920	0,893	0,906	0,030	0,686	0,812	0,744	–0,006	0,974	0,993	0,983	–0,003	0,797	0,869	0,831	–0,004	0,895	0,839	0,866	0,004
	8	0,879	0,879	0,879	0,003	0,699	0,803	0,748	–0,002	0,984	0,986	0,985	–0,001	0,783	0,881	0,829	–0,006	0,872	0,845	0,858	–0,004
	9	0,910	0,879	0,894	0,018	0,727	0,752	0,739	–0,011	0,986	0,982	0,984	–0,002	0,783	0,867	0,823	–0,012	0,875	0,870	0,873	0,011
	Mittelwert	0,885	0,886	0,885	0,009	0,716	0,799	0,755	0,005	0,980	0,990	0,985	–0,001	0,799	0,870	0,832	–0,003	0,886	0,846	0,865	0,003

B Plots zur Bestimmung des Unsicherheitsindexes

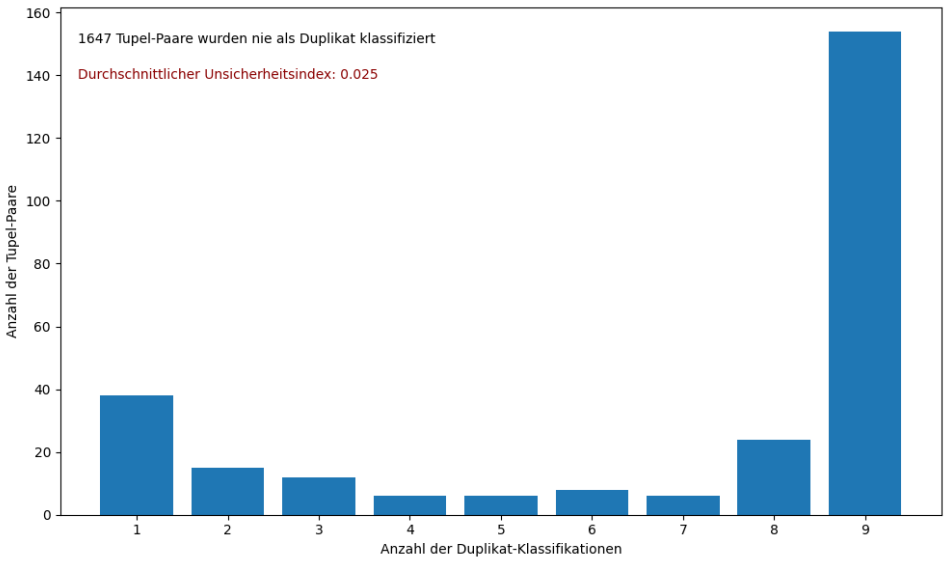

Abbildung B.1: Ditto Standard-Konfiguration auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

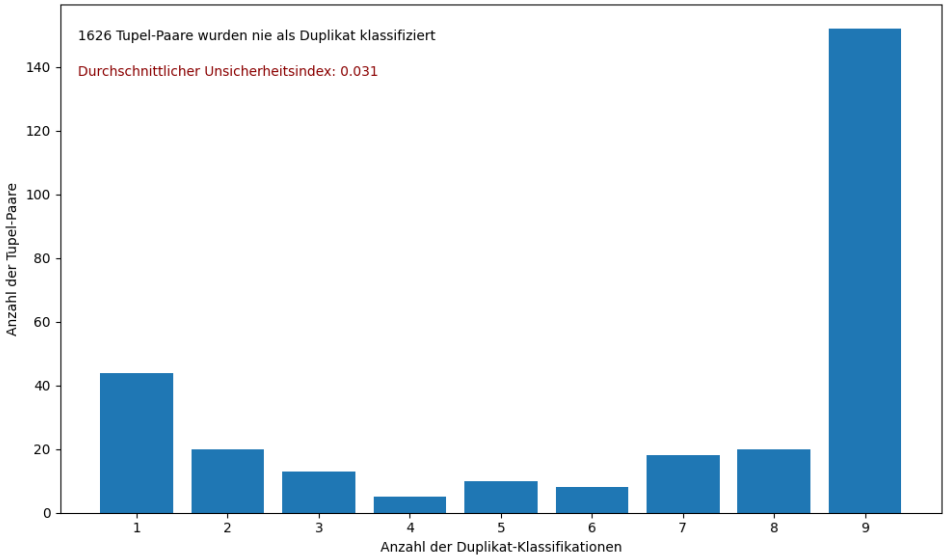

Abbildung B.2: HierGAT Standard-Konfiguration auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

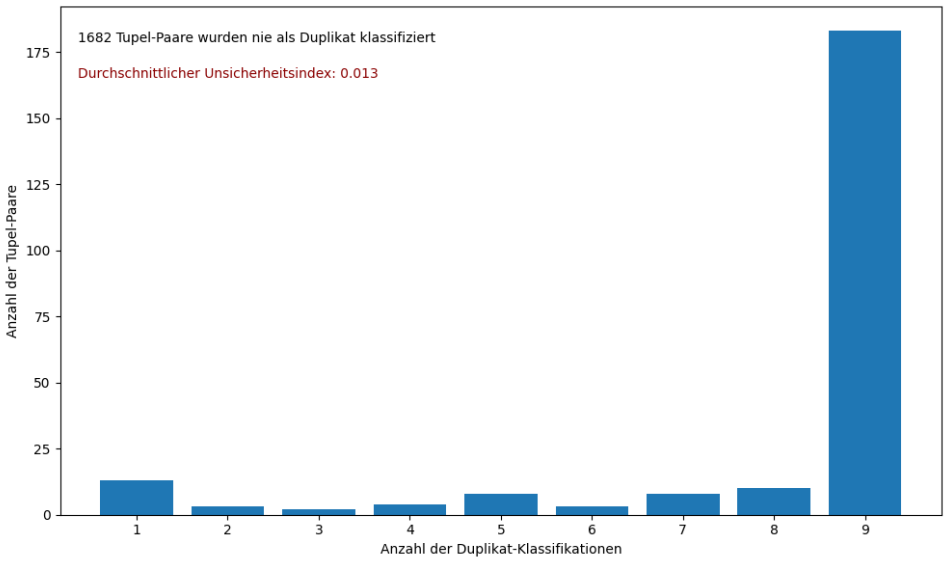

Abbildung B.3: Unicorn Standard-Konfiguration auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

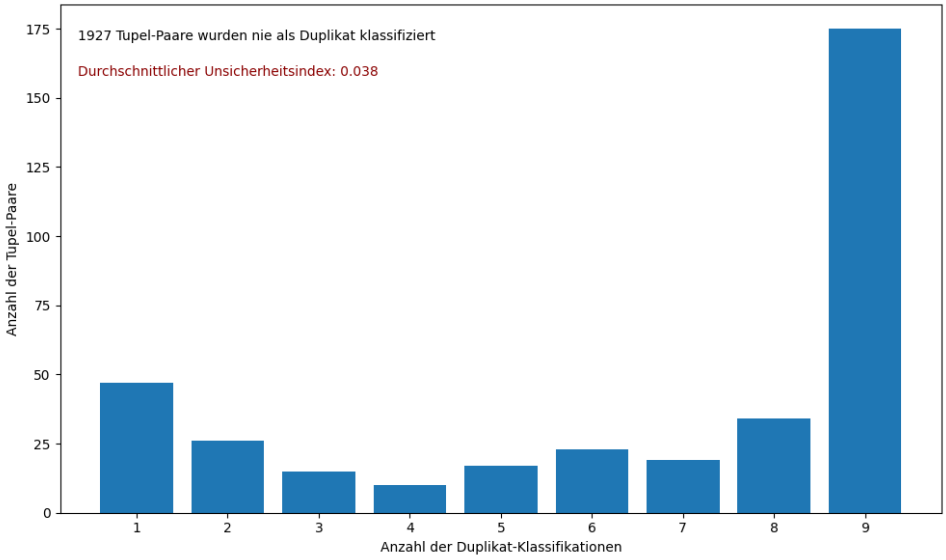

Abbildung B.4: Ditto Standard-Konfiguration auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

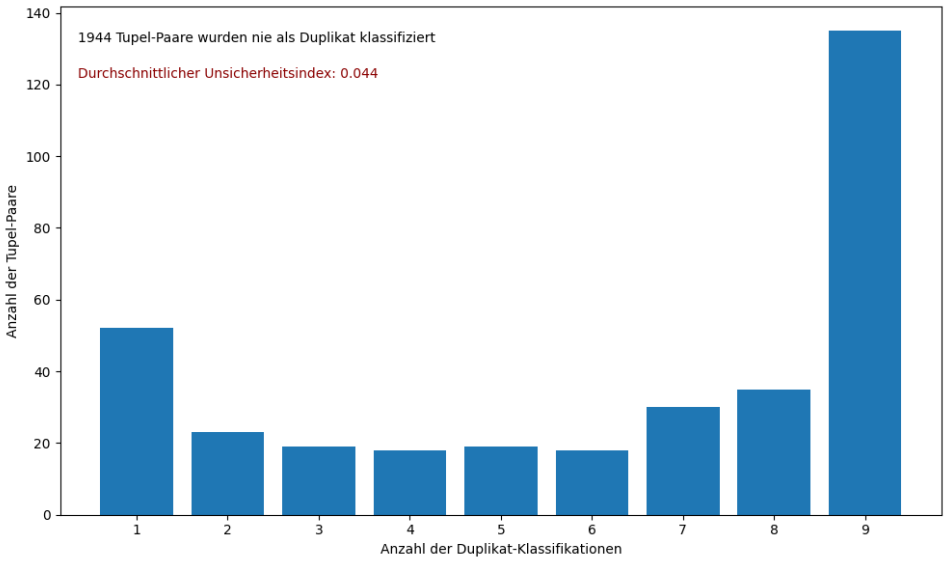

Abbildung B.5: HierGAT Standard-Konfiguration auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

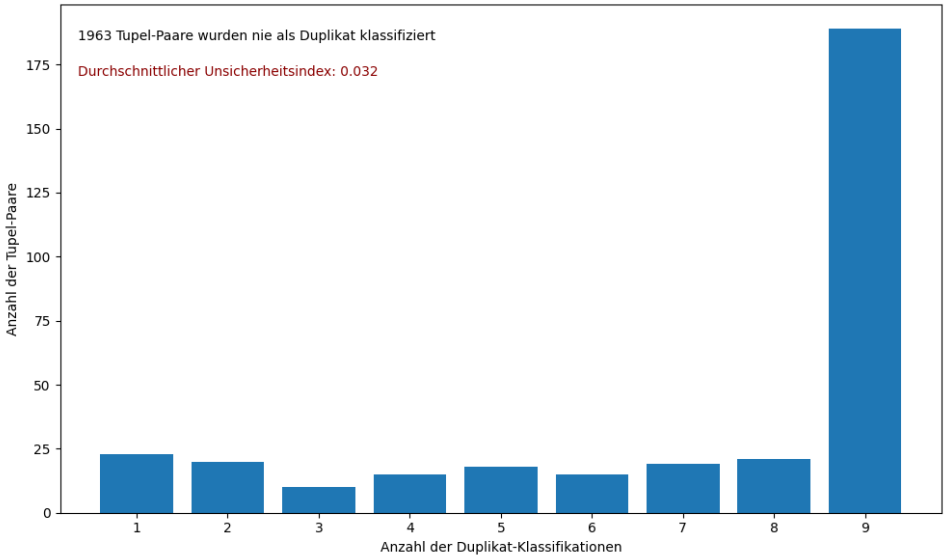

Abbildung B.6: Unicorn Standard-Konfiguration auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

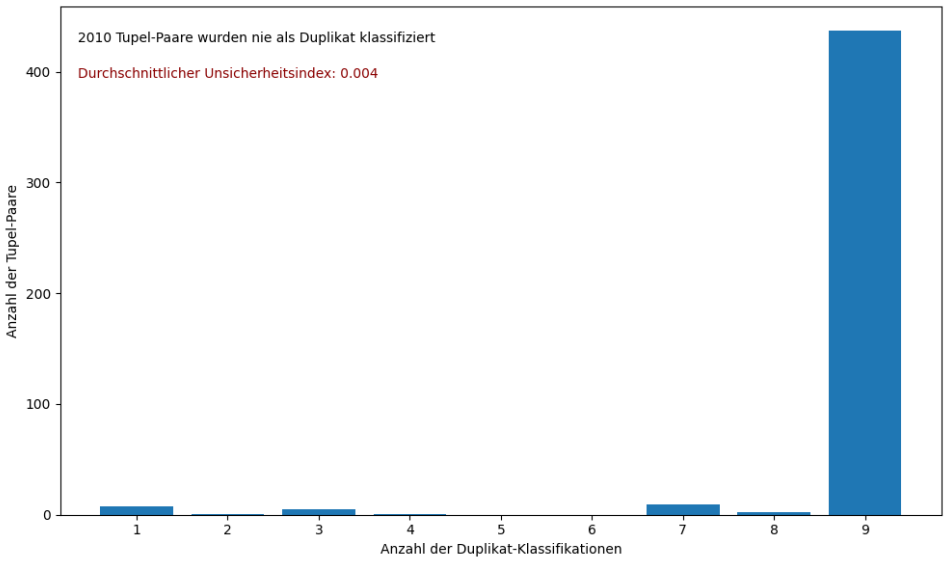

Abbildung B.7: Ditto Standard-Konfiguration auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

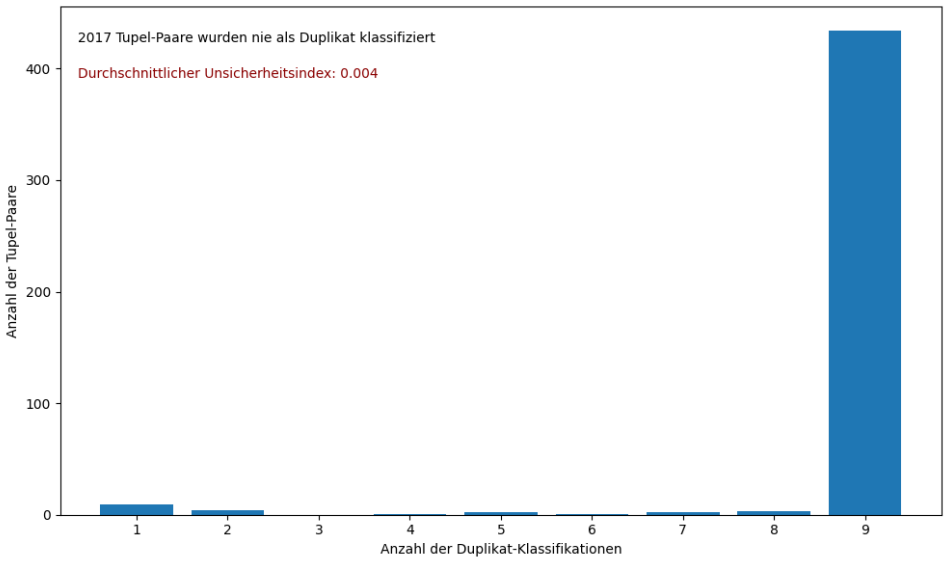

Abbildung B.8: HierGAT Standard-Konfiguration auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

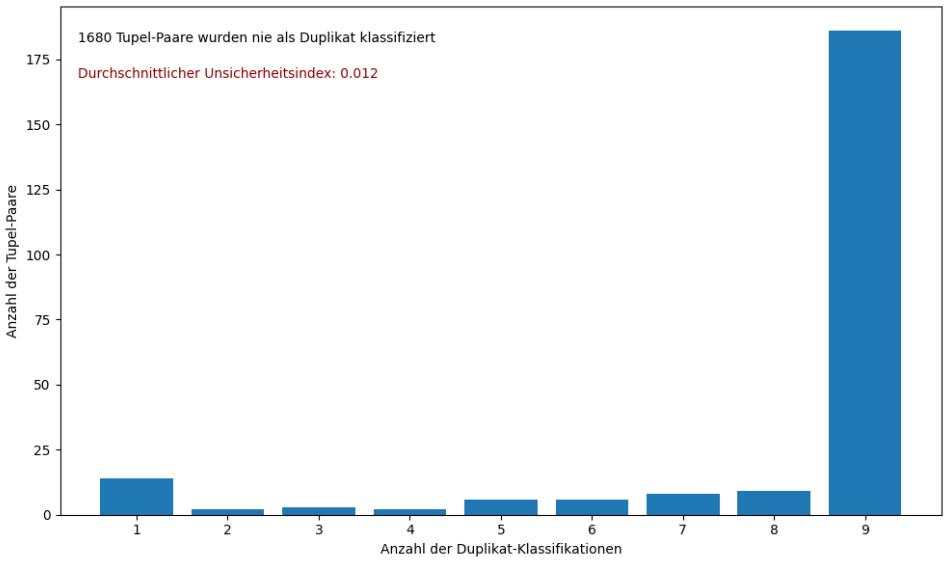

Abbildung B.9: Unicorn Standard-Konfiguration auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

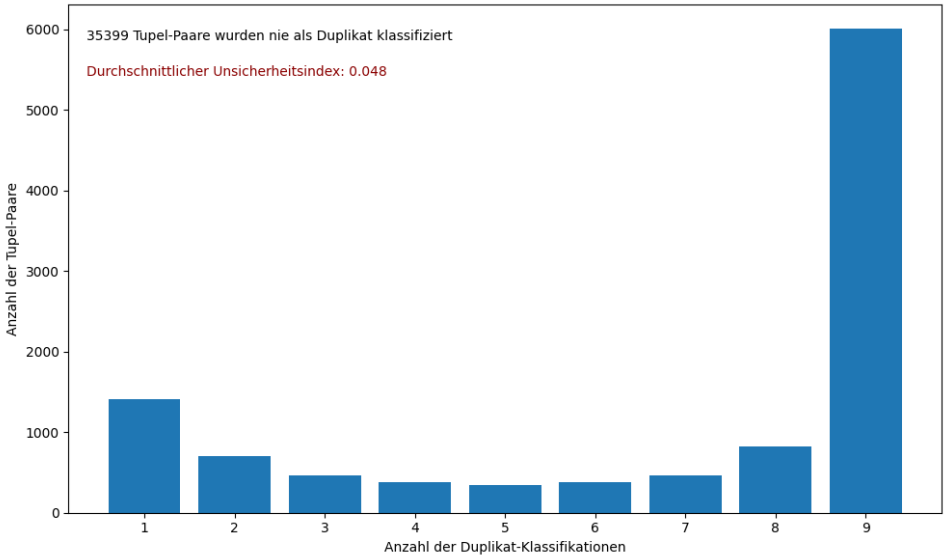

Abbildung B.10: Ditto Standard-Konfiguration auf Organizations - Plot der Übereinstimmung der Durchläufe

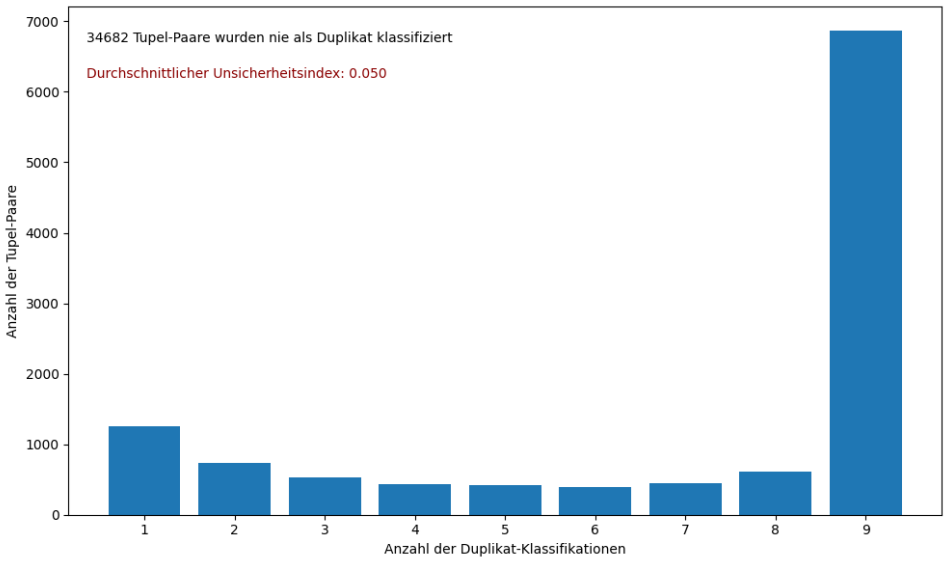

Abbildung B.11: HierGAT Standard-Konfiguration auf Organizations - Plot der Übereinstimmung der Durchläufe

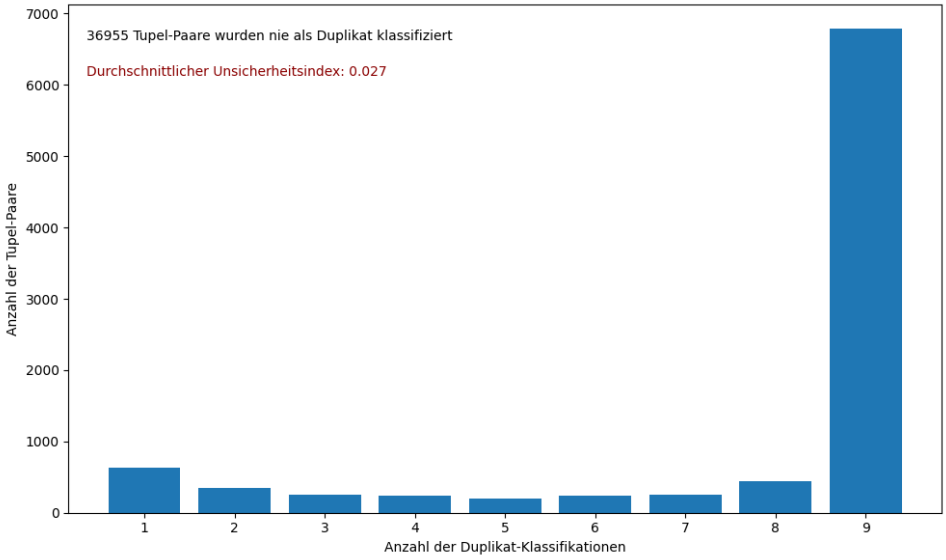

Abbildung B.12: Unicorn Standard-Konfiguration auf Organizations - Plot der Übereinstimmung der Durchläufe

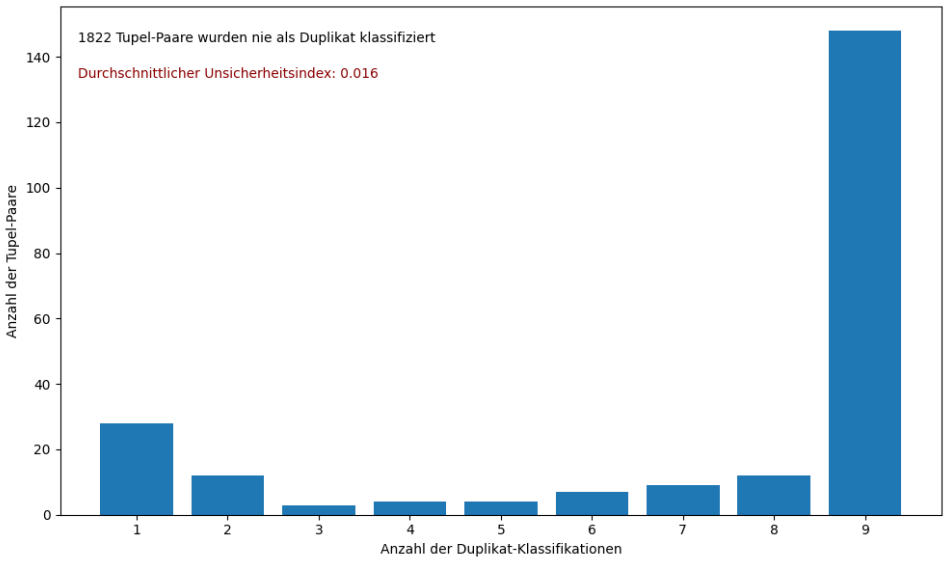

Abbildung B.13: Ditto Standard-Konfiguration auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

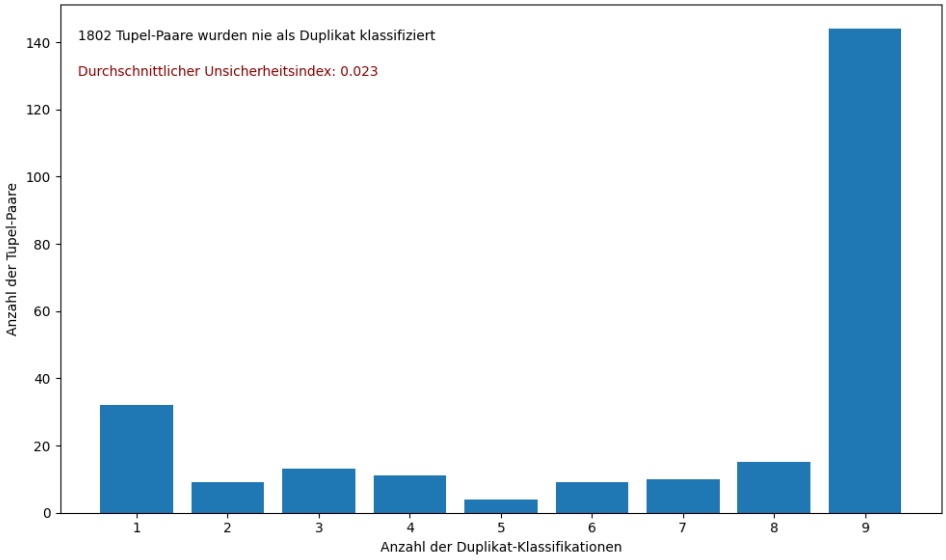

Abbildung B.14: HierGAT Standard-Konfiguration auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

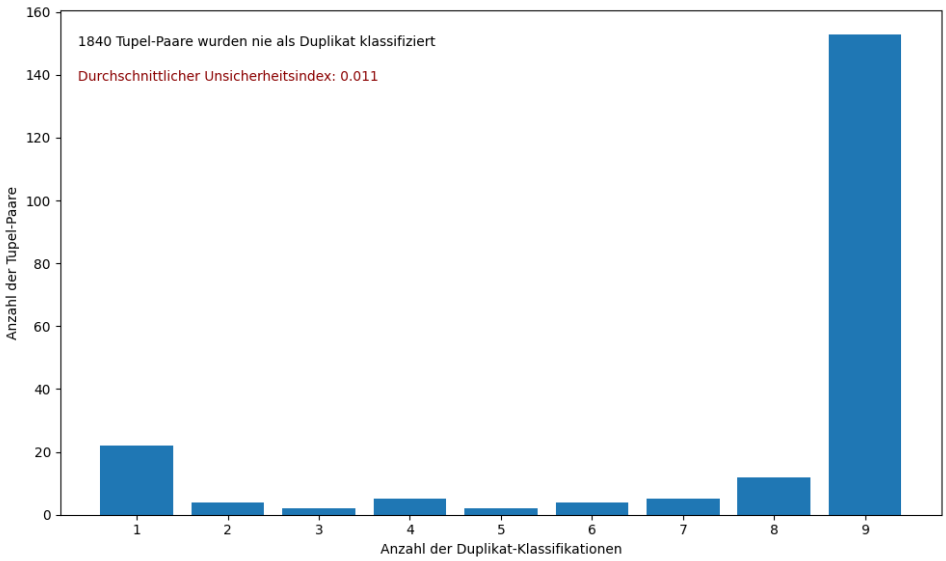

Abbildung B.15: Unicorn Standard-Konfiguration auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

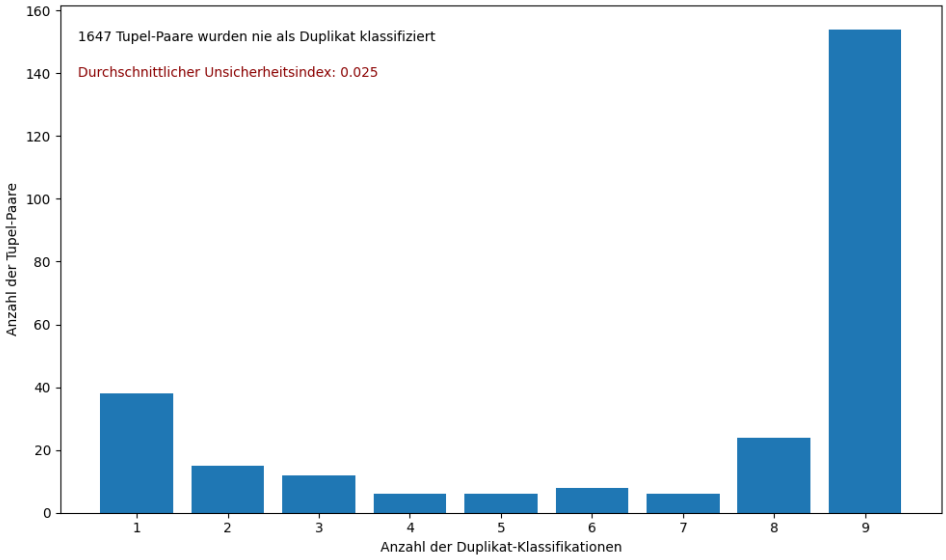

Abbildung B.16: Ditto Standard-Konfiguration auf Abt-Buy auf verschiedenen Knoten - Plot der Übereinstimmung der Durchläufe

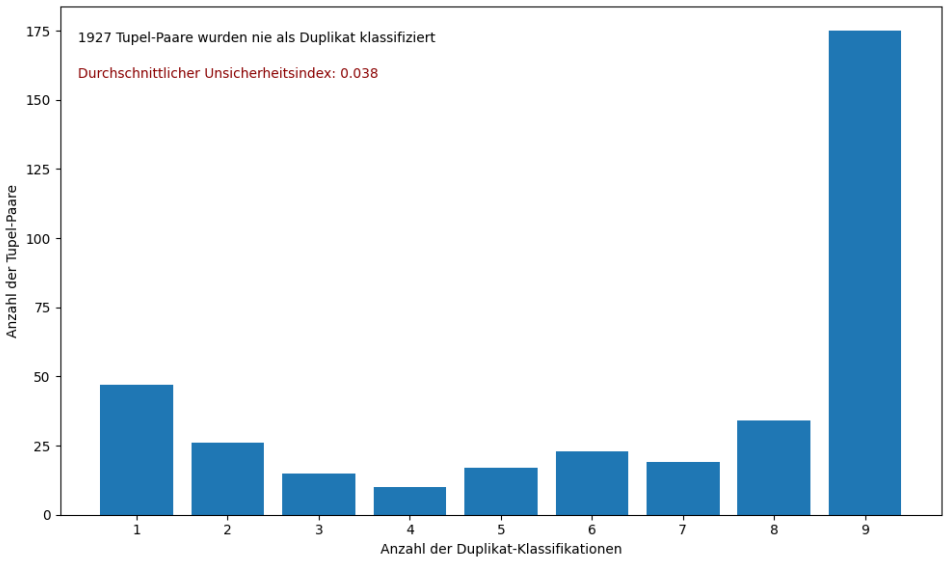

Abbildung B.17: Ditto Standard-Konfiguration auf Amazon-Google auf verschiedenen Knoten - Plot der Übereinstimmung der Durchläufe

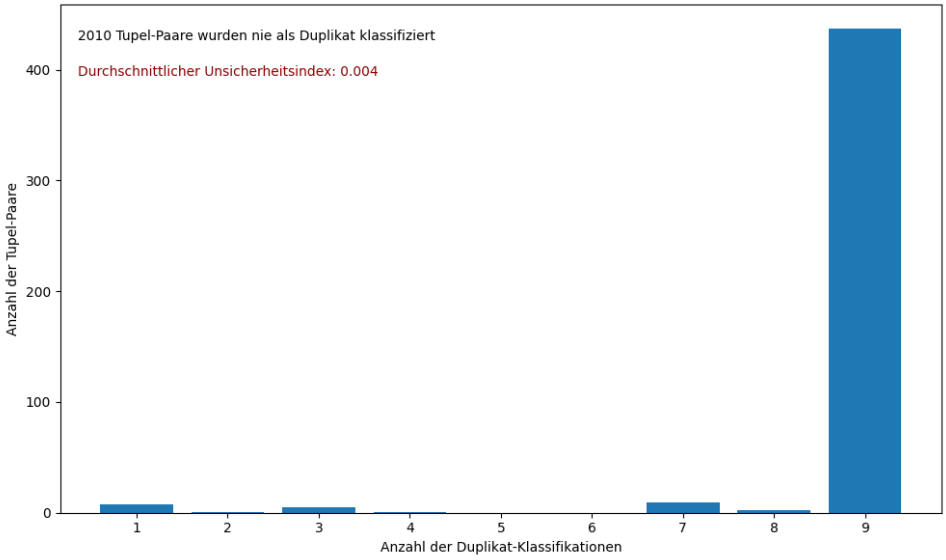

Abbildung B.18: Ditto Standard-Konfiguration auf DBLP-ACM auf verschiedenen Knoten - Plot der Übereinstimmung der Durchläufe

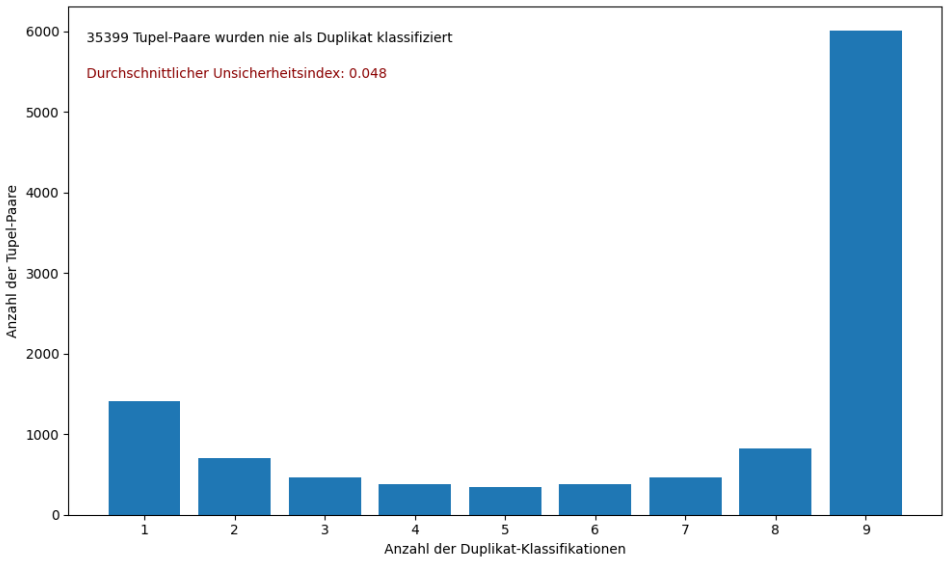

Abbildung B.19: Ditto Standard-Konfiguration auf Organizations auf verschiedenen Knoten - Plot der Übereinstimmung der Durchläufe

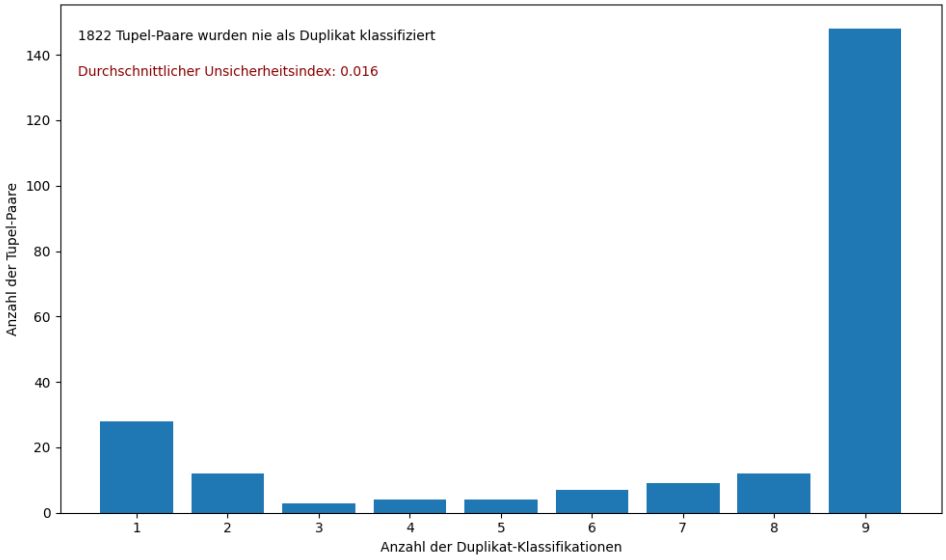

Abbildung B.20: Ditto Standard-Konfiguration auf Walmart-Amazon auf verschiedenen Knoten - Plot der Übereinstimmung der Durchläufe

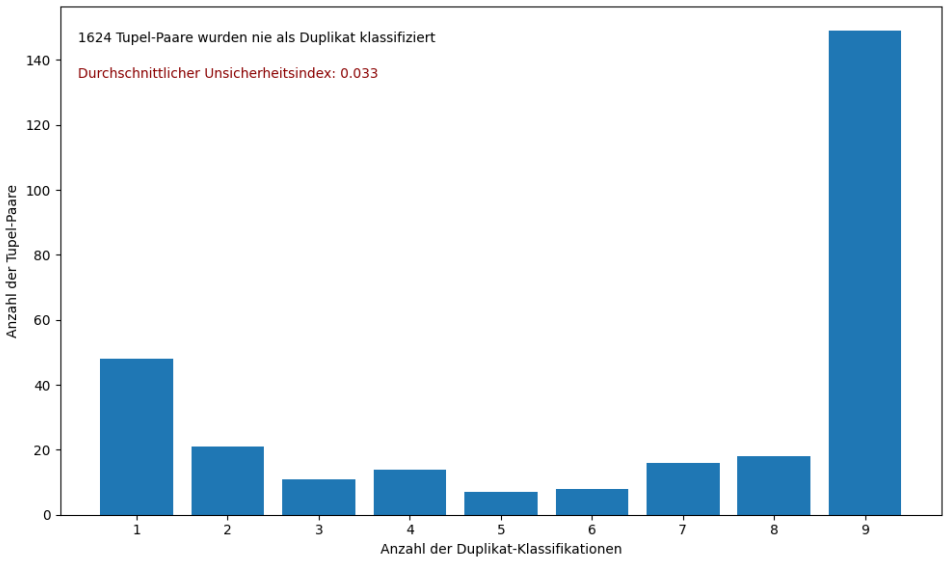

Abbildung B.21: Ditto batch size 16 auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

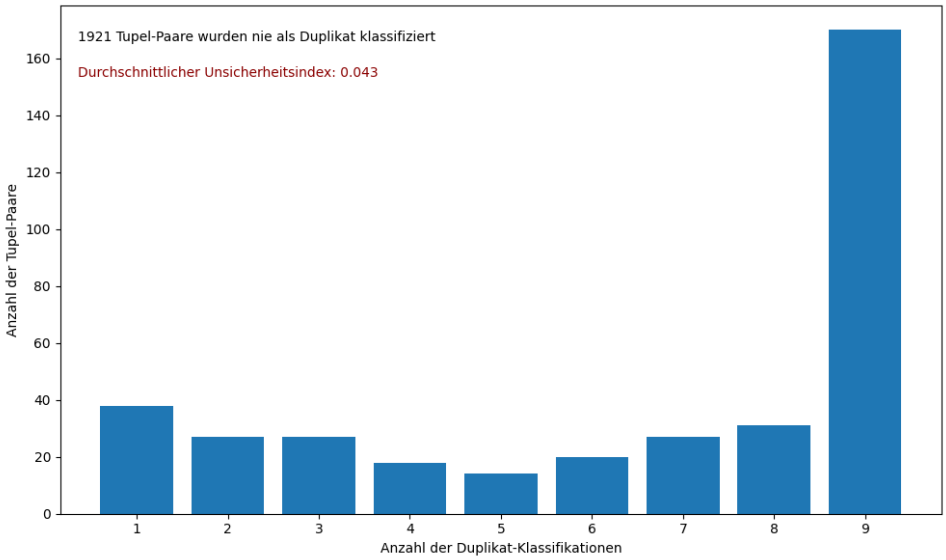

Abbildung B.22: Ditto batch size 16 auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

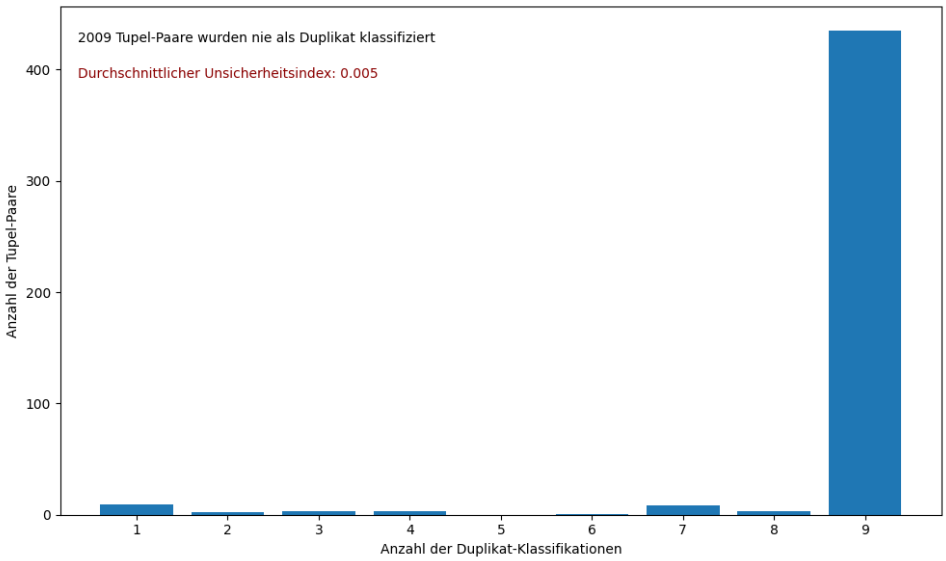

Abbildung B.23: Ditto batch size 16 auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

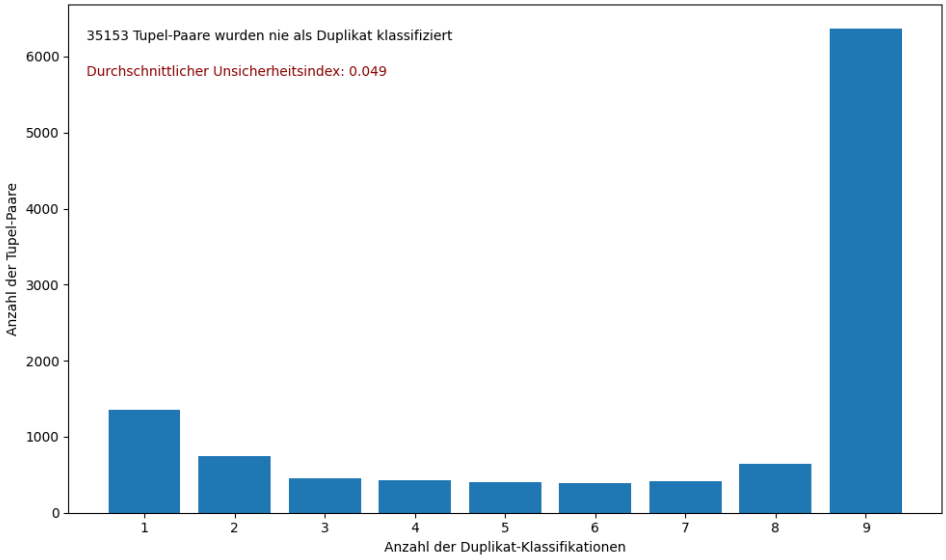

Abbildung B.24: Ditto batch size 16 auf Organizations - Plot der Übereinstimmung der Durchläufe

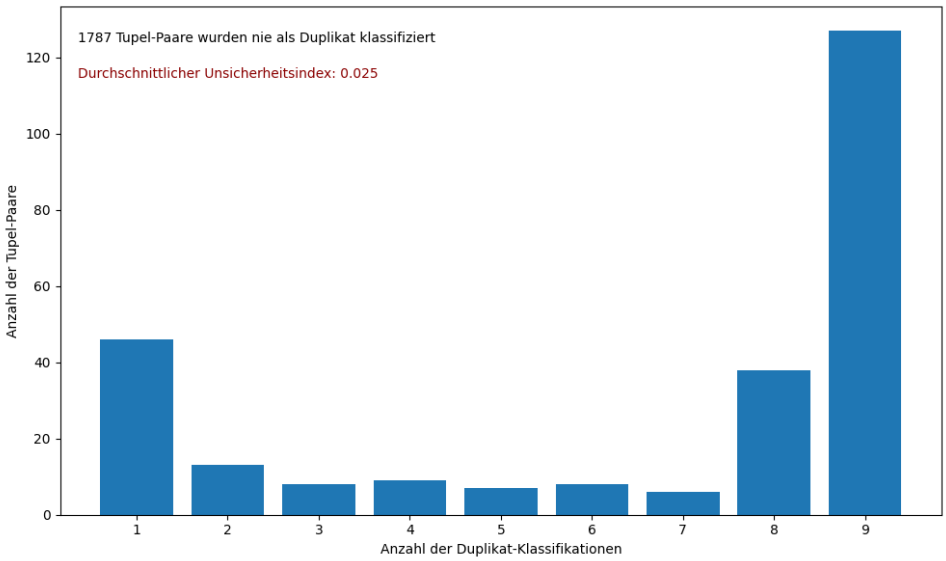

Abbildung B.25: Ditto batch size 16 auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

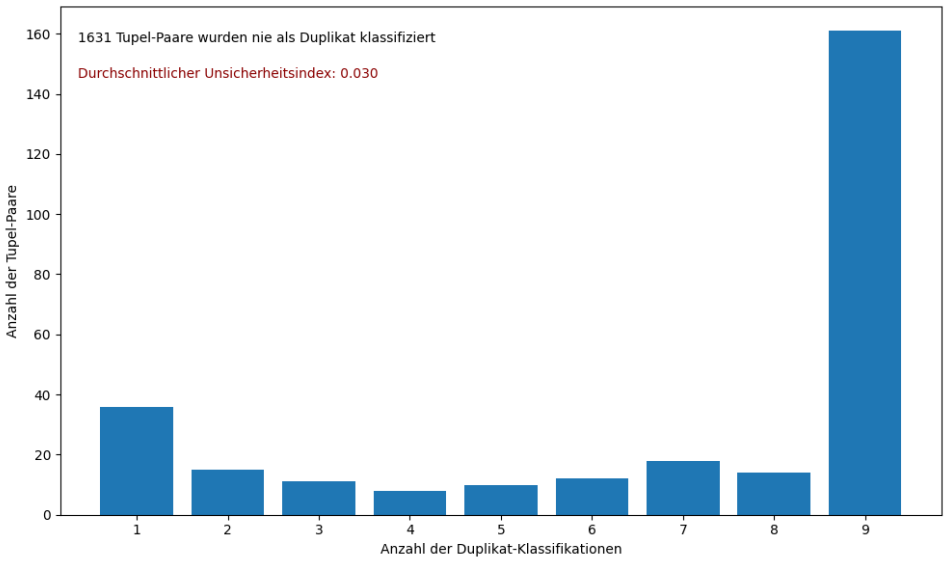

Abbildung B.26: Ditto bert auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

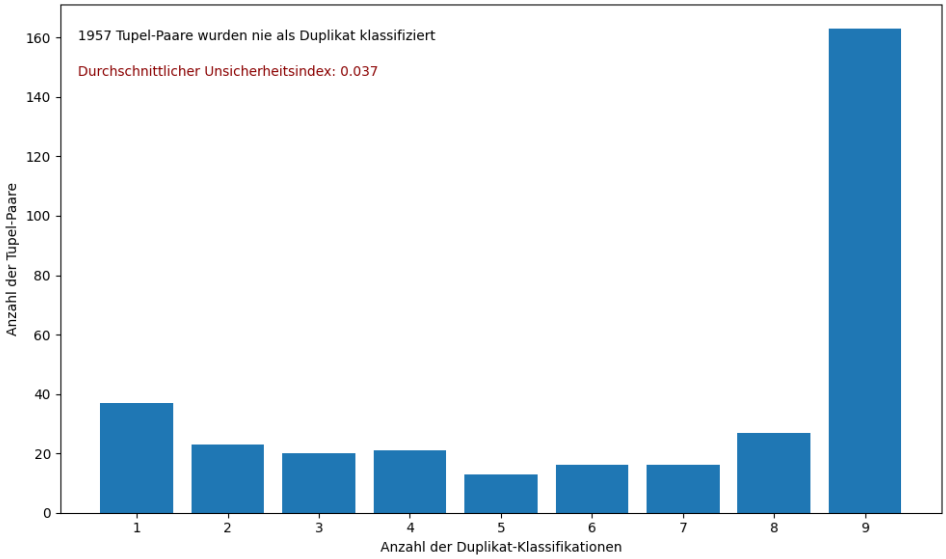

Abbildung B.27: Ditto bert auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

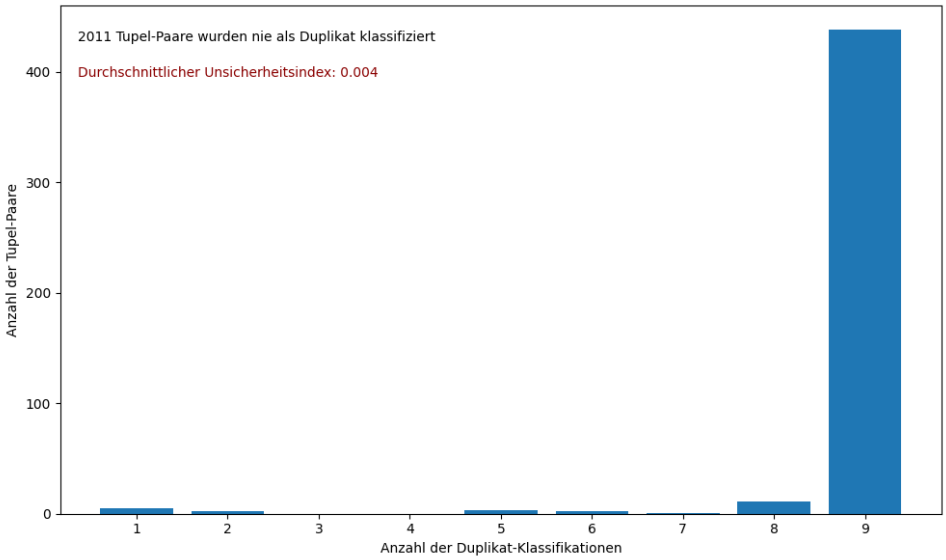

Abbildung B.28: Ditto bert auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

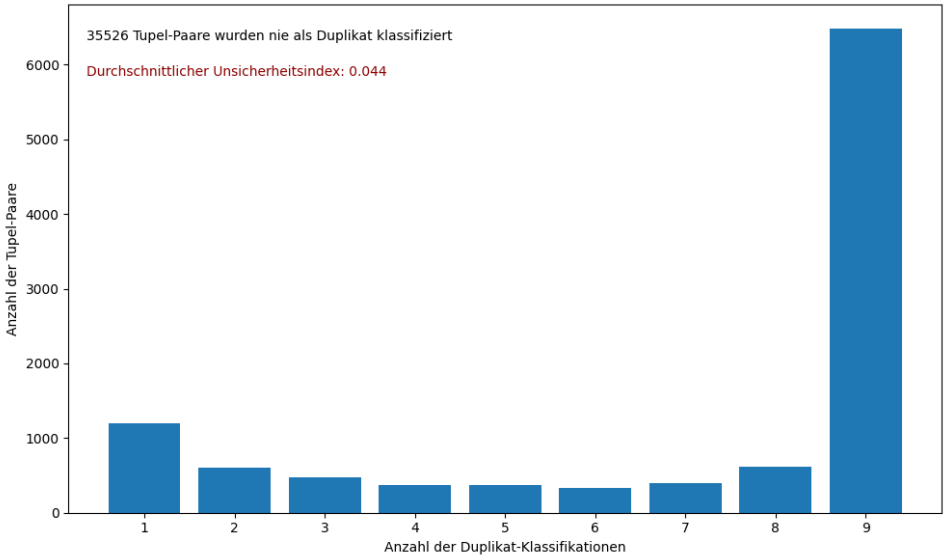

Abbildung B.29: Ditto bert auf Organizations - Plot der Übereinstimmung der Durchläufe

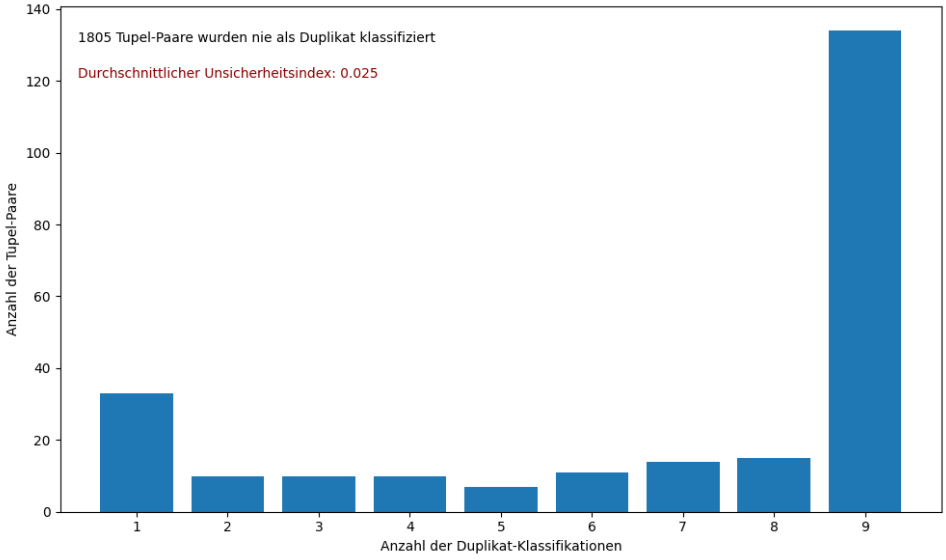

Abbildung B.30: Ditto bert auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

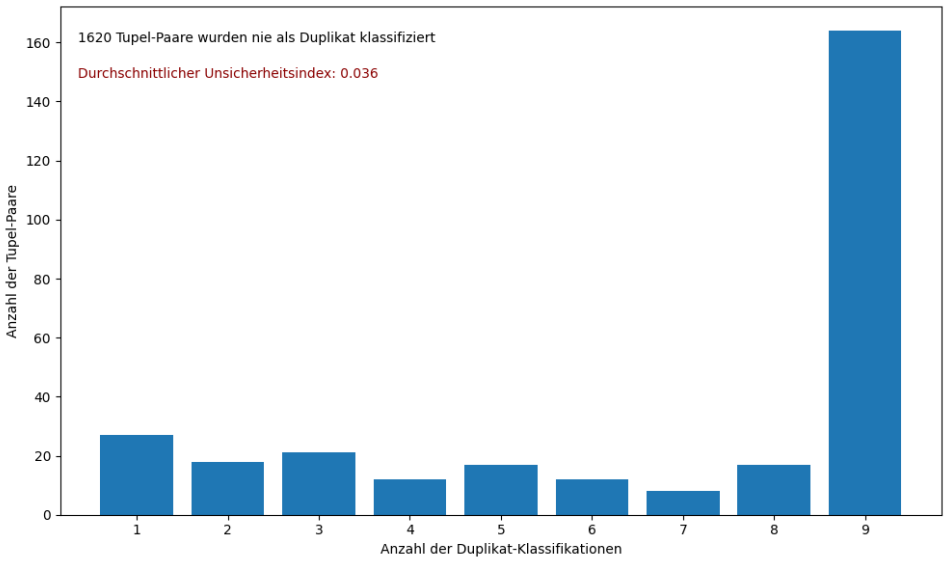

Abbildung B.31: Ditto distilbert auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

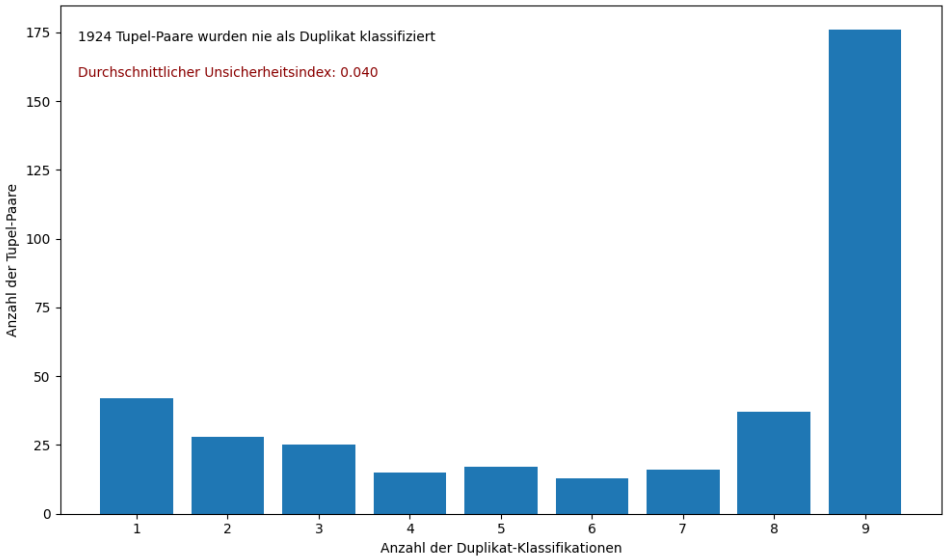

Abbildung B.32: Ditto distilbert auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

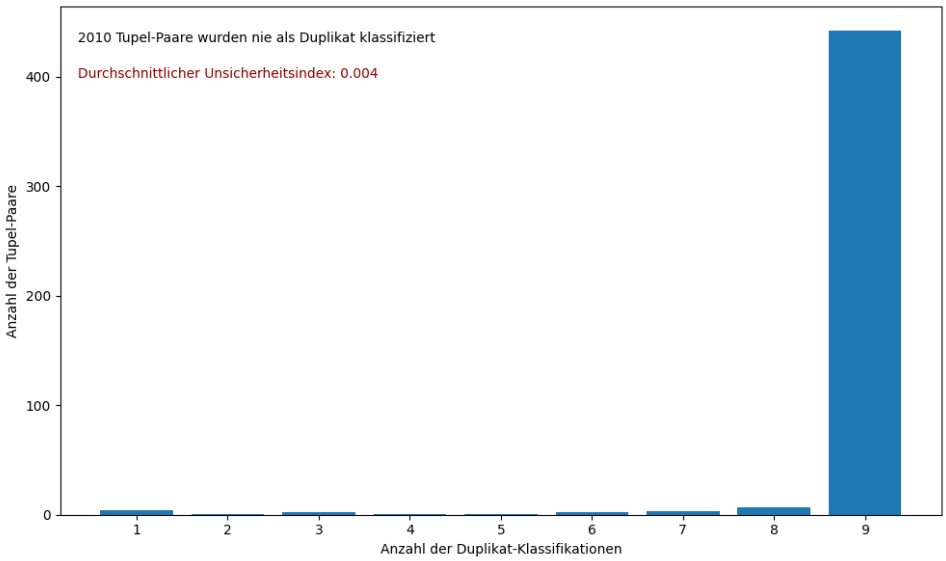

Abbildung B.33: Ditto distilbert auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

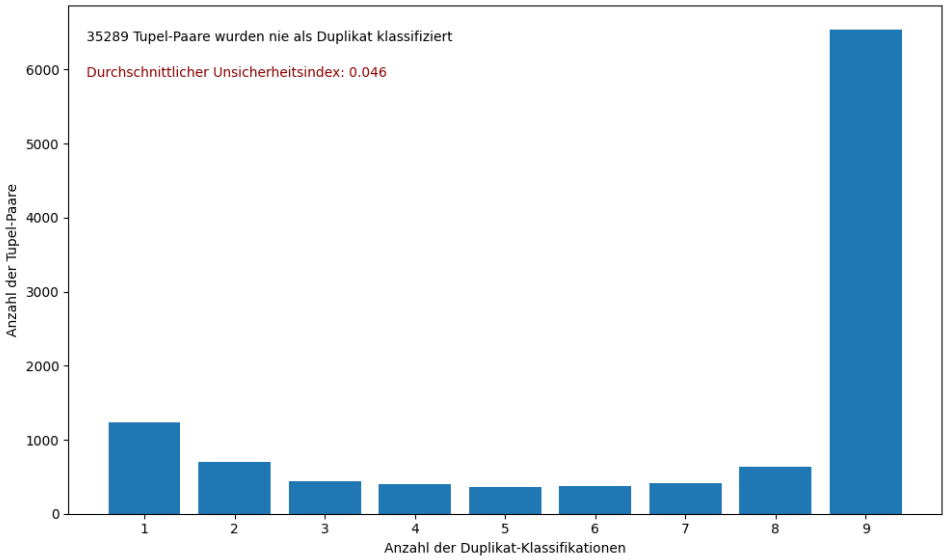

Abbildung B.34: Ditto distilbert auf Organizations - Plot der Übereinstimmung der Durchläufe

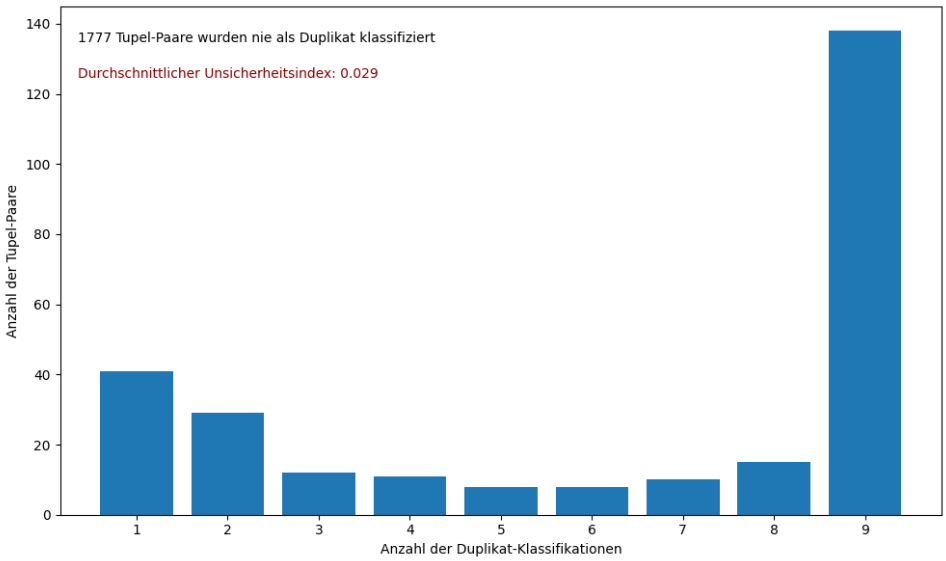

Abbildung B.35: Ditto distilbert auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

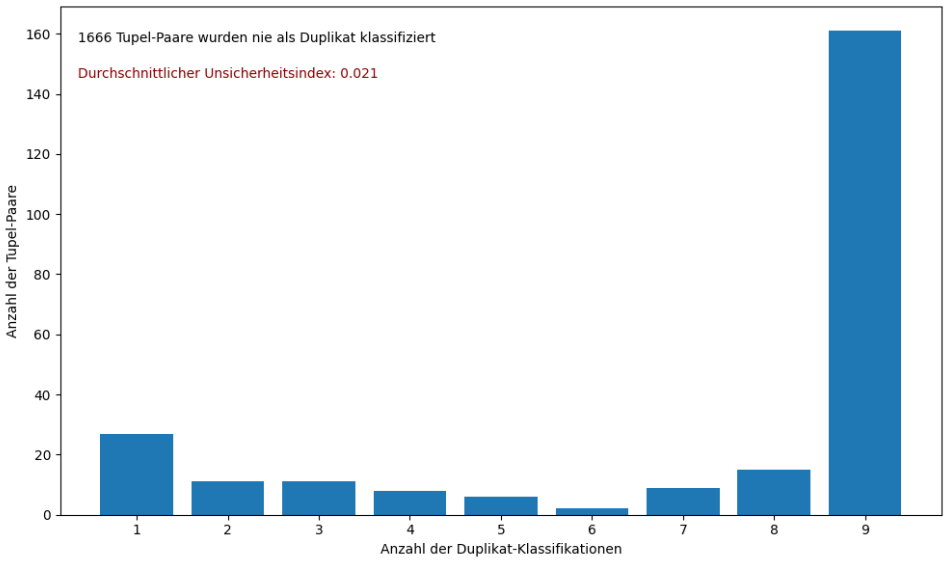

Abbildung B.36: Ditto 30 epochs auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

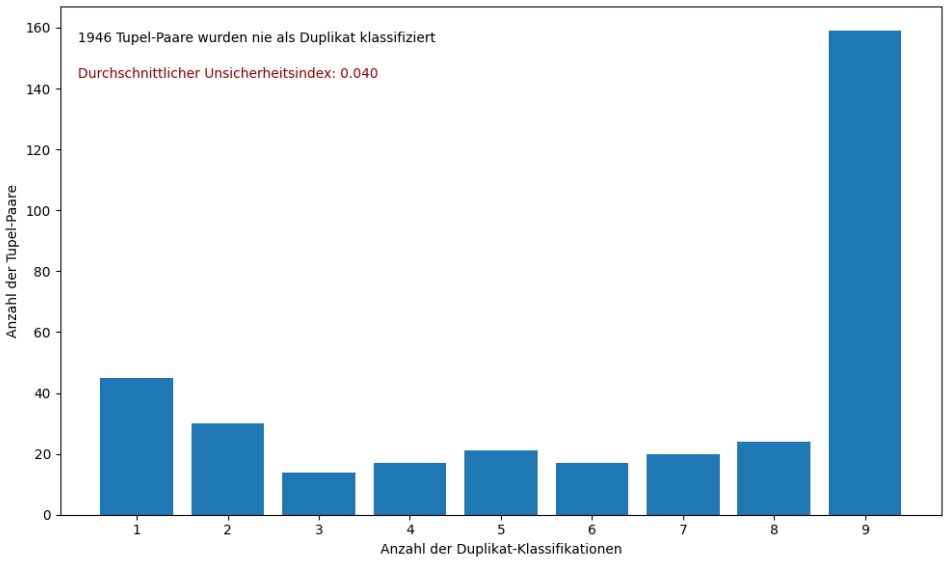

Abbildung B.37: Ditto 30 epochs auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

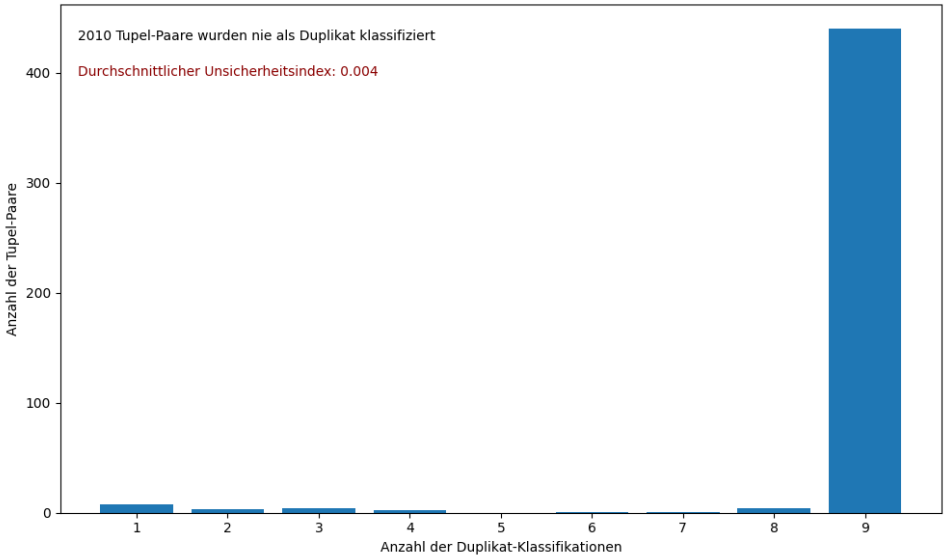

Abbildung B.38: Ditto 30 epochs auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

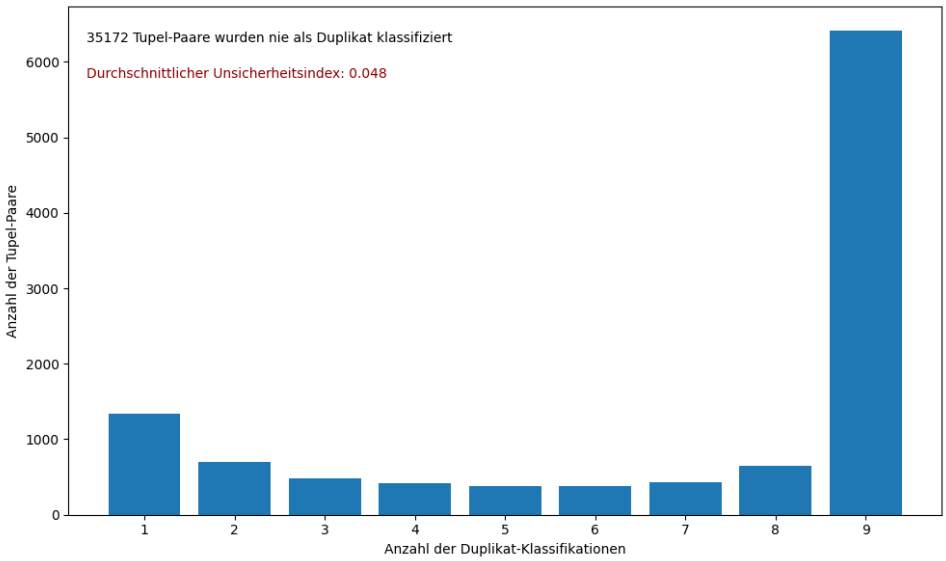

Abbildung B.39: Ditto 30 epochs auf Organizations - Plot der Übereinstimmung der Durchläufe

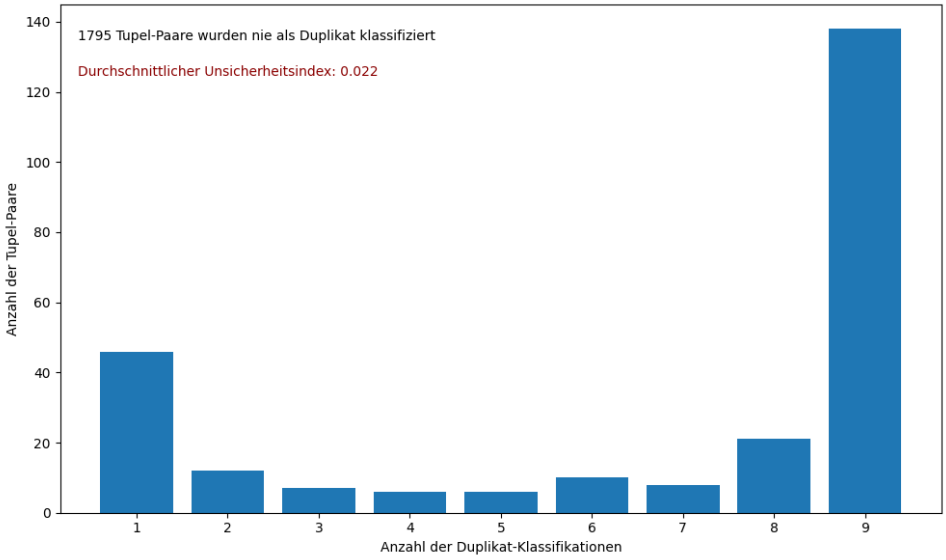

Abbildung B.40: Ditto 30 epochs auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

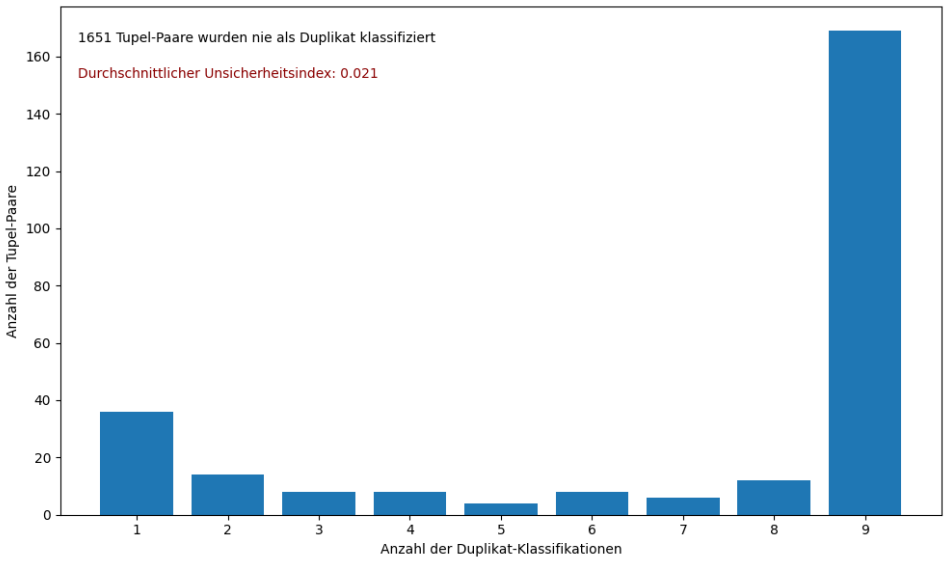

Abbildung B.41: Ditto low learning rate auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

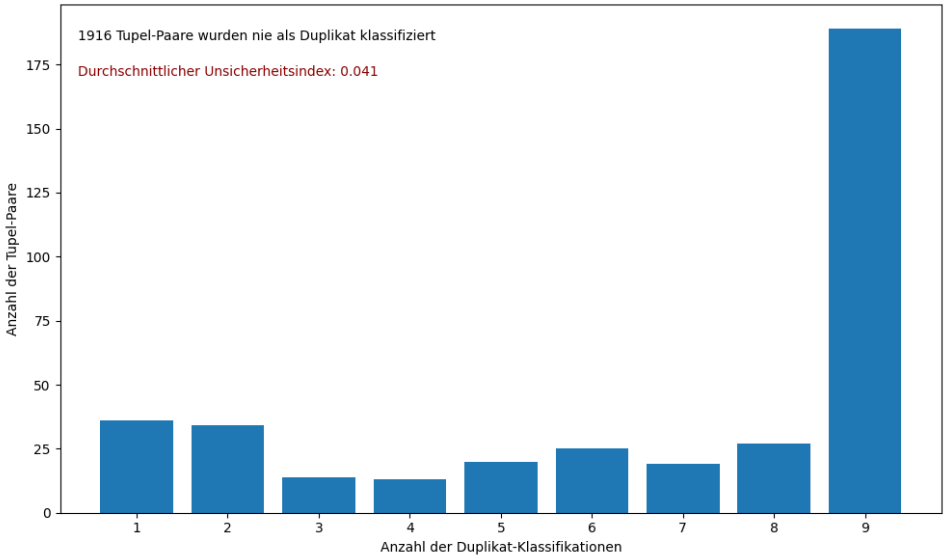

Abbildung B.42: Ditto low learning rate auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

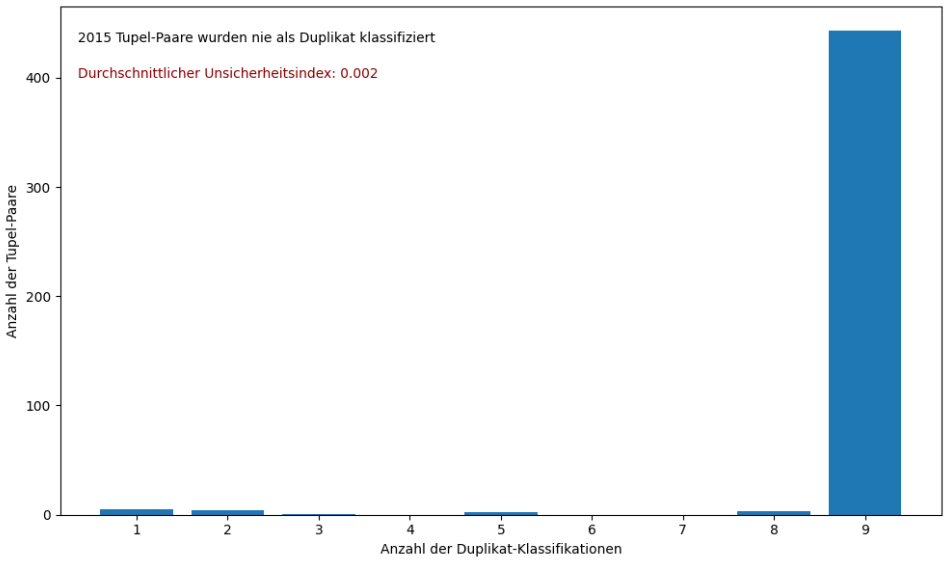

Abbildung B.43: Ditto low learning rate auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

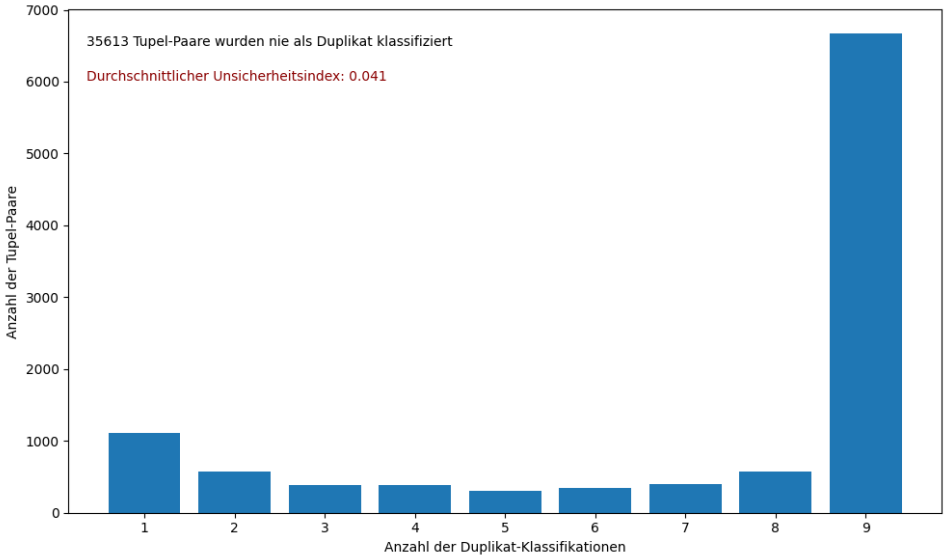

Abbildung B.44: Ditto low learning rate auf Organizations - Plot der Übereinstimmung der Durchläufe

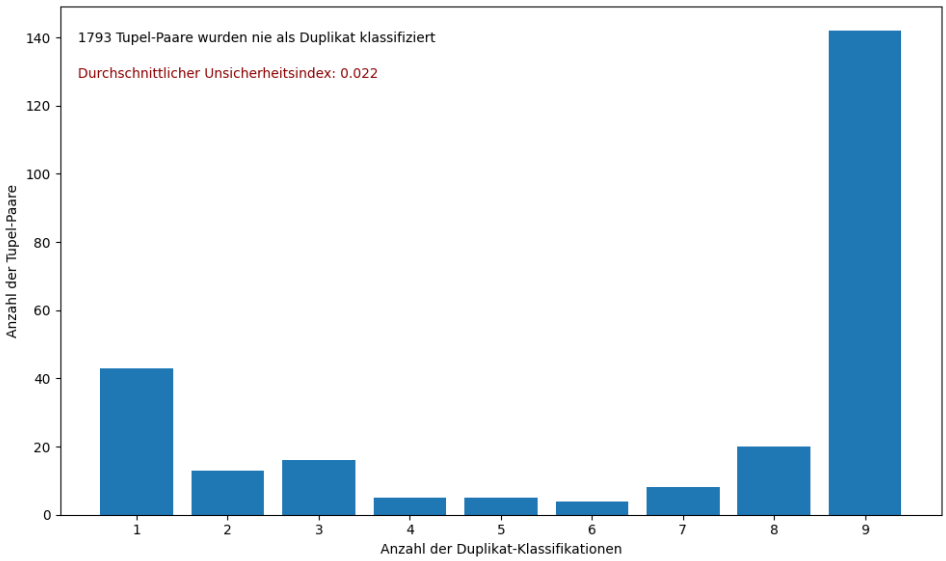

Abbildung B.45: Ditto low learning rate auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

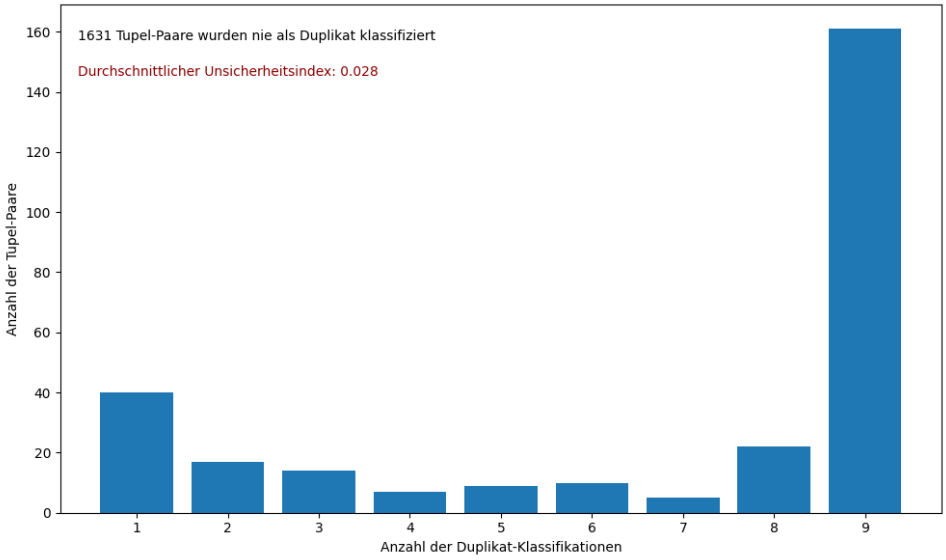

Abbildung B.46: Ditto no data augmentation auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

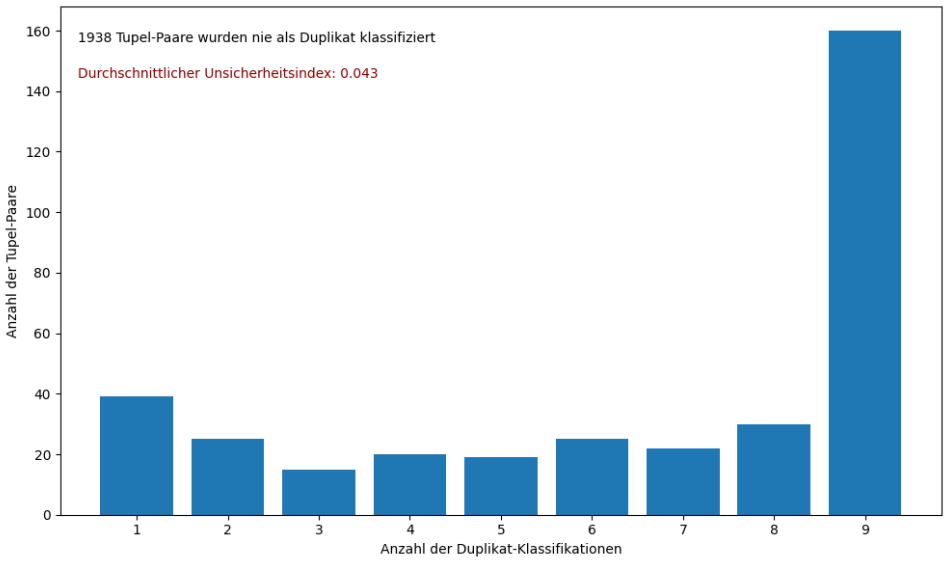

Abbildung B.47: Ditto no data augmentation auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

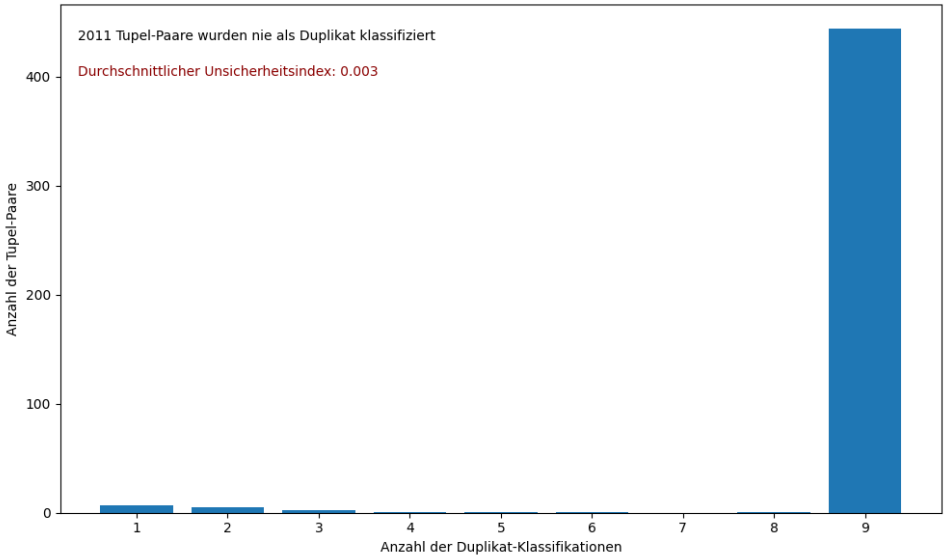

Abbildung B.48: Ditto no data augmentation auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

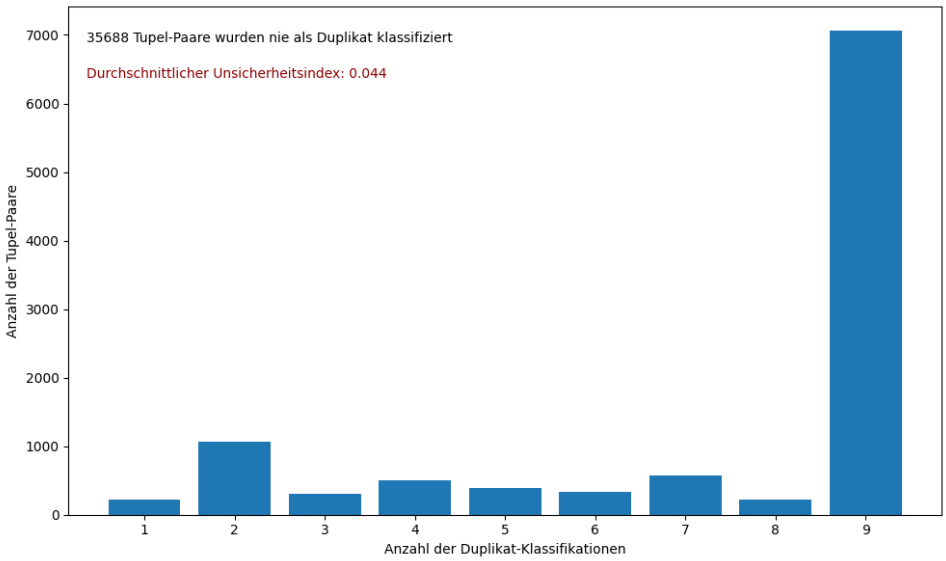

Abbildung B.49: Ditto no data augmentation auf Organizations - Plot der Übereinstimmung der Durchläufe

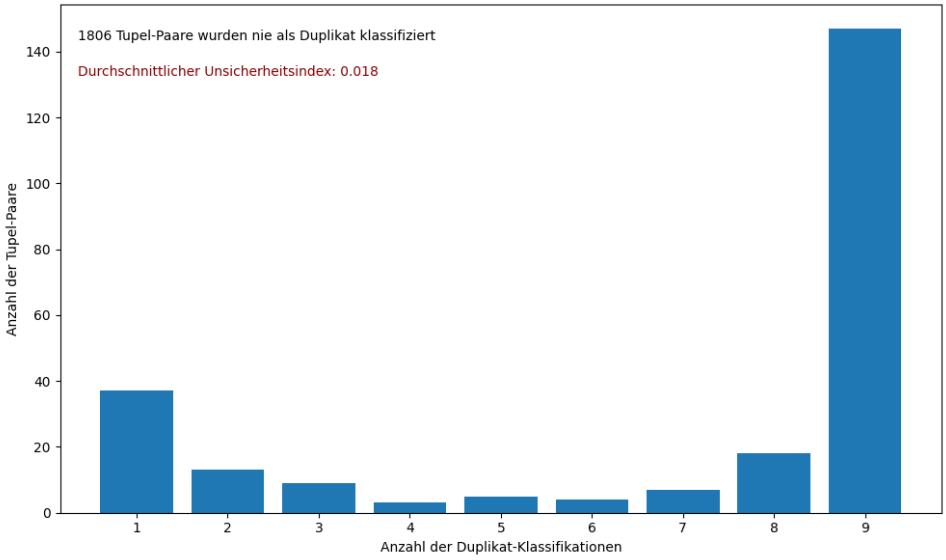

Abbildung B.50: Ditto no data augmentation auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

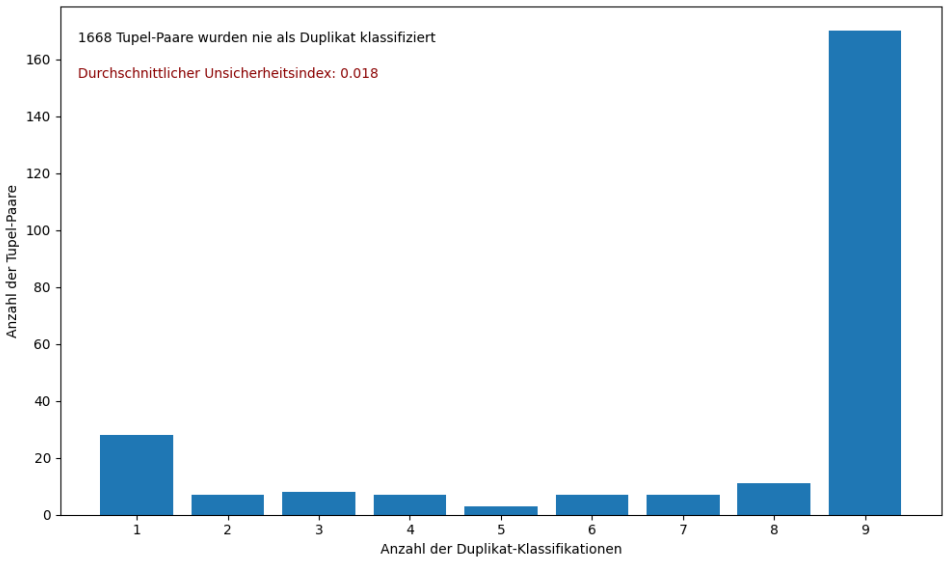

Abbildung B.51: Ditto no data augmentation and no domain knowledge auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

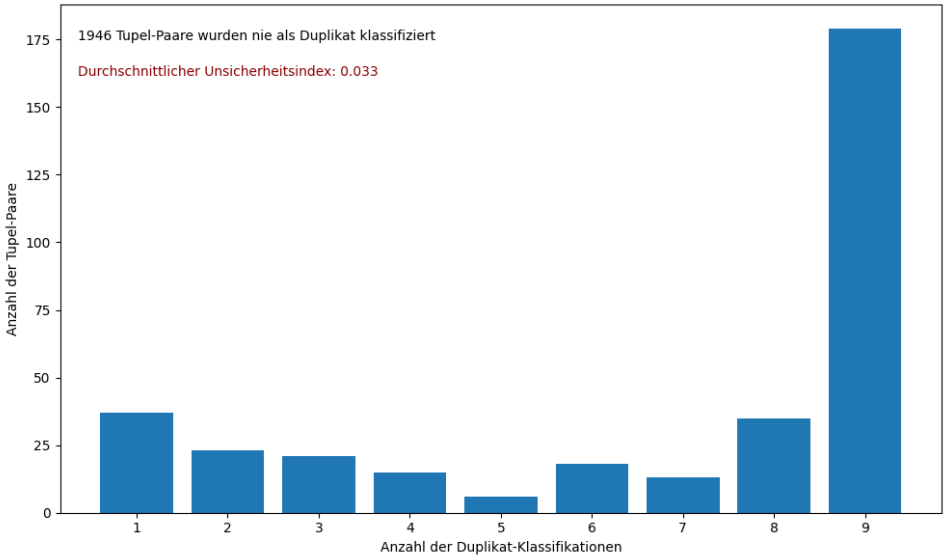

Abbildung B.52: Ditto no data augmentation and no domain knowledge auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

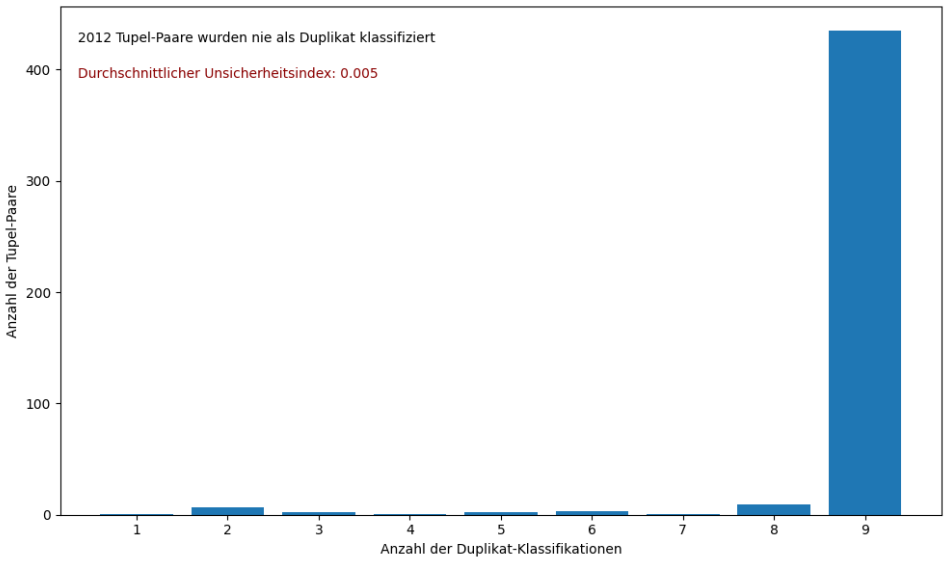

Abbildung B.53: Ditto no data augmentation and no domain knowledge auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

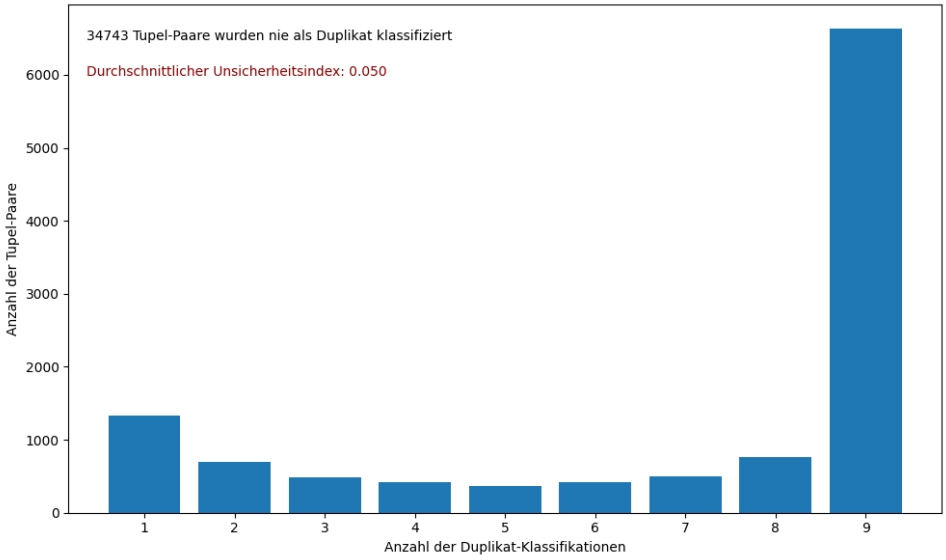

Abbildung B.54: Ditto no data augmentation and no domain knowledge auf Organizations - Plot der Übereinstimmung der Durchläufe

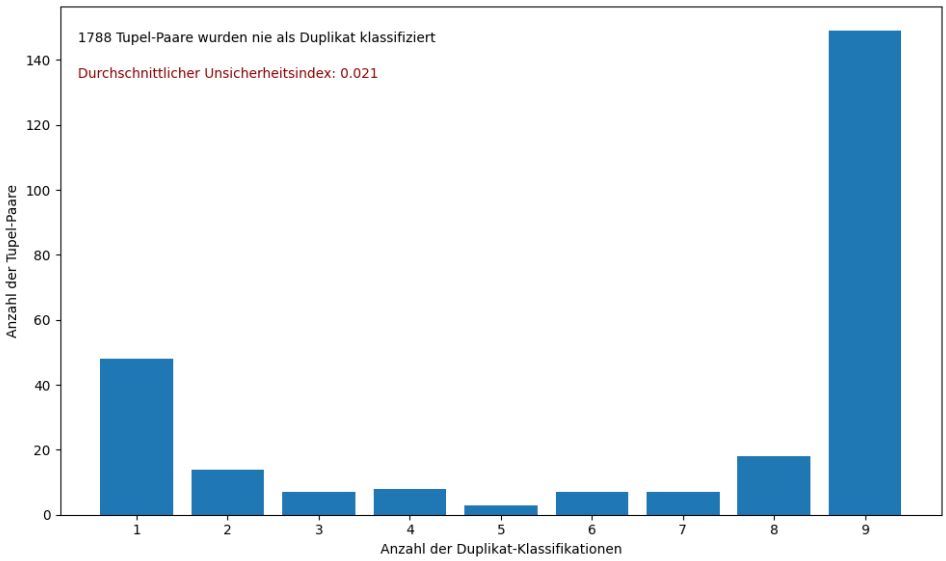

Abbildung B.55: Ditto no data augmentation and no domain knowledge auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

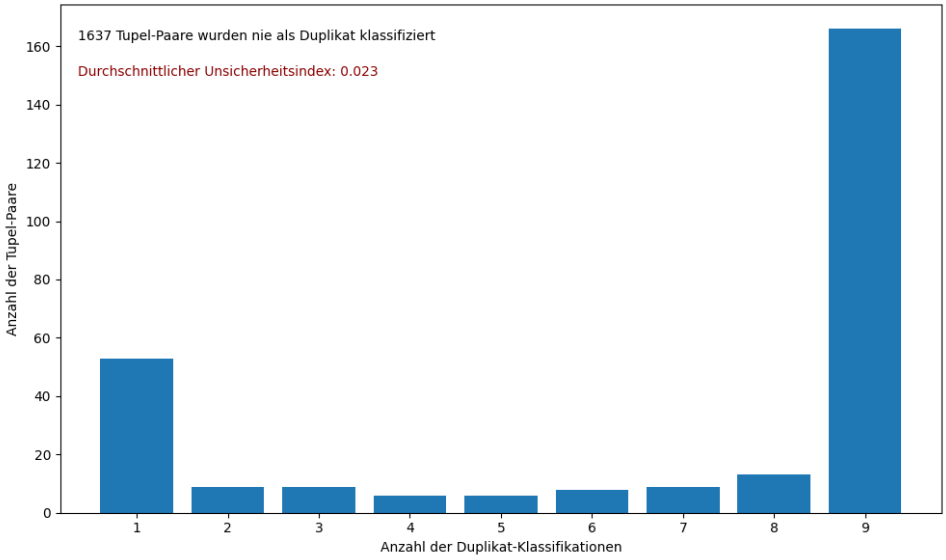

Abbildung B.56: Ditto no domain knowledge auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

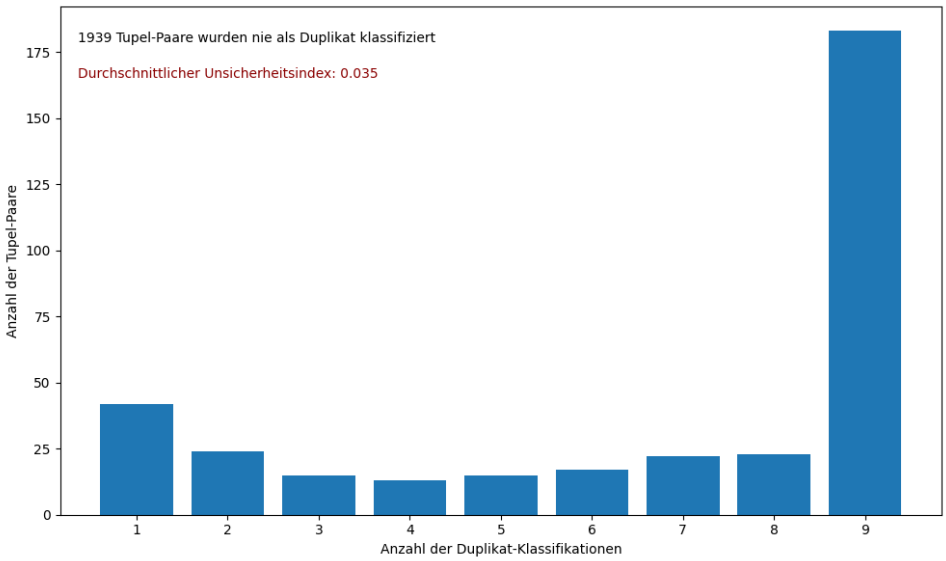

Abbildung B.57: Ditto no domain knowledge auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

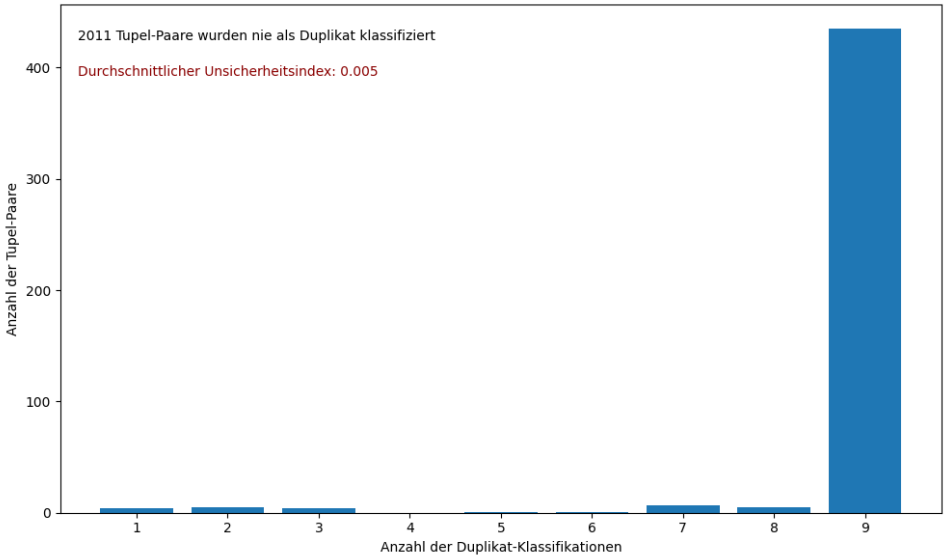

Abbildung B.58: Ditto no domain knowledge auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

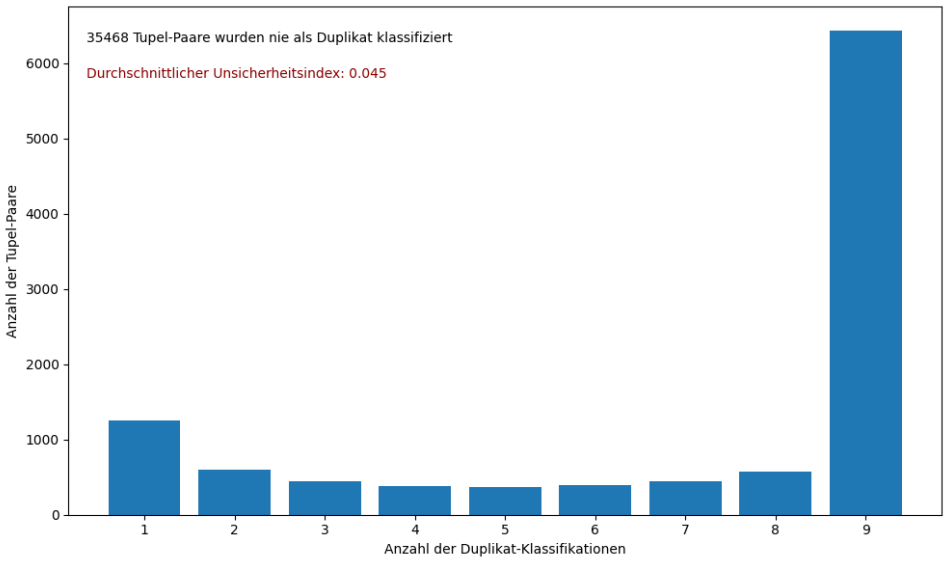

Abbildung B.59: Ditto no domain knowledge auf Organizations - Plot der Übereinstimmung der Durchläufe

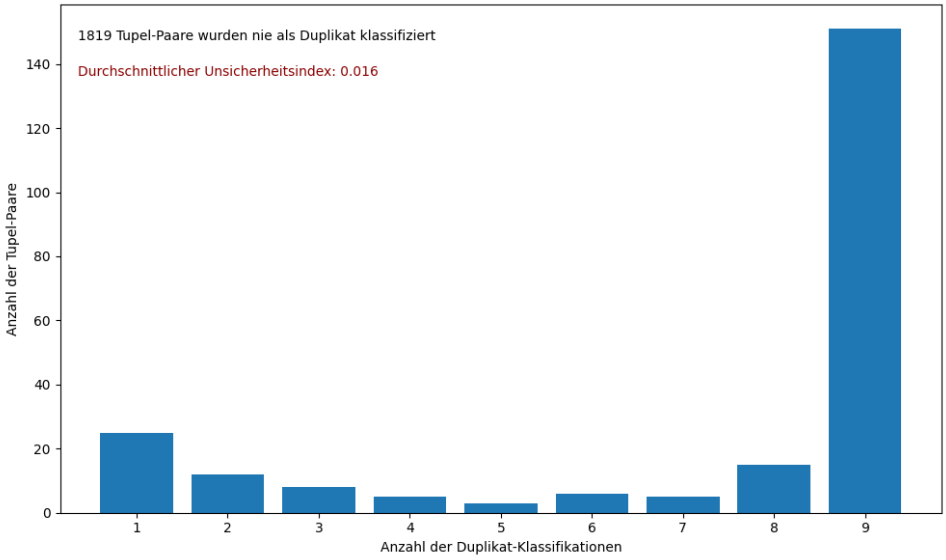

Abbildung B.60: Ditto no domain knowledge auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

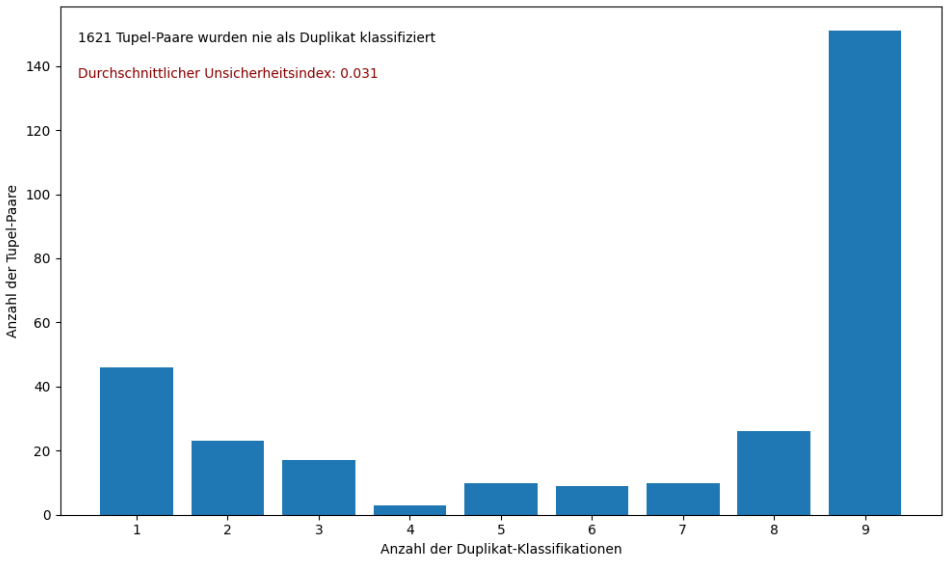

Abbildung B.61: Mix der Verfahren in Standard-Konfiguration auf Abt-Buy - Plot der Übereinstimmung der Durchläufe

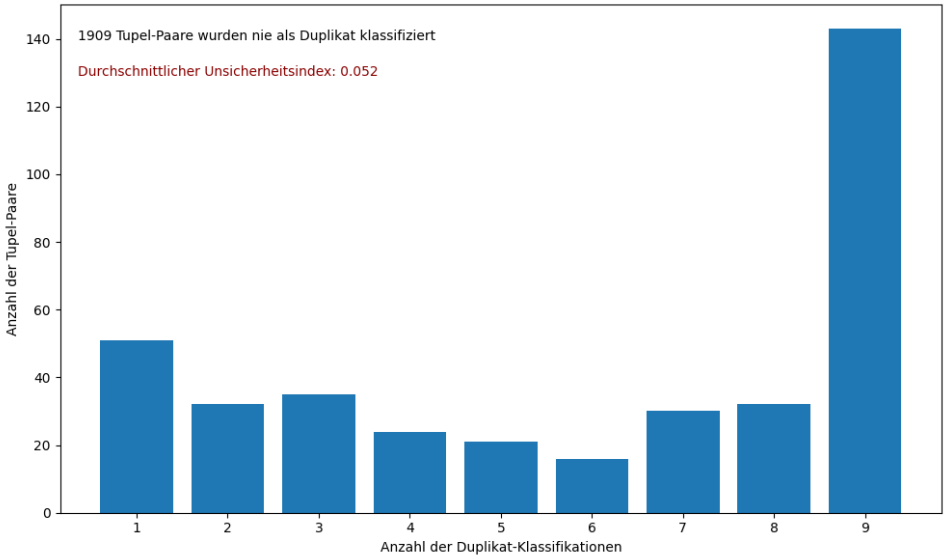

Abbildung B.62: Mix der Verfahren in Standard-Konfiguration auf Amazon-Google - Plot der Übereinstimmung der Durchläufe

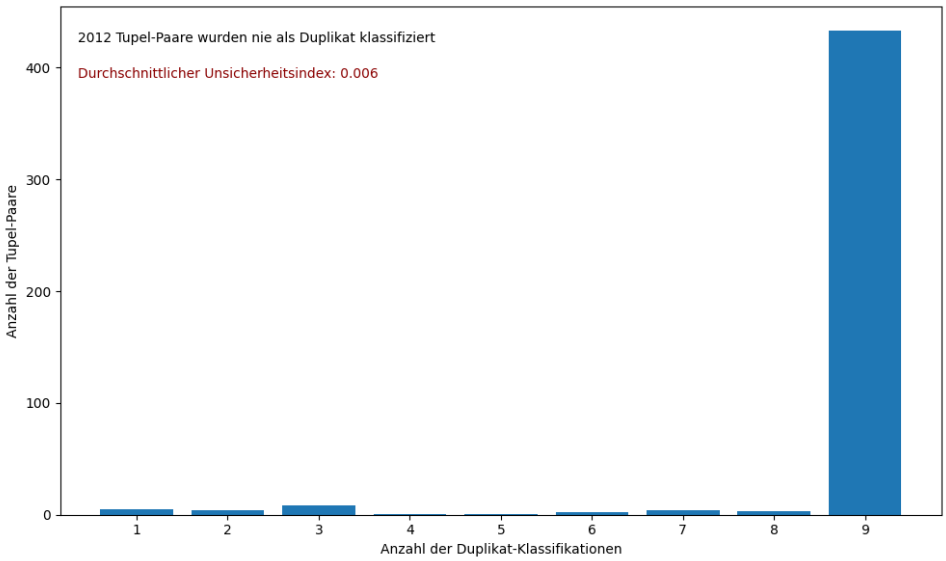

Abbildung B.63: Mix der Verfahren in Standard-Konfiguration auf DBLP-ACM - Plot der Übereinstimmung der Durchläufe

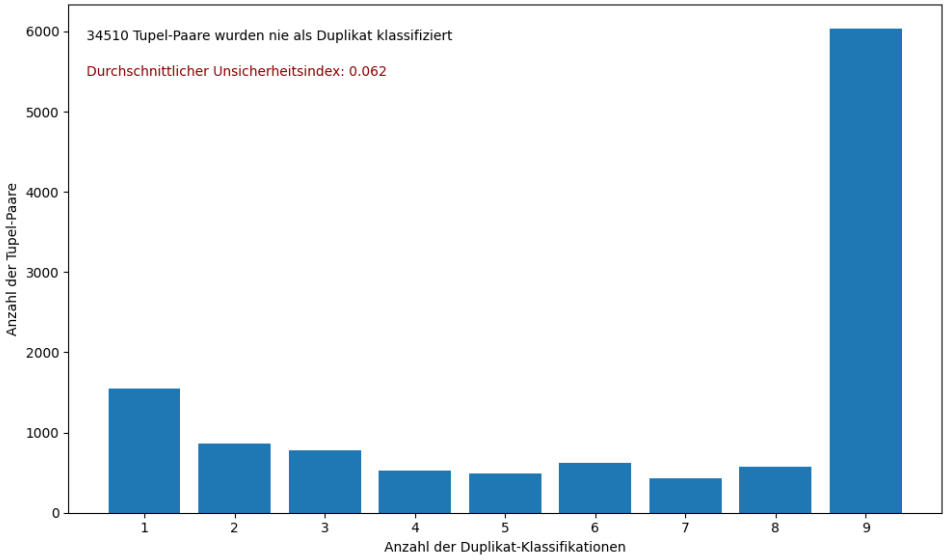

Abbildung B.64: Mix der Verfahren in Standard-Konfiguration auf Organizations - Plot der Übereinstimmung der Durchläufe

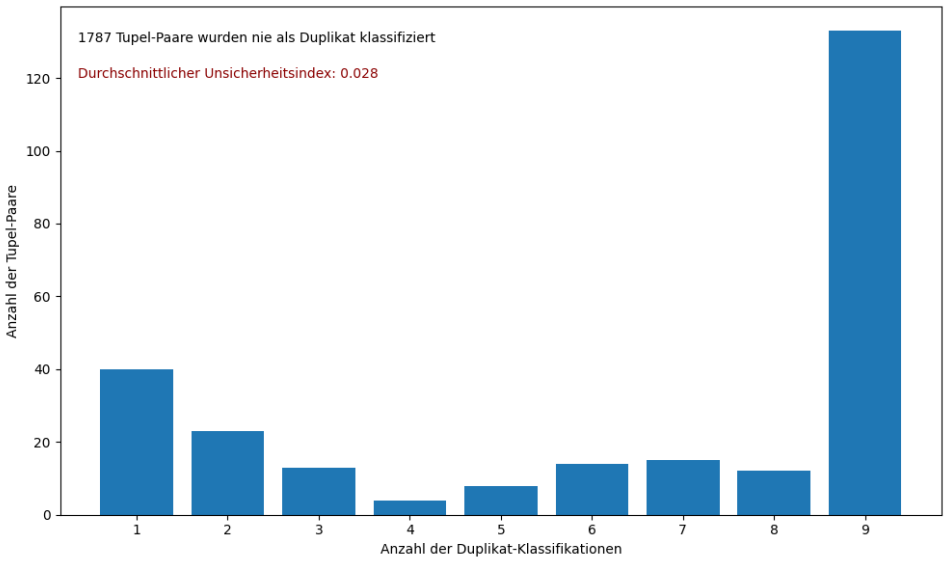

Abbildung B.65: Mix der Verfahren in Standard-Konfiguration auf Walmart-Amazon - Plot der Übereinstimmung der Durchläufe

Versicherung an Eides statt

(nach § 59 Abs. 3 HmbHG)

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Bachelorstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe. Sofern im Zuge der Erstellung der vorliegenden Abschlussarbeit generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten Wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate."

Hamburg, 14.09.2025
Ort, Datum

S. Söker
Unterschrift