

Amazon Book Sales Data

Team - 50 : Snigdha Raghavan Pradhipa, Vaishnavi Narasimhaiah Sathish,
Samrudhi Ramesh Rao, Kongarasan Sathiya Moorthy

July 6, 2025

1 Schema Matching

1.1 Mediated Schema

We use three publicly available datasets to analyze and integrate information about books. The first dataset is the Amazon Kindle Books Dataset (2023), with data containing title, author, price, ratings, and publication year. The second dataset is the Goodreads Choice Awards 2023, which includes award-winning books along with attributes like title, author, average ratings, number of reviews, and additional descriptive features. The third dataset covers the Top 100 Best-Selling Books on Amazon from 2009 to 2021, providing sales rankings, reviews, and pricing information.

To facilitate schema matching and data integration across these heterogeneous sources, we define a Mediated Schema consisting of the core attributes: `title`, `author`, `rating`, `review_count`, `price`, and `publication_year`. These attributes represent the key information required for consistent comparison and analysis across the datasets. The image below shows the snapshot of the mediated schema.



	title	author	rating	review_count	price	publication_year
0	The Lost Symbol	Dan Brown	4.4	16118.0	12.49	2009
1	The Shack: Where Tragedy Confronts Eternity	William P. Young	4.7	23392.0	13.40	2009
2	Liberty and Tyranny: A Conservative Manifesto	Mark R. Levin	4.8	5036.0	9.93	2009
3	Breaking Dawn (The Twilight Saga, Book 4)	Stephenie Meyer	4.7	16912.0	14.30	2009
4	Going Rogue: An American Life	Sarah Palin	4.6	1572.0	9.99	2009
...
134687	NaN	NaN	4.7	0	0.00	2023-09-15
134688	NaN	NaN	4.4	0	9.99	2021-09-30
134689	NaN	NaN	4.8	0	1.99	2023-09-15
134690	NaN	NaN	4.7	0	0.00	2023-08-25
134691	NaN	NaN	4.7	0	0.00	2023-10-01

Figure 1: Snapshot of the Mediated Schema

1.2 String based Matching

To align columns across different datasets with our mediated schema, we first performed string-based schema matching using the Jaro-Winkler similarity metric.

Jaro-Winkler similarity is a string comparison technique that measures the similarity between two strings based on the number and order of matching characters, as well as the number of transpositions. The base algorithm, known as the Jaro similarity, computes a score between 0 and 1, where 1 indicates an exact match. The Winkler extension gives higher scores to strings that match from the beginning, making it especially effective for short strings like attribute names.

In our use case, we used Jaro-Winkler similarity to compare each column name in the mediated schema against the column names in each of the source datasets (Amazon, Goodreads, and Kindle). We implemented this using the `jellyfish` library in Python, which computes the similarity between two strings in a case-insensitive manner.

The output of this process is a similarity matrix for each dataset, where each row corresponds to a mediated schema attribute (e.g., `title`, `author`), and each column corresponds to an attribute from the source dataset. Each matrix cell contains a value between 0 and 1 indicating how similar the two column names are. These similarity matrices help us identify which columns from the original datasets most closely match the mediated schema, and thus guide the schema mapping process.

The output of the Jaro-Winkler similarity metric is shown in the image below.

Jaro-Winkler similarity with Amazon dataset:

	Unnamed: 0	price	ratings	title	no. of reviews	author	cover type	year	genre
title	0.433333	0.600000	0.000000	1.000000	0.425641	0.561905	0.455556	0.366667	0.000000
author	0.344444	0.000000	0.455556	0.455556	0.495726	0.539683	1.000000	0.488889	0.611111
rating	0.344444	0.577778	0.760000	0.577778	0.414530	0.971429	0.555556	0.511111	0.472222
review_count	0.394444	0.627778	0.427778	0.355556	0.382479	0.484127	0.333333	0.588889	0.505556
price	0.433333	1.000000	0.466667	0.600000	0.517949	0.561905	0.000000	0.522222	0.000000
publication_year	0.441667	0.595833	0.420833	0.484722	0.352564	0.523810	0.555556	0.537500	0.437500

Jaro-Winkler similarity with Goodreads dataset:

	source URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
title	0.433333	0.416667	0.329986	1.000000	0.455556	0.595833	0.419608	0.419608	0.422222	0.561905	0.420833	0.583333	0.575000	0.616667	0.595833
author	0.488889	0.455556	0.466667	0.455556	1.000000	0.555556	0.447712	0.447712	0.455556	0.396603	0.486111	0.538889	0.395556	0.716667	0.581944
rating	0.422222	0.550000	0.549478	0.577778	0.555556	0.451389	0.316993	0.316993	0.322222	0.531746	0.409722	0.428889	0.430556	0.520000	0.486111
review_count	0.572222	0.549206	0.533989	0.355556	0.333333	0.361111	0.436928	0.356209	0.372222	0.634921	0.513889	0.422222	0.482143	0.388889	0.361111
price	0.633333	0.583333	0.574638	0.600000	0.000000	0.000000	0.339216	0.339216	0.344444	0.447619	0.341667	0.583333	0.463889	0.522222	0.420833
publication_year	0.470833	0.387500	0.503120	0.484722	0.555556	0.482143	0.531863	0.520425	0.481944	0.675595	0.50185	0.559722	0.444444	0.578373	0.583333

Jaro-Winkler similarity with Kindle dataset:

	asin	title	author	soldBy	imgURL	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate	category_name
title	0.483333	1.000000	0.455556	0.455556	0.577778	0.433333	0.466667	0.561905	0.600000	0.592157	0.527273	0.544444	0.499145	0.505882	0.499145	0.517949
author	0.472222	0.455556	1.000000	0.000000	0.555556	0.522222	0.588889	0.000000	0.000000	0.408497	0.676768	0.416667	0.576923	0.483660	0.495726	0.658120
rating	0.750000	0.577778	0.555556	0.000000	0.000000	0.511111	0.411111	0.539683	0.577778	0.483660	0.590909	0.333333	0.329060	0.467320	0.414530	0.574786
review_count	0.444444	0.355556	0.333333	0.000000	0.333333	0.494444	0.427778	0.871429	0.627778	0.436928	0.396465	0.388889	0.380342	0.436928	0.493590	0.463675
price	0.483333	0.600000	0.000000	0.000000	0.455556	0.633333	0.000000	0.676190	1.000000	0.505882	0.527273	0.522222	0.517949	0.481046	0.610256	0.499145
publication_year	0.534722	0.484722	0.555556	0.451389	0.451389	0.617262	0.341667	0.401786	0.595833	0.473564	0.548431	0.513889	0.499038	0.464869	0.663462	0.677528

Figure 2: Snapshot of the String based Matching using Jaro-Winkler similarity metric

1.3 Semantic based Matching

To improve the matching of schema attributes, especially when column names differ in wording but have similar meanings, we used semantic similarity with the help of the `spaCy` library. Unlike string-based methods, which focus on how similar two words look, semantic similarity aims to understand how similar the meanings of two words or phrases are.

For this, we used `spaCy`'s pre-trained language model, which includes vector representations of words. These vectors allow us to compare the meaning of two attribute names. For example, while `rating` and `stars` may not look similar, they are semantically related and would receive a high similarity score.

In our code, we first processed all attribute names from both the mediated schema and all three source datasets by converting underscores to spaces to make them more natural language-like (e.g., `review_count` becomes `review count`). Each name was then turned into a `spaCy` document. We computed the similarity between each pair using `spaCy`'s built-in `similarity()` function, which outputs a score between 0 and 1.

This resulted in a similarity matrix for each dataset (Amazon, Goodreads, Kindle), where each cell shows how closely the meanings of two attributes align. This approach helped us make better matching decisions when exact string matches were not possible but the intent of the columns was still the same.

The below Image shows the resulted matrix of semantic based similarity index.

Semantic similarity with Amazon dataset:

	Unnamed: 0	price	rank	title	no. of reviews	ratings	author	cover_type	year	genre
title	0.197435	0.182185	0.274463	1.000000	0.348178	0.179478	1.000000	0.470557	0.116806	0.176670
author	0.197435	0.182185	0.274463	1.000000	0.348178	0.179478	1.000000	0.470557	0.116806	0.176670
rating	0.219100	0.368238	0.489637	0.170101	0.637476	0.816872	0.170101	0.258596	0.188736	0.059141
review_count	0.240654	0.357592	0.491957	0.442599	0.606636	0.410101	0.442599	0.436643	0.460118	0.047429
price	0.252831	1.000000	0.236469	0.182185	0.491697	0.345526	0.182185	0.319655	0.336872	0.042610
publication_year	0.041870	0.197092	0.303843	0.284768	0.316379	0.240577	0.284768	0.277165	0.741094	0.021456

Semantic similarity with goodreads dataset:

	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
title	0.268200	0.602178	0.624171	1.000000	1.000000	0.209238	0.301707	0.301707	0.611881	0.328353	0.518614	0.671481	0.320590	0.282663	0.673901
author	0.268200	0.602178	0.624171	1.000000	1.000000	0.209238	0.301707	0.301707	0.611881	0.328353	0.518614	0.671481	0.320590	0.282663	0.673901
rating	0.139320	0.507274	0.355419	0.170101	0.170101	0.744673	0.558230	0.558230	0.279250	0.042204	0.120570	0.348021	0.355914	0.251501	0.357396
review_count	0.303489	0.546372	0.503708	0.442599	0.442599	0.641303	0.613798	0.613798	0.615606	0.159701	0.343764	0.642183	0.536200	0.308465	0.620913
price	0.366592	0.376949	0.356397	0.182185	0.182185	0.474273	0.490494	0.490494	0.418552	0.247511	0.295785	0.444600	0.718403	0.768161	0.454068
publication_year	0.274571	0.304255	0.426613	0.284768	0.284768	0.319298	0.384762	0.384762	0.395821	0.214990	0.302087	0.372220	0.313379	0.290848	0.270767

Semantic similarity with Kindle dataset:

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate	category_name
title	0.192773	1.000000	1.000000	0.0	0.0	0.0	0.170101	0.179478	0.182185	0.0	0.237677	0.0	0.0	0.0	0.0	0.305816
author	0.192773	1.000000	1.000000	0.0	0.0	0.0	0.170101	0.179478	0.182185	0.0	0.237677	0.0	0.0	0.0	0.0	0.305816
rating	0.148880	0.170101	0.170101	0.0	0.0	0.0	1.000000	0.816872	0.368238	0.0	0.262992	0.0	0.0	0.0	0.0	0.252249
review_count	0.110212	0.442599	0.442599	0.0	0.0	0.0	0.391012	0.410101	0.357592	0.0	0.282719	0.0	0.0	0.0	0.0	0.339599
price	0.327039	0.182185	0.182185	0.0	0.0	0.0	0.368238	0.345526	1.000000	0.0	0.250011	0.0	0.0	0.0	0.0	0.180116
publication_year	0.113380	0.284768	0.284768	0.0	0.0	0.0	0.163626	0.240577	0.197092	0.0	0.237728	0.0	0.0	0.0	0.0	0.313044

Figure 3: Snapshot of the Semantic-Based Matching using spaCy

1.4 Instance based Matching

To complement schema-based matching techniques such as Jaro-Winkler similarity, an instance-based schema matching approach was also implemented to enhance the alignment accuracy of attributes across heterogeneous book datasets. This method relies on the actual instance values under each column rather than just the column names.

The goal is to match semantically similar columns from different datasets (Amazon, Goodreads, and Kindle) to a common mediated schema based on the similarity of their instance-level values.

For instance-based schema matching, the column values from each dataset were first preprocessed and flattened into single text strings. These were then transformed into TF-IDF vectors using Scikit-learn's 'TfidfVectorizer', capturing the importance of terms within and across columns. Pairwise cosine similarity was computed between each mediated schema column and source dataset column to assess semantic similarity. The results were organized into similarity matrices, allowing clear visualization of attribute-level matches for each dataset.

The instance-based matching approach using TF-IDF and cosine similarity provides a data-driven way to identify semantically similar attributes, even when column names differ significantly. By leveraging the actual values within each column, it captures contextual meaning and implicit relationships that name-based methods might miss. This enhances the robustness of schema matching, especially in heterogeneous datasets with inconsistent or ambiguous labeling. It is particularly useful when working with real-world data sources like Amazon, Goodreads, and Kindle, where schema variations are common.

The output is shown in the image below.

TF-IDF Cosine similarity with Amazon dataset:

	Unnamed: 0	price	ratings	title	no_of_reviews	ratings	author	cover_type	year	genre
title	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
author	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
rating	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
review_count	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
price	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
publication_year	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

TF-IDF Cosine similarity with Goodreads dataset:

	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
title	0.0	0.0	0.0	1.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
author	0.0	0.0	0.0	0.0	1.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.492501
rating	0.0	0.0	0.0	0.0	0.0	0.528612	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
review_count	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000
price	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.431112	0.670547	0.000000
publication_year	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000

TF-IDF Cosine similarity with Kindle dataset:

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsChoice	publishedDate	category_name
title	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
author	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
rating	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
review_count	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
price	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
publication_year	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 4: Snapshot of the Instance Based matching with TF-IDF Cosine similarity

2 Combiners

In the context of schema matching or entity resolution, multiple similarity scores can be computed between attributes based on different criteria (e.g., name, value, type). To derive a single unified similarity score, combiners are used to aggregate these individual similarity measures. The available combiners for aggregating multiple similarity scores include **Max**, **Minimum**, **Average**, **Weighted Average**, each offering a different strategy for integrating diverse similarity metrics into a unified score. These combiners were applied to the similarity matrices, followed by binarization based on a defined threshold to predict matches. Their effectiveness was evaluated using precision, recall, and F1-score against ground truth datasets.

In this project, we explored and evaluated two such combiners:

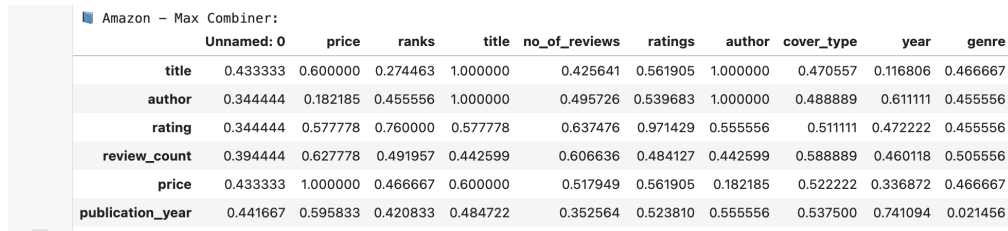
2.1 Maximum-combiner

The Maximum Combiner selects the highest similarity score for each attribute pair across all matching techniques. It captures the strongest signal, even if only one matcher shows a high confidence in the mapping. This approach prioritizes recall, allowing more potential matches to be considered in schema integration. While it increases coverage, it may also introduce false positives due to over-reliance on a single method. The resulting matrix reflects the most optimistic view of attribute similarity across datasets.

2.2 Weighted Average-Combiner

The Weighted Average Combiner calculates a similarity score by assigning different weights to each matcher based on their reliability. For example, semantic similarity may be given more weight than string-based similarity if meaning matters more than appearance. This allows fine-tuning of the matching process, emphasizing certain perspectives over others. It balances flexibility with control, making it ideal for datasets where some matchers outperform others. The output reflects a weighted confidence score, tailored to the strengths of each matching method.

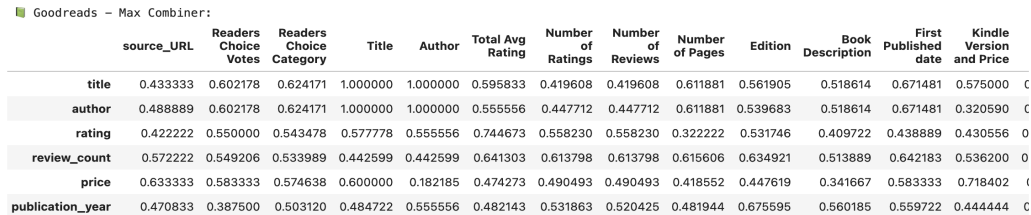
The output of Max Combiner is shown in the images below:



Amazon - Max Combiner:

	Unnamed: 0	price	ranks	title	no_of_reviews	ratings	author	cover_type	year	genre
title	0.433333	0.600000	0.274463	1.000000	0.425641	0.561905	1.000000	0.470557	0.116806	0.466667
author	0.344444	0.182185	0.455556	1.000000	0.495726	0.539683	1.000000	0.488889	0.611111	0.455556
rating	0.344444	0.577778	0.760000	0.577778	0.637476	0.971429	0.555556	0.511111	0.472222	0.455556
review_count	0.394444	0.627778	0.491957	0.442599	0.606636	0.484127	0.442599	0.588889	0.460118	0.505556
price	0.433333	1.000000	0.466667	0.600000	0.517949	0.561905	0.182185	0.522222	0.336872	0.466667
publication_year	0.441667	0.595833	0.420833	0.484722	0.352564	0.523810	0.555556	0.537500	0.741094	0.021456

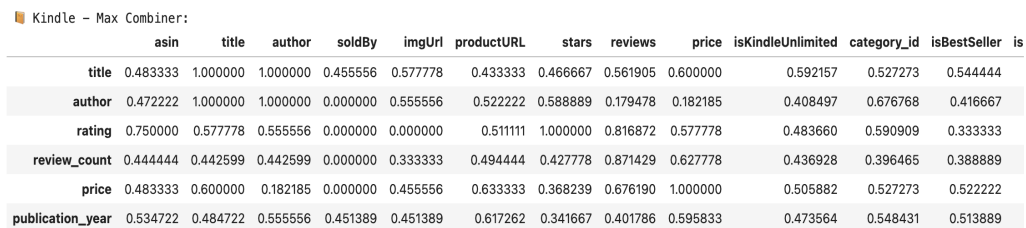
Figure 5: Snapshots of the Maximum Combiner similarity matrix for Amazon Dataset



Goodreads - Max Combiner:

	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price
title	0.433333	0.602178	0.624171	1.000000	1.000000	0.595833	0.419608	0.419608	0.611881	0.561905	0.518614	0.671481	0.575000
author	0.488889	0.602178	0.624171	1.000000	1.000000	0.555556	0.447712	0.447712	0.611881	0.539683	0.518614	0.671481	0.320590
rating	0.422222	0.550000	0.543478	0.577778	0.555556	0.744673	0.558230	0.322222	0.531746	0.409722	0.438889	0.430556	0
review_count	0.572222	0.549206	0.533989	0.442599	0.442599	0.641303	0.613798	0.613798	0.615606	0.634921	0.513889	0.642183	0.536200
price	0.633333	0.583333	0.574638	0.600000	0.182185	0.474273	0.490493	0.490493	0.418552	0.447619	0.341667	0.583333	0.718402
publication_year	0.470833	0.387500	0.503120	0.484722	0.555556	0.482143	0.531863	0.520425	0.481944	0.675595	0.560185	0.559722	0.444444

Figure 6: Snapshots of the Maximum Combiner similarity matrix for GoodReads Dataset

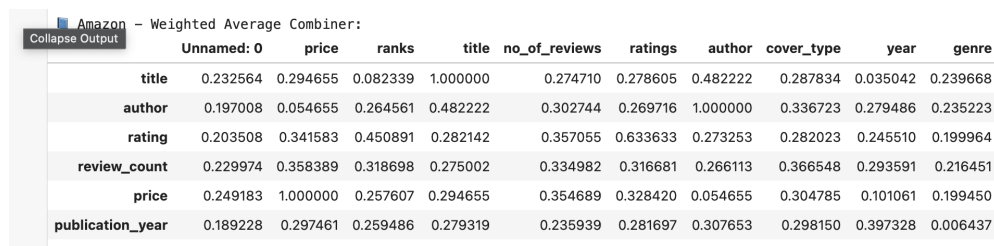


Kindle - Max Combiner:

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	is
title	0.483333	1.000000	1.000000	0.455556	0.577778	0.433333	0.466667	0.561905	0.600000	0.592157	0.527273	0.544444	
author	0.472222	1.000000	1.000000	0.000000	0.555556	0.522222	0.588889	0.179478	0.182185	0.408497	0.676768	0.416667	
rating	0.750000	0.577778	0.555556	0.000000	0.000000	0.511111	1.000000	0.816872	0.577778	0.483660	0.590909	0.333333	
review_count	0.444444	0.442599	0.442599	0.000000	0.333333	0.494444	0.427778	0.871429	0.627778	0.436928	0.396465	0.388889	
price	0.483333	0.600000	0.182185	0.000000	0.455556	0.633333	0.368239	0.676190	1.000000	0.505882	0.527273	0.522222	
publication_year	0.534722	0.484722	0.555556	0.451389	0.451389	0.617262	0.341667	0.401786	0.595833	0.473564	0.548431	0.513889	

Figure 7: Snapshots of the Maximum Combiner similarity matrix for Kindle Dataset

The output of Weighted Average Combiner is shown in the images below:



Amazon - Weighted Average Combiner:

	Unnamed: 0	price	ranks	title	no_of_reviews	ratings	author	cover_type	year	genre
title	0.232564	0.294655	0.082339	1.000000	0.274710	0.278605	0.482222	0.287834	0.035042	0.239668
author	0.197008	0.054655	0.264561	0.482222	0.302744	0.269716	1.000000	0.336723	0.279486	0.235223
rating	0.203508	0.341583	0.450891	0.282142	0.357055	0.633633	0.273253	0.282023	0.245510	0.199964
review_count	0.229974	0.358389	0.318698	0.275002	0.334982	0.316681	0.266113	0.366548	0.293591	0.216451
price	0.249183	1.000000	0.257607	0.294655	0.354689	0.328420	0.054655	0.304785	0.101061	0.199450
publication_year	0.189228	0.297461	0.259486	0.279319	0.235939	0.281697	0.307653	0.298150	0.397328	0.006437

Figure 8: Snapshots of the Weighted Average Combiner similarity matrix for Amazon Dataset



Goodreads - Weighted Average Combiner:

	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price
title	0.253793	0.347320	0.318846	1.000000	0.482222	0.301105	0.258355	0.258355	0.352453	0.323268	0.323918	0.434778	0.326177
author	0.276016	0.362876	0.365995	0.482222	1.000000	0.284994	0.269597	0.269597	0.365786	0.314379	0.350029	0.417000	0.218399
rating	0.210685	0.372182	0.324017	0.282142	0.273253	0.562541	0.294266	0.294266	0.212664	0.225360	0.200060	0.279962	0.278996
review_count	0.319936	0.383594	0.364708	0.275002	0.266113	0.336835	0.358911	0.326623	0.333571	0.301879	0.308685	0.361544	0.353717
price	0.363311	0.346388	0.336774	0.294655	0.054655	0.142282	0.282834	0.282834	0.263343	0.253301	0.225402	0.366713	0.530410
publication_year	0.270705	0.246277	0.329232	0.279319	0.307653	0.288647	0.328174	0.323599	0.311524	0.334735	0.314700	0.335555	0.271791

Figure 9: Snapshots of the Weighted Average Combiner similarity matrix for GoodReads Dataset

Kindle - Weighted Average Combiner:												
	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller
title	0.251165	1.000000	0.482222	0.182222	0.231111	0.173333	0.237697	0.278605	0.294655	0.236863	0.282212	0.217778
author	0.246721	0.482222	1.000000	0.000000	0.222222	0.208889	0.286586	0.053843	0.054655	0.163399	0.342010	0.166667
rating	0.344664	0.282142	0.273253	0.000000	0.000000	0.204444	0.464444	0.460935	0.341583	0.193464	0.315261	0.133333
review_count	0.210841	0.275002	0.266113	0.000000	0.133333	0.197778	0.288415	0.471602	0.358389	0.174771	0.243402	0.155556
price	0.291445	0.294655	0.054655	0.000000	0.182222	0.253333	0.110472	0.374134	1.000000	0.202353	0.285912	0.208889
publication_year	0.247903	0.279319	0.307653	0.180556	0.180556	0.246905	0.185754	0.232887	0.297461	0.189426	0.290691	0.205556

Figure 10: Snapshots of the Weighted Average Combiner similarity matrix for Kindle Dataset

3 Ground Truth Table

A ground truth table is a manually created binary matrix that shows the correct mappings between a mediated schema and one or more source datasets. It acts as the "answer key" for evaluating how well your matchers and combiners performed.

Why Create a Ground Truth Table?

Because automated matchers (Jaro, SpaCy, TF-IDF, etc.) can make mistakes, we need a reliable way to measure accuracy. The ground truth lets you calculate:

Precision – how many predicted matches are correct

Recall – how many correct matches were found

F1-score – balance between precision and recall

The images below display the Ground Truth tables corresponding to the Max combiners for the datasets.

	Unnamed: 0	price	ranks	title	no_of_reviews	ratings	author	cover_type
1	title	0	1	0	1	0	1	0
2	author	0	0	0	0	0	1	1
3	rating	0	1	1	1	0	1	1
4	review_count	0	1	0	0	0	1	0
5	price	0	1	0	1	1	1	0
6	publication_year	0	1	0	0	0	1	1

Figure 11: Ground Truth Table obtained using Maximum combination methods for Amazon Dataset

		source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings
1	title	0	0	0	1	0	1	0
2	author	0	0	0	0	1	1	0
3	rating	0	1	1	1	1	1	1
4	review_count	1	1	1	0	0	0	0
5	price	1	1	1	1	0	0	0
6	publication_year	0	0	1	0	1	0	1

Figure 12: Ground Truth Table obtained using Maximum combination methods for GoodReads Dataset

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews
1	title	0	1	0	0	1	0	0
2	author	0	0	1	0	1	1	1
3	rating	1	1	1	0	0	1	0
4	review_count	0	0	0	0	0	0	0
5	price	0	1	0	0	0	1	0
6	publication_year	1	0	1	0	0	1	0

Figure 13: Ground Truth Table obtained using Maximum combination methods for Kindle Dataset

The images below display the Ground Truth tables corresponding to the Weighted Average combiners for the datasets.

	Unnamed: 0	price	rank	title	no_of_reviews	ratings	author	cover_type
1	title	0	0	0	1	0	0	0
2	author	0	0	0	0	0	0	1
3	rating	0	0	0	0	0	1	0
4	review_count	0	0	0	0	0	0	0
5	price	0	1	0	0	0	0	0
6	publication_year	0	0	0	0	0	0	0

Figure 14: Ground Truth Table obtained using Weighted Average combination methods

		source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings
1	title	0	0	0	1	0	0	0
2	author	0	0	0	0	1	0	0
3	rating	0	0	0	0	0	1	0
4	review_count	0	0	0	0	0	0	0
5	price	0	0	0	0	0	0	0
6	publication_year	0	0	0	0	0	0	0

Figure 15: Ground Truth Table obtained using Weighted Average combination methods

	asin	title	author	soldBy	imgUrl	productURL	stars	reviews
1	title	0	1	0	0	0	0	0
2	author	0	0	1	0	0	0	0
3	rating	0	0	0	0	0	0	0
4	review_count	0	0	0	0	0	0	0
5	price	0	0	0	0	0	0	0
6	publication_year	0	0	0	0	0	0	0

Figure 16: Ground Truth Table obtained using Weighted Average combination methods

4 Threshold

What is Thresholding?

Thresholding is the process of deciding how similar two attributes must be to be considered a match. Your matchers (like Jaro-Winkler, SpaCy, TF-IDF) output similarity scores between 0 and 1. These scores show how close one attribute is to another — but they are still fuzzy. You apply a threshold (like 0.5) and binarize the outputs between 0 and 1.

If similarity $\geq 0.5 \rightarrow$ treat it as a match (1)

If similarity $< 0.5 \rightarrow$ not a match (0)

Amazon - Max Combiner (Binarized):										
	Unnamed: 0	price	rank	title	no_of_reviews	ratings	author	cover_type	year	genre
title	0	1	0	1	0	1	1	0	0	0
author	0	0	0	1	0	1	1	0	1	0
rating	0	1	1	1	1	1	1	1	0	0
review_count	0	1	0	0	1	0	0	1	0	1
price	0	1	0	1	1	1	0	1	0	0
publication_year	0	1	0	0	0	1	1	1	1	0

Figure 17: Binary representations obtained after applying thresholding to the Maximum similarity matrices.

Goodreads - Max Combiner (Binarized):															
	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
title	0	1	1	1	1	1	0	0	1	1	1	1	1	1	1
author	0	1	1	1	1	1	0	0	1	1	1	1	0	0	1
rating	0	1	1	1	1	1	1	1	0	1	0	0	0	0	0
review_count	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1
price	1	1	1	1	0	0	0	0	0	0	0	1	1	1	0
publication_year	0	0	1	0	1	0	1	1	0	1	1	1	0	1	1

Figure 18: Binary representations obtained after applying thresholding to the Maximum similarity matrices.

Kindle - Max Combiner (Binarized):														
	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsC
title	0	1	1	0	1	0	0	1	1	1	1	1	0	
author	0	1	1	0	1	1	1	0	0	0	1	0	1	
rating	1	1	1	0	0	1	1	1	1	0	1	0	0	
review_count	0	0	0	0	0	0	0	1	1	0	0	0	0	
price	0	1	0	0	0	1	0	1	1	1	1	1	1	
publication_year	1	0	1	0	0	1	0	0	1	0	1	1	0	

Figure 19: Binary representations obtained after applying thresholding to the Maximum similarity matrices.

Amazon - Weighted Avg Combiner (Binarized):										
	title	author	rating	review_count	price	publication_year	genre	year	cover_type	author
title	0	0	0	1	0	0	0	0	0	0
author	0	0	0	0	0	0	0	1	0	0
rating	0	0	0	0	0	0	1	0	0	0
review_count	0	0	0	0	0	0	0	0	0	0
price	0	1	0	0	0	0	0	0	0	0
publication_year	0	0	0	0	0	0	0	0	0	0

Figure 20: Binary representations obtained after applying thresholding to the Weighted Average similarity matrices.

Goodreads - Weighted Avg Combiner (Binarized):															
	source_URL	Readers Choice Votes	Readers Choice Category	Title	Author	Total Avg Rating	Number of Ratings	Number of Reviews	Number of Pages	Edition	Book Description	First Published date	Kindle Version and Price	Kindle Price	About the Author
title	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
author	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
rating	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
review_count	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
price	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
publication_year	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 21: Binary representations obtained after applying thresholding to the Weighted Average similarity matrices.

Kindle - Weighted Avg Combiner (Binarized):														
	asin	title	author	soldBy	imgUrl	productURL	stars	reviews	price	isKindleUnlimited	category_id	isBestSeller	isEditorsPick	isGoodReadsC
title	0	1	0	0	0	0	0	0	0	0	0	0	0	0
author	0	0	1	0	0	0	0	0	0	0	0	0	0	0
rating	0	0	0	0	0	0	0	0	0	0	0	0	0	0
review_count	0	0	0	0	0	0	0	0	0	0	0	0	0	0
price	0	0	0	0	0	0	0	0	1	0	0	0	0	0
publication_year	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 22: Binary representations obtained after applying thresholding to the Weighted Average similarity matrices.

5 Quality Measure

To evaluate how accurately the schema matching techniques identified correct attribute correspondences, we compared their output against a manually created ground truth table.

Method: Similarity matrices were generated using different matchers: Jaro-Winkler, SpaCy (semantic), and TF-IDF (textual). These were combined using techniques like Max, Min, Average, and Weighted Average. A threshold of 0.5 was

applied to convert similarity scores to binary match results. The resulting binary matrices were compared against the ground truth using standard evaluation metrics:

Precision: Correct matches out of all predicted matches

Recall: Correct matches found out of all actual matches

F1-Score: Balance between precision and recall.

```
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.5 Threshold
Amazon Max Combiner:
Precision: 0.8276
Recall: 0.9600
F1 Score: 0.8889
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.5 Threshold
Amazon Weighted_Average Combiner:
Precision: 1.0000
Recall: 1.0000
F1 Score: 1.0000
```

Figure 23: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.5 Threshold

```
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.5 Threshold
Goodreads Max Combiner:
Precision: 0.6724
Recall: 1.0000
F1 Score: 0.8041
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.5 Threshold
Goodreads Weighted_Average Combiner:
Precision: 0.8333
Recall: 1.0000
F1 Score: 0.9091
```

Figure 24: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.5 Threshold

```
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.5 Threshold
Kindle Max Combiner:
Precision: 0.8696
Recall: 0.9756
F1 Score: 0.9195
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.5 Threshold
Kindle Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.7500
F1 Score: 0.8571
```

Figure 25: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.5 Threshold

```
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.6 Threshold
Amazon Max Combiner:
Precision: 0.6667
Recall: 0.3200
F1 Score: 0.4324
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.6 Threshold
Amazon Weighted_Average Combiner:
Precision: 1.0000
Recall: 1.0000
apse Output : 1.0000
```

Figure 26: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.6 Threshold

```

Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.6 Threshold
Goodreads Max Combiner:
Precision: 0.4444
Recall: 0.3077
F1 Score: 0.3636
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.6 Threshold
Goodreads Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.6000
F1 Score: 0.7500

```

Figure 27: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.6 Threshold

```

Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.6 Threshold
Kindle Max Combiner:
Precision: 0.7778
Recall: 0.3415
Output: 0.4746
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.6 Threshold
Kindle Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.7500
F1 Score: 0.8571

```

Figure 28: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.6 Threshold

```

Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.7 Threshold
Amazon Max Combiner:
Precision: 0.7500
Recall: 0.2400
F1 Score: 0.3636
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.7 Threshold
Amazon Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.7500
F1 Score: 0.8571

```

Figure 29: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.7 Threshold

```

Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.7 Threshold
Goodreads Max Combiner:
Precision: 0.5714
Recall: 0.1026
F1 Score: 0.1739
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.7 Threshold
Goodreads Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.4000
F1 Score: 0.5714

```

Figure 30: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.7 Threshold

```

Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Max Combiner with 0.7 Threshold
Kindle Max Combiner:
Precision: 0.6667
Recall: 0.1463
F1 Score: 0.2400
Evaluation Scores with Jaro-winkler, spaCy,TF-IDF Cosine Similarity + Weighted_avg Combiner with 0.7 Threshold
Kindle Weighted_Average Combiner:
Precision: 1.0000
Recall: 0.7500
F1 Score: 0.8571

```

Figure 31: Evaluation Scores with Jaro-Winkler, SpaCy, and TF-IDF Cosine Similarity + Max and Weighted Avg combiner with 0.7 Threshold

To assess the effectiveness of various entity matching strategies, a comprehensive evaluation was performed using three quality measures: Precision, Recall, and F1-score. These metrics were computed for each dataset—Amazon, Kindle, and GoodReads—under different configurations of combiners (Max and Weighted Average) and threshold values (**0.5**, **0.6**, **0.7**). All combinations of matchers, **Jaro-Winkler**, **SpaCy**, and **TF-IDF with Cosine Similarity**, were used to compute pairwise similarity scores, which were then aggregated using the respective combiners.

The results were tabulated into a structured format to compare the influence of different thresholds and combiner functions on the final performance.

DATASET	MATCHER	COMBINER	THRESHOLD	PRECISION	RECALL	F1 - SCORE
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.960000	1.000000	0.980000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	1.000000	1.000000	1.000000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	1.000000	0.360000	0.520000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	1.000000	1.000000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	1.000000	0.240000	0.380000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.750000	0.850000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.930000	0.970000	0.950000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	1.000000	1.000000	1.000000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	0.920000	0.370000	0.470000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	0.750000	0.850000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	1.000000	0.120000	0.210000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.750000	0.850000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.670000	1.000000	0.800000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	0.830000	1.000000	0.900000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	0.440000	0.300000	0.360000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	0.600000	0.750000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	0.570000	0.100000	0.170000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.400000	0.570000

Table 1: Evaluation of different combiners on Amazon, Kindle, and GoodReads datasets across multiple thresholds.

5.1 Possibilities

DATASET	MATCHER	COMBINER	THRESHOLD	PRECISION	RECALL	F1 - SCORE
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.960000	1.000000	0.980000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	1.000000	1.000000	1.000000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	1.000000	0.360000	0.520000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	1.000000	1.000000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	1.000000	0.240000	0.380000
Amazon Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.750000	0.850000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.930000	0.970000	0.950000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	1.000000	1.000000	1.000000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	0.920000	0.370000	0.470000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	0.750000	0.850000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	1.000000	0.120000	0.210000
Kindle Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.750000	0.850000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.500000	0.670000	1.000000	0.800000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.500000	0.830000	1.000000	0.900000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.600000	0.440000	0.300000	0.360000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.600000	1.000000	0.600000	0.750000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Max	0.700000	0.570000	0.100000	0.170000
GoodReads Dataset	Jaro-Winkler, SpaCy, TF-IDF + Cosine Similarity	Weighted Average	0.700000	1.000000	0.400000	0.570000

Figure 32: Proposed Possibilities

After evaluating various combinations of matching strategies, combiners, and threshold values across the Amazon, Kindle, and GoodReads datasets, the optimal configurations were selected based on achieving the highest possible F1-score, while also ensuring high precision and recall.

POSSIBILITY 1: For the Amazon dataset, the combination using Max combiner with a threshold of 0.5 delivered an excellent balance between precision (0.96) and perfect recall (1.0), resulting in a strong F1-score of 0.98. Although a Weighted Average combiner also gave a perfect F1-score of 1.0, it may be slightly more permissive, whereas the Max combiner ensured a more conservative match decision without significant performance loss.

POSSIBILITY 2: For the Kindle dataset, the configuration using the Weighted Average combiner at threshold 0.5 stood out by achieving perfect scores across all metrics—precision, recall, and F1-score—all at 1.0. This indicates a highly effective match quality without any trade-off between false positives or false negatives.

POSSIBILITY 3: For the GoodReads dataset, the configuration with the Weighted Average combiner at a threshold of 0.5 again proved to be the most reliable, achieving a strong F1-score of 0.90, with high recall (1.0) and solid precision (0.83). This indicates that the approach was able to identify most true matches while keeping the number of false matches relatively low.

Thus, the selected configurations represent the most effective trade-off between precision and recall for each dataset, ensuring robust performance tailored to their respective data distributions.

6 Contributions

Snigdha Raghavan Pradhipa

- Schema Matching
- Matcher Development
- Combiner Strategies
- Threshold-Based Selection
- Cardinality Constraints
- Ground Truth Creation
- TP, FP, FN Calculation
- Precision, Recall, F1
- Evaluation and Comparison

Vaishnavi Narasimhaiah Sathish

- Schema Matching

- Matcher Development
- Combiner Strategies
- Threshold-Based Selection
- Cardinality Constraints
- Ground Truth Creation
- TP, FP, FN Calculation
- Precision, Recall, F1
- Evaluation and Comparison

Samrudhi Ramesh Rao

- Schema Matching
- Matcher Development
- Combiner Strategies
- Threshold-Based Selection
- Cardinality Constraints
- Ground Truth Creation
- TP, FP, FN Calculation
- Precision, Recall, F1
- Evaluation and Comparison

Kongarasan Sathiya Moorthy

- Schema Matching
- Matcher Development
- Combiner Strategies
- Threshold-Based Selection
- Cardinality Constraints
- Ground Truth Creation
- TP, FP, FN Calculation
- Precision, Recall, F1
- Evaluation and Comparison