BrandNERD: An Extensive Brand Dataset and Analysis Pipeline for Named Entity Resolution

Nicholas Caporusso[a], Alina Campan[b], Ayush Bhandari, Stephen Kroeger and Sarita Gautam
Northern Kentucky University, Highland Heights, Kentucky, U.S.A.

Keywords: Named Entity Resolution, Brand Canonicalization, String Pattern Matching, Text Mining.

Abstract: Named entity resolution (NER) comprises several steps to address multifaceted challenges, including canonicalization, aggregation, and validation. Nonetheless, NER research is hindered by the scarcity of realistic, labeled corpora that capture the spelling noise and brand proliferation found in data from multiple sources, from e-commerce to social media. In this paper, we introduce the Brand Name Entity Resolution Dataset (BrandNERD), an extensive dataset of real-world brand names extracted from an existing high-traffic retail marketplace. BrandNERD consists of multiple datasets along the entity resolution pipeline: raw surface forms, unique canonical entities, similarity clusters, validated brands, and a lookup table reconciling multiple canonical forms with a list of validated preferred brand labels. In addition to the BrandNERD dataset, our contribution includes an analysis of adequacy of various text similarity measures to the brand NER task at hand, the processing algorithms used in each step of the resolution process, and user interfaces and data visualization tools for manual reviews, resulting in a modular, fully reproducible, and extensible pipeline that reflects the complete NER workflow. BrandNERD, which is released as a public repository, contains the dataset and processing pipeline for over 390,000 raw brand names. The repository is continuously updated with new data and improved NER algorithms, making it a living resource for research in marketing and machine learning, and for enabling more complex downstream tasks such as entity disambiguation and brand sentiment analysis.

1 INTRODUCTION

Named entity resolution (NER) constitutes a critical information extraction task that seeks to locate and classify named entities mentioned in unstructured text into predefined categories. The field has traditionally focused on detecting and classifying entities such as persons, organizations, and locations within well-structured text corpora. However, the scope of NER extends beyond simple classification to encompass the more complex challenge of entity resolution, which involves identifying when different textual mentions refer to the same real-world entity.

NER still represents a fundamental challenge in modern natural language processing, particularly when dealing with brand names, which exhibit significant variation in their surface forms across different platforms and data sources. The challenge is compounded by the inherent noise present in real-world data, including spelling variations, abbreviations, and inconsistent formatting practices that are prevalent in e-commerce and social media environments. As brand surface form variations can be extensive and context-dependent, disambiguation becomes essential for accurate entity linking.

Brand identifiers often mutate in spelling, length, and even language as they move across markets and media. A single company may appear as an acronym (“IBM”), its full legal name (“International Business Machines”), or a colloquial nickname (“Big Blue”). Typos (“Addidas”) and stylistic tinkering with punctuation or capitalization (“McDonald’s” vs. “Mcdonalds”) cause the same entity to manifest in dozens of orthographic guises. Furthermore, individual digital venues might introduce naming quirks. For instance, e-commerce catalogs aggregate or shorten brand and model names into SKUs (e.g., “Nike AM2090”) and use localized names (“Unilever Indonesia”). Also, brand landscapes are anything but static, with thousands of new labels debuting each year, while legacy giants periodically rebranding (e.g., “Dunkin”’ dropping “Donuts,” or Facebook morphing into Meta), or consolidating as a result of mergers and acquisitions (e.g., “Whatsapp by Facebook”), leaving labels that

[a] https://orcid.org/0000-0002-8661-4868
[b] https://orcid.org/0000-0002-9296-3036

Caporusso, N., Campan, A., Bhandari, A., Kroeger, S. and Gautam, S.
BrandNERD: An Extensive Brand Dataset and Analysis Pipeline for Named Entity Resolution.
DOI: 10.5220/0013830600004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 481-488

no longer match current registries.

While there are existing methods for handling name disambiguation in other contexts, such as cleaning up corporate registries or standardizing author names in bibliographic databases, these methods often rely on external validation sources or verified dictionaries. NER corpora, while valuable for general entity recognition tasks, lack the comprehensive coverage needed for brand name resolution and fail to handle diverse naming conventions, abbreviations, and the continuous evolution of brand nomenclature in digital spaces. Contemporary brand NER research faces significant limitations due to the scarcity of realistic, labeled corpora that adequately capture the complexity of brand names. This is particularly pressing in cases where a brand has a narrower scope than a trademark, which makes conventional resources like the USPTO database (of Public Affairs (OPA), 2025) not suitable, as the database may not filter out service-oriented or expired trademarks and can thus present too many potential matches that introduce further confusion. As a result, researchers must devise procedures to systematically recognize, reconcile, and cluster variations.

In this paper, we present BrandNERD, an extensive open-source named entity resolution dataset and pipeline specifically focusing on brands. Built from an online retail marketplace featuring brands also available on popular stores and e-commerce platforms, BrandNERD contains: (1) over 390,000 raw brand names harvested from product listings; (2) a curated set of canonical entities produced using extensible canonicalization rules; (3) similarity clustering tools to expose near-duplicate names; (4) a validation corpus derived from name search and manual review; (5) a lookup table that maps competing canonical forms to a preferred, manually vetted brand label. BrandNERD also includes the tools accompanying the entire NER pipeline, organized in a modular and reproducible set of steps and including convenient user interfaces for data exploration and visualization.

2 RELATED WORK

NER has been extensively studied across multiple research communities, and the foundational tasks of entity resolution have been well-established in the literature. The authors of (Brizan and Tansel, 2006; Saeedi et al., 2021) provide a systematic overview of the processing steps and execution strategies, identifying three primary components: deduplication (i.e., eliminating exact duplicate copies), record linkage (i.e., identifying records referencing the same entity across different sources), and canonicalization (i.e., converting data into standardized forms). These core tasks form the backbone of most ER pipelines and directly relate to the challenges addressed in brand name resolution, where surface form variations must be standardized and linked to canonical representations. Different approaches have been proposed to tackle each task in the NER pipeline. For instance, a recent literature review (Barlaug and Gulla, 2021) surveyed deep learning techniques for traditional ER challenges.

Product entity resolution has emerged as a specialized domain within the broader ER landscape, with unique challenges arising from the heterogeneous nature of product descriptions across different platforms. In (Vermaas et al., 2014), the authors proposed an ontology-based approach for product entity resolution on the web, using the descriptive power of product ontologies to improve matching accuracy. Their method employed domain-specific similarity measures for different product feature types and utilized genetic algorithms for feature weight optimization, achieving F1-measures of 59% and 72% across different product categories. This work demonstrates the importance of domain-specific approaches to entity resolution, particularly relevant for brand name matching, where product context significantly influences resolution accuracy. Unfortunately, in many scenarios, context information is unavailable or unreliable. A study from Microsoft Research (Liu et al., 2007) addressed the challenge of resolving product feature references within customer reviews. Their approach combined edit distance measures with context similarity to group references related to the same product feature, highlighting the importance of contextual information in product-related entity resolution tasks. (Jin et al., 2020) used a combination of neural network models and human annotators to detect and label with brand information a set of over 1.4 million images. Although the work falls into the broader scope of NER, it tackles a different set of challenges. The study focuses on extracting brand names from images instead of processing text. Furthermore, it involves brand recognition rather than resolution tasks. Wang et. al. (Wang et al., 2012) presented a hybrid human-machine approach to entity resolution. The proposed workflow uses machine-based techniques to find pairs or clusters of records likely to refer to the same entity. Then, only these most likely matches are crowd-sourced to humans for review. This approach saves resources while still leading to accurate results. While our approach to ER is also hybrid, there are some differences. The data in our ER problem consists of texts that are usually shorter than product descriptions or restaurant infor-

mation used to evaluate the methodology proposed in (Wang et al., 2012).

The availability of high-quality benchmark datasets has been crucial for advancing entity resolution research. In their work, (Lovett et al., 2014) Lovett et. al. made available a set of 700 of the top U.S. national brands from 16 categories and a large number of descriptive characteristics (such as brand personality, satisfaction, age, complexity, and brand equity. This dataset is appropriate for marketing research, but an ER approach is not proposed in this work. The dataset presented in (Jin et al., 2020) introduced a dataset of 1,437,812 images that contain brands and 50,000 images without brands. The images containing brands are annotated with brand name and logo information. The authors of (Lamm and Keuper, 2023) released the first publicly available large-scale dataset for visual entity matching. They provide 786,000 manually annotated product images containing around 18,000 different retail products, which are grouped into about 3,000 entities. The annotation of these products is based on a price comparison task, where each entity forms an equivalence class of comparable products. The Database Group at Leipzig University (Christophides et al., 2020) has contributed several widely-used benchmark datasets for binary entity resolution evaluation. These include the Amazon-Google Products dataset (Christophides et al., 2020; Saeedi et al., 2021), containing 1,363 Amazon entities and 3,226 Google products with 1,300 known matches, and the Abt-Buy dataset comprising 1,081 and 1,092 entities from the respective e-commerce platforms with 1,097 matching pairs. These datasets have become standard evaluation benchmarks, focusing primarily on general product matching rather than brand-specific resolution challenges. Additional specialized datasets have emerged for different aspects of entity resolution. The Web Data Commons project has created training and test sets for large-scale product matching using schema.org marked-up data from e-commerce websites, covering product categories including computers, cameras, watches, and shoes.

However, despite this variety of available datasets, none specifically address the unique challenges of brand name resolution with the scale and real-world complexity required for comprehensive evaluation.

3 BrandNERD

BrandNERD addresses the critical challenge of NER for brand names by providing an extensive brand dataset of over 394,000 unique raw brand names extracted from a high-traffic retail marketplace, making it significantly larger than existing brand-focused datasets, together with a lookup table that pairs surface names with their resolved names. By doing this, BrandNERD provides researchers with a large-scale dataset that can be utilized for developing and benchmarking machine learning approaches for text similarity, clustering, and resolution tasks, as a trusted source of resolved brand names for sentiment analysis, or as a disambiguation tool for obtaining unique product information in the context of auction and e-commerce websites. In addition, the BrandNERD pipeline implements a comprehensive, modular workflow consisting of six main steps. Although the pipeline itself does not introduce any novel contribution, it makes it convenient for researchers to intervene in any step in the process, where they can replace or extend the algorithms.

BrandNERD, including its datasets and algorithms, lives in a public GitHub repository available at https://bit.ly/3VCc2Sn, and is constantly curated and expanded by the research team. In addition, the repository contains detailed technical documentation about the algorithms in the pipeline, the format of the datasets, and other information that we could not include in this paper. The dataset is released under the Creative Commons BY 4.0 license, so researchers are free to download, fork, integrate, and redistribute the corpus and code, provided they give appropriate attribution. The dataset is continuously updated as new data are processed along the pipeline. Also, the repository accepts pull requests to encourage community-driven enhancements and continuous expansion.

3.1 NER Processing, Pipeline, and Tools

3.1.1 Data Acquisition

The list of raw brand names was acquired from the publicly available product pages of an online marketplace primarily featuring consumer products also available on various popular e-commerce websites, including Amazon, Walmart, Poshmark, Home Depot, and Target. The name of the data source is kept undisclosed to protect the business's anonymity, and any identifiers linking brands with the original marketplaces have been removed from the dataset. Also, although the list was pre-processed to remove other irrelevant information and standardize the data, BrandNERD does not include data acquisition tools, as the pipeline assumes that the surface names have already been acquired. By working with brand strings in isolation, that is, without assuming access to product descriptions, model numbers, category tags, or any other

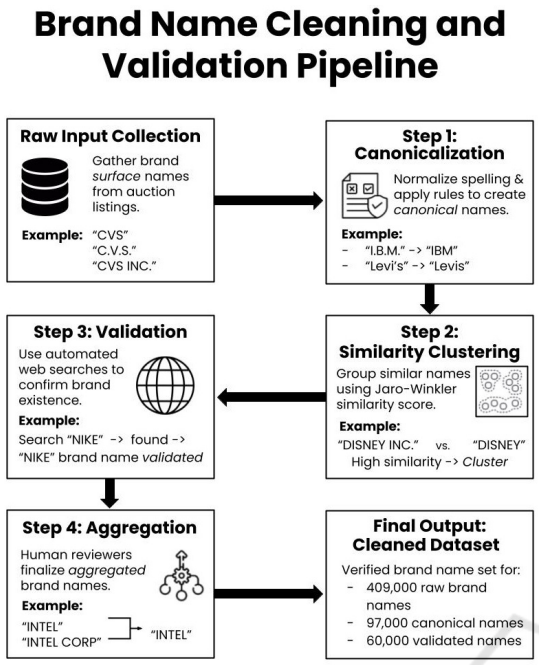

Figure 1: BrandNERD's processing pipeline.

curated metadata that might simplify matching, our methodology deliberately tackles the worst-case scenario in brand NER.

3.1.2 Canonicalization

After acquisition and pre-processing, the first step in our NER pipeline is canonicalization. This process reconciles multiple surface names with a unique canonical name, and it involves the following steps:

1. *Cleaning*, where names are sanitized to remove characters that are not usually part of a brand such as single and double quotes, encoding-specific non-alphanumerical symbols (e.g., Unicode characters), or company designators (e.g., LLC, INC, and CO) that do not commonly appear in the brand name.
2. *Normalization*, which involves processing the sanitized name to remove all non-alphanumerical characters, including whitespace.

To this end, we initially analyzed the dataset of acquired brands to define a set of rules that address most cleaning and normalization requirements. Our tool implements canonicalization rules using regular expressions, which enable parsing strings efficiently and maintain a high level of interpretability. Given the iterative nature of NER, downstream processing steps (i.e., similarity clustering, name search, validation, and review) might inform new rules that can be conveniently implemented in the script to further clean or normalize surface names into more effective canonical forms. Subsequently, the list can be processed to aggregate multiple surface forms with their canonical form.

3.1.3 Similarity Clustering

The second step in our processing pipeline consists in clustering surface or canonical brand names by their similarity. In addition to surface forms, multiple canonical names can also relate to the same entity. Thus, to resolve them, first, we identify and resolve similar canonical names that refer to the same entity.

To this end, we considered several established text similarity algorithms, including Levenshtein and Damerau-Levenshtein (i.e., based on edit distance), Jaro-Winkler, the Jaccard and Sørensen–Dice coefficients on bi- and tri-grams, cosine similarity on TF-IDF character n-grams, the phonetic pair Soundex and Double-Metaphone, the probabilistic Monge-Elkan composite scorer, and the affine-gap Needleman–Wunsch sequence aligner. Subsequently, we implemented and compared four algorithms particularly suitable for our scenario, that is, Levenshtein distance, Jaro-Winkler, phonetic (metaphone) similarity, and cosine similarity with sentence transformer embeddings. Also, we evaluated two combinations of these measures: a hybrid combination of phonetic similarity and Levenshtein distance, and a hybrid combination of sentence transformer embeddings and Levenshtein distance.

To evaluate the performance of these metrics and select one, we manually validated 786 brand names that served as the ground truth to compare the performance of these text metrics in terms of the detection accuracy of the target, ground-truth brand name. For each surface name, we found the three most similar matches according to each text measure from a set of 80,902 validated (through automated searches on retail websites) brand names. We operated with the top three matches rather than the top match only, because the best match according to a text comparison metric is not guaranteed to actually be the name to which the candidate brand needs to be resolved. The chances of finding the correct resolution name increase when considering more top matches, as shown below in Table 1. Also, in practice, the text comparison metrics have sufficient errors that we cannot fully automate the brand resolution process; instead, manual resolution needs to be used, and each brand name can be selected from a series of possible verified target names, which are chosen to be a set of the highest matches – and not just the highest match, according to a text metric.

The Jaro-Winkler similarity was the most accurate

Table 1: Percentage of target matches found at each rank (Top 1–3) or missed entirely using various similarity methods: JW = Jaro-Winkler, Lev = Levenshtein, Phon = Phonetic, Cosine = Cosine with sentence embeddings, P+E = Hybrid (Phonetic + Edit), E+E = Hybrid (Embedding + Edit).

	#1	#2	#3	Not in Top 3
JW	74.81	14.25	5.09	5.85
Lev	61.58	13.49	4.96	19.97
Phon	13.99	1.15	0.00	84.86
Cosine	52.04	14.12	7.63	26.21
P+E	58.65	7.00	3.18	31.17
E+E	56.23	14.38	4.71	24.68

in matching the surface forms in our test dataset to their validated name, as shown in Table 1. This table reports the percentage of candidate brand names in our test dataset, for which the target resolution name was found as the first, second, and third match, respectively, or when the target brand was not found among the closest three matches, when using the respective text distance or similarity measure.

Based on this experimental evaluation, we decided to use the Jaro-Winkler similarity in our validation tool for calculating pairwise name similarity. One of the advantages of this text metric is that strings having similar initial segments are assigned a higher similarity score, which is especially suitable for brand names.

To better understand the relationship between brand name resolution and the pairwise text similarity of brand names, we constructed and analyzed a graph representing these connections. Specifically, we constructed a multigraph $G = (B,(S,R))$, where B is the set of all brand names, S is a set of undirected edges connecting brand name pairs with Jaro-Winkler similarity ≥ 0.9; each edge has a label the respective similarity value, and R is a set of directed edges, where each edge connects a source node, which is one of the 786 brand names that were resolved manually, to a target node, which is the verified brand name that replaces the source brand name node. We must note that not all nodes that had an edge in R also had a corresponding edge in S. For example, the surface name "PHILIPS SONICARE" was resolved to the verified brand name "PHILIPS", but the two brand names did not have similarity ≥ 0.9, and therefore they were not connected by an edge in S. There were 38 surface names in the set of validated brands in the test dataset that were resolved to verified brand names with which they had a Jaro-Winkler similarity of < 0.9. For the remaining 748 resolved brand names, we compared their neighborhoods in S and R, and we found, as expected, that the resolution for surface names, among

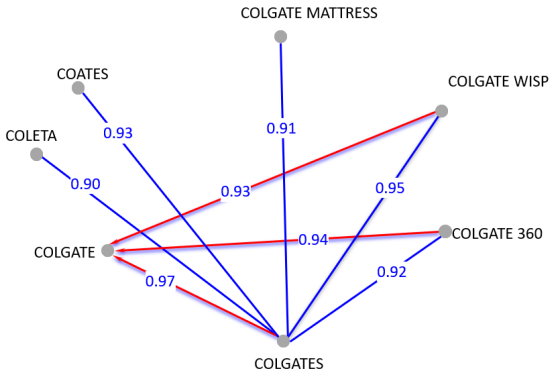

Figure 2: Renamed Nodes' Neighborhood.

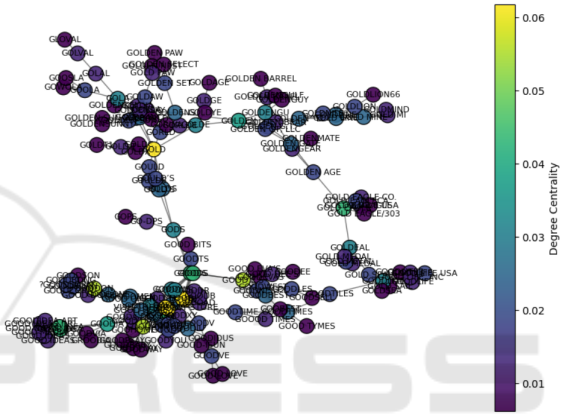

Figure 3: Connected Component in the Similarity Graph. (Large number of nodes and edges make visual inspection of overall component difficult. Included to illustrate scale).

the list of verified brand names, is not always the match with the highest Jaro-Winkler similarity. Figure 2 illustrates one of these situations, where the brand name "Colgate Wisp" was resolved to the verified name "Colgate", with which it has a similarity of 0.93, although there is another brand, "Colgates", with which it has a higher similarity of 0.95. In this case, the more similar brand was not a verified name, but that situation could occur in a large dataset with many similar verified brand names.

Another confirmation that high text similarity is not foolproof and cannot be used alone to determine clusters of similar names was obtained by looking at connected components in the graph (B, S) above. One such connected component, which consists of nodes that have pairwise similarity over 0.9, is shown in Figure 3, where it is visible how some well-formed clusters of similar names are combined together by weak edges forming incorrect bridges between these dense clusters.

An essential aspect of similarity clustering is the choice of a similarity threshold, which is particularly

relevant for the subsequent NER steps. We empirically found 0.9 as a similarity threshold that captures significant but reasonably small clusters. Also, such a high threshold helps identify edge cases that reveal the need for additional rules. For instance, clustering identified the presence of a cluster with more than 3,000 surface names in the form of "VISIT THE [BRANDNAME] STORE", a typical pattern that we addressed with a canonicalization rule.

Our scripts support the integration of other text metrics, instead of Jaro-Winkler. For example, (1) the weighted-trigram Jaccard could be utilized to shrink the impact of boilerplate tokens (e.g., "Manufacturing", "Products") and let the distinctive parts of a brand name dominate the similarity score, and (2) the Inverse Document Frequency (IDF) can be used to increase similarity in case of typos. However, we decided not to implement these at this stage because validating would have required a larger corpus. Also, our normalization process, particularly removing white space, might have an impact on the IDF calculation.

3.1.4 Name Search

The third step of our NER pipeline consists in using an external source of information to check whether the brand exists or not. As most brands involve commercial products sold on popular e-commerce platforms, we decided to use a web search engine to query the brand names in their surface forms and analyze the results. To this end, we developed a JavaScript application for NodeJS environments that leverages the Selenium package to control an instance of the Chrome browser. Specifically, we utilized browser automation to query the Brave search engine for surface names. After storing the first 10 results, the application verifies whether the brand has a presence on popular online e-commerce websites (e.g., Amazon, Walmart, or Home Depot) to validate its canonical name automatically. Although APIs can be utilized for this purpose, unfortunately, the large number of brands in the dataset and the cost of API calls hinder the feasibility of NER research. Also, we used Brave as a search engine instead of more popular ones (e.g., Google search) because it provides easier search automation over a large number of queries. In our processing pipeline, name search and similarity clustering can be executed simultaneously because they consume separate input and produce independent output data.

3.1.5 Brand Validation

The goal of this step is to determine whether any of the surface forms of a canonical name represent a valid brand based on the results obtained from the search engine (i.e., if the brand actually exists). In addition to the checks realized during name search, which allow for a quick evaluation of whether the brand has a store on popular e-commerce websites, this phase performs a more in-depth analysis of the search results. To this end, the validation process utilizes automated string processing and manual review. First, a series of algorithms tokenize the text and URLs to identify the most common strings and compare them with a brand's surface and canonical names. Then, similarity algorithms are utilized to identify clusters around brand names, automatically processing the ones with sufficiently high similarity. Subsequently, the results are reviewed manually thanks to a user interface where the reviewer can mark the brand as valid, invalid, or unsure. In addition to confirming the validity of brands, this step also helps identify brands initially flagged as invalid due to incorrect spelling, use of uncommon surface forms, or not having a storefront on e-commerce websites (e.g., brands sold outside digital marketplaces). This process also helps discover new rules or correct brand names that were not initially part of the dataset. Consequently, new canonical names can be added, and the upstream pipeline and datasets are updated accordingly.

3.1.6 Resolution

In this step, similar canonical forms representing the same brand name are aggregated to validated brand names. This task is similar to what is realized in step one after canonicalization, where multiple surface forms are automatically associated with the same canonical form. However, resolution involves analyzing the semantics of different canonical names (and their associated raw labels) to address typos and misspellings and reconcile brands using multiple names (e.g., "Samsung" and "Samsung electronics", or "Starbucks" and "Starbucks coffee"). To this end, for each similarity cluster found in the second step of our pipeline, our algorithm compares the similarity score between the central node and its direct connections (one-hop neighbors) with the similarity between each of the adjacent nodes and their two-hop neighbors (excluding the central node), to obtain a ranking of the best matches. Moreover, the algorithm assigns a different weight to edges connecting invalid or valid brands. A user interface presents this information to a reviewer who can manually confirm the results. The output of this process is a lookup table with pairs where canonical names representing an invalid brand or being one of the less common canonical representations of a valid brand name (i.e., lookup

value) are reconciled with their corresponding valid and most representative canonical name (i.e., target value).

3.1.7 Manual Review

In the final step, results are reviewed manually to check for potential errors. This is done via a user interface where an agent can observe a sample of the dataset and confirm or reject the results of the validation and resolution. The output of the user interface is a list of canonical names where each brand resolution is flagged as correct, incorrect, or needs review.

3.2 Dataset

As the dataset is designed to continuously grow, the numerical details about the dataset mentioned later in this section reflect the situation at the time of writing.

3.2.1 Raw Brand Names

The dataset contains a total of 394,542 unique raw brand names. This number reflects publicly available surface names collected from the websites mentioned earlier. After obtaining the list using web scraping techniques, we removed syntactically invalid names, including 389 brand names that consisted of numbers only. Raw brand names appear sorted by item count on the platform in descending order (the brands with the most item occurrences appear first).

3.2.2 Canonical Names

The canonicalized dataset comprises 376,613 unique cleaned surface names and 368,703 unique canonical names (93.45% of surface names), as 25,839 brand names from the original dataset were merged with their canonical forms after applying canonicalization rules. Canonical names are sorted following the same criterion as their raw surface forms, even though canonicalization could inherently disrupt distribution sorting, for instance, in case some canonical names represent raw brand names at the top and bottom of the list.
As shown in Table 2 360,919 canonical names (approximately 97.8%) are associated with only one surface name, 7,668 (2.1%) with two, 108 (0.03%) with three, 7 (0.002%) with four, and just one canonical name (0.0003%) is linked to six different surface names, with an average surface name count per canonical name of 1.02 ± 0.15. In addition to individual canonical names, the dataset also retains the relationship with the original surface names to maintain consistency and reference across the datasets.

Table 2: Canonical name counts by surface name count (descending).

Index	surfaceNameCount	canonicalCount
0	6	1
1	4	7
2	3	108
3	2	7668
4	1	360919

3.2.3 Similarity Clusters

This dataset contains 782,299 pairs featuring 273,845 unique canonical names (74.27% of the total) showing a Jaro-Winker similarity score higher than 0.9 with other brands. The dataset consists of two components, one with one-way similarity matches with unique $< brand_i, brand_j >$ pairs, and one with two-way matches, that is, with occurrences featuring both $< brand_i, brand_j >$ and $< brand_j, brand_i >$ pairs, together with their similarity scores.
This dataset can be used to identify misspellings, resolve brand names having different canonical forms, or identify new canonicalization rules. Additionally, this dataset can serve as a benchmark for other similarity clustering algorithms to enhance the accuracy of identifying false positives and false negatives.

3.2.4 Brand Search Results

This dataset consists of the following components:

- The list of 368,703 canonical names (100%) checked with the search engines.
- The list of 72,303 brand names (19.61% of the total) found through web search that are potentially valid candidates, comprising the canonical name, the raw brand name used as the search term (i.e., query), and the first URL resulting from a match.
- For each canonical name checked with the search engine, the list of the top 10 results obtained from the web search, including the title, source URL, and short description. The results obtained for each canonical name are sorted using the same criterion used by the search engine, that is, by relevance.

3.2.5 Validated Brands

This dataset contains the list of 32,114 brand names (8,7% of the total) that were processed in the validation step, where each canonical name is associated with a value of 1 if the brand is validated, -1 if

the brand was not validated, and zero if the canonical name could not be determined to be valid or not. 31,622 brand names (i.e., 98.47% of the total) were found to be valid, while the remaining were invalid.

3.2.6 Entity Resolution Lookup Tables

This dataset contains:

- The list of 768 canonical names (2.40% of the total) processed.
- The list of 824 pairs of canonical brand names where a brand name was reconciled to another canonical form.

4 CONCLUSION

Our work aims to address a long-standing gap in brand-oriented NER research by offering a large-scale, openly licensed dataset and pipeline that mirrors the complexity of commercial data. By anchoring BrandNERD in nearly 400,000 raw surface forms scraped from a high-traffic marketplace, we provide a benchmark whose size, noise profile, and continual growth far exceed those of prior corpora, offering scholars and practitioners an authentic testbed for research and experimentation in various tasks of interest to several scientific communities.

BrandNERD is a large-scale, open-source dataset for brand NER, supported by an end-to-end, modular workflow for disambiguating, deduplicating, and validating brand names. The framework combines interpretable rule-based canonicalization, modular similarity metrics, browser-automated web search, and user interfaces and data visualization tools for human curation. The dataset includes multiple interconnected components: 376,613 cleaned surface names mapped to 368,703 canonical names, 782,299 similarity pairs covering 273,845 unique canonical names, search results for all canonical names with 72,303 potentially valid candidates identified, 32,114 validated brands (with 98.47% confirmed as valid), and 824 entity resolution pairs in lookup tables (with new data being added regularly), resulting in the most extensive datasets currently available for research. The GitHub repository of the project is continuously updated and shared under Creative Commons licensing, providing researchers with an authentic, large-scale testbed that reflects the complexity and noise profile of real commercial brand data.

In our future work, we will (1) finalize validation of all current canonical entries; (2) manually review brands to expand the lookup table into a gold-standard; (3) refine canonicalization rules and trial misspelling-aware similarity measures; (4) enrich resolution by incorporating item-level descriptions from auction listings; and (5) explore density-based clustering and centrality metrics on the similarity graph to surface latent brand groups and identify authoritative canonical labels.

REFERENCES

Barlaug, N. and Gulla, J. A. (2021). Neural networks for entity matching: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–37.

Brizan, D. G. and Tansel, A. U. (2006). A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 6(3):5.

Christophides, V., Efthymiou, V., Palpanas, T., Papadakis, G., and Stefanidis, K. (2020). An overview of end-to-end entity resolution for big data. *ACM Computing Surveys (CSUR)*, 53(6):1–42.

Jin, X., Su, W., Zhang, R., He, Y., and Xue, H. (2020). The open brands dataset: Unified brand detection and recognition at scale. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4387–4391. IEEE.

Lamm, B. and Keuper, J. (2023). Retail-786k: a large-scale dataset for visual entity matching. *arXiv preprint arXiv:2309.17164*.

Liu, J.-J., Cao, Y.-B., and Huang, Y.-L. (2007). Effective entity resolution in product review domain. In *2007 International Conference on Machine Learning and Cybernetics*, volume 1, pages 106–111. IEEE.

Lovett, M., Peres, R., and Shachar, R. (2014). A data set of brands and their characteristics. *Marketing Science*, 33(4):609–617.

of Public Affairs (OPA), U. O. (2025). United states patent and trademark office.

Saeedi, A., David, L., and Rahm, E. (2021). Matching entities from multiple sources with hierarchical agglomerative clustering. In *KEOD*, pages 40–50.

Vermaas, R., Vandic, D., and Frasincar, F. (2014). An ontology-based approach for product entity resolution on the web. In *International Conference on Web Information Systems Engineering*, pages 534–543. Springer.

Wang, J., Kraska, T., Franklin, M. J., and Feng, J. (2012). Crowder: Crowdsourcing entity resolution. *arXiv preprint arXiv:1208.1927*.