Bulletin of Yerevan University: Economics
2026, Vol. 17, No. 1(47), June, 83-91
https://doi.org/10.46991/BYSU.G.2026.17.1.083

IDENTIFICATION AND HARMONIZATION OF MATERIAL VALUES AND PRODUCT NAMES IN A GROUP OF COMPANIES USING NLP METHODS

RAFIK MASHURYAN*
Yerevan State University

Abstract: The article examines the problem of heterogeneous material-value and product names in a group of companies. The same physical material may be registered under different abbreviations, spellings, languages, internal codes, or incomplete descriptions in the accounting and enterprise systems of separate subsidiaries. This reduces the quality of consolidated reporting, complicates procurement analysis, inventory control, price comparison, and managerial decision-making at the group level. The problem is formulated as an Entity Resolution and product-matching task and is addressed through Natural Language Processing and machine learning methods. A dataset of 17,258 material and product names was annotated manually and used to train a domain-specific Named Entity Recognition model. The proposed pipeline extracts structured components from free-text descriptions and creates a basis for unified material classification, centralized procurement, and analytical control in a group of companies. The article also adds a model-evaluation framework based on the confusion matrix, precision, recall, and F1-score.

Key words: *material classification, data standardization, ERP systems, master data management, entity resolution, named entity recognition, NLP, precision, recall, F1-score*

Introduction

Groups of companies that operate simultaneously in construction, production, and related business areas frequently purchase and use the same or very similar materials, such as cement, rebar, electrical cables, fittings, fasteners, and other inventory items. However, these materials are often recorded differently in local accounting or enterprise-resource-planning systems. A single physical item can appear under different textual descriptions, abbreviations, internal codes, measurement formats, or language variants.

From a management perspective, such heterogeneity prevents the group from answering basic but important questions: how much of a specific material was purchased during a period across all subsidiaries, what part of it was consumed in projects and what part remains in stock, whether there are opportunities for centralized procurement and

* **Rafik Mashuryan** – PhD Student at the YSU Chair of Mathematical Modeling in Economics
E-mail: rafikmashuryan@gmail.com, ORCID ID: https://orcid.org/0009-0007-4782-7389

Received: 12.05.2026
Revised: 19.05.2026
Accepted: 19.06.2026

price negotiation, and whether the same item is purchased at materially different prices by different companies.

This article defines the task as the identification and harmonization of material values and product names stored in the data repositories of different companies. The proposed solution is based on NLP and machine learning tools that transform noisy free-text descriptions into structured attributes and support further matching at group level.

Problem Statement and Managerial Motivation

The core problem is the absence of a unified classifier for material values and products. Each company in the group may preserve its own accounting and operational information system, which leads to duplicated and inconsistent representations of the same real-world object. In a construction and production environment, the inconsistencies typically include:

- different abbreviations, for example 'A500C rebar', 'A500 armature', or 'armature 12 mm';
- language and transliteration differences across Armenian, Russian, and English records;
- incomplete or noisy descriptions, for example cement grades or cable descriptions with missing units;
- internal codes that are meaningful inside one subsidiary but are not aligned with codes used by another subsidiary;
- different numerical and measurement formats for diameter, length, voltage, package size, or weight.

From a management perspective, the group needs to answer a number of simple but important questions:

- What is the total quantity of a given material (e.g., M500 cement) purchased over a specific period by all companies?
- What portion of this material has been used in projects, and what portion remains in inventory?
- Are there opportunities for centralized procurement, price negotiations, or standardization of material specifications?
- Comparison of the purchase price of the same material across any two companies, with the aim of optimizing prices and avoiding artificially inflated prices and fraud.

If each company uses its own naming system, answering group-level analytical questions requires manual examination and matching of thousands of material rows. This process is time-consuming, error-prone, and difficult to scale. Therefore, the task can be narrowed to the design of a practical pipeline that integrates with corporate data systems and produces a unified group-level material classifier.

Theoretical Background

The inconsistency of material names is a central issue of Master Data Management (Reddy, 2025). The quality of material master data affects procurement, warehouse management, reporting, and strategic control, especially when data are distributed across several ERP systems. In computer science and data-management literature, this type of problem is usually formulated as Entity Resolution or record linkage. Two entity profiles match if they refer to the same real-world object (Christen, 2012). The probabilistic

foundations of record linkage are commonly traced to the Fellegi-Sunter model, while later ER research extends this logic to modern database, machine learning, and NLP settings (Fellegi & Sunter, 1969; Getoor & Machanavajjhala, 2012).

For a group consisting of n companies, each company C_k maintains a set of material or product records R_k = {r_k1, r_k2, ..., r_km}. The research task is to construct a matching function that estimates whether two records from different systems refer to the same real material. If r_i and r_j are two records, the model estimates a similarity function s(r_i, r_j), where larger values indicate a higher probability of equivalence. A binary matching decision is then made by applying a threshold τ: two records are treated as a match when s(r_i, r_j) ≥ τ, and as a non-match otherwise.

Threshold selection. In the empirical run reported in Table 5, in the experimental specification, a threshold of τ = 0.70 was used as an illustrative operational threshold. This value was selected on the validation subset as a compromise between avoiding incorrect material merges and preserving a reasonable ability to detect repeated materials. A higher threshold would increase precision but would miss more true matches, while a lower threshold would increase recall but would also merge more descriptions that refer to different specifications.

Similarity aggregation. The final score s(r_i, r_j) was formed as a weighted combination of the extracted components: the core noun received the highest weight because it identifies the material family; adjectives and complements captured type or grade; numerical values and measurement units captured technical specifications; and the remaining normalized text similarity captured minor spelling or transliteration differences. In practical terms, the aggregation can be written as s = 0.45*s_noun + 0.20*s_modifier + 0.25*s_number-unit + 0.10*s_text. Numerical-unit mismatches were treated as a strong penalty rather than a minor difference, because materials with similar names but different diameters, voltage, length, weight, or grade usually refer to different inventory items. The use of weighted and learnable string-similarity components is consistent with adaptive duplicate-detection approaches proposed in earlier data-mining research (Bilenko & Mooney, 2003).

Pairwise grouping and non-transitive cases. After pairwise matches are detected, records are grouped through connected components, but the resulting groups are checked for specification consistency. If r_i is similar to r_j and r_i is similar to r_k, while r_j is not similar to r_k, the group is not accepted automatically. Such cases are treated as ambiguous clusters and are either split according to the numerical-unit attributes or sent to manual review. This rule is important for material harmonization because a general name can serve as a bridge between two specific but different materials.

$$r_i \equiv r_j \Leftrightarrow s(r_i, r_j) \geq \tau$$

This formulation is important because it converts a managerial harmonization problem into a measurable task. It also makes it possible to define evaluation criteria such as precision, recall, and F1-score. In large datasets, naive pairwise comparison has quadratic complexity, so blocking or filtering methods may be required to reduce the number of candidate pairs before detailed matching (Papadakis et al., 2020). Duplicate-record detection studies also emphasize that approximate matching requires both suitable similarity metrics and scalable search strategies (Elmagarmid et al., 2007).

Recent research also emphasizes the role of learning-based similarity models in Entity Resolution. Neural approaches can support product matching by learning not only

direct lexical overlap, but also deeper contextual and structural similarities between heterogeneous descriptions (Cicco & Firmani, 2019). In particular, BERT-based similarity learning has shown effectiveness for product-matching tasks, where semantically similar product descriptions may differ in wording, abbreviations, or attribute order (Trącz et al., 2020). More recent entity-matching studies show that deep-learning models are especially useful for textual and dirty entity-matching tasks, while BERT provides contextual bidirectional representations that improve semantic comparison of text descriptions (Mudgal et al., 2018; Devlin et al., 2019).

Natural Language Processing is the set of methods through which computers process and analyze text written in natural language. In groups of companies, much management-relevant data are not available as clean structured tables but as short free-text descriptions, purchase requests, warehouse comments, technical specifications, or product names. NLP therefore acts as an intermediate layer that converts unstructured text into structured features suitable for economic analysis and managerial decision-making.

Named Entity Recognition is one of the fundamental NLP tasks. Its goal is to detect and classify meaningful spans in text (Yadav & Bethard, 2018). In the context of material names, NER can be used to extract the core noun, modifiers, numerical-measurement segments, units, and complementary descriptors. Once these elements are separated, material descriptions become more comparable across different subsidiaries. Neural NER architectures, including character-aware and sequence-labeling models, are particularly relevant when domain-specific entity spans must be learned from annotated examples (Lample et al., 2016).

Data and Methodology

The empirical part of the work applies a Named Entity Recognition approach to material descriptions collected from different companies in the group. A total of 17,258 material-value and product names were collected and manually annotated. The annotations identify the relevant components inside each description, such as nouns, adjectives, numerical values, measurement units, and complements. These annotations form the training data for the neural model.

Data source and sampling. The dataset was compiled from material master data, warehouse, and procurement nomenclature records used by construction and production companies within the group. The sampling unit was the individual material or product name as recorded in the local enterprise or accounting system. The sample included all available non-empty material-name records from the selected corporate databases at the time of extraction; purely technical blank rows, duplicate empty entries, and records without a textual description were removed before annotation. Because the objective was to harmonize operational nomenclature, the dataset intentionally preserved noisy spelling, abbreviations, mixed languages, and inconsistent measurement formats instead of normalizing them manually before training.

For NER implementation, the study uses spaCy, an open-source NLP framework developed in Python (*spaCy Usage Documentation*, n.d.). A blank multilingual model was created and trained directly on the domain corpus of material descriptions. This approach is appropriate for a group of companies because material descriptions may contain several languages or writing systems. The annotations for

each record were converted into spaCy training format: (text, {'entities': [(start_char, end_char, label), ...]}) (*spaCy Usage Documentation*, n.d.). The software implementation is also aligned with the published spaCy system description for industrial-strength NLP workflows (Honnibal et al., 2020).

The model was trained through mini-batch stochastic gradient descent for 100 iterations. Dynamic batch sizes were gradually increased from 4 to 32, dropout was used to reduce overfitting, problematic records were skipped with warnings instead of stopping the training process, and loss values were printed after each iteration to monitor convergence. In Table 2, the sequence of the model training steps is presented.

Table 1. **Examples of annotated material descriptions**

rebar 8 mm	{'noun': 'rebar', 'unit and numbers': '8 mm'}
rebar 10 mm	{'noun': 'rebar', 'unit and numbers': '10 mm'}
nail 80 mm	{'noun': 'nail', 'adj': '', 'unit and numbers': '80 mm'}
nail 100 mm	{'noun': "nail", 'adj': "", 'unit and numbers': "100 mm"}
mesh 4 mm 20x20	{'noun': 'mesh', 'adj': '', 'unit and numbers': '4mm 20*20'}
0.4 kV current transformer 1200/5	{'noun': current transformer', 'adj': '', 'unit and numbers': ['0,4kv' ,'1200/5']}
tuff building stone straight cut	{'noun': 'stone', ' adj ': 'tuff',' noun complement': straight cut '}
nut 14	{'noun': "nut", 'adj': "", 'unit and numbers': "14"}
washer 14x30	{'noun': " washer ", 'adj': "", 'unit and numbers': "14x30"}
bolt 14*50	{'noun': bolt', 'unit and numbers': '14*50'}

Source: Author's processing based on the manually annotated material-name dataset of 17,258 descriptions.

Table 2. **Sequence of model-training steps**

Step	***Component***	***Description***	***Result***
1	Data export	Import of material descriptions	Raw text and labels
2	Annotation conversion	Mapping labels to text offsets	spaCy format
3	Validation	Checking token-aligned offsets	Clean dataset
4	Model initialization	Blank multilingual model with NER	Prepared model
5	Label registration	Nouns, adjectives, complements, units, and numerical segments	Complete label set
6	Training	100 iterations, mini-batches, dropout	Trained NER model
7	Evaluation	Confusion matrix, precision, recall, F1-score	Quality indicators
8	Application	Automatic extraction of material-name components	Structured data

Source: the spaCy training pipeline and NER component documentation, as well as general NER model-training methodology (Explosion, n.d.; Honnibal et al., 2020; Yadav & Bethard, 2018; Lample et al., 2016).

Validation strategy. To evaluate generalization rather than memorization, the annotated corpus was divided into training, validation, and test subsets in an 80/10/10 proportion. The training subset was used to fit the NER model, the validation subset was used to monitor convergence and choose the threshold τ, and the hold-out test subset was used only for the final calculation of precision, recall, and F1-score. During the split, records with similar material families were distributed across subsets where possible, so that the test results would reflect performance on unseen descriptions rather than repeated copies of the same text.

Evaluation Framework: Confusion Matrix, Precision, Recall, and F1

After training, model quality must be evaluated. For classification and entity extraction tasks, evaluation is based on a comparison between predicted labels and true labels. The basic instrument is the confusion matrix (Table 3), which describes classification outcomes quantitatively. In a binary interpretation of matching or entity detection, the following quantities are distinguished.

***Table 3.*Confusion-matrix quantities**

TP	True positives; Correctly detected matches or entity spans
FP	False positives; Records or spans incorrectly classified as positive
FN	False negatives; True matches or entity spans missed by the model
TN	True negatives; Correctly rejected non matches or negative cases

Source: the standard confusion-matrix framework for classification evaluation (Sokolova & Lapalme, 2009; Powers, 2011; Christen, 2012).

Precision describes what share of the examples classified as positive by the model are actually correct. It expresses the reliability of positive predictions. High precision means that the number of false positives is small, which is important when incorrect harmonization of two different materials may distort procurement, inventory, or price-comparison decisions (Table 3).

Recall describes what share of all true positive examples were successfully identified by the model. It expresses the detection capability of the model. High recall means that the model misses few real matches or valid entity spans, which is important when undetected duplicates keep the material classifier fragmented (Table 3).

Precision and recall are often in a trade-off relationship. Increasing the decision threshold usually increases precision but reduces recall; lowering the threshold usually increases recall but may reduce precision. Because these two indicators may move in opposite directions, F1-score is used as a combined measure. F1-score is the harmonic mean of precision and recall and penalizes imbalance: if either precision or recall is low, the F1-score also becomes low (Table 4).

Table 4. **Evaluation indicators used for the NER and matching task**

Precision	TP / (TP + FP); Reliability of positive predictions; lower false-positive risk
Recall	TP / (TP + FN); Ability to find true matches or entity spans; lower false- negative risk
F1 score	2 * Precision * Recall / (Precision + Recall); Balanced combined measure of precision and recall

Source: Standard classification-evaluation measures for precision, recall, and F1-score (Sokolova & Lapalme, 2009; Powers, 2011)

In the context of a group of companies, these indicators are more informative than accuracy alone. Material data are typically imbalanced: the number of non-matching pairs can be much larger than the number of true matches, and accuracy may remain high even when the system misses many important duplicate materials. Precision, recall, and F1-score therefore provide the theoretical basis for evaluating whether the proposed NLP pipeline is suitable for operational harmonization. In our discussed case, the results are presented in Table 5.

Interpretation of the obtained results. The reported precision value of 0.75 means that three quarters of the positive matches or extracted entity spans accepted by the model were correct. The recall value of 0.64 is lower than precision, which indicates that the model was more conservative in accepting matches and missed part of the true positive cases. This is expected in the material-harmonization setting because many descriptions are short, contain different abbreviations, or differ mainly by numerical specifications. The resulting F1-score of 0.69 can therefore be interpreted as satisfactory for an initial applied model, but it also shows that recall should be improved through more annotated examples, better normalization of measurement units, and additional domain-specific rules.

Table 5. **Indicators**

Indicator	Values
Precision	0.75
Recall	0.64
F1	0.69

Source: Author's calculation on the hold-out test subset of the annotated material-description dataset.

The proposed approach supports the creation of a unified data environment at the group level. Once material names are transformed into structured components and comparable representations, the group can identify repeated materials, standardize descriptions, compare purchase prices across subsidiaries, estimate opportunities for centralized procurement, and improve inventory planning. The same approach can also support internal-process KPIs such as the number of identified materials, the number of unified names, precision, recall, F1-score, the number of incorrect matches, and the potential volume of centralized purchases.

Implementation limitation. The proposed system should be used as a decision-support tool rather than as a fully autonomous replacement for master data governance. High-confidence matches can be standardized automatically, low-confidence non-matches can remain separate, and borderline or non-transitive cases should be reviewed by a responsible data specialist. This mixed workflow reduces manual effort while protecting the company from costly errors caused by incorrectly merging different material specifications.

Conclusion

The article investigated the problem of inconsistent material value and product names in a group of companies as a data management and Entity Resolution task. An NLP-based approach was proposed, combining text preprocessing, Named Entity Recognition, and extraction of structured components from material descriptions. The practical implementation used a manually annotated dataset of 17,258 material and product names and a domain-specific spaCy NER model trained for 100 iterations.

The results and methodological framework show that NLP methods can automate the standardization of material descriptions and create a foundation for further matching and classification. By adding the evaluation framework based on the confusion matrix, precision, recall, and F1-score, the approach becomes measurable and suitable for managerial use. It can serve as a basis for centralized master data management, procurement optimization, inventory control, price comparison, and improved analytical processes in groups of companies.

The main contribution of the study is that it connects a practical managerial problem with a measurable data-science procedure. In groups of companies, material master data are often fragmented across subsidiaries, and this fragmentation directly affects procurement planning, inventory control, price comparison, and the reliability of consolidated reporting. By treating material-name harmonization as a combination of Named Entity Recognition and Entity Resolution, the proposed approach makes it possible to move from manual comparison of thousands of heterogeneous descriptions toward a more systematic and reproducible workflow. The extraction of nouns, modifiers, numbers, units, and complementary attributes creates a structured basis for comparing records that were originally stored as short, noisy, and multilingual text strings.

The empirical indicators also show both the usefulness and the current limitations of the model. A precision value of 0.75 indicates that the model is reasonably reliable when it accepts a match or identifies a relevant entity span, while the recall value of 0.64 shows that a portion of true matches is still missed. This balance is acceptable for an initial applied model because incorrectly merging different material specifications may create higher managerial risk than leaving some potential duplicates for manual review. At the same time, the F1-score of 0.69 confirms that further improvement is necessary before the model can be used as a fully automated component of master data governance.

Future development should therefore focus on expanding the annotated dataset, standardizing measurement units before model training, testing alternative threshold values, and improving similarity aggregation for cases where the same material family has different technical specifications. The proposed pipeline can also be strengthened by adding a manual validation layer for borderline and non-transitive matches. In this form,

the approach can serve as a practical decision-support tool for centralized procurement, inventory optimization, and unified analytical reporting, while still preserving human control over cases where incorrect harmonization may lead to financial or operational errors.

References

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer.

Cicco, V., & Firmani, D. (2019). Interpreting deep learning models for entity resolution: An experience report using LIME.

Papadakis, G., Skoutas, D., & Thanos, E. (2020). *A Survey of Blocking and Filtering Techniques for Entity Resolution*.

Reddy, A. (2025). An indepth guide to materials master data management. Verdantis. https://www.verdantis.com/materials-master-data-management

Explosion. (n.d.). spaCy usage documentation. Retrieved from https://spacy.io/usage

Trącz, J., et al. (2020). BERT-based similarity learning for product matching. Proceedings of the Workshop on Natural Language Processing in E-Commerce (EComNLP), 66-75.

Yadav, V., & Bethard, S. (2018). A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models. *Proceedings of COLING*.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. Proceedings of NAACL-HLT 2016.

Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., & Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. Proceedings of the 2018 International Conference on Management of Data, 19-34.

Bilenko, M., & Mooney, R. J. (2003). Adaptive duplicate detection using learnable string similarity measures. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 39-48. https://doi.org/10.1145/956750.956759

Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. Proceedings of the VLDB Endowment, 5(12), 2018-2019. https://doi.org/10.14778/2367502.2367564

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1-16. https://doi.org/10.1109/TKDE.2007.250581

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210. https://doi.org/10.1080/01621459.1969.10501049

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies, 2(1), 37–63.