Predicting CVE Enrichment Fields from Descriptions with a Publication-Year Holdout

Tara Dixit
Stanford University
Stanford, CA
tdixit@stanford.edu

Abstract

Public vulnerability records pair natural-language descriptions with structured fields such as severity, Common Weakness Enumeration (CWE), Common Vulnerability Scoring System (CVSS) base score, and CVSS vector components. This study measures how reliably simple lexical models reconstruct those fields when evaluated on later records. The data contain 86,784 usable complete-case records from a frozen copy of the FKIE NVD JSON feeds for CVE-ID years 2022 through 2024. The main experiment trains on records published in 2022 and 2023 and tests on 36,539 records published in 2024. Word unigram and bigram TF-IDF features feed logistic-regression classifiers for severity, CWE, and eight CVSS v3.x components, while Ridge regression predicts the base score. The model reaches 66.05% severity accuracy, 67.15% closed-set CWE accuracy, 0.997 CVSS mean absolute error, and 86.13% average component accuracy. These aggregate results hide important weaknesses. Original-label CWE accuracy is 63.60%; unseen labels and labels seen one to nine times in training have zero original-label recovery. Predicted score and severity disagree for 19.63% of test records. Performance is lower in test groups with less TF-IDF similarity to training descriptions, but this observational pattern does not establish that similarity caused the difference. Exact-description grouping changes the main metrics only slightly, but semantic near-duplicates remain unmeasured. The results support a limited conclusion: lexical models are useful baselines for common labels and repeated patterns, but they do not reliably reconstruct every enrichment field from description text alone.

1 Introduction

A CVE record provides a shared identifier and a text description for a publicly known vulnerability. Security workflows often also use structured metadata, including CVSS scores and vectors and CWE labels. CVSS summarizes vulnerability severity and technical characteristics, while CWE describes the underlying weakness type [1, 2]. The National Vulnerability Database (NVD) distributes standards-based vulnerability data through its website, APIs, and data feeds [3, 4].

Predicting structured fields from a description is a natural text-classification problem, but it is easy to evaluate in a misleading way. A random split can place repeated product templates in both training and test data. The year in a CVE identifier is not necessarily its publication year. Rare CWE labels can be collapsed into an `OTHER` class, causing an apparently correct prediction without recovery of the original weakness. Finally, separate models for score, severity, and vector components can generate outputs that contradict one another.

This project asks:

> *How reliably can lexical models reconstruct structured CVE enrichment fields in a publication-year holdout, and how sensitive are the results to split design, duplicates, and label frequency?*

This project builds a transparent baseline and examines how its results change under several evaluation choices. The evaluation records the source snapshot, exact split membership, row-level predictions, uncertainty, duplicate sensitivity, label-frequency behavior, nearest-training similarity, cross-field consistency, and target provenance. Here, target provenance means the organization or record entry that supplied the selected CVSS or CWE target.

2 Background and Related Work

CVSS is an open framework for communicating vulnerability characteristics and severity. Version 3.1 defines eight Base metrics covering attack conditions and impacts, from which a numeric score and qualitative severity can be derived [1]. CWE is a community-developed taxonomy of software and hardware weakness types [2].

Prior work has treated CVE-to-CWE mapping and CVSS prediction as natural-language tasks. ThreatZoom combines statistical and semantic features with a hierarchical classifier for CVE-to-CWE mapping [5]. V2W-BERT uses hierarchical multiclass learning and reports both random and temporal evaluations [6]. CVSS-BERT predicts CVSS vector metrics with one classifier per component [7]. These studies show the value of learned text representations, but their reported scores are not directly comparable with this project because the datasets, label spaces, time periods, and evaluation rules differ.

This study instead uses TF-IDF with linear models as a transparent baseline. The models are implemented with scikit-learn [8]. The design tests what can be recovered from lexical patterns before introducing a larger representation model.

3 Data

3.1 Frozen source cohort

The source snapshot is FKIE release `v2026.08.29-000010`. FKIE reconstructs legacy-style JSON feeds from NVD API data and states that its repository is not endorsed or certified by NVD [9]. The local compressed and decompressed source files were verified with SHA-256 hashes before training.

The three source files contain 98,162 records from CVE-ID years 2022, 2023, and 2024. A record is usable when it contains:

- an English description of at least 20 characters;
- CVSS v3.x data with a base score; and
- a usable CWE label.

These rules produce 86,784 usable complete-case records. Of 11,378 rejected records, 3,759 lack usable CVSS v3.x data and 7,619 lack a usable CWE. The complete-case filter is an important limitation because excluded records may differ systematically from included records.

3.2 Publication-year holdout

The source files are organized by CVE-ID year, not publication year. Within the usable cohort, 17,323 records were published in 2022, 24,952 in 2023, 36,539 in 2024, 7,473 in 2025, and 497 in 2026. The main experiment uses only publication years 2022 through 2024. It is therefore a frozen source cohort, not a complete census of every CVE published in those years.

The 2022 and 2023 publication-year pool is divided into training and validation sets with a deterministic 80/20 stratified split using seed 42. All 2024 records form the test set. Table 1 shows the final sizes. I kept validation data for evaluating future changes without using the test set. Because this configuration was fixed in advance, validation was not used for fitting, tuning, early stopping, or model selection.

4 Models

Descriptions are represented using word unigram and bigram TF-IDF features. The vocabulary is fitted only on the training split. The vectorizer uses at most 50,000 features, a minimum document frequency of two, and sublinear term frequency.

Table 1: Main publication-year split.

Split	Records	Publication years	Use
Training	33,820	2022–2023	Fit representation and models
Validation	8,455	2022–2023	Preserved, not used for selection
Test	36,539	2024	Final evaluation and audits

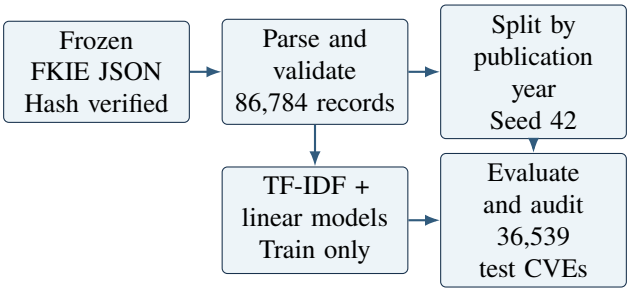

Figure 1: Overview of the reproducible pipeline.

Separate logistic-regression classifiers predict severity, CWE, and each of the eight CVSS v3.x Base components. The selected targets include both CVSS 3.1 and 3.0 records: training contains 33,757 version 3.1 and 63 version 3.0 targets, validation contains 8,442 and 13, and test contains 36,042 and 497. All classifiers use $C = 1.0$ and random seed 42. Severity and CWE classifiers allow up to 1,000 iterations, while component classifiers allow up to 500. Ridge regression with $\alpha = 1.0$ predicts the numeric CVSS base score; predictions are clipped to the valid interval from 0 to 10.

The CWE classifier uses the 100 most frequent training labels plus `OTHER`. Training and test labels outside those 100 classes are mapped to `OTHER` for closed-set evaluation. Original test labels are retained separately, preventing successful `OTHER` prediction from being mistaken for recovery of a rare original CWE.

The simple baseline predicts the most frequent training label for every classification task and the mean training score for regression. All hyperparameters were fixed before the reported runs. All logistic-regression models converged without recorded warnings.

5 Evaluation

Severity is evaluated with accuracy, macro F1, and balanced accuracy. The main four-class scope includes CRITICAL, HIGH, MEDIUM, and LOW. A single 2024 record with severity NONE and score 0.0 is kept in the all-label result but excluded from the four-class comparison because NONE does not occur in training.

CWE evaluation includes closed-set accuracy and macro F1, original-label exact accuracy, metrics restricted to true top-100 labels, and precision, recall, and F1 for `OTHER`. CVSS score uses mean absolute error (MAE). Vector performance is the mean fraction of eight correctly predicted components per row.

Because score, severity, and components are modeled separately, the evaluation also checks their agreement. Score/severity consistency asks whether the predicted Ridge score falls in the interval implied by predicted severity. A second audit applies the CVSS 3.1 scoring formula to the eight predicted components and compares the recomputed score with the independent score and severity predictions [1]. This audit uses the CVSS 3.1 formula even though the source targets include both CVSS 3.1 and 3.0 records.

Uncertainty is estimated with 1,000 ordinary row-level bootstrap samples using a fixed seed. This treats rows as independent. Repeated and related descriptions weaken that assumption and may make the intervals too narrow. The intervals also do not measure variation from alternate split seeds or model configurations.

6 Main Results

Table 2 compares the fixed model with the simple baseline on the complete publication-year test set.

Table 2: Main publication-year holdout results. Higher is better except for MAE.

Metric	Majority/mean baseline	TF-IDF logistic/Ridge
Four-class severity accuracy	47.70%	**66.05%**
Four-class severity macro F1	16.15%	**46.99%**
Closed-set CWE accuracy	19.76%	**67.15%**
Closed-set CWE macro F1	0.33%	**34.63%**
CVSS MAE	1.450	**0.997**
Vector partial match	63.11%	**86.13%**
Score/severity consistency	0.00%	**80.37%**

The lexical model improves over the simple baseline on every reported predictive metric. The row-bootstrap 95% intervals are 65.55–66.54% for four-class severity accuracy, 66.68–67.60% for closed-set CWE accuracy, 0.987–1.005 for CVSS MAE, and 85.97–86.29% for average component accuracy.

The aggregate metrics require qualification. Four-class balanced accuracy is 47.05%, close to macro F1 and much lower than ordinary accuracy. LOW severity recall is 0% despite 525 LOW examples in training. This shows that the overall severity accuracy is driven mainly by the more common classes.

Closed-set CWE accuracy is 67.15%, but original-label exact accuracy is 63.60%. The 3.54-point gap comes from predictions that count as correct after rare labels are collapsed to `OTHER` but do not recover the original weakness. The model predicts `OTHER` for 15.39% of records even though its true mapped prevalence is 7.62%. `OTHER` precision is 23.02%, recall is 46.52%, and F1 is 30.80%.

7 Sensitivity and Error Audits

7.1 Year definition and exact descriptions

A sensitivity run splits by the year embedded in each CVE ID. Its main scores are slightly higher, but the design has a larger training set and different membership. Publication year differs from CVE-ID year for 9,381 training, 2,304 validation, and 5,463 test records in that design. The result is therefore not described as a publication-time evaluation.

The main test set contains 544 rows whose lowercase, whitespace-collapsed description occurs in training. Removing those rows leaves 35,995 test records. Severity accuracy becomes 65.92%, closed-set CWE accuracy 67.82%, CVSS MAE 1.001, and vector partial match 86.12%. The changes are small and do not all move in the same direction.

A separate group-aware run keeps exact normalized-description groups within one split and excludes test groups found in the earlier-year pool. It contains 33,744 training, 8,531 validation, and 35,972 test records, with zero exact normalized-description groups shared across splits. Its results remain close to the main run: 66.19% four-class severity accuracy, 67.74% CWE accuracy, 1.003 CVSS MAE, and 86.15% vector partial match. This test addresses exact descriptions, not semantic near-duplicates.

Table 3: Sensitivity to split and exact-description handling.

Design	Test	Four-class severity	CWE	CVSS MAE
CVE-ID year	36,306	67.30%	68.72%	0.976
Publication year	36,539	66.05%	67.15%	0.997
Group-aware publication year	35,972	66.19%	67.74%	1.003

The severity column uses the four-class scope, which excludes the single NONE record.

7.2 Label frequency

Table 4 shows that CWE performance depends strongly on training frequency and on the definition of `OTHER`. Unseen and one-to-nine-frequency labels have nonzero closed-set accuracy because they are mapped to `OTHER`, but their original-label exact accuracy is zero.

Table 4: CWE results by original-label training frequency.

Training frequency	Test records	Closed-set accuracy	Original-label accuracy
Unseen	361	45.98%	0.00%
1–9	1,262	41.68%	0.00%
10–49	2,003	30.95%	0.85%
50–199	4,464	25.45%	25.45%
200 or more	28,449	77.64%	77.64%

Frequency alone does not explain all minority-class errors. LOW availability recall is 0.13% with 624 training examples, HIGH attack-complexity recall is 9.43% with 1,219 examples, PHYSICAL attack-vector recall is 19.16% with 253 examples, and HIGH privileges-required recall is 29.80% with 3,584 examples. Possible explanations include class imbalance, weak textual evidence, label ambiguity, and target noise; this experiment does not isolate them.

7.3 Nearest-training similarity

Every test description is compared with the nearest training description in the training-fitted TF-IDF space. Performance is lower in the least similar band, which contains almost half of the test set (Table 5). TF-IDF similarity of 1.0 means identical retained feature vectors, not necessarily identical source text.

Table 5: Performance by nearest-training TF-IDF similarity.

Similarity	Test share	Severity	CWE	CVSS MAE
1.0	2.14%	77.49%	43.35%	0.814
0.9 to <1.0	4.34%	86.19%	89.97%	0.698
0.7 to <0.9	25.64%	75.45%	85.21%	0.809
0.5 to <0.7	21.48%	64.58%	73.21%	1.018
<0.5	46.40%	59.13%	53.32%	1.127

The pattern is observational. Similarity bands can differ in label distribution, product families, and description style, so the table does not estimate a causal effect of similarity.

7.4 Cross-field consistency

The independent Ridge score and severity classifier agree under the CVSS severity boundaries for 29,366 records, or 80.37% of the test set. They disagree for 7,173 records. True-score MAE is 0.952 for consistent outputs and 1.179 for inconsistent outputs.

Scores recomputed from the eight predicted vector components differ from the independent Ridge score by 0.791 points on average. Only 41.75% are within 0.5 points, and 70.66% are within 1.0 point. Severity derived from predicted components agrees with independently predicted severity for 73.69% of rows. These contradictions are a direct cost of training each output independently.

7.5 Target provenance and qualitative review

The parser follows a fixed preference rule for selecting CVSS and CWE targets when multiple entries exist. Target provenance records which organization and record entry supplied the selected value. A separate audit re-applied the selection rule to all 78,814 split rows and reproduced every selected score and CWE target. The source composition differs across splits. NVD supplied 31,680 of 33,820 selected training CVSS targets but 20,804 of 36,539 test targets. Secondary entries supplied 2,137 training targets and 15,735 test targets. These counts show a source-composition difference, but the audit does not establish that it caused a performance difference.

The qualitative review uses 24 examples selected by predeclared, hash-based rules. The reviewed cases include rare-label mappings, frequent-CWE errors, severity errors, component errors, and high-similarity pairs. Some failures occur despite direct lexical cues, while other descriptions omit information needed to infer every structured field. The review is diagnostic and is not used to estimate population frequencies.

8 Discussion

The main result is that lexical evidence provides a useful baseline for reconstructing common vulnerability fields under a later-year holdout. The model performs better than majority and mean predictions on the reported metrics. It uses a small dependency set and records the code, data, split, and output hashes needed to inspect each run.

However, the audits change the interpretation of the headline numbers. Closed-set CWE accuracy partly rewards predicting `OTHER`; rare original labels remain largely unrecoverable. Ordinary accuracy hides zero recall for LOW severity. Similarity to training text is strongly associated with performance. Independently modeled fields contradict one another often enough that field-level metrics alone are incomplete.

The pipeline should therefore be treated as a baseline or decision-support component, not an automatic source of truth. A practical system could route rare-label, low-similarity, and inconsistent outputs for human review. It should also preserve the source record and model version with each prediction.

9 Limitations

This study has several limitations:

- The data are one frozen, complete-case source cohort rather than a complete publication-year census.
- The experiments use one split seed and one fixed lexical configuration.
- Keeping an unused validation set reduced the earlier data available for fitting even though no model selection was performed.
- A publication-year holdout does not show that publication year alone caused differences between results.
- Row-level bootstrap intervals may be too narrow because related descriptions are not fully independent.
- Exact-description grouping does not remove every semantic near-duplicate.
- Top-100-plus-`OTHER` is a closed CWE task and cannot recover the original rare label through `OTHER`.
- CVSS and CWE target sources and assessment types vary across records and splits.
- Descriptions sometimes omit attack prerequisites or impacts needed to reconstruct the structured target.
- The project does not compare lexical models with language models, retrieval systems, or fine-tuned adapters.

10 Reproducibility and Artifact Design

Each authoritative training run preserves a manifest, dataset statistics, duplicate audit, metrics, exact split IDs, and one prediction per test CVE. Manifests record the frozen release, source hashes, executed-code hashes, output hashes, row counts, environment versions, runtime, convergence status, and iteration counts. Raw predictions and split IDs remain outside version control because of size, while their hashes and row counts are retained.

The repository contains 25 synthetic unit tests covering parsing, timestamps, deterministic splits, group separation, CWE mapping, severity scope, malformed outputs, CVSS consistency, similarity-band boundaries, and analysis helpers. The tests do not download data or train the full models. The complete experiment, not the unit suite, evaluates tens of thousands of real CVEs.

An executed-source archive preserves the exact versions of code named by authoritative manifests. This matters because later code cleanup should not silently change the record of what generated an earlier result.

11 Conclusion

A fixed TF-IDF logistic/Ridge pipeline performs better than simple baselines on common CVE enrichment fields under a publication-year holdout. The result is useful but incomplete. Rare CWEs, minority values, dissimilar descriptions, and cross-field contradictions remain important weaknesses. The evaluation also shows that the year used for splitting, the duplicate definition, the CWE label space, and the source of selected targets all affect what the reported metrics mean.

The evidence supports a narrow conclusion: lexical models are useful, inexpensive baselines for common patterns, but they do not reliably reconstruct every enrichment field from description text alone. Future work should evaluate hierarchical CWE prediction, constrained joint CVSS modeling, semantic duplicate control, provenance-aware analysis, repeated time windows, and stronger text representations using the same fixed data and evaluation setup.

References

[1] FIRST.Org, Inc. Common Vulnerability Scoring System v3.1: Specification Document. https://www.first.org/cvss/v3.1/specification-document.

[2] MITRE. Common Weakness Enumeration. https://cwe.mitre.org/.

[3] National Institute of Standards and Technology. National Vulnerability Database. https://www.nist.gov/itl/nvd.

[4] National Institute of Standards and Technology. NVD Data Feeds. https://nvd.nist.gov/vuln/data-feeds.

[5] Ehsan Aghaei, Waseem Shadid, and Ehab Al-Shaer. ThreatZoom: CVE2CWE using Hierarchical Neural Network. arXiv:2009.11501, 2020. https://arxiv.org/abs/2009.11501.

[6] Siddhartha Shankar Das, Edoardo Serra, Mahantesh Halappanavar, Alex Pothen, and Ehab Al-Shaer. V2W-BERT: A Framework for Effective Hierarchical Multiclass Classification of Software Vulnerabilities. arXiv:2102.11498, 2021. https://arxiv.org/abs/2102.11498.

[7] Mustafizur R. Shahid and Hervé Debar. CVSS-BERT: Explainable Natural Language Processing to Determine the Severity of a Computer Security Vulnerability from its Description. arXiv:2111.08510, 2021. https://arxiv.org/abs/2111.08510.

[8] F. Pedregosa et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011. https://www.jmlr.org/papers/v12/pedregosa11a.html.

[9] Fraunhofer FKIE. NVD JSON Data Feeds. https://github.com/fkie-cad/nvd-json-data-feeds.