Management Summary — German Financial NER

Project: Named Entity Recognition for German official financial/legal documents

Team: Tim Greß, Jan Tölken, Viktor Hoffmann, Lars Böhmer / Group 23 **Date:** 2026-01-25

Executive Summary

- **Goal:** Automatically extract key information from German financial/legal documents to support faster review and analysis.
- **What we extracted:** Organization names, monetary amounts, and references to laws and regulations.
- **Main finding:** There is no single “best” solution—methods trade off accuracy, speed, cost, and transparency depending on the use case.
- **Practical takeaway:** Simple methods work well for structured information (amounts and legal references), while organization names are more variable and therefore harder.

1 Task Overview

German financial and legal documents contain key pieces of information - company names, monetary amounts, and legal citations, buried in dense, unstructured text. Manually finding and extracting these is slow and error-prone. Our task was to build and compare automated systems that can reliably identify and extract these three categories from raw document text. We evaluated a range of approaches, from simple pattern-matching rules to modern AI models, to understand the practical trade-offs between extraction quality, speed, cost, and transparency.

2 Data, Labeling, and Evaluation Setup

- **Source data:** FinCorpus-DE10k, a collection of 10,000 German financial and legal PDF documents (e.g., prospectuses, central bank reports, and legal texts), released via Hugging Face.
- **Preprocessing:** We selected a subset of 1,000 documents while ensuring coverage across all document types. The raw text was standardized into a consistent, structured format suitable for downstream modeling.
- **Human-labeled reference data:** We created a high-quality set of 170 manually labeled sentences. Each sentence was double-checked by at least two team members to improve annotation consistency.
- **Train/test split:** The labeled data were split into 80% training and 20% test data using a stratified split to preserve label distributions. Reproducibility was ensured by fixing the random seed.
- **Evaluation:** Methods were compared on their ability to detect entities while minimizing false positives.

3 Approaches Evaluated

- **RuleChef (main focus):** An approach that attempts to automatically derive human-readable extraction rules from a small number of labeled examples using AI, iteratively refining them based on observed errors. The main goal is transparency and reusability.
- **OpenAI (reference approach):** A direct extraction approach where documents are submitted to OpenAI models, which return entity labels for ORG, MON, and LEG.
- **Baseline and comparison methods:** Improved hand-written rules, a simple statistical baseline, a sequence model, a lightweight hybrid system, and a modern German language model fine-tuned on the labeled data.

4 External Resources

- **Dataset:** FinCorpus-DE10k (see *Source Data*, Section 2).
- **AI assistance (engineering):** GitHub Copilot for code completion and partial AI assistance for the GUI labeling tool and documentation drafts.

5 Key Challenges Encountered

- **Labeling ambiguity:** Decisions about entity boundaries (e.g., whether a currency symbol belongs to a monetary value) are subjective and can strongly affect measured performance. A more forgiving “partial match” evaluation often better reflects real-world usefulness.
- **Class imbalance:** Approximately 93% of tokens are non-entities, making it difficult for statistical models to reliably learn rare entity patterns.
- **Unstructured ORG entities:** Organization names vary widely and lack consistent surface patterns, limiting the effectiveness of pure rule-based approaches.
- **RuleChef instability:** Iterative refinement sometimes overgeneralized or produced malformed rules, leading to noticeable variability across runs.

6 Results (see Figure 1)

- **Highest extraction quality:** Direct OpenAI extraction (best overall results, but slowest runtime).
- **Best practical trade-off:** spaCy combined with heuristics (strong quality with very fast runtime).
- **Best candidate for improvement with more data:** German BERT (trainable, but currently limited by the small labeled dataset).
- **Promising candidate with further improvements:** RuleChef shows potential through reusable, human-readable rules, but requires further stabilization and tuning to be competitive in this setting.

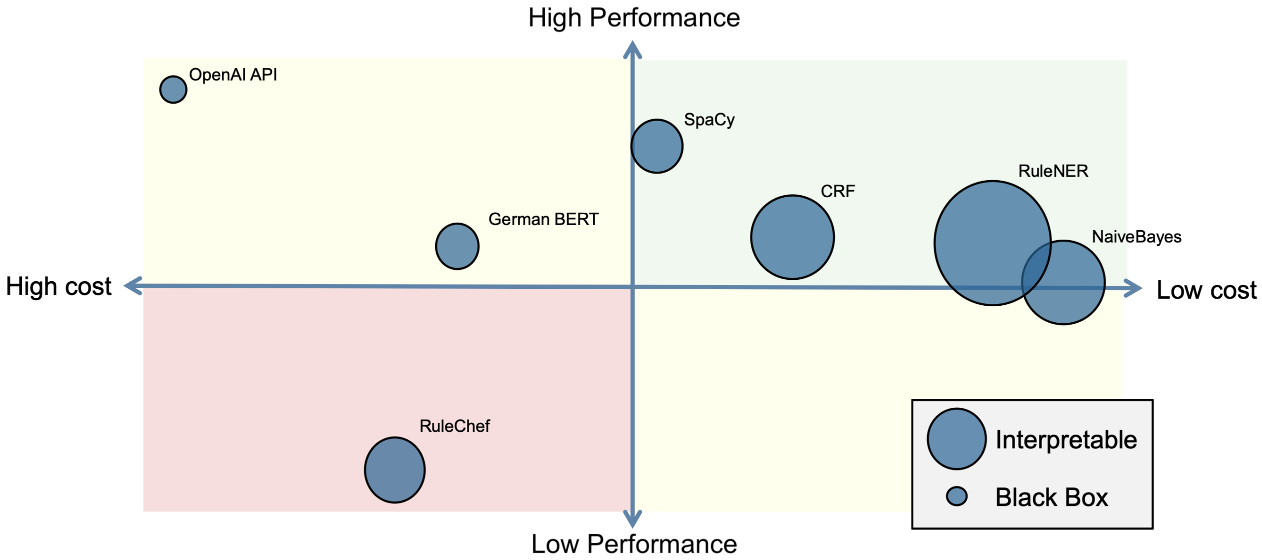

Figure 1: Evaluation of different approaches on the test set

7 Limitations and Risks

- **Limited labeled data:** With only 170 labeled sentences, generalization to unseen documents—especially for ORG entities—is limited.
- **Limited evaluation size:** The test set contains only 30 sentences, so results should be interpreted as indicative rather than definitive.
- **Labeling subjectivity:** Some annotation decisions are inherently ambiguous and affect measured scores without changing practical usefulness.
- **Operational constraints:** Although highest-performing, the OpenAI-based approach is costly, not easily scalable, and may raise compliance or privacy concerns.

8 Possible Next Steps

- **Grow the labeled dataset:** Expand the gold standard—especially ORG-heavy cases—and further tighten annotation guidelines.
- **RuleChef stabilization:** Introduce stronger output validation, tighter constraints, and more targeted feedback to reduce variability.
- **Plan for deployment trade-offs:** No single method optimizes quality, speed, cost, and transparency simultaneously; practical systems will likely combine approaches.
- **Improve ORG detection:** Apply a hybrid strategy combining domain-specific name lists, simple rules, and trainable models.

9 Team Contributions

Most tasks were completed collaboratively, there were no unforeseen issues. Primary contributions were:

- **Tim Greß**: Data preprocessing; RuleChef experiments and integration.
- **Jan Tölken**: Data preprocessing; OpenAI prompting and experimentation.
- **Viktor Hoffmann**: Data labeling; baseline model implementations and comparisons.
- **Lars Böhmer**: Data labeling; evaluation setup, scoring, and analysis.