**Rebar Manufacturer Dataset – Extraction & Analysis Summary Report**

**Summary**

This project consolidated a validated dataset of 442 unique rebar manufacturers by integrating data from four key industry sources and employing a multi-phase, API-driven contact enrichment process. The most accurate method involved Google-based profile filtering (Serper API) and email verification (EmailDetective API). This final approach enabled scalable, role-specific contact discovery ideal for business intelligence and lead generation.

**Data Sources & Collection**

Collected from four primary industry bodies:

* Bar UK (12) – UK-based suppliers
* CARES UK (270) – International certified producers
* Eurofer (49) – European steel producers
* Eurometal (111) – Steel traders/distributors

Three-phase approach:

1. Web Scraping (Selenium/BeautifulSoup) – Initial data retrieval
2. API Trials (Hunter.io, SerpAPI) – Proved inaccurate
3. Serper + EmailDetective – Accurate, scalable, and current contact enrichment

**Methodology Overview**

* Extraction:

Company data was collected using a hybrid approach based on the number of entries per source.

Sources with fewer than 50 companies (e.g., Bar UK and Eurofer) were processed manually, ensuring accuracy and completeness.

Sources with more than 50 companies (e.g., CARES UK and Eurometal) were extracted using automated web scraping via Selenium and BeautifulSoup to handle pagination and structured data retrieval efficiently.

* Standardization:

All extracted data was cleaned and normalized into a unified format consisting of four key columns: Company Name, Email Domain, Email, and Source. Encoding issues were corrected using the ftfy library, and duplicate records were removed based on company name and domain.

* Enhancement:

Using LinkedIn profiles discovered via the Serper API, role-specific contacts were targeted (e.g., Finance Director, Head of Operations). An email generation algorithm produced likely addresses using standard patterns (e.g., firstname.lastname@domain.com). Only profiles showing current employment were retained, excluding any with terms like “former” or “previously at”.

* Validation:

Generated emails were processed through the EmailDetective API, which assigned deliverability, confidence, and suspicion scores. Low-confidence entries and undeliverable addresses were excluded to ensure data quality.

**Key Assumptions & Limitations**

* Assumptions:

The process assumes that most companies use standard email formats (e.g., firstname.lastname@domain.com) and that LinkedIn profiles are up-to-date, accurately reflecting current roles and employers. These assumptions are essential for reliable email generation and contact relevance.

* Limitations:

The dataset focuses on publicly available information from four EU/UK-centric associations, meaning global or unaffiliated manufacturers are excluded. API rate limits (from Serper and EmailDetective) occasionally slowed processing, and despite multiple validation layers, some false positives may remain (e.g., outdated roles or incorrect domains).

* Challenges:

Non-English characters and localized company naming conventions created minor difficulties during parsing and standardization. Additionally, private companies or those with minimal online presence lacked sufficient data for contact extraction or validation.

**Insights**

* Market: UK & European concentration; overlapping memberships suggest strong industry links
* Data Quality: Proprietary domains improve accuracy; finance/ops roles most accessible
* Business Value: Verified contacts support B2B engagement and industry analysis
* Data source: Cares was able to provide us with most companies, we should focus on it more.

**Future Scopes**

* Maintain: Quarterly updates; automate contact monitoring
* Expand: Include global data; gather firmographics (size, capacity)
* Improve: Use ML for smarter email generation; real-time scoring; enhanced batch pipelines

**Conclusion**

This project evolved from unreliable scraping to a robust, API-driven method combining Serper for role-targeted search and EmailDetective for validation. The resulting dataset is accurate, scalable, and actionable, ideal for market mapping and outreach.