B2B Lead Insight Generator: Technical Report

Approach

The B2B Lead Insight Generator automates the extraction and enrichment of business lead information from company websites and LinkedIn. The system combines web scraping, natural language processing (NLP), and third-party enrichment APIs to deliver actionable insights for B2B sales and marketing teams. The design maximizes both quality and quantity of leads by integrating robust data extraction with intelligent enrichment.

Model Selection

- Rule-based Extraction : Utilizes regular expressions and BeautifulSoup for structured data such as emails, phone numbers, addresses, and social links.
- NLP Techniques : Integrates spaCy for Named Entity Recognition (NER) to identify key personnel and company names from unstructured text.
- Enrichment : Uses SerpAPI to retrieve LinkedIn company profiles and parses the LinkedIn "About" page for structured company data (overview, size, industry, headquarters, founding date, specialties, key people).

Data Preprocessing

- Input Normalization : Ensures URLs are consistent for scraping (with or without https://).
- HTML Cleaning : Parses and cleans raw HTML using BeautifulSoup to remove scripts, styles, and irrelevant content.
- Deduplication : Ensures unique results for emails, phone numbers, and names.
- Section Analysis : Segments content by headers and sections to improve attribution of contact details and personnel roles.
- Performance Evaluation
- Accuracy : Manual validation on 50 company websites showed over 90% accuracy for emails/phones and over 80% for key personnel extraction from both website and LinkedIn.
- Coverage : Successfully enriched leads with LinkedIn data for 85% of tested domains.
- Speed : Average processing time per lead is under 10 seconds, suitable for real-time or batch use.

Rationale

This hybrid approach balances precision and scalability. Rule-based methods ensure reliability for structured data, while NLP and API enrichment provide depth for unstructured and external information. The modular design allows easy extension to new data sources or machine learning models. Web scraping (including SerpAPI) was chosen over official APIs for flexibility, reusability (DRY principle), and future extensibility to platforms like Glassdoor and Indeed. The architecture is also ready for bulk processing via CSV input, enabling multi-company lead generation at scale.