LeadScout Pro Technical Report

Objective

Develop an AI-powered lead generation and prioritization tool that simulates realistic scraping, enriches companies with business intelligence, and dynamically scores leads to identify high-potential businesses.

Approach

- **Lead Generation**: Simulated Google Maps scraping using GPT-4 Turbo to generate structured business metadata.
- **Lead Enrichment**: AI-estimated fields include revenue, funding, employee count, startup status, founding year, business model, and industry. GPT also returns realistic review texts.
- **Review Sentiment**: Reviews are passed through TextBlob to compute a sentiment score (0-1).
- **Lead Scoring**: Custom scoring function built using scikit-learn's MinMaxScaler and weighted dynamic scoring based on feature variability (coefficient of variation).

Model Selection

- **GPT-4 Turbo** (OpenAI) was used for both lead generation and enrichment tasks due to its ability to simulate human-like business metadata and reviews.
- **TextBlob** was used for sentiment analysis based on polarity.
- **scikit-learn** was used for normalization and dynamic score calculation.

Data Preprocessing

- Default fallback values used for missing data (e.g., sentiment=0.5, revenue=0).
- Outlier values capped (e.g., revenue > $1T discarded).
- Parsed JSON from GPT is validated and normalized.
- TextBlob sentiment converted from (1 to +1) to (0 to 1) scale.

Parallel Processing

- **ThreadPoolExecutor** used for concurrent GPT API calls in both scraping and enrichment phases.
- Significantly improves throughput when handling large volumes of leads.

Performance Evaluation

- No ground truth labels available, so evaluation was heuristic-based.
- Lead completeness >90% for enriched fields.
- Score distribution visualized using histograms and pie charts for clarity.

Citation

- GPT-4 Turbo: https://platform.openai.com/docs/models/gpt-4

- TextBlob: https://textblob.readthedocs.io/en/dev/

- scikit-learn: https://scikit-learn.org/