Country Intelligence Web Mining

Armand Debray, Christopher Stein, Keerthi Anand, Krupasankari Ragunathan, Namra Shah, Navya Kondaveeti, Shashank Sridhar, Dr. Shoaib Khan

Purdue University, Department of Management, 403 W. State Street, West Lafayette, IN 47907

debray@purdue.edu; stein64@purdue.edu; ksangee@purdue.edu; kragunat@purdue.edu; shah1041@purdue.edu; nkondav@purdue.edu; sridha70@purdue.edu; khan180@purdue.edu

ACKNOWLEDGEMENTS

We would like to thank Professor Shoaib Amjad Khan and our industry partner for this opportunity, their guidance, and support on this project.

OUR DILEMMA

Web-scraping is tedious...

Bugs, runtimes, and data loads can take up all of a data analyst's time.

How does applying **automation** in critical **labor-intensive** web-scraping processes free up company *manpower* and *resources?*

PAIN POINTS

Country Web-Scraping Currently **COSTS...**

20+	*$1500+*	*4-5*
hours per project	monthly avg cost	team members

HOW DO WE ADDRESS THIS?

Step 1 *Find websites and important data*

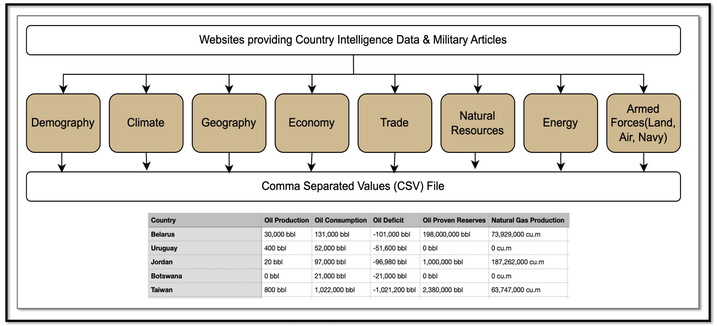
Step 2 *Create efficient code to extract data*

Step 3 *Automate extraction process for ppt*

PURDUE UNIVERSITY | Mitch Daniels School of Business

DATA

We leveraged **web scraping techniques to extract country-specific data across multiple domains**, including defense activity, climate, natural resources, energy, economics and trade. This data was cleaned, structured, and visualized using dashboards to uncover key trends. By integrating data from **sources such as the World Bank, CIA Factbook, and defense reports**, we provide a **comprehensive perspective on global intelligence.**



Country	Oil Production	Oil Consumption	Oil Deficit	Oil Proven Reserves	Natural Gas Production
Belarus	30,000 bbl	131,000 bbl	-101,000 bbl	198,000,000 bbl	73,929,000 cu.m
Uruguay	400 bbl	52,000 bbl	-51,600 bbl	0 bbl	0 cu.m
Jordan	20 bbl	97,000 bbl	-96,980 bbl	1,000,000 bbl	187,262,000 cu.m
Botswana	0 bbl	21,000 bbl	-21,000 bbl	0 bbl	0 cu.m
Taiwan	800 bbl	1,022,000 bbl	-1,021,200 bbl	2,380,000 bbl	63,747,000 cu.m

Fig 1. Sourcing the data

METHODOLOGY

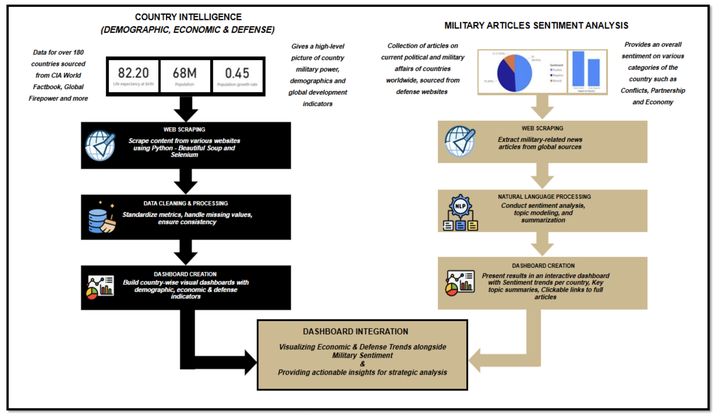


Fig 2. Data Pipeline

MODEL BUILDING

The second part of the project requires the scraping and analysis of the news articles and webpages to gather information on the happenings and how the client can leverage this to gain strategic business benefits.

- **Evaluation Metrics:** The ML model using NLTK, SERPAPI, VADER and Classification is evaluated using accuracy and MAE (Mean Absolute Error) to rank countries based on defense activity. Cross-validation ensures robustness.
- **Defense Activity Ranking:** The model identifies areas of interest, such as rising military tensions or increased defense spending and attaches importance rank to articles using the TF-IDF method.
- **Classification & Insights:** The model provides rankings or classifications based on defense data, highlighting key factors that shape a country's defense profile.
- **Future Improvements:** Enhancing data quality, integrating more features (e.g., arms trade data), and ensuring real-time updates will improve accuracy.

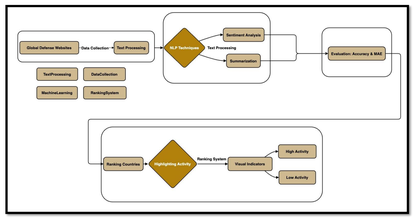


Fig 3. Breaking down the model

RESULTS

PART - 1

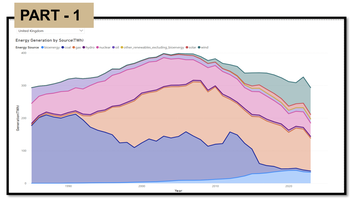


Fig 4. UK Energy

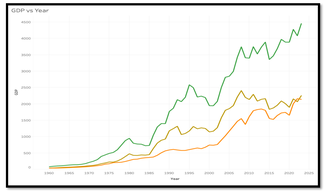


Fig 6.UK GDP

PART - 2

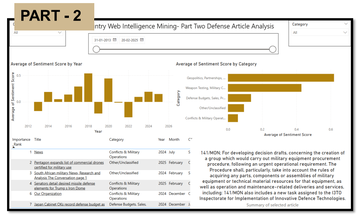


Fig 5. Defense Articles

- The energy and GDP charts shown are for UK and are a part of a dashboard *with data for all countries web scraped using Python*.
- Similarly, we have covered other parameters like demographics, natural resources, climate etc. for a comprehensive analysis.
- The defense dashboard displays the scraped data of the countries along with the sentiment analysis and a short summary of the article.
- This contributes towards the end goal of the analyst and helps in *devising a strategic business plan to pitch to potential customers*.

DEPLOYMENT & LIFE CYCLE MANAGEMENT

Our process ensures *continuous, up-to-date* intelligence for the client across all countries by leveraging automated web scraping, combined with manually curated dashboards that allow country-specific filtering. By integrating Python scrapers, APIs, Power Automate, and Power BI, **we established a scalable, efficient pipeline that reduces manual analyst effort by 30%, while delivering faster and more reliable insights.**

Key Steps:

- Scrape data from public sources for all countries using Python and APIs.
- Store the data in CSV files, organized by country, for easy access.
- Manually create Power BI dashboards using the latest scraped data.
- Country filter added to dashboards, enabling users to switch between countries dynamically.
- Conduct quarterly reviews to update scrapers for website changes or new client requirements.
- Collect stakeholder feedback for continuous process improvement.

This lifecycle minimizes manual data collection while ensuring flexibility for client-specific analysis across all countries of interest.

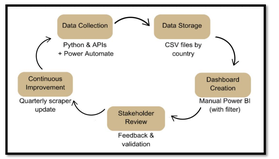


Fig 7. End-to-end process from data collection to dashboard creation, review, and updates

CONCLUSIONS

Our automated intelligence pipeline addresses the high costs, time requirements, and team effort currently associated with country web scraping. By automating manual processes, we reduce the 20+ hours per project, lower the $1500+ monthly cost, and optimize the workload of the 4-5 team members involved.

By integrating Python web scrapers, APIs, Power Automate, and Power BI, we ensure real-time, scalable insights across all countries while cutting manual effort by 30%. Additionally, we provided the client with our Power Automate flow, enabling independent automation of data extraction and dashboard updates.

To further streamline operations, we explored RPA and Agentic AI, ensuring adaptability to evolving data sources and client needs through quarterly reviews and stakeholder feedback. This AI-driven workflow enhances decision-making with minimal analyst effort, providing a cost-effective, efficient, and sustainable approach to global intelligence