CASE STUDY

SCRAPING DATA OFF THE WEB WITH ARTFUL INTELLIGENCE

Abstract

Sophia Almeida, AVP at a Europe-based investment research firm, had data on her mind all the time. Wanting to expand her data acquisition coverage, she was hamstrung by the lack of technology innovation within the firm – specifically the dependency on manual processes used by her teams to collect market information. Read this case study to discover how a partnership with Infosys BPM brought in much needed Artificial Intelligence to scrape information off the web, delivering benefits that exceeded all of Sophia's expectations.

Infosys®
Navigate your next

Data, data, all around

Sophia Almeida is an AVP at a privately held Europe-based investment research company. Responsible for data metrics and data standardisation, Sophia oversees the collection, storage, retrieval, and usage of all the data within the organisation. She is thus involved in developing and implementing data management strategies, managing data quality and accuracy, publishing insights, leading data analysis efforts, and ensuring data security and privacy.

Sophia's large team manually extracted financial data through extensive searches on the web, using sources such as company websites, company filings, and LinkedIn. However, their manual processes not only tended to take long, but they also caused data, accuracy, and standardisation issues. When Sophia wanted to expand the scale of her research coverage, she had a limited resource pool to pick from and balked at the costs of adding to her already large team. The last straw was the lack of technology innovation in the investments-focused firm, so she finally started mulling over outsourcing as a possible option to resolve all her challenges.

The art and intelligence of data gathering

In early 2023, Sophia struck up a conversation with Avantika Naidu, Infosys BPM's data and analytics expert, outlining her challenges. Avantika responded with how Infosys' Generative AI (GenAI) technology would be a great solution, detailing how it could scrape unstructured web content automatically to deliver numerous benefits. Through minimising human effort, it would not only help scale Sophia's data acquisition efforts and improve accuracy but deliver better data enrichment and accelerate time-to-market.

Convinced, Sophia struck a partnership with Infosys BPM, and Avantika directed her team of AI experts to first build a proof of concept (PoC). The team proceeded to exceed the brief and by September had built three distinct PoCs. One PoC garnered data on ~250 companies for the company's contacts database. Another fetched the financial filings on ~5,500 companies within minimum time. Yet another one monitored the web, tracking any updates to the latest filings.

Approach summay

Impressed with the PoCs, Sophia approved the deployment of these tools into her live operations. Avantika's team then proceeded to customise the PoCs for Sophia's specific requirements. They enabled the Gen AI engine to automatically source data from multiple sources and in multiple formats. They also developed GPT-based custom models to accurately identify, extract, and validate data — given a firm name — from its website sub-URLs and other sources such as news articles, contact bios, and financial websites. The validation included checks for duplicate contacts at a firm and database level, to maintain data integrity and to handle a single contact associated with multiple firms. Then studying the business rules of the organisation, the team orchestrated the platform's ability to process and publish the data onto the organisation's data servers, after the requisite quality audits. Finally, Avantika directed them to bake in auto-curation and enrichment modules as well as capabilities for automatic report generation.

During the development of the platform, Avantika's team had to overcome several challenges. For instance, the platform had to be integrated with all the existing tools being used by Sophia's teams, which could have led to delays in implementation. The team also made frequent enhancements to the GenAI code to improve extraction accuracy, incorporating multiple scenarios to fortify the data acquisition models.

Finally, Avantika and her team proceeded to deploy the platform in the production environment of Sophia's operations. During this initial phase of deployment, the team instituted multiple human-led validation checks to ensure that the data acquired and enriched using the platform met the quality standards expected by the organisation.

The benefits of artful intelligence

After successfully deploying Infosys' GenAI platform on the operations floor, Avantika's team worked hard to deliver sustained outcomes. Thus, for example, they steadily increased the contacts in the company's database from just 5,000 at the platform's launch to over 50,000 by January 2024. This was possible, as Avantika's team had designed the solution to be highly scalable, geared to manage high volumes per month. They also refined the Gen AI platform's abilities to ingest unstructured web content, delivering improved entity identification.

Key benefits

Sophia witnessed an astonishing 150% improvement in data scaling as the new AI platform steadily updated the contacts database, which grew by ~20 times from only 5000 entries to more than 90,000 within six months. Of these, over 4,000 were from ~390 firms with non-English websites. Also, with Avantika's team currently testing the optimal maintenance cycle, she looked forward to a targeted 35+% improvement in the database maintenance time-to-market cycle. What was more, Avantika had assured her that over the next six months, she could expect to see the database grow further to cross 150,000 entries.

In her quarterly review meeting with the company's top leadership, Sophia presented other impressive outcomes of the project. The average handling time of the team improved by 60%, reducing the earlier ~10-12 min taken for all their data sourcing and processing tasks to just around 4 minutes. Besides, the coverage of her contacts data had improved by 10 times with job title coverage increasing from just around 440 earlier to over 5,000 unique titles.

With the platform's useful automation features, Sophia no longer needed to worry about the high costs of scaling up her team. It not only automatically gathered data but directly fed it into the databases, had mechanisms for automated validation and compliance checking, also automatically building reports using metadata. Further, Avantika's team kept tweaking the AI models to reduce the platform's average handling times, and to improve its accuracy and quality scores.

The partnership with Infosys BPM had some other unexpected outcomes. Not long after her review meeting, the company rewarded Sophia and some of the other stakeholders overseeing the project with promotions and significant portfolio additions.

**Names have been altered to preserve the identities of the people involved.*

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next