Title: AI-Driven Financing Opportunity Prescription: Implementation and Validation of the FARO System

Lorena Rojo, Joana Paredes, Carmen Medrano, Fernando Ibañez

Correspondence: Fernando.ibanez@gestinver.es

Abstract

This communication presents the implementation and validation of FARO (Financing Automaton for Resource Optimization), an artificial intelligence-driven system for automating the prescription of financing opportunities. By integrating web scraping, natural language processing (NLP), and machine learning, FARO addresses the challenges of data heterogeneity, scalability, and contextual understanding in public fund allocation. Our results demonstrate high accuracy (75-89%) in extracting budgetary and temporal information across different administrative levels, while highlighting areas for improvement, particularly in sector identification (30.56% accuracy). This work contributes to the growing field of AI applications in public administration and offers insights into the development of intelligent systems for financial opportunity matching.

Introduction

The allocation of public funds through grants and subsidies is crucial for economic growth, innovation, and social development. However, the complexity and volume of financing opportunities, coupled with the diverse landscape of potential beneficiaries, pose significant challenges for both funding agencies and applicants. Traditional methods of identifying and matching financing opportunities with eligible companies are often time-consuming, inefficient, and prone to human error.

Recent advancements in artificial intelligence and related technologies have opened new avenues for addressing these challenges. For instance, Lagoze et al. [1] demonstrated the potential of machine learning algorithms in analyzing and classifying grant proposal texts. Similarly, Feldman and Sanger [2] explored the use of text mining techniques to analyze large volumes of textual data, which has implications for processing funding documents and company information.

The FARO project aims to address these challenges by developing an integrated system that leverages cutting-edge AI technologies. At its core, the project seeks to create a robust, scalable, and intelligent platform capable of automatically capturing, processing, and analyzing information about both financing opportunities and potential beneficiaries.

Methodology

Our approach integrates several key technologies:

- **Web crawling and scraping**: We implemented custom web scraping tools using KNIME (Konstanz Information Miner) for efficient and targeted extraction of information from diverse online sources [3].
- **Natural Language Processing (NLP)**: We developed an NLP pipeline using spaCy and custom rules for Named Entity Recognition (NER), fine-tuned on manually annotated financing documents. This pipeline was enhanced with advanced language models, specifically a combination of SentenceTransformers for semantic understanding and GPT-3 for context-aware information extraction [4].
- **Machine learning**: We employed XGBoost for company classification and opportunity matching, developing separate models for state, regional, and local level subsidies [5].
- **Robotic Process Automation (RPA)**: We implemented RPA workflows using UiPath to automate the capture and update of company information from various public sources [6].
- **Big Data technologies**: We utilized Hadoop for distributed storage and processing, and Apache Kafka for real-time data streaming, to handle large volumes of heterogeneous data [7].

We developed and tested the FARO system using a dataset of approximately 5,000 companies and their corresponding subsidy records. Key steps in our methodology included:

Implementing a custom version of Tesseract OCR, optimized for Spanish language documents, resulting in a 15% increase in text recognition accuracy.

Fine-tuning our NER model on a dataset of 20 manually annotated financing opportunity documents, with an additional 20 documents used for validation.

Employing SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in our dataset.

Developing a robust preprocessing pipeline to handle varied formats and structures of financing opportunity documents, including text normalization, sentence segmentation optimized for legal and financial documents, and custom tokenization rules.

Results

Our results demonstrate the system's effectiveness in extracting key information from subsidy documents:

Budgetary information:

Maximum budget per project: 75.00% accuracy

Total budget of calls: 88.89% - 91.67% accuracy

Application deadlines:

Start date: 80.56% accuracy

End date: 80.56% accuracy

The system's performance varied across different administrative levels:

State-level subsidies: Accuracy ranging from 75% to 89% across different sectors

Regional-level subsidies: Accuracy ranging from 80% to 91% across different sectors

Local-level subsidies: Accuracy ranging from 86% to 93% across different sectors

However, sector identification remains a significant challenge, with only 30.56% accuracy.

The impact of data preprocessing techniques was also evaluated:

Base model (no special preprocessing): 79% accuracy in both training and test sets

With SMOTE: 85% accuracy in the training set, 79% in the test set

Discussion

The FARO system shows promising results in automating the initial screening of financing opportunities, potentially reducing manual workload and improving accessibility for businesses. The high accuracy in extracting budgetary information and application deadlines (consistently above 75%) suggests that the system could be immediately valuable in processing subsidy documents.

The varying performance across administrative levels indicates that the system can adapt to different document structures and terminologies. However, it also suggests the need for further refinement to achieve consistent performance across all levels. This variability could be due to more standardized formats or consistent terminology used in regional and local subsidy documents compared to state-level documents.

The poor performance in sector identification (30.56% accuracy) represents a critical area for improvement. This low accuracy might be due to the complexity and variability in how sectors are described in subsidy documents, or it could indicate a limitation in our current NLP models in understanding and categorizing more abstract or varied textual descriptions.

The use of SMOTE for handling class imbalance showed promising results in improving training set accuracy. However, the consistent test set accuracy suggests that while SMOTE helps in addressing class imbalance, it may not necessarily improve the model's generalization to new data.

Conclusion and Future Work

The FARO project demonstrates the potential of AI in revolutionizing the prescription of financing opportunities. While challenges remain, particularly in sector identification and ensuring consistent performance across different contexts, the system's high accuracy in extracting key financial and temporal data marks a significant step forward.

Future work will focus on:

- Improving sector identification through more sophisticated NLP techniques or the incorporation of domain-specific knowledge bases.
- Expanding the dataset to include a wider range of subsidy types and business profiles to enhance the system's generalizability.

- Conducting real-world trials to validate the system's effectiveness in practical scenarios.
- Developing methods to enhance the explainability of the AI's decision-making process, thereby increasing transparency and trust.
- Investigating potential biases in the system's recommendations to ensure fairness in financing opportunity prescription.

This research lays the groundwork for more efficient and accurate matching of businesses with financing opportunities, promising to streamline the process of public fund allocation and potentially democratize access to financial support for businesses.

Acknowledgments

This research was funded by the European Union - NextGenerationEU, as part of the 2021 call for proposals for research and development projects in artificial intelligence and other digital technologies and their integration into value chains. Project Name: "Automatic System for Financing Opportunity Prescription Based on Artificial Intelligence. FARO" Project ID: 2021/C005/00146850

References

Lagoze, C., Velden, T., Lowe, B., Brady, H., & Derry, T. (2022). Enhancing grant proposal evaluation with machine learning-based citation analysis. Quantitative Science Studies, 3(1), 229-253.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

Zhao, Y. (2017). Web scraping. In Encyclopedia of Big Data (pp. 1-3). Springer, Cham.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

van der Aalst, W. M., Bichler, M., & Heinzl, A. (2018). Robotic process automation. Business & Information Systems Engineering, 60(4), 269-272.

Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. Journal of King Saud University-Computer and Information Sciences, 30(4), 431-448.