BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Analytical Tool for European Parliament Voting Contextualization

Bc. Petr Smažinka*

Abstract

Despite the abundance of EU open data, a semantic gap persists between raw availability and actionable intelligence. To bridge this, we develop a modular framework that transforms fragmented records into structured knowledge by integrating SPARQL, REST APIs, and LLM-based extraction into an automated data mining pipeline. The system facilitates multidimensional deep analysis by classifying legislative votes and quantifying metrics such as party cohesion, MEP loyalty, and alignment with macroeconomic indicators. This approach reveals statistically significant correlations between political behavior and socio-economic factors like GDP or energy composition. By increasing democratic transparency, this research provides a scalable methodology for the public and researchers to interpret legislative motivations through objective, multi-source data.

*xsmazi00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

European Parliament offers large datasets [1], however, the transparency is limited by the disconnect between legislative votes and the socio-economic conditions of member states. Raw datasets currently lack the context needed to identify the actual drivers of political decisions. This project addresses this by linking parliamentary behavior directly to national economic realities.

The challenge is to synchronize heterogeneous sources—APIs, SPARQL, and web data—into a unified relational model. The system must automate vote classification and correlate results with Eurostat [2] metrics while maintaining data integrity across electoral periods.

The closure of VoteWatch Europe [3] has left a fragmented analytical landscape. While HowTheyVote.eu [4] and Parltrack [5] provide excellent data infrastructure, they lack integrated socio-economic metrics. Contextual analysis currently requires manual data merging, as no ready-to-use tool exists for interactive correlation of MEP voting and Eurostat indicators.

The proposed platform utilizes a unified relational model to integrate SPARQL queries and web scraping [6] with Large Language Models (LLM). This architecture extracts political context from unstructured [7] sources like Wikipedia and applies semantic classification to legislative votes. By synthesizing these results with Eurostat socio-economic indicators, the tool enables multidimensional deep analysis of party cohesion and member loyalty through an interactive visualization layer.

The primary achievement is an automated pipeline that bridges raw legislation with socio-economic context. Key contributions include a unified schema for EP and Eurostat data and a functional analytical tool that identifies correlations between voting trends and national indicators like GDP or energy composition.

The core logic is implemented in Python, leveraging its robust data processing libraries to manage the pipeline.

2. Data Extraction and Loading

The ETL pipeline Fig. 2 is designed to automate the aggregation of structured and unstructured data from diverse sources, including the EP Open Data Portal, Eurostat, and Wikipedia. A significant challenge in this phase is correctly handling the great volume of data, particularly the roll-call votes, which exceed 12.6 million rows per term.

The system fetches basic lists of members, parties, and legislative terms from the EP portal. However, some metadata such as party colors are retrieved via SPARQL

from Wikidata, while information regarding national government memberships is extracted from Wikipedia. Since Wikipedia data is often unstructured, the pipeline utilizes a combination of HTML cleaning and LLM-based extraction to ensure accuracy. Furthermore, for each final vote, an LLM-based classification is performed to categorize the motion into predefined thematic areas. Socio-economic indicators from Eurostat are synchronized directly. For current-year data not yet officially released, the system computes estimates using regression [8] or weighted averages based on the indicator's nature.

To ensure efficient data ingestion, the system implements a dual-stream approach for voting records. While the pipeline can interface directly with the EP Open Data API, it also supports a quicker import from the HowTheyVote dataset for the last two parliamentary terms. This flexibility allows the system to bypass the latency of official APIs when bulk historical data is already available.

The tool supports two LLM providers: Gemini and Ollama (tested with models Gemma 3.2 [9] and Llama 3.2 [10]). This architecture allows users to choose between remote API-based extraction or local execution. Furthermore, the system provides the flexibility to select specific models, balancing computational cost with extraction precision.

3. Data Mining

The data mining module Fig. 3 computes behavioral metrics and socio-economic correlations directly from MySQL tables. Due to the massive data size, implementing reasonable database indexing was crucial. However, even with indexing, some computations can't be performed in real time and have to be precomputed. The results are exported into JSON files.

The system pre-computes several key metrics:

- **Agreement Index (AI) [11]:** Measures the cohesion within parties, countries, or inter-faction groups.
- **Loyalty and Participation [11]:** Tracks how often MEPs vote with their party plurality and their overall attendance.
- **Correlation Analysis:** Uses Pearson r [12] to find statistical relationships between country voting patterns and Eurostat indicators.

4. Visualisation

The final stage of the pipeline transforms precomputed JSON files into an interactive web interface Fig. 4. The front-end is implemented as a Single Page Application (SPA) using Apache, JavaScript, and PHP.

By rendering precomputed files rather than querying the database directly, the interface maintains acceptable response times when displaying complex behavioral patterns. As demonstrated in the results Fig. 5–8, the application visualizes agreement matrices, loyalty charts, and party cohesion. This approach allows users to observe correlations between specific voting categories and socio-economic indicators, such as a member state's energy mix, without the latency associated with real-time large-scale data processing.

5. Conclusions

The developed tool provides a functional framework for bridging the gap between raw legislative records and political insights. While the application is stable, its interpretability is inherently limited by the quality of the input data.

Despite automated cleaning efforts, source datasets contain persistent inconsistencies. For instance, approximately 4 out of 2,600 MEP records remained malformed or incomplete in the original portals. These discrepancies, combined with the fact that data is aggregated from multiple independent sources, mean that results may contain minor errors, though they do not significantly distort overall statistical trends. Future work will focus on more extensive cleaning of the source datasets, restoration of missing information to further improve data integrity and the automated extraction of association rules to identify complex co-voting patterns between political factions.

Acknowledgements

I would like to thank my supervisor, Ing. Jiří Hynek, Ph.D., for his mentorship and guidance throughout this project. I especially appreciate his personal approach and willingness to help at any time; his insightful feedback and valuable advice often saved me a significant amount of time during the development and research process.

References

[1] European Parliament. European parliament open data portal, 2026.

[2] Eurostat. About eurostat - who we are, 2024.

[3] VoteWatch Europe. Votewatch europe: Tracking european parliament votes and political influence, 2022. [Platform ceased operations in 2022; accessed via archival records].

[4] Brenden Vantyghem. Howtheyvote: European parliament voting records, 2024.

[5] Stefan Marsiske. Parltrack: European parliament open data tracking, 2026.

[6] Ritu Banerjee. Website scraping. *Happiest Minds Technologies*, (1), 2014.

[7] Suyash Mishra and Anuranjan Misra. Structured and unstructured big data analytics. In *Proceedings of the 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*, pages 740–746, Mysuru, India, 2017. IEEE.

[8] Alan O. Sykes. An introduction to regression analysis. Working Paper 20, Coase-Sandor Institute for Law & Economics, 1993.

[9] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al. Gemma 3 technical report, 2025.

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024.

[11] Simon Hix, Abdul G Noury, and Gérard Roland. *Democratic politics in the European Parliament*. Cambridge University Press, Cambridge, 1 edition, 2007.

[12] J. Cohen, P. Cohen, S.G. West, and L.S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Taylor & Francis, 2013.