Report 1
Lynx: System for Financial News Analysis Using NLP and Graph Databases

Andrei Zhdanov Mikhail Sofin

September 16, 2025

A. Team

Team Akira:
Andrei Zhdanov (an.zhdanov@innopolis.university)
Mikhail Sofin (m.sofin@innopolis.university)

B. Project idea

Project Title: Lynx: System for Financial News Analysis Using NLP and Graph Databases.

In the context of an intensive information flow, financial market participants—such as traders, analysts, and risk managers—require timely identification and analysis of corporate events (mergers and acquisitions, executive appointments, investment rounds). Manual monitoring of news sources is labor-intensive, slow, and prone to human error. Existing commercial solutions (e.g., Bloomberg Terminal) are costly and often inaccessible to small and medium-sized businesses. This project proposes the development of an accessible prototype system for automatic knowledge extraction from texts, leveraging modern Natural Language Processing (NLP) techniques and graph technologies (Neo4j).

To develop a prototype software platform that automatically extracts structured information about companies and their relationships from news articles, represents it as an interactive graph, and provides users with tools for semantic search and analysis.

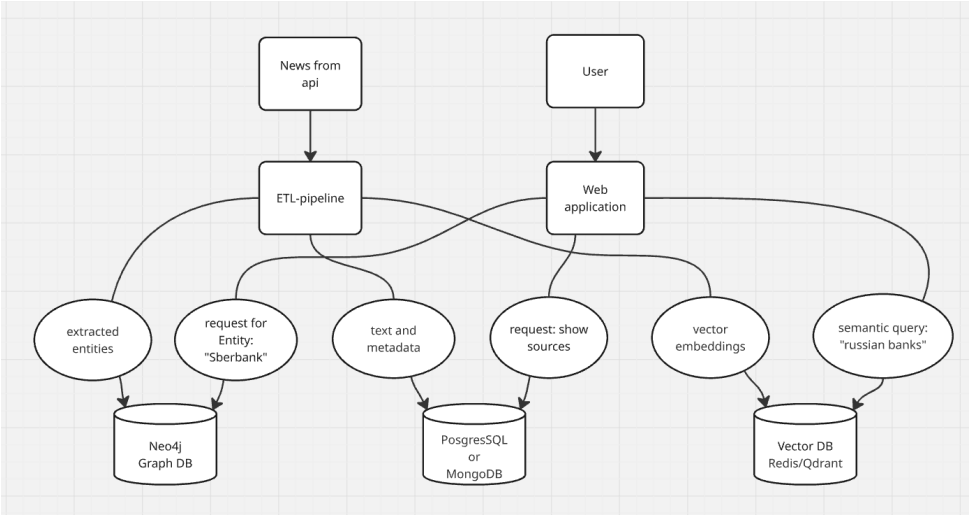

Fig. 1. App flow and user stories

C. Links

Link to Github repository: Lynx
Link to Miro: Miro
Link to Design docs: Design Document

D. Research

Recent research on Knowledge Graph (KG) extraction highlights both general and domain-specific challenges. Core issues include linguistic ambiguity (e.g., synonymy, polysemy), scalable relation extraction, data quality assurance, and integration with existing ontologies [1].

Advances in NLP have shifted the field from feature-based methods to deep learning approaches. In particular, transformer architectures form the basis for modern Named Entity Recognition (NER) and Relation Extraction (RE) tasks, as surveyed in [2].

Nevertheless, general-purpose models often underperform in specialized domains. For instance, [3] introduced FinRED, a financial relation extraction dataset, and demonstrated that even strong models such as BERT and RoBERTa struggle with domain-specific terminology and relations. This underscores the need for domain adaptation through fine-tuning and specialized datasets.

Recent surveys provide comprehensive insights into the evolution and current state of KG extraction techniques:

On the Evolution of Knowledge Graphs: A Survey and Perspective: This survey offers a comprehensive overview of various types of knowledge graphs (static, dynamic, temporal, and event-based) and techniques for knowledge extraction and reasoning. It also discusses the integration of large language models (LLMs) with KGs, highlighting future directions in knowledge engineering. [4]

A Comprehensive Survey on Relation Extraction: This survey delves into the latest advancements in Relation Extraction (RE), focusing on models that leverage language models. It analyzes 137 papers presented at the Association for Computational Linguistics (ACL) conferences over the past four years, providing insights into the evolution and current state of RE techniques. [5]

In summary, effective KG extraction requires:

1. Leveraging transformer-based architectures.
2. Addressing domain-shift challenges with specialized datasets.
3. Ensuring scalability and data quality in extracted knowledge.

E. Data

For training and evaluating Named Entity Recognition (NER) and Relation Extraction (RE) models, the following data sources are planned to be used.

Primary Dataset: English FinRED: At the initial stage, the publicly available FinRED dataset will be used as the primary dataset for experiments and pipeline debugging.

Future Perspective: Creation of a Russian-language Dataset: Considering the target domain and potential system users, a key future step is the creation of a Russian-language financial dataset.

1. **Collection of raw texts**: Automatic parsing of news feeds from Russian financial media (such as RBC, Kommersant, Vedomosti), as well as official company press releases and regulator communications (Bank of Russia).
2. **Development of annotation guidelines**: Adapting the FinRED annotation scheme to the specifics of the Russian financial market and legal terminology. Work will be conducted to unify entity types (`ORGANIZATION`, `PERSON`, `AMOUNT`, `ASSET`) and relation types (`INVESTS_IN`, `ACQUIRES`, `IS_SUBSIDIARY_OF`).
3. **Annotation process**: The texts will be manually annotated by the project team.

F. Metrics

The key metrics for assessing the quality of the Named Entity Recognition (NER) and Relationship Extraction (RE) models were chosen to be classical metrics for sequence classification and segmentation problems: Precision, Recall and F1-score, which is their harmonic mean. The target value for the final system is to achieve $F_1 \geq 0.80$ for both tasks (NER and RE). This threshold is chosen based on the following considerations: The specifics of the domain. Working in a narrow financial subject area is more challenging than in general domains. The model must learn to recognize specific entities (names of companies, funds, financial instruments) and relationships (acquisitions, investments, appointments), which are often expressed in complex and variable terminology. Research shows that even modern transformer models show slightly lower results in such specialized areas compared to general news buildings.

G. Architecture

We have defined our preliminary architecture for the non-ML part and the ML part.

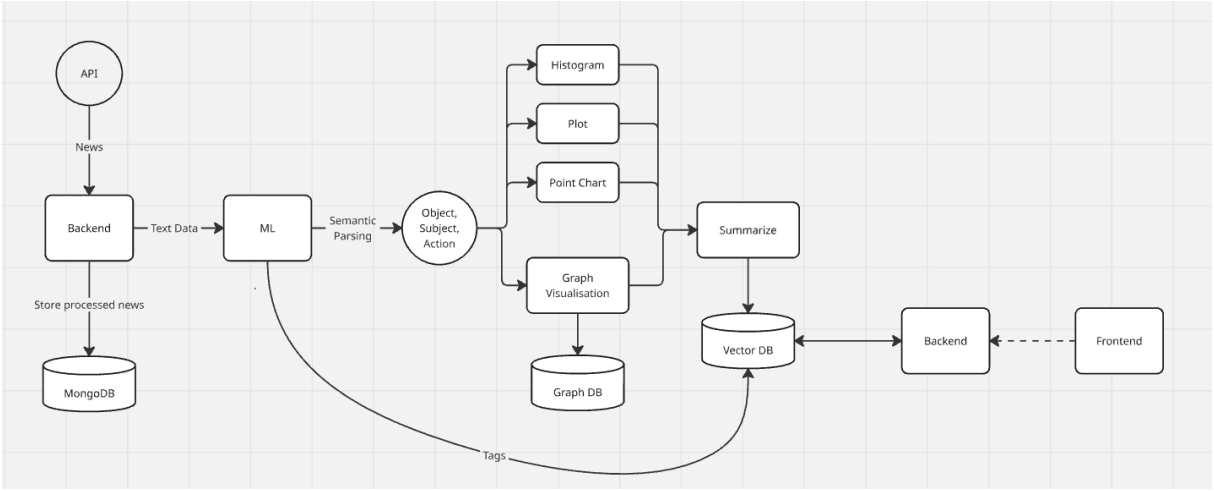

Fig. 2. Non-ML Architecture

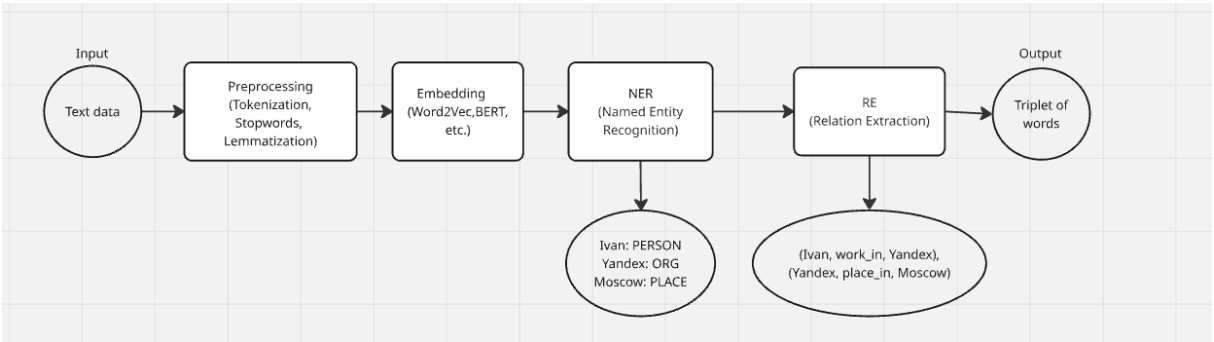

Fig. 3. ML Architecture

H. Work Distribution

The roles for our duo:
- Andrei Zhdanov: ML, data preparation, research, product manager.

- Mikhail Sofin: Fullstack (Backend, frontend, devops).

References

[1] D. Kaustubh. "Knowledge Graph Extraction and Challenges." [Accessed: 16.09.2025]. [Online]. Available: https://neo4j.com/blog/developer/knowledge-graph-extraction-challenges.

[2] N. Zhang et al., "Deepke: A deep learning based knowledge extraction toolkit for knowledge base population," *CoRR*, vol. abs/2201.03335, 2022. arXiv: 2201.03335. [Online]. Available: https://arxiv.org/abs/2201.03335.

[3] S. Sharma et al., *Finred: A dataset for relation extraction in financial domain*, 2023. arXiv: 2306.03736 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2306.03736.

[4] Y. Shi, X. Li, and H. Zhang, "On the evolution of knowledge graphs: A survey and perspective," *arXiv preprint arXiv:2310.04835*, 2023. [Online]. Available: https://arxiv.org/abs/2310.04835.

[5] Y. Zhang, L. Wang, and J. Liu, "A comprehensive survey on relation extraction," *arXiv preprint arXiv:2411.18157*, 2024. [Online]. Available: https://arxiv.org/abs/2411.18157.