Report 1.2
Lynx: System for Financial News Analysis Using NLP and Graph Databases

Andrei Zhdanov Mikhail Sofin

October 7, 2025

A. Team

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

B. Project topic

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

C. Links

Link to Github repository with all source code and notebooks: Lynx
Link to Miro: Miro

D. What has been done so far

Over the past three weeks, our work has been focused primarily on two main areas: developing part of the backend infrastructure and conducting experiments in the machine learning component of the project. In the architecture diagram below, the parts we concentrated on are highlighted in red to clearly indicate our scope of work during this period.

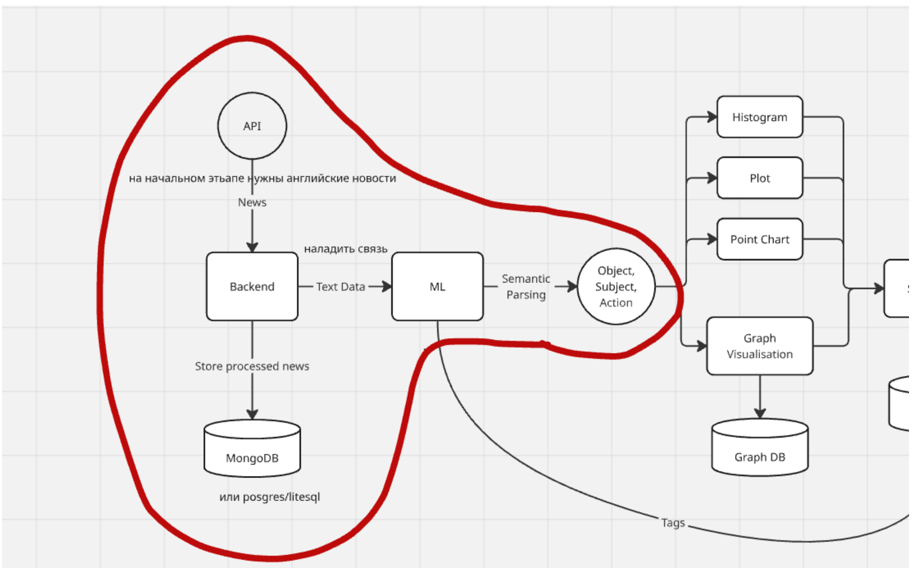
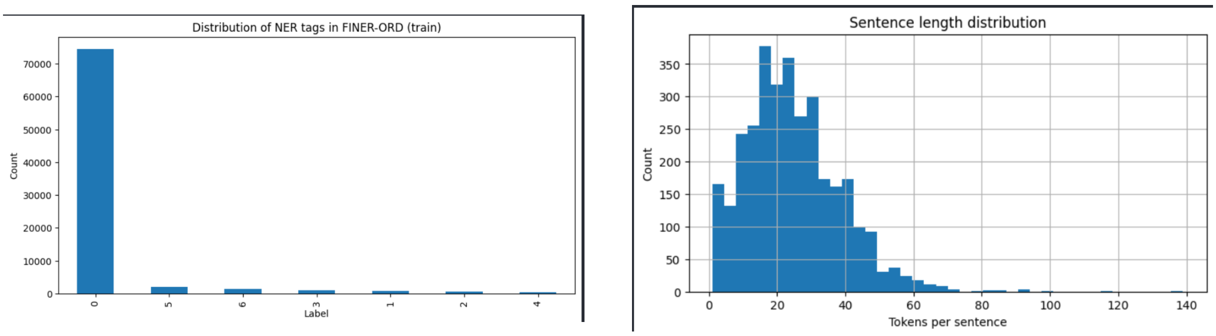


Fig. 1. Architecture

- **Backend Development**: We initiated the development of the backend system. This included the initialization of a MongoDB database for storing raw news articles and a Qdrant vector database for storing news text embeddings to enable semantic search capabilities.

Fig. 2. backend

- **Exploratory Data Analysis (EDA)**: We conducted thorough exploratory data analysis on two datasets:
  - **FINER-ORD** for Named Entity Recognition (NER) tasks.

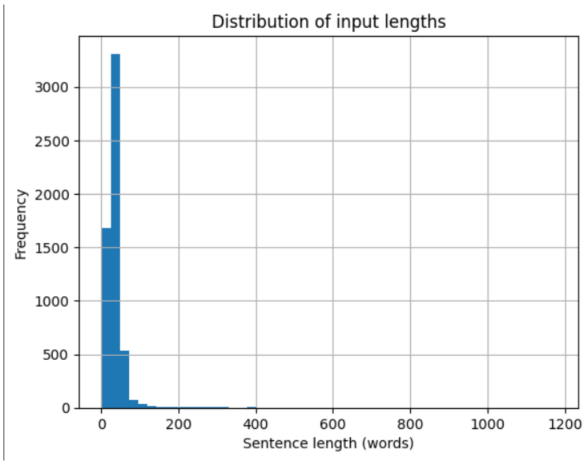


(a) Distribution of NER tags

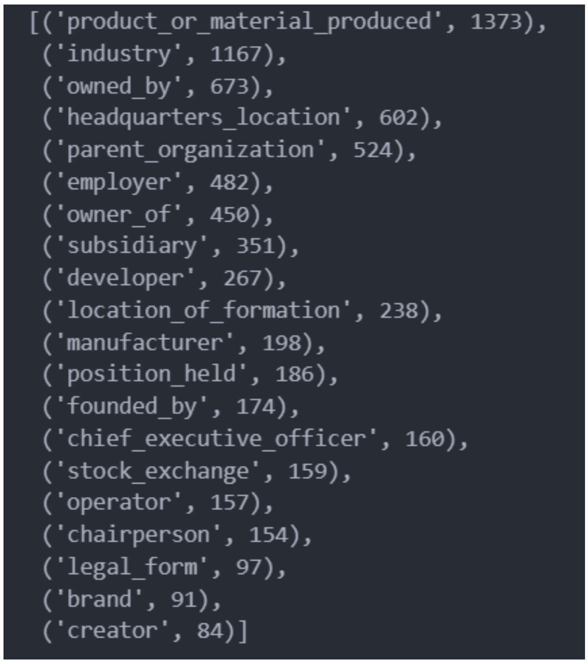
(b) Sentence length statistics

Fig. 3. Statistical analysis of the FINER-ORD dataset

  - **FINRED** for Relation Extraction (RE) tasks.



(a) Sentence length statistics

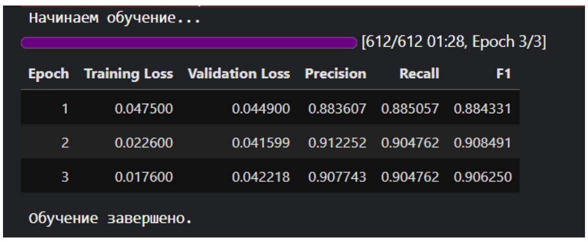


(b) Classes distribution

Fig. 4. Statistical analysis of the FINRED dataset

The EDA included statistical analysis of data distribution, class imbalance checking, and inspection of entity and relation labels to ensure data quality and consistency.

- **NER Model Fine-tuning**: For the Named Entity Recognition (NER) task, we selected the pre-trained model `dslim/bert-base-NER` as the foundation for fine-tuning. This model is based on the BERT architecture and has been pre-trained specifically for NER tasks using the CoNLL-2003 dataset, which includes common entity types such as organizations, locations, and persons. By fine-tuning `dslim/bert-base-NER` on the FINER-ORD dataset, we were able to effectively adapt the model to the financial domain, achieving an accuracy of 90% on the validation set, demonstrating strong performance in identifying named entities in text.


Начинаем обучение...
[612/612 01:28, Epoch 3/3]

Epoch	Training Loss	Validation Loss	Precision	Recall	F1
1	0.047500	0.044900	0.883607	0.885057	0.884331
2	0.022600	0.041599	0.912252	0.904762	0.908491
3	0.017600	0.042218	0.907743	0.904762	0.906250

Обучение завершено.

Fig. 5. Training NER model

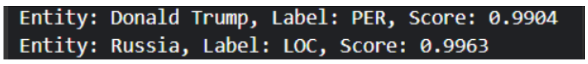


Fig. 6. Example of NER

- **RE Model Fine-tuning**: For the Relation Extraction (RE) task, we fine-tuned the `microsoft/deberta-v3-small` model. This model belongs to the DeBERTa (Decoding-enhanced BERT with Disentangled Attention) family, which improves upon BERT and RoBERTa by separating content and positional embeddings and using an enhanced mask decoder mechanism. We selected this model because it achieves strong performance across various NLP benchmarks while maintaining relatively low computational cost due to its smaller size compared to larger DeBERTa variants. Its strong contextual understanding makes it particularly effective for identifying semantic relationships between entities in complex financial texts. After fine-tuning on the FINRED dataset, the model achieved an accuracy of 96%, demonstrating its ability to accurately capture relational structures in financial news.

[2883/2883 1:01:59, Epoch 3/3]

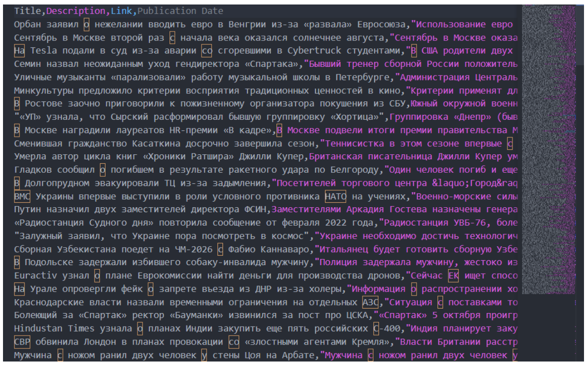
Epoch	Training Loss	Validation Loss	Precision	Recall	F1
1	0.393400	0.395047	0.885242	0.883748	0.879814
2	0.278800	0.255683	0.932399	0.932357	0.931658
3	0.215700	0.228898	0.943726	0.942020	0.941618

Fig. 7. Training

- **Dataset Construction for Russian News**: We have begun collecting news articles in Russian and preparing a dataset with annotations using large language models (LLMs). This dataset will be used to train and evaluate models for NER and RE in Russian-language news, enabling multilingual capabilities for our system.

```
our_feeds = {
    'Kommersant Economy': 'https://www.kommersant.ru/RSS/economy.xml',
    'Vedomosti Politics': 'https://www.vedomosti.ru/rss/politics',
    'Vedomosti Finance': 'https://www.vedomosti.ru/rss/finance',
    'RBC News': 'https://rssexport.rbc.ru/rbcnews/news/30/full.rss',
    'Interfax': 'https://www.interfax.ru/rss.asp',
    'RIA Novosti': 'https://ria.ru/export/rss2/index.xml',
    'Lenta Politics': 'https://lenta.ru/rss/news/politics/',
    'TASS': 'https://tass.ru/rss/v2.xml',
    'Rossiyskaya Gazeta': 'https://rg.ru/xml/index.xml',
    'Gazeta Ru': 'https://www.gazeta.ru/export/rss/first.xml'
}
```

(a) News RSS collection pipeline



(b) Example of collected news item

Fig. 8. Dataset creation process: collecting and preparing Russian financial news

E. Work Distribution

- Andrei Zhdanov: Data EDA, fine-tune models, ml architecture
- Mikhail Sofin: backend architecture, data collection

F. Plan for the Next Weeks

In the next stages of the project, we plan to focus on completing the integration of the backend and machine learning pipelines into a unified system. This includes connecting the data ingestion, processing, and model inference components.

Additionally, we aim to start developing the frontend part of the application and to experiment with graph-based visualizations of entities and relations using Neo4j.

Another important direction will be the continuation of dataset development — expanding the corpus with new annotated examples and improving data quality. We also intend to conduct further experiments with machine learning models, including the possibility of designing and training our own custom model for financial text understanding.