Final Technical Report
Lynx: System for Financial News Analysis Using NLP and Graphs

Andrei Zhdanov Mikhail Sofin

November 25, 2025

I. Team

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

II. Project topic

Lynx: System for Financial News Analysis Using NLP and Graphs. In today's financial world, news comes from thousands of sources every second. Traders, analysts, and investors face the problem of information overload, how can they quickly isolate the essence of the news stream and understand the connections between events?

We offer a solution: visualization of financial news in the form of knowledge graphs. Instead of reading long texts, the user sees a visual diagram where:

- **Nodes** — companies, persons, assets
- **Connections** — financial relations between them
- **Structure** — shows who influences whom

For example, instead of an article about a merger, you immediately see: *"Company A → acquired → Company B for $X"*

This approach allows you to:

- Instantly understand key events
- See hidden connections between organizations
- Analyze network effects in the economy
- To make more informed investment decisions

In this project, we present a system that automatically converts text news into knowledge graphs using natural language processing and machine learning methods.

III. Links

Github repository: Lynx
Miro board(architecture planning): Miro
Design documentation: design_doc
Datasets: datasets
Models: models
Dataset Collecting Scripts: collection
Notebook with Modelling Experiments: notebooks

IV. Methods

A. Solved Problems

We addressed two primary tasks: Named Entity Recognition (NER) and Relation Extraction (RE). The challenge was twofold: first, to adapt these tasks to the specialized domain of finance, and second, to develop a solution for the Russian language, as the project was based on a corpus of Russian financial news.

Key Challenges and Objectives:

- Domain Adaptation: Moving beyond general-purpose models to accurately identify financial entities (e.g., ORG, PERSON, MONEY, SECURITY) and relations (e.g., CEO_OF, ACQUIRED_BY) specific to the financial sector.
- Language Specificity: Building a robust pipeline for Russian, which involves handling its complex morphology, case system, and lack of capitalization cues compared to English.
- Data-Driven Approach: The solution was designed to be trained on and evaluated against a dedicated dataset of Russian financial news articles.

B. Approaches Taken

Our initial approach involved tackling the problem in English to establish a baseline using existing resources. We utilized two specialized financial datasets:

- FinRED for Named Entity Recognition (NER).
- Finer for Relation Extraction (RE).

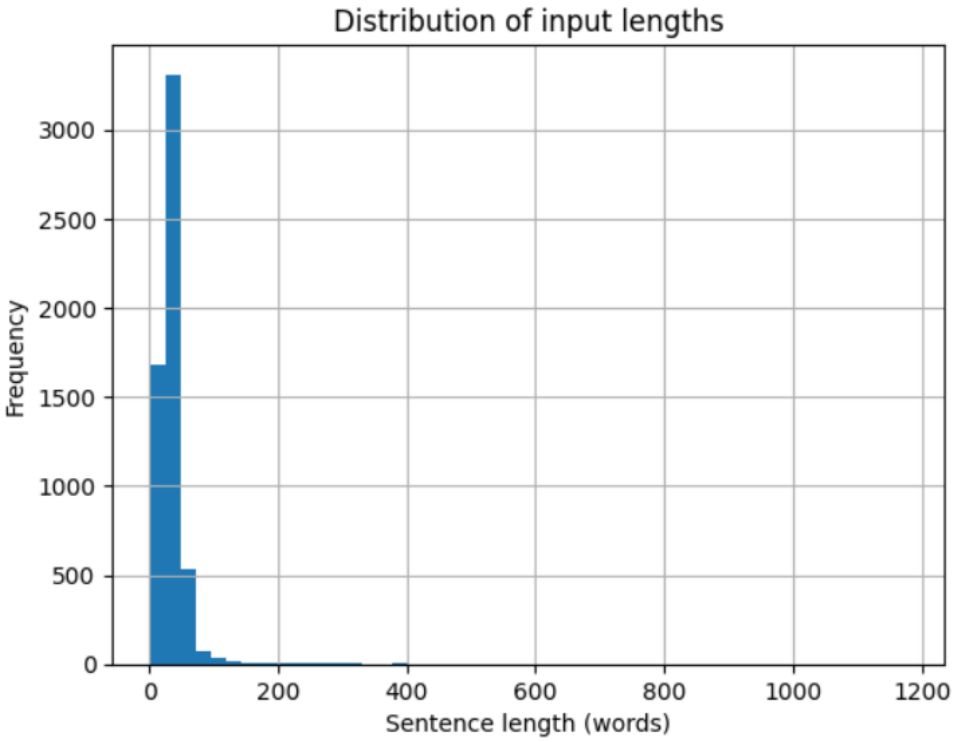


Fig. 1. NER dataset sentence lengths

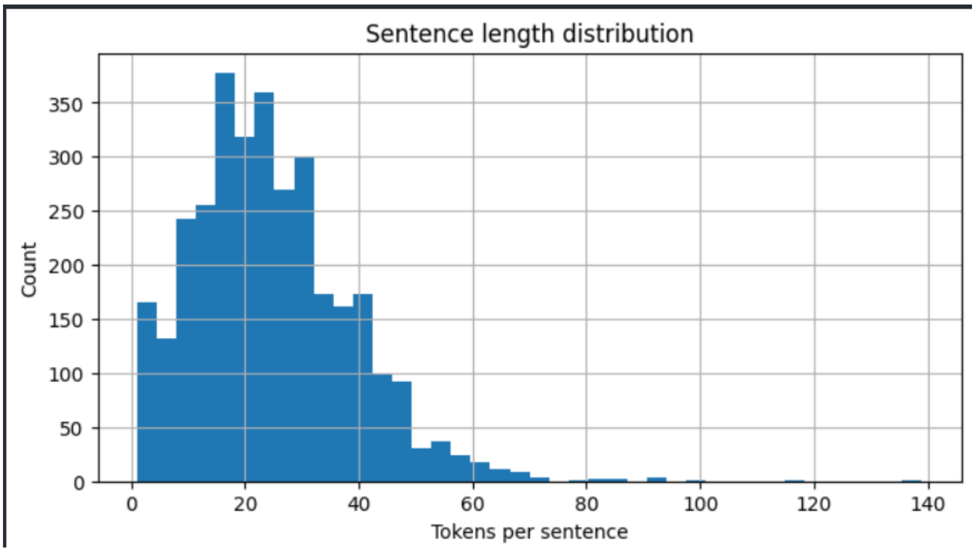
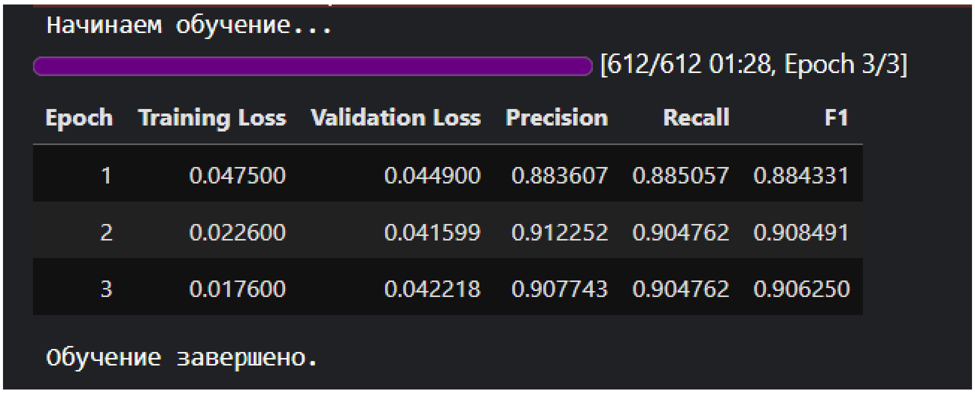


Fig. 2. RE dataset sentence lengths

We fine-tuned two separate BERT-based models on these respective datasets. While each model individually achieved high accuracy on its specific task, we encountered significant challenges when combining them into a single pipeline. The primary issue was a terminology mismatch between the datasets—they used slightly different definitions and labels for financial entities, which caused errors and inconsistencies when the RE model processed the output of the NER model.



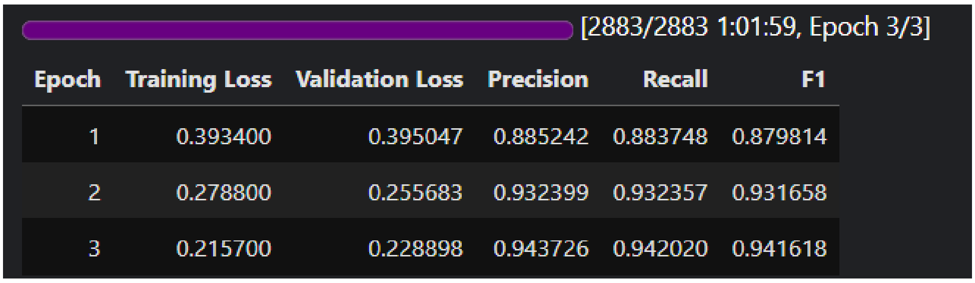
Начинаем обучение...

[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. 3. NER model accuracy



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

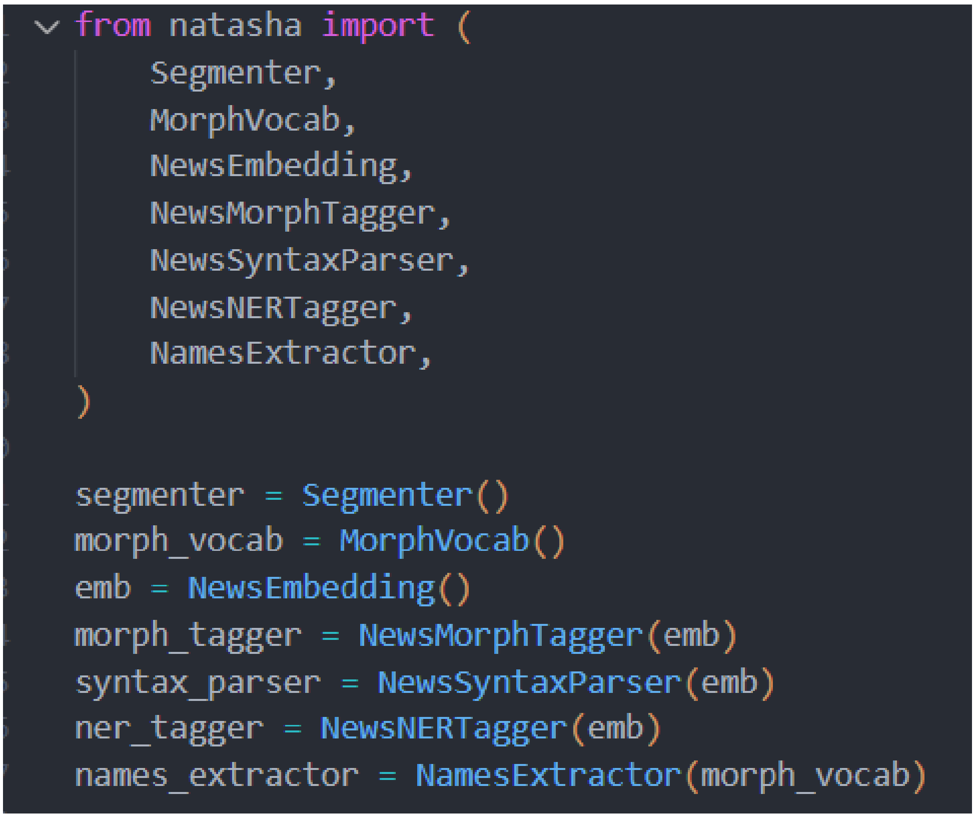
Epoch	Training Loss	Validation Loss	Precision	Recall	F1
1	0.393400	0.395047	0.885242	0.883748	0.879814
2	0.278800	0.255683	0.932399	0.932357	0.931658
3	0.215700	0.228898	0.943726	0.942020	0.941618

Fig. 4. RE model accuracy

Following the challenges with the English pipeline, we pivoted to our primary objective: building a solution for Russian financial texts. Our initial research revealed a significant obstacle—the absence of large, high-quality, and publicly available datasets for Russian NER and RE in the financial domain.

Faced with this data scarcity, we shifted our strategy from supervised training on existing datasets to leveraging and adapting pre-existing models for the Russian language. This led us to explore the ecosystem of Russian NLP libraries, where we identified Natasha as the most promising candidate for our base NER component.

Natasha is a robust and widely-used open-source project specifically designed for Russian natural language processing. Its strengths are particularly evident in Named Entity Recognition.



```python
from natasha import (
    Segmenter,
    MorphVocab,
    NewsEmbedding,
    NewsMorphTagger,
    NewsSyntaxParser,
    NewsNERTagger,
    NamesExtractor,
)

segmenter = Segmenter()
morph_vocab = MorphVocab()
emb = NewsEmbedding()
morph_tagger = NewsMorphTagger(emb)
syntax_parser = NewsSyntaxParser(emb)
ner_tagger = NewsNERTagger(emb)
names_extractor = NamesExtractor(morph_vocab)
```

Fig. 5. Natasha library

Key features include:

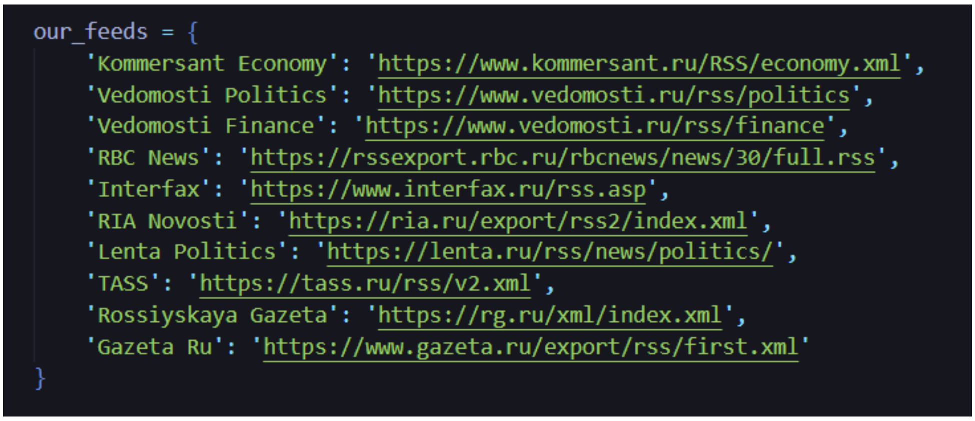
- Architecture: It is built upon BERT-like architectures that have been pre-trained on large Russian corpora, giving it a deep understanding of the language's grammar and syntax.

- Entity Types: Out-of-the-box, Natasha is optimized for recognizing common entity types such as Names (PER), Locations (LOC), and Organizations (ORG).

- Performance: It achieves state-of-the-art accuracy on standard Russian NER benchmarks for these general domains.

- It was train on finance domain

For the task of Relation Extraction (RE), we explored two distinct methodologies to overcome the lack of pre-existing datasets for Russian financial relations.

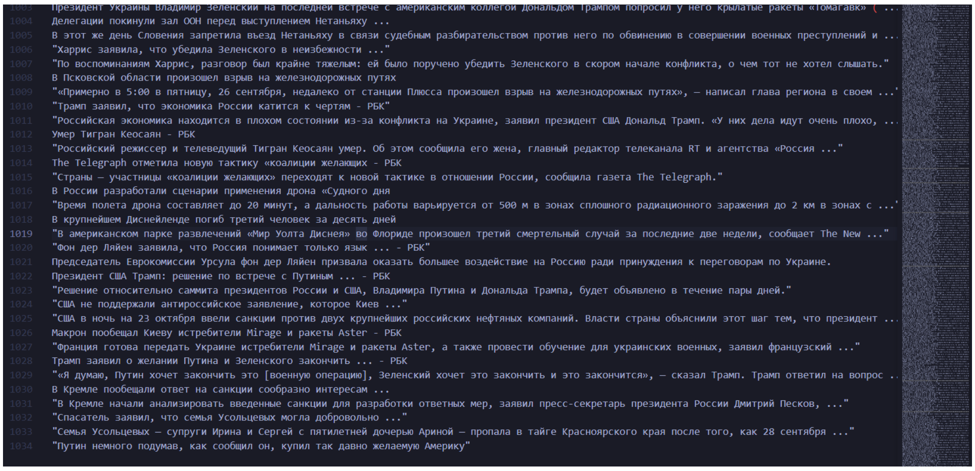


```
our_feeds = {
    'Kommersant Economy': 'https://www.kommersant.ru/RSS/economy.xml',
    'Vedomosti Politics': 'https://www.vedomosti.ru/rss/politics',
    'Vedomosti Finance': 'https://www.vedomosti.ru/rss/finance',
    'RBC News': 'https://rssexport.rbc.ru/rbcnews/news/30/full.rss',
    'Interfax': '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'
}
```

Fig. 6. Feeds

Our initial approach involved a rule-based system. This method leveraged the Natasha model as a robust entity extractor to identify the core subjects and objects in a text. We then developed a set of handmade, domain-specific rules designed to detect financial actions and relationships between these entities based on lexical and syntactic patterns. While this approach allowed for rapid prototyping and provided high precision on pattern-matched sentences, it ultimately proved to be non-scalable, suffering from low recall and an inability to generalize to complex or unforeseen linguistic structures.

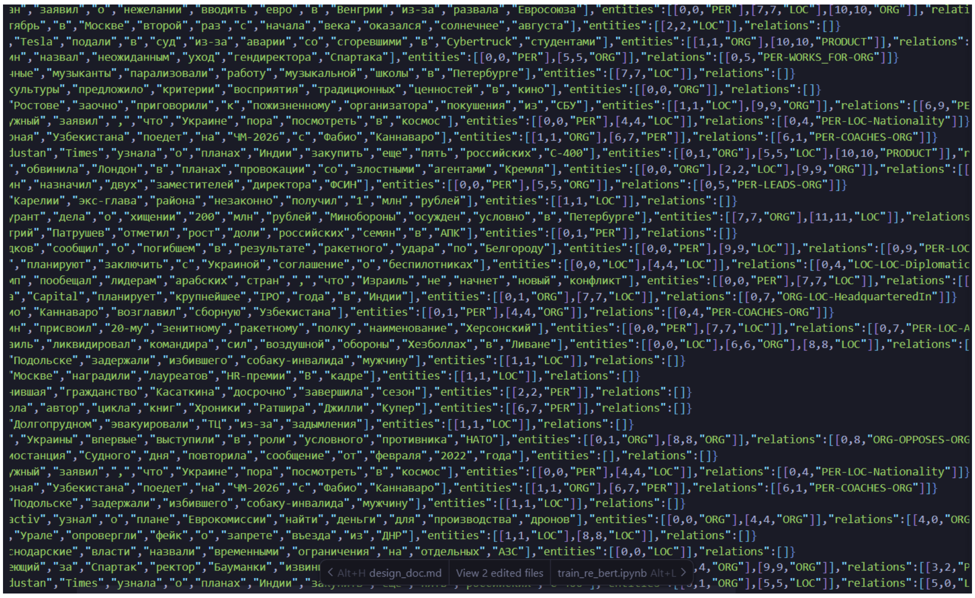
To develop a more robust and scalable solution, we pivoted to a data-driven, supervised learning approach. This initiative began with a large-scale data collection effort, gathering thousands of raw Russian news articles from various API sources.



```
Президент Украины Владимир Зеленский на последней встрече с американским коллегой Дональдом Трампом попросил у него крылатые ракеты «Томагавк» ( ...
Делегации покинули зал ООН перед выступлением Нетаньяху ...
В этот же день Словения запретила въезд Нетаньяху в связи судебным разбирательством против него по обвинению в совершении военных преступлений и ...
"Харрис заявила, что убедила Зеленского в неизбежности ..."
"По воспоминаниям Харрис, разговор был крайне тяжелым: ей было поручено убедить Зеленского в скором начале конфликта, о чем тот не хотел слышать."
В Псковской области произошел взрыв на железнодорожных путях
"«Примерно в 5:00 в пятницу, 26 сентября, недалеко от станции Плюсса произошел взрыв на железнодорожных путях», — написал глава региона в своем ...'
"Трамп заявил, что экономика России катится к чертям - РБК"
"Российская экономика находится в плохом состоянии из-за конфликта на Украине, заявил президент США Дональд Трамп. «У них дела идут очень плохо, ...
Умер Тигран Кеосаян - РБК
"Российский режиссер и телеведущий Тигран Кеосаян умер. Об этом сообщила его жена, главный редактор телеканала RT и агентства «Россия ..."
The Telegraph отметила новую тактику «коалиции желающих - РБК
"Страны — участницы «коалиции желающих» переходят к новой тактике в отношении России, сообщила газета The Telegraph."
В России разработали сценарии применения дрона «Судного дня
"Время полета дрона составляет до 20 минут, а дальность работы варьируется от 500 м в зонах сплошного радиационного заражения до 2 км в зонах с ...'
В крупнейшем Диснейленде погиб третий человек за десять дней
"В американском парке развлечений «Мир Уолта Диснея» во Флориде произошел третий смертельный случай за последние две недели, сообщает The New ..."
"Фон дер Ляйен заявила, что Россия понимает только язык ... - РБК"
Председатель Еврокомиссии Урсула фон дер Ляйен призвала оказать большее воздействие на Россию ради принуждения к переговорам по Украине.
Президент США Трамп: решение по встрече с Путиным ... - РБК
"Решение относительно саммита президентов России и США, Владимира Путина и Дональда Трампа, будет объявлено в течение пары дней."
"США не поддержали антироссийское заявление, которое Киев ..."
"США в ночь на 23 октября ввели санкции против двух крупнейших российских нефтяных компаний. Власти страны объяснили этот шаг тем, что президент ...
Макрон пообещал Киеву истребители Mirage и ракеты Aster - РБК
"Франция готова передать Украине истребители Mirage и ракеты Aster, а также провести обучение для украинских военных, заявил французский ..."
Трамп заявил о желании Путина и Зеленского закончить ... - РБК
"«Я думаю, Путин хочет закончить это [военную операцию], Зеленский хочет это закончить и это закончится», — сказал Трамп. Трамп ответил на вопрос ..
В Кремле пообещали ответ на санкции сообразно интересам ...
"В Кремле начали анализировать введенные санкции для разработки ответных мер, заявил пресс-секретарь президента России Дмитрий Песков, ..."
"Спасатель заявил, что семья Усольцевых могла добровольно ..."
"Семья Усольцевых — супруги Ирина и Сергей с пятилетней дочерью Ариной — пропала в тайге Красноярского края после того, как 28 сентября ..."
"Путин немного подумав, как сообщил он, купил так давно желаемую Америку"
```

Fig. 7. news collections

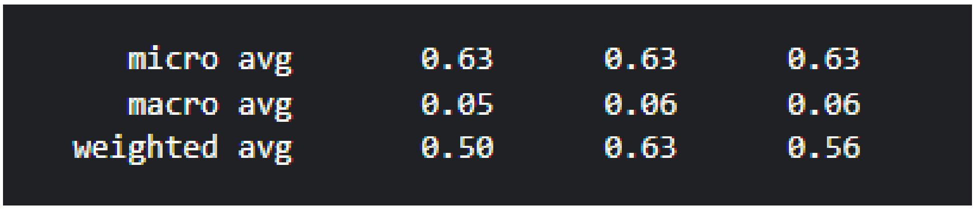
The subsequent and most critical step was the creation of a labeled RE dataset. Following TA's advice, we utilized a Large Language Model to assist in the annotation process, generating potential relation triplets (subject, relation, object) from the collected news text. This methodology allowed us to efficiently create the specialized dataset, which was necessary for training.



```
эн","заявил","о","нежелании","вводить","евро","в","Венгрии","из-за","развала","Евросоюза"],"entities":[[0,0,"PER"],[7,7,"LOC"],[10,10,"ORG"]],"relati
гябрь","в","Москве","второй","раз","с","начала","века","оказался","солнечнее","августа"],"entities":[[2,2,"LOC"]],"relations":[]}
,"Tesla","подали","в","суд","из-за","аварии","со","сгоревшими","в","Cybertruck","студентами"],"entities":[[1,1,"ORG"],[10,10,"PRODUCT"]],"relations":
ин","назвал","неожиданным","уход","гендиректора","Спартака"],"entities":[[0,0,"PER"],[5,5,"ORG"]],"relations":[[0,5,"PER-WORKS_FOR-ORG"]]}
чные","музыканты","парализовали","работу","музыкальной","школы","в","Петербурге"],"entities":[[7,7,"LOC"]],"relations":[]}
культуры","предложило","критерии","восприятия","традиционных","ценностей","в","кино"],"entities":[[0,0,"ORG"]],"relations":[]}
"Ростове","заочно","приговорили","к","пожизненному","организатора","покушения","из","СБУ"],"entities":[[1,1,"LOC"],[9,9,"ORG"]],"relations":[[6,9,"PE
ужный","заявил",",","что","Украине","пора","посмотреть","в","космос"],"entities":[[0,0,"PER"],[4,4,"LOC"]],"relations":[[0,4,"PER-LOC-Nationality"]]}
эная","Узбекистана","поедет","на","ЧМ-2026","с","Фабио","Каннаваро"],"entities":[[1,1,"ORG"],[6,7,"PER"]],"relations":[[6,1,"PER-COACHES-ORG"]]}
dustan","Times","узнала","о","планах","Индии","закупить","еще","пять","российских","С-400"],"entities":[[0,1,"ORG"],[5,5,"LOC"],[10,10,"PRODUCT"]],"r
","обвинила","Лондон","в","планах","провокации","со","злостными","агентами","Кремля"],"entities":[[0,0,"ORG"],[2,2,"LOC"],[9,9,"ORG"]],"relations":[[
ин","назначил","двух","заместителей","директора","ФСИН"],"entities":[[0,0,"PER"],[5,5,"ORG"]],"relations":[[0,5,"PER-LEADS-ORG"]]}
"Карелии","экс-глава","района","незаконно","получил","1","млн","рублей"],"entities":[[1,1,"LOC"]],"relations":[]}
урант","дела","о","хищении","200","млн","рублей","Минобороны","осужден","условно","в","Петербурге"],"entities":[[7,7,"ORG"],[11,11,"LOC"]],"relations
грий","Патрушев","отметил","рост","доли","российских","семян","в","АПК"],"entities":[[0,1,"PER"]],"relations":[]}
дков","сообщил","о","погибшем","в","результате","ракетного","удара","по","Белгороду"],"entities":[[0,0,"PER"],[9,9,"LOC"]],"relations":[[0,9,"PER-LOC
","планируют","заключить","с","Украиной","соглашение","о","беспилотниках"],"entities":[[0,0,"LOC"],[4,4,"LOC"]],"relations":[[0,4,"LOC-LOC-Diplomatic
мп","пообещал","лидерам","арабских","стран",",","что","Израиль","не","начнет","новый","конфликт"],"entities":[[0,0,"PER"],[7,7,"LOC"]],"relations":[[
a","Capital","планирует","крупнейшее","IPO","года","в","Индии"],"entities":[[0,1,"ORG"],[7,7,"LOC"]],"relations":[[0,7,"ORG-LOC-HeadquarteredIn"]]}
мо","Каннаваро","возглавил","сборную","Узбекистана"],"entities":[[0,1,"PER"],[4,4,"ORG"]],"relations":[[0,4,"PER-COACHES-ORG"]]}
ин","присвоил","20-му","зенитному","ракетному","полку","наименование","Херсонский"],"entities":[[0,0,"PER"],[7,7,"LOC"]],"relations":[[0,7,"PER-LOC-A
эиль","ликвидировал","командира","сил","воздушной","обороны","Хезболлах","в","Ливане"],"entities":[[0,0,"LOC"],[6,6,"ORG"],[8,8,"LOC"]],"relations":[
"Подольске","задержали","избившего","собаку-инвалида","мужчину"],"entities":[[1,1,"LOC"]],"relations":[]}
"Москве","наградили","лауреатов","HR-премии","В","кадре"],"entities":[[1,1,"LOC"]],"relations":[]}
чившая","гражданство","Касаткина","досрочно","завершила","сезон"],"entities":[[2,2,"PER"]],"relations":[]}
эла","автор","цикла","книг","Хроники","Ратшира","Джилли","Купер"],"entities":[[6,7,"PER"]],"relations":[]}
"Долгопрудном","эвакуировали","ТЦ","из-за","задымления"],"entities":[[1,1,"LOC"]],"relations":[]}
","Украины","впервые","выступили","в","роли","условного","противника","НАТО"],"entities":[[0,1,"ORG"],[8,8,"ORG"]],"relations":[[0,8,"ORG-OPPOSES-ORG
иостанция","Судного","дня","повторила","сообщение","от","февраля","2022","года"],"entities":[],"relations":[]}
ужный","заявил",",","что","Украине","пора","посмотреть","в","космос"],"entities":[[0,0,"PER"],[4,4,"LOC"]],"relations":[[0,4,"PER-LOC-Nationality"]]}
эная","Узбекистана","поедет","на","ЧМ-2026","с","Фабио","Каннаваро"],"entities":[[1,1,"ORG"],[6,7,"PER"]],"relations":[[6,1,"PER-COACHES-ORG"]]}
"Подольске","задержали","избившего","собаку-инвалида","мужчину"],"entities":[[1,1,"LOC"]],"relations":[]}
activ","узнал","о","плане","Еврокомиссии","найти","деньги","для","производства","дронов"],"entities":[[0,0,"ORG"],[4,4,"ORG"]],"relations":[[4,0,"ORG
,"Урале","опровергли","фейк","о","запрете","въезда","из","ДНР"],"entities":[[1,1,"LOC"],[8,8,"LOC"]],"relations":[]}
снодарские","власти","назвали","временными","ограничения","на","отдельных","АЗС"],"entities":[[0,0,"LOC"]],"relations":[]}
еющий","за","Спартак","ректор","Бауманки","извин                                              ,4,"ORG"],[9,9,"ORG"]],"relations":[[3,2,"P
dustan","Times","узнала","о","планах","Индии","за                                              ,1,"ORG"],[5,5,"LOC"]],"relations":[[5,0,"L
```

Fig. 8. self made RE dataset

This custom dataset was then used to fine-tune the DeepPavlov/rubert-base-cased model for the relation classification task. However, the final performance of the trained model was suboptimal.



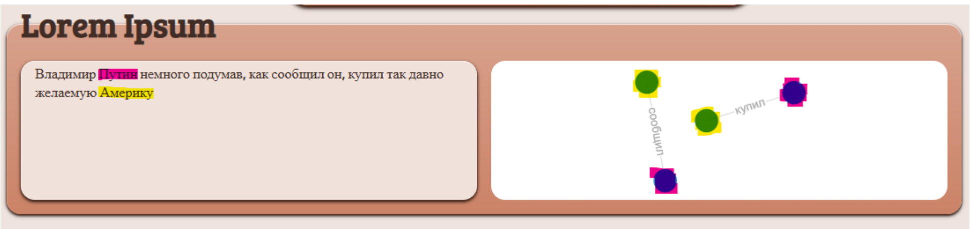
```
   micro avg       0.63      0.63      0.63
   macro avg       0.05      0.06      0.06
weighted avg       0.50      0.63      0.56
```

Fig. 9. Russian RE model

We attribute this primarily to three interconnected factors: the fundamental difficulty and time cost of collecting and annotating high-quality financial text, the relatively small size of the resulting dataset, and a lack of deep domain expertise during the LLM-assisted annotation process, which likely introduced noise and inaccuracies that the model subsequently learned.

C. Frontend Implementation

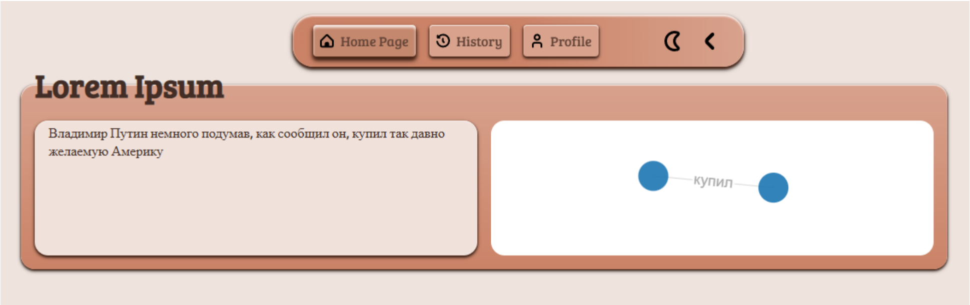
The frontend application was built using a modern technology stack comprising React, Vite, and React Router to create a dynamic and responsive user experience. A central focus was placed on developing a cohesive design system with theming support. A library of reusable components, such as data cards and graph widgets, was created, all designed with a warm color palette to ensure visual comfort. All interface elements are fully adaptive, providing an optimal layout and usability across a wide range of screen sizes, from desktop monitors to mobile devices. The application features two primary data visualization contexts: a layout for presenting structured entity information and an interactive force-graph for rendering relational networks.



Lorem Ipsum

Владимир Путин немного подумав, как сообщил он, купил так давно желаемую Америку

Fig. 10. Relation Rendering



Lorem Ipsum

Владимир Путин немного подумав, как сообщил он, купил так давно желаемую Америку

Fig. 11. layout and card

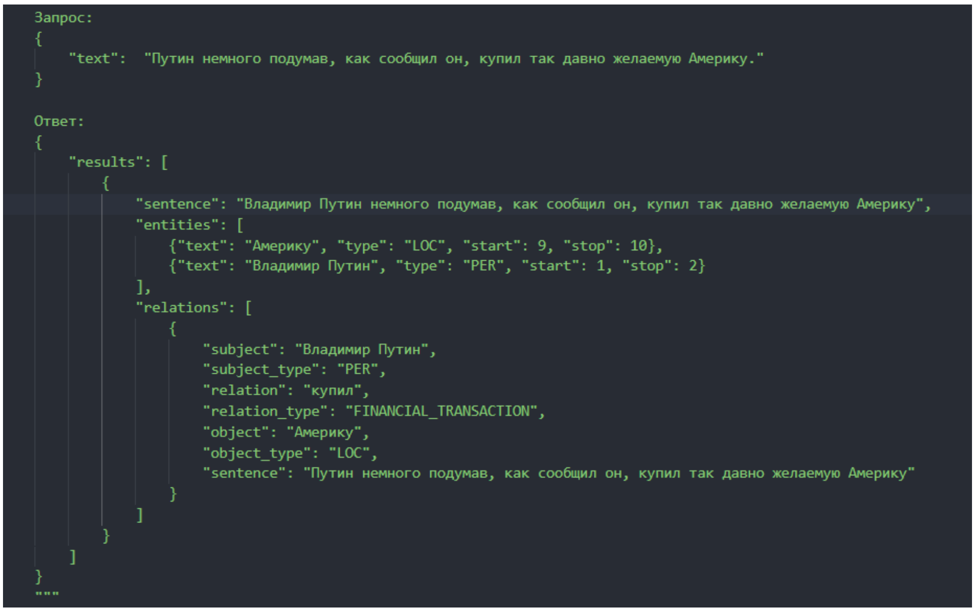
D. Backend and MLOps Architecture

The backend system follows a containerized microservices architecture to ensure modularity, scalability, and ease of deployment. The core logic is split into two distinct services. The first is a dedicated Machine Learning microservice, which is responsible for all natural language processing tasks. It receives raw text, performs the Named Entity Recognition and Relation Extraction, and returns a structured JSON output containing the identified entities and their relationships. The second service is the main API, built with the high-performance FastAPI framework. It acts as an orchestrator, handling client requests from the frontend, communicating with the ML service for processing, and managing business logic and data validation before sending the finalized data back to the client.

For deployment and DevOps, the entire system is containerized using Docker. This involves creating two separate container images: one for the ML microservice, which includes all necessary deep learning dependencies, and another for the FastAPI backend. This container-based approach provides critical isolation between the services, preventing dependency conflicts and allowing for independent scaling. For instance, multiple instances of the ML container can be spawned to handle high computational load without affecting the API layer. The use of Docker ensures environment consistency, making the application portable and reliably deployable from a local development machine to a production cloud server.

V. Results

The primary outcome of this work is a functional, end-to-end information extraction pipeline for Russian financial news. The final implemented solution utilizes a hybrid approach, combining the robust entity recognition capabilities of the Natasha model with a rule-based system for relation extraction. This pipeline successfully processes raw text input, identifies key entities (such as persons, organizations, and locations), and then applies a set of handcrafted grammatical and lexical rules to establish relationships between them. The system is fully containerized, with a React-based frontend for user interaction and a FastAPI backend that orchestrates the process, providing a clear and accessible endpoint for text analysis.



```
Запрос:
{
    "text":  "Путин немного подумав, как сообщил он, купил так давно желаемую Америку."
}

Ответ:
{
    "results": [
        {
            "sentence": "Владимир Путин немного подумав, как сообщил он, купил так давно желаемую Америку",
            "entities": [
                {"text": "Америку", "type": "LOC", "start": 9, "stop": 10},
                {"text": "Владимир Путин", "type": "PER", "start": 1, "stop": 2}
            ],
            "relations": [
                {
                    "subject": "Владимир Путин",
                    "subject_type": "PER",
                    "relation": "купил",
                    "relation_type": "FINANCIAL_TRANSACTION",
                    "object": "Америку",
                    "object_type": "LOC",
                    "sentence": "Путин немного подумав, как сообщил он, купил так давно желаемую Америку"
                }
            ]
        }
    ]
}
"""
```

Fig. 12. example

While the rule-based RE method demonstrated high precision for sentences matching its predefined patterns, its overall performance is limited by its inability to generalize. The recall is low for semantically complex or syntactically diverse sentences that fall outside the scope of the manually created rules. Consequently, the pipeline's effectiveness is currently constrained to texts that align closely with the expected grammatical structures and keywords defined during development.

A. Limitations

The main limitations of the current system are directly tied to the chosen methodology for relation extraction:

- **Low Recall and Generalization:** The rule-based RE component fails to identify relationships expressed through unconventional phrasing or complex syntax, leading to a significant number of false negatives.

- **Development Scalability:** Expanding the system's domain coverage or adapting it to new types of financial relations requires manual creation and validation of new rules, which is a time-consuming and expert-dependent process.

- **Dependence on NER Quality:** The entire RE process is contingent on the flawless identification of entities by the Natasha model. Any errors in the NER stage are directly propagated and often amplified in the relation extraction phase.

B. Future Work

Future efforts will be directed towards overcoming the current limitations by replacing the rule-based RE component with a robust, data-driven model. The immediate next step is the creation of a high-quality, expertly annotated dataset for Russian financial relations. This will involve:

- **Targeted Data Collection:** Focusing on specific financial sub-domains to ensure data consistency and depth.
- **Expert-Assisted Annotation:** Collaborating with domain experts to annotate data, ensuring high accuracy and relevance of the relation labels.
- **Advanced Model Fine-Tuning:** Utilizing this improved dataset to fine-tune a state-of-the-art transformer model, such as a further adapted rubert-base-cased, for the specific task of relation classification. This approach promises a significant increase in recall and the ability to generalize to a much wider variety of textual expressions.

This transition from a rule-based to a model-based RE system is crucial for developing a scalable, accurate, and production-ready application.

VI. Timeline of the project

Name	**Assign**	**Date**
Idea brainstorming	Andrei Zhdanov and Mikhail Sofin	September 10 → September 12
Architecture design	Andrei Zhdanov and Mikhail Sofin	September 12 → September 16
Research SOTA solutions	Mikhail Sofin	September 12 → September 15
Report D1 composing	Andrei Zhdanov	September 15
Initial backend implementation	Mikhail Sofin	September 16 → October 1
Exploratory Data Analysis	Andrei Zhdanov	September 16 → October 1
NER and RE bert model fine-tune	Andrei Zhdanov	October 1 → October 7
Collecting Russian News	Mikhail Sofin	October 1 → October 7
Report D2 composing	Andrei Zhdanov	October 5 → October 7
Creating Russian RE news dataset	Andrei Zhdanov	October 7 → October 15
Exploring Graph Visualization	Mikhail Sofin	October 7 → October 15
Russian NER task	Andrei Zhdanov	October 16 → October 25
Connect models with backend	Mikhail Sofin	October 16 → October 25

TABLE I. Project Timeline (Part 1)

Name	Assign	Date
Frontend design implemented via layout and context cards with force graphs	Mikhail Sofin	October 25 → November 4
Russian RE task	Andrei Zhdanov	October 25 → November 2
Report D3 composing	Andrei Zhdanov	November 2 → November 4
Fixing model bugs	Andrei Zhdanov	November 4 → November 15
Finish Frontend part	Mikhail Sofin	November 4 → November 15
Final Technical report composing	Andrei Zhdanov and Mikhail Sofin	November 15 → November 24
Presentation slides composing	Andrei Zhdanov and Mikhail Sofin	November 15 → November 24

TABLE II. Project Timeline (Part 2)

VII. Contributions

Zhdanov Andrei - Dataset collection, model implementation, report composing.
Sofin Mikhail - Frontend with Graphs, Article reading, backend implementation, presentation composing.