

Saint Petersburg State University

Tomin Denis Valerievich

Bachelor Diploma Thesis

Understanding the Aspect Structure of Financial Publications Using Deep Neural Networks

Level of education: Bachelor's degree

Direction 01.03.02 "Applied Mathematics and Informatics"

Basic educational program CB.5005.2015 «Management»

Graduated School of Management

Supervisor:

Professor, Research Center for Market Efficiency and Applied Finance,

Dr. Darko Vuković

Peer reviewer:

Senior Lecturer, Department of Finance and Accounting,

Vitaly Leonidovich Okulov

Saint Petersburg
2025

Содержание

Введение	4
ГЛАВА 1. ТЕОРЕТИЧЕСКИЙ ОБЗОР	7
1.1. Искусственный интеллект в финансах	7
1.1.1 Прогнозирование стоимости	7
1.1.2 Анализ тональности	8
1.1.3 Аспектный анализ	10
1.2. Алгоритмы машинного обучения	11
1.2.1 Понижение размерности	11
1.2.2 Кластеризация	14
1.2.3 Метрики оценивания	16
1.3. Глубокие нейронные сети	18
1.3.1 Модели	18
1.3.2 Техники	20
1.3.3 Бенчмарки	22
ГЛАВА 2. ПРАКТИЧЕСКОЕ РЕШЕНИЕ	25
2.1. Ограничение	25
2.2. Работа с данными	28
2.2.1 Постановка задачи	28
2.2.2 Сбор данных	31
2.2.3 Анализ данных	32
2.2.4 Предобработка данных	37
2.3. Разработка моделей	39
2.3.1 Извлечение векторных представлений	39
2.3.2 UMAP и HDBSCAN	41
2.3.3 Оптимизация гиперпараметров	43
2.4. Разработка системы	44
ГЛАВА 3. РАЗУЛЬТАТЫ	49
3.1. Спроектированная архитектура	49
3.1.1 Обзор	49
3.1.2 Эмбеддинговая система	51
3.1.3 Смесь тематических экспертов	51
3.1.4 Механизм кэширования признаков	53
3.1.5 Предиктивная модель	54
3.1.6 Аналитический графический интерфейс	56
3.2. Разработанные системные компоненты	58
3.2.1 Эмбеддинговая система и роутер	58

3.2.2	Семантическая карта	60
3.2.3	Динамическое тематическое моделирование	62
3.2.4	Прикладное значение	65
3.3.	Аналитическое решение проблемы семантической дедубликации	68
3.3.1	Математическая постановка	68
3.3.2	Преимущества и недостатки	69
	Заключение	72

Введение

In recent years, the use of data analytics and artificial intelligence, specifically machine learning (ML) and deep learning (DL), to make investment decisions has become an integral part of many companies' and funds' strategies. However, today's financial markets are characterized by high volatility and high speed of information dissemination, which creates significant challenges for analyzing its impact on stock prices.

Events such as news releases, regulatory changes and analysts' reviews can have both immediate and cumulative effects on market performance. However, traditional approaches to analyzing data often ignore the dynamics of these influences, resulting in poor forecast accuracy and, consequently, ineffective investment strategies.

To compete in a rapidly changing financial environment, companies need to continuously optimize their approaches to data analysis. This requires the development of tools that can not only account for the dynamic nature of information, but also provide forecasts based on an in-depth analysis of events and their cumulative effect. This paper responds to these challenges by proposing a methodology and a technological solution for more accurate stock price forecasting based on deep neural networks, namely large language models (LLM).

Classical ML algorithms have demonstrated their effectiveness in financial forecasting in numerous studies. However, DL and natural language processing (NLP) architectures have fundamentally shifted the paradigm following the emergence of the Transformer architecture in 2017 [Vaswani (и др.), 2017]. Since then, LLMs have gained wide acceptance and proven their applicability across various applied tasks, including asset price forecasting [Halder, 2022; Jiang, Zeng, 2023; J. Kim, H. S. Kim, Choi, 2023].

Contemporary research demonstrates the high efficacy of LLMs in addressing a range of tasks related to asset evaluation and forecasting. Nevertheless, unresolved issues remain regarding the integration of LLMs with classic quantitative models, the scarcity of open-source solutions for the financial domain, and the limitations of current models in processing long textual sequences (see Section 1.3.1). In December 2024, a new state-of-the-art (SoTA) model, ModernBERT, was introduced, capable of processing texts that are 16 times longer than those handled by previous architectures [Devlin (и др.), 2019; Warner (и др.), 2024]. This model extends analytical capabilities from isolated headlines or posts to full news articles, press releases, interview transcripts, and analytical reviews. Nonetheless, processing comprehensive financial reports (e.g., 10-Q, 10-K) remains a challenging task (see Section 2.1.4). Moreover, focusing on full articles helps to mitigate issues related to clickbait and insufficient context.

Pre-trained language models on general text corpora do not always effectively address financial forecasting tasks [Jiang, Zeng, 2023]. This is due to the unique nature of financial information, characterized by specialized terminology and jargon, which complicates the application of general-purpose models. Consequently, there is a need to adapt baseline LLMs for the financial domain.

From a management perspective, when making investment decisions in volatile markets, it is crucial to promptly analyze the cumulative impact of various events (news, regulatory changes, analyses, etc.) on asset dynamics. Experts are unable to process such an immense volume of information within extremely short time frames. Conversely, the absence of a comprehensive analysis tool leads to delayed or inaccurate decisions, reducing the effectiveness of investment strategies and increasing the risk of missing lucrative opportunities.

The development of such a tool is complex and demanding, and it can be conceptually divided into the following stages:

1. Development of an effective architecture for LLMs.
2. Adaptation of the model to the specifics of the financial domain.
3. Fine-tuning of the model to address particular tasks.
4. Integration of the model into a system operating with both quantitative and qualitative data, encompassing training, testing, and deployment processes.

Since the basic architecture of ModernBERT has already been established, the present study focuses on its domain adaptation. Among the available adaptation methods — fine-tuning, complete retraining, and domain-adaptive pretraining — the latter is emphasized in this work (see Section 1.3.2), as it minimizes time, computational, and financial costs.

Within the scope of this research, hierarchical aspect-based analysis of financial publications (see Section 3.2) is considered an effective task for financial forecasting. The object of the study is investment strategies based on the use of artificial intelligence, while the subject is the integration of language models into asset price forecasting processes. The goal of this work is to develop a practice-oriented toolkit for dynamic multimodal forecasting using aspect-based sentiment analysis. It is important to emphasize that a key requirement for the proposed solution is its interpretability, in contrast to other approaches that employ deep neural networks as black-box models.

Throughout the study, the following key tasks were undertaken:

- Selection and analysis of SoTA architectures and models (see Section 1.3.1).
- Investigation of cutting-edge techniques and algorithms to enhance the performance of deep neural networks (see Section 1.3.2).
- Evaluation of various metrics, datasets, and benchmarks to assess the efficiency of the final solution (see Section 1.4).
- Determination of the technology stack for developing the solution (see Section 1.5).

In total, more than 6 GB of exclusively financial texts were collected for the domain adaptation of ModernBERT (see Section 2.2.2). Following comprehensive analysis and

preprocessing of the data (see Sections 2.2.3 and 2.2.4, respectively), the domain adaptation of the model was performed (see Section 2.3) and inference was conducted on the task of hierarchical clustering of texts (see Section 2.4). The final outcome includes a comparison of key benchmarks between the original and the domain-adapted models (see Section 3.1), analysis and interpretation of the results obtained from the hierarchical clustering of financial publications (see Section 3.2), as well as the development of a multimodal architecture for dynamic asset forecasting based on aspect-based sentiment analysis (see Section 3.3) and a mathematical framework for semantic deduplication of texts (see Section 3.4).

All the results of this study, including data collection code, training code, analyses, and results are available in the official repository of the project¹. The domain-adapted model² and the collected corpus³ are also publicly available.

¹URL: <https://github.com/denisalpino/FinABYSS>

²URL: None

³URL: <https://huggingface.co/datasets/denisalpino/YahooFinanceNewsRaw>

ГЛАВА 1. ТЕОРЕТИЧЕСКИЙ ОБЗОР

1.1 Искусственный интеллект в финансах

1.1.1 Прогнозирование стоимости

Существует множество подходов, демонстрирующих эффективность прогнозирования стоимости активов с помощью глубоких нейронных сетей. Наиболее распространёнными в этой задаче являются рекуррентные нейронные сети (Recurrent Neural Networks, RNN), тогда как сверточные нейронные сети (Convolutional Neural Networks, CNN) используются преимущественно в качестве вспомогательного компонента.

Наиболее характерным представителем семейства RNN является архитектура долгой кратковременной памяти (Long Short-Term Memory, LSTM) [Hochreiter, Schmidhuber, 1997], часто применяемая для прогнозирования цен. Её модификации, такие как двунаправленная LSTM (Bidirectional LSTM, Bi-LSTM) и гибридные модели CNN-LSTM, также показывают высокую результативность. Более того, с появлением трансформерной архитектуры [Vaswani (и др.), 2017] всё активнее исследуются методы её адаптации к специфике временных рядов [Wen (и др.), 2022].

В числе последних экспериментов особенно выделяются:

- репозиторий, в котором демонстрируется потенциал трансформерной архитектуры на примере прогнозирования цены Биткоина⁴;
- исследование, сравнивающее производительность моделей Bi-LSTM, гибридных CNN-Bi-LSTM и трансформера при прогнозировании стоимости акций IBM⁵;
- торговый робот на базе LSTM, ориентированный на извлечение краткосрочной прибыли в боковом движении цены актива CGEN, показавший при бэктестинге доходность в размере 4% депозита за сутки⁶.

Тем не менее отдельные модели, будь то LSTM или методы на основе деревьев решений, обладают ограниченной адаптивностью при смене рыночных режимов и плохо реагируют на их динамику [Vuković (и др.), 2024]. Несмотря на очевидность этого факта, исследователи нередко недооценивают его значение. Согласно теории эффективного рынка (Efficient Market Theory, EMT), предложенной Фама в 1970 году, цены активов отражают всю доступную рыночную информацию, что ставит под сомнение возможность точного прогнозирования на основе лишь исторических количественных показателей — цены открытия, закрытия, максимума, минимума (Open High Low Close, OHLC), объёма торгов и классических индикаторов.

К тому же большинство моделей не учитывают такие тонкие аспекты, как заявки лимитного порядка, влияние других торговых роботов, а также качественную внебиржевую

⁴URL: <https://github.com/baruch1192/-Bitcoin-Price-Prediction-Using-Transformers>

⁵URL: <https://github.com/JanSchm/CapMarket>

⁶URL: <https://github.com/roeeben/Stock-Price-Prediction-With-a-Bot>

информацию. Новейшие исследования показывают, что интеграция анализа информационного поля за пределами биржи в прогнозные модели заметно повышает качество предсказаний, о чём будет подробно рассказано в последующих подразделах.

1.1.2 Анализ тональности

Как отмечалось во введении, появление трансформерной архитектуры позволило моделям глубокого обучения достичь значительного прогресса в понимании естественного языка (Natural Language Understanding, NLU). Это имеет особую ценность для финансовой сферы, где традиционные количественные данные недостаточны для точного прогнозирования.

Еще до распространения современных языковых моделей для анализа тональности текста (Sentiment Analysis, SA) применялись LSTM-модели, ULMfit, авторегрессионные сети и другие методы [Hochreiter, Schmidhuber, 1997; Howard, Ruder, 2018]. Например, в одном исследовании сравнивали авторегрессионную модель без учета тональности с аналогичной моделью, интегрировавшей тональностные признаки. Результаты показали, что в 77.8% случаев модель с учетом тональности превосходила версию, обученную только на количественных данных [Vanstone, Gepp, Harris, 2019].

Современные предобученные трансформерные модели, известные как большие языковые модели (Large Language Model, LLM), открывают новые возможности для финансового прогнозирования. Наиболее очевидное применение — анализ тональности новостей. Для этого требуется сбор значительного объема текстовых данных с последующей разметкой. Метки классов обычно включают три категории: -1 (негативный), 0 (нейтральный) и 1 (позитивный) [Pathak (и др.), 2020].

Ручная разметка предполагает привлечение отраслевых экспертов, способных оценить влияние текста на финансовые рынки. Для повышения качества данных применяется метод перекрестного консенсуса [Bogdan, Biklen, 1997]: несколько экспертов независимо аннотируют одни и те же тексты, после чего метки согласуются. Именно этот подход использовался при создании популярного датасета FinancialPhraseBank [Malo (и др.), 2014] для тонкой настройки моделей под задачу финансового анализа тональностей (Financial SA, FSA).

Алгоритмические методы, основанные на динамике стоимости активов, менее распространены из-за субъективности и нестабильности. В частности по той причине, что сложно определить пороговое значение изменения цены для классификации тональности и гарантировать, что фактический прирост не обусловлен любым другим фактором.

После разметки данных, на что уходит большая часть ресурсов, выполняется тонкая настройка предобученных языковых моделей путем модификации весов на последних слоях нейросети.

Таксономия SA включает три уровня [Pathak (и др.), 2020]:

- **Документальный уровень.** Тональность оценивается для всего документа (новости, отчета и т.д.). Данный уровень предполагает наличие в документе мнений об одной сущности, что редко соответствует реальности. Более того, существует и техническое ограничение — современные трансформеры технически не способны обрабатывать многостраничные документы (например, отчеты 10-K) или новостные статьи, вследствие ограниченных возможностей обработки токенов. Тем не менее, в последнее время стали появляться модели, способствующие обработке хотя бы новостных статей, однако отчеты все еще не могут быть обработаны ими.
- **Уровень предложений.** Тональность относится к одному или нескольким предложениям (обычно 1-8 предложений), при этом сохраняется предпосылка об одном субъекте. Именно работы данного уровня преобладают в общем числе благодаря доступности датасетов и технической совместимости с современными LLM, из которых практически все способны обработать тексты такого размера. Типичные источники — заголовки новостей, публикации в социальных сетях X (бывший Twitter) и Reddit.
- **Аспектный уровень (Aspect-Based SA, ABSA).** Выявление тональности по отношению к отдельным аспектам сущности. ABSA на данный момент является наиболее продвинутым уровнем, который довольно часто рассматривают как отдельную и достаточно обширную область. ABSA включает четыре подзадачи: извлечение аспектов, определение их полярности, категоризация аспектов и оценка полярности категорий. Отдельно выделяют таргетированный ABSA, который допускает несколько субъектов, но предполагает не более одного сентимента на каждого из них. Более подробно ABSA рассматривается в Section 1.1.3.

Большинство методов SA требуют разметки текстов, что является существенным недостатком, ввиду больших ресурсных затрат, а также имеет естественные ограничения:

- Не существует универсального и точного определения тональности, подходящего для каждой задачи. Это вынуждает формулировать уникальное определение в ходе решения практических задач, что затруднительно из-за многочисленности паттернов и сценариев, встречающихся в текстах.
- Невозможно без кратного увеличения затрат на контроль (например, при помощи метода перекрёстного консенсуса [Bogdan, Biklen, 1997]) обеспечить надежность и объективность оценок, так как на качество разметки сильно сказывается человеческий фактор.

Как было отмечено, постановка SA может несколько отклоняться от своей классической постановки в контексте различных доменов. Так, особенностью FSA является то, что

он фокусируется не исключительно на обособленных текстах, но так же и на количественной информации, которая в финансовых статьях крайне много [Du (и др.), 2024].

Тем не менее, FSA терпит неудачи в нескольких типовых сценариях среди которых: нереальные настроения (условные настроения, сослагательное наклонение, повелительное наклонение), риторика (негативные заявления, персонификация, сарказм), зависимые мнения, неопределенные аспекты, нераспознанные слова (жаргонизм, микротекст, сущности) и external references (то есть отсылки к тем знаниям, которые не заключены в модель) [Xing (и др.), 2020].

1.1.3 Аспектный анализ

ABSA представляет собой специализированную задачу выявления и оценки сентимента по отношению к конкретным «аспектам» — характеристикам или свойствам рассматриваемого объекта. В финансовой сфере такими аспектами могут выступать рисковые факторы, кредитная политика, макроэкономические индикаторы и другие элементы, упоминаемые в публикациях. Классический подход к ABSA включает три этапа: выделение аспектов из текста, определение полярности сентимента по каждому аспекту, и агрегацию результатов для построения итоговой картины мнений экспертов или рынка .

Параллельно с развитием ABSA в лингвистическом сообществе активно развалось тематическое моделирование, главной целью которого является выявление скрытых «тем» (topics) в корпусах документов. Одним из первых и наиболее влиятельных методов стал Latent Dirichlet Allocation (LDA) [Blei, Ng, Jordan, 2003], где темы формулируются как распределения слов, а документы рассматриваются как смеси таких распределений. В контексте финансовых текстов LDA позволяет выделять основные направления обсуждения (например, «монетарная политика», «корпоративные риски» и т.д.), но испытывает сложности с моделированием синтагматических связей и динамичной лексики.

На стыке эмбеддинговых моделей и тематического анализа сформировалась идея top2vec [Angelov, 2020], согласно которой темы представляются в том же векторном пространстве, что и слова, что позволяет объединить преимущества распределённых представлений и тематических структур. Далее эволюция привела к появлению BERTopic [Grootendorst, 2022], где для построения тем используются плотностные алгоритмы поверх эмбеддингов, а затем для каждого кластера извлекаются наиболее характерные ключевые слова. Эта модель демонстрирует высокую адаптивность к изменениям лексики и позволяет работать с динамическими, даже мультиязычными корпусами.

Связь между ABSA и тематическим моделированием очевидна: аспекты в широком смысле являются темами, по отношению к которым измеряется сентимент. В традиционном ABSA аспекты задаются или извлекаются на основе лингвистических паттернов (например, правилами на зависимостях или словарями), тогда как тематическое моделирование открывает путь к автоматическому выявлению аспектов как скрытых латентных переменных. Пре-

имущество такого подхода в финансовом домене заключается в том, что заранее неизвестное множество аспектов («тем») может быть извлечено без ручной разметки, после чего к выявленным кластерам документов применяется стандартная процедура оценки сентимента (например, с помощью LLM), что позволяет получать более полный и интерпретируемый анализ мнений рынка.

Таким образом, в рамках данной работы аспекты рассматриваются как темы в тематическом моделировании. А работа предлагает способ первоначальной кластеризации векторных представлений, которая формирует «тематические кластеры», для того, чтобы агрегировать по ним сентимент. Такое объединение позволяет (1) выявлять как глобальные тренды, так и локальные, нишевые аспекты финансового рынка, (2) минимизировать необходимость ручной разметки аспектов, и (3) обеспечить интерпретируемость результатов за счёт явной привязки сентимента к темам, выраженным через ключевые термины каждого кластера. Это сочетание взяло лучшее от обоих направлений: структуры тем при LDA/topic2vec/BERTopic и точности полярности при ABSA.

1.2 Алгоритмы машинного обучения

1.2.1 Понижение размерности

Метод главных компонент (Principal Component Analysis, PCA) является линейным методом понижения размерности (dimensionality reduction, DR), основанным на разложении ковариационной матрицы исходных данных. При нулевом среднем по признакам он находит собственные векторы ковариации (компоненты), отвечающие за максимальную дисперсию: $\Sigma = X^T X$, $\Sigma w_i = \lambda_i w_i$, проекцию данных $Y = XW_{d'}$ на первые d' собственных векторов.

PCA стремится сохранить как можно больше дисперсии исходных данных и, как следствие, хорошо сохраняет «глобальную» структуру кластеров [Amid, Warmuth, 2019]. На практике PCA часто применяется как промежуточный этап: например, при работе с высокоразмерными текстовыми эмбеддингами (несколько сотен признаков, например, BERT-эмбеддинги размерности 768) перед нелинейными методами понижения размерности. Использование PCA позволяет значительно сократить размерность (до десятков — сотен компонент) и ускорить последующие вычисления [H. Huang (и др.), 2022].

Однако PCA — линейный метод, и он не отражает более сложных нелинейных зависимостей, присущих текстовым эмбеддингам. Вследствие этого тонкие локальные отношения слов и документов могут быть потеряны, хотя общая структура («глобальный ландшафт» данных) при этом чаще всего сохраняется.

However, PCA is a linear method, and it does not capture the more complex nonlinear dependencies inherent in language embeddings. As a consequence, subtle local word-document relationships may be lost, although the overall structure (the “global landscape” of the data) is most often preserved.

Стохастическое вложение соседей с t-распределением (t-distribution Stochastic Neighbor Embedding, t-SNE) нелинейный стохастический метод, ориентированный на сохранение локальной структуры данных. В высокоразмерном пространстве он вычисляет условные вероятности $p_{j|i} \propto \exp(-|x_i - x_j|^2 / 2\sigma_i^2)$, отражающие близость соседей; затем симметризует их: $p_{ij} = (p_{j|i} + p_{i|j})/2n$. В низкоразмерном отображении задаётся похожая мера (распределение Стьюдента с 1 степенью свободы). Алгоритм оптимизирует расстановку точек Y , минимизируя дивергенцию Кульбака–Лейблера $C = \sum_{i,j} p_{ij} \ln \frac{p_{ij}}{q_{ij}}$. В результате в низкоразмерном пространстве у близких в исходном пространстве точек будет сохраняться локальная кластерная структура.

Гиперпараметр перплексии определяет число эффективных соседей. t-SNE показывает впечатляющую визуализацию локальных кластеров, но плохо воспроизводит глобальное расстояние между кластерами. Он вычислительно дорог при больших выборках и обычно используется лишь для финального перехода в двумерное пространство (или трехмерное), а не для промежуточного снижения размерности. Для текстовых эмбеддингов t-SNE часто применяют после предварительной обработки (например, снижения размерности PCA), поскольку напрямую на несколько сотен признаков он масштабируется плохо.

Равномерная аппроксимация и проекция многообразий (Uniform Manifold Approximation and Projection, UMAP) метод нелинейного DR на основе предположения о многообразии. Теоретически UMAP опирается на риманову геометрию и теорию «нечетких симплексиальных множеств» (fuzzy simplicial sets). Алгоритм строит граф k -ближайших соседей в исходном пространстве, затем каждой паре точек присваивает «членство» в нечетком множестве по формуле вида:

$$\mu_{ij} = \exp \left(- \frac{\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i} \right), \quad (1)$$

где ρ_i учитывает плотность соседей. После чего эти локальные симплексиальные множества объединяются и симметризуются, получается взвешенный граф данных. Далее строится аналогичный «нечеткий» граф в низкоразмерном пространстве и оптимизируется расположение точек, минимизируя энергию перекрёстной энтропии между двумя графиками.

UMAP удерживает локальную структуру данных, при этом стремясь равномерно распределить точки на многообразии; в отличие от t-SNE, он может лучше сохранять некоторые глобальные особенности (обусловлено используемой кросс-энтропией). Основные гиперпараметры UMAP – число соседей ‘n_neighbors’ (задаёт масштаб локальности) и ‘min_dist’ (минимальное расстояние точек в отображении).

UMAP демонстрирует высокую скорость и масштабируемость (метод может работать сразу на произвольном числе выходных измерений). В эксперименте было показано, что UMAP даёт качество визуализации, сопоставимое или лучше, чем у t-SNE, при существенно меньшем времени работы [McInnes, Healy, Melville, 2018]. К преимуществам UMAP мож-

но отнести также сохранение более крупномасштабной структуры данных и возможность понижать размерность вплоть до многомерных векторов, а не только двумерных.

Тем не менее, UMAP может некорректно отображать сильно разреженные кластеры, «выравнивая» плотные и разреженные области (алгоритм фактически стремится к равномерному распределению данных на предполагаемом многообразии). Кроме того, результаты UMAP чувствительны к выбору гиперпараметров и степени шума в выборке (алгоритм использует приближённый поиск соседей и сэмплирование отрицательных примеров) [H. Huang (и др.), 2022]. Впрочем, в большинстве прикладных задач по кластеризации текстовых данных UMAP показал себя устойчивым и эффективным инструментом.

Существует несколько Python-реализаций данного метода. Классическая реализация — 'umap-learn' (CPU) — имеет крайне широкое распространение [McInnes, Healy, Saul (и др.), 2018]. Для ускорения процесса обучения существует GPU-реализации ('cuML') [Raschka, Patterson, Nolet, 2020]. GPU-реализация UMAP в 'cuML' может давать ускорение до 10-100× по сравнению с CPU-версией на больших объёмах данных. Однако в ранних версиях 'cuML' в целях скорости были введены приближения, что иногда влечёт небольшую разницу в качестве отображения по сравнению с оригиналом. Также GPU реализация UMAP может понизить сохранность локальной структуры относительно эталонной версии. Так, 'cuML' UMAP подходит для очень больших объемов данных, а реализация из 'umap-learn' более универсален и стабилен, но работает медленно медленнее.

Триплетная аппроксимация и проекция многообразий (Triplet Manifold Approximation and Projection, TriMAP) метод обучения вложения с упором на глобальную структуру данных [Amid, Warmuth, 2019]. Он формулирует задачу через тройки точек (i, j, k) : «точка i должна быть ближе к j , чем к k » в низкоразмерном представлении. Выбор таких триплетов базируется на ближайших и дальних соседях в исходном пространстве; каждому триплету придаётся вес, отражающий относительную близость пар в исходном пространстве. Оптимизация выполняется по большой выборке информативных триплетов с помощью градиентного спуска.

TriMAP сохраняет глобальную структуру гораздо лучше, чем t-SNE и часто лучше, чем UMAP. Также TriMAP хорошо масштабируется и демонстрирует малое время работы для больших и высокоразмерных выборок [Там же]. С точки зрения текстовых эмбеддингов TriMAP может дать более читаемую картину расположения кластеров документов на двумерную плоскость (хотя при этом детали локальных сообществ могут сглаживаться).

Аппроксимация парных управляемых многообразий (Pairwise Controlled Manifold Approximation and Projection, PaCMAP) более новый метод, специально сконструированный для баланса локальной и глобальной структуры [Y. Wang (и др.), 2021]. Как и в TriMAP, здесь используются выборки пар точек разных типов: «близкие» пары (соседи), «средние» пары (между кластерами) и «дальние» пары. Для каждого типа пар заданы соответствующие веса и силы притяжения/отталкивания. В результате оптимизируемая функция

потерь стремится одновременно сжимать локально близкие точки и раздвигать удалённые, сохраняя глобальную форму распределения.

PaCMAP устойчив к выбору гиперпараметров и сокращению размерности при предобработке, хорошо сохраняет как локальную, так и глобальную структуру. Недостатки PaCMAP — относительная новизна и необходимость подбора долей разных типов пар.

Систематические сравнения указывают на характерное разделение свойств этих алгоритмов. Так, PCA, TriMAP и PaCMAP хорошо сохраняют глобальные расстояния (крупномасштабную структуру кластеров), в то время как t-SNE и UMAP лучше улавливают локальные детали [H. Huang (и др.), 2022]. PCA традиционно используется для предобработки: снижение размерности до десятков компонент ускоряет дальнейший анализ и делает его более стабильным. Однако заметно, что полная предобработка PCA может исказить исходные расстояния, поэтому результаты итогового вложения (например, визуализации t-SNE/UMAP) часто зависят от числа компонент PCA. В экспериментальных оценках метода PaCMAP и TriMAP показали наилучшее согласование глобальных расстояний, тогда как UMAP и t-SNE в среднем уступали им в этой задаче. И наоборот, в задачах классификации на векторных признаках (проверка локальной согласованности) лучше всего проявили себя t-SNE и UMAP.

В целом выбор метода понижения размерности для высокоразмерных эмбеддингов документов зависит от задачи: для последующей кластеризации и тематического часто применяют PCA или UMAP (для более «стабильного» представления кластеров), а для финальной двумерной визуализации и детального анализа локальных кластеров — t-SNE, UMAP или PaCMAP. Тщательный выбор гиперпараметров и, возможно, сочетание методов (например, PCA+UMAP) позволяет добиться лучшего отображения структуры текстовых данных.

1.2.2 Кластеризация

В рассматриваемой схеме тематического анализа финансовых новостных статей после получения текстовых эмбеддингов применяется алгоритм понижения размерности. После такого сокращения размерности алгоритмы кластеризации работают в менее разреженном пространстве, что может улучшить выделение плотных областей и снизить влияние шумовых факторов.

K-Средних один из классических партиционных алгоритмов кластеризации [Jain, 2010]. Он ищет разбиение данных на K кластеров, минимизируя внутрикластерный разброс точек. Обозначим кластеры C_1, \dots, C_K и центроиды кластеров μ_k ; тогда оптимизируемая целевая функция (метрика «суммы квадратов расстояний» от точек до соответствующих центроидов) задаётся как

$$J(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \quad (2)$$

и минимизируется при разбиении по Евклидовой метрике. Нахождение глобального

минимума этой функции является NP-полной задачей, поэтому метод К-Средних выполняет жадную итеративную процедуру переклассификации точек и пересчёта центроидов (обычно со случайной инициализацией), которая сходится к локальному минимуму. Важной особенностью К-Средних является требование заранее задать число кластеров K и начальные приближения.

Поскольку алгоритм обычно использует Евклидову метрику, он образует в основном сферические кластеры. Все объекты автоматически относятся к какому-либо кластеру (жёсткое «принадлежность» каждой точки одному кластеру), и алгоритм не выделяет явно выбросы или шум. Среди преимуществ метода — простота реализации, низкие вычислительные затраты и широкая распространённость. Однако К-Средних неустойчив к выбросам, и плохо выделяет вложенные или сильно неоднородные по форме кластеры.

Пространственная кластеризация приложений с шумом на основе плотности (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) метод кластеризации на основе плотности [Ester (и др.), 1996]. Он определяет кластеры как регионы высокой плотности данных, отделённые низкоплотными областями. Алгоритм использует два параметра: радиус ϵ и минимальное число точек min_{pts} . Точка называется «ядром» кластера, если в её ϵ -окрестности содержится по крайней мере min_{pts} точек. Точки, достижимые по плотности от ядра, принадлежат тому же кластеру.

DBSCAN автоматически отделяет точки, не попавшие в плотные области, как шум, и не требует задания числа кластеров. Благодаря этому метод находит кластеры произвольной формы и хорошо подходит для наборов данных с неравномерно распределёнными объектами. Однако DBSCAN имеет значимые ограничения: выбор единого порога ϵ критичен, и при совмещении кластеров разной плотности алгоритм либо объединяет их в одно целое, либо разбивает на слишком мелкие фрагменты. Кроме того, с ростом размерности пространства данные разрежаются, и отличить высокоплотную область от низкоплотной становится затруднительно. Вследствие этого DBSCAN в высокоразмерных эмбеддингах текста часто демонстрирует пониженную эффективность. Также сложность классического DBSCAN при отсутствии оптимизации может достигать $O(n^2)$, хотя практические реализации с индексами обычно имеют существенно меньшую скорость работы.

Иерархический DBSCAN (HDBSCAN) иерархическое расширение метода DBSCAN [Campello, Moulavi, Sander, 2013]. В отличие от DBSCAN, HDBSCAN не требует фиксированного ϵ : вместо этого он рассчитывает расстояние взаимной достижимости между точками x и y как

$$d_{\text{mreach}}(x, y) = \max\{d_{\text{core}}(x), d_{\text{core}}(y), d(x, y)\}, \quad (3)$$

где $d_{\text{core}}(x)$ — расстояние от точки x до её k -го ближайшего соседа (то есть минимальное ϵ , при котором x становится «ядром» кластера). Затем строится граф полных взаимных достижимости (или напрямую минимальное оствовное дерево на таких расстояниях). Удаляя

рёбра этого дерева в порядке убывания веса, алгоритм формирует древовидную иерархию кластеров, отражающую вложенную структуру данных при разных порогах плотности. Из этого дерева по критерию стабильности выделяются финальные кластеры. Такая процедура эквивалентна проведению множества запусков DBSCAN при всех возможных ϵ и выбору наиболее значимых «стабильных» групп. Важной чертой HDBSCAN является то, что он сам находит оптимальное число кластеров, требуя лишь указать минимальный размер кластера ('min_cluster_size').

Благодаря иерархическому подходу HDBSCAN способен выявлять вложенные кластеры различной плотности и более гибко адаптироваться к распределению данных по сравнению с DBSCAN и К-Средних. Алгоритм устойчив к флуктуациям плотности: если в данных имеются области разной однородности, HDBSCAN выделит крупные разреженные кластеры и более мелкие плотные кластеры одновременно. Он автоматически помечает выбросы, аналогично DBSCAN, но без жёсткой привязки к одному порогу. Из-за дополнительной обработки (поиск k -ближайших соседей, построение MST и анализ иерархии) HDBSCAN несколько сложнее вычислительно, но современные реализации с эффективными структурами соседства обычно обеспечивают сопоставимую или даже лучшую производительность [McInnes, Healy, Astels, 2017]. В целом HDBSCAN предоставляет более богатое описание структуры данных на разных уровнях детализации и чаще даёт более осмысленные кластеры в сложных многомерных пространствах.

Существующие экспериментальные исследования показывают, что в задаче тематического анализа длинных текстов каждый метод имеет свои плюсы и минусы. К-Средних часто используется как базовый метод ввиду простоты и масштабируемости, однако он даёт сравнительно грубое разбиение тем, так как ограничен сферической формой кластеров и требует заранее указать число тем. DBSCAN позволяет выявлять кластеры произвольной формы и отделять шум, но его эффективность снижается на высокоразмерных эмбеддингах текстов из-за разреженности и необходимости настройки глобального порога плотности.

Во многих сравнительных экспериментах показано, что HDBSCAN превосходит оба упомянутых метода по качеству тематических кластеров: он автоматически адаптируется к вариативности плотности эмбеддингов, выявляет вложенные темы разной детальности и надёжно исключает нерелевантный шум [Campello, Moulavi, Sander, 2013; McInnes, Healy, Astels, 2017]. Эти выводы подтверждаются практическими применениями (например, при кластеризации новостных статей), где HDBSCAN чаще всего даёт более интерпретируемые и стабильные результаты по сравнению с К-Средних или «плоским» DBSCAN [Grootendorst, 2022].

1.2.3 Метрики оценивания

В задачах совместной оптимизации алгоритмов понижения размерности и кластеризации критически важно применять метрики, способные одновременно оценивать качество

проекции и чистоту выделенных групп. Две наиболее распространённые метрики в этом контексте — коэффициент силуэта [Rousseeuw, 1987] и индекс DBCV [Moulavi (и др.), 2014].

Коэффициент силуэта характеризует для каждого объекта соотношение средней внутрикластерной и ближайшей внекластерной дистанций [Rousseeuw, 1987]. Для объекта i вычисляются

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i \setminus \{i\}} d(i, j), \quad b(i) = \min_{C \neq C_i} \frac{1}{|C|} \sum_{j \in C} d(i, j), \quad (4)$$

где C_i — кластер, содержащий i , а $d(\cdot, \cdot)$ — метрика (обычно Евклидова). Тогда

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]. \quad (5)$$

Среднее значение $s = \frac{1}{N} \sum_i s(i)$ отражает, насколько объекты внутри кластеров компактны и удалены от соседних кластеров.

Коэффициент силуэта хорошо подходит для оценки разделимости при «сферических» кластерах, однако при плотностном разделении (разной плотности и формы кластеров) он может дать завышенные или искажённые оценки, поскольку при этом внутрикластерная и межкластерная дистанции не отражают качества плотностных областей. В частности, то же справедливо и для других индексов подобных метрик, например, индекс Дэвиса–Болдена и Калински–Харабаша [G. Liu, 2023]. Все они опираются на средние расстояния до центров или дисперсию по кластерам и не учитывают неоднородность плотности и выделение шума [G. Liu, 2024].

Так, была разработана метрика на основе плотностей специально для DBSCAN/HDBSCAN-кластеризации — индекс валидации кластеризации на основе плотности (Density-Based Clustering Validation Index, DBCV) [Moulavi (и др.), 2014]. Данная метрика измеряет среднее отношение плотностей «внутри — между» кластерами, а также явно обрабатывает шум и произвольную форму кластеров. В совокупности, она позволяет корректно оценивать кластеры произвольных форм, которые строго говоря называются нерегулярными.

С другой стороны, у DBCV есть и ключевой недостаток — при перекрытии кластеров он дает большую оценку их объединению, нежели рассечению, что может негативно сказаться именно на иерархической плотностной кластеризации. Еще одна проблема с существующими индексами валидности кластера заключается в предположении, что данные внутри кластера имеют однородное распределение, даже если форма кластера произвольна

Также существуют и другие метрики, например, такие, как VIASCKDE [Şenol, 2022] и Min-Max-Jump Silhouette coefficient (MMJ-SC) [G. Liu, 2023; 2024], они явно обладают потенциалом в достаточно широком спектре задач, однако все равно уступают DBCV в стабильности на большом спектре различных задач и тестовых кейсов. Более того, например, MMJ-SC абсолютно новая метрика и еще не была достаточно протестирована на практике.

На тестовых данных кластеров сложных вложенных форм в трехмерном пространстве,

DBCV индекс, единственная метрика, которая корректно отработала, показав максимальное значение 0.53 на истинном варианте кластеризации, в то же время коэффициент силуэта хоть и показал большее максимальное значение 0.62, но данный максимум метрики пришелся на категорически неверную разметку кластеров [G. Liu, 2024].

1.3 Глубокие нейронные сети

1.3.1 Модели

Изначально в аналитике финансовых текстов применялись классические рекуррентные [Rumelhart, Hinton, R. J. Williams, 1986] и сверточные [LeCun (и др.), 1998] сети. Рекуррентные сети, включая улучшенные варианты LSTM [Hochreiter, Schmidhuber, 1997], способны обрабатывать текст как последовательность, моделируя контекст слова на основе предыдущих состояний. Поворотным этапом стал переход к трансформерам [Vaswani (и др.), 2017].

Представления двунаправленного кодировщика трансформера (Bidirectional Transformer Encoder Representations, BERT). Модель BERT [Devlin (и др.), 2019] предложила двунаправленный кодировщик на основе механизмов внимания, что позволило учитывать контекст слова одновременно слева и справа. BERT задаёт новую парадигму: модель предобучается на больших корпусах на задаче моделирования маскированного языка (Masked Language Modeling, MLM), а затем тонко настраивается под конкретные задачи.

Ключевой архитектурной особенностью BERT является полно связная модель трансформера с механизмом Multi-Head Attention и скрытыми слоями Feed-Forward (в BERT-base – 12 слоёв по 768 нейронов). При этом BERT базово ограничен длиной входной последовательности (до 512 токенов), что ставит ограничения на обработку длинных документов. Зато архитектура BERT оказалась крайне удачной для тонкой настройки: она хорошо зарекомендовала себя в классификации, вопросно-ответных системах и других задачах при минимальных изменениях слоёв вывода.

Финансовый BERT (FinBERT). Для финансового домена были разработаны адаптированные версии BERT — FinBERT. Так, в 2019 году была разработана модель FinBERTv1 [Araci, 2019], дообученный BERT на финансовом датасете для задачи анализа тональности. Его модель показала улучшение всех метрик. FinBERTv1 использует стандартную архитектуру BERT, но имеет специфический словарь финансовых текстов — например, термины из отчетов и новостей. Несмотря на размер обучающей выборки меньший, чем у моделей общего назначения, Успех FinBERTv1 прежде всего обусловлен тем, что модель «понимает» контекст финансовой терминологии лучше, чем модели общего назначения.

Позже была разработана другая версия FinBERTv2, специально предобучая её на финансовых корпусах (новости, отчёты компаний, финансовые форумы) и используя мультизадачное обучение [Z. Liu (и др.), 2020]. Это позволило FinBERTv2 выучить богатые представления, учитывающие специфику финансовой лексики. FinBERTv2 демонстрирует заметно

лучшие результаты на задачах финансового сентимента и вопросно-ответных системах.

Еще одна работа — FinBERTv3 [A. Huang, H. Wang, Y. Yang, 2023; Y. Yang, UY, A. Huang, 2020] — также адаптировала BERT под финансы, но с акцентом на интерпретируемость и извлечение информации из аналитических отчётов. Авторы развивают идею учета контекста: их FinBERTv1 «интегрирует финансовые знания» и лучше «резюмирует контекстуальную информацию в финансовых текстах». При этом модель значительно превосходит как словарные подходы, так и классические алгоритмы машинного обучения, а также свёрточные и LSTM-модели на задачах классификации тональности в экспериментах с разметкой аналитиков. Особенно сильным преимуществом FinBERTv3 [Y. Yang, UY, A. Huang, 2020] является обработка предложений, где другие модели ошибочно ставят метку «нейтрально» — благодаря лучшему учёту контекста модель выявляет позитивные/негативные оттенки там, где неглубокие методы промахиваются.

Таким образом, все три варианта FinBERT основаны на архитектуре трансформера BERT, но различаются стратегией обучения: модель FinBERTv1 просто тонко настроена на специализированном датасете, FinBERTv2 — предобучена смешанных данных и задачах, а FinBERTv3 — упор сделан на извлечение финансовой информации, контекстный анализ и интерпретацию. Во всех случаях финансовая специализация даёт превосходство над «чистой» моделью BERT в финансовом домене.

ModernBERT [Warner (и др.), 2024] — это новейшее поколение моделей, кодировщиков, эта архитектура сохраняет идею BERT, но включает несколько усовершенствований. Во-первых, ModernBERT обучен на очень большом объёме данных (≈ 3 триллиона токенов) и изначально поддерживает длину последовательности до 8192 токенов, в то время как исходный BERT ограничен 512 токенами. Это качественно меняет способности модели — она может захватывать длинный контекст документа целиком без специальной пост-обработки. Во-вторых, архитектура ModernBERT включает технологические новшества: так, в ней применены гейтированные слои GeGLU, использованы ротационные позиционные кодировки (RoPE) для лучшего кодирования положения на больших расстояниях, а также внедрён чередующийся механизм локально-глобального внимания вместо полного механизма внимания у BERT. Это позволяет эффективнее обрабатывать длинные тексты с точки зрения вычислительных ресурсов. Кроме того, ModernBERT оптимизирован для скорости и памяти: он спроектирован с учётом аппаратных особенностей GPU.

По сравнению с BERT и его «каноническими» модификациями, ModernBERT достигает новых рекордов эффективности. Так, эта модель превосходит BERT на всех стандартных задачах бенчмарка GLUE (см. Раздел 1.3.3) и значительно улучшает показатели в задачах извлечения и поиска по длинным контекстам. Её показатели значительно выше конкурентов в задачах с длинным контекстом. Например, при построении индексов ColBERT ModernBERT имеет метрику 80.2, в то время, как для BERT она составляет всего 28.1, а для предыдущей модели с аналогичным контекстом — 69.3. При этом ModernBERT также обрабатывает ко-

роткий контекст быстрее предыдущих лидирующих моделей, сохраняя при этом экономию памяти [Devlin (и др.), 2019; Warner (и др.), 2024]. Иными словами, ModernBERT представляет собой «новое поколение» кодировщиков. Он не только эффективен, но и специально заточен под длинные тексты и большие масштабы данных.

Таким образом, ModernBERT — это эволюционная ступень архитектуры кодировщика. Он сохраняет концепцию двунаправленного трансформера BERT, но адаптирует её к масштабным данным и длинным последовательностям. Благодаря сочетанию новых архитектурных приёмов и гигантского корпуса обучения ModernBERT значительно расширяет возможности по обработке финансовых (и любых других) текстов большого объёма по сравнению с оригинальным BERT.

1.3.2 Техники

Тонкая настройка. В современных NLP-системах, основанных на трансформерах, предобученная модель служит обобщённым языковым ядром, однако она не всегда готова давать оптимальные представления эмбеддингов для конкретных задач. Тонкая настройка представляет собой дополнительную фазу обучения, в ходе которой уже предобученная модель корректирует свои внутренние веса под узконаправленные цели. Тонкая настройка задействует размеченные данные именно для целевой задачи (классификация, регрессия и т. д.). В финансовом тематическом анализе тонкая настройка критична для того, чтобы модель не просто генерировала универсальные представления текста, а акцентировала внимание на релевантных паттернах, интенсивности финансовых терминов, синтаксических конструкциях и даже количественных упоминаниях, что обеспечивает большую точность кластеризации и прогнозных модулей.

Одной из наиболее востребованных downstream-задач является оценка семантической схожести текстов (Semantic Textual Similarity, STS) — определение смыслового сходства двух текстовых фрагментов. В отличие от бинарных задач классификации, STS формулируется как задача регрессии: модели требуется выдать непрерывную оценку близости, обычно в диапазоне от 0 до 5 [Cer (и др.), 2017] или от 0 до 1 [Gao, Yao, D. Chen, 2021].

STS-задача предъявляет высокие требования к качеству эмбеддингов: они должны сохранять тонкие семантические нюансы и одновременно быть инвариантны к синтаксическим перестановкам и стилевым вариациям. Так, была предложена архитектура Sentence-BERT, где два экземпляра BERT-сети обмениваются градиентами через сиамскую конфигурацию, а их выходные векторы сравниваются по косинусному расстоянию, что позволило значительно ускорить и улучшить оценку STS [Reimers, Gurevych, 2019].

Таким образом, тонкая настройка эмбеддинговой модели на STS играет ключевую роль в построении пайплайна, где комбинируются модели понижения размерности и кластеризации. Тонкая настройка обеспечивает эмбеддинги, в которых локальные и глобальные семантические связи выстроены таким образом, чтобы и нелинейное понижение размерно-

сти сохраняло релевантные соседства, и плотностная кластеризация автоматически выделяла тематические объединения без избыточного шума.

Доменно-адаптированное предобучения (Domain-Adaptive Pretraining, DAPT). В современной практике работы с LLM оказывается недостаточным одноразовое предобучение на обширных общекорпусных данных. DAPT представляет собой дополнительную фазу предобучения на текстах целевого домена, предшествующую этапу тонкой настройки на определенной задаче [Gururangan (и др.), 2020].

В отличие от классической тонкой настройки, при которой модель лишь корректирует веса в последних слоях под конкретный датасет и задачу, DAPT продлевает исходную задачу языкового моделирования, позволяя модели «освежить» и углубить своё представление о лексике, синтаксисе и семантике специфического домена — например, финансовых публикаций. Такое поэтапное наращивание преобучения демонстрирует существенные приросты качества на доменных задачах как в условиях дефицита размеченных данных, так и при их изобилии.

Исходя из расчетов, DAPT в среднем обеспечивает прирост бенчмарков на 4% в относительной шкале по сравнению с базовой моделью, которая не была адаптирована под определенный домен. Данная цифра является весьма весомой, учитывая, что для некоторых специфичных задач прирост может составлять вплоть до 20% [Там же].

Необходимость доменной адаптации особенно очевидна в финансовом секторе: общие корпуса, такие как BookCorpus [Zhu (и др.), 2015] или Wikipedia [Merity (и др.), 2016], не отражают терминологию, метафоры и стилистические паттерны пресс-релизов, аналитических отчётов и биржевых новостей. Множество исследований подтвердило преимущества DAPT в специализированных областях. Так, разработчики SciBERT продемонстрировали, что предобучение BERT на корпусе научных статей приводит к улучшению показателей на задачах извлечения научных сущностей и классификации статей по дисциплинам, обеспечивая рост F1-метрики до 2-3 пунктов по сравнению с базовой моделью [Beltagy, Lo, Cohan, 2019]. Аналогично BioBERT, адаптированный к биомедицинскому тексту, повысил точность извлечения медицинских терминов и отношений на 10-20% по сравнению с исходным BERT [Lee (и др.), 2020].

В финансовом домене доменная адаптация позволяет LLM лучше улавливать нюансы упоминаний акций, процентных ставок и регуляторных терминов, что вкупе с задачей тематической кластеризации повышает качество эмбеддингов и делает выходные представления более «разборчивыми» при последующей кластеризации методом HDBSCAN. Таким образом, DAPT выполняет роль «моста» между общими языковыми знаниями и потребностями предметной области, позволяя добиться значительного улучшения результатов на доменных задачах.

Механизм слияния. В многомодальных и многокомпонентных системах механизм слияния отвечает за объединение информации из разных источников или представлений,

обеспечивая совместную обработку гетерогенных сигналов.

Первый — раннее слияние — предполагает подачу всех признаков (биржевых котировок, технических индикаторов и текстовых эмбеддингов) сразу в единую модель CNN-LSTM [Dutt, Zare, Gader, 2022; Karpathy (и др.), 2014]. Преимущество метода — простота реализации и возможность немедленно учить кросс-модальные зависимости. Однако на практике раннее слияние подвержен «захлёбыванию» в шуме одной из модальностей и утрачивает гибкость при динамической оценке вклада каждой из них [Dutt, Zare, Gader, 2022].

Второй метод — позднее слияние — объединяет предсказания отдельных каналов (каждая модальность обрабатывается своей CNN-LSTM-ветвью) лишь на завершающем этапе [Karpathy (и др.), 2014; Ortega (и др.), 2019]. Такой подход отличается модульностью (простота замены или дообучения отдельного канала), но исключает извлечение низкоуровневых кросс-модальных закономерностей и требует обучения всех ветвей по-отдельности, что влечет кратное увеличение вычислительных ресурсов [Dutt, Zare, Gader, 2022].

Третий, компромиссный механизм — медленное слияние — обеспечивает поэтапное слияние каналов на разных уровнях сети [Dutt, Zare, Gader, 2022; Feichtenhofer, Pinz, Zisserman, 2016]. Ключевые преимущества метода: баланс между автономной переработкой каждой модальности и возможностью учёта их взаимодействия, сохранение «чистоты» низкоуровневых признаков и гибкость настройки числа и глубины этапов интеграции [Karpathy (и др.), 2014]. Главные недостатки — сложность выбора оптимального уровня слияния и повышенные вычислительные затраты из-за параллельных ветвей на ранних слоях, но тем не менее, меньшие нежели при позднем слиянии.

1.3.3 Бенчмарки

Бенчмарк общего понятия языка (General Language Understanding Evaluation, GLUE) представляет собой набор из девяти разнообразных задач на понимание естественного языка, предназначенных для оценки и сравнения производительности моделей в различных задачах обработки естественного языка (Natural Language Processing, NLP) [A. Wang (и др.), 2018]. Предоставляя стандартизированную основу, GLUE облегчает разработку моделей, которые хорошо обобщаются при решении различных задач, способствуя прогрессу в создании надежных и универсальных систем понимания языка.

Бенчмарк GLUE представляет собой набор ресурсов для обучения, оценки и анализа систем понимания естественного языка. GLUE состоит из:

- **Набор из девяти задач**, на понимание языка по предложениям или парам предложений, созданный на основе существующих наборов данных и отобранный таким образом, чтобы охватить разнообразный диапазон размеров наборов данных, жанров текстов и степеней сложности.
- **Диагностический набор данных**, предназначенный для оценки и анализа работы

моделей в отношении широкого спектра лингвистических явлений, встречающихся в естественном языке;

- **Публичная таблица лидеров** для отслеживания результатов выполнения эталона и приборная панель для визуализации работы моделей на диагностическом наборе.

Ниже представлена таблица, иллюстрирующая ключевые характеристики датасетов, используемых в GLUE:

Таблица 1: Обзор датасетов, входящих в бенчмарк GLUE.

Название	Задача	Источник	Размер	Метрика
CoLA	Классификация по одному предложению	[Dudy (и др.), 2018]	~8 500	Коэффициент корреляции Мэттьюса
SST-2	Бинарная классификация по одному предложению	[Socher (и др.), 2013]	~67 000	Точность
MRPC	Идентификация перефразирования	[Dolan, Brockett, 2005]	~3 700	Точность, F1
STS-B	Семантическое текстовое сходство (Регрессия)	[Cer (и др.), 2017]	~7 000	Корреляция Пирсона/Спирмена
QQP	Обнаружение повторяющихся вопросов (пары вопросов Quora)	[Z. Chen (и др.), 2017]	~364 000	Точность, F1
MNLI	Многожанровый вывод на естественном языке	[A. Williams, Nangia, Bowman, 2018]	~393 000	Точность
QNLI	Задача вывода	[Rajpurkar (и др.), 2016]	~105 000	Точность
RTE	Распознавание текстовых фрагментов	[Bentivogli (и др.), 2009]	~2,500	Точность
WNLI	Задача вывода	[Levesque, Davis, Morgenstern, 2012]	634	Точность

В целом, бенчмарк GLUE представляет собой надежную основу для оценки как стандартных, так и адаптированных к конкретной области моделей NLP. Его комплексный дизайн и включение различных лингвистических задач позволяют провести тонкий анализ возможностей модели. После этого обзора будет рассмотрен эталон FLUE, предназначенный для оценки моделей в финансовом контексте, будет рассмотрен для дальнейшего дополнения оценки доменно-адаптивных стратегий предварительного обучения.

Бенчмарк Financial Language Understanding Evaluation (FLUE) представляет собой аналог бенчмарка GLUE, но специализированный под домен финансов [Shah (и др.), 2022]. Данный бенчмарк был создан совсем недавно на основе 5 разнообразных датасетов. Его разработка обусловлена необходимостью оценки моделей, способных эффективно обрабатывать финансовый текст, поскольку стандартные универсальные наборы данных зачастую не отражают специфические особенности финансовой лексики и задач, присущих данной области.

Таблица 2: Обзор датасетов, входящих в бенчмарк FLUE.

Название	Задача	Источник	Рамер (Обучающая/Валидационная/Тестовая)	Метрика	Лицензия
FPB	Классификация тональности	[Malo (и др.), 2014]	3488 / 388 / 969	Точность	CC BY-SA 3.0
FiQA SA	Анализ тональности (Регрессия)	[Macedo (и др.), 2018; Shah (и др.), 2022]	822 / 117 / 234	MSE	Публичный
NHC	Классификация заголовков новостей	[Sinha, Khandait, 2021]	7989 / 1141 / 2282	Усредненная F1	CC BY-SA 3.0
FinNER	Распознавание именованных сущностей	[Alvarado, Verspoor, Baldwin, 2015]	932 / 232 / 302	F1	CC BY-SA 3.0
FinSBD3	Обнаружение границ структуры	[Au, Ait-Azzi, Kang, 2021]	460 / 165 / 131	F1	CC BY-SA 3.0
FiQA QA	Ответы на вопросы	[Macedo (и др.), 2018; Shah (и др.), 2022]	5676 / 631 / 333	nDCG, MRR	Публичный

FLUE охватывает 5 различных финансовых задач, что позволяет проводить комплексную оценку качества моделей на разнообразных аспектах финансового языка. Статистика, представленная в Таблице 2, демонстрирует величину и разнообразие датасетов. При этом все датасеты, входящие в состав FLUE, характеризуются низким уровнем этических рисков и не содержат конфиденциальной информации ни о каких организациях или отдельных

лицах. Дополнительно, для включения каждого датасета в бенчмарк был получен соответствующий запрос на согласие авторов, что подчеркивает его легитимность и корректность с точки зрения этики.

Возникновение бенчмарка FLUE продиктовано потребностью стандартизировать оценку моделей в области финансового понимания языка. Финансовая сфера предъявляет особые требования к обработке текстовых данных: высокая терминологическая сложность, динамичность рынка, специфические задачи (например, анализ тональности новостных заголовков, извлечение информации и другие). Именно эти факторы способствуют формированию разнородного набора задач, объединённых в FLUE, что позволяет всесторонне оценивать эффективность различных моделей. Таким образом, FLUE служит важным инструментом для исследователей, способствуя объективному сравнению моделей и выявлению областей для дальнейшего совершенствования подходов в финансовом NLP.

ГЛАВА 2. ПРАКТИЧЕСКОЕ РЕШЕНИЕ

2.1 Ограничение

Перед началом практической реализации на основе предварительного семантического анализа и структурного проектирования была проведена всесторонняя аналитическая оценка, призванная выявить и формализовать технические ограничения, влияющие на архитектуру FinABYSS. В результате было выделено пять ключевых проблемных областей, из которых трем удалось предложить устойчивые решения, а две менее критичные оставлены для последующих исследований.

Дедубликация текстов. Дедубликация является необходимым элементом конвейера текстового анализа тональности, поскольку финансовые сигналы, распространяющиеся с задержкой, могут неоднократно попадать в корпус, и их многократное учёт искажает результаты прогнозирования. Так, первичная публикация о нарушениях в выдаче ипотечных кредитов в сентябре 2008 г. формировалась сильный негативный сентимент в момент возникновения события, в то время как её републикации через годы, сопровождающие реtроспективные обзоры, не оказывают аналогичного влияния на цену активов. Подобная дискретность временного контекста не учитывается при классической текстовой дедубликации, основанной на лингвистическом или синтаксическом сходстве.

Две новости могут иметь практически полное сходство текста и при этом нести кардинально различные смысловые оттенки. Теоретический пример — если бы в исходной статье о штрафе Citi Group (Раздел 3.2.4), «79 млн» было бы ошибочно заменено на «79 тыс.», — это бы повлекло за собой принципиально разный сентимент и потенциально противоположное торговое решение.

Особая категория дублирующих материалов представлена коррекциями на платформе Yahoo! Finance. Такие публикации начинаются с маркера «/CORRECTION/», затем в статье указывают на исправленные фрагменты и повторяют основной текст почти дословно. Линейная дедубликация на уровне строк или *n*-грамм в таких случаях удаляет последнюю версию, что нарушает логику потокового анализа.

Таким образом, дедубликация в финансовом медиапотоке должна отвечать двум требованиям: во-первых, различать семантически эквивалентные, но контекстуально разных по времени публикации; во-вторых, корректно обрабатывать «correction»-версии, извлекая непересекающуюся семантику, если она значимо отличается от предыдущей версии. Каждый документ следует считать уникальным, если он содержит новую информацию, даже при высокой доле текстового совпадения.

Поскольку в задаче отсутствует исходная разметка, требуется формальная постановка проблемы семантической дедубликации, выходящей за рамки простого текстового сравнения. Семантическое пространство текстов, в отличие от лексического, непрерывно и неограниченно, и поэтому для проверки уникальности одного документа недостаточно по-

парного сопоставления строк или текстов посредством косинусного расстояния: необходимо оперировать объемами векторных представлений и учитывать распределение идей в более широком контексте.

Полное описание математической формулировки задачи и предложенного аналитического решения представлено в разделе 3.3.1. Именно там даётся строгое определение семантической уникальности, приводится алгоритм обнаружения близких по смыслу, но различающихся по информационной ценности документов.

Вычислительные ресурсы. Для обучения моделей на больших данных требуется крайне мощная вычислительная инфраструктура. Так, одна модель FinBERT обучалась на четырех NVIDIA Tesla P100 в течение двух суток на 5 миллиардах токенов [Y. Yang, UY, A. Huang, 2020]. Стоит уточнить, что здесь приводится время обучения лишь одной модели, а в процессе экспериментирования обычно обучается еще несколько моделей. Говоря о ModernBERT — он был обучен на восьми NVIDIA H100 за 10 суток, а общий объем корпуса составил около 3 триллионов токенов [Warner (и др.), 2024].

В контексте нашего исследования в качестве базовой модели используется именно ModernBERT, соответственно ориентироваться стоит именно на него. В ходе исследования был собран корпус из приблизительно 1 миллиарда токенов, что хоть и меньше корпуса ModernBERT, однако хватило бы для доменно-адаптированного предобучения, то есть дополнительной четвертой стадии предобучения [Gururangan (и др.), 2020; Warner (и др.), 2024].

Обучение даже базовой модели ModernBERT требует крайне мощного вычислительного оборудования. Минимальные требования NVIDIA Tesla T4 из серверного сегмента и NVIDIA RTX 3090 — из пользовательского [Warner (и др.), 2024]. Тем не менее, в контексте исследования была доступна только NVIDIA RTX 3060, на которой было бы невозможно обчинить ModernBERT. Поэтому, из-за труднодоступности вычислительных ресурсов, в контексте исследования используется базовая версия ModernBERT тонко настроенная для задачи кластеризации векторных представлений.

Ограничность данных. Отсутствие репрезентативного текстового корпуса в финансовом домене стало одной из ключевых проблем данного исследования. При попытках найти открытые наборы данных на платформах HuggingFace, Kaggle, GitHub и аналогичных ресурсах выяснилось, что доступные финансовые корпуса либо удалены по причинам нарушения авторских прав, либо закрыты для общего пользования, либо представляют собой слишком короткие фрагменты текста, непригодные для тематического моделирования и сентимент-анализа [Daudert, 2022; Macedo (и др.), 2018; Malo (и др.), 2014; Wiebe, Wilson, Cardie, 2005; Xing (и др.), 2020]. Кроме того, многие из оставшихся наборов предназначены исключительно для тонкой настройки нейросетевых моделей (fine-tuning), но не содержат необходимого объема или разнообразия данных.

Самостоятельный сбор финансовых новостных статей осложняется тем, что ключевые

источники разрознены и зачастую противодействуют массовому скрейпингу. Лишь немногие провайдеры предоставляют доступ к платным и дорогостоящим API (например, X, ранее Twitter), где цены на исторические данные могут исчисляться тысячами долларов. При этом новостные площадки обычно фокусируются на узких сегментах отраслевого контента, что делает корпус с одного источника предвзятым и фрагментарным.

С финансовой точки зрения важна не только широта охвата, но и наличие достоверных метаданных: временных меток, информации об авторе и других. Многие сайты лишены открытых архивов, предоставляют лишь RSS-ленты, а часть ресурсов попросту не размещает даты публикации в HTML, что делает автоматизированный парсинг невозможным без глубокого анализа и регулярного обновления скриптов. Различные структуры HTML-шаблонов и динамические генераторы контента (JavaScript-рендеринг) увеличивают сложность разработки конвейеров извлечения данных.

Таким образом, попытка собрать по-настоящему репрезентативный корпус потребует интеграции множества источников с разнородными схемами сайтов, что приводит к высокой технической нагрузке. Ориентация на один-единственный поставщик данных рискует ввести дисбаланс в тематических и региональных представлениях финансовых событий. При этом существующие платные альтернативы (например, коммерческие датасеты News от Bright Data⁷) оцениваются в несколько тысяч долларов за объём, идентичный с собранным ходе исследования корпусом, что выходит за рамки бюджетных и исследовательских ограничений проекта.

Следовательно, отсутствие открытых, крупных и однородных финансовых корпусов остаётся серьёзным препятствием для создания масштабируемых и надёжных моделей тематического моделирования и анализа тональности в сфере финансов. В дальнейшем разделе 2.2.2 описывается выбранный компромиссный подход к сбору и агрегации данных, учитывающий выявленные ограничения и требования к качеству корпуса.

Ограниченност контекстного окна. Базовая версия модели ModernBERT обеспечивает контекстное окно в 8 192 токена, что существенно превосходит традиционные BERT-подобные модели с ограничением в 512 токенов [Devlin (и др.), 2019; Warner (и др.), 2024]. Однако даже при таком расширенном ресурсе финансовые документы формата 10-K и 10-Q, зачастую превышающие десятки страниц, не умещаются полностью. Хотя существуют техники скользящего окна или фрагментирования текста на перекрывающиеся чанки, они выходят за рамки исходной задачи оценки «из коробки» возможностей современных LLM. В контексте данного исследования было принято решение ограничиться анализом новостных статей, размеры которых вписываются в 8 192 токена, а многостраничные аналитические отчёты не рассматривать. Это позволило сохранить фокус на сопоставлении эмбеддинговых и тематических моделей без усложнения препроцессинга за счет агрегирования отрывков длинных документов.

⁷URL: <https://brightdata.com/products/datasets/news>

Нерелевантность существующих бенчмарков. Сравнение производительности модели ModernBERT и её доменно-адаптированной версии на общепринятом бенчмарке FLUE кажется естественным шагом для оценивания прогресса. Однако датасеты FLUE состоят преимущественно из коротких фрагментов (до 512 токенов), предназначенных для типовых заданий понимания текста [Shah (и др.), 2022]. Поскольку эти датасеты заточены под 512-токенное окно, они не отражают преимущества расширенного контекста ModernBERT и, наоборот, будут занижать его итоговые показатели. Таким образом, использование FLUE в текущем исследовании приведёт к искаженному восприятию качеств модели: задачи на коротких текстах не демонстрируют её способность захватывать долгосрочные зависимости и синтезировать информацию из больших объёмов данных.

Обе проблемы — ограниченность контекстного окна при анализе длинных документов и нерепрезентативность стандартных 512-токеновых бенчмарков — диктуют необходимость создания специализированных методик предобработки и валидации, адаптированных под финансовый домен. За исключением данных двух проблем, все остальные были разрешены и их решения предлагаются в дальнейших разделах работы.

2.2 Работа с данными

2.2.1 Постановка задачи

Одной из ключевых задач исследования являлся сбор данных. Как отмечалось, корпус формировался с учётом требований универсальности и перспективы дальнейшего использования в смежных исследованиях.

Требования к данным. В качестве основного источника качественных данных выбраны тексты. В финансовом домене наиболее ёмкими и значимыми по влиянию на стоимость активов являются:

- новостные статьи;
- посты в социальных сетях;
- официальные отчёты (годовые, квартальные, стратегические);
- пресс-релизы;
- аналитические обзоры и статьи;
- транскрипты интервью, конференций и веб-трансляций публичных компаний.

Предыдущие исследования подтвердили эффективность этих типов текстов. Так, FinBERTv3 [A. Huang, H. Wang, Y. Yang, 2023; Y. Yang, UY, A. Huang, 2020] обучался на официальных отчётах и аналитических статьях, а его вторая версия включала пресс-релизы [Z. Liu (и др.), 2020]. Другие модели тонко настраивались под задачу анализ тональности

на заголовках новостей и публикациях из социальных сетей [Araci, 2019]. Тем не менее в рамках нашего исследования целенаправленно обрабатываются полные тексты без их фрагментации. Официальные отчёты часто превышают 8192 токена, которые способен обработать ModernBERT, что затрудняет их интеграцию, тогда как слишком короткие форматы (посты в социальных сетях) не раскрывают преимущества работы с длинным контекстом.

Кроме того, нет единого источника, агрегирующего все перечисленные типы текстов. Ввиду чего, обращая внимание на ограниченность ресурсов, для первичной фокусировки были выбраны наиболее значимые виды:

- новостные статьи;
- аналитические обзоры и статьи;
- пресс-релизы;
- транскрипты финансовых мероприятий.

В будущем планируется расширить корпус другими категориями контента.

Для обеспечения универсальности и удобства дальнейшего использования для каждого текста были собраны следующие метаданные:

- заголовок (например, «Covestro board enters formal talks on \$12 billion ADNOC approach»);
- источник (правообладатель), например, Reuters, Simply Wall St., PR Newswire, @ilyasut, Max Gottich;
- ресурс, на котором был опубликован текст (например, Twitter, Yahoo! Finance, Reddit, Seeking Alpha);
- дата и время публикации по UTC с точностью до секунды (например, 2024-09-01T01:48:13);
- тематические метки от автора (например, ["M&A "Cryptocurrency "Tech"]);
- список тикеров, ассоциированных автором с текстом (например, ["9626.HK "BILI"]).

В заключении был установлен оптимальный период для сбора данных. Нижняя и верхняя границы дат подбирались исходя из предположения о том, что данный корпус в дальнейшем будет использоваться для обучения модели предсказания стоимости, что обязывает синхронизировать первоначальные знания ModernBERT с теми, на которых он позднее будет тонко настраиваться. По причине того, что в публикации ModernBERT не раскрываются данные на которых была обучена модель, невозможно точно судить о том, за какой период были

взяты данные [Warner (и др.), 2024]. Посему в нашем исследовании, ориентируясь на дату публикации ModernBERT [Там же] и классического соотношения обучающей, валидационной и тестовой выборок как 75/15/10 был принят диапазон дат с 17.09.2023 по 18.03.2025, то есть 548 дней, из которых 374 рабочие по календарю США.

Требования к источникам. Источники данных должны быть открытыми, бесплатными, англоязычными и авторитетными, поскольку широкий охват и оперативность публикаций напрямую влияют на реакцию рынков. В числе рассматриваемых источников оказались:

- Новостные издания: Bloomberg⁸, The New York Times⁹, Reuters¹⁰, etc;
- Аналитические платформы: Seeking Alpha¹¹, TradingView¹²;
- Официальные сайты: корпоративные и правительственные порталы.

При анализе более 50 корпоративных и более 100 правительственные ресурсов установлено, что благодаря наличию RSS-лентам, пресс-релизы централизованно агрегируются через PR Newswire¹³ и GlobeNewswire¹⁴. Имеется и ряд других автоматических агрегаторов, например, Business Wire¹⁵), ACCESS Newswire и других, однако эмпирически было выявлено, что среди подобных агрегаторов доминируют PR Newswire и GlobeNewswire. Тем не менее, данные агрегаторы не дают разнообразия жанров.

Из новостных изданий по скорости реакции и охвату рынков оптимальной признана служба Reuters. Также было изучено несколько более узких, но качественных источников (например, The Information¹⁶, Epoch AI¹⁷), однако они содержат слишком узкоспециализированную информацию, в то время как более традиционные источники по типу Bloomberg, Wall Street Journal¹⁸, The Economist¹⁹ и другие покрывают более обширную часть рынков, однако технически ограничены.

Среди аналитических платформ Seeking Alpha уступает из-за большого объёма платного контента, а TradingView не обеспечивает доступ к историческим публикациям.

Таким образом, основными первичными источниками корпуса стали PR Newswire и Reuters. Однако, чтобы снизить риск систематической предвзятости и обеспечить дополнительный охват, принято решение обогатить корпус другими издательствами, но в силу их разрозненности, отсутствия API, а также зачастую и полной информации о статье (включая

⁸URL: <https://www.bloomberg.com/>

⁹URL: <https://www.nytimes.com/>

¹⁰URL: <https://www.reuters.com/>

¹¹URL: <https://seekingalpha.com/>

¹²URL: <https://tradingview.com/>

¹³URL: <https://www.cision.com/>

¹⁴URL: <https://www.globenewswire.com/>

¹⁵URL: <https://www.businesswire.com/>

¹⁶URL: <https://www.theinformation.com/>

¹⁷URL: <https://epoch.ai/>

¹⁸URL: <https://www.wsj.com/>

¹⁹URL: <https://www.economist.com/>

время публикации с точностью о секунды), агрегировать и синхронизировать данные не представляется возможным. Так, дополнительно были рассмотрены агрегаторы Google Finance²⁰, Yahoo! Finance²¹ и FinURLs²².

В итоге предпочтение отдано Yahoo! Finance благодаря унифицированной схеме сайта и широкому отражению разнообразных источников, тогда как FinURLs использует переадресацию на сайты источников, каждый из которых имеет свою собственную схему сайта. С другой стороны, Google Finance больше похож на Yahoo! Finance, однако не предоставляет возможности для сбора исторических данных.

2.2.2 Сбор данных

Предваряя сбор данных, был проведен поиск и анализ существующих открытых решений для сбора данных с Yahoo! Finance. В ходе анализа выявлено пять наиболее подходящих инструментов. При этом два — 'yahoocquery' и 'yahoo-stock-api' — не поддерживают извлечение статей, ещё два — 'yahoo_fin' и 'fin-news' — оказались заброшенными и не работают должным образом, а 'yfinance' предоставляют доступ исключительно к последним двадцати новостным материалам в реальном времени. В связи с этим была разработана собственная программа-парсер на языке Python. Архитектура парсера состоит из двух ключевых этапов.

1. Сбор ссылок. Рекурсивный обход официальной карты сайта позволяет собрать ссылки на статьи за заданный период. Каждая «страница дня» содержит 50 ссылок на новости и указатель на следующую страницу, причём доступ к странице n возможен только через страницу $n - 1$, что являлось узким местом процесса парсинга, имплементируя связанный список с высокой сетевой задержкой.
2. Извлечение содержимого. По собранным URL-адресам производится парсинг текстов статей. Важно отметить, что, как и для обучения классической модели BERT, таблицы и изображения не обрабатывались [Devlin (и др.), 2019].

В ходе парсинга было выявлено множество ограничений, которые затем были учтены при разработке архитектуры программы-парсера:

- IP-блокировки и cookies. Инфраструктура Yahoo! Finance ограничена 14 параллельными запросами с одного IP с интервалом не менее 4 секунд; при нарушении интервалов ответы содержат коды 404, 429 или 200 с пустым телом. Тем не менее, как оказалось, даже с выполнением всех инфраструктурных требований, данная проблема все равно может возникать. Для обхода блокировок использовался пул из 50 прокси-серверов, а неудачные запросы автоматически обрабатывались в дополнительной итерации.

²⁰URL: <https://www.google.com/finance/>

²¹URL: <https://finance.yahoo.com/>

²²URL: <https://finurls.com/>

- Региональные ограничения. Из разных стран одни и те же ссылки могут быть недоступны или работать некорректно.
- Технические неполадки. Встречались ссылки с переадресацией на внешние источники, битые и платные ссылки, которые игнорировались в ходе сбора корпуса.

Для ускорения обработки больших объёмов данных была использована библиотека реализованная на С — 'selectolax' — которая в 30 раз быстрее 'BeautifulSoup' и в 5 раз — 'lxml'.

В результате на первом этапе было собрано 1 362 103 ссылки, из которых 1 360 761 принадлежали домену Yahoo! Finance. Успешно спаршено 1 304 717 статей благодаря модульной архитектуре парсера, использованию множества прокси и повторным итерациям. Итоговый корпус составил 15,4 ГБ в формате CSV и 8,6 ГБ в более оптимизированном формате Parquet.

Итоговый класс реализующий программу-парсер, находится в официальном репозитории проекта, называясь 'YahooFinanceParser'²³.

2.2.3 Анализ данных

Перед началом этапа предобработки было принято решение провести всесторонний анализ собранного корпуса новостных статей. Такой предварительный анализ позволил не только выявить характерные особенности данных, но и сформировать базу для последующей автоматизации очистки и структурирования текстов. Более того, результаты анализа, в том числе, сказались на качестве обученной модели.

Локальный анализ охватывал точечное изучение различных срезов корпуса с целью обнаружения паттернов, характерных для нерепрезентативных или «шумовых» статей. При этом были выделены ключевые сигналы, такие как характерные ключевые слова в заголовках и первых абзацах, которые позволяют автоматически отсеять нежелательные тексты. Кроме того, локальное исследование выявило потенциальные правила для удаления маркетинговых фрагментов, метаданных и прочих артефактов, негативно влияющих на качество данных. Все полученные правила были затем формализованы (см. Раздел 2.2.4).

Глобальный анализ направлен на изучение центральных тенденций корпуса через описательные статистики и анализ различных репрезентаций данных — как метаданных, так и непосредственно текстовой информации. Такой подход позволил оценить распределение ключевых характеристик, выявить сезонные и тематические закономерности, а также подготовить агрегированные результаты, служащие базисом для дальнейшего улучшения методики предобработки.

Ниже представлены агрегированные результаты глобального анализа, которые в совокупности с локальными выводами позволяют глубже понять природу собранного датасета и

²³URL: <https://github.com/denisalpino/FinABYSS>

определить направления для его дальнейшей оптимизации.

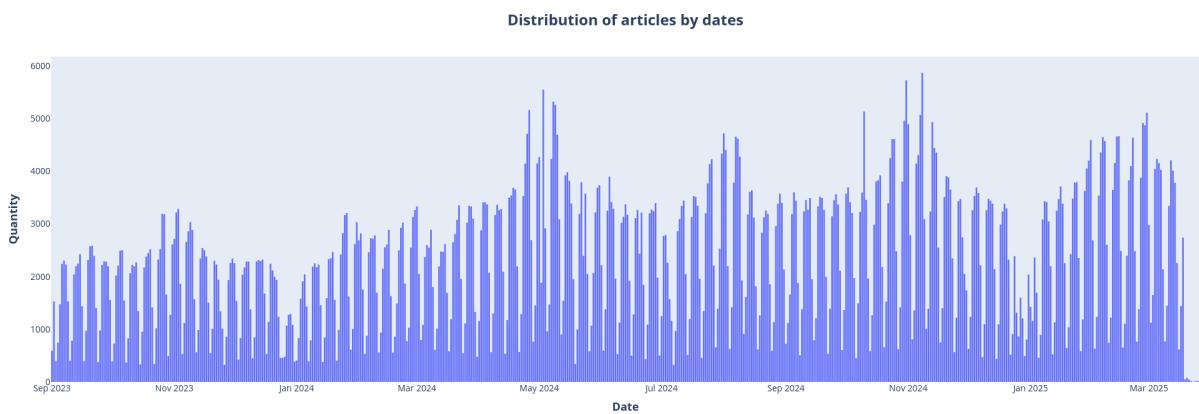


Рис. 1: Распределение содержащихся в корпусе публикаций по датам.

Распределение публикаций по датам. На Рис. 1 видно, что количество публикаций колеблется по дням с определённой периодичностью. Детальное изучение показало, что минимумы приходятся на воскресные и праздничные дни, когда публикуется меньше финансовых новостей. Это естественно отражает специфику рынка: в выходные и праздничные дни деловая активность снижается, поэтому и публикаций становится меньше.

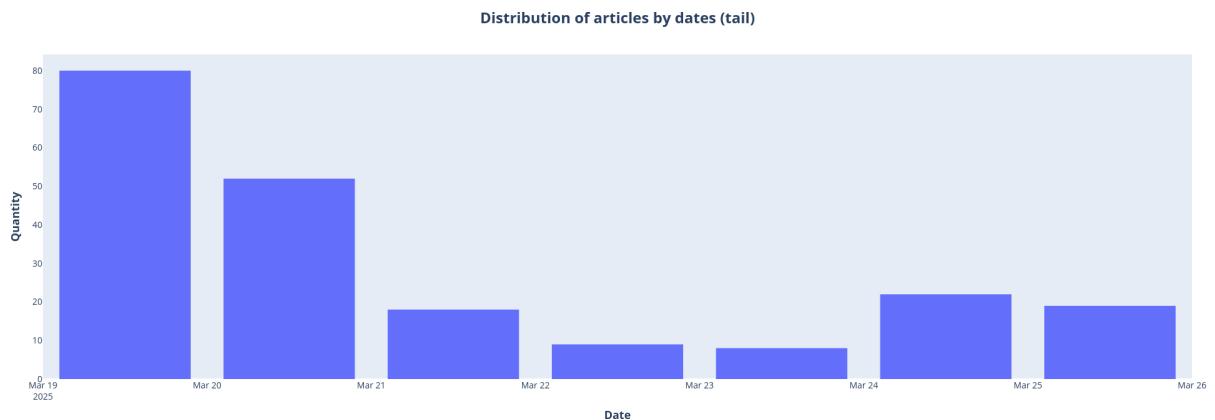


Рис. 2: Распределение содержащихся в корпусе публикаций по датам (хвост).

При этом часть статей, по формальным признакам, выходит за границы периода сбора (1 сентября 2023 года — 18 марта 2025 года). На Рисунок 2 показаны эти «хвостовые» публикации, численность которых слегка превышает 200 единиц. Более детальный анализ установил, что публикации фактически были сделаны в установленный интервал, однако их содержание редактировали или дополняли позже. В результате дата и время публикации на соответствующем сайте обновились, а старая версия (со старой датой) утрачена. Если бы ссылки парсились не через неделю, а спустя несколько недель и более, таких случаев было бы несколько больше.

С точки зрения краткосрочного прогнозирования рынка это обстоятельство может привести к искажению временных меток: часть статей будто бы опубликована позже, чем на самом деле. Поэтому датасет может оказаться менее эффективен в краткосрочных исследованиях, чем в средне- и долгосрочных (где сдвиг в пару дней уже не так принципиален). Тем не менее, для данной работы это не играет критической роли, поскольку модель всё равно опирается лишь на текст статьи и не учитывает точные временные метки публикаций.

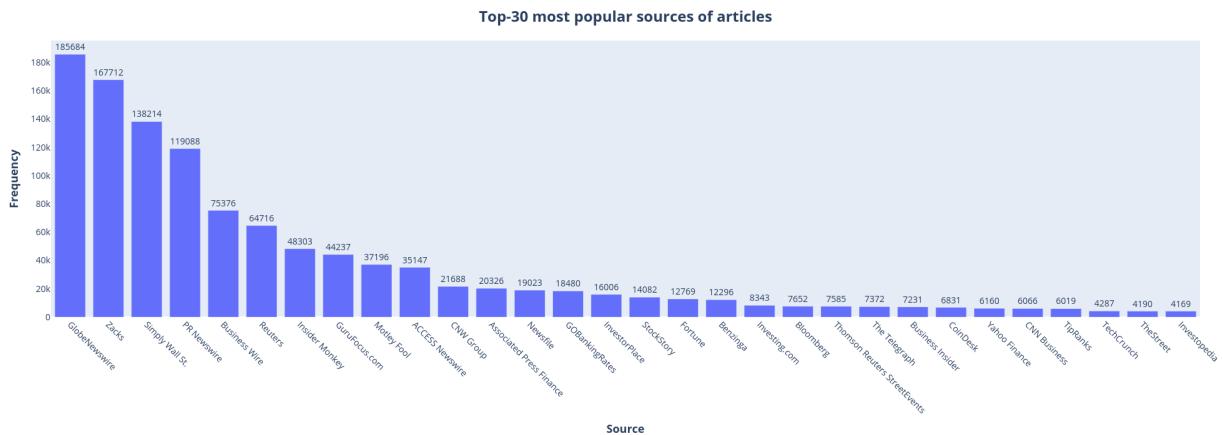


Рис. 3: 30 наиболее популярных источников финансовых публикаций в корпусе.

Новостные источники. Рисунок 3 иллюстрирует распределение публикаций по 30 наиболее частым источникам. Анализ показал, что преобладающая доля статей (потенциально 69.2%) была опубликована полу-автоматизированными агрегаторами: GlobeNewswire, Zacks, Simply Wall St., PR Newswire, Business Wire, GuruFocus.com, Motley Fool и др. Они фокусируются на автоматическом сборе ключевых данных с различных ресурсов (регуляторы, официальные сайты компаний и т.п.), публикуют пресс-релизы, краткие обзоры отчётов и приглашения на корпоративные мероприятия.

В топ-15 источников лишь некоторые можно условно считать «традиционными» новостными ресурсами, такими как Reuters, Insider Monkey, CNW Group, Associated Press Finance, InvestorPlace. При этом за пределами топ-30 превалируют именно классические издания, которые в основном публикуют авторские статьи. В реальности же размытая грань между авторскими и полу-автоматическими материалами усложняет попытки чёткого разграничения.

По приблизительной оценке, из 1 300 000 статей около 900 000 (69.2%) являются полу-автоматическими. Это важный фактор для обучения языковой модели, поскольку:

1. Качество таких материалов нередко ниже: тексты содержат артефакты, ломаются разметки и вставляются некорректные символы.
2. Их объём велик, что, с одной стороны, даёт большую семплирующую способность, но с другой — затрудняет очистку и нормализацию без потери значимой информации.

Тем не менее, даже «нечистые» тексты из агрегаторов несут полезную информацию о финансовом рынке и компаниях. Однако, крайне важно разработать корректные правила очистки и предобработки (см. Раздел 2.2.4) текстов, чтобы не повредить их семантическую целостность.

Кроме того, и, что более важно, эти полу-автоматические тексты составляют примерно такое же общее количество токенов, как и «авторские» статьи (30.8%), несмотря на их количественное доминирование. Следовательно, при правильной обработке данная группа полу-автоматических статей может дать значимый вклад в обучение языковой модели, не размывая значимость «авторских» текстов.

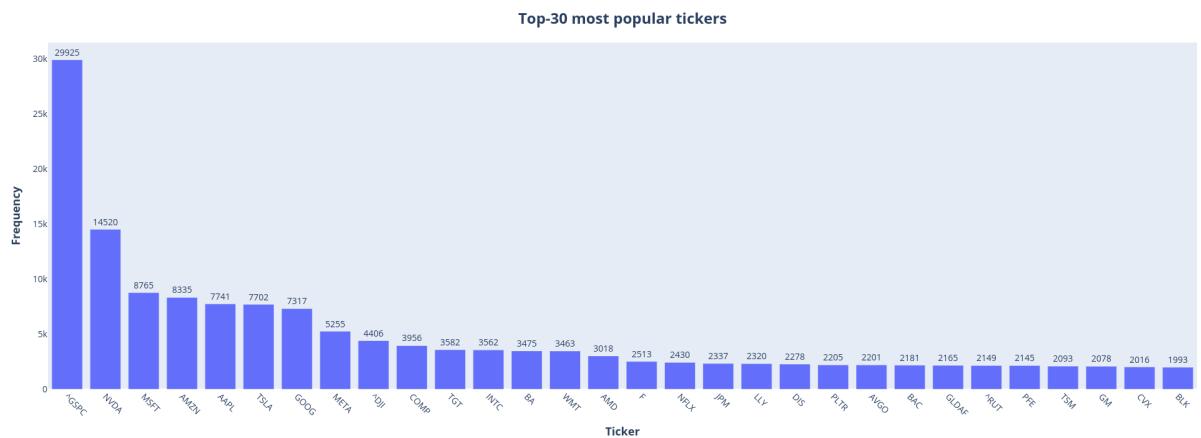


Рис. 4: 30 наиболее частых тикеров в датасете.

Анализ тикеров. На Рисунке 4 представлено распределение публикаций по 30 наиболее часто встречающимся тикерам. Лидером является индекс S&P 500, однако в выборку также попали Dow Jones и Russell 2000. Примечательно, что в топ-10 оказались главным образом IT-компаний, причём с существенным отрывом лидирует Nvidia.

При этом около 574 000 (44,2%) публикаций не содержат вообще никаких тикеров в описании статьи, которое находится в шапке страницы. Более того, даже когда тикеры присутствуют, они могут отображать не все фактически упомянутые в статье компании или индексы. Это говорит о том, что данный столбец в датасете, пусть и достаточно репрезентативен, но не даёт полного покрытия всех потенциальных тикеров, и часть новостей формально «выпадает» из рассмотрения. Следовательно, для задач, выходящих за рамки данной исследовательской работы стоит создать словарь терминов и наименований, относящихся к каждому конкретному тикеру, а затем алгоритмически дополнить столбец с тикерами, используя соответствующие тексты.



Рис. 5: Облако слов, составленное по всем текстам, содержащимся в собранном корпусе.

Качество текста. На Рисунке 5 представлено облако слов, построенное по всему корпусу собранных текстов. Из этой визуализации можно сделать следующие ключевые выводы:

- 1. Репрезентативность данных.** Облако демонстрирует широкий спектр финансовых терминов, что указывает на достаточную репрезентативность датасета в финансовой тематике. Это свидетельствует о том, что материал охватывает разнообразные аспекты рыночной деятельности и экономических событий.
- 2. Специфичность финансовой терминологии.** Распределение частот финансовых терминов существенно отличается от того, что наблюдается в популярных корпусах для обучения языковых моделей (например, English Wikipedia или BookCorpus). Это различие обуславливает необходимость проведения доменной адаптации DAPT для корректного обучения модели на специфических данных финансовой тематики.
- 3. Уровень шума и наличие нерелевантной информации.** Облако слов включает такие элементы, как «Zacks», «click», «please», «free» и «source». Это указывает на значительное присутствие шумовых, рекламных или автоматически генерированных фрагментов, что требует разработки специальных методов очистки данных без нарушения семантической целостности текстов.

Дополнительно можно отметить, что выявленные шум и рассеянность терминов могут негативно сказаться на качестве downstream-задач, таких как классификация или извлечение эмбеддингов, если данные не будут корректно обработаны на этапе предобработки.

Вывод. Собранный датасет новостей характеризуется некоторыми особенностями. Во-первых, наблюдается ярко выраженная сезонность публикаций — минимумы приходятся на выходные и праздничные дни, также зафиксированы так называемые «хвостовые» статьи. Во-вторых, анализ источников показывает, что около 69% текстов поступают от полу-автоматизированных агрегаторов, что может усложнить процесс очистки данных, поскольку такие источники нередко порождают тексты с нарушенной разметкой, встроенными артефактами и нерелевантной информацией. Наконец, выяснено, что датасет обладает высокой вариативностью финансовой терминологии, но в то же время содержит значительный уровень шума, что в совокупности подтверждает необходимость проведения доменной адаптации (DAPT) и разработки методов для эффективной очистки текстов.

С одной стороны, выявленные особенности (временные сдвиги, шум, доминирование полу-автоматизированных источников) могут снижать пригодность датасета для краткосрочного прогнозирования или задач, требующих строгой временной разметки. С другой стороны, для задач, ориентированных на семантическое содержание текста, данные проблемы не оказывают критического влияния. Надлежащая предобработка, включающая очистку текстов и устранение нерелевантных элементов, позволит существенно повысить качество обучаемой модели и расширить ее возможности для обобщения на различные типы публикаций.

2.2.4 Предобработка данных

После проведения локального и глобального анализов текстовых данных, в ходе которых были выявлены как шумовые паттерны внутри отдельных документов, так и системные аномалии, свидетельствующие о нерепрезентативности некоторых публикаций в корпусе в целом, был спроектирован многоэтапный потоковый пайплайн предобработки на лету с использованием чанков по 100 000 публикаций. Ключевым ограничением при разработке решения стала ограниченность оперативной памяти: при объёме исходного корпуса порядка 15 ГБ стандартные операции фильтрации и полного прохода по текстам требовали выделения значительного константного буфера памяти. Это породило необходимость адаптации алгоритма к потоковой обработке порциями фиксированного размера.

В ходе первого этапа каждый чанк загружался из хранилища и сразу подвергался первичной фильтрации: документы, содержащие менее 100 символов, отбраковывались как недостаточно информативные. Эмпирически установленный порог в 100 символов оказался достаточен для устранения «пустых» артефактов, возникающих из-за нестабильности разметки на стороне источников. Так, на популярных площадках вроде Yahoo! Finance текст порой оказываются зашитым внутри нетипичных HTML-тегов — например, в ‘`<tbody>`’ вместо привычного ‘`<p>`’ — что приводит к появлению почти полностью пробельных записей или маркетинговых вставок в конце документа.

Следующим шагом реализован фильтр по заголовку: на его основе формализованы 66 правил, охватывающих как общие шаблоны («Form 8», «Net Asset Value», «Holdings

in company»), так и более таргетированные для шумных 12 источников (например, в GlobeNewswire удалялись заголовки, содержащие «Declaration» или начинающиеся с «Key digital»). Такая селекция гарантировала удаление публикаций, состоящих преимущественно из табличных данных или кратких заполнителей.

После нормализации всех видов пробельных символов — слияния подряд идущих пробелов или переносов строк в единичный символ, замены неразрывных пробелов ‘xa0’ на стандартный пробел и унификации специальных знаков в тексте — применялись ещё 12 правил, направленных на отбраковку нерепрезентативных документов по содержанию основного тела публикации. В частности, всё, что начиналось с маркера «(Repeat)», а также приветственные шаблоны специфичных источников вроде «Dear madam, sir, please find hereunder the links» для GlobeNewswire, автоматически исключалось из дальнейшей обработки.

Далее текст очищался от контактных данных и типовых футеров: ссылки, адреса электронной почты, а также фразы, однозначно указывающие на конец документа («Forward-looking statements», «Contact Details»), либо нерепрезентативность абзаца («Source:», «See More:», «Sponsored:»). Кроме того, из начала публикаций удалялись стандартные вводные фразы вроде «The recommendations of Wall Street analysts» и «When deciding whether to buy, sell, or hold a stock, investors often rely on analyst recommendations». Всего в различных группах было сформировано 92 правила очистки.

Такое агрессивное очищение текстов обуславливается тем, что отфильтрованные паттерны встречаются настолько часто, что крайне негативноказываются на кластеризации искусственно завышенная метрику кластеризации. После проведенном агрессивном очищении корпуса, метрика качества действительно уменьшилась, однако кластеры стали гораздо более репрезентативными и основанными на семантике самой новости, а не ее источниках, маркетинговой и правовой информации в текстах.



Рис. 6: Облако слов, составленное по всем по текстам, содержащимся в предобработанном корпусе.

В качестве валидации качества очистки текстов можно обратиться к облаку слов корпуса после предобработки (Изображение 6), на котором отчетливо видно улучшение: отсутствуют такие слова, как 'forwardlooking', 'link', 'click', 'simply' (название источника Simply Wall St.) и другие.

Наконец, перед сохранением чанков происходила сортировка по дате публикации и удаление дубликатов на основе полного текста документа. С учётом описанных операций объём корпуса сократился с 15,4 ГБ в формате CSV (8,6 ГБ в Parquet) до 5,9 ГБ (2,0 ГБ в Parquet): из 1 304 717 исходных записей осталось 1 267 416, то есть удалено было всего 2,8% документов. Данный результат свидетельствует о высокой доле шумов в тексте в собранном датасете и подчёркивает эффективность многоступенчатого потокового подхода к предобработке, который сохраняет оперативную память и обеспечивает качественную очистку текстовых данных перед последующими этапами тематического моделирования и анализа.

2.3 Разработка моделей

2.3.1 Извлечение векторных представлений

После предобработки корпуса текстов и удаления фонового шума, для ускорения последующих этапов тематического моделирования, включая обучение моделей понижения размерности и кластеризации, были извлечены их эмбеддинги.

Базовая модель ModernBERT не оптимальна для этой задачи, поскольку её векторные представления оказываются чрезмерно разреженными. Высокая степень разреженности эмбеддингов ухудшает качество кластеризации, в частности методов, основанных на оценке плотностей (например, DBSCAN). Несмотря на более высокую устойчивость HDBSCAN

к различиям плотностей, разреженность всё равно негативно сказывается на результатах кластерного разбиения.

Чтобы устранить указанную проблему, обычно применяют тонкую настройку модели под задачу STS. При этом модель получает на вход пару текстов и возвращает оценку их сходства [Muennighoff (и др.), 2023], чаще всего вычисляемую через косинусное расстояние (Формула 6).

$$D_{\cos}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}. \quad (6)$$

Именно STS-модели демонстрируют наиболее плотные и информативные эмбеддинги.

Так, для базовой модели ModernBERT было рассмотрено 2 тонко настроенные модели — ‘modernbert-embed’ от Nomic AI [Nussbaum (и др.), 2024] и ‘gte-modernbert-base’ от Alibaba [Z. Li (и др.), 2023; X. Zhang (и др.), 2024]. Для оценки данных двух моделей был использован бенчмарк Massive Text Embedding Benchmark (MTEB) [Muennighoff (и др.), 2023], охватывающий восемь классов задач и 58 датасетов. По результатам:

- В задаче кластеризации на 12 датасетах ‘gte-modernbert-base’ превзошла ‘modernbert-embed’ на 1.5 процентных пункта (44.98% против 44.47%).
- В задаче STS на 10 датасетах их результаты близки (81.78% и 81.57% соответственно).
- В среднем по всем восьми задачам на 56 датасетах ‘gte-modernbert-base’ опережает ‘modernbert-embed’ на 1.76 процентных пункта (64.38% против 62.62%).

В связи с этим для извлечения эмбеддингов была выбрана модель gte-modernbert-base из библиотеки ‘sentence_transformers’ [Reimers, Gurevych, 2019]. При получении эмбеддингов для всего документа вместо токена [CLS] применялась более продвинутая техника — Mean Pooling, которая заключается в усреднении значений эмбеддингов по всем токенам последовательности (Формула 7).

$$\mathbf{h}_{\text{mean}} = \frac{1}{T} \sum_{t=1}^T \mathbf{h}_t. \quad (7)$$

где T — длина токенизированной последовательности, а \mathbf{h}_t — эмбеддинг t -го токена.

Для оптимизации скорости обучения моделей понижения размерности и кластеризации, эмбеддинги предварительно приводились к единичной L_2 -норме (Формула 8):

$$\widehat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \quad \|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^n u_i^2}. \quad (8)$$

Подобная предобработка позволяет пользоваться GPU-оптимизированной евклидовой метрикой расстояния (Формула 9) при обучении модели DR, не повторяя при этом на

каждой итерации оптимизации гиперпараметров вычисление L_2 -нормы, которое происходит при вычислении метрики косинусного расстояния.

$$D_2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2 = \sqrt{\sum_{i=1}^n (u_i - v_i)^2}. \quad (9)$$

Таким образом, после вычисления L_2 -нормы и евклидова расстояния, мы, на самом деле, в некотором смысле можно считать, что мы работаем с косинусным расстоянием, так как приведенная мера становится монотонно связанной с косинусным и отражает тот же порядок близости точек (Формула 10), однако все вычисления ускоряются за счет GPU-оптимизации.

$$D_2(\hat{\mathbf{u}}, \hat{\mathbf{v}}) = \|\hat{\mathbf{u}} - \hat{\mathbf{v}}\|_2 = \sqrt{2(1 - \hat{\mathbf{u}} \cdot \hat{\mathbf{v}})}. \quad (10)$$

Для построения и обучения алгоритмов понижения размерности и кластеризации была сформирована тренировочная выборка из 200 000 эмбеддингов и сопутствующих метаданных, что составляет приблизительно 16% от всего корпуса. Такой объем обучающей подвыборки был выбран, исходя из доступных вычислительных ресурсов. Так, 200 000 эмбеддингов использовались для подбора оптимальных гиперпараметров, а остальные 1 050 000 эмбеддингов были зарезервированы для этапов валидации и инференса. При этом, перед инференсом конвейер моделей понижения размерности и кластеризации были обучены на всем корпусе с линейным увеличением значений гиперпараметров для алгоритма HDBSCAN, выбор которого обуславливается в Секции 2.3.2.

Кроме того, при извлечении эмбеддингов была задействована смешанная точность (float16) и ускоренный механизм внимания FlashAttention [Dao (и др.), 2022], что существенно сократило требования к вычислительным ресурсам и время обработки.

2.3.2 UMAP и HDBSCAN

Итак, после формирования выборки эмбеддингов мы приступили к экспериментам с моделями DR и кластеризации, проводя их совместную оптимизацию. Такой подход обусловлен многокритериальным характером задачи: требуется не только сохранить структурные (глобальные) и локальные взаимосвязи из исходного 768-мерного пространства, но и обеспечить разделимость («кластерность») эмбеддингов в низкоразмерной проекции, пригодной для визуализации на плоскости.

В качестве ключевых условий было обозначено:

1. Сохранение кластерной структуры. Эмбеддинги после понижения размерности должны оставаться разделимыми, то есть сохранять группировку по смыслу и тематике.
2. Готовность к двумерной визуализации. Итоговое пространство должно быть пригодно для наглядного отображения на плоскости без потери интерпретируемости.

Для одновременного поиска оптимальных гиперпараметров выбора алгоритмов понижения размерности и методов кластеризации использовался единый мета-процесс.

В качестве критерия оптимальности был принят индекс DBCV, поскольку он не предполагает заранее заданной формы кластеров (в отличие от индексов, ориентированных на сферические или эллипсоидальные структуры) и эффективно оценивает плотностные методы кластеризации.

В качестве основы был выбран алгоритм HDBSCAN [Campello, Moulavi, Sander, 2013], отвечающий двум важным требованиям:

- Отсутствие предположений о форме кластеров. В отличие от K-Means HDBSCAN не предполагает, что кластеры являются гауссовыми сферами, что критически для представления тем.
- Иерархическая и плотностная природа. Позволяет выявлять как крупные тематические группы, так и мелкие, высококонцентрированные ниши.

Кроме того, благодаря реализации на GPU, HDBSCAN показал высокую скорость обработки как больших выборок, так и высокоразмерных данных.

В контексте алгоритма HDBSCAN для оптимизации рассматривается несколько ключевых гиперпараметров. Минимальное число соседей — количество точек в окрестности, необходимое для признания точки «ядром» кластера. Минимальный размер кластера — число наблюдений, определяющее порог для формирования кластера, что позволяет учитывать редкие, узкотематические группы.

В ходе пилотных экспериментов было обнаружено, что одновременно существуют плотные области («горячие темы») и редкие, но значимые по содержанию кластеры. При малом минимальный размер кластера удается захватывать редкие темы, однако возрастает число микрокластеров, разделение между которыми семантически не всегда оправдано.

Для смягчения этой проблемы рассматривалась техника объединения кластеров по порогу ϵ [Malzer, Baum, 2020], позволяющая консолидировать соседние микрокластеры в высококонцентрированных регионах. Однако такой подход усложняет инференс модели: при появлении новых наблюдений ϵ -объединение не может быть учтено, что вынуждает полностью переобучать модель.

Так как исследование ставит своим приоритетом практичность, было принято решение отказаться от использования ϵ , но использовать меньшие значения минимального размера кластера, принимая некоторую фрагментацию, но сохраняя возможность последующей интерпретации и агломерации кластеров на более высоком уровне иерархии.

Ещё один гиперпараметр — метод выделения кластеров — определяет, будут ли они формироваться на основе «избытка массы» или «листьев дерева». Практика показала, что второй вариант даёт более мелкозернистые и однородные группы, что и было избрано для окончательной настройки.

Таким образом, на финальной стадии оптимизации HDBSCAN оставлены лишь два подлежащих настройке параметра: минимальное число соседей и минимальный размер кластера.

Также стоит отметить, что в архитектуре, описанной в Разделе 3.3, планируется фиксированное количество кластеров для статичного числа экспертов, поэтому в качестве базовой реализации используется именно HDBSCAN из cuML, а не его адаптивная версия [Vijayan, Aziz, 2022].

Переходя к алгоритмам понижения размерности, для предварительного отбора рассматривались t-SNE, PCA, UMAP, TriMap и PaCMAP, причём ключевыми критериями были точность, возможность балансировать отображения локальных и глобальных отношений и скорость обучения. Исходя из опыта других исследователей в качестве базового алгоритма для сравнения был выбран UMAP [Grootendorst, 2022].

Так, исключён из-за недостаточной производительности на больших объёмах данных. PCA рассматривался как быстрое глобальное приближение, но для конечного понижения размерности не обеспечивает локальную точность. Тем не менее, был проведён эксперимент с предварительным применением PCA перед основным алгоритмом, направленный на ускорение вычислений. Эксперимент показал, что пайпайн PCA + UMAP уступает по индексу DBCV «чистому» UMAP примерно на 37%, что сделало применение PCA неоправданным.

Дальнейшее сравнение UMAP, TriMap и PaCMAP продемонстрировало схожую точность при проекции в промежуточные размерности, однако для окончательной двумерной визуализации PaCMAP оказался предпочтительнее за счёт более равномерного распределения кластеров на плоскости. Тем не менее, поскольку для кластеризации на промежуточном этапе была важна возможность GPU-ускоренной оптимизации и проверенная устойчивость, выбор пал на UMAP реализованном в cuML [Raschka, Patterson, Nolet, 2020].

2.3.3 Оптимизация гиперпараметров

После этапа инициализации моделей последовала комплексная оптимизация гиперпараметров, затрагивающая семь ключевых параметров: пять для UMAP (число соседей 'n_neighbors', размер выходного пространства 'n_components', параметры 'min_dist' и 'spread', коэффициент 'negative_sample_rate') и два для HDBSCAN ('min_cluster_size' и 'min_samples'). В качестве фреймворков для поиска оптимума были апробированы Ray Tune [Liaw (и др.), 2018] и Optuna [Akiba (и др.), 2019], причём акцент был сделан на ресурсно-ориентированном подходе, где «ресурсом» выступал размер обучающей подвыборки. Экспериментально были выбрано 5 этапов обучения с долями от 1.5% до 10% общего корпуса, что позволило оценить масштабируемость методов и устойчивость полученных гиперпараметров.

В контексте Ray Tune применялась схема Байесовской оптимизации с HyperBand (ВОНВ) [Bergstra (и др.), 2011; Falkner, Klein, Hutter, 2018; L. Li (и др.), 2018; Shahriari (и др.), 2015], однако она не обеспечила должного соотношения скорости и качества при

работе с UMAP/HDBSCAN. Переключившись на Optuna, были реализованы три разновидности «прунеров» — механизмы для досрочного отсечения бесперспективных комбинаций при наращивании ресурса. Первый, 'AdaptiveStablePercentilePruner', удалял худшие по заранее заданному перцентилю в каждом шаге ресурса, второй, 'CustomPatientPruner', минимизировал риск преждевременного исключения, отсекая комбинации после n последовательных неудачных шагов с приростом менее δ , а третий, 'NormalPruner', использовал Z-статистику (Формула 11) и заданный перцентиля стандартного нормального распределения ($\mathcal{N}(0, 1)$).

$$z\text{-score} = \frac{m - \mathbb{E}[M]}{\mathbb{V}[M]}, \quad (11)$$

где m — текущая метрика и M — распределение метрик на данном этапе.

Метрикой оценки в ходе итераций стал взвешенный кумулятивный DBCV-индекс (WCDBCV):

$$WCDBCV_j = \frac{1}{j \sum_{i=1}^j p_i} \sum_{i=1}^j p_i \cdot DBCV_i, \quad (12)$$

где j — текущий этап ресурса, а p_i — доля выборки, использованная на i -м шаге.

Из всех сочетаний наилучшую эффективность продемонстрировал алгоритм Древовидной оценки Парсена (TPE) [Bergstra (и др.), 2011; Shahriari (и др.), 2015; Watanabe, 2023] в связке с 'AdaptiveStablePercentilePruner', однако прирост качества оказался недостаточно существенным относительно возросших вычислительных затрат. В связи с этим последующий поиск гиперпараметров был выполнен без использования прунинга. В качестве глобальной стратегии применялся классический TPE с 400 испытаниями, из которых половина выполнялась в «разминочном» режиме случайных конфигураций. Для локальной донастройки возникших перспективных областей пространства гиперпараметров использовалась адаптивная эволюционная стратегия (CMAES) [Auger, Hansen, 2005; Nomura, Shibata, 2024] с механизмом повторного старта IPOB, адаптивной скоростью обучения и предварительным «тёплым запуском» на базе 15 лучших комбинаций, найденных на этапе глобальной оптимизации [Auger, Hansen, 2005; Nomura, Akimoto, Ono, 2023; Nomura, Shibata, 2024; Nomura, Watanabe (и др.), 2021]. Бюджет локальной фазы составил 250 испытаний, что обеспечило глубокий поиск вокруг уже высокоэффективных конфигураций.

В результате такого двухуровневого подхода — сначала широкого исследования методом TPE, затем углублённого локального поиска CMAES — удалось сбалансировать широту и глубину оптимизации, повысив устойчивость конвейера моделей.

2.4 Разработка системы

После завершения этапа семантической кластеризации и построения тематических групп, необходимо обработать сформированные кластеры на лингвистическом уровне. В те-

матическом моделировании качество представлений тем является ключевым для интерпретации тем, передачи результатов и понимания закономерностей. Так, необходимо убедиться, что каждое множество документов действительно характеризуется уникальным набором ключевых терминов и не содержит артефактов неправильной сегментации. Данная валидация выходит за пределы чисто метрических оценок и требует эмпирического, «человеческого» взгляда на то, какие слова действительно определяют содержание темы. В контексте неразмеченного корпуса такая проверка особенно важна: без исходных меток о тематике единственным источником информации о семантической однородности кластеров остается текст самой публикации. Более того, существует необходимость в присвоении названий каждому из кластеров, поэтому крайне важно разработать пайплайн репрезентации тем.

С другой стороны, результатом кластеризации являются метки нижнего уровня, а в контексте системы подразумевается иерархический подход к репрезентации тем, что ведет к необходимости извлечения из текущей модели дополнительной информации и ее постобработки.

Для упрощения и формализации процесса в FinABYSS был реализован специализированный пайплайн репрезентации тем, основанный на следующих инструментах:

- BERTopic — для удобства и упрощения процесса обработки лингвистических признаков, визуализации взвешенных частот слов, а также построения иерархии.
- OpenAI API — для автоматизации разметки полученных кластеров на основе их лингвистических признаков.
- DataMapPlot для упрощения работы с графическим интерфейсом и создания интерактивной семантической карты.
- Собственная реализация:
 - над BERTopic — для реализации возможности выбора определенного количества иерархических тематических уровней.
 - между BERTopic и OpenAI — для реализации возможности присвоения тематических названий на основе лингвистических признаков на более высоких иерархических уровнях, нежели только на самом нижнем.
 - между BERTopic и DataMapPlot — для тонкой настройки как статической, так и интерактивной семантической карты.

Таким образом, первым этапом в пайплайне стала векторизация текстов, то есть приведение их к матричному виду. С этой целью использовался 'CountVectorizer' из библиотеки 'sklearn', который был настроен настроенная на извлечение униграмм и биграмм. Это полезно для более точной репрезентации тем, так как ведет к рассмотрению таких терминов как "central bank", "monetary policy" и "New York" и совместно, а не отдельно по словам..

Отныне, слова в контексте документов такие уни- и биграммы будут называться «терминами» для предотвращения путаницы, так как, на самом деле, в контексте разработанной системы рассматриваются не слова, а n -граммы, то есть короткие последовательности из n слов.

Также на основе свободного англоязычного словаря стоп-слов [Nothman, Qin, Yurchak, 2018] был настроен этап фильтрации артиклей, предлогов, местоимений, союзов и других частей речи, мешающих отбору по-настоящему репрезентативных терминов.

Наконец была настроена минимальная частота термина в документе для его включения в матрицу. Несложно представить ситуацию, когда определенный термин встречается во всех документах только один раз. Маловероятно, что данный термин отражает определенную тематику. С другой стороны, в корпусе собрано более миллиона статей и, если не установить минимальную частоту термина, матрица станет огромной и непрактичной в использовании. Поэтому термины, появляющиеся менее 15 раз во всём корпусе, отсекаются как статистически нерепрезентативные.

Наконец, система подразумевает функционирование в режиме реального времени, поэтому принципиально важно уметь инкрементально обновлять матрицу терминов, для этого была использована онлайн версия 'CountVectorizer' — 'OnlineCountVectorizer'.

После формирования матрицы терминов, необходимо сгруппировать их по тематическим кластерам и определить релевантность терминов для каждой тематической группы, чтобы в дальнейшем, на основе наиболее репрезентативных терминов сгенерировать названия кластеров.

Частота термина, обратная частота документа (Term Frequency Inverse Document Frequency, TF-IDF) и Best Match (BM-25) являются аддитивными функциями релевантности. Они используются в большинстве поисковых систем как основные метрики релевантности. Обе метрики показывают релевантность документа. Чем выше значение метрик, тем более релевантен документ. При этом, важно отметить, что само значение метрик не имеет какой-либо значимой интерпретации, кроме как относительная разность релевантности документов или терминов друг с другом.

В контексте поисковых систем метрика отражает релевантность документа или термина поисковому запросу, но в текущей системе используется модифицированная TF-IDF метрика, которая выражает релевантность термина, не по отношению к запросу, а по отношению к теме.

Такая метрика, называется c-TF-IDF [Grootendorst, 2022]. Лучше всего ее можно объяснить, как формулу TF-IDF принятую для всей темы, в целом. То есть все документы внутри темы не рассматриваются по-отдельности, а объединяются в один большой документ, на основе которого происходит подсчет. Так, c-TF-IDF учитывает то, что отличает документы в одном тематическом кластере от документов в другом.

Таким образом, мы сначала извлекаем частоту термина x в тематическом кластере c ,

которому и принадлежит термин x . В результате получается представление tf на основе классов. А для того, чтобы учесть различия в размере тематических кластеров, по полученным частотам вычисляется L_1 -норма.

Затем, вычисляется логарифм суммы единицы и частного от деления среднего количества слов в каждом из тематических кластеров на частоту слова x во всех кластерах. Так, получается представление idf , которое помогает определить насколько данное слово x редко встречается среди всех остальных классов. Единица нужна для того, чтобы логарифм всегда был положительным.

Наконец, как и в случае с обычным TF-IDF, полученные представления tf и idf перемножаются (Формула 13):

$$w_{x,c} = ||tf_{x,c}||_1 \times \log \left(1 + \frac{\frac{1}{||\mathbb{C}||} \sum_{i \in \mathbb{C}} ||\mathbb{X}_i||}{\sum_{i \in \mathbb{C}} x} \right), \quad (13)$$

где $w_{x,c}$ — релевантность терма x тематическому кластеру c , \mathbb{X}_i — множество всех термов в тематическом кластере i .

Тем не менее, вместо классической формулы с-TF-IDF, было принято решение использовать ее модифицированный аналог. Некоторые слова или термины встречаются слишком часто в каждой из тем, однако не считаются типичными стоп-словами для исключения из текста. Чтобы сгладить самые частые термины в теме, в системе применяется нормализация частот терминов, то есть извлекается квадратный корень из tf , после применения схемы взвешивания.

С другой стороны, несмотря на то, что собранный финансовый корпус велик, он может не вполне исчерпывающе отражать все лексическое изобилие генеральной совокупности. Поэтому, для более стабильного результата, к idf было применено преобразование BM-25. Итоговая формула релевантности выглядит следующим образом:

$$w_{x,c} = \sqrt{||tf_{x,c}||_1} \times \log \left(1 + \frac{\frac{1}{||\mathbb{C}||} \sum_{i \in \mathbb{C}} ||\mathbb{X}_i|| - \sum_{i \in \mathbb{C}} x + 0.5}{\sum_{i \in \mathbb{C}} x + 0.5} \right), \quad (14)$$

Коэффициенты 0.5 основаны на подгонке под теоретически более чистую форму. Такая форма дает меньший вес слишком часто встречающимся терминам.

В итоге, мы получаем для каждого из тематических кластеров мешки слов, которые достаточно представительно отражают лексическое богатство каждой темы. Тем не менее, в мешках слов все еще могут присутствовать некоторые не совсем представительные слова или же мешок может быть излишне однородным. Данные проблемы могут негативно сказаться на процессе генерации названий тем, поэтому при разработке системы был предусмотрен более сложный пайплайн, который добавляет еще 2 шага перед финальной презентацией.

Первый шаг заключается в семантическом сравнении наиболее представительных слов, с наиболее представительными документами. Сначала из мешка слов извлекается 30

наиболее репрезентативных терминов, а из самого тематического кластера — 5 наиболее репрезентативных документов. Затем, мы собираем уже извлеченные эмбеддинги по соответствующим документам, и при помощи той же эмбеддинговой модели — 'gpt2-modetnbert-base' — извлекаем эмбеддинги из выбранных терминов. После чего, мы сравниваем эмбеддинги терминов-кандидатов с наиболее репрезентативными документами и ранжируем их по полученному значению метрики косинусного расстояния.

На втором шаге мы применяем алгоритм Максимальной Предельной Релевантности (Maximal Marginal Relevance, MMR) — это техника для реферирования по запросу, которая максимизирует похожесть фрагментов ответа на запрос и минимизирует схожесть с уже выбранными в ответ фрагментами. Аналогично TF-IDF, в контексте текущей системы рассматривается не запрос, а термин. То есть мы максимизируем внутреннее разнообразие наиболее репрезентативных тем терминов. MMR вычисляется по следующей формуле:

$$MMR = \operatorname{argmax}_{t_i \in \mathbb{R} \setminus \mathbb{S}} [\lambda(sim_1(t_i, \mathbb{T})) - (1 - \lambda) \max_{t_j \in \mathbb{S}} (sim_2(t_i, t_j))], \quad (15)$$

где \mathbb{T} — тема, для которой рассчитывается MMR , \mathbb{R} — множество всех терминов t_i — рассматриваемый термин, а \mathbb{S} — множество уже выбранных терминов. Получается, что мы ищем новый термин из $\mathbb{R} \setminus \mathbb{S}$ так, чтобы он был максимально похож на тему, но минимально на уже присутствующие в мешке слов термины. В Формуле 15 коэффициент λ является гиперпараметром и балансирует похожесть терминов на тему с непохожестью терминов друг с другом. Чем меньше λ , тем термины менее похожи друг на друга, а чем больше, тем похоже термины на тему.

Так, у нас получается максимально репрезентативный набор терминов, относительно которых мы уже можем генерировать названия тем. Непосредственно для генерации тем в системе предусмотрено использование text2text моделей. Конкретно в текущей реализации использовалась модель GPT-4o, для которой был задан соответствующий промпт с инструкциями. Модель использовалась с помощью OpenAI API. А все сгенерированные метки позднее были провалидированы вручную.

Наконец, заключительным этапом является визуализация семантической карты. Визуализация была выполнена в двух форматах: в статичном — для работы и презентации, и в интерактивном — для функционирования системы. Визуализация осуществлялась при помощи Python-библиотеки DataMapPlot, а также кастомных дополнений на HTML, CSS и JavaScript. Также на системном уровне были разработаны функциональности для поиска по текстам, построения облака слов, фильтрации по источникам, активам, дате публикации и другим количественным признакам.

ГЛАВА 3. РАЗУЛЬТАТЫ

3.1 Спроектированная архитектура

3.1.1 Обзор

В рамках данного исследования была разработана архитектура предиктивной системы нового поколения. Основное внимание в настоящей работе уделено созданию системы тематического моделирования и аналитического интерфейса для финансовой предметной области. Логичным продолжением станет построение интерпретируемой и эффективной модели прогнозирования на основе тематической тональности финансовых публикаций (новостных статей, пресс-релизов, постов, транскриптов и др.). Таким образом, настоящая работа не только демонстрирует принципы функционирования базового модуля, но и формулирует видение того, каким образом можно расширить её возможностями предсказания цен активов.

Предложенная архитектура опирается на доказавшие свою надёжность подходы смешанного прогнозирования — сочетание сверточных и рекуррентных нейронных сетей (CNN-LSTM) [Hochreiter, Schmidhuber, 1997; LeCun (и др.), 1998; Lu (и др.), 2020]. Это решение позволяет одновременно учитывать глобальные долгосрочные тенденции (например, общий тренд рынка) и локальные краткосрочные паттерны (такие как «голова и плечи», «двойное дно» и другие структуры, обусловленные поведенческой экономикой).

При этом важно помнить о разнообразии исходных данных. Для прогнозирования обычно используют биржевые метрики (OHLCV и т.п.) и производные технические индикаторы (RSI, MACD и др.). В нашем исследовании к ним добавляются текстовые данные как внебиржевой источник. В дальнейшем в качестве дополнительных модальностей могут быть привлечены графовые представления связей, изображения, аудио- и видеозаписи. Таким образом, на текущем этапе мы работаем с тремя модальностями, каждая из которых несёт самостоятельную смысловую нагрузку и способна объяснить динамику цен независимо, но их взаимодействие может как усиливать полезный сигнал, так и вносить шум.

В данном исследовании архитектура была построена на основе CNN-LSTM с использованием медленного слияния. Основной инженерный вызов состоял в согласовании пространственно-временных форм признаков разных модальностей. Для решения этой задачи в архитектуру введён Механизм Кэширования Признаков (Feature Caching Mechanism, FCM), который синхронизирует тональностные векторы текстовой модальности с биржевыми временными рядами.

Дополнительно текстовая ветвь подвергается предобработке: на основе разработанной системы тематического моделирования вычисляются тематические тональности публикаций. Это позволяет получить специализированные оценки по каждой теме, что даёт преимущество перед использованием единой общей тональности.

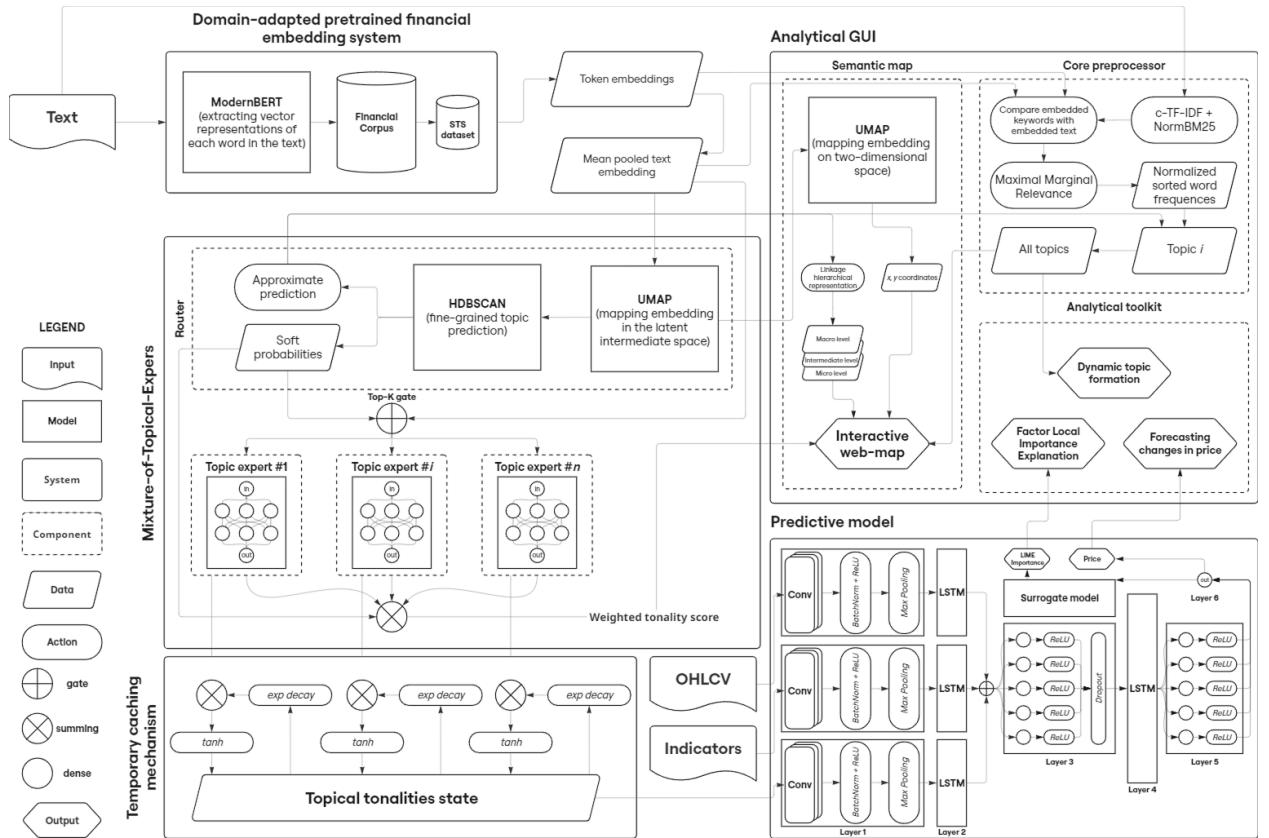


Рис. 7: Обзор спроектированной архитектуры FinABYSS.

Спроектированная система состоит из пяти основных модулей (Изображение 7):

- Финансовая языковая модель (доменно-адаптированная, тонко настроенная для задачи STS), выдающая векторные представления текста целиком и для каждого токена.
- Смесь тематических экспертов (Mixture-of-Topical-Experts, MoTE), формирующая тональность публикации по каждой теме.
- TCM, выравнивающий тематические векторы по временным меткам биржевых данных.
- Ядерная предиктивная модель с CNN-LSTM и медленным слиянием, обрабатывающая все модальности и предсказывающая цену актива.
- Аналитический графический интерфейс (GUI), агрегирующий промежуточные результаты для последующего финансового анализа.

Таким образом, созданная архитектура представляет собой мощный и адаптируемый инструмент для прогнозирования стоимости финансовых активов на основе комплексной оценки тематических сентиментов публикаций.

3.1.2 Эмбеддинговая система

Приступая к рассмотрению отдельных блоков разработанной архитектуры, следует начать с этапа предобработки данных. Наиболее ресурсоёмким и сложным из них является подготовка внебиржевых текстовых источников.

Сразу после поступления в систему текст проходит первичную обработку с помощью доменно-адаптированной финансовой эмбеддинговой модели, настроенной на задачу STS. В результате работы этой модели извлекаются векторные представления для каждого токена, что позволяет сохранить тонкие контекстные связи внутри предложений. По мере появления новых версий эмбеддинговых моделей важно регулярно обновлять базовую модель и повторно адаптировать её к специфике финансового домена.

Простейший путь доменной адаптации включает дополнительную стадию обучения на задаче MLM. В качестве корпуса для этой цели может использоваться корпус финансовых публикаций, сформированный в ходе настоящего исследования, либо его расширенная версия, а также другие релевантные текстовые корпусы. Для тонкой настройки под задачу STS иногда применяют небольшие размеченные датасеты, однако для экономии ресурсов возможно обойтись и без них, используя методы контрастивного обучения на случайных подвыборках основного корпуса [Gao, Yao, D. Chen, 2021].

После извлечения эмбеддингов токенов производится их агрегирование: обычно применяется усреднение, что даёт векторное представление всего документа. Далее документ и его эмбеддинги (как токенов, так и усреднённый вектор) передаются в аналитический модуль для визуализации и интерпретации результатов. Параллельно вектор документа направляется в блок MoTE, где на его основе формируются тематические тональности публикации.

В качестве конкретной реализации в данном исследовании используется модель 'gtemodernbert-base', построенная на архитектуре ModernBERT. Несмотря на то, что изначально она не проходила доменную адаптацию, она демонстрирует высокую эффективность: при кластеризации эмбеддингов удалось получить DBCV-индекс 0.407, а визуальный и контекстуальный анализ показали чёткое разделение тематических групп.

Таким образом, предложенный этап предобработки обеспечивает:

- Сохранение глубоких семантических связей на уровне токенов и всего документа.
- Гибкость и воспроизводимость процесса адаптации модели к финансовому корпусу.
- Интеграцию результатов в соседние блоки архитектуры — MoTE и аналитический GUI.

3.1.3 Смесь тематических экспертов

Данный блок использует современную архитектуру Mixture-of-Experts (MoE), успешно применяемую в крупных языковых моделях (например, Mixtral 8×7B, DeepSeek R1) и в

исследованиях Google (Switch Transformer) [Fedus, Zoph, Shazeer, 2022]. Основные компоненты данного блока — обучаемый роутер и множество экспертов — позволяют динамически активировать лишь небольшую часть параметров при инференсе, обеспечивая высокую пропускную способность и эффективность [Shazeer (и др.), 2017]. Специально для анализа финансовых текстов эксперты тематически специализированы, что повышает интерпретируемость и точность оценки тональности публикаций.

Архитектура MoE построена на принципе условных вычислений: лишь часть подсетей («экспертов») активна для каждого входного примера, что снижает вычислительные затраты при огромном общем числе параметров [Там же].

- Роутер получает на вход эмбеддинг документа и вычисляет распределение «весов» для всех экспертов, после чего выбирает либо топ- K экспертов, либо тех, чей вес превышает заданный порог, являющийся гиперпараметром. [Fedus, Zoph, Shazeer, 2022].
- Эксперты представляют собой неглубокие feed-forward сети, каждая из которых специализируется на своей тематике (в контексте работы — на одной из финансовых тем) [Shazeer (и др.), 2017].

При обработке каждого эмбеддинга активизируется лишь малая часть экспертов — обычно по Топ K (выбор K наиболее релевантных экспертов) или по пороговому значению (вес эксперта выше заданного порога) — что делает MoE чрезвычайно экономным с точки зрения вычислений при колоссальном общем числе параметров [Fedus, Zoph, Shazeer, 2022]. Значения K и θ выступают гиперпараметрами и настраиваются на валидационных данных.

Для анализа финансовых текстов MoE-модель является логичным выбором, поскольку публикации часто содержат смесь смежных тематик (макроэкономика, корпоративные отчёты, geopolitika и др.), и требуется оценить тональность под разными углами. Эксперты позволяют формировать специализированные оценки по каждой теме одновременно, сохраняя «чистоту» низкоуровневой обработки и обеспечивая интерпретируемость результатов [Jacobs (и др.), 1991]. Тем не менее, стоит отметить и высокую сложность подбора K или θ гиперпараметров [Там же].

Вместо обучения экспертов на субъективно размеченных данных их параметры оптимизируются обратным распространением ошибки, поступающей из блока предсказания цен активов. Это гарантирует, что каждый эксперт обучается напрямую на влиянии публикаций на стоимость активов, а не на внешних аннотациях [Shazeer (и др.), 2017]. После получения тематических тональностей они агрегируются взвешенным суммированием с учётом вероятностей тем, нормализуются и направляются в аналитическую систему для визуализации, а также в механизм кэширования признаков для синхронизации с биржевыми данными.

Таким образом, блок MoTE на базе MoE обеспечивает сочетание масштабируемости, эффективности и интерпретируемости при анализе финансовых текстов.

Динамическая активация небольшого числа экспертов позволяет обрабатывать огромные модели без пропорционального роста вычислений, а обучение экспертов через сигнал от модели прогнозирования цен делает оценку тональностей объективной и близкой к экономической реальности. Таким образом, предлагаемая архитектура открывает новые возможности для точного прогнозирования и глубокого анализа влияния текстовых публикаций на стоимость финансовых активов.

3.1.4 Механизм кэширования признаков

FCM представляет собой центральный компонент архитектуры, обеспечивающий синхронность между биржевыми данными (регулярно поступающими временными рядами) и внебиржевыми текстовыми «тиками», разбросанными во времени. Когда из блока Смеси Тематических Экспертов поступает новая оценка тональности публикации по теме i , она моментально добавляется в матрицу накопительных экспоненциально сглаженных значений по Формуле 16:

$$x_{i,t} = x_{i,t-1} + x_{i,0} \cdot e^{-\lambda t}, \quad (16)$$

где $x_{i,0}$ — исходная тональность по теме i , $x_{i,t-1}$ — предыдущее сглаженное значение, а λ — гиперпараметр скорости экспоненциального «затухания» сигнала. Такой подход отражает запаздывающую и сглаженную во времени реакцию участников рынка на публикации.

Данный подход учитывает, что реакция участников рынка на информационные события задерживается и распределена во времени: первая «волна» восприятия быстро затухает, за ней следуют более плавные и длительные эффекты, что в финансах традиционно моделируется экспоненциальным спадом.

Так как исходные тональности нормированы в диапазоне $[-1, 1]$, накопительное суммирование по первой формуле может выходить за эти границы. Чтобы сохранить скалирование в допустимых интервалах и придать плавность накоплению, предлагается повторно нормировать результаты с помощью гиперболического тангенса (Формула 17):

$$tone_{i,t} = \tanh(tone_{i,t-1} + x_{i,t}) = \frac{\tanh(tone_{i,t-1}) + \tanh(x_{i,t})}{1 + \tanh(tone_{i,t-1}) \cdot \tanh(x_{i,t})}. \quad (17)$$

что эквивалентно Формуле 18:

$$tone_{i,t} = \frac{(1 - e^{-2 \cdot tone_{i,t}}) + (1 - e^{-2 \cdot x_{i,t}} - (1 - e^{-2 \cdot tone_{i,t}}) \cdot (1 - e^{-2 \cdot x_{i,t}})}{(1 + e^{-2 \cdot tone_{i,t}}) + (1 + e^{-2 \cdot x_{i,t}} - (1 + e^{-2 \cdot tone_{i,t}}) \cdot (1 + e^{-2 \cdot x_{i,t}})}, \quad (18)$$

где $tone_{i,t}$ — текущая нормированная экспоненциально сглаженная тональность по теме

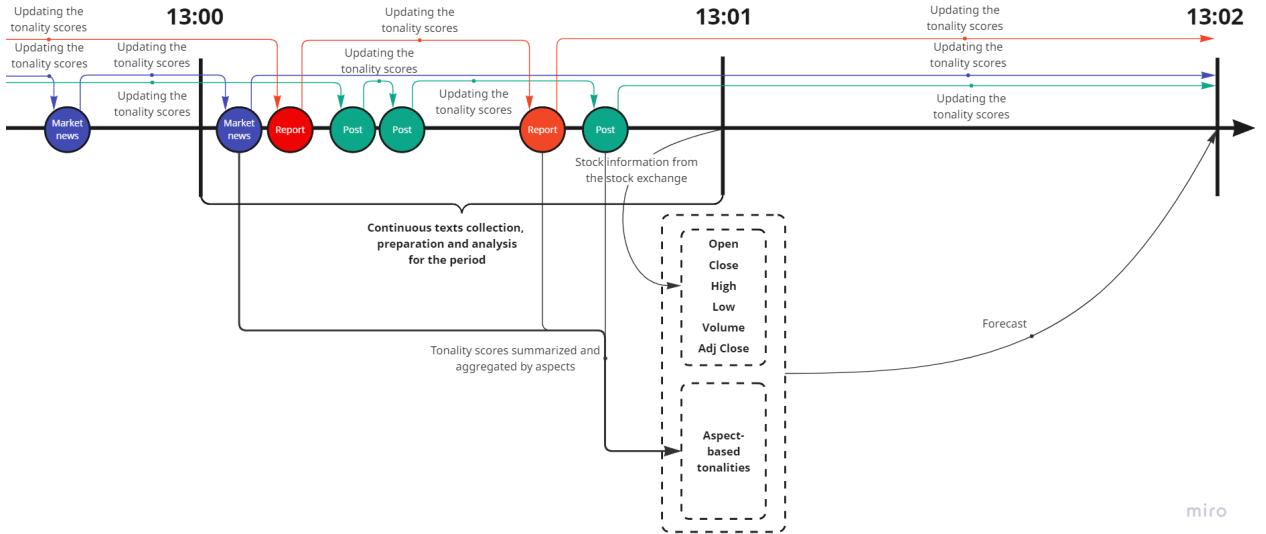


Рис. 8: Динамическая репрезентация синхронизации внебиржевых данных при помощи FCM на минутном свечевом интервале.

Таким образом, FCM выполняет следующие функции: преобразует нерегулярные текстовые события в непрерывный многомерный временной ряд, применяет адекватное финансовым реалиям сглаживание и масштабирование тональностей, а также гарантирует синхронизацию с биржевыми данными. Это обеспечивает согласованность и стабильность входных признаков в предиктивной модели и способствует повышению точности прогнозирования стоимости финансовых активов.

3.1.5 Предиктивная модель

В предлагаемой гибридной архитектуре мы стремимся объединить преимущества свёрточных слоёв для локального выделения признаков и рекуррентных модулей для моделирования длительных зависимостей, одновременно обеспечив адаптивное слияние количественных и качественных сигналов. На вход подаётся многомерный временной ряд длины T , в котором каждому временному шагу t соответствует вектор $x_t = [x_t^{\text{price}}, x_t^{\text{ind}}, x_t^{\text{sent}}]$, где $x_t^{\text{price}} \in \mathbb{R}^{C_p}$ содержит OHLCV, $x_t^{\text{ind}} \in \mathbb{R}^{C_i}$, а $x_t^{\text{sent}} \in \mathbb{R}^{C_s}$ — набор скользящих и сглаженных значений тональностей новостей.

Сначала данные разбиваются на три ветви: «Price», «Indicator» и «Sentiment». Каждая ветвь проходит два базовых уровня обработки. На первом уровне в каждой ветви последовательность $\{x_t^{(\cdot)}\}_{t=1}^T$ пропускается через одномерную свёртку с небольшими ядрами и операциями пулинга (Формула 19)

$$y_t^{(\ell+1)} = \text{ReLU}(W^{(\ell)} * y_t^{(\ell)} + b^{(\ell)}), \quad y_t^{(\ell+\frac{1}{2})} = \max(y_t^{(\ell+1)}, y_{t+1}^{(\ell+1)}). \quad (19)$$

где символ $*$ обозначает свёртку по времени, функция активации ReLU определяется следующим образом (Формула 20):

$$\text{ReLU}(x) = \max(0, x). \quad (20)$$

Затем операция \max пулинга сокращает временную размерность вдвое. Такая последовательность позволяет в начальных слоях выделить локальные шаблоны внутри каждого потока, будь то короткие ценовые колебания, импульсы индикаторов или флуктуации сентимента.

Далее на выходе пулинга в каждой ветви обучается рекуррентный модуль LSTM, который моделирует накопление и забывание информации с учётом длительных временных зависимостей. Пусть h_t и c_t — скрытое состояние и ячейка памяти LSTM; их эволюция задаётся стандартными уравнениями входных-, забывающих- и выходных-гейтов. Благодаря этому каждая ветвь самостоятельно учится определять, какие локальные признаки значимы для прогнозирования на более отдалённых временных горизонтах.

После того как каждая ветвь выдала своё скрытое состояние $h_t^{\text{price}}, h_t^{\text{ind}}, h_t^{\text{sent}}$ на каждом временном шаге, происходит этап адаптивного слияния. Для этого векторная конкатенация u_t пропускается через гейтовый слой (Формула 21)

$$g_t = \sigma(W_g u_t + b_g). \quad (21)$$

где σ — сигмоида и $g_t \in (0, 1)^{\dim u}$ задаёт поэлементные коэффициенты, контролирующие относительный вклад каждого потока. Само же объединённое представление вычисляется как

$$m_t = g_t \odot u_t + (1 - g_t) \odot \mu(u_t). \quad (22)$$

где $\mu(u_t)$ может быть усреднением или другим агрегатором компонентов вектора u_t . Такой механизм позволяет модели динамически переключаться между акцентом на ценовых паттернах, индикаторных сигналах или сентименте в зависимости от контекста конкретного временного отрезка.

Объединённая последовательность $\{m_t\}_{t=1}^{T/2}$ затем вновь поступает в рекуррентный модуль LSTM, который, подобно предыдущим, аккумулирует информацию о совместном развитии трёх модальностей на более высоком уровне абстракции. Его выходом служит последнее скрытое состояние $h_{T/2}^{\text{merge}}$, содержащее сжатое резюме всей истории. Для придания нелинейности и дополнительного измерения выразительности это состояние может быть пропущено через полносвязный слой с функцией ReLU, а затем — через слой Dropout для регуляризации и смягчения переобучения.

Наконец, заключительный слой без активации переводит полученный вектор в одномерное предсказание \hat{y} , которое интерпретируется как новая цена (в режиме регрессии).

Также, стоит отметить, что, в целях, локальной интерпретируемости оценок тональности важную роль в архитектуре играет наличие отдельной суррогатной модели, которая

на основе выборки ряда тональностей, полученного из Feature Caching Mechanism, с дополнительными волнениями, оценивает вклад каждой тематической тональности в итоговое предсказание стоимости актива.

Таким образом, предложенная гибридная архитектура объединяет локальную детекцию паттернов (одномерной свёртки), способность LSTM моделировать длительные зависимости, а также механизм адаптивного слияния, позволяющий динамически регулировать вклад ценовых, индикаторных и сентиментальных сигналов. Это обеспечивает высокую точность прогнозов при умеренной вычислительной сложности и прозрачности модели относительно используемых признаков, при этом оставаясь более лёгкой и прозрачной по сравнению с трансформерными аналогами.

3.1.6 Аналитический графический интерфейс

В завершение описания архитектуры FinABYSS необходимо остановиться на ключевом пользовательском компоненте — Аналитическом GUI. Этот модуль аккумулирует результаты всех предыдущих этапов: от предобработки внебиржевых текстов и их тематической интерпретации в блоке MoTE до конкатенации с ценовыми и индикаторными рядами в предиктивной модели. На вход GUI поступают промежуточные и финальные данные, а его задача — привести их в наглядный, интерактивный и аналитически релевантный вид. GUI состоит из трёх взаимосвязанных компонент, реализующих постобработку данных, представление инструментов глубокого анализа и визуализацию семантических связей между документами.

Первый компонент — ядерный постпроцессор — отвечает за трансформацию «сырых» выходов предыдущих блоков в упорядоченные и компактные репрезентации. На этапе инициализации он вычисляет относительные частоты терминов с помощью BM25-модификации TF-IDF и дополнительной нормализации, что удаляет шумовые и часто встречающиеся, но семантически незначимые токены (например, названия агрегаторов типа Zacks). Далее, после получения эмбеддингов токенов и всего документа из доменно-адаптированной языковой модели, выполняется ранжирование n -грамм по частотам с учётом косинусного сходства между векторами токенов и документного вектора. Для обеспечения баланса между релевантностью и диверсификацией подсистемы применяется алгоритм MMR: при разбиении на биграммы MMR устраняет дублирующиеся или избыточно близкие по смыслу фрагменты, оставляя наиболее информативные. Например, из сырых биграмм «ai» и «ai stocks» в зависимости от относительной частоты и семантической близости могут сохраняться либо «ai» и «stocks», либо только «ai stocks», либо только «ai», что повышает однородность итогового словаря и уменьшает избыточность.

После применения MMR словарь сортируется по итоговым частотам, и постпроцессор получает от блока MoTE метку кластера, к которому принадлежит текущий документ. На основе этой кластерной принадлежности осуществляется инкрементальное обновление

c-TF-IDF статистики: накопленные частоты слов внутри каждого кластера корректируются с учётом новой публикации, что обеспечивает адаптивность и учёт эволюции тематических трендов во времени и используется для динамического тематического моделирования. Параллельно от постпроцессора поступают два потока данных: обогащённые частотные признаки и другие извлечённые метаданные текста направляются в аналитический инструментарий, а сам очищенный и классифицированный текст с эмбеддингами — в Семантическую карту.

Второй компонент — Аналитический инструментарий — представляет собой набор визуальных и интерактивных элементов, позволяющих пользователю углублённо исследовать динамику тематических трендов и их влияние на цены активов. Через дашборды можно получать:

- исторические графики тональностей по каждой теме, совмещённые с ценовыми рядами и техническими индикаторами;
- динамические корреляционные матрицы между темами и движениями цен;
- сводные отчёты по влиянию макро-, мезо- и микротем на волатильность и трендовые движения;
- агрегированные показатели рыночной тональности (взвешенная суммарная тональность по всем темам) с трендовыми линиями.

Инструментарий разработан с учётом будущего расширения: при добавлении новых модальностей (изображения, графовые данные, аудио- или видеопотоки) к нему легко интегрируются модальностно-специфичные виджеты — например, интерактивные облака ключевых сущностей, графы взаимосвязей эмитентов или тепловые карты тональностей в голосовых и видео-форматах.

Третий компонент — Семантическая карта — реализует двумерную визуализацию эмбеддингов документов. На основе заранее обученной модели понижения размерности каждое документное представление из латентного пространства, где проводилась кластеризация, отображается в координатах (x, y) . Параллельно используется аппроксимальный предсказатель темы из блока Смеси Тематический Экспертов: по каждому документу определяется иерархически организованная тройка тематических меток — на макро-, мезо- и микроверсиях. Цвет, размер и форма маркеров на карте кодируют принадлежность к кластеру, силу тональности и объём текста. Это позволяет анализировать не только пространственное соседство публикаций, но и их тематическую многоуровневую структуру, быстро выявлять периферийные и центральные документы, а также отслеживать эволюцию тематических сообществ.

В совокупности GUI создаёт единую интерактивную панораму, в которой результаты глубочайших вычислительных блоков — от FCM и MoTE до гибридной CNN-LSTM предиктивной модели — превращаются в интуитивно понятный и аналитически ёмкий инструмент.

Пользователь получает возможность не просто просматривать прогнозы цен, но и детально исследовать причинно-следственные связи между медиасигналом и движением рынка.

Аналитический графический интерфейс FinABYSS обеспечивает сквозную визуализацию и интерпретацию выходных данных системы, интегрируя частотные, тематические и эмбеддинговые признаки в три взаимосвязанных подсистемы. Ядровой постпроцессор трансформирует «сырые» текстовые и числовые данные в компактные и репрезентативные признаки, аналитический инструментарий предлагает гибкие средства глубокого исследования влияния тем на стоимость активов, а семантическая карта даёт наглядное отображение многомерных эмбеддингов с учётом иерархической тематики. Благодаря такому сочетанию технологической мощности и удобства взаимодействия GUI выступает завершающим, но не менее важным звеном в конвейере прогнозирования, обеспечивая прозрачность, адаптивность и наглядность результатов для экспертов в области финансов.ⁱ на момент t .

В результате непрерывной работы FCM формируется матрица накопительных тематических тональностей, обновляемая при каждом поступлении текстового «тика». Пока биржевые данные (OHLCV и индикаторы) отсутствуют, этот временной ряд накапливает сигналы «рыночного сентимента». Как только в систему приходит новый временной ряд биржевых измерений (с учётом сетевой задержки δ), накопленные и нормированные тональности конкатенируются с ценовыми и индикаторными данными и передаются в предиктивную модель для совместной обработки.

Динамический процесс работы механизма проиллюстрирован на Изображении 8. Во входной момент кэш инициализируется нулевыми значениями, затем при каждом поступлении текстовой оценки обновляется по первой формуле, после чего нормируется гиперболическим тангенсом. По достижении триггера — момента t или поступления биржевых данных в момент $t + \delta$ — объединённый временной ряд передаётся дальше.

3.2 Разработанные системные компоненты

3.2.1 Эмбеддинговая система и роутер

Основным итогом настоящего исследования стало создание двух ключевых компонентов: гибкой эмбеддинговой подсистемы и обучаемого роутера для блока MoTE. Эти наработки не только обеспечивают надёжный фундамент для дальнейшего развития FinABYSS, но и открывают широкий простор для применения в смежных прикладных проектах.

В рамках исследования мы провели серию экспериментов по отбору оптимальных комбинаций эмбеддинговых моделей, методов понижения размерности и алгоритмов кластеризации, используя разные метрики для настройки гиперпараметров. На первом этапе базовая цепочка состояла из ModernBERT, UMAP (косинусная метрика), HDBSCAN (евклидова метрика) оптимизировалась по коэффициенту силуэта. Несмотря на удовлетворительные результаты, мы отметили сильную зависимость от структуры кластеров и чувствительность силуэта к форме кластеров.

Второй этап повторил ту же связку, но в качестве целевой метрики был выбран DBCV-индекс. Благодаря свойству DBCV учитывать плотностные особенности и неоднородность распределения точек, гиперпараметры, подобранные под DBCV, обеспечили более семантически согласованные и плотные кластеры.

Наконец, мы протестировали тонко настроенную на задачу STS версию ModernBERT ('gte-modernbert-base') [Warner (и др.), 2024; X. Zhang (и др.), 2024] в сочетании с UMAP, но уже с L_2 -евклидовой метрикой, и тем же HDBSCAN, также настроенным под L_2 -евклидову дистанцию (Таблица 3). Однако это сочетание уступило предыдущему: полученный DBCV-индекс составил лишь 0.405 против 0.476 у пары "cosine + euclidean" (Таблица 4).

Таблица 3: Description of the final configurations of experiments conducted on a subsample of 10,000 embeddings.

Model	Configuration	
	(I)	(II)
UMAP	L_2 -Euclidean	Cosine
HDBSCAN	L_2 -Euclidean	Euclidean

Дополнительный анализ шумовых точек и структуры кластеров не подтвердил это преимущество (Таблица 4), однако в ходе анализа выяснилось, что подобная ситуация происходит по причине внутренней нестабильности GPU реализации [Raschka, Patterson, Nolet, 2020]. На метрике же, данный факт не сказался существенно негативным образом. При использовании L_2 -метрики доля шумовых точек достигала 39.54%, число кластеров — 73 (при максимальном размере 3 824 точек и минимальном — 365). Для пары "cosine + euclidean" больше точек было отмечено как шум 43.38%, а число кластеров увеличилось до 141; при этом максимальный кластер вырос до 2 558 точек, а минимальный сократился до 147. Так как целевая метрика именно DBCV, можно сделать вывод, что подход "cosine + euclidean" продемонстрировал лучшую устойчивость кластеров.

Экспериментальные результаты однозначно свидетельствуют о превосходстве схемы ModernBERT + UMAP с косинусной метрикой и HDBSCAN с евклидовой дистанцией в задачах тематической кластеризации финансовых текстов. В сочетании с оптимизацией по DBCV-индексу данный подход формирует наиболее семантически связные группы, минимизирует долю шумовых экземпляров и предоставляет более стабильную основу для последующего обучения роутера MoTE.

Тем не менее, к сожалению, были выявлены и явные проблемные места текущей реализации. Полученные кластеры и их структура не могут инкрементально обновляться при потоковых поступлениях новых публикаций, что происходит по нескольким причинам. Во-первых, обученная модель является GPU-реализацией из библиотеки 'cuML', в которой лишь после проведения исследования были обнаружены критические ошибки в коде версии 25.02 [Там же]. Во-вторых, выбранный алгоритм UMAP достаточно плохо работает в ин-

Таблица 4: Summary table of the results of hyperparameter optimization performed for three specified configurations on a subsample of 10,000 embeddings.

Best trial		Configuration	
		(I)	(II)
Hyperparameters	n_components	47	41
	n_neighbors	70	75
	min_dist	0,0	0,065
	spread	6,5	6,4
	negative_sample_rate	10	11
	min_cluster_size	360	145
	min_samples	180	170
Statistics	Max cluster size	3824	2558
	Min cluster size	365	147
	Total clusters	73	141
	Noise %	39,54%	43,38%
	DBCV Index	0,397	0,407

крементальном режиме по своей природе, именно поэтому существуют другие реализации, как например AlignedUMAP [McInnes, Healy, Saul (и др.), 2018] или ParametricUMAP, основанный на нейронной сети в качестве базовой модели [Sainburg, McInnes, Gentner, 2020]. К сожалению, обе данные реализации доступны исключительно для обучения на центральном, а не графическом процессоре. А для обучения на центральном процессоре в контексте текущего исследования недостаточно вычислительных ресурсов.

Таким образом, узким местом данного исследования и предложенного решения является недостаток вычислительных мощностей для обучения моделей, что будет доработано в дальнейших исследованиях.

3.2.2 Семантическая карта

Наконец, на базе разработанной эмбеддинговой подсистемы и роутера MoTE была обучена отдельная модель UMAP, призванная переводить векторы из промежуточного латентного пространства кластеризации в двумерное представление, удобное для визуального анализа. Именно эта проекция лежит в основе Семантической карты финансовых публикаций — одного из ключевых компонентов интерфейса FinABYSS, обеспечивающего глубокое и интуитивно понятное исследование тематической структуры новостных потоков.

Semantic Map of Financial News

News articles from Yahoo! Finance (sample of 10,000 instances)

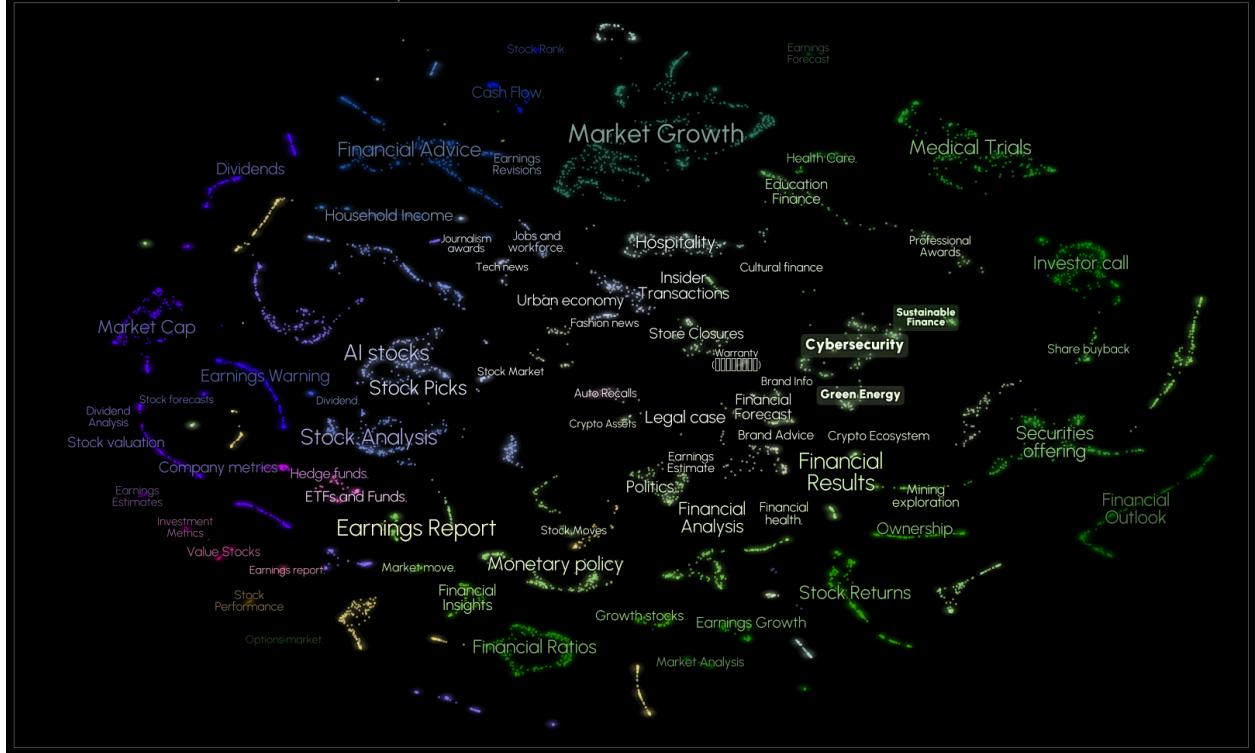


Рис. 9: Статичная емантическая карта, разработанная на основе подвыборки из 10 000 финансовых публикаций, опубликованных в период с 17.09.2023 по 18.04.2025, и кластеризованных по темам.

Для начала была реализована статическая версия карты, на которой отображены 100000 финансовых статей периода с сентября 2023 по март 2025 года, разбитые на кластеры по тематике (Изображение 9). Каждый кластер промаркирован автоматически сгенерированным словом-меткой без какой-либо ручной аннотации. Несмотря на автоматический характер присвоения, метки получились репрезентативными: плотные группы статей, посвящённые здравоохранению, «Устойчивым финансам», «Кибербезопасности» и «Зелёной энергетике», расположены рядом, что отражает их семантическую близость; аналогично происходит с кластерами «Политика» и «Монетарная политика».

Однако статическая карта демонстрирует потенциал лишь UMAP-проекции. Интерактивная Семантическая карта FinABYSS выходит далеко за рамки простой визуализации точек:

- Наведение курсора на любую точку раскрывает метаданные статьи: заголовок, дата и время публикации, иерархическую тему (макро-/мезо-/микротема), автора и источник. При этом доступен предпросмотр полного текста и прямая ссылка на оригинал публикации.
- Механизм поиска по ключевым словам позволяет быстро отфильтровать статьи, содержащие специальные термины. Поиск можно комбинировать с фильтрами по диа-

пазону дат, объёму публикаций и другим числовым признакам (например, объёму текста или количеству просмотров), что облегчает точечное обнаружение исторических событий и триггеров на графике.

- Раздел источники и темы даёт возможность включать и исключать источники новостей или целевые кластеры, помогая сфокусироваться на релевантных публикациях в сложных аналитических сценариях.
- Также для статей доступна функция облака слов, формируемого на основе наиболее частых слов, встречающихся в выделанной группе текстов. Облако слов мгновенно отображает доминирующие термины и паттерны обсуждения, что дополняет количественную канву графика качественными характеристиками.

Таким образом, Семантическая карта FinABYSS представляет собой полноценную аналитическую систему, объединяющую мощь выбранных моделей UMAP и HDBSCAN, а при дальнейшем развитии и гибридной CNN-LSTM-архитектуры и MoE-подхода к оценке тональностей. Она обеспечивает исследователю возможность не только визуально различать именованные кластеры и их взаимное расположение, но и глубоко погружаться в содержание каждой публикации, комбинируя автоматические и ручные методы анализа.

Разработанная интерактивная Семантическая карта выступает естественным продолжением предыдущих модулей FinABYSS. Все звенья конвейера связаны в единую цепь. Инструмент предлагает пользователю прозрачный, масштабируемый и гибко настраиваемый интерфейс для семантического исследования финансовых новостей, где каждый элемент — от кластерных меток до облака слов — отражает результаты вычислительной логики системы и поддерживает экспериментальные решения при анализе рыночных процессов.

3.2.3 Динамическое тематическое моделирование

Помимо семантической карты, FinABYSS предоставляет мощный инструментарий для анализа лингвистических особенностей сформированных тематических групп и динамического темпорального моделирования тем. Все входящие тексты проходят описанный в Разделе 2.4 пайплайн постобработки в контексте Аналитического GUI (см. Раздел 3.1.6), где каждому документу присваивается превалирующая тема, а затем извлекаются и агрегируются соответствующие лексические признаки.

Так, система визуализирует частотное распределение самых релевантных слов в рамках выбранной темы (Изображение 10). Такая лингвистическая панель служит двум целям:

- Во-первых, это позволяет после разработки системы в ручном режиме провалидировать качество сформированных тем и оценить насколько они уникальны и семантически однородны.

- Во-вторых, эта функциональность может быть весьма полезна при первичном знакомстве финансового аналитика с системой. Финансовый аналитик, впервые работающий с FinABYSS, получает мгновенное представление о содержании каждой темы без глубокого чтения текстов.

Именно последний факт послужил причиной включения данной функциональности в набор основных инструментов GUI.



Рис. 10: Ранжированные по релевантности термины выборки из 12 тем.

Для иллюстрации данного функционала случайным образом была сформирована выборка из 125 000 публикаций и внутри них было выбрано 12 тем для отображения. Среди выборки все темы несомненно несут прикладное значение для финансовых рынков. Причем есть более общие отраслевые темы, такие как:

- Тема №2, связанная с фармацевтической отраслью;
- Тема №13, связанная с сектором искусственного интеллекта;
- Тема №28, связанная с кибербезопасностью;
- Тема №33, связанная с горнодобывающей отраслью;
- Тема №54, связанная с авиационной промышленностью.

С другой стороны в выборку попала и общая тема, которая стоит на повестке дня и в финансах — Тема №56, которая явно относится к ESG. Наконец, есть и узкоспециализированные финансовые темы, которые также были выявлены автоматически:

- Тема №4, связанная с судебными процессами и разбирательствами, которая потенциально является крайне важной и влиятельной в контексте ценообразования стоимости актива;
- Тема №6, связанная с внешнеэкономическими факторами и экономикой, в общем;
- Тема №11, напрямую связанная с рынком активов, а именно криптовалют;
- Тема №15, связанная с правами собственности на компании;
- Тема №21, связанная с валютным рынком;
- Тема №22, связанная с денежными потоками, включая будущие, настоящие и дисконтированные денежные потоки.

Так, мы можем наблюдать за лингвистическими особенностями тематических групп, а также крайне быстро и эффективно изучать как различия между ними, так и конкретные темы вглубь.

С другой стороны, было бы крайне полезно понимать как темы меняются сквозь время, ведь информационное медиапространство крайне нестабильно, а повестка дня в современном обществе меняется крайне быстро. Так, FinABYSS реализует динамическое тематическое моделирование через интерактивный график тематических временных рядов.

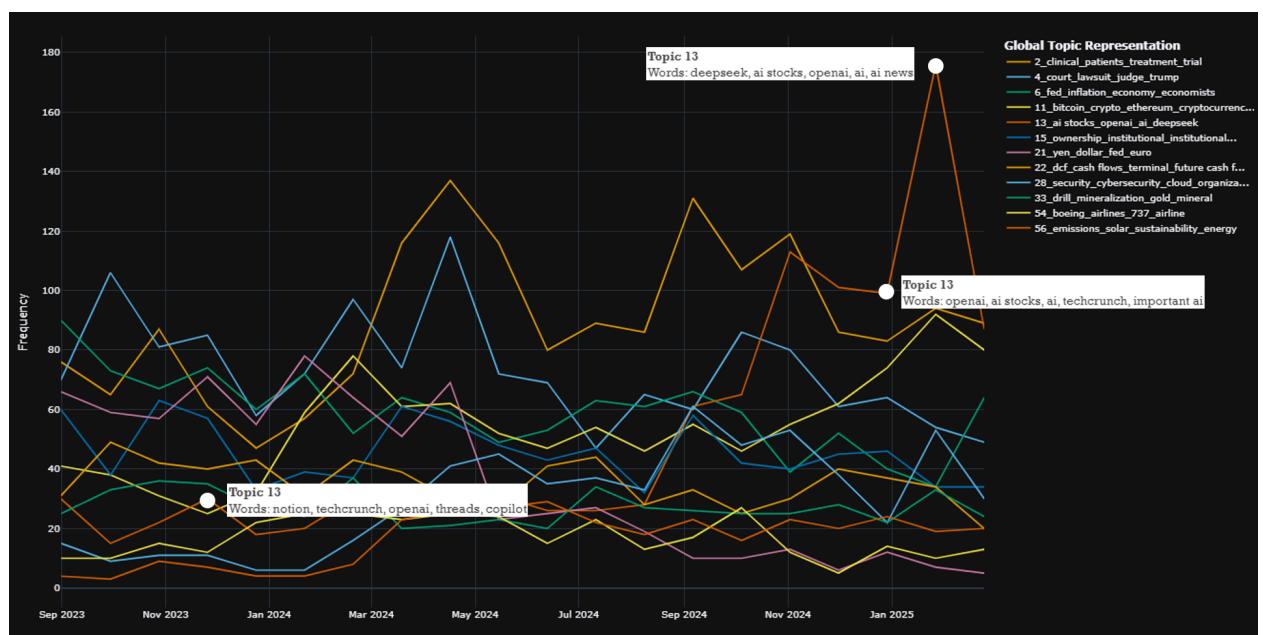


Рис. 11: Эволюционное тематическое моделирование и наиболее релевантные термины выборки из 12 тем за период с 17.09.2023 по 18.04.2025.

На Изображении 11 представлена та же выборка из 12 тем. Ось абсцисс отражает интервал публикаций (настраиваемый: от дневного до месячного или годового), ось ординат — число статей по каждой теме за соответствующий период. При наведении курсора выводятся пять наиболее репрезентативных слов, характеризующих тему в данном временном окне. Пользователь может изменить число отображаемых слов.

Преимущества такого подхода очевидны:

- Раннее обнаружение трендов. Аналитик может наблюдать за появлением новых лексических маркеров или ростом публикационной активности в тематических группах, что служит сигналом зарождающихся событий.
- Отслеживание циклических явлений. Так, для Темы 13 (ИИ) на графике и в декабре 2023 года, и в декабре 2024 года появляется слово TechCrunch — название одной из крупнейших и самых престижных конференций в сферах ИТ и AI. Ежегодно конференция выпускает множество крупнейших стартапов, которые в дальнейшем получают крупное финансирования. Таким образом, имея данную визуализацию, аналитик может меньшими усилиями оставаться в курсе значимых для финансов циклических событий.
- Реакция на форс-мажорные события. В феврале 2025 по той же Теме 13 наблюдается резкий пик из-за нонса китайской LLM DeepSeek R1. Этот инцидент крайне важен, так как в последствии он стал причиной обрушения акций Nvidia, крупнейшей компании поставляющей вычислительное оборудование.

Таким образом, FinABYSS не ограничивается статическим построением тематических кластеров: взаимодействие с лингвистическими метриками и темпоральными трендами превращает систему в универсальную платформу для финансовой аналитики. Эксперт может переходить от макро-трендов к микро-лексическим деталям в несколько кликов, комбинировать фильтры по дате, источнику и тематике, изучать эволюцию терминологии и оперативно реагировать на появление новых ключевых слов или аномальное изменение частот. Это делает FinABYSS не просто инструментом кластеризации, а полноценной экосистемой для семантического мониторинга и прогнозирования воздействия медийных сигналов на динамику финансовых активов.

3.2.4 Прикладное значение

Подводя итоги, стоит подчеркнуть, что разработанная система FinABYSS представляет собой не просто набор моделей, а полноценный инструмент для оперативного обнаружения критических сигналов на финансовых рынках. Семантическая карта, глубокая тематическая кластеризация и временные ряды новостей позволяют финансовому аналитику мгновенно реагировать на неожиданные события, снижать риски и извлекать прибыль.

Так, рассматривая исключительно функциональность Семантической Карты, можно найти несколько критически важных событий, которые сразу же после попадания в сеть своевременно отразились в тематическом кластере «Судебные разбирательства».

Так, первый иллюстративный пример (Изображение 12) демонстрирует, как статья от 22 мая 2024 года повлияла на стоимость акций целевой компании. Данная новость осветила скандал с обвалом стоимости европейских акций по причине недостаточного контроля за трейдинговыми операциями со стороны одного из 4 крупнейших банков в мире — Citi Group (C.NYSE) — акции которого торгуются на Нью-Йоркской фондовой бирже. После данной новости, которая также сообщала о назначении рекордного штрафа банку со стороны британского правительства, за ближайшие 28 часов акции компании упали почти на 4%. В последующие 23 дня, пока шли судебные разбирательства, появлялись дополнительные репортажи и происходила отложенная рыночная реакция, цена уменьшилась почти на 10%.

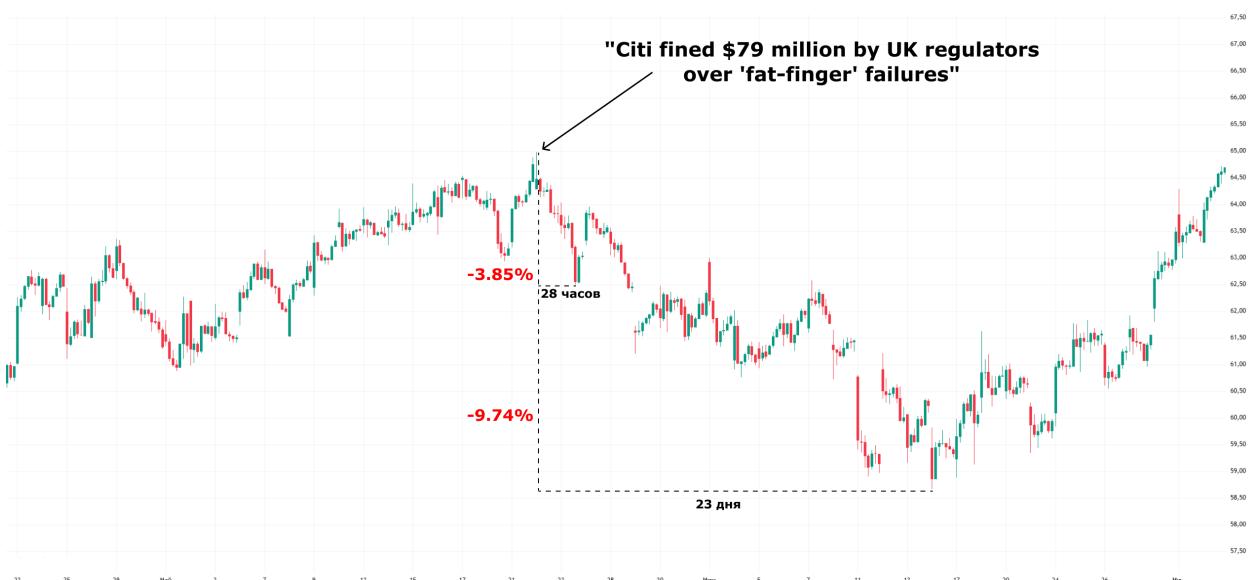


Рис. 12: Пример падения цены акций одного из крупнейших банков США, Citi Group (C.NYSE), из-за обвинений в недостаточном контроле за торговыми операциями, что вызвало падение европейских акций

Сама новостная статья от Reuters послужила сигналом для продажи актива, а на Изображении 12 явно виден переломный момент с резким падением цены, небольшим откатом и дальнейшей сменой глобального тренда почти на месяц.

Без FinABYSS подобные сигналы часто теряются в потоке новостей, в то время как разработанная система упрощает обнаружение триггерных событий, посредством реализации возможности отслеживания значимых для конкретного портфеля кластеров.

Второй кейс связан с публикацией от 30 августа 2024 года о возможном отзыве лицензии и штрафе AU\$ 67 млн в отношении крупнейшего австралийской компании по азартным играм — The Star Entertainment Group LTD (рис. 13). Новость сообщала об обвинениях в отмывании денег. Акции оказались заморожены на месяц, а при возобновлении торгов их

стоимость рухнула на 55%.



Рис. 13: Пример того, как акции австралийской компании Star Entertainment Group LTD (SGR.ASX) обвалились и торговля ими была временно приостановлена из-за судебного разбирательства по отмыванию денег.

Важно отметить, что торговля была приостановлена через сутки после публикации новости, что хоть и является достаточным интервалом для обнаружения сигнала о продаже. Однако для рядового не вовлеченного во внутридневную торговлю инвестора, данный сигнал весьма вероятно был бы незаметен, что привело бы к ужасающей ситуации, когда инвестор не может продать обесценившийся актив и вынужден держать очевидный убыток неопределенно длительный срок.

В обоих случаях, имея систему FinABYSS финансовый аналитик или инвестор мог бы одним из первых узнать о данных инцидентах и незамедлительно отреагировать на рыночные сигналы. Простейшим способом своевременно узнавать о триггерах является модуль активо-ориентированного отслеживания, который позволяет задать фильтры по конкретному тикеру, источнику и тематике, чтобы получать только таргетированные сигналы.

В перспективе, после полной реализации и обучения архитектуры, предложенной в Разделе 3.1, станет возможным настройка фильтрации новостей по их сентименту, а также настройка пользовательского порога для обнаружения значимо позитивных или негативных новостей.

Таким образом, FinABYSS выводит финансовую аналитику на совершенно новый уровень: вместо бессистемного мониторинга СМИ и статей он получает готовые к действию алерты, визуализацию и прогнозную поддержку. Семантическая карта уже сегодня становится неотъемлемой частью рабочего процесса, а по мере расширения модальностей и внедрения предиктивного механизма её ценность будет лишь расти.

It is important to note that trading was halted 24 hours after the news was published, which

is a sufficient interval for a sell signal to be detected. However, for the average investor not involved in intraday trading, this signal would very likely have been undetectable, leading to the dire situation of an investor unable to sell a depreciating asset and forced to hold an obvious loss indefinitely.

In both cases, with FinABYSS, a financial analyst or investor would be among the first to recognize these incidents and react immediately to market signals. The simplest way to learn about triggers in a timely manner is the asset-based tracking module, which allows you to set filters by specific ticker, source, and topic to receive only targeted signals.

Eventually, once the architecture proposed in Section 3.1 is fully implemented and trained, it will be possible to customize the filtering of news by its sentiment, as well as setting a custom threshold to detect meaningfully positive or negative news.

FinABYSS thus takes financial analytics to a whole new level: instead of haphazard monitoring of media and articles, it gets ready-to-action alerts, visualization, and predictive support. The semantic map is already becoming an integral part of the workflow, and its value will only grow with the expansion of modalities and the introduction of a predictive mechanism.

3.3 Аналитическое решение проблемы семантической дедубликации

3.3.1 Математическая постановка

В рамках исследования был разработан новый подход к дедубликации, основанный на анализе семантического одержимого объектов. Хотя в настоящей работе сущностью выступает текст, метод легко обобщается на любые объекты, допускающие векторное представление в семантическом пространстве.

Каждая статья представляется в виде последовательности эмбеддингов:

$$x_i \subset \mathbb{R}^{t \times d} \quad (23)$$

где t — число токенов, а d — размерность семантического векторного пространства. Для последующего анализа вместо набора эмбеддингов используется их выпуклая оболочка, обозначаемая как $\text{CH}(x_i)$ или, сокращенно, CH_i . В качестве меры уникальности текста применяется объем этой оболочки, $\text{vol}(\text{CH}_i)$.

Пересечение выпуклых оболочек. При прямом вычитании пересечений между CH_i и оболочками других текстов может возникнуть проблема множественного учета. Чтобы устранить кратное вычитание, применяется метод включений–исключений.

Обозначим множество всех статей, кроме i , как

$$\mathbb{I} = \{1, \dots, N\} \setminus \{i\}. \quad (24)$$

Пересечение CH_i с оболочками статей, индексированных подмножествами $\mathbb{J} \subseteq \mathbb{I}$, задается выражением:

$$\text{vol} \left(\text{CH}_i \cap \bigcap_{j \in \mathbb{J}} \text{CH}_j \right). \quad (25)$$

Тогда объем пересечения CH_i с объединением оболочек остальных статей вычисляется по формуле:

$$\text{vol} \left(\text{CH}_i \cap \bigcup_{j \in \mathbb{I}} \text{CH}_j \right) = \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J} \subseteq \mathbb{I} \\ |\mathbb{J}|=k}} \text{vol} \left(\text{CH}_i \cap \bigcap_{j \in \mathbb{J}} \text{CH}_j \right). \quad (26)$$

Уникальность статьи i определяется как доля объема её выпуклой оболочки, не задействованная в пересечениях с оболочками других статей:

$$\mu_i = \frac{\text{vol} \left(\text{CH}_i \setminus \bigcup_{j \in \mathbb{I}} \text{CH}_j \right)}{\text{vol} \left(\text{CH}_i \right)}. \quad (27)$$

Преобразуя это определение с учетом разбиения CH_i на область пересечения и её дополнения, получаем:

$$\mu_i = 1 - \frac{\text{vol} \left(\text{CH}_i \cap \bigcup_{j \in \mathbb{I}} \text{CH}_j \right)}{\text{vol} \left(\text{CH}_i \right)}. \quad (28)$$

Подставляя выражение по принципу включений–исключений, окончательная Формула 26 имеет вид:

$$\mu_i = 1 - \frac{1}{\text{vol} \left(\text{CH}_i \right)} \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J} \subseteq \mathbb{I} \\ |\mathbb{J}|=k}} \text{vol} \left(\text{CH}_i \cap \bigcap_{j \in \mathbb{J}} \text{CH}_j \right). \quad (29)$$

Значение $\mu_i \in [0, 1]$ характеризует уникальность текста: $\mu_i = 1$ означает отсутствие пересечений с другими текстами (полная уникальность), а $\mu_i = 0$ указывает на то, что семантический объём текста полностью занят пересечениями с оболочками других текстов.

3.3.2 Преимущества и недостатки

Предлагаемый метод основан на теоретически обоснованном представлении текста: каждая статья трактуется как выпуклая оболочка эмбеддингов токенов, что позволяет четко определить семантическое содержимое объекта, применить принцип включения–исключения, унаследованный из теории множеств, для корректного вычисления объема пересечений, а также нормировать результат так, чтобы итоговая мера уникальности находилась в интервале $[0, 1]$.

У данного подхода помимо теоретической стройности есть целый ряд дополнительных преимуществ:

- Применение эмбеддингов для каждого токена позволяет уловить тонкие различия в семантическом содержании, а агрегирование посредством выпуклой оболочки обеспечивает обобщённое представление о содержании текста. Это даёт возможность сравнивать тексты различной длины и тематики в едином векторном пространстве.
- Нормировка метрики, выраженной числом в интервале $[0, 1]$, упрощает интерпретацию.

С другой стороны, у метода есть и ряд основательных недостатков:

- Выпуклая оболочка в высокоразмерном пространстве (например, 768-мерном) может существенно «растянуться». Это приводит к неинформативности получаемых объёмов, а геометрия оболочек может не отражать сложные распределения эмбеддингов.
- Эмбеддинги имеют сложную, зачастую нелинейную структуру. Выпуклая оболочка, являясь минимально необходимым выпуклым множеством, может включать экстремальные точки, что приводит к переоценке занимаемого пространства и, как следствие, к искажению оценок.
- Из-за того, что эмбеддинги могут содержать случайные шумовые компоненты или артефакты, выпуклая оболочка может быть чувствительна к выбросам. Это приводит к тому, что небольшие неточности в эмбеддингах могут непропорционально увеличить объём выпуклой оболочки, и, соответственно, искажить оценку уникальности.
- Построение выпуклой оболочки и вычисление объёмов в высокоразмерном пространстве является ресурсозатратной задачей. Применение принципа включения–исключения для корректного вычисления пересечений между оболочками текстов усложняет расчёты, особенно при большом количестве документов.

Все перечисленные проблемы могут критичным образом сказаться на использования данного метода на практике, однако некоторых проблем можно избежать инженерным образом.

Проблему чувствительности к шуму (пункт 3) можно частично нивелировать использованием токена [CLS] в качестве центроида выпуклой оболочки. Введение коэффициента δ для нормирования «вогнутости» оболочки в направлении эмбеддинга [CLS] позволяет снижать влияние шумовых компонентов.

Проблему вычислительных затрат (пункт 4) можно решать различными способами:

- Регулирование числа включающих–исключающих пар (гиперпараметр N в суммировании) позволяет получить приближённую оценку уникальности с уменьшением вычислительной нагрузки.

- Применение алгоритмов, таких как UMAP, t-SNE, PCA и других, может привести исходное пространство к более низкой размерности, что значительно сократит затраты на вычисления. При этом необходимо учитывать возможную потерю точности.
- Аппроксимация объёма методом Монте–Карло позволяет получить оценку при уменьшении вычислительных ресурсов.

Метод представления семантической уникальности текста через выпуклые оболочки эмбеддингов обладает рядом теоретических преимуществ (сильная нормировка, применимость принципа включения–исключения и единообразная интерпретация результата). Однако практическое применение требует решения проблем, связанных с высокой размерностью, нелинейностью распределения эмбеддингов и существенными вычислительными затратами. Дальнейшие исследования могут быть направлены на разработку более устойчивых и эффективных методов оценки уникальности с учетом указанных ограничений.

Заключение

Disparity in Access to Financial Resources. During the study, it was found that there exists a significant barrier for individual researchers who lack the financial resources required for expensive data collection, infrastructure rental, and the time needed to develop a system entirely from scratch.

The financial community — which includes news outlets, data aggregators, professional traders, and investment funds — often does not facilitate the development of publicly available tools for extracting value from financial instruments. On the contrary, several market participants deliberately create additional obstacles to free data access, while failing to utilize existing resources efficiently. Examples include:

- **Infrastructure limitations.** Restrictions imposed by aggregators and news services (e.g., Yahoo! Finance) impede large-scale data collection.
- **Closed APIs and high tariffs.** Services such as Google Finance and Yahoo! Finance, along with platforms like Twitter and Seeking Alpha, offer limited functionality or charge high fees for access.
- **Restrictions on access to analytical tools.** Cases such as BloombergGPT illustrate the deliberate concealment of general-purpose tools.
- **Strict copyright policies.** Tighter copyright conditions result in restricted access to various datasets [Wu (и др.), 2023].

Thus, it can be concluded that the financial community contributes to a scarcity of open informational resources by artificially raising the barriers to access with the aim of reducing competition and limiting the number of independent market players.

This issue is not new — it has been repeatedly highlighted in several studies (including by the creators of FinBERT [Y. Yang, UY, A. Huang, 2020]); however, over the past five years the situation has remained virtually unchanged. A crisis also persists in the open-source segment of financial tools.

Despite the widespread restrictive practices, there are proactive participants in the financial sector who strive to distribute information more equitably. For instance, the financial data provider Alpha Vantage²⁴ offers a free and open API that grants access to a vast array of valuable data, including intraday OHLCV. Although Reddit²⁵ is less popular than platform X (ex-Twitter)²⁶ in the financial community, it also provides an open API and can serve as an alternative channel for publishing announcements, opinions, and insider information.

²⁴URL: <https://www.alphavantage.co/>

²⁵URL: <https://www.reddit.com/>

²⁶URL: <https://x.com/>

In addition, aggregators such as FinURLs²⁷ and MarketWatch²⁸ represent important information sources. FinURLs compiles links to historical news from 24 sources over several years. Despite the lack of a dedicated API and certain interface inconveniences for data extraction, this resource remains valuable. At the same time, MarketWatch boasts a more advanced infrastructure by offering not only links to news articles but also quantitative data, as well as the ability to obtain information on specific markets, assets, or indices.

Individual yet significant sources, such as the websites of certain companies and government agencies, also deserve attention. For example, the SEC²⁹ provides free access to historical financial reports (e.g., 10-K and 10-Q) via an RSS feed, thereby promoting more equitable access to information. However, even these open datasets are frequently accompanied by technical challenges: precise timestamps are often missing or the website structure is disrupted, which complicates automated data extraction.

It should be noted that nearly all real-time data are available without significant restrictions, as most services promptly provide such information. Nevertheless, the collection of both historical and real-time data regularly encounters ethical and copyright issues, which remain an important aspect in the practical use of these resources.

In summary, despite various initiatives aimed at expanding access to financial data, the overall landscape is still characterized by artificially high barriers. These restrictions contribute to a shortage of open tools, which in turn reduces market competition and limits opportunities for independent researchers. Therefore, the development of methodologies aimed at the free and equitable dissemination of information remains an urgent task, requiring a comprehensive approach that takes technical, ethical, and legal aspects into account.

Источники литературы

Akiba, T. Optuna: A Next-Generation Hyperparameter Optimization Framework / T. Akiba [и др.] // The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — 2019. — C. 2623—2631.

Alvarado, J. C. S. Domain adaption of named entity recognition to support credit risk assessment / J. C. S. Alvarado, K. Verspoor, T. Baldwin // Proceedings of the australasian language technology association workshop 2015. — 2015. — C. 84—90.

Amid, E. TriMap: Large-scale Dimensionality Reduction Using Triplets / E. Amid, M. K. Warmuth // arXiv preprint arXiv:1910.00204. — 2019. — arXiv: 1910.00204.

Angelov, D. Top2vec: Distributed representations of topics / D. Angelov // arXiv preprint arXiv:2008.09470. — 2020.

²⁷URL: <https://finurls.com/>

²⁸URL: <https://www.marketwatch.com/>

²⁹URL: <https://www.sec.gov/>

- Araci, D.* FinBERT: Financial Sentiment Analysis with Pre-trained Language Models / D. Araci. — 2019. — Авг. — URL: <http://arxiv.org/abs/1908.10063>.
- Au, W.* FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain / W. Au, A. Ait-Azzi, J. Kang // Companion Proceedings of the Web Conference 2021. — Ljubljana, Slovenia : Association for Computing Machinery, 2021. — C. 276—279. — (WWW '21). — ISBN 9781450383134. — DOI: 10.1145/3442442.3451378. — URL: <https://doi.org/10.1145/3442442.3451378>.
- Auger, A.* A restart CMA evolution strategy with increasing population size / A. Auger, N. Hansen // 2005 IEEE congress on evolutionary computation. T. 2. — IEEE. 2005. — C. 1769—1776.
- Beltagy, I.* SciBERT: A pretrained language model for scientific text / I. Beltagy, K. Lo, A. Cohan // arXiv preprint arXiv:1903.10676. — 2019.
- Bentivogli, L.* The Fifth PASCAL Recognizing Textual Entailment Challenge. / L. Bentivogli [и др.] // TAC. — 2009. — Т. 7, № 8. — С. 1.
- Bergstra, J.* Algorithms for hyper-parameter optimization / J. Bergstra [и др.] // Advances in neural information processing systems. — 2011. — Т. 24.
- Blei, D. M.* Latent dirichlet allocation / D. M. Blei, A. Y. Ng, M. I. Jordan // Journal of machine Learning research. — 2003. — Т. 3, Jan. — С. 993—1022.
- Bogdan, R.* Qualitative research for education. T. 368 / R. Bogdan, S. K. Biklen. — Allyn & Bacon Boston, MA, 1997.
- Campello, R. J.* Density-based clustering based on hierarchical density estimates / R. J. Campello, D. Moulavi, J. Sander // Pacific-Asia conference on knowledge discovery and data mining. — Springer. 2013. — С. 160—172.
- Cer, D.* SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation / D. Cer [и др.] // Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). — Association for Computational Linguistics, 2017. — DOI: 10.18653/v1/s17-2001. — URL: <http://dx.doi.org/10.18653/v1/S17-2001>.
- Chen, Z.* Quora question pairs. / Z. Chen [и др.]. — 2017.
- Dao, T.* FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness / T. Dao [и др.]. — 2022. — arXiv: 2205.14135 [cs.LG]. — URL: <https://arxiv.org/abs/2205.14135>.
- Daudert, T.* A multi-source entity-level sentiment corpus for the financial domain: the FinLin corpus / T. Daudert // Language Resources and Evaluation. — 2022. — Т. 56, № 1. — С. 333—356.
- Devlin, J.* Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin [и др.] // 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). — 2019. — С. 4171—4186.

- Dolan, W. B.* Automatically Constructing a Corpus of Sentential Paraphrases / W. B. Dolan, C. Brockett // Proceedings of the Third International Workshop on Paraphrasing (IWP2005). — 2005. — URL: <https://aclanthology.org/I05-5002/>.
- Du, K.* Financial Sentiment Analysis: Techniques and Applications / K. Du [и др.] // ACM Computing Surveys. — 2024. — Окт. — Т. 56, вып. 9. — ISSN 15577341. — DOI: 10.1145/3649451.
- Dudy, S.* A Multi-Context Character Prediction Model for a Brain-Computer Interface / S. Dudy [и др.] // Proceedings of the Second Workshop on Subword/Character LEvel Models / под ред. M. Faruqui [и др.]. — New Orleans : Association for Computational Linguistics, 06.2018. — С. 72—77. — DOI: 10.18653/v1/W18-1210. — URL: <https://aclanthology.org/W18-1210/>.
- Dutt, A.* Shared manifold learning using a triplet network for multiple sensor translation and fusion with missing data / A. Dutt, A. Zare, P. Gader // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. — 2022. — Т. 15. — С. 9439—9456.
- Ester, M.* A density-based algorithm for discovering clusters in large spatial databases with noise / M. Ester [и др.] // kdd. Т. 96. — 1996. — С. 226—231.
- Falkner, S.* BOHB: Robust and efficient hyperparameter optimization at scale / S. Falkner, A. Klein, F. Hutter // International conference on machine learning. — PMLR. 2018. — С. 1437—1446.
- Fedus, W.* Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity / W. Fedus, B. Zoph, N. Shazeer // Journal of Machine Learning Research. — 2022. — Т. 23, № 120. — С. 1—39.
- Feichtenhofer, C.* Convolutional two-stream network fusion for video action recognition / C. Feichtenhofer, A. Pinz, A. Zisserman // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 1933—1941.
- Gao, T.* Simcse: Simple contrastive learning of sentence embeddings / T. Gao, X. Yao, D. Chen // arXiv preprint arXiv:2104.08821. — 2021.
- Grootendorst, M.* BERTopic: Neural topic modeling with a class-based TF-IDF procedure / M. Grootendorst // arXiv preprint arXiv:2203.05794. — 2022.
- Gururangan, S.* Don't stop pretraining: Adapt language models to domains and tasks / S. Gururangan [и др.] // arXiv preprint arXiv:2004.10964. — 2020.
- Halder, S.* FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis / S. Halder. — 2022. — Нояб. — URL: <http://arxiv.org/abs/2211.07392>.
- Hochreiter, S.* Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. — 1997. — Нояб. — Т. 9, вып. 8. — С. 1735—1780. — ISSN 08997667. — DOI: 10.1162/neco.1997.9.8.1735.
- Howard, J.* Universal language model fine-tuning for text classification / J. Howard, S. Ruder // arXiv preprint arXiv:1801.06146. — 2018.

- Huang, A.* FinBERT: A Large Language Model for Extracting Information from Financial Text* / A. Huang, H. Wang, Y. Yang // Contemporary Accounting Research. — 2023. — Май. — Т. 40, вып. 2. — С. 806—841. — ISSN 19113846. — DOI: 10.1111/1911-3846.12832.
- Huang, H.* Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization / H. Huang [и др.] // Communications biology. — 2022. — Т. 5, № 1. — С. 719.
- Jacobs, R.* Adaptive Mixtures of Local Experts / R. Jacobs [и др.] // Neural Computation. — 1991. — Март. — Т. 3. — С. 79—87. — DOI: 10.1162/neco.1991.3.1.79.
- Jain, A. K.* Data clustering: 50 years beyond K-means / A. K. Jain // Pattern recognition letters. — 2010. — Т. 31, № 8. — С. 651—666.
- Jiang, T.* Financial sentiment analysis using FinBERT with application in predicting stock movement / T. Jiang, A. Zeng. — 2023. — Июнь. — URL: <http://arxiv.org/abs/2306.02136>.
- Karpathy, A.* Large-scale Video Classification with Convolutional Neural Networks / A. Karpathy [и др.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 06.2014.
- Kim, J.* Forecasting the S&P 500 Index Using Mathematical-Based Sentiment Analysis and Deep Learning Models: A FinBERT Transformer Model and LSTM / J. Kim, H. S. Kim, S. Y. Choi // Axioms. — 2023. — Сент. — Т. 12, вып. 9. — ISSN 20751680. — DOI: 10.3390/axioms12090835.
- LeCun, Y.* Gradient-based learning applied to document recognition / Y. LeCun [и др.] // Proceedings of the IEEE. — 1998. — Т. 86, № 11. — С. 2278—2324.
- Lee, J.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining / J. Lee [и др.] // Bioinformatics. — 2020. — Т. 36, № 4. — С. 1234—1240.
- Levesque, H. J.* The Winograd schema challenge. / H. J. Levesque, E. Davis, L. Morgenstern // KR. — 2012. — Т. 2012. — 13th.
- Li, L.* Hyperband: A novel bandit-based approach to hyperparameter optimization / L. Li [и др.] // Journal of Machine Learning Research. — 2018. — Т. 18, № 185. — С. 1—52.
- Li, Z.* Towards general text embeddings with multi-stage contrastive learning / Z. Li [и др.] // arXiv preprint arXiv:2308.03281. — 2023.
- Liaw, R.* Tune: A Research Platform for Distributed Model Selection and Training / R. Liaw [и др.] // arXiv preprint arXiv:1807.05118. — 2018.
- Liu, G.* A New Index for Clustering Evaluation Based on Density Estimation / G. Liu. — 2024. — arXiv: 2207.01294 [cs.LG]. — URL: <https://arxiv.org/abs/2207.01294>.
- Liu, G.* Min-Max-Jump distance and its applications / G. Liu // arXiv preprint arXiv:2301.05994. — 2023.
- Liu, Z.* FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining : тех. отч. / Z. Liu [и др.]. — 2020. — URL: <http://commoncrawl.org/>.

- Lu, W.* A CNN-LSTM-based model to forecast stock prices / W. Lu [и др.] // Complexity. — 2020. — Т. 2020, № 1. — С. 6622927.
- Macedo, M.* WWW'18 Open Challenge: Financial Opinion Mining and Question Answering / M. Macedo [и др.] // Companion Proceedings of The Web Conference 2018. — Lyon, France : International World Wide Web Conferences Steering Committee, 2018. — С. 1941—1942. — (WWW '18). — ISBN 9781450356404. — DOI: 10.1145/3184558.3192301. — URL: <https://doi.org/10.1145/3184558.3192301>.
- Malo, P.* Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts / P. Malo [и др.] // Journal of the Association for Information Science and Technology. — 2014. — Т. 65.
- Malzer, C.* A Hybrid Approach To Hierarchical Density-based Cluster Selection / C. Malzer, M. Baum // 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). — IEEE, 09.2020. — С. 223—228. — DOI: 10.1109/mfi49285.2020.9235263. — URL: <http://dx.doi.org/10.1109/MFI49285.2020.9235263>.
- McInnes, L.* hdbscan: Hierarchical density based clustering / L. McInnes, J. Healy, S. Astels // The Journal of Open Source Software. — 2017. — Март. — Т. 2, № 11. — DOI: 10.21105/joss.00205. — URL: <https://doi.org/10.21105%2Fjoss.00205>.
- McInnes, L.* Umap: Uniform manifold approximation and projection for dimension reduction / L. McInnes, J. Healy, J. Melville // arXiv preprint arXiv:1802.03426. — 2018.
- McInnes, L.* UMAP: Uniform Manifold Approximation and Projection / L. McInnes, J. Healy, N. Saul [и др.] // The Journal of Open Source Software. — 2018. — Т. 3, № 29. — С. 861.
- Merity, S.* WikiText: A Large Language Modeling Dataset / S. Merity [и др.] // arXiv preprint arXiv:1602.01376. — 2016.
- Moulavi, D.* Density-based clustering validation / D. Moulavi [и др.] // Proceedings of the 2014 SIAM international conference on data mining. — SIAM. 2014. — С. 839—847.
- Muennighoff, N.* MTEB: Massive Text Embedding Benchmark / N. Muennighoff [и др.]. — 2023. — arXiv: 2210.07316 [cs.CL]. — URL: <https://arxiv.org/abs/2210.07316>.
- Nomura, M.* CMA-ES with learning rate adaptation: Can CMA-ES with default population size solve multimodal and noisy problems? / M. Nomura, Y. Akimoto, I. Ono // Proceedings of the Genetic and Evolutionary Computation Conference. — 2023. — С. 839—847.
- Nomura, M.* cmaes: A simple yet practical python library for cma-es / M. Nomura, M. Shibata // arXiv preprint arXiv:2402.01373. — 2024.
- Nomura, M.* Warm starting CMA-ES for hyperparameter optimization / M. Nomura, S. Watanabe [и др.] // Proceedings of the AAAI conference on artificial intelligence. Т. 35. — 2021. — С. 9188—9196.
- Nothman, J.* Stop word lists in free open-source software packages / J. Nothman, H. Qin, R. Yurchak // Proceedings of workshop for NLP open source software (NLP-OSS). — 2018. — С. 7—12.

- Nussbaum, Z.* Nomic Embed: Training a Reproducible Long Context Text Embedder / Z. Nussbaum [и др.]. — 2024. — arXiv: 2402.01613 [cs.CL].
- Ortega, J. D.* Multimodal fusion with deep neural networks for audio-video emotion recognition / J. D. Ortega [и др.] // arXiv preprint arXiv:1907.03196. — 2019.
- Pathak, A. R.* Application of Deep Learning Approaches for Sentiment Analysis / A. R. Pathak [и др.] // Deep Learning-Based Approaches for Sentiment Analysis / под ред. В. Agarwal [и др.]. — Singapore : Springer Singapore, 2020. — С. 1—31. — ISBN 978-981-15-1216-2. — DOI: 10.1007/978-981-15-1216-2_1. — URL: https://doi.org/10.1007/978-981-15-1216-2_1.
- Rajpurkar, P.* Squad: 100,000+ questions for machine comprehension of text / P. Rajpurkar [и др.] // arXiv preprint arXiv:1606.05250. — 2016.
- Raschka, S.* Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence / S. Raschka, J. Patterson, C. Nolet // arXiv preprint arXiv:2002.04803. — 2020.
- Reimers, N.* Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks / N. Reimers, I. Gurevych // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics, 11.2019. — URL: <https://arxiv.org/abs/1908.10084>.
- Rousseeuw, P.* Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis / P. Rousseeuw // Journal of Computational and Applied Mathematics. — Amsterdam, The Netherlands, The Netherlands, 1987. — Нояб. — Т. 20, № 1. — С. 53—65. — ISSN 0377-0427. — DOI: 10.1016/0377-0427(87)90125-7. — URL: http://svn.donarmstrong.com/don/trunk/projects/research/papers_to_read/statistics/silhouettes_a_graphical_aid_to_the_interpretation_and_validation_of_cluster_analysis_rousseeuw_j_comp_app_math_20_53_1987.pdf.
- Rumelhart, D. E.* Learning representations by back-propagating errors / D. E. Rumelhart, G. E. Hinton, R. J. Williams // nature. — 1986. — Т. 323, № 6088. — С. 533—536.
- Sainburg, T.* Parametric UMAP: learning embeddings with deep neural networks for representation and semi-supervised learning / T. Sainburg, L. McInnes, T. Q. Gentner // ArXiv e-prints. — 2020. — arXiv: 2009.12981 [stat.ML].
- Şenol, A.* VIASCKDE Index: A Novel Internal Cluster Validity Index for Arbitrary-Shaped Clusters Based on the Kernel Density Estimation / A. Şenol // Computational Intelligence and Neuroscience. — 2022. — Т. 2022, № 1. — С. 4059302.
- Shah, R. S.* When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain / R. S. Shah [и др.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2022.

- Shahriari, B.* Taking the human out of the loop: A review of Bayesian optimization / B. Shahriari [и др.] // Proceedings of the IEEE. — 2015. — Т. 104, № 1. — С. 148—175.
- Shazeer, N.* Outrageously large neural networks: The sparsely-gated mixture-of-experts layer / N. Shazeer [и др.] // arXiv preprint arXiv:1701.06538. — 2017.
- Sinha, A.* Impact of news on the commodity market: Dataset and results / A. Sinha, T. Khandait // Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2. — Springer. 2021. — С. 589—601.
- Socher, R.* Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher [и др.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — Seattle, Washington, USA : Association for Computational Linguistics, 10.2013. — С. 1631—1642. — URL: <https://www.aclweb.org/anthology/D13-1170>.
- Vanstone, B. J.* Do news and sentiment play a role in stock price prediction? / B. J. Vanstone, A. Gepp, G. Harris // Applied Intelligence. — 2019. — Нояб. — Т. 49, вып. 11. — С. 3815—3820. — ISSN 15737497. — DOI: 10.1007/s10489-019-01458-9.
- Vaswani, A.* Attention is all you need / A. Vaswani [и др.] // Advances in neural information processing systems. — 2017. — Т. 30.
- Vijayan, D.* Adaptive hierarchical density-based spatial clustering algorithm for streaming applications / D. Vijayan, I. Aziz // Telecom. Т. 4. — MDPI. 2022. — С. 1—14.
- Vuković, D. B.* Predictive Patterns and Market Efficiency: A Deep Learning Approach to Financial Time Series Forecasting / D. B. Vuković [и др.] // Mathematics. — 2024. — Окт. — Т. 12, вып. 19. — ISSN 22277390. — DOI: 10.3390/math12193066.
- Wang, A.* GLUE: A multi-task benchmark and analysis platform for natural language understanding / A. Wang [и др.] // arXiv preprint arXiv:1804.07461. — 2018.
- Wang, Y.* Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization / Y. Wang [и др.] // Journal of Machine Learning Research. — 2021. — Т. 22, № 201. — С. 1—73. — URL: <http://jmlr.org/papers/v22/20-1061.html>.
- Warner, B.* Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference / B. Warner [и др.]. — 2024. — Дек. — URL: <http://arxiv.org/abs/2412.13663>.
- Watanabe, S.* Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance / S. Watanabe // arXiv preprint arXiv:2304.11127. — 2023.
- Wen, Q.* Transformers in time series: A survey / Q. Wen [и др.] // arXiv preprint arXiv:2202.07125. — 2022.
- Wiebe, J.* Annotating expressions of opinions and emotions in language / J. Wiebe, T. Wilson, C. Cardie // Language resources and evaluation. — 2005. — Т. 39. — С. 165—210.

- Williams, A.* A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference / A. Williams, N. Nangia, S. R. Bowman. — 2018. — arXiv: 1704 . 05426 [cs . CL]. — URL: <https://arxiv.org/abs/1704.05426>.
- Wu, S.* Bloomberggpt: A large language model for finance / S. Wu [и др.] // arXiv preprint arXiv:2303.17564. — 2023.
- Xing, F.* Financial sentiment analysis: An investigation into common mistakes and silver bullets / F. Xing [и др.] // Proceedings of the 28th international conference on computational linguistics. — 2020. — С. 978—987.
- Yang, Y.* FinBERT: A Pretrained Language Model for Financial Communications / Y. Yang, M. C. S. UY, A. Huang. — 2020. — Июнь. — URL: <http://arxiv.org/abs/2006.08097>.
- Zhang, X.* mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval / X. Zhang [и др.] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. — 2024. — С. 1393—1412.
- Zhu, Y.* Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books / Y. Zhu [и др.] // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — 2015.