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

Contents

Introduction

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 (et al.), 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 (et al.), 2019; Warner (et al.), 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 project[1]. The domain-adapted model[2] and the collected corpus[3] are also publicly available.

[1]URL: `https://github.com/denisalpino/FinABYSS`
[2]URL: None
[3]URL: `https://huggingface.co/datasets/denisalpino/YahooFinanceNewsRaw`

CHAPTER 1. THEORETICAL OVERVIEW

1.1 Artificial Intelligance in Finance

1.1.1 Price Prediction

There exists a variety of approaches demonstrating the effectiveness of asset price prediction using deep neural networks. Recurrent neural networks (RNN) are undeniably the most prevalent for this task, while convolutional neural networks (CNN) are often employed as auxiliary components.

The long short-term memory (LSTM) [Hochreiter, Schmidhuber, 1997] architecture is the most representative member of the RNN family and is frequently applied to price forecasting. Its adaptations — such as bidirectional LSTM (Bi-LSTM) and hybrid LSTM+CNN models — also exhibit high performance. Furthermore, with the advent of the Transformer [Vaswani (et al.), 2017] architecture, research has increasingly focused on adapting it to the characteristics of time series [Wen (et al.), 2022].

Among recent experiments, the following are particularly noteworthy:

- A repository demonstrating the potential of the Transformer architecture for Bitcoin price forecasting[4];
- A study comparing the performance of Bi-LSTM, hybrid Bi-LSTM+CNN, and Transformer models in predicting IBM stock prices[5];
- An LSTM-based trading bot designed to capture short-term profits during sideways price movements of the CGEN asset, which achieved a 4% daily return on deposit in backtesting[6].

Nevertheless, individual models — whether LSTM-based or decision tree–based — have limited adaptability to changing market regimes and struggle to adjust to their dynamics [Vuković (et al.), 2024]. Although this limitation is evident, it is often underestimated in the literature. According to the Efficient Market Theory (EMT) [Fama, 1970], asset prices incorporate all available market information, which casts doubt on the feasibility of precise forecasting using only historical quantitative indicators (OHLC data, trading volume, and classical technical indicators).

Moreover, most models fail to account for nuanced factors such as limit order book dynamics, interactions with other trading algorithms, and qualitative off-exchange information. Recent studies indicate that integrating analysis of the off-exchange information environment into predictive models significantly enhances forecast accuracy, as will be detailed in the subsequent sections.

[4]URL: `https://github.com/baruch1192/-Bitcoin-Price-Prediction-Using-Transformers`
[5]URL: `https://github.com/JanSchm/CapMarket`
[6]URL: `https://github.com/roeeben/Stock-Price-Prediction-With-a-Bot`

1.1.2 Sentiment Analysis

As noted in the Introduction, the emergence of the Transformer architecture has enabled deep learning models to achieve significant progress in natural language understanding (NLU). This capability is of particular value in finance, where traditional quantitative data alone prove insufficient for precise forecasting.

Prior to the widespread adoption of modern language models for sentiment analysis (SA), various approaches were employed — LSTM–based networks, ULMFiT, autoregressive architectures, and others [Hochreiter, Schmidhuber, 1997; Howard, Ruder, 2018]. For example, one study compared an autoregressive model without sentiment features to an otherwise identical model augmented with sentiment inputs; it demonstrated that in 77.8% of cases the sentiment–aware version outperformed its purely quantitative counterpart [Vanstone, Gepp, Harris, 2019].

Modern pre-trained Transformer models, commonly referred to as large language models (LLMs), open new horizons for financial forecasting. The most intuitive application is the SA of news. This process requires the collection of a substantial corpus of texts, which are then annotated with one of three labels: -1 (negative), 0 (neutral), or 1 (positive) [Pathak (et al.), 2020].

Manual annotation typically involves domain experts capable of assessing a text's impact on financial markets. To ensure high data quality, the cross-consensus method [Bogdan, Biklen, 1997] is often applied: multiple experts independently label the same texts, and their labels are reconciled. This methodology underpinned the creation of the widely used FinancialPhraseBank dataset for fine-tuning models to financial sentiment analysis (FSA) tasks [Malo (et al.), 2014].

Algorithmic labeling methods based on asset price dynamics are less common, owing to their subjectivity and instability. In particular, it is difficult to define a reliable pric–change threshold for sentiment classification and to guarantee that observed gains are not driven by extraneous factors.

Once annotation is complete — a process consuming the majority of resources — pre-trained language models are fine-tuned by adjusting the weights of the final neural layers.

The SA taxonomy comprises three levels [Pathak (et al.), 2020]:

- **Document level.** Sentiment is evaluated across an entire document (e.g., a news article or report). This level assumes opinions pertain to a single entity, which seldom reflects reality. Furthermore, LLMs are technically constrained by maximum token counts and therefore cannot process multi-page filings (such as 10-K reports) or lengthy news articles in full. Recent advances have improved the handling of standard news articles, yet comprehensive reports remain beyond reach.

- **Sentence level.** Sentiment is assigned to individual sentences or short passages (typically 1–3 sentences), still under the assumption of a single subject. Most SA research operates at this level, supported by abundant datasets and well within the token processing capacity of modern LLMs. Common sources include news headlines and social-media posts on

platforms such as X (formerly Twitter) and Reddit.

- **Aspect level (ABSA).** This level captures sentiment toward specific aspects of an entity. ABSA is the most advanced SA tier and is treated as a distinct field encompassing four subtasks: aspect extraction, aspect-polarity assignment, aspect-category detection, and category-polarity assignment. A specialized variant, targeted ABSA, permits multiple entities but restricts sentiment to one per entity. Further discussion of ABSA appears in Section 1.1.3.

Most SA methodologies require extensive text annotation, which represents a major drawback due to high resource demands and inherent limitations:

- No universal, precise definition of "sentiment" fits every application, forcing bespoke definitions for each task — a challenge given the diversity of textual patterns and contexts.
- Without significant investments in quality control (e.g., cross-consensus [Bogdan, Biklen, 1997]), it is impossible to guarantee annotation reliability and objectivity, since human factors heavily influence label consistency.

As mentioned, SA's formulation may vary across domains. FSA is distinguished by its reliance not only on textual data but also on the abundant quantitative information present in financial publications [Du (et al.), 2024].

Nevertheless, FSA encounters failures in several common scenarios, including: irrealis moods (conditional, subjunctive or imperative moods), rhetorical devices (negative assertions, personification, sarcasm), dependent opinions, unspecified aspects, out-of-vocabulary terms (jargon, microtext, named entities), external references (allusions to knowledge not encoded in the model) [Xing (et al.), 2020].

1.1.3 Aspect Analysis

«To be done later»

1.2 Machine Learning Algorithms

1.2.1 Dimensionality Reduction

UMAP & TriMAP | PCA | t-SNE

1.2.2 Clustrering

HDBSCAN | Agglomerative k-Means

1.3 Deep Neural Networks

1.3.1 Models

LSTM [Hochreiter, Schmidhuber, 1997]

«...»

BERT [Devlin (et al.), 2019]

«...»

FinBERT (2019 — first) [Araci, 2019]

«...»

FinBERT (2020 — good) [A. Huang, H. Wang, Y. Yang, 2023; Y. Yang, UY, A. Huang, 2020]

«...»

FinBERT (2020 — best) [Liu (et al.), 2020]

«...»

ModernBERT [Warner (et al.), 2024]

«...»

1.3.2 Techniques

Domain-Adaptive Pretraining (DAPT). — [Gururangan (et al.), 2020]

Based on the calculations, DAPT provides on average a 4% increase in benchmarks on a relative scale compared to the baseline model, which was not domain-specific. This figure is quite significant, considering that for some specific tasks the gains can be as high as 20%, as shown in the paper [Ibid.]

Fusion Mechanisms. In early fusion, features are integrated as soon as they are extracted using Joint representation or Coordinated representation.

In late fusion, integration is performed only after each unimodal network produces a prediction (classification, regression). Late fusion usually uses voting schemes, weighted averages and other techniques

There are also hybrid fusion methods. They combine early fusion results and results from unimodal predictors using late fusion.

Approaches to representation and clustering of density embeddings. [CLS] token and Mean-pooling

1.4 Evaluation

1.4.1 General Language Understanding Evaluation (GLUE)

The General Language Understanding Evaluation (GLUE) benchmark constitutes a standardized framework for assessing the language comprehension capabilities of natural language

processing (NLP) models [A. Wang (et al.), 2018]. It comprises 9 tasks encompassing classification, semantic similarity evaluation, and textual entailment recognition. Through the diversity of these tasks, GLUE facilitates the identification of models' ability to generalize and effectively transfer learned representations across a range of linguistic challenges.

The primary components of GLUE include:

- **A suite of nine tasks**, each derived from pre-existing corpora and targeting distinct aspects of language understanding (e.g., linguistic acceptability, sentiment analysis, paraphrase detection).
- **A diagnostic dataset** for an in-depth evaluation of model performance in capturing various linguistic phenomena.
- **A public leaderboard and dashboard** that enable continuous tracking of benchmark performance and provide visualization of model results on the diagnostic tasks.

Below is a summary table of the key characteristics of the datasets included in the GLUE benchmark:

Table 1: Overview of GLUE Benchmark Datasets

Name	Task	Source	Size	Metric
CoLA	Single-Sentence Classification	[Dudy (et al.), 2018]	∼8 500	Matthews Correlation Coefficient
SST-2	Binary Single-Sentence Sentiment Classification	[Socher (et al.), 2013]	∼67 000	Accuracy
MRPC	Paraphrase Identification	[Dolan, Brockett, 2005]	∼3 700	Accuracy, F1
STS-B	Semantic Textual Similarity (Regression)	[Cer (et al.), 2017]	∼7 000	Pearson/Spearman Correlation
QQP	Duplicate Question Detection (Quora Question Pairs)	[Chen (et al.), 2017]	∼364 000	Accuracy, F1
MNLI	Multi-Genre Natural Language Inference	[Williams, Nangia, Bowman, 2018]	∼393 000	Accuracy
QNLI	Inference Tasks	[Rajpurkar (et al.), 2016]	∼105 000	Accuracy
RTE	Recognizing Textual Entailment	[Bentivogli (et al.), 2009]	∼2,500	Accuracy
WNLI	Inference Tasks	[Levesque, Davis, Morgenstern, 2012]	634	Accuracy

In summary, the GLUE benchmark provides a robust foundation for evaluating both standard and domain-adapted NLP models. Its comprehensive design and the inclusion of diverse linguistic tasks allow for a nuanced analysis of model capabilities. Following this overview, the FLUE benchmark, which is tailored for the evaluation of models in the financial context, will be discussed to further complement the assessment of domain-adaptive pre-training strategies.

1.4.2 Financial Language Understanding Evaluation (FLUE)

The Financial Language Understanding Evaluation (FLUE) benchmark is a domain-specific analog to the GLUE benchmark, tailored specifically for the financial domain [Shah (et al.), 2022]. This benchmark was developed very recently based on 5 diverse datasets. Its creation was driven by the need to evaluate models capable of effectively processing financial texts, as standard general-purpose datasets often fail to capture the unique characteristics of financial lexicon and the specific tasks inherent to this domain.

Table 2: Overview of FLUE Benchmark Datasets

Name	Task	Source	Size (Train/Val/Test)	Metric	License
FPB	Sentiment Classification	[Malo (et al.), 2014]	3488 / 388 / 969	Accuracy	CC BY-SA 3.0
FiQA SA	Sentiment Analysis	[Macedo (et al.), 2018; Shah (et al.), 2022]	822 / 117 / 234	MSE	Public
NHC	News Headlines Classification	[Sinha, Khandait, 2021]	7989 / 1141 / 2282	Avg F1 Score	CC BY-SA 3.0
FinNER	Named Entity Recognition	[Alvarado, Verspoor, Baldwin, 2015]	932 / 232 / 302	F1 Score	CC BY-SA 3.0
FinSBD3	Structure Boundary Detection	[Au, Ait-Azzi, Kang, 2021]	460 / 165 / 131	F1 Score	CC BY-SA 3.0
FiQA QA	Question Answering	[Macedo (et al.), 2018; Shah (et al.), 2022]	5676 / 631 / 333	nDCG, MRR	Public

FLUE covers 5 distinct financial tasks, which allow for a comprehensive evaluation of model performance across various aspects of financial language. The statistics presented in Table 2 demonstrate the scale and diversity of the included datasets. Moreover, all datasets that comprise FLUE are characterized by low ethical risks and do not contain confidential information regarding any organization or individual. In addition, explicit consent was obtained from the authors of each dataset prior to their inclusion in the benchmark, underscoring its legitimacy and ethical soundness.

The emergence of the FLUE benchmark is driven by the necessity to standardize the evaluation of models in the field of financial language understanding. The financial sector imposes unique requirements for processing textual data, such as high terminological complexity, market dynamism, and specific tasks (e.g., sentiment analysis of news headlines, information extraction, etc.). These factors have led to the creation of a heterogeneous set of tasks unified within FLUE, thereby enabling a holistic assessment of different models. Thus, FLUE serves as an essential tool for researchers, facilitating objective model comparison and the identification of areas for further improvement in financial NLP approaches.

In the context of this work, FLUE provides an objective benchmark for assessing model quality. However, despite the aforementioned advantages of this benchmark, it still suffers from a general limitation: it is designed for models with a context window of 512 tokens. Consequently, FLUE may not fully reveal the true potential of models that are capable of processing longer contexts compared to the more constrained models of previous generations such as BERT, FinBERT, ELECTRA, and others.

1.4.3 Clustering Evaluation Metrics

Silhouette Index + Stability Index

CHAPTER 2. PRACTICAL SOLUTION

2.1 Limitaions

2.1.1 De-duplication

«...»

2.1.2 Computing Hardware

«...»

2.1.3 Data

«...»

2.1.4 Context (Attention Mechanism)

«...»

2.2 Data Governance

2.2.1 Data Requirements

One of the study's key objectives was data collection. As noted, the corpus was constructed to meet requirements for universality and to support future research in related fields.

Data Requirements. Texts were chosen as the primary source of qualitative data. In the financial domain, the most content-rich and impactful on asset prices are:

- news articles;
- social-media posts;
- official reports (annual, quarterly, strategic);
- press releases;
- analytical reviews and articles;
- transcripts of interviews, conferences, and public-company webcasts.

Previous studies have validated the effectiveness of these text types. FinBERT [A. Huang, H. Wang, Y. Yang, 2023; Y. Yang, UY, A. Huang, 2020], for example, was trained on official reports and analytical articles, and its later versions incorporated press releases [Liu (et al.), 2020]. Other models have been successfully pre-trained on fragmented news articles and fine-tuned for sentiment analysis on news headlines and social-media publications [Araci, 2019]. Nevertheless,

our research deliberately processes full texts without fragmentation: official reports frequently exceed the 8,192 token limit of ModernBERT, complicating their integration, while very short formats (social-media posts) fail to exploit the advantages of a long-context model.

Furthermore, no single source aggregates all of the above text types. Given limited resources, the following most significant categories were selected for initial focus:

- news articles;
- press releases;
- analytical reviews and articles;
- transcripts of financial events.

Expansion of the corpus to include additional content categories is planned for future work.

To ensure the corpus's versatility and facilitate subsequent use, the following metadata were collected for each text:

- Headline (e.g., "Covestro board enters formal talks on $12 billion ADNOC approach");
- Source (copyright holder), e.g., Reuters, Simply Wall St., PR Newswire, @ilyasut, Max Gottich;
- Publication platform, e.g., Twitter, Yahoo! Finance, Reddit, Seeking Alpha;
- UTC timestamp with second-level precision (e.g., 2024-09-01T01:48:13);
- Author-assigned topical tags (e.g., ["M&A", "Cryptocurrency", "Tech"])'
- List of tickers mentioned by the author (e.g., ["9626.HK", "BILI"]).

Finally, the optimal period for data collection was established. The lower and upper bounds of the dates were selected based on the assumption that this corpus will be used to train the value prediction model in the future, which requires that the initial knowledge of ModernBERT be synchronized with those on which it will be fine-tuned later. Due to the fact that the ModernBERT publication does not disclose the data on which the model was trained [Warner (et al.), 2024], it is impossible to accurately judge for which period the data was taken. Therefore, in our study, focusing on the date of ModernBERT publication [Ibid.] and the classical ratio of training, validation and test samples as 75/15/10, we took the date range from September 17, 2023 to March 18, 2025, i.e. 548 days, of which 374 are working days according to the US calendar.

Source Requirements. Data sources must be open, free, English-language, and authoritative, since wide dissemination and timely publication directly affect market reactions. Considered sources included:

- News outlets: Bloomberg[7], The New York Times[8], Reuters[9], etc;
- Analytical platforms: Seeking Alpha[10], TradingView[11];
- Official sites: Corporate and government portals.

An analysis of over 50 corporate and more than 100 government resources showed that, thanks to RSS feeds, press releases are centrally aggregated via PR Newswire[12] and GlobeNewswire[13]. Other automated aggregators (e.g., Business Wire[14]) exist, but PR Newswire and GlobeNewswire empirically dominate; however, they offer press releases only, without genre diversity.

Among news outlets, Reuters proved optimal in responsiveness and market coverage. Niche but high-quality sources (e.g., The Information[15], Epoch AI[16]) were also examined; they provide overly specialized content, whereas traditional outlets such as Bloomberg, The Wall Street Journal[17], and The Economist[18] cover a broader market spectrum, albeit with technical constraints.

Of the analytical platforms, Seeking Alpha was excluded due to its large volume of paywalled content, and TradingView was unsuitable because it does not grant access to historical publications.

Thus, PR Newswire and Reuters became the corpus's primary sources. To mitigate the risk of systematic bias and ensure broader coverage, it was nevertheless decided to enrich the corpus with additional publishers. Due to their fragmentation, lack of APIs, and often incomplete metadata (including sub-second timestamps), aggregation and synchronization proved unfeasible. Consequently, the aggregators Google Finance[19], Yahoo! Finance[20], and FinURLs[21] were also considered.

Ultimately, Yahoo! Finance was selected thanks to its unified site structure and comprehensive aggregation of diverse sources, whereas FinURLs redirects to individual source sites, each with its own layout. Google Finance, though similar to Yahoo! Finance, does not support the collection of historical data.

[7] URL: `https://www.bloomberg.com/`
[8] URL: `https://www.nytimes.com/`
[9] URL: `https://www.reuters.com/`
[10] URL: `https://seekingalpha.com/`
[11] URL: `https://tradingview.com/`
[12] URL: `https://www.cision.com/`
[13] URL: `https://www.globenewswire.com/`
[14] URL: `https://www.businesswire.com/`
[15] URL: `https://www.theinformation.com/`
[16] URL: `https://epoch.ai/`
[17] URL: `https://www.wsj.com/`
[18] URL: `https://www.economist.com/`
[19] URL: `https://www.google.com/finance/`
[20] URL: `https://finance.yahoo.com/`
[21] URL: `https://finurls.com/`

2.2.2 Data Collecting

Prior to data collection, a survey and analysis of existing open-source tools for harvesting data from Yahoo! Finance were conducted. The analysis identified five candidate libraries. Two — yahooquery and yahoo-stock-api — do not support article extraction; two others — yahoo_fin and fin-news — are abandoned and no longer function correctly; and yfinance affords access only to the latest twenty news items in real time. Consequently, a custom Python parser was developed. Its architecture comprises two principal stages:

1. Link Collection. A recursive traversal of the official sitemap is used to gather article URLs for a specified period. Each "daily page" lists 50 news links and a pointer to the next page; critically, page n can only be accessed via page $n-1$, creating a bottleneck akin to traversing a linked list under high network latency.

2. Content Extraction. The gathered URLs are then parsed to extract each article's text. It should be noted that, as with training the original BERT model, tables and images are not processed [Devlin (et al.), 2019].

During development, several constraints were encountered and subsequently addressed in the parser's design:

- IP-blocking and Cookies. Yahoo! Finance limits to 14 concurrent requests per IP at minimum 4-second intervals; violations yield HTTP 404, 429, or 200 responses with empty bodies. Even when these constraints are met, blocks may still occur. To mitigate this, a pool of 50 proxy servers was employed, and failed requests were automatically retried in subsequent iterations.

- Regional Restrictions. Identical URLs may be inaccessible or behave inconsistently when requested from different countries.

- Technical Errors. Redirects to external sources, broken links, and paywalled URLs were encountered and excluded during corpus assembly.

To accelerate processing of large datasets, the C-based library selectolax was used, offering roughly 30× the speed of BeautifulSoup and 5× that of lxml.

As a result, the link-collection stage yielded 1,362,103 URLs, of which 1,360,761 belonged to the Yahoo! Finance domain. Thanks to the parser's modular architecture, extensive proxy usage, and multiple iterations, 1,304,717 articles were successfully parsed. The final corpus occupies 6.5 GB in CSV format and 2.2 GB in the more compact Parquet format.

The final class implementing the parser program can be found in the official repository of the project, called YahooFinanceParser[22].

[22]URL: `https://github.com/denisalpino/FinABYSS`

2.2.3 Data Analysis

Before commencing the pre-processing stage, it was decided to conduct a comprehensive analysis of the collected corpus of news articles. This preliminary analysis not only revealed the characteristic features of the data but also established the foundation for subsequent automation of text cleaning and structuring. Moreover, the analysis results have also impacted the quality of the trained model.

Local analysis encompassed a detailed examination of various subsets of the corpus aimed at identifying patterns characteristic of non-representative or "noisy" articles. In this process, key signals—such as specific keywords in the titles and opening paragraphs—were identified that allow for the automatic filtering out of undesirable texts. Furthermore, the local investigation uncovered potential rules for removing marketing fragments, metadata, and other artifacts that adversely affect data quality. All the obtained rules were subsequently formalized (see Section 2.2.4 for further details).

Global analysis is dedicated to studying the central tendencies of the corpus through descriptive statistics and the analysis of various data representations—both metadata and the textual content itself. This approach enabled the evaluation of the distribution of key characteristics, the identification of seasonal and thematic patterns, and the preparation of aggregated results that serve as the basis for further refinement of the pre-processing methodology.

Below are the aggregated results of the global analysis, which, together with the local findings, allow for a deeper understanding of the nature of the collected dataset and help determine directions for its further optimization.

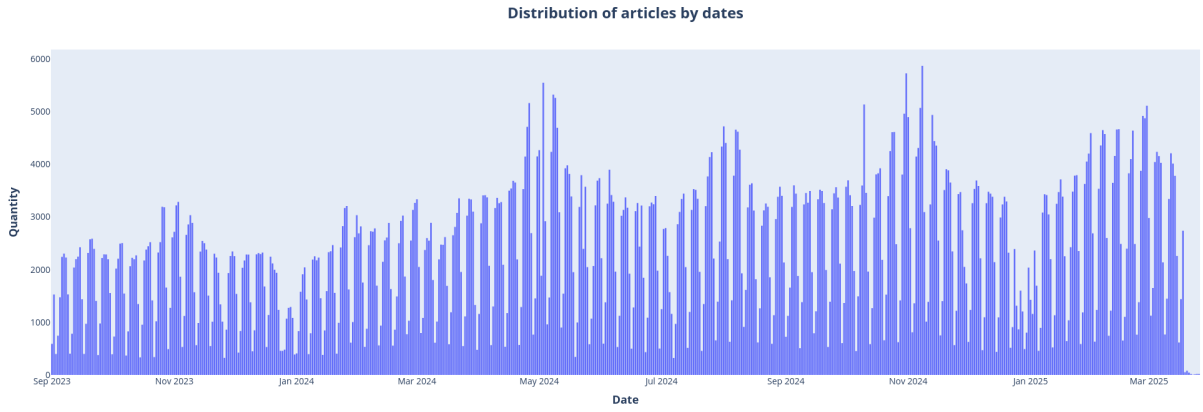

Figure 1: Distribution of publications by dates.

Distribution of publications by dates. Figure 1 shows that the number of publications fluctuates daily with a certain periodicity. A detailed examination revealed that the minimums occur on Sundays and public holidays, when fewer financial news items are published. This naturally reflects the market's characteristics: on weekends and holidays, business activity declines, leading to fewer publications.

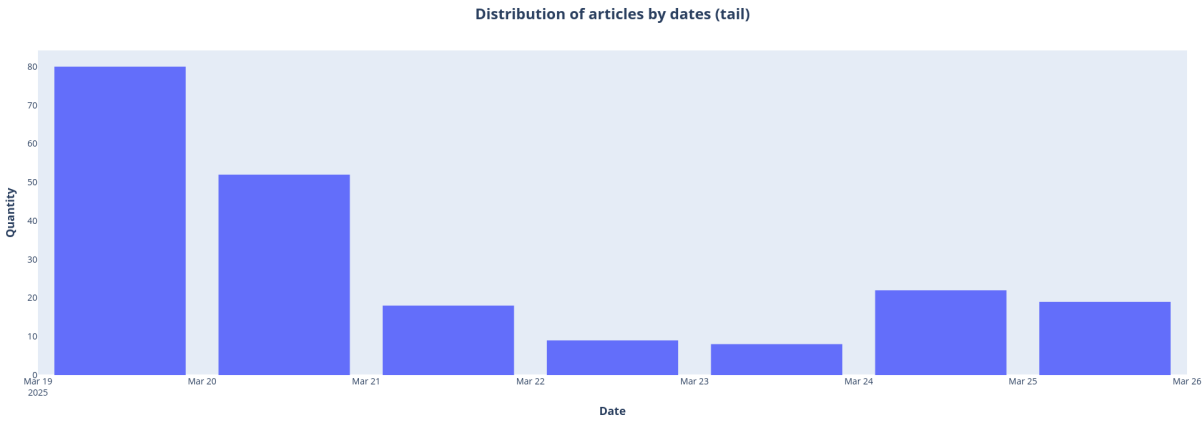

Figure 2: Distribution of publications by dates (tail).

At the same time, some articles, by formal characteristics, fall outside the collection period (September 17, 2023 - March 18, 2025). Figure 2 displays these "tail" publications, whose count slightly exceeds 200. A more detailed analysis determined that these articles were indeed published within the specified interval, but their content was later edited or supplemented. As a result, the publication date and time on the corresponding website were updated, and the old version (with the original date) was lost. Had the links been parsed not after one week but several weeks later, more such cases would have been observed.

From the perspective of short-term market forecasting, this circumstance may lead to distorted timestamps, making some articles appear to have been published later than they actually were. Therefore, the dataset might prove less effective for short-term studies compared to medium- and long-term ones (where a shift of a couple of days is less critical). Nevertheless, for this work it does not play a crucial role, as the model relies solely on the text of the article and does not take into account the precise publication timestamps.

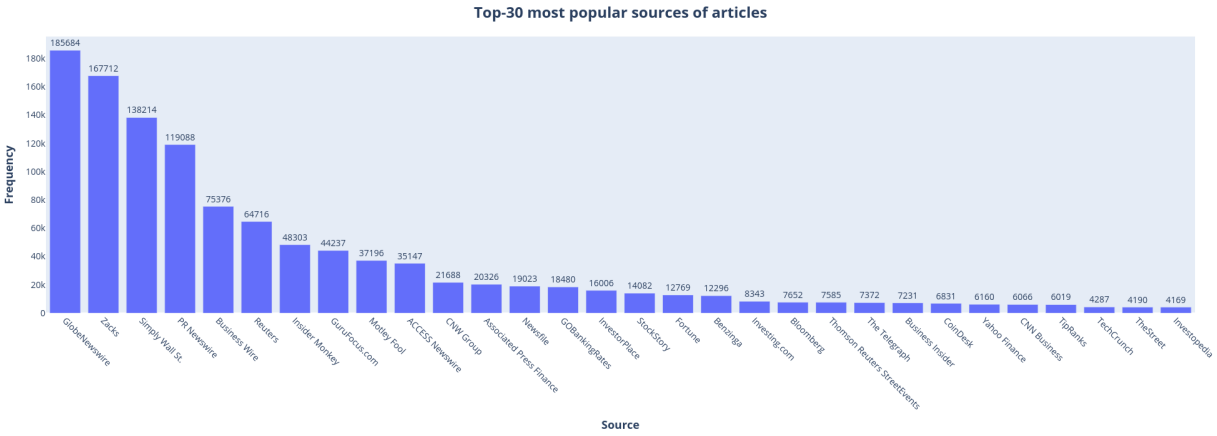

Figure 3: Top-30 most popular sources of publications.

News Sources. Figure 3 illustrates the distribution of publications across the 30 most frequent sources. The analysis showed that the predominant share of articles (potentially 69.2%)

was published by semi-automated aggregators: GlobeNewswire, Zacks, Simply Wall St., PR Newswire, Business Wire, GuruFocus.com, Motley Fool, among others. These aggregators focus on the automatic collection of key data from various resources (regulators, official company websites, etc.), publishing press releases, brief report summaries, and invitations to corporate events.

Among the top 15 sources, only some can be conditionally considered as "traditional" news outlets, such as Reuters, Insider Monkey, CNW Group, Associated Press Finance, and InvestorPlace. Meanwhile, outside the top 30, classic publications that primarily publish original articles prevail. In reality, the blurred boundary between original and semi-automatically generated content complicates efforts to clearly differentiate them.

According to approximate estimates, out of 1 300 000 articles, about 900 000 (69.2%) are semi-automated. This is an important factor for training a language model because:

1. The quality of such materials is often lower: texts contain artifacts, broken formatting, and incorrectly inserted characters.

2. Their volume is large, which, on one hand, provides a substantial sampling capacity, but on the other, complicates cleaning and normalization without the loss of significant information.

Nevertheless, even "imperfect" texts from aggregators convey useful information about the financial market and companies. However, it is extremely important to develop appropriate cleaning and pre-processing rules (discussed in detail in Section 2.2.4) to preserve the semantic integrity of the texts.

Furthermore, and perhaps more importantly, these semi-automated texts contribute roughly the same total number of tokens as the "original" articles (30.8%), despite their numerical dominance. Consequently, with proper processing, this group of semi-automated articles can make a significant contribution to training the language model without diminishing the value of the original texts.

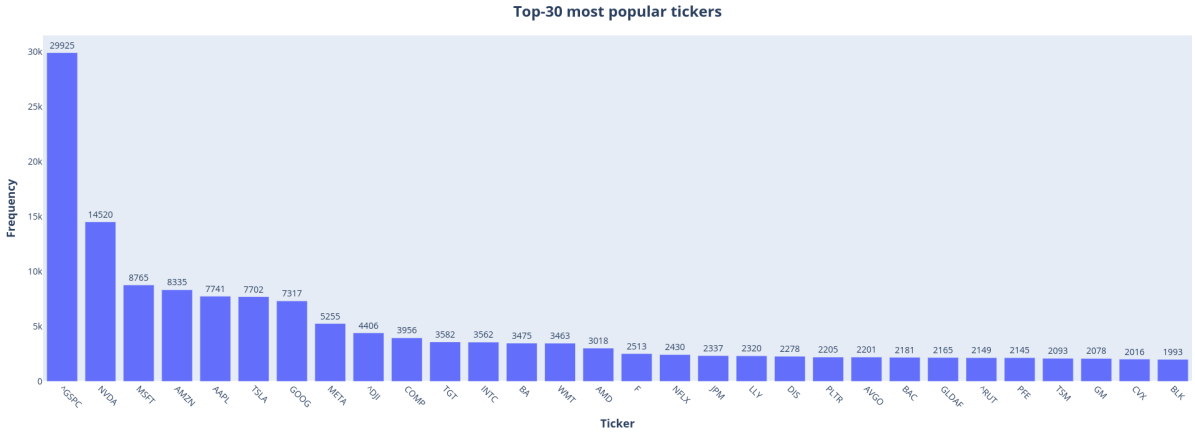

Figure 4: Top-30 most popular tickers.

Analysis of tickers. Figure 4 presents the distribution of publications by the 30 most frequently mentioned tickers. The leader is the S&P 500 index, although the sample also includes the Dow Jones and Russell 2000. Notably, the top 10 are predominantly IT companies, with Nvidia leading by a significant margin.

At the same time, approximately 574 000 (44.2%) publications do not contain any tickers in the article header. Moreover, even when tickers are present, they may not reflect all the companies or indices mentioned in the article. This indicates that although this dataset column is fairly representative, it does not provide complete coverage of all potential tickers, and some news items are formally omitted from consideration. Therefore, for tasks beyond the scope of this research, it would be advisable to create a dictionary of terms and names associated with each specific ticker and then algorithmically augment the ticker column using the corresponding texts.

Figure 5: Wordcloud of the whole collected corpus.

Text Quality. Figure 5 shows a word cloud generated from the entire corpus of collected texts. From this visualization, the following key conclusions can be drawn:

1. **Data Representativeness**. The word cloud demonstrates a wide range of financial terms, indicating that the dataset is sufficiently representative of financial topics. This suggests that the material covers various aspects of market activity and economic events.

2. **Specificity of Financial Terminology**. The frequency distribution of financial terms significantly differs from that observed in popular corpora used for training language models (e.g., English Wikipedia or BookCorpus). This discrepancy necessitates the application of DAPT to effectively train the model on domain-specific financial data.

3. **Level of Noise and Presence of Irrelevant Information**. The word cloud includes elements such as “Zacks”, “click”, “please”, “free”, and “source”. This indicates a significant

presence of noisy, promotional, or automatically generated fragments, which calls for the development of specialized methods for data cleaning without compromising the semantic integrity of the texts.

Additionally, it can be noted that the identified noise and dispersion of terms may negatively affect the quality of downstream tasks, such as classification or embedding extraction, if the data is not properly processed during the pre-processing stage.

Summary. The collected news dataset is characterized by several notable features. Firstly, there is pronounced seasonality in the publications—the minimums occur on weekends and public holidays, and so-called "tail" articles have also been recorded. Secondly, the analysis of sources indicates that about 69% of the texts originate from semi-automated aggregators, which can complicate the data cleaning process, as such sources often yield texts with broken formatting, embedded artifacts, and irrelevant information. Finally, it has been determined that the dataset exhibits a high variability of financial terminology while also containing a significant level of noise, which altogether confirms the need for DAPT and the development of effective text cleaning methods.

On one hand, the identified features (time shifts, noise, dominance of semi-automated sources) may reduce the suitability of the dataset for short-term forecasting or tasks that require precise timestamping. On the other hand, for tasks oriented toward the semantic content of the text, these issues do not have a critical impact. Proper pre-processing, including text cleaning and the removal of irrelevant elements, will substantially improve the quality of the trained model and expand its ability to generalize across various types of publications.

2.2.4 Data Preprocessing

«...»

2.3 Model Development

2.3.1 Feature Extraction

After preprocessing the text corpus and removing background noise, embeddings were extracted to accelerate subsequent stages of topic modeling, including dimensionality reduction and clustering.

The base ModernBERT model is not optimal for this task, since its vector representations are excessively sparse. High sparsity of embeddings degrades clustering quality, particularly for density-based methods (e.g., DBSCAN). Although HDBSCAN is more robust to density variations, sparsity still adversely affects clustering outcomes.

To address this issue, it is common to fine-tune the model on a semantic textual similarity (STS) task. In this setting, the model receives a pair of texts and returns a similarity score [Muennighoff (et al.), 2023], typically computed via cosine distance (Formulation 1).

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

STS models consistently produce denser and more informative embeddings.

Accordingly, for the base ModernBERT we evaluated two fine-tuned variants: modernbert-embed from Nomic AI [Nussbaum (et al.), 2024] and gte-modernbert-base from Alibaba [Li (et al.), 2023; X. Zhang (et al.), 2024]. Evaluation employed the Massive Text Embedding Benchmark (MTEB), which spans eight task categories and 58 datasets [Muennighoff (et al.), 2023]. Results were as follows:

- Clustering (12 datasets): gte-modernbert-base outperformed modernbert-embed by 1.5 percentage points (44.98% vs. 44.47%).
- STS (10 datasets): Their performances were comparable (81.78% vs. 81.57%).
- Overall (eight tasks, 56 datasets): gte-modernbert-base led by 1.76 percentage points on average (64.38% vs. 62.62%).

Consequently, gte-modernbert-base from the sentence_transformers library was chosen for embedding extraction [Reimers, Gurevych, 2019]. Instead of using the [CLS] token, a more advanced Mean Pooling technique was applied, averaging token embeddings across the sequence (Formulation 2).

$$\mathbf{h}_{\mathbf{mean}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t. \tag{2}$$

where T is the tokenized sequence length and $\mathbf{h}_t$ denotes the embedding of the t-th token.

To accelerate the training of dimensionality reduction and clustering models, the embeddings were first reduced to the unit L_2-norm (Formulation 3):

$$\widehat{\mathbf{u}} = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \qquad \|\mathbf{u}\|_2 = \sqrt{\sum_{i=1}^{n} u_i^2}. \tag{3}$$

Such preprocessing allows us to use the GPU-optimized Euclidean distance metric (Equation 4) when training the dimensionality reduction model, without having to repeat the L_2-norm computation that occurs when computing the cosine distance metric at each iteration of the hyperparameter optimization. so that Euclidean distance could be used for both dimensionality reduction and clustering.

$$D_2(\mathbf{u},\mathbf{v}) = \|\mathbf{u}-\mathbf{v}\|_2 = \sqrt{\sum_{i=1}^{n}\left(u_i - v_i\right)^2}. \tag{4}$$

Thus, after computing the L_2-norm and the Euclidean distance, we can actually, in a sense, be considered to be working with the cosine distance, since the reduced measure becomes mono-

tonically related to the cosine and reflects the same order of proximity of the points (Equation 5), but all computations are accelerated by GPU optimization.

$$D_2(\widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \left\|\widehat{\mathbf{u}} - \widehat{\mathbf{v}}\right\|_2 = \sqrt{2\left(1 - \widehat{\mathbf{u}} \cdot \widehat{\mathbf{v}}\right)}. \tag{5}$$

To build and train the dimensionality reduction and clustering algorithms, a training subsample of 200,000 embeddings and associated metadata was generated, which is approximately 16% of the entire corpus. This size of the training subsample was chosen based on the available computational resources. Thus, 200,000 embeddings were used to select the optimal hyperparameters, while the remaining 1,050,000 embeddings were reserved for the validation and inference phases. At the same time, before inference, the pipeline of dimensionality reduction and clustering models were trained on the entire corpus with a linear increase in hyperparameter values for the HDBSCAN algorithm, the choice of which is conditioned in Section 2.3.2.

Moreover, mixed precision (float16) and the FlashAttention mechanism [Dao (et al.), 2022] were employed during embedding extraction, substantially reducing computational resource requirements and runtime.

2.3.2 Dimensionality Reduction and Clustering

Thus, after assembling the embedding sample, we proceeded to experiments with dimensionality reduction and clustering models, performing their joint optimization. This approach reflects the multi-criteria nature of the task: it is necessary not only to preserve the structural (global) and local relationships from the original 768-dimensional space, but also to ensure that embeddings remain separable ("clusterable") in a low-dimensional projection suitable for two-dimensional visualization.

As key requirements we identified:

1. Preservation of cluster structure. Embeddings after dimensionality reduction must remain separable, preserving groupings by semantic and topical similarity.

2. Suitability for two-dimensional visualization. The resulting space must support clear and interpretable planar display.

To search simultaneously for optimal hyperparameters of both dimensionality reduction algorithms and clustering methods, we employed a unified meta-optimization process.

We used the DBCV index as our optimization metric, since it does not assume any predefined cluster shape (unlike as example silhouette coefficient favoring spherical or ellipsoidal structures) and effectively evaluates density-based clustering methods.

As our base clustering algorithm we selected HDBSCAN [Campello, Moulavi, Sander, 2013], which meets two crucial requirements:

- No shape assumptions. Unlike K-Means, HDBSCAN does not assume clusters are Gaussian spheres, which is critical for representing topics.
- Hierarchical, density-based nature. It can identify both large thematic groups and small, highly concentrated niches.

Moreover, GPU acceleration of HDBSCAN yielded high processing speed on both large samples and high-dimensional data.

Within the HDBSCAN framework, we tuned two key hyperparameters: the minimum number of neighbors — the count of points in a neighborhood required to consider a point a cluster "core" — and the minimum cluster size — the threshold number of observations for forming a cluster, which allows capturing rare, narrowly topical groups.

Pilot experiments revealed the coexistence of very dense regions ("hot topics") and rare but semantically significant clusters. A small minimum cluster size captures these rare topics but also increases the number of micro-clusters, some of which lack clear semantic distinction.

To mitigate this, we considered an ε-based cluster-merging technique [Malzer, Baum, 2020], consolidating adjacent micro-clusters in high-density regions. However, this approach complicates inference: when new observations arrive, ε-merging cannot be incrementally updated, necessitating full retraining.

Prioritizing practicality, we therefore abandoned ε-merging in favor of smaller minimum cluster sizes, accepting some fragmentation while preserving the ability to interpret and agglomerate clusters at higher hierarchical levels.

Another hyperparameter — the cluster selection method — determines whether clusters form based on excess of mass or tree leaves. We found that the latter yields finer-grained, more homogeneous groups, and used it for our final configuration.

Thus, in the final optimization stage, only two HDBSCAN parameters remained tunable: minimum neighbors and minimum cluster size.

It is noteworthy that, as will be described in Section 3.3, the future architecture assumes a fixed number of clusters corresponding to a static number of experts; hence, we employ the cuML implementation of HDBSCAN rather than its adaptive variant [Vijayan, Aziz, 2022].

Turning to dimensionality reduction algorithms, our preliminary selection included t-SNE, PCA, UMAP, TriMap, and PaCMAP, with key evaluation criteria of fidelity, the ability to balance local and global relationships, and training time complexity. Based on other researchers' experience, UMAP was chosen as the baseline algorithm [Grootendorst, 2022].

t-SNE was excluded due to its insufficient scalability on large datasets. PCA, although fast as a global approximation, did not preserve local structure in the final low-dimensional embedding, and an experiment combining PCA with UMAP fell short of standalone UMAP by approximately 37% in DBCV score. While TriMap and PaCMAP achieved similar performance in intermediate dimensions — and PaCMAP produced a more uniform distribution for two-dimensional visualization — the GPU-accelerated implementation and demonstrated robustness of UMAP in cuML led

us to select it as the definitive method for both intermediate dimensionality reduction and final 2D projection.

2.3.3 Hyperparameters Optimization

«To be done later»

2.4 System Development

«To be done later»

CHAPTER 3. RESULTS

3.1 Aspect-Based Representation

«To be done later»

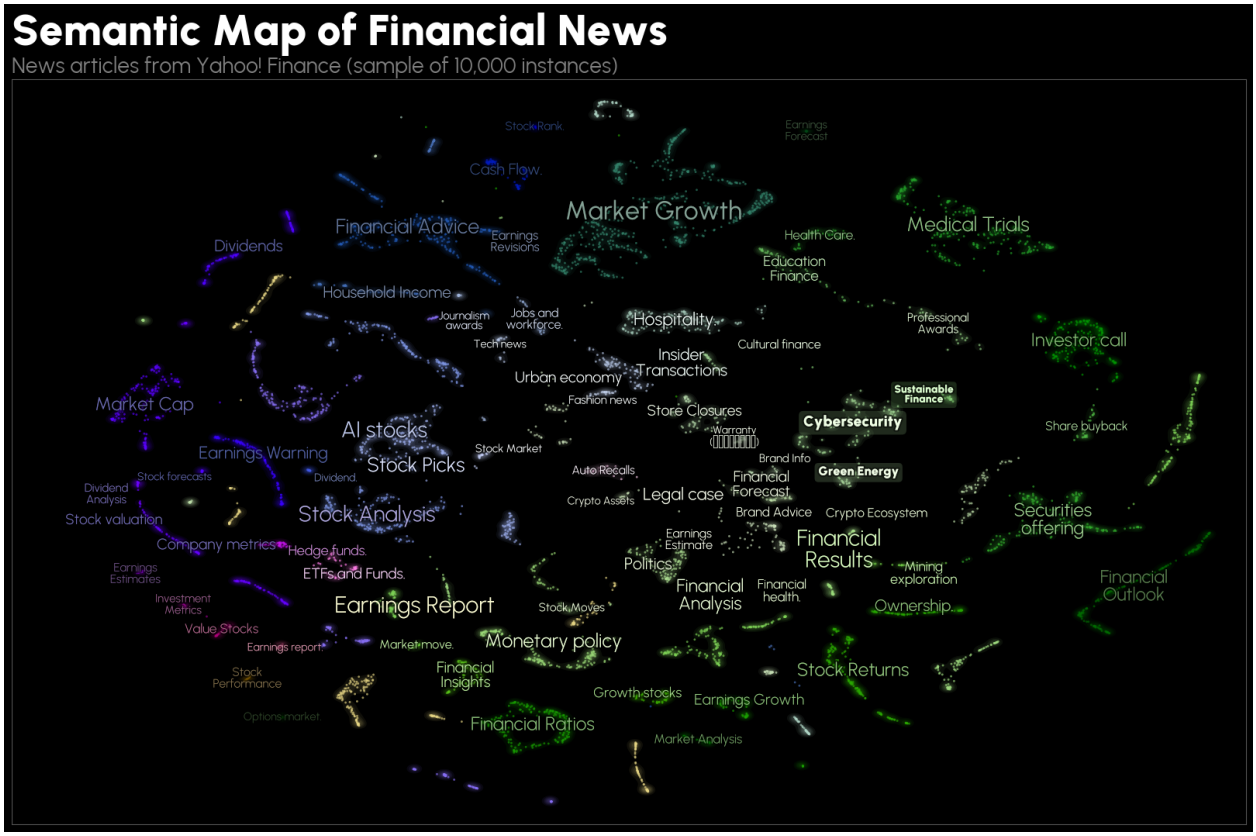

Figure 6: Semantic map (early demo version) of a sample of 10,000 financial publications from September 17, 2023 to March 1, 2025, clustered by financial topics.

3.2 Practical Importance

«To be done later»

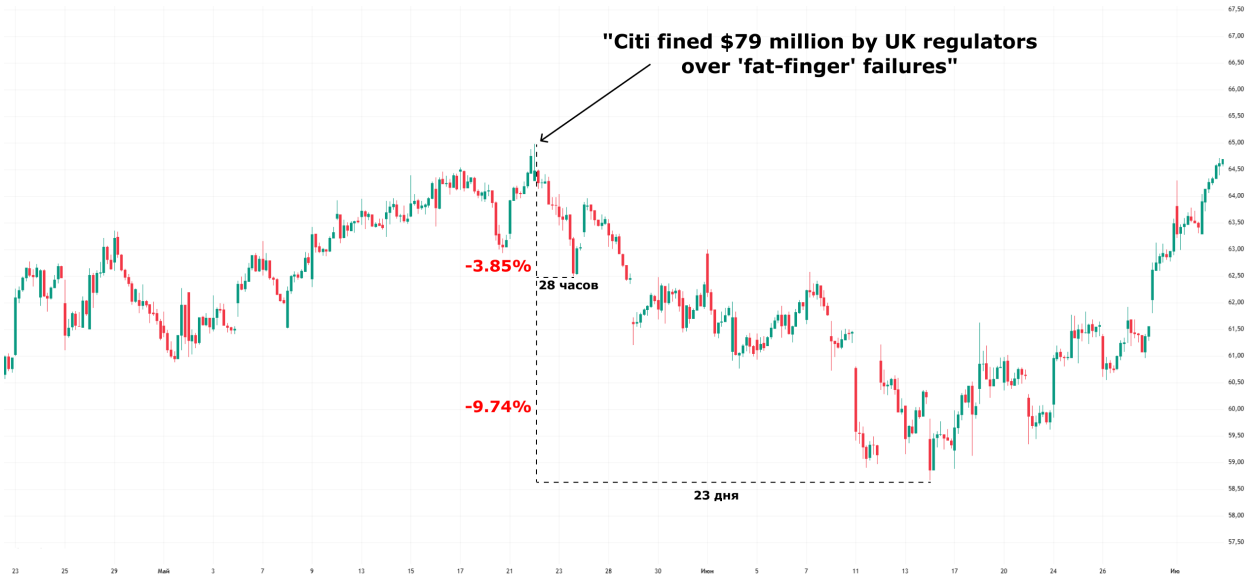

Figure 7: An example of the fall in the share price of one of the largest US banks, Citi Group (C.NYSE), due to allegations of insufficient control over trading operations, which caused a fall in European shares

«To be done later»

Figure 8: An example of Australian company The Star Entertainment Group LTD (SGR.ASX) shares collapsing and being temporarily halted from trading due to money laundering litigation.

3.3 Invented Architecture

3.3.1 Overview

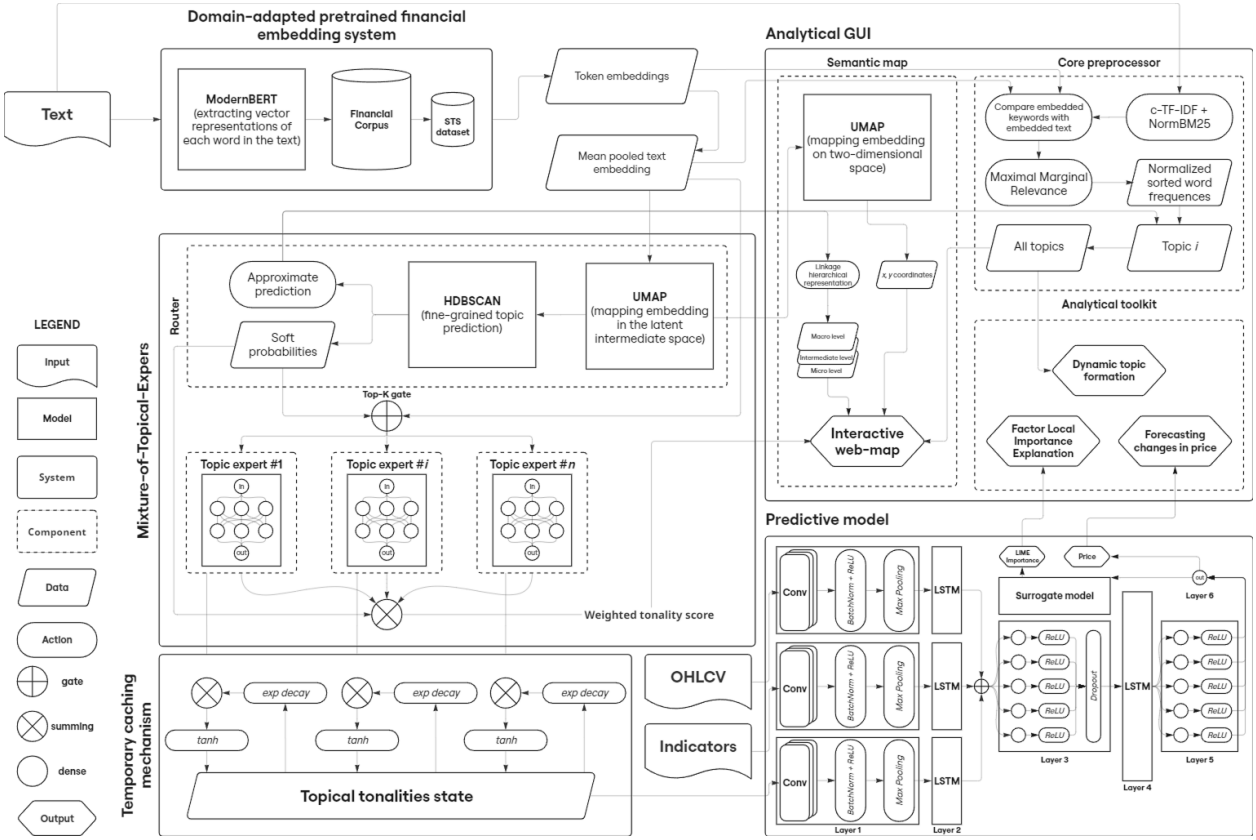

Figure 9: Architecture.

3.3.2 Embedding System

«To be done later»

3.3.3 Aspect-Based Sentimental Block

«To be done later»

3.3.4 Feature Caching Machine (FCM)

«To be done later»

3.4 Semantic De-duplication Solution

3.4.1 Mathematical Formulaion

Within the framework of this study, a novel deduplication approach was developed based on the analysis of the semantic content of objects. Although in the present work the entity is a

text, the method can be readily generalized to any objects that admit a vector representation in a semantic space.

Each article is represented as a sequence of embeddings:

$$x_i \subset \mathbb{R}^{t \times d} \tag{6}$$

where t is the number of tokens and d is the dimensionality of the semantic vector space. For subsequent analysis, instead of the raw set of embeddings, their convex hull is used, denoted as $\mathrm{CH}(x_i)$ or, for brevity, CH_i. The uniqueness of a text is quantified by the volume of this convex hull, $\mathrm{vol}(\mathrm{CH}_i)$.

Accounting for the Intersections of Convex Hulls. Direct subtraction of the intersections between CH_i and the hulls of other texts may lead to multiple counting. To eliminate this issue, the inclusion–exclusion principle is applied.

Let the set of all articles except i be denoted by

$$\mathbb{I} = \{1, \ldots, N\} \setminus \{i\}. \tag{7}$$

The intersection of CH_i with the hulls of articles indexed by subsets $\mathbb{J} \subseteq \mathbb{I}$ is expressed as:

$$\mathrm{vol}\Big(\mathrm{CH}_i \cap \bigcap_{j \in \mathbb{J}} \mathrm{CH}_j\Big). \tag{8}$$

Then, the volume of the intersection of CH_i with the union of the hulls of the remaining articles is computed as:

$$\mathrm{vol}\left(\mathrm{CH}_i \cap \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\right) = \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J} \subseteq \mathbb{I} \\ |\mathbb{J}|=k}} \mathrm{vol}\left(\mathrm{CH}_i \cap \bigcap_{j \in \mathbb{J}} \mathrm{CH}_j\right). \tag{9}$$

The uniqueness of article i is defined as the fraction of its convex hull's volume that is not occupied by intersections with the hulls of other articles:

$$\mu_i = \frac{\mathrm{vol}\Big(\mathrm{CH}_i \setminus \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\Big)}{\mathrm{vol}\Big(\mathrm{CH}_i\Big)}. \tag{10}$$

By decomposing CH_i into the intersection region and its complement, we obtain:

$$\mu_i = 1 - \frac{\mathrm{vol}\left(\mathrm{CH}_i \cap \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\right)}{\mathrm{vol}\left(\mathrm{CH}_i\right)}. \tag{11}$$

Substituting the inclusion–exclusion formulation 9, the final expression becomes:

$$\mu_i = 1 - \frac{1}{\mathrm{vol}\big(\mathrm{CH}_i\big)} \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J}\subseteq\mathbb{I} \\ |\mathbb{J}|=k}} \mathrm{vol}\Big(\mathrm{CH}_i \cap \bigcap_{j\in\mathbb{J}} \mathrm{CH}_j\Big). \tag{12}$$

The value $\mu_i \in [0,1]$ characterizes the text's uniqueness: $\mu_i = 1$ indicates no intersections with other texts (complete uniqueness), while $\mu_i = 0$ implies that the semantic volume of the text is entirely occupied by intersections with the hulls of other texts.

3.4.2 Pros and Cons

The proposed method is founded on a theoretically sound representation of text: each article is treated as the convex hull of its token embeddings. This representation enables a precise definition of an object's semantic content, facilitates the application of the inclusion-exclusion principle (inherited from set theory) for accurate calculation of intersection volumes, and normalizes the result so that the final uniqueness measure lies within the interval $[0,1]$.

In addition to its theoretical rigor, the method offers several advantages:

- The use of embeddings for each token allows for capturing subtle distinctions in semantic content, while aggregation via the convex hull yields a generalized representation of the text. This approach enables the comparison of texts of varying lengths and topics within a unified vector space.
- Normalization of the metric to a $[0,1]$ interval simplifies interpretation.

Conversely, the method has several notable drawbacks:

- In high-dimensional spaces (e.g., 768 dimensions), the convex hull may become excessively "stretched," resulting in uninformative volume measurements, and its geometry may fail to accurately reflect the complex distribution of embeddings.
- Embeddings typically possess a complex, often non-linear structure. As the convex hull is the minimal convex set containing the data, it may enclose extreme points, leading to an overestimation of the occupied space and, consequently, to skewed evaluations.
- Since embeddings can include random noise or artifacts, the convex hull is sensitive to outliers. Minor inaccuracies in embeddings may disproportionately enlarge the convex hull's volume, thus distorting the uniqueness assessment.
- Constructing the convex hull and computing volumes in high-dimensional spaces is computationally intensive. Moreover, applying the inclusion–exclusion principle to accurately compute intersections between the hulls of texts further complicates calculations, particularly with a large number of documents.

These challenges can critically affect the practical application of the method; however, some can be mitigated through engineering solutions.

The sensitivity to noise (item 3) can be partially alleviated by using the [CLS] token as the centroid of the convex hull. Introducing a coefficient δ to normalize the "concavity" of the hull in the direction of the [CLS] embedding helps to diminish the impact of noisy components.

The computational burden (item 4) can be addressed through various strategies:

- Regulating the number of inclusion–exclusion pairs (the hyperparameter N in the summation) allows for an approximate evaluation of uniqueness while reducing computational demands.
- Employing dimensionality reduction algorithms, such as UMAP, t-SNE, or PCA, can project the original space onto a lower-dimensional one, substantially decreasing computational costs, though potentially at the expense of some accuracy.
- Approximating the volume using Monte Carlo methods offers an alternative that lessens computational load.

The method of representing semantic uniqueness of text via the convex hulls of embeddings boasts several theoretical advantages (robust normalization, applicability of the inclusion–exclusion principle, and consistent interpretability of the result). Nevertheless, its practical deployment necessitates addressing challenges related to high dimensionality, non-linear distributions of embeddings, and substantial computational costs. Future research may focus on developing more robust and computationally efficient methods for assessing text uniqueness while accommodating these limitations.

Conclusion

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 (et al.), 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 Vantage[23] offers a free and open API that grants access to a vast array of valuable data, including intraday OHLCV. Although Reddit[24] is less popular than platform X (ex-Twitter)[25] in the financial community, it also provides an open API and can serve as an alternative channel for publishing announcements, opinions, and insider information.

[23]URL: `https://www.alphavantage.co/`
[24]URL: `https://www.reddit.com/`
[25]URL: `https://x.com/`

In addition, aggregators such as FinURLs[26] and MarketWatch[27] 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 SEC[28] 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.

References

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. — P. 84–90.

Araci, *D.* FinBERT: Financial Sentiment Analysis with Pre-trained Language Models / D. Araci. — 2019. — Aug. — 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. — P. 276–279. — (WWW '21). — ISBN 9781450383134. — DOI: 10.1145/3442442.3451378. — URL: https://doi.org/10.1145/3442442.3451378.

[26]URL: https://finurls.com/

[27]URL: https://www.marketwatch.com/

[28]URL: https://www.sec.gov/

Bentivogli, *L.* The Fifth PASCAL Recognizing Textual Entailment Challenge. / L. Bentivogli [et al.] // TAC. — 2009. — Vol. 7, no. 8. — P. 1.

Bogdan, *R.* Qualitative research for education. Vol. 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. — P. 160–172.

Cer, *D.* SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation / D. Cer [et al.] // 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 [et al.]. — 2017.

Dao, *T.* FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness / T. Dao [et al.]. — 2022. — arXiv: 2205.14135 [cs.LG]. — URL: https://arxiv.org/abs/2205.14135.

Devlin, *J.* Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin [et al.] // 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. — P. 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 [et al.] // ACM Computing Surveys. — 2024. — Oct. — Vol. 56, issue 9. — ISSN 15577341. — DOI: 10.1145/3649451.

Dudy, *S.* A Multi-Context Character Prediction Model for a Brain-Computer Interface / S. Dudy [et al.] // Proceedings of the Second Workshop on Subword/Character LEvel Models / ed. by M. Faruqui [et al.]. — New Orleans : Association for Computational Linguistics, 06/2018. — P. 72–77. — DOI: 10.18653/v1/W18-1210. — URL: https://aclanthology.org/W18-1210/.

Fama, *E. F.* Efficient Capital Markets: A Review of Theory and Empirical Work / E. F. Fama // The Journal of Finance. — 1970. — Vol. 25, no. 2. — P. 383–417. — ISSN 00221082, 15406261. — URL: http://www.jstor.org/stable/2325486 (visited on 04/30/2025).

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 [et al.] // arXiv preprint arXiv:2004.10964. — 2020.

Halder, *S.* FinBERT-LSTM: Deep Learning based stock price prediction using News Sentiment Analysis / S. Halder. — 2022. — Nov. — URL: http://arxiv.org/abs/2211.07392.

Hochreiter, *S.* Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. — 1997. — Nov. — Vol. 9, issue 8. — P. 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. — May. — Vol. 40, issue 2. — P. 806–841. — ISSN 19113846. — DOI: 10.1111/1911-3846.12832.

Jiang, *T.* Financial sentiment analysis using FinBERT with application in predicting stock movement / T. Jiang, A. Zeng. — 2023. — June. — URL: http://arxiv.org/abs/2306.02136.

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. — Sept. — Vol. 12, issue 9. — ISSN 20751680. — DOI: 10.3390/axioms12090835.

Levesque, *H. J.* The Winograd schema challenge. / H. J. Levesque, E. Davis, L. Morgenstern // KR. — 2012. — Vol. 2012. — 13th.

Li, *Z.* Towards general text embeddings with multi-stage contrastive learning / Z. Li [et al.] // arXiv preprint arXiv:2308.03281. — 2023.

Liu, *Z.* FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining : tech. rep. / Z. Liu [et al.]. — 2020. — URL: http://commoncrawl.org/.

Macedo, *M.* WWW'18 Open Challenge: Financial Opinion Mining and Question Answering / M. Macedo [et al.] // Companion Proceedings of The Web Conference 2018. — Lyon, France : International World Wide Web Conferences Steering Committee, 2018. — P. 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 [et al.] // Journal of the Association for Information Science and Technology. — 2014. — Vol. 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. — P. 223–228. — DOI: 10.1109/mfi49285.2020.9235263. — URL: http://dx.doi.org/10.1109/MFI49285.2020.9235263.

Muennighoff, *N.* MTEB: Massive Text Embedding Benchmark / N. Muennighoff [et al.]. — 2023. — arXiv: 2210.07316 [cs.CL]. — URL: https://arxiv.org/abs/2210.07316.

Nussbaum, *Z.* Nomic Embed: Training a Reproducible Long Context Text Embedder / Z. Nussbaum [et al.]. — 2024. — arXiv: 2402.01613 [cs.CL].

Pathak, *A. R.* Application of Deep Learning Approaches for Sentiment Analysis / A. R. Pathak [et al.] // Deep Learning-Based Approaches for Sentiment Analysis / ed. by B. Agarwal [et al.]. — Singapore : Springer Singapore, 2020. — P. 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 [et al.] // arXiv preprint arXiv:1606.05250. — 2016.

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.

Shah, *R. S.* When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain / R. S. Shah [et al.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2022.

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. — P. 589–601.

Socher, *R.* Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank / R. Socher [et al.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. — Seattle, Washington, USA : Association for Computational Linguistics, 10/2013. — P. 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. — Nov. — Vol. 49, issue 11. — P. 3815–3820. — ISSN 15737497. — DOI: 10.1007/s10489-019-01458-9.

Vaswani, *A.* Attention is all you need / A. Vaswani [et al.] // Advances in neural information processing systems. — 2017. — Vol. 30.

Vijayan, *D.* Adaptive hierarchical density-based spatial clustering algorithm for streaming applications / D. Vijayan, I. Aziz // Telecom. Vol. 4. — MDPI. 2022. — P. 1–14.

Vuković, *D. B.* Predictive Patterns and Market Efficiency: A Deep Learning Approach to Financial Time Series Forecasting / D. B. Vuković [et al.] // Mathematics. — 2024. — Oct. — Vol. 12, issue 19. — ISSN 22277390. — DOI: 10.3390/math12193066.

Wang, *A.* GLUE: A multi-task benchmark and analysis platform for natural language understanding / A. Wang [et al.] // arXiv preprint arXiv:1804.07461. — 2018.

Warner, *B.* Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference / B. Warner [et al.]. — 2024. — Dec. — URL: http://arxiv.org/abs/2412.13663.

Wen, *Q.* Transformers in time series: A survey / Q. Wen [et al.] // arXiv preprint arXiv:2202.07125. — 2022.

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 [et al.] // arXiv preprint arXiv:2303.17564. — 2023.

Xing, *F.* Financial sentiment analysis: An investigation into common mistakes and silver bullets / F. Xing [et al.] // Proceedings of the 28th international conference on computational linguistics. — 2020. — P. 978–987.

Yang, *Y.* FinBERT: A Pretrained Language Model for Financial Communications / Y. Yang, M. C. S. UY, A. Huang. — 2020. — June. — URL: http://arxiv.org/abs/2006.08097.

Zhang, *X.* mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval / X. Zhang [et al.] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. — 2024. — P. 1393–1412.