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

«…»

CHAPTER 1. THEORETICAL OVERVIEW

1.1 Problem statement

«…»

1.2 Methodology

«…»

1.2.1 Literature review

«…»

1.2.2 Tool review

«…»

CHAPTER 2. PRACTICAL SOLUTION

2.1 Limitaions

2.1.1 De-duplication

«...»

2.1.2 Computing Hardware

«...»

2.1.3 Data

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.

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

In addition, aggregators such as FinURLs and MarketWatch 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 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.

2.1.4 Context (Attention Mechanism)

«...»

2.2 Data Governance

2.2.1 Selection of Information Resources

«...»

2.2.2 Data Collecting

«...»

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.

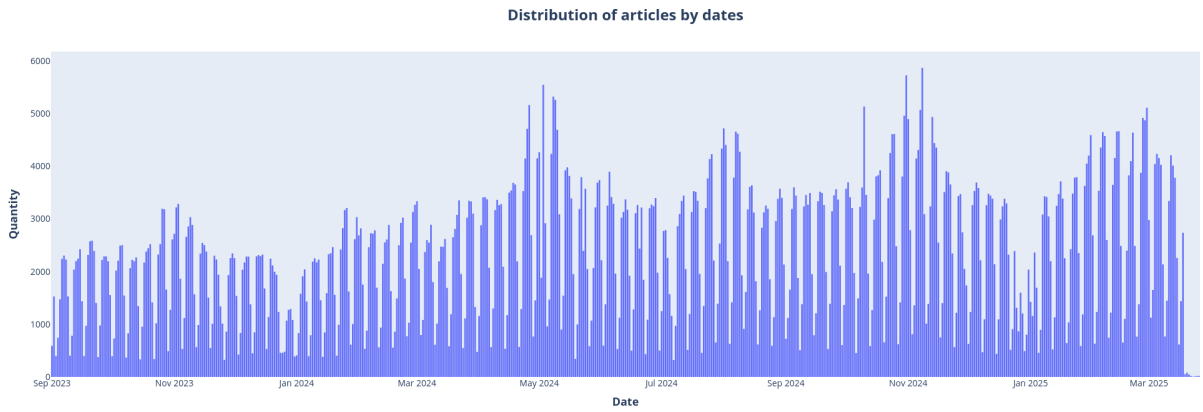

Figure 1: Distribution of publications by dates.

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 oc-

cur 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.

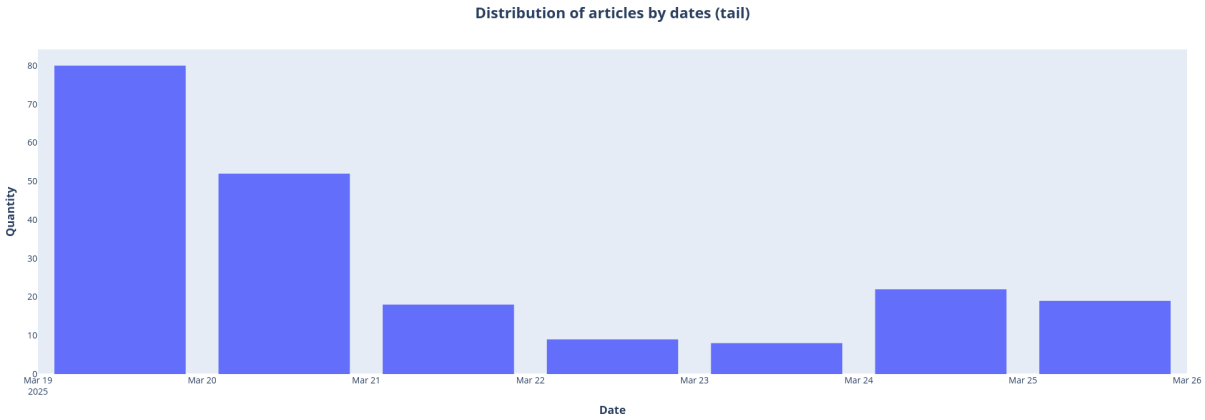

Figure 2: Distribution of publications by dates (tail).

At the same time, some articles, by formal characteristics, fall outside the collection period (September 1, 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.

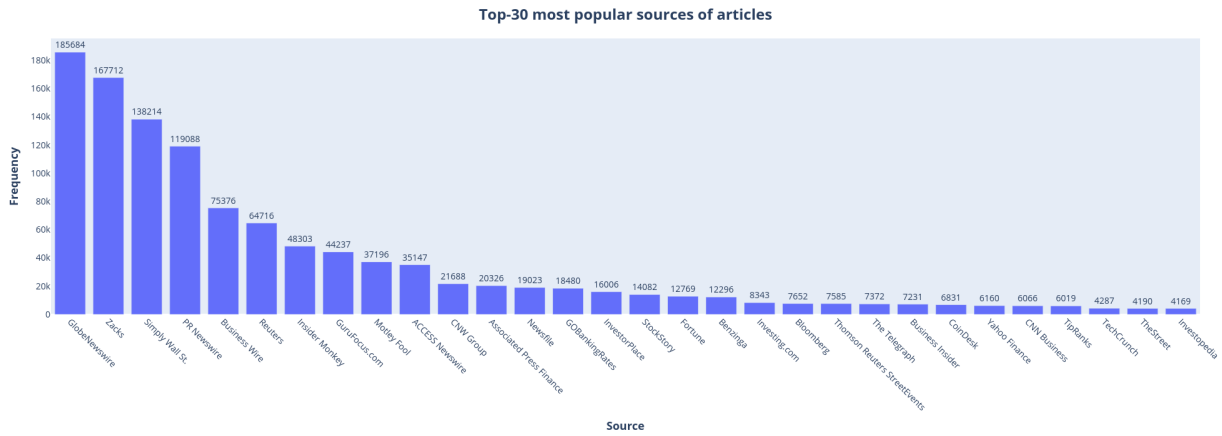

Figure 3: Top-30 most popular sources of publications.

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.

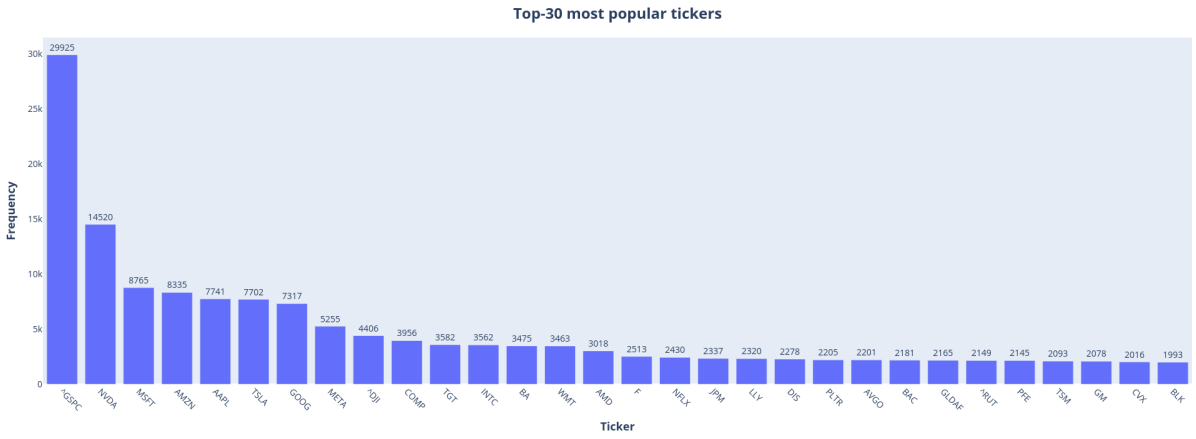

Figure 4: Top-30 most popular tickers.

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

«...»

2.2.5 Retrieval of Embeddings

«...»

CHAPTER 3. RESULTS

3.1 Benchmarks

3.2 Aspect-Based Representation

3.3 Invented Architecture

3.3.1 Overview

«...»

3.3.2 Embedding System

«...»

3.3.3 Aspect-Based Sentimental Block

«...»

3.3.4 Feature Caching Machine (FCM)

«...»

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:

$$x_i \subset \mathbb{R}^{t \times d} \tag{1}$$

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

$$\mathbb{I} = \{1, \ldots, N\} \setminus \{i\}. \tag{2}$$

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

$$\mathrm{vol}\Big(\mathrm{CH}_i \cap \bigcap_{j \in \mathbb{J}} \mathrm{CH}_j\Big). \tag{3}$$

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

$$\mathrm{vol}\left(\mathrm{CH}_i \cap \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\right) = \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J} \subseteq \mathbb{I} \\ |\mathbb{J}|=k}} \mathrm{vol}\left(\mathrm{CH}_i \cap \bigcap_{j \in \mathbb{J}} \mathrm{CH}_j\right). \tag{4}$$

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:

$$\mu_i = \frac{\mathrm{vol}\Big(\mathrm{CH}_i \setminus \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\Big)}{\mathrm{vol}\Big(\mathrm{CH}_i\Big)}. \tag{5}$$

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

$$\mu_i = 1 - \frac{\mathrm{vol}\left(\mathrm{CH}_i \cap \bigcup_{j \in \mathbb{I}} \mathrm{CH}_j\right)}{\mathrm{vol}\left(\mathrm{CH}_i\right)}. \tag{6}$$

Substituting the inclusion–exclusion formulation 4, the final expression becomes:

$$\mu_i = 1 - \frac{1}{\mathrm{vol}\big(\mathrm{CH}_i\big)} \sum_{k=1}^{N-1} (-1)^{k-1} \sum_{\substack{\mathbb{J} \subseteq \mathbb{I} \\ |\mathbb{J}|=k}} \mathrm{vol}\Big(\mathrm{CH}_i \cap \bigcap_{j \in \mathbb{J}} \mathrm{CH}_j\Big). \tag{7}$$

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

«...»

References

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