Semantic Signal Mining for Oil Price Forecasting Using News Embeddings

Sara Borello[1] **and Keita Jacopo Viganò**[1]

[1]MSc Students in Data Science, University of Milano-Bicocca (Unimib)

This project investigates the integration of textual data extracted from market news with natural language processing techniques to enhance crude oil price forecasting. Through automated scraping of relevant articles and reports, up-to-date textual data were collected and transformed into numerical representations using embeddings. These embeddings were then used as inputs for machine learning models to capture hidden informational signals within the news and improve the accuracy of oil price predictions. The study demonstrates how semantic analysis of financial news can effectively complement historical market data, providing an innovative and data-driven approach to forecasting in the energy sector.

Contents

1 Introduction

The increasing availability of online financial news has opened new opportunities for extracting latent information relevant to market dynamics, particularly in volatile and geopolitically sensitive sectors such as energy. Among these, crude oil stands out as a globally traded commodity whose price is heavily influenced by qualitative factors, such as geopolitical tensions, production decisions, and technological shifts, disseminated through unstructured news content. Capturing and quantifying these signals represents a significant challenge for traditional financial models.

This work addresses the problem of forecasting Brent crude oil prices by integrating advanced natural language processing techniques into a financial prediction framework. Specifically, multiple neural embedding models ranging from traditional static representations like Word2Vec to contextualized Transformer-based encoders such as DistilBERT and MiniLM were employed to convert oil-related news articles into dense semantic representations. These embeddings were then used to extract a rich set of features including daily semantic means, dispersion metrics, and measures of temporal drift, offering a quantitative lens into the evolving narrative structure of the energy domain.

In parallel, a comparative topic modeling analysis was conducted through clustering and BERTopic, with the aim of exploring the thematic composition of the news corpus and tracking its evolution from 2011 to 2025. Finally, the derived embedding features were integrated into multiple predictive architectures including Random Forest, XGBoost, and a Transformer-Autoencoder hybrid to assess their contribution to next-day price forecasting.

Data Acquisition and Dataset Description

The data acquisition phase for this project involved automated web scraping of news articles related to the oil market from the website *OilPrice.com*, specifically from the section dedicated to the latest world energy news (URL: `https://oilprice.com/Latest-Energy-News/World-News/`)(1). Using a Python-based pipeline built with the `requests` and `BeautifulSoup` libraries, the scraper systematically navigated through multiple pages (up to 1171 pages) of the website. For each page, the HTML content was parsed to extract structured information on individual articles contained within specific HTML elements and CSS classes.

The scraper extracted five key fields for each news article: the **title** (headline), the **URL** linking to the original article, the **publication date** as reported by the website, the **author** name if available, and a short **excerpt** summarizing the article content. These fields were carefully targeted by locating the corresponding tags and classes within the page structure, ensuring consistent and accurate data extraction.

To avoid overwhelming the server and to comply with ethical scraping practices, a delay of 1.5 seconds was introduced between requests. The extracted data from all pages was aggregated into a single `pandas.DataFrame` and saved locally for further processing.

The final dataset used in this project consists of 23,420 news articles from June 18, 2011, to June 25, 2025. After the initial data acquisition, the dataset was refined by removing less relevant fields such as the `author` and `URL`, focusing on three primary columns: `title`, `date`, and `excerpt`.

The `title` contains the headline of each article, providing a concise summary of the news content. The `date` field records the publication date of the article, spanning several years and allowing for temporal analysis. The `excerpt` offers a brief textual summary or preview of the article's main points.

2 Embeddings

In this phase, the raw textual data from oil news articles is transformed into numerical vector representations. Embeddings capture the semantic and contextual information of the texts in a high-dimensional continuous space, enabling downstream quantitative analysis. To explore different semantic representations, multiple embedding models are employed. Once the news articles are converted into embeddings, cluster analysis is applied to investigate the latent structures and thematic groupings within the textual data. From the embeddings, statistical features are extracted to summarize the information in a compact and informative manner.

A. SentenceTransformer's all-MiniLM-L6-v2 Embeddings. The model used for sentence representation in this work is all-MiniLM-L6-v2, developed as part of the Sentence-Transformers framework and released by Reimers and colleagues (2). This model is designed to produce dense vector representations for sentences and short paragraphs that reflect their semantic content. The core architecture is based on a distilled Transformer variant, MiniLM (3), featuring 6 Transformer layers, a hidden size of 384, and 12 self-attention heads. Pretraining was conducted using a standard masked language modeling objective over general-purpose corpora (e.g., Wikipedia, BookCorpus), with the aim of learning contextualized token-level embeddings.

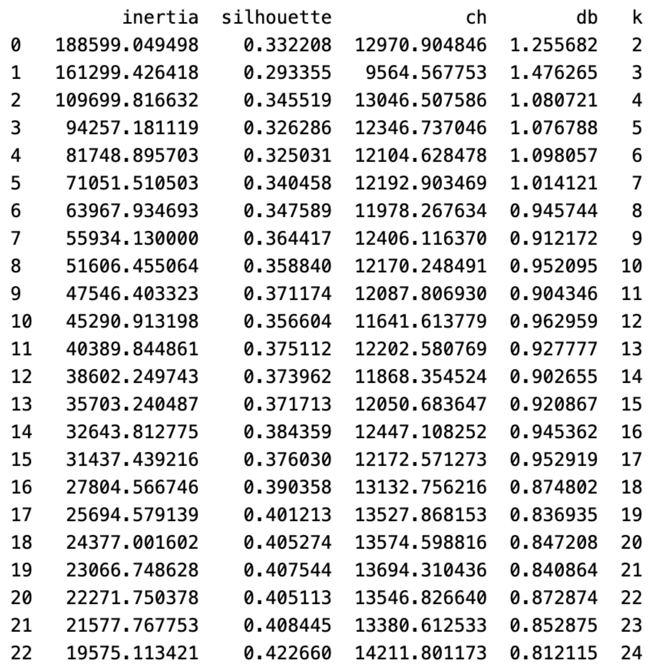
The model was later fine-tuned on a large-scale dataset containing over one billion sentence pairs, using a contrastive learning objective. The fine-tuning process employed a Siamese network architecture (4) in which both elements of a sentence pair are passed through a shared encoder. Embeddings are extracted using mean pooling over token representations, and training is guided by maximizing cosine similarity between true sentence pairs (anchor and positive) while minimizing it for in-batch negatives. Cross-entropy loss is applied over a similarity matrix computed across all pairs in the batch. Optimization was performed using the AdamW optimizer with a learning rate of 2e-5, a batch size of 1024 (distributed across TPU cores), and a sequence length limit of 128 tokens.

The final model, all-MiniLM-L6-v2, takes as input a sentence or a short text sequence of up to 256 tokens (with longer inputs truncated) and outputs a 384-dimensional dense embedding vector. These embeddings are suitable for downstream tasks such as semantic search, clustering, or sentence similarity.

To visualize and compress the high-dimensional embeddings, UMAP (5) was applied. UMAP is a non-linear dimensionality reduction technique that preserves the local and global structure of high-dimensional data in a lower-dimensional space. It operates under manifold learning assumptions and constructs a weighted k-nearest neighbor graph in the original space, which is then optimized in a lower-dimensional layout using stochastic gradient descent. UMAP was applied to project the 384-dimensional sentence embeddings into both two-dimensional and three-dimensional spaces.

A.1. Cluster Analysis. a cluster analysis was performed using sentence embeddings obtained from the pre-trained transformer model `all-MiniLM-L6-v2`. To determine the optimal number of clusters (k), three commonly used internal evaluation metrics were computed: the **Silhouette Score**, the **Calinski–Harabasz Index**, and the **Davies–Bouldin Index**.

The two candidate values of k are k = 20 and k= 24, reported in Figure 1 there are corresponding metrics. As shown, $k = 24$ achieves the best overall performance across all metrics, indicating the formation of compact and well-separated clusters. However, $k = 20$ offers a favorable trade-off, achieving a silhouette score of 0.4053 (only 4% lower), while still maintaining a strong Calinski–Harabasz Index and a reasonably low Davies–Bouldin Index. Therefore, $k = 20$ was selected as the final configuration in order to reduce model complexity while preserving interpretability and cluster quality.



```
          inertia  silhouette            ch        db   k
0   188599.049498    0.332208  12970.904846  1.255682   2
1   161299.426418    0.293355   9564.567753  1.476265   3
2   109699.816632    0.345519  13046.507586  1.080721   4
3    94257.181119    0.326286  12346.737046  1.076788   5
4    81748.895703    0.325031  12104.628478  1.098057   6
5    71051.510503    0.340458  12192.903469  1.014121   7
6    63967.934693    0.347589  11978.267634  0.945744   8
7    55934.130000    0.364417  12406.116370  0.912172   9
8    51606.455064    0.358840  12170.248491  0.952095  10
9    47546.403323    0.371174  12087.806930  0.904346  11
10   45290.913198    0.356604  11641.613779  0.962959  12
11   40389.844861    0.375112  12202.580769  0.927777  13
12   38602.249743    0.373962  11868.354524  0.902655  14
13   35703.240487    0.371713  12050.683647  0.920867  15
14   32643.812775    0.384359  12447.108252  0.945362  16
15   31437.439216    0.376030  12172.571273  0.952919  17
16   27804.566746    0.390358  13132.756216  0.874802  18
17   25694.579139    0.401213  13527.868153  0.836935  19
18   24377.001602    0.405274  13574.598816  0.847208  20
19   23066.748628    0.407544  13694.310436  0.840864  21
20   22271.750378    0.405113  13546.826640  0.872874  22
21   21577.767753    0.408445  13380.612533  0.852875  23
22   19575.113421    0.422660  14211.801173  0.812115  24
```

Fig. 1. Value of Clustering metics for K

Following the clustering step, a two-stage topic modeling procedure was applied to interpret the semantic content of each cluster. For each cluster, the titles of all documents were concatenated into a single string. Two distinct approaches were used to extract meaningful keywords and generate topic labels:

- **TF-IDF-based method**: This approach measured the importance of each word within the cluster using Term Frequency–Inverse Document Frequency. After calculating TF-IDF scores, the top 20 terms were selected as representative keywords for each cluster. This method captures statistically frequent but informative terms.

- **BERTopic-based method**: BERTopic integrates transformer-based embeddings with class-based TF-IDF to extract coherent and semantically rich topics. This method identified the top 10 representative keywords per cluster. It is particularly effective at leveraging contextual information in textual data.

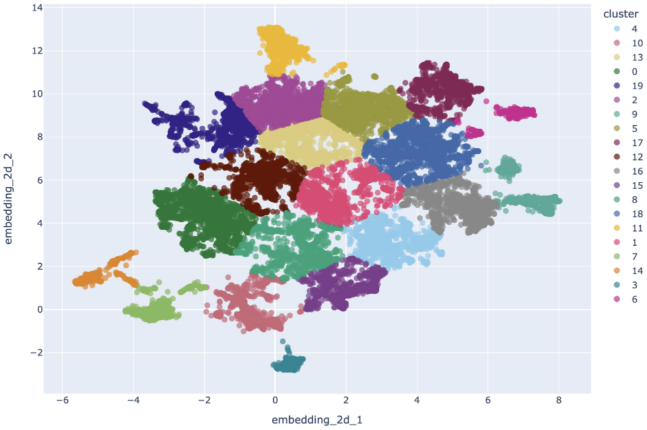


Fig. 2. Results of Cluster Analyisis with K= 20 with all-MiniLM-L6-v2 embeddings

The keyword sets produced by both methods were then processed through a LLM, which synthesized the keywords into concise and human-readable topic labels.

The following are examples of topic definitions obtained through this process visible in Figure 2 :

- **Cluster 0**: Russian oil and gas sector, including Gazprom, Nord Stream pipeline, EU sanctions, export dynamics, and pricing amid the Ukraine conflict.
- **Cluster 3**: Electric vehicle and battery market developments, with a focus on Tesla, lithium, EV sales in China, and global energy markets.
- **Cluster 8**: Global energy transition, including solar and wind power, emissions reduction, coal use, and climate change policy.
- **Cluster 13**: Global oil price dynamics, covering gasoline demand, OPEC production forecasts, and reports from agencies such as the IEA.

B. Distilbert-base-uncased. For the second method of embedding extraction, the `distilbert-base-uncased` model (6) was employed from Huggingface. DistilBERT is a distilled version of the original BERT-base model, designed to be smaller and faster while retaining a high level of performance. The model is pretrained using a self-supervised knowledge distillation technique, where the larger BERT model acts as the teacher and DistilBERT as the student. During training, DistilBERT consists of 6 Transformer layers compared to BERT's 12 layers, reducing the parameter count from 110 million to 66 million.

The training procedure involves masking tokens in the input sequence and predicting them, with 15% of tokens selected for masking. The loss function is a weighted sum of three components: the distillation loss measured by the Kullback-Leibler divergence between the teacher and student logits, the masked language modeling loss computed as the cross-entropy on masked tokens using teacher labels, and a cosine embedding loss to maximize the similarity between the hidden states of the teacher and student models.

At inference, DistilBERT processes tokenized input sequences through its 6 Transformer layers, which include multi-head self-attention, feed-forward networks, and normalization layers. The final embedding for each input sentence is obtained by pooling the sum of token embeddings and position embeddings, producing a fixed-length vector of 768 dimensions. The input to the model usage consisted of concatenated title and excerpt texts. These inputs were tokenized and processed by DistilBERT to obtain dense embeddings representing the semantic content of each document. Subsequently, these high-dimensional embeddings were projected into lower-dimensional spaces (2D and 3D) using UMAP.

B.1. Cluster Analysis. Cluster analysis was then applied to the reduced embeddings to identify groups of semantically similar documents. Multiple values of k were tested to determine the optimal number of clusters. While lower values of k (such as 3, 4, and 8) exhibited strong performance according to silhouette scores, the choice of k=18 was preferred due to its ability to maintain higher semantic granularity and produce more meaningful topic differentiation. This selection was supported by a superior Calinski-Harabasz index (26,416.59), a low Davies-Bouldin score (0.7619), and a competitive silhouette coefficient (0.4002), indicating a robust balance between cluster cohesion and separation.

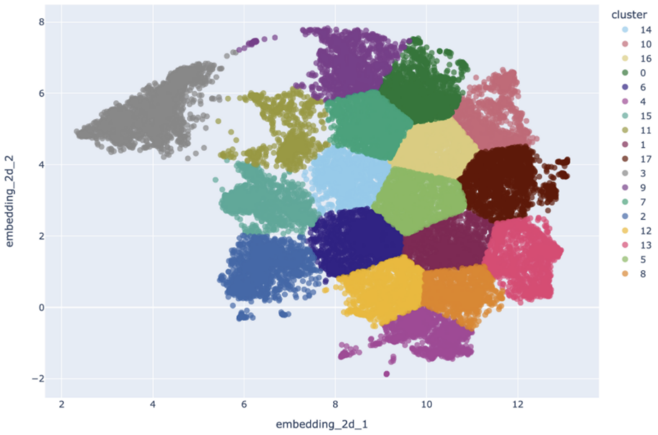


Fig. 3. Visualization of clusters in the 2D embedding space obtained via UMAP reduction. Each color corresponds to a different cluster.

The resulting clusters from the analysis are visualized in the 2D embedding space in Figure 3, where each point represents a document colored according to its assigned cluster. The segmentation clearly shows well-defined groups, reflecting semantic similarity within clusters and distinction between different topics.

Following the clustering, topic extraction was performed on each cluster using the same methodology described previously. Table 1 summarizes the main themes identified for each of the 18 clusters, providing descriptive labels based on the most relevant keywords and document content.

Among the identified topics, several clusters stand out for their thematic significance. For example, Cluster 0 focuses on geopolitical energy dynamics and sanctions, covering countries and organizations such as Russia, Iran, Ukraine, and OPEC, which are central to understanding global energy markets under political tension. Cluster 6 highlights the European energy transition, with discussions about renewable energy sources, carbon emissions, and shifts from coal to gas, reflecting ongoing structural changes in the energy sector. Cluster 13 addresses clean energy and electric vehicle revolution topics, including key players like Tesla and issues related to batteries and wind power, indicative of technological innovation trends. Additionally, Cluster 10 concerns major upstream projects and partnerships involving companies like Gazprom, Rosneft, and Exxon, emphasizing investment and development activities in the oil and gas industry.

Cluster	Topic
0	Geopolitical energy dynamics and sanctions (focus on Russia, Iran, Ukraine, OPEC, pipelines, nuclear energy)
1	Global oil and energy markets (price movements, climate considerations, major producers like Shell and Canada, LNG/shale)
2	Large-scale energy investments and deals (billions in oil/gas projects, funds, Aramco plans, India and Shell)
3	Oil and gas market fluctuations under sanctions (Iran, Venezuela, Russia) and export/production shifts
4	Inventory reports and crude price drivers (API builds/draws, gasoline stocks, weekly supply surprises)
5	China and India’s energy demand (coal, nuclear, LNG), global power mix and pricing
6	European energy transition (UK/EU renewables like wind/solar, emissions, carbon, coal-to-gas shifts)
7	OPEC production decisions (Saudi cuts, Russian output, Iran/Venezuela export policies, ministerial actions)
8	Q-series earnings and profit reports in oil majors (Shell, Exxon, record profits vs. estimates)
9	Attacks on oil infrastructure (Libya tankers, pipelines, Houthi/ISIS threats, Saudi/Nigerian fields)
10	Major upstream projects and partnerships (Gazprom, Rosneft, Exxon offshore fields, LNG pipelines)
11	Production/output metrics and OPEC court rulings (barrels per day, “000” figures, Gulf storms, Nigeria/Libya)
12	OPEC price forecasts and demand outlooks (IEA, Goldman Sachs, growth projections, cuts)
13	Clean energy and EV revolution (Tesla solar, batteries, wind power, Model 3/EV sales, Musk vs. Aramco)
14	China’s crude trade flows (exports/imports, records, India, Russia, Saudi pricing trends)
15	North American pipeline politics (Keystone XL, Trump/Biden energy policies, Canadian courts, Alberta drilling)
16	Global crude trade and import dependencies (Russia, China, India, Iran, Venezuela refiners)
17	Integrated oil-major strategies (Aramco, Exxon, Shell, Petrobras asset deals, refinery stakes, offshore projects)

Table 1. Main topics extracted for each cluster using keyword and document content analysis.

C. Word2Vec. To assess the comparative performance of contextualized versus non-contextualized embeddings, it was decided to incorporate a traditional word embedding technique. Due to the limited availability of computational resources and training data, training a model from scratch was deemed unfeasible. Consequently, the focus was directed towards leveraging a pre-trained model. However, given the declining popularity of traditional embedding methods, few viable pre-trained options were available. Among them, the model word2vec-google-news-300 was selected. This model was trained on approximately 100 billion tokens sourced from Google News using the Skip-gram architecture combined with Negative Sampling, as outlined in the foundational works Efficient Estimation of Word Representations in Vector Space (7) and Distributed Representations of Words and Phrases and their Compositionality (8). The Skip-gram model aims to predict surrounding context words for a given target word within a fixed-size window, effectively capturing both semantic and syntactic relationships. Negative Sampling serves as an efficient approximation of the softmax function by balancing observed word pairs with randomly sampled negative examples. The output of the model is a static embedding vector of 300 dimensions. In this analysis, these embeddings were computed for each document by concatenating the title and excerpt fields, subsequently applying UMAP for dimensionality reduction to two and three dimensions.

C.1. Cluster Analysis. To explore latent semantic structures within the embedded representations, a clustering analysis was conducted on the two-dimensional UMAP projections. The optimal number of clusters k was selected by evaluating values from 2 to 24 using the KMeans algorithm. For each value of k, three clustering quality metrics were computed: the Silhouette Score, which evaluates the compactness and separation of clusters; the Calinski-Harabasz Index, which reflects the ratio of between- to within-cluster dispersion; and the Davies-Bouldin Index, which penalizes overlapping or poorly separated clusters. Among all values tested, k=16 yielded the best balance between these criteria, with a Silhouette Score of 0.401, a Calinski-Harabasz Index of 28,276.38, and a Davies-Bouldin Index of 0.758. This combination indicated high-quality and semantically meaningful clusters, making k=16 the most suitable choice for subsequent analyses.

After clustering the UMAP-reduced word2vec embeddings into k=16 groups, a topic extraction step was performed to interpret the semantic content of each cluster. The resulting clusters revealed clear and coherent thematic groupings. Particularly notable is Cluster 0, which groups documents concerning global oil gas production and pricing trends, including shale output and demand in key economies such as China and Russia.

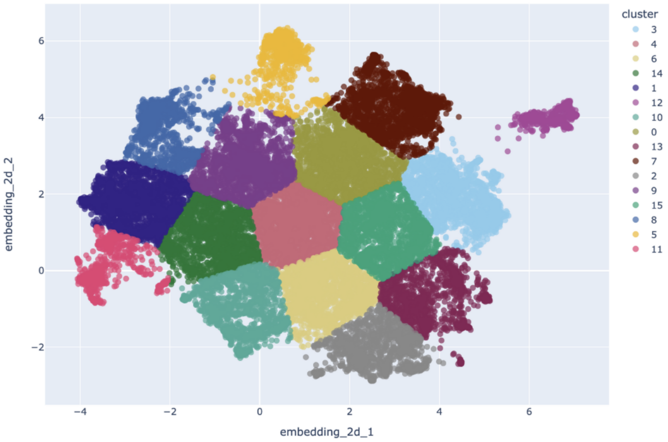


Fig. 4. 2D visualization of clusters obtained from word2vec embeddings using UMAP and KMeans (k = 16).

Cluster 2 focuses on oil infrastructure attacks and security incidents, reflecting geopolitical tensions and operational vulnerabilities. Cluster 6 captures the effects of sanctions on global markets, while Cluster 10 encompasses pipeline politics and legal battles, a key issue for transnational energy transport. Finally, Cluster 15 centers on OPEC production decisions and export quotas, highlighting the role of institutional governance in shaping market dynamics. The quality and interpretability of the clusters suggest that even static word embeddings, when combined with dimensionality reduction and clustering, can effectively uncover latent thematic structures in energy-related news.

Cluster	Topic
0	Global oil & gas production and pricing trends (shale output, demand in China/Russia, industry records)
1	Power generation and renewables transition (solar, wind, coal, emissions, capacity in UK/China/India)
2	Oil infrastructure attacks and security incidents (tanker/pipeline strikes, Iran/Libya/ISIS/Houthis)
3	Crude export metrics and trade flows (OPEC/Russia output, China/India imports, barrels-per-day records)
4	Upstream LNG & offshore project deals (new field developments, Shell/Exxon exploration and production)
5	Oil majors' financial results (quarterly earnings, profit beats/misses, refining segment performance)
6	Sanctions' impact on oil & gas markets (Russia/Iran restrictions, EU LNG deals, pipeline shifts)
7	Oil price dynamics and demand analysis (gasoline trends, supply/demand balance, OPEC signals)
8	Electric vehicles and clean mobility (Tesla/EV sales, batteries, market growth in China/UK)
9	Aramco asset transactions and investments (stakes, IPOs, Saudi fund deals, Shell/Exxon participations)
10	Pipeline politics and legal battles (Nord Stream, Keystone XL, Trans Mountain, court rulings)
11	Nuclear & power-plant developments (new reactors, grid resilience, Japan/Iran/UK energy tariffs)
12	Crude inventory reports & price drivers (API builds/draws, surprise stock changes, rally expectations)
13	Refinery operations and disruptions (exports, pipeline flows, strikes in Libya/Mexico, hurricane impact)
14	Energy policy & taxation debates (UK/EU climate bills, fracking, windfall taxes, natural gas levies)
15	OPEC production decisions and cuts (Saudi/Russia output, India/Iran imports, export quotas)

Table 2. Main topics extracted for each cluster using keyword and document content analysis.

D. Comparative Analysis of Clustered Topics Derived from DistilBERT and Word2Vec Embeddings. The use of DistilBERT embeddings resulted in the identification of 18 thematic clusters, characterized by high semantic specificity. The topics are finely segmented along geopolitical, financial, and technological dimensions. Conversely, clustering based on Word2Vec embeddings yielded 16 broader clusters, often aggregating related but distinct sub-themes into unified semantic areas.

For instance, geopolitical dynamics and sanctions-related themes are segmented in DistilBERT into:

- Cluster 0: *Geopolitical energy dynamics and sanctions*
- Cluster 3: *Market fluctuations under sanctions*
- Cluster 9: *Infrastructure attacks and threats*

In contrast, Word2Vec groups these aspects mainly into:

- Cluster 2: *Infrastructure attacks*
- Cluster 6: *Sanctions' impact on oil & gas markets*

The contextual nature of DistilBERT allows it to disambiguate between types of geopolitical influence (economic, physical, legal), while Word2Vec, based on fixed co-occurrence windows, captures a more generalized representation. Both models converge on identifying core themes in global energy discourse:

- **OPEC policy and production metrics:** DistilBERT Clusters 7, 11, 12 vs. Word2Vec Clusters 3, 15
- **Renewable transition and power generation:** DistilBERT Cluster 6 vs. Word2Vec Cluster 1
- **Electric vehicles and clean energy:** DistilBERT Cluster 13 vs. Word2Vec Cluster 8
- **Oil majors' financial performance:** DistilBERT Cluster 8 vs. Word2Vec Cluster 5

However, DistilBERT's segmentation is more fine-grained. For example, financial reporting is treated separately from long-term strategic investments (Cluster 17), which Word2Vec tends to merge into the same cluster.

Some topics emerge exclusively or more distinctly in one model:

- **Only in DistilBERT:**
  - Cluster 14: China's crude trade flows
  - Cluster 10: Upstream partnerships
- **Only in Word2Vec:**
  - Cluster 11: Nuclear and power plant developments
  - Cluster 14: Energy policy and taxation
  - Cluster 13: Refinery operations and disruptions

These differences can be attributed to the underlying architectures: contextual transformers can disambiguate similar terms across contexts, while static models tend to generalize over frequent co-occurrences.

E. Temporal Evolution of Topics (2011–2025). The longitudinal analysis reveals strong convergence between the two models in identifying structural and exogenous events that shaped the energy discourse.

Geopolitics and Sanctions (Post-2016 and 2022).

- In DistilBERT, Cluster 0 exhibits a marked increase starting in 2016 and peaks in 2022, corresponding to the Russia–Ukraine conflict and associated sanctions.
- In Word2Vec, the semantically aligned Cluster 6 peaks similarly in 2022.

Energy Transition and Electric Mobility.

- DistilBERT Cluster 6 and Word2Vec Cluster 1 show rising trends from 2018 onward, reflecting increasing policy attention to decarbonization.
- Cluster 13 (DistilBERT) and Cluster 8 (Word2Vec), relating to EVs, expand steadily from 2020 to 2024, consistent with global growth in EV adoption and battery investment.

Corporate Strategy and Financial Reporting. Clusters addressing oil majors' earnings and strategic transactions (Clusters 8, 17 in DistilBERT; 5, 9 in Word2Vec) show recurrent spikes during periods of heightened market volatility, such as 2015 (price crash), 2020 (COVID-19), and 2022 (energy crisis).

This comparative study highlights fundamental differences in the semantic structuring capabilities of contextual vs. static embeddings:

- **DistilBERT** enables a more nuanced and context-sensitive topic segmentation, suitable for identifying emerging sub-themes and disambiguating actors, actions, and geographies within the same thematic domain.
- **Word2Vec**, while semantically coarser, offers robust aggregation of major thematic axes, suitable for macro-level topic monitoring.

3 Feature Creation

A. Mean. To construct a coherent representation of each day based on the published news articles, the daily average of the 2D-projected embeddings was computed. Each embedding corresponds to a single article and reflects its semantic location in the reduced-dimensional space. Given that multiple articles are typically published on the same day, each potentially covering different yet related topics, the objective was to derive a single embedding that summarizes the collective semantic content of the day.

Averaging the embeddings for articles published on the same date serves two main purposes. First, it provides a semantic aggregation mechanism: rather than analyzing each article separately, the central tendency of their embeddings is assumed to approximate the dominant thematic content of the day. This is based on the notion that the mean of semantically similar vectors tends to fall within the shared topic cluster, while in the presence of diverse content, the mean still captures an interpretable balance of themes. Second, this aggregation ensures temporal alignment with the target variable in the downstream forecasting task. In this context, the goal is to predict daily fluctuations in the price of Brent crude oil. Since the target variable is defined at a daily resolution, it is necessary to construct input features that follow the same temporal granularity (Results in Figure 5).

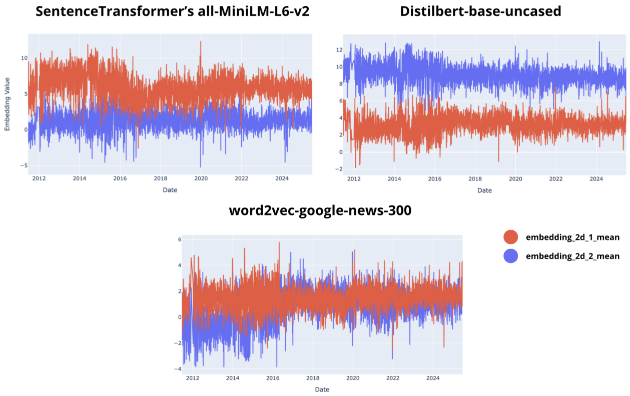


Fig. 5. Average daily values of the first two components of 2D-projected embeddings from three models: `SentenceTransformer's all-MiniLM-L6-v2`, `DistilBERT-base-uncased`, and `word2vec-google-news-300`. Each line represents the temporal mean of one embedding dimension across news articles published each day.

SentenceTransformer's all-MiniLM-L6-v2. Among the three models, MiniLM exhibits the largest dynamic range and variability, particularly in the first component. The daily averages fluctuate significantly over time, suggesting that this model is highly sensitive to shifts in the semantic content of the daily news corpus. This behavior can be attributed to the fact that MiniLM is a transformer-based sentence embedding model fine-tuned on semantic similarity tasks. As such, it is optimized to capture nuanced contextual differences between textual inputs. The observed volatility in the temporal embedding signal indicates that MiniLM is effective in encoding topical changes and emerging events.

DistilBERT-base-uncased. In contrast, the embeddings derived from DistilBERT show a more stable pattern over time. The average values of the first two components vary less, especially the first component, which remains within a relatively narrow band. Interestingly, the second component displays a consistently elevated baseline, indicating a potential directional bias in the reduced embedding space. This model's relative smoothness may reflect its architecture: while DistilBERT retains much of BERT's language modeling capacity, it is not explicitly fine-tuned for sentence-level semantics. As a result, it provides a robust but less reactive representation.

word2vec-google-news-300. The word2vec-based embeddings, finally, demonstrate the lowest variance and the narrowest value range across both dimensions. The corresponding time series appear noisy but semantically diluted, with limited structure or discernible trends. This behavior is consistent with the limitations of static word embeddings, which assign a fixed vector to each word regardless of context. When aggregated across multiple words and articles, these embeddings tend to average out, resulting in less informative document-level representations. Consequently, word2vec embeddings are less capable of capturing day-to-day thematic shifts.

B. Daily Embedding Dispersion as a Measure of Semantic Heterogeneity. In addition to the computation of the daily mean of the 2D-projected embeddings, a complementary feature was introduced: the *daily embedding variance*, aimed at capturing the semantic dispersion of news articles published on the same day. Specifically, for each date, the variance of the *embedding_2d_1* and *embedding_2d_2* components was calculated across all articles grouped by day. This procedure yields two additional features: *embedding_2d_1_var* and *embedding_2d_2_var*.

From a semantic standpoint, these variance-based indicators quantify the degree of intra-day heterogeneity in the news corpus. A low variance suggests that most articles on a given day are clustered in close proximity within the reduced semantic space, indicative of thematic convergence around one or a few dominant topics. In contrast, a high variance implies that articles are widely dispersed across the embedding space, reflecting the coexistence of multiple, potentially unrelated or loosely connected topics.

This feature proves particularly valuable for identifying complex or fragmented news cycles, such as those associated with geopolitical crises, regulatory shifts, or the temporal overlap of several global events. Moreover, it may serve as a proxy for informational uncertainty or cognitive load, insofar as diverse and conflicting narratives demand increased interpretative effort from both market participants and the general public.

Notably, the inclusion of daily dispersion features complements the information provided by the mean embeddings, offering a second-order semantic signal. Whereas the mean captures the central tendency of the semantic content, the variance highlights the degree of consistency or fragmentation in the discourse (Results in Figure 6).

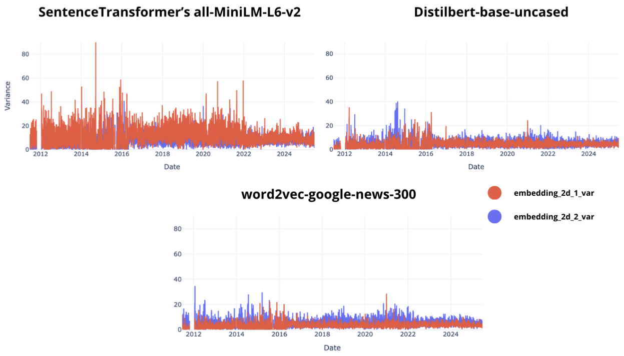


Fig. 6. Daily variance of the two 2D-projected embedding components (`embedding_2d_1_var` and `embedding_2d_2_var`) across models. Higher values indicate greater semantic dispersion in the news content published on the same day.

SentenceTransformer's all-MiniLM-L6-v2. The MiniLM-based embeddings exhibit the highest and most sustained levels of daily variance, particularly in the first component. This indicates that the model is highly sensitive to the presence of multiple co-occurring themes within a single day. The variance signals reach values exceeding 80 in certain periods, suggesting that the semantic space derived from MiniLM is both expressive and finely attuned to subtle topic differentiation. This behavior aligns with MiniLM's architecture and training objective: as a transformer-based sentence encoder fine-tuned on semantic similarity tasks, it is designed to detect nuanced contextual distinctions, making it particularly effective in representing days with fragmented or multifaceted discourse.

DistilBERT-base-uncased. In contrast, the variance profiles for DistilBERT are more moderate and less volatile. While the second component occasionally exhibits localized spikes, overall dispersion remains lower than in MiniLM. The lower variance reflects a more compressed semantic space, where articles are represented in a way that smooths over fine-grained thematic differences. This can be attributed to the fact that DistilBERT, although based on the BERT architecture, is not explicitly fine-tuned for capturing sentence-level semantic variation. As a result, it encodes documents in a more generalized and stable manner, which may be advantageous for tasks prioritizing consistency over granularity.

word2vec-google-news-300. Word2Vec embeddings display the lowest levels of daily dispersion, with both components remaining below 20 for the entire period. The reduced variance is a consequence of the model's static nature: each word is represented by a single fixed vector, regardless of context, and document embeddings are typically obtained through averaging. This results in a loss of contextual richness and limited sensitivity to semantic diversity, causing the daily variance to flatten. Consequently, Word2Vec appears less suitable for capturing thematic fragmentation or complex news cycles, as its representational power is fundamentally constrained.

C. Semantic Change Metrics Based on Cosine Similarity. To quantify the temporal dynamics of semantic content in the news corpus, a set of features was derived from the cosine similarity between the mean embedding vectors of consecutive days. These features aim to measure not only the degree of semantic change, but also its temporal progression, offering a multi-scale characterization of discourse evolution.

C.1. Cosine Similarity. The *cosine similarity* between two consecutive days, defined as

$$\text{cosine\_sim}_t = \frac{\mathbf{e}_t \cdot \mathbf{e}_{t-1}}{\|\mathbf{e}_t\| \|\mathbf{e}_{t-1}\|},$$

where $\mathbf{e}_t$ is the mean embedding vector for day t, captures the angular closeness between the semantic representations of adjacent days. Values close to 1 indicate strong similarity and thematic continuity, while values approaching 0 denote orthogonal or highly divergent discourse content. This metric serves as a baseline indicator of semantic stability over time.

C.2. Semantic Drift. To directly express the magnitude of semantic change, the notion of *semantic drift* was introduced as the complement of cosine similarity:

$$\text{drift}_t = 1 - \text{cosine\_sim}_t.$$

This transformation converts a similarity measure into a dissimilarity metric bounded in $[0, 1]$, where higher values correspond to more substantial changes in content. Semantic drift provides a more interpretable scale for detecting disruptions or thematic transitions in the discourse.

C.3. Drift Velocity. Building on Semantic Drift, the notion of *drift velocity* was defined as a rolling average of semantic drift over a fixed temporal window. Formally,

$$v_t = \frac{1}{w} \sum_{i=t-w+1}^{t} \text{drift}_i,$$

with $w = 3$ days in the present study. This quantity reflects the average speed at which semantic content evolves, smoothing out short-term noise and highlighting persistent changes in narrative structure.

C.4. Drift Velocity Difference. To capture the local dynamics of change, the first-order temporal difference of the drift velocity, referred to as *drift velocity difference*, was computed as

$$\Delta v_t = v_t - v_{t-1}.$$

This feature indicates whether the semantic shift is accelerating or decelerating. Positive values suggest that the pace of change is increasing, while negative values denote a return to semantic stability.

C.5. Drift Acceleration. Finally, the second-order difference, or *drift acceleration*, was introduced to quantify the curvature of the semantic trajectory:

$$a_t = v_t - 2v_{t-1} + v_{t-2}.$$

This higher-order feature captures sudden changes in the dynamics of discourse. High positive acceleration values may signal the onset of semantic shocks or abrupt topic shifts, whereas negative acceleration suggests a deceleration of narrative volatility, often observed during post-crisis stabilization phases.

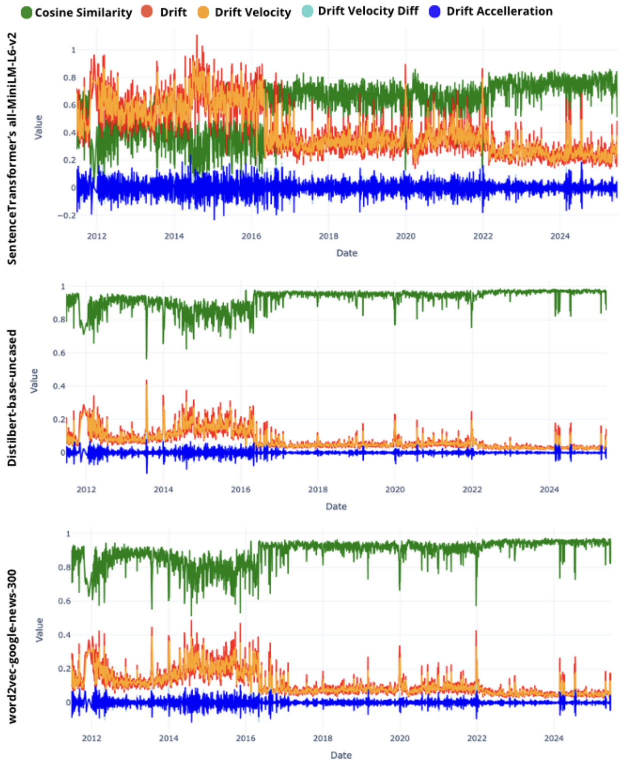


Fig. 7. Temporal evolution of semantic change metrics for three embedding models: `SentenceTransformer's all-MiniLM-L6-v2`, `DistilBERT-base-uncased`, and `word2vec-google-news-300`. The plotted metrics include cosine similarity (green), drift (orange), drift velocity (red), drift velocity difference (cyan), and drift acceleration (blue). The visualization highlights model-specific sensitivity to semantic shifts and multi-scale discourse dynamics.

SentenceTransformer's all-MiniLM-L6-v2. The MiniLM-based embeddings exhibit the most pronounced temporal variability across all semantic drift metrics. Cosine similarity fluctuates considerably over time, with frequent dips indicating sensitivity to daily thematic changes. Correspondingly, the drift values (1 -

cosine similarity) are persistently high, often exceeding 0.4, suggesting that this model consistently detects non-trivial semantic shifts. The drift velocity is relatively elevated and exhibits long-lasting oscillations, implying that content variation is sustained rather than transient.

Notably, the first- and second-order derivatives (velocity difference and acceleration) are well structured and non-flat, revealing coherent patterns of semantic momentum and curvature. This indicates that MiniLM is capable of capturing multi-scale transitions in discourse, including both smooth evolutions and sudden shocks.

DistilBERT-base-uncased. DistilBERT embeddings reveal a much more stable semantic trajectory. Cosine similarity values are predominantly high and consistent over time, reflecting high semantic continuity. The associated drift values remain low throughout the timeline, with only minor peaks corresponding to sporadic changes in discourse. As a result, the drift velocity is also low and relatively flat.

The derivative metrics remain close to zero, indicating that the model is largely insensitive to fine-grained semantic fluctuations.

word2vec-google-news-300. Word2Vec-based embeddings show similar behavior to DistilBERT in terms of high and stable cosine similarity, with limited day-to-day variability. However, some isolated spikes in drift and velocity are visible, particularly in earlier years. These are likely the result of changes in surface-level word distributions rather than true contextual shifts, since Word2Vec does not account for polysemy or sentence-level meaning.

The low and noisy nature of the higher-order derivatives further confirms that Word2Vec embeddings lack the semantic resolution needed to track discourse transitions effectively.

4 Sentiment Analysis of Oil News Titles

Another feature incorporated into the model is sentiment, which has been extensively documented to enhance the accuracy of financial market forecasts (9). Accordingly, sentiment scores were derived from news headlines related to the oilprice.com, with the objective of capturing the prevailing market tone and expectations that frequently anticipate short-term price fluctuations.

To extract sentiment signals, the transformer-based model FinBERT was initially adopted. FinBERT is a domain-specific adaptation of BERT, designed for financial sentiment analysis (10). It is pre-trained on general English corpora to learn language structure, then further adapted to the financial domain through continued pre-training on financial news datasets such as the Thomson Reuters TRC2 corpus. The model is subsequently fine-tuned on the Financial Phrase Bank, a labeled dataset of financial sentences annotated with sentiment classes (positive, neutral, negative). In the context of crude oil news, however, FinBERT produced suboptimal results. Sentiment labels were often misaligned with actual market implications: headlines anticipating price increases were frequently labeled as negative, whereas headlines related to oil's price reductions, were labeled as neutral or positive. This discrepancy arises from FinBERT's limited capacity to internalize commodity-specific price mechanisms, particularly those governed by the economic principles of supply and demand.

Due to the unavailability of labeled datasets tailored to crude oil sentiment, the existing literature was surveyed to identify alternative approaches. A suitable candidate was found in CrudeBERT, a model developed by Kaplan et al. (2023) (11), which explicitly addresses the limitations of FinBERT in the oil domain. CrudeBERT is fine-tuned using a methodology that incorporates economic theory to better align sentiment labels with price impacts in the crude oil market. The authors document that FinBERT consistently misclassifies headlines such as "Fire at major oil platform" as negative, despite the fact that such events reduce supply and are typically associated with upward price movements. CrudeBERT corrects these inconsistencies by learning the directional relationships between events and price responses specific to the oil sector.

A. Crudebert Fine-tuning approach. Given the absence of publicly available, labeled sentiment datasets specific to crude oil, the authors of CrudeBERT constructed a silver-standard dataset using a semi-automated labeling strategy grounded in the classical theory of price formation, namely the law of supply and demand (Smith, 1776). The initial corpus consisted of approximately 46,000 headlines relevant to the oil market, collected from the RavenPack Realtime News Discovery platform and spanning the period from 2000 to 2021.

To assign sentiment labels, the headlines were first categorized into a set of frequently recurring topics, including accidents, oil discoveries, import/export changes, pipeline disruptions, and supply/demand variations. Each topic was then associated with its likely directional effect on crude oil prices. For example, a headline describing a refinery accident was labeled as positive due to the expected price increase resulting from supply disruption, while news of increased oil production or exports was labeled as negative due to the anticipated surplus. Headlines not expected to significantly affect price dynamics were labeled neutral.

This labeling approach resulted in a silver-standard dataset comprising approximately 15,000 positive, 14,000 negative, and 500 neutral headlines. Label assignment was driven by a combination of keyword-based topic detection and directional polarity inference using lexical patterns and numerical cues ("+10%," "decline," "drop in exports").

Using this dataset, FinBERT was fine-tuned to produce CrudeBERT through a supervised training regime. The model architecture consisted of the pre-trained BERT-base encoder with an added softmax classification head. The dataset was split into training (60%), validation (20%), and test (20%) sets, with stratified sampling to preserve class distribution. Despite the relatively small number of neutral examples, they were retained to provide the model with a broader range of semantic signals and improve generalization beyond binary classification.

B. Aggregation of News-Based Sentiment Scores for Crude Oil Forecasting . To incorporate qualitative information from news media into a quantitative forecasting framework, a sentiment analysis pipeline was applied to a dataset consisting of daily headlines related to the crude oil market. Each headline was classified using CrudeBERT which assigns one of three sentiment labels: positive, neutral, or negative. These categorical labels were subsequently mapped to scalar values of +1, 0, and 1, respectively. Out of the total analyzed news headlines, 11,816 were classified as positive, 11,297 as negative, and only 307 as neutral. Then for each calendar day, the sentiment scores of all associated headlines were aggregated using the arithmetic mean, yielding a daily sentiment index. This scalar captures the average tone of the news coverage on that day, where positive values indicate predominantly optimistic sentiment, and negative values reflect bearish or pessimistic tones.

5 Modelling

A. Transformer-based Stock Price Prediction with Autoencoder Denoising. In this study, was develop a hybrid modeling approach that integrates a denoising autoencoder with a Transformer encoder to predict the next-day closing prices of Brent crude oil based on multivariate time series data following the paper Stock Price Prediction with Denoising Autoencoder and Transformers (12). Financial time series are notoriously volatile and noisy, posing significant challenges for traditional statistical methods and recurrent neural networks to effectively capture complex temporal dependencies and abrupt market shifts. Although RNN architectures like LSTM and GRU have shown strong capabilities in sequential financial modeling, they are often limited by issues such as vanishing gradients and inherently sequential processing, which restrict their ability to leverage parallel computation efficiently. To enrich the input data, the original OHLCV features Open, High, Low, Close, and Volume which were complemented with a comprehensive set of embedding-based features, including sentiment indicators and drift-related statistics, all of which were normalized using min-max scaling to maintain consistent and stable ranges across the dataset.

Time series were segmented into fixed-length sliding windows of 64 hours as input sequences, with the model tasked to predict the subsequent closing price. The autoencoder architecture comprises bidirectional LSTM layers encoding the high-dimensional multivariate input into a compact latent representation, followed by a decoder reconstructing the original input to enforce feature denoising and dimensionality reduction in a self-supervised manner. This latent encoding effectively filters noise and extracts salient temporal patterns relevant for price movement.

The Transformer predictor consumes these latent embeddings, employing multi-headed scaled dot-product attention mechanisms within a stacked Transformer encoder architecture. This design enables the model to capture complex temporal dependencies and contextual relationships within the denoised latent space. Dropout regularization mitigates overfitting observed during training, and hyperparameters such as latent dimensionality, number of attention heads, and Transformer layers were empirically tuned for optimal trade-off between model complexity and generalization.

Training proceeded jointly on the autoencoder reconstruction loss and the Transformer's mean squared error for price prediction, optimizing both components simultaneously. Evaluation metrics including MSE, MAE, RMSE, and R^2 score were computed on validation and test splits, with results indicating improved convergence and predictive performance compared to baseline approaches.

A.1. Model Evaluation. The Transformer-Autoencoder architecture was trained using three different textual embedding methods: `SentenceTransformer's all-MiniLM-L6-v2`, `DistilBERT-base-uncased`, and `Word2Vec-GoogleNews-300`. Each embedding was used to encode the semantic content of news articles, generating the features described in the previous chapter, including sentiment-related information. These features, along with the corresponding day's price, were then fed into the autoencoder-based model to predict the next day's price, with the aim of capturing temporal dependencies in Brent crude oil price dynamics.

Among the evaluated models, the variant using `MiniLM` embeddings exhibited the best overall performance, achieving a Mean Squared Error (MSE) of 6.70, a Mean Absolute Error (MAE) of 2.05, and a Root Mean Squared Error (RMSE) of 2.59. The corresponding coefficient of determination ($R^2 = 0.866$) confirms the model's strong ability to explain the variance in the target variable.

The models trained on `DistilBERT` and `Word2Vec` embeddings also performed competitively, with R^2 values above 0.86.

Table 3. Performance metrics of Transformer-Autoencoder models with different embeddings

Embedding	MSE	MAE	RMSE	R^2
MiniLM	6.7049	2.0538	2.5894	0.8658
DistilBERT	6.9156	1.9939	2.6298	0.8616
Word2Vec	6.8662	2.0490	2.6203	0.8626

When applying a Transformer-based predictive model with autoencoder denoising, MiniLM embeddings outperformed both Word2Vec and DistilBERT across most evaluation metrics. This performance gain can be attributed to the architectural alignment between the embedding model and the downstream architecture: both are Transformer-based and operate in similar representation spaces. Furthermore, the autoencoder mechanism appears to mitigate the noise and redundancy typically associated with contextual embeddings, allowing the model to effectively extract relevant temporal patterns from the rich semantic signals encoded by MiniLM. In contrast, Word2Vec embeddings, while more stable and interpretable, lack the contextual sensitivity required to fully exploit the model's capacity for sequence-level representation learning.

B. Random forest. The second model implemented in this study is a Random Forest regression model, designed to forecast next-day Brent crude oil prices. Historical daily closing prices for Brent crude oil were obtained from Yahoo Finance covering the period from June 2011 to June 2025. The core dataset was augmented with the feature discussed in the chapter before: embedding-based features, drift-related statistics, and sentiment metrics.

The target variable was defined as the closing price of the following day, thereby framing a one-step-ahead forecasting task.

Data was split chronologically into training (90%) and testing (10%) subsets to preserve temporal integrity and avoid data leakage. A Random Forest regressor with 200 trees was trained on the training set, leveraging parallel processing for efficiency. Model evaluation employed a variety of metrics including root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), symmetric MAPE (sMAPE), coefficient of determination (R^2), explained variance score, Theil's U statistic, and directional accuracy, providing a multifaceted assessment of predictive performance.

Evaluation of Random Forest Models with Varying Embedding Architectures. To assess the impact of different embedding representations on predictive performance, three Random Forest models were trained using identical feature engineering and hyperparameters, each differing only by the embedding model used to encode the news corpus and extract the features. The three architectures evaluated were: (i) `all-MiniLM-L6-v2`, (ii) `distilbert-base-uncased`, and (iii) `word2vec-google-news-300`

As shown in Table 4, the model using `all-MiniLM-L6-v2` achieved the best overall performance with an RMSE of 1.4577 and an R^2 of 0.9581, outperforming the others across most metrics. The model using `distilbert-base-uncased` followed closely, with an RMSE of 1.5681 and an R^2 of 0.9515. Meanwhile, the model based on `word2vec` embeddings showed slightly lower accuracy (RMSE = 1.5969; R^2 = 0.9497), but still delivered consistent performance.

Table 4. Random Forest Performance with Different Embeddings

Embedding	RMSE	MAE	MAPE%	sMAPE (%)	R^2	Dir. Acc.%
`MiniLM`	1.4577	1.0895	1.4488	1.4451	0.9581	55.49
`distilbert`	1.5681	1.1689	1.5471	1.5432	0.9515	51.93
`word2vec`	1.5969	1.2053	1.6052	1.6009	0.9497	46.5875

All models maintained symmetric MAPE (sMAPE) below 1.6%, confirming the robustness of percentage-based error handling. Theil's U statistics remained close to zero across models, further validating their predictive advantage over naïve benchmarks. Directional Accuracy ranged from 46.5875% to 55.49%, suggesting that while embedding choice influences trend detection capacity, improvements remain marginal in that regard.

Interpreting SHAP Values: The Added Value of Embedding-Derived Features. Once the Random Forest model is trained, local explainability can be performed using SHAP values. This approach enables the identification of the most influential features for individual predictions, allowing for a detailed instance-level interpretation of model behavior. In the context of Brent crude oil forecasting, this means understanding which signals contributed most to the predicted price for a given day. In addition to highlighting individual feature contributions, SHAP values also allow for a semantic comparison across features derived from different types of embeddings. This makes it possible to investigate how distinct embedding models interpret the same input data and what thematic patterns they prioritize.

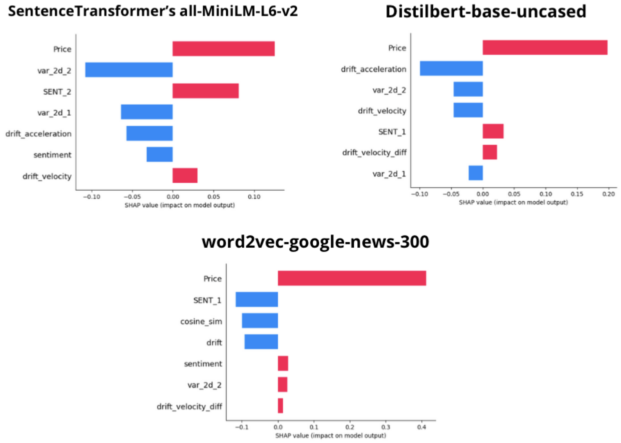


Fig. 8. SHAP value explanation for the prediction of June 18 using features from June 17, across three embedding models: MiniLM, DistilBERT, and Word2Vec

To exemplify this, the prediction for June 18 is analyzed, based on input features from June 17. The SHAP analysis highlights significant differences in how each embedding model contributes to the forecast:

- Price is consistently identified as the most impactful feature across all models, confirming its strong autoregressive influence on Brent crude oil price movements.

- The model using MiniLM embeddings exhibits a more balanced distribution of contributions across features such as `SENT_2` (embedding2d_mean), `var_2d_1`, and `drift_acceleration`, indicating a nuanced integration of semantic signals. The predicted value (76.93) slightly overshoots the actual target (76.70).

- The DistilBERT-based model assigns higher importance to fewer features, with a strong positive impact from `Price` and a negative contribution from `drift_acceleration`. The prediction (76.38) underestimates the target, likely due to heightened sensitivity to volatility-related variables.

- The model based on Word2Vec embeddings relies more heavily on `Price` and `SENT_1` (embedding2d_mean), producing the highest SHAP magnitude among all models. Its prediction (76.52) is closer to the actual value than that of DistilBERT, but less accurate than MiniLM.

Furthermore, the semantic context, as captured by the dominant topic cluster on that day, appears to influence model behavior:

- MiniLM embeddings focus on *global macroeconomic signals and supply-demand balance* (Cluster 13),

- DistilBERT captures *geopolitical risks and infrastructure-related events* (Cluster 9),

- Word2Vec highlights *gasoline demand and OPEC-related dynamics* (Cluster 7).

C. XGBoost. A complementary modeling approach was developed using the XGBoost algorithm, a gradient boosting framework well-suited for structured regression tasks with high-dimensional and potentially correlated features. The same historical dataset of daily Brent crude oil closing prices (June 2011–June 2025) was enriched with a suite of features derived from financial news embeddings and sentiment analysis. In this context, XGBoost's ability to model non-linear relationships and capture complex feature interactions proved particularly advantageous. To ensure comparability, the same set of input features adopted in the Random Forest model was employed. These included two-dimensional embedding means and variances, cosine similarity with the previous day (`cosine_sim`), semantic drift measures (`drift`, `drift_velocity`, `drift_acceleration`), and an aggregated daily sentiment score derived from financial news content.

Training and evaluation adhered to a time-aware design: the first 90% of the time series was used for training, while the final 10% served as the test set. The XGBoost model was configured with 200 boosting rounds, a learning rate of 0.07, and regularization via subsampling and feature sampling to prevent overfitting and improve generalization.

Evaluation of XGBoost Models with Varying Embedding Architectures. A comprehensive evaluation of the XGBoost model was conducted using a diverse set of performance metrics, including both absolute and relative error measures, as well as directional indicators. The model was tested in three different types of text embeddings, `MiniLM`, `DistilBERT`, and `Word2Vec` to assess the impact of each representation on the predictive precision of the movements in the prices of Brent crude oil.

Among the configurations, the model leveraging `Word2Vec` embeddings demonstrated the best performance overall. It achieved the lowest RMSE (1.5870), MAE (1.1876), and sMAPE (1.5616%), along with a competitive R^2 score of 0.9503 and Directional Accuracy of 52.23%. Surprisingly, despite its static nature, `Word2Vec` outperformed both transformer-based models in all absolute and relative error metrics.

The `DistilBERT` based model ranked second, showing improved Directional Accuracy (53.71%) and a relatively high R^2 score (0.9417), though with higher errors than `Word2Vec`. The model using `MiniLM` embeddings showed the weakest results across all metrics, with the highest RMSE (1.9432), MAE (1.4981), and sMAPE (1.9905%), as well as the lowest R^2 (0.9255). This counterintuitive outcome suggests that more compact transformer models may underperform in settings with limited training data or noisy financial targets.

Interestingly, the static Word2Vec embeddings outper-

Table 5. XGBoost Performance with Different Embeddings

Embedding	RMSE	MAE	MAPE (%)	sMAPE (%)	R^2	Dir. Acc. (%)
MiniLM	1.9432	1.4981	2.0060	1.9905	0.9255	54.5994
DistilBERT	1.7188	1.2609	1.6570	1.6508	0.9417	53.71
Word2Vec	1.5870	1.1876	1.5695	1.5616	0.9503	52.23

formed their contextualized counterparts (MiniLM and DistilBERT) across most evaluation metrics. This result can be attributed to several factors. First, the handcrafted features derived from the embeddings, such as cosine similarity, drift, and 2D variance, already encapsulate much of the semantic signal needed for forecasting, possibly reducing the added value of more complex representations. Second, static embeddings offer greater temporal stability and are inherently more compatible with averaging operations across multiple news articles per day. Finally, XGBoost tends to benefit from lower-variance, smoother features, which Word2Vec naturally provides. These findings suggest that, in certain forecasting contexts, simpler embedding models can be more effective when paired with carefully engineered semantic and sentiment-based features.

6 Conclusion and Future Developments

The integration of semantic information derived from financial news into quantitative models for crude oil price forecasting has proven to be both effective and informative. Embedding-based features, such as sentiment orientation, semantic drift, and distributional variance, allow models to capture dimensions of market dynamics that are inaccessible through historical price data alone. These features encode forward-looking signals embedded in the information ecosystem, particularly useful in markets influenced by exogenous shocks, geopolitical tensions, and supply chain disruptions.

This approach bridges the gap between qualitative narratives and quantitative prediction, demonstrating that news content, when transformed into appropriate vector representations, can enrich the feature space with anticipatory signals. The added interpretability through SHAP analysis further shows how embedding-based variables contribute to predictions, providing decision-makers with a clearer understanding of the underlying drivers behind model outputs.

Different modeling architectures benefit in different ways: Transformer-based models are best suited to capture the full potential of contextual embeddings like MiniLM, while simpler static embeddings such as Word2Vec show strong performance in ensemble tree methods like XGBoost due to their temporal stability and compatibility with feature aggregation strategies.

Future Developments. Future research may extend this work in several directions. First, the inclusion of multimodal data, such as macroeconomic indicators, energy inventory reports, or satellite-based supply chain signals, could further improve forecasting performance and generalization. Second, more advanced time-aware sentiment models could be explored, especially those trained with causal supervision to better align sentiment with future price directionality. Third, a real-time system could be implemented to test this framework in production-like settings, validating its utility in decision support for trading or policy monitoring. Lastly, the embedding pipeline could be expanded to incorporate cross-lingual news sources, enabling a global perspective on oil market sentiment.

7 Bibliography

1. OilPrice.com. Latest world energy news. https://oilprice.com/Latest-Energy-News/World-News/, 2025. Accessed: 2025-07-04.
2. Nils Reimers and Iryna Gurevych. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021. Accessed: 2025-07-03.
3. Wenhui Wang, Furu Wei, Li Dong, Hang Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*, 2020.
4. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
5. Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
6. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
7. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient

estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, 2013.

8. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, 2013.
9. Bin Qian and Khalid Rasheed. Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1):25–33, 2007.
10. Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
11. Himmet Kaplan, Ralf-Peter Mundani, Heiko Rölke, and Albert Weichselbraun. Crudebert: Applying economic theory towards fine-tuning transformer-based sentiment analysis models to the crude oil market. *arXiv preprint arXiv:2305.06140*, 2023.
12. Zhiyang Chen. Stock price prediction with denoising autoencoder and transformers. *Highlights in Science, Engineering and Technology CSIC 2023, Volume 85*, pages 803–810, 2024.