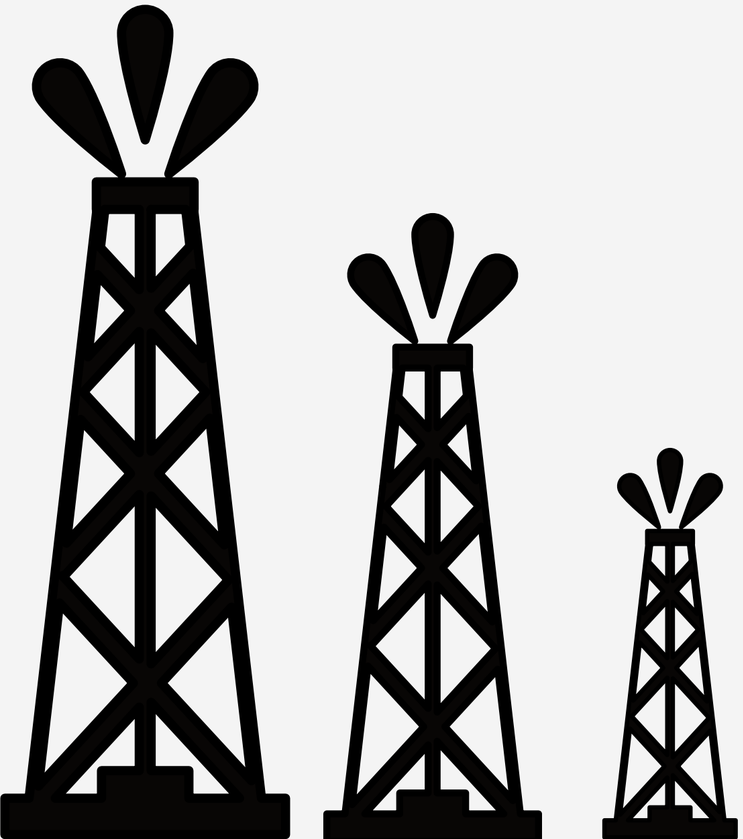
SignalMining-OilEmbeddings

Sara Borello - 882793
Keita Jacopo Viganò - 870980



The Goal

OIL NEWS

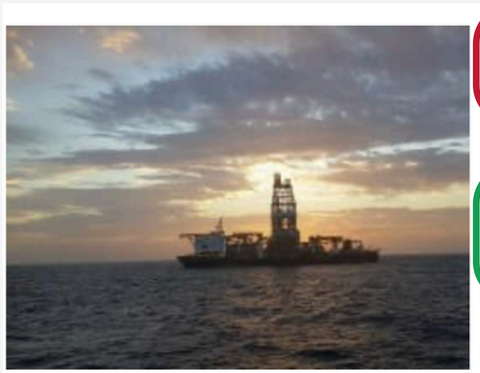
Transform news articles into quantitative signals that capture sentiment shifts and latent dynamics potentially anticipating market movements

FIT DIFFERENT MODELS

- Random Forest
- XGBoost
- Transformers

The Source of Data

- Data obtained via scraping from the website
- Data retrieved from **June 18, 2011,** to **June 25, 2025**, consisting of a total of **23,419 news** articles.



Floating LNG Port in German Baltic Sea Supplies Record-High Gas Volume in Q2

← **Title**

Jul 03, 2025 at 11:05 | Charles Kennedy

The floating LNG import terminal in Mukran in the German part of the Baltic Sea achieved in the second quarter record-high gas deliveries for Germany's gas grid, the terminal operator...

← **Abstract**

EY Audit Breach Forces Shell to Amend U.S. Filings

Jul 03, 2025 at 08:26 | City A.M

← **Date**

Shell is to file amended regulatory filings after Big Four firm EY told the oil giant it broke audit rules after its lead...

Oil News Today

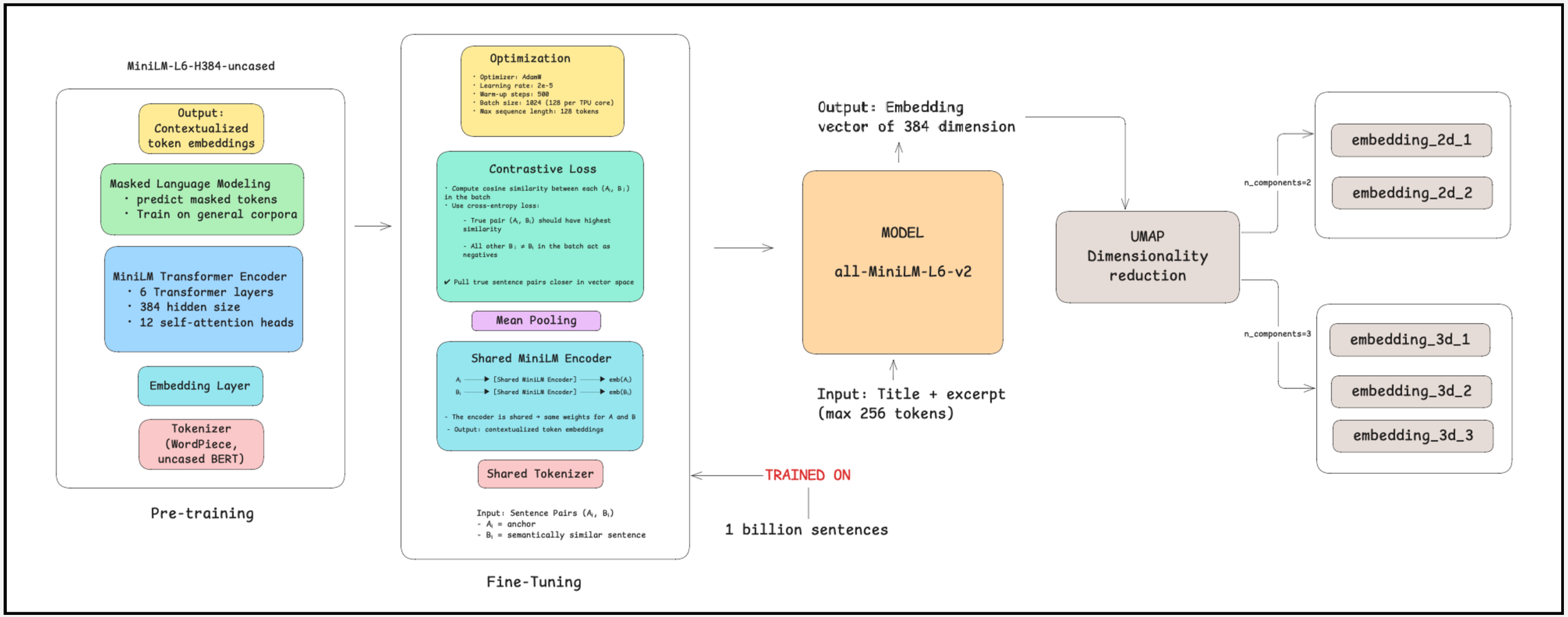
Latest world news from the energy sector. Our news analysis covers Fossil fuels, alternative energy and environmental developments.

OilPrice.com

← **Source of Data**

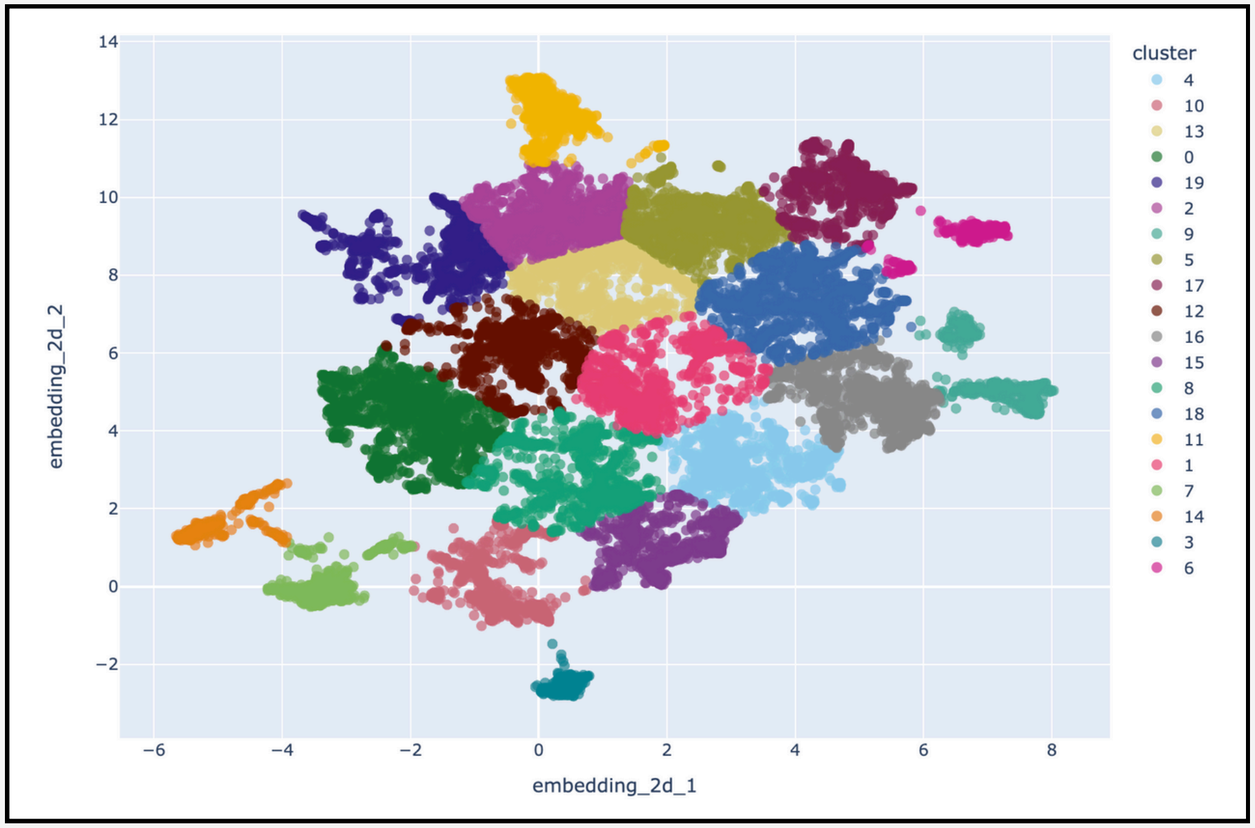
From Text to Embeddings

SentenceTransformer’s all-MiniLM-L6-v2



Cluster Analysis

SentenceTransformer's all-MiniLM-L6-v2



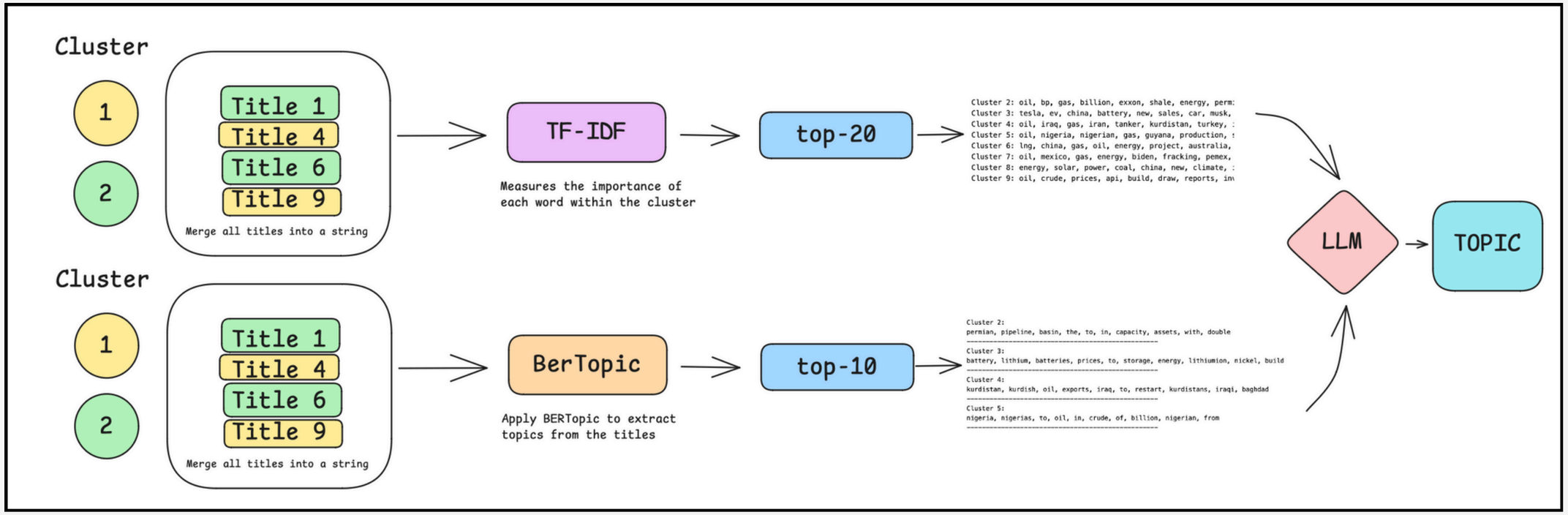
k	Silhouette Score ↑	Calinski–Harabasz Index ↑	Davies–Bouldin Index ↓
24	**0.4227**	**14211.80**	**0.8121**
20	0.4053	13574.60	0.8472

- k = 24 clearly optimizes all metrics (highest silhouette and CH, lowest DB), indicating the most compact and well-separated clusters.
- k = 20 is an attractive trade-off:
  - ***Silhouette*** remains high at 0.405 (only a ~4% drop),
  - ***Calinski–Harabasz*** is strong at 13 574.6,
  - ***Davies–Bouldin*** is still low at 0.847.

Choosing 20 clusters gives nearly the same clustering quality as 24, while reducing complexity and making the results easier to interpret.

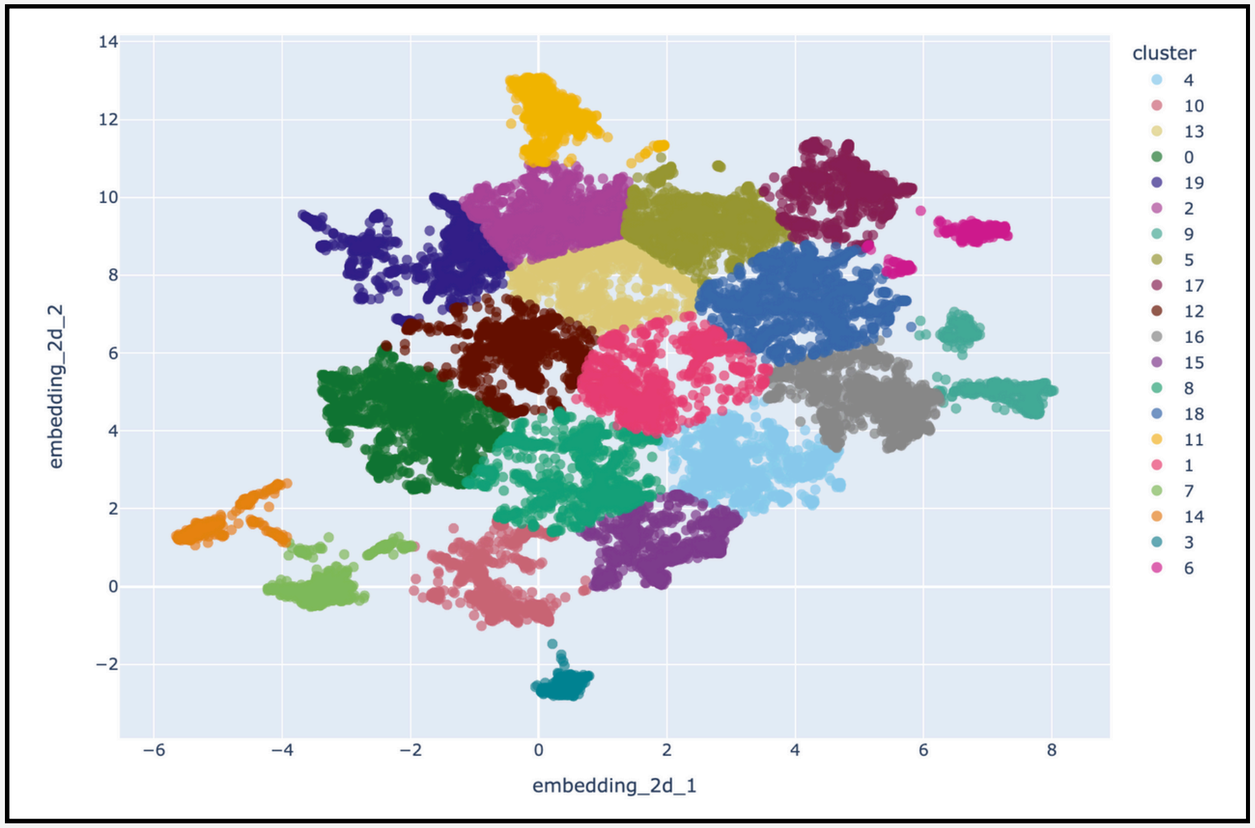
Cluster Analysis-Topic

SentenceTransformer's all-MiniLM-L6-v2



Cluster Analysis

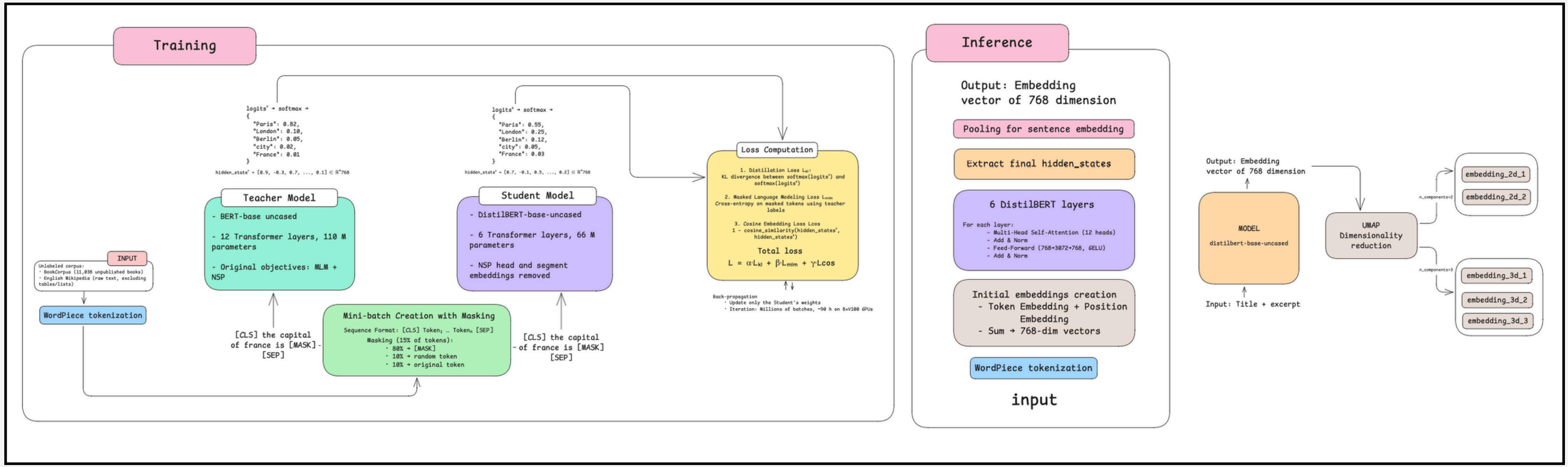
SentenceTransformer's all-MiniLM-L6-v2



Cluster	Topic Definition
0	Russian oil and gas sector including Gazprom, Nord Stream pipeline, EU sanctions, export dynamics, pricing and deals amid Ukraine conflict.
1	Major energy companies (Shell, Eni, Glencore, Chevron) operations in Africa, including oil, gas, coal production, assets, and legal matters.
2	Large oil and gas producers like BP, Exxon, Chevron active in Permian Basin shale production, earnings, and asset management.
3	Electric vehicles and battery market developments focusing on Tesla, lithium, EV sales in China and global markets.
4	Oil exports and pipeline security in Iraq, Kurdistan, and surrounding regions with geopolitical tensions involving Iran, Turkey, and ISIS.
5	Nigeria's oil and gas production including offshore fields, pipelines, OPEC participation, and regional output dynamics.
6	LNG trade and energy projects involving Qatar, China, Australia, coal markets, and international supply contracts.
7	Mexico's oil and gas sector including Pemex operations, US-Mexico energy relations, fracking, drilling bans, and regulatory policies.
8	Global energy transition issues, solar power, coal use, climate change impacts, emissions, and fossil fuel demand worldwide.
9	US oil market inventory dynamics, crude oil price fluctuations, API reports, gasoline stock movements, and supply surprises.
10	Saudi Arabia and OPEC production, Aramco's IPO and billion-dollar valuations, UAE and Kuwait quota deals and market impact.
11	China and India's crude oil demand and imports, Russian energy relations, refinery operations, and trade under sanctions.
12	Venezuela's PDVSA oil production, sanctions impact, Maduro government, Citgo operations, and international oil deals.
13	Global oil price trends including gasoline demand, OPEC production forecasts, EIA and IEA reports, and supply-demand analysis.
14	European energy sector with emphasis on UK, Germany, nuclear and wind power, offshore wind farms, natural gas, and coal markets.
15	Canadian oil production and pipeline infrastructure (Keystone, Trans Mountain), trade relations with US, and energy export issues.
16	Iranian and Russian oil and gas exports, OPEC quotas, sanctions effects, nuclear talks, and production cuts with geopolitical context.
17	Libyan oil production, export challenges due to protests, port closures, force majeure declarations, and field operations.
18	Petrobras and South American energy sector including Brazil, Argentina, shale plays like Vaca Muerta, and regional oil deals.
19	North Sea oil and gas industry, UK and Norway energy production, tax policies, Arctic drilling, Equinor activities, and windfall taxes.

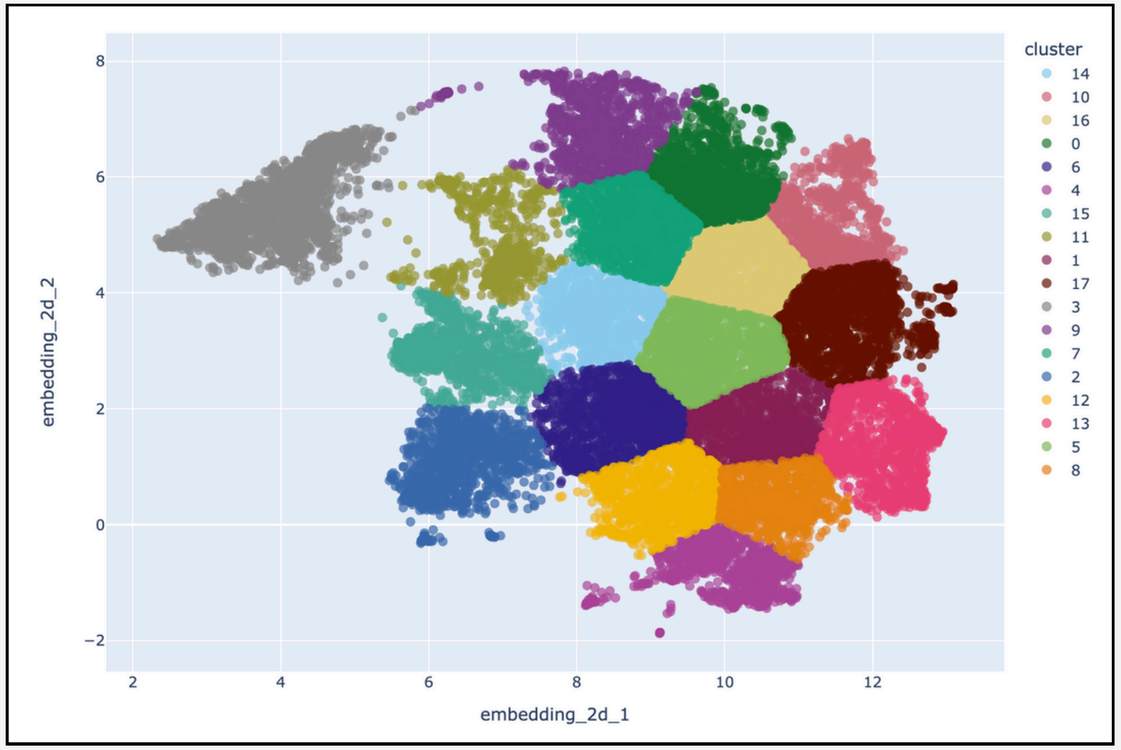
From Text to Embeddings

Distilbert-base-uncased



Cluster Analysis

Distilbert-base-uncased



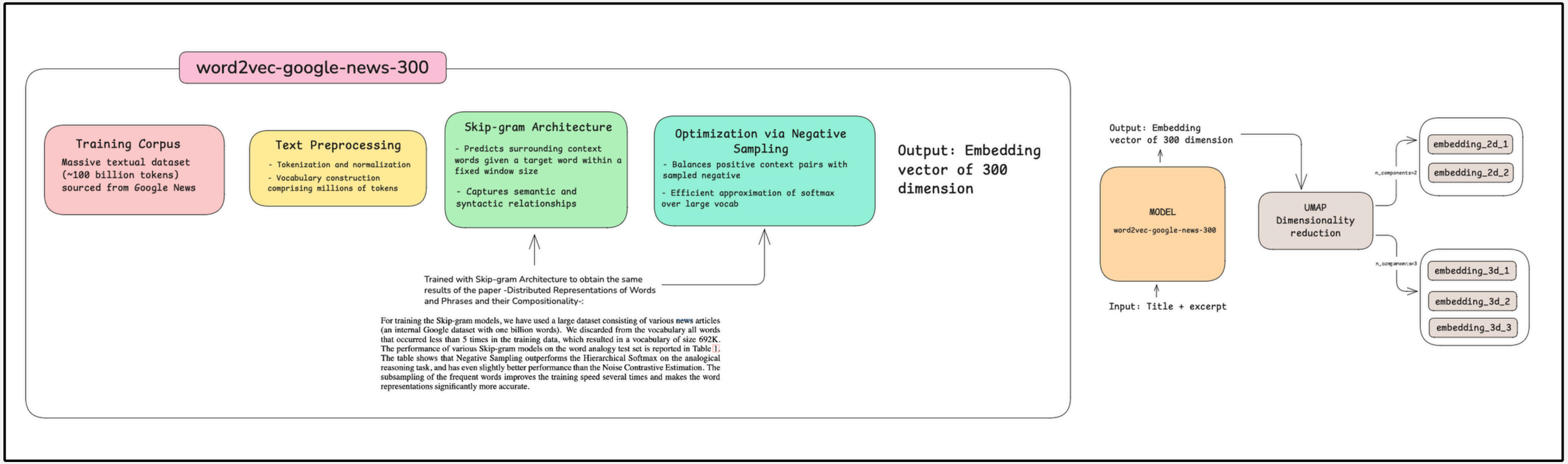
k	Silhouette Score ↑	Calinski–Harabasz Index ↑	Davies–Bouldin Index ↓
18	0.4002	26,416.59	0.7619

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)

To extract topics, the same procedure previously described was applied

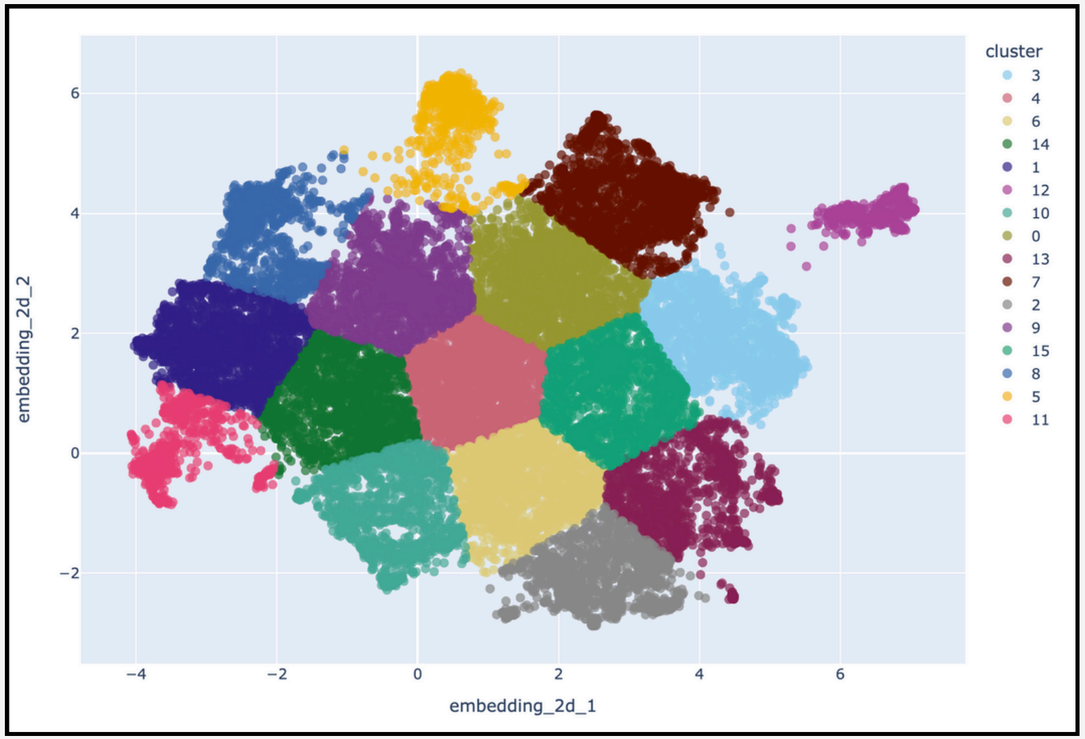
From Text to Embeddings

word2vec-google-news-300



Cluster Analysis

word2vec-google-news-300



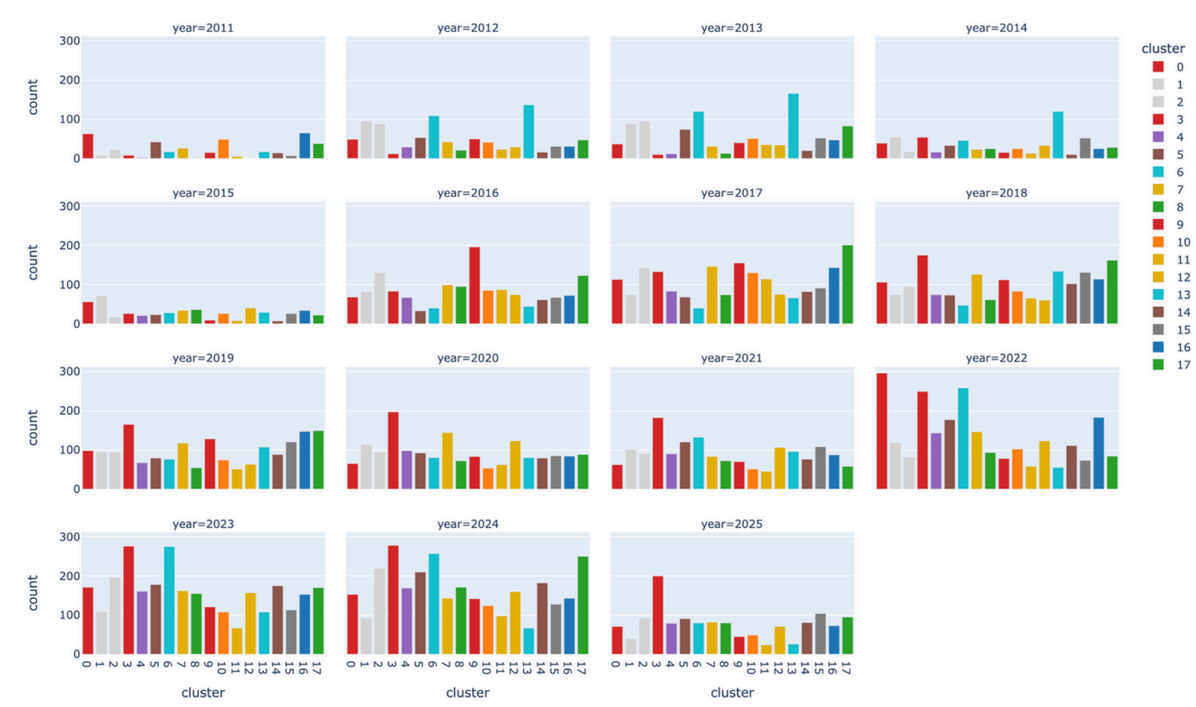
k	Silhouette Score ↑	Calinski–Harabasz Index ↑	Davies–Bouldin Index ↓
16	0.4053	24337.18	0.7241

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)

To extract topics, the same procedure previously described was applied

Comparative Analysis of Clustered Topics

Distilbert-base-uncased



Only Distilbert-base-uncased

- Cluster 10: Upstream partnerships
- Cluster 14: China’s crude trade flows

Geopolitics & Sanctions Cluster: 0, 3, 9

OPEC Production Cluster: 7, 11, 12

Global Crude Trade Cluster: 16

Clean Energy & EVs Cluster: 6, 13

Financial Results Cluster: 8, 17

word2vec-google-news-300

Geopolitics & Sanctions Cluster: 2,6

OPEC Production Cluster: 7

Global Crude Trade Cluster: 0, 3

Clean Energy & EVs Cluster: 1, 8

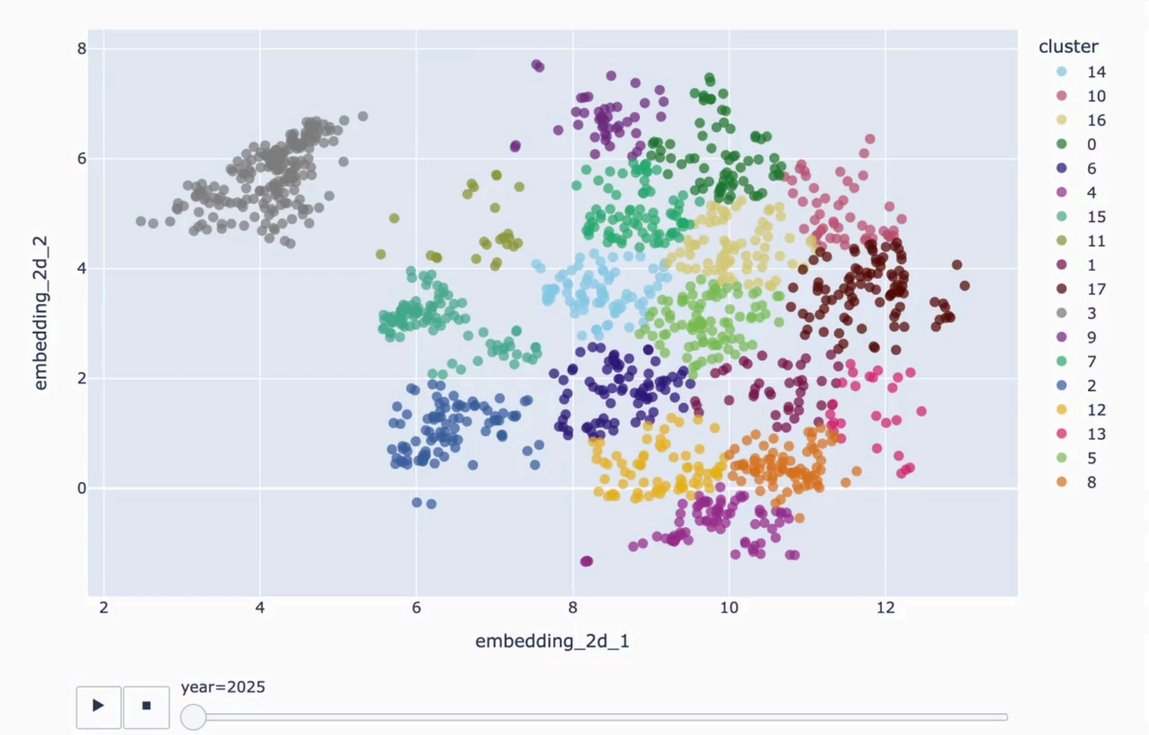
Financial Results Cluster: 9, 5

Only word2vec-google-news-300

- Cluster 11: Nuclear & Power Plants
- Cluster 13: Refinery Disruptions
- Cluster 14: Energy Policy & Taxation

Topics over Years

Distilbert-base-uncased



Video

Cluster 3 – Oil and Gas Market Fluctuations under Sanctions

Temporal Peak: 2022–2023

This cluster experienced a pronounced surge during the Russia-Ukraine war, particularly in response to the sanctions imposed on Russian oil and gas exports by Western nations. The EU's phased embargo, alongside U.S. and U.K. restrictions, significantly disrupted global supply chains.

Cluster 9 – Attacks on Oil Infrastructure

Temporal Growth: Post-2020

This cluster became increasingly salient after 2020, driven by a series of attacks targeting critical oil infrastructure in the Middle East and North Africa. Notably:

- Houthi drone and missile strikes on Saudi oil facilities (Ras Tanura, Jeddah, 2021–2022),
- Continued instability and militia activity around Libyan export terminals,
- Broader concerns about energy infrastructure security, including cyber vulnerabilities.

Cluster 0 – Geopolitical Energy Dynamics and Sanctions

Temporal Explosion: Post-2016, peaking in 2022–2023

Cluster 0 shows a clear expansion beginning in 2016, coinciding with the U.S. withdrawal from the Iran nuclear deal (JCPOA) and the reinstatement of sanctions on Iranian oil exports. This marked a shift toward greater energy weaponization in foreign policy. The cluster intensified further as Russia's global positioning became increasingly adversarial, culminating in the 2022 invasion of Ukraine.

Topics over Years

word2vec-google-news-300

Video

Cluster 1 – Energy Transition and Renewables

Temporal Surges: 2012, 2023–2024

This cluster exhibits two major spikes of media attention. The first surge in 2012 reflects growing global interest in renewable energy technologies in the wake of the Fukushima disaster (2011) and early policy signals from Germany's Energiewende and other EU climate commitments. A second, more sustained expansion occurs in 2023–2024, coinciding with aggressive decarbonization goals, record investments in wind and solar, and the mainstreaming of ESG frameworks in energy finance.

Cluster 6 – Attacks on Oil Infrastructure

Temporal Explosion: 2022

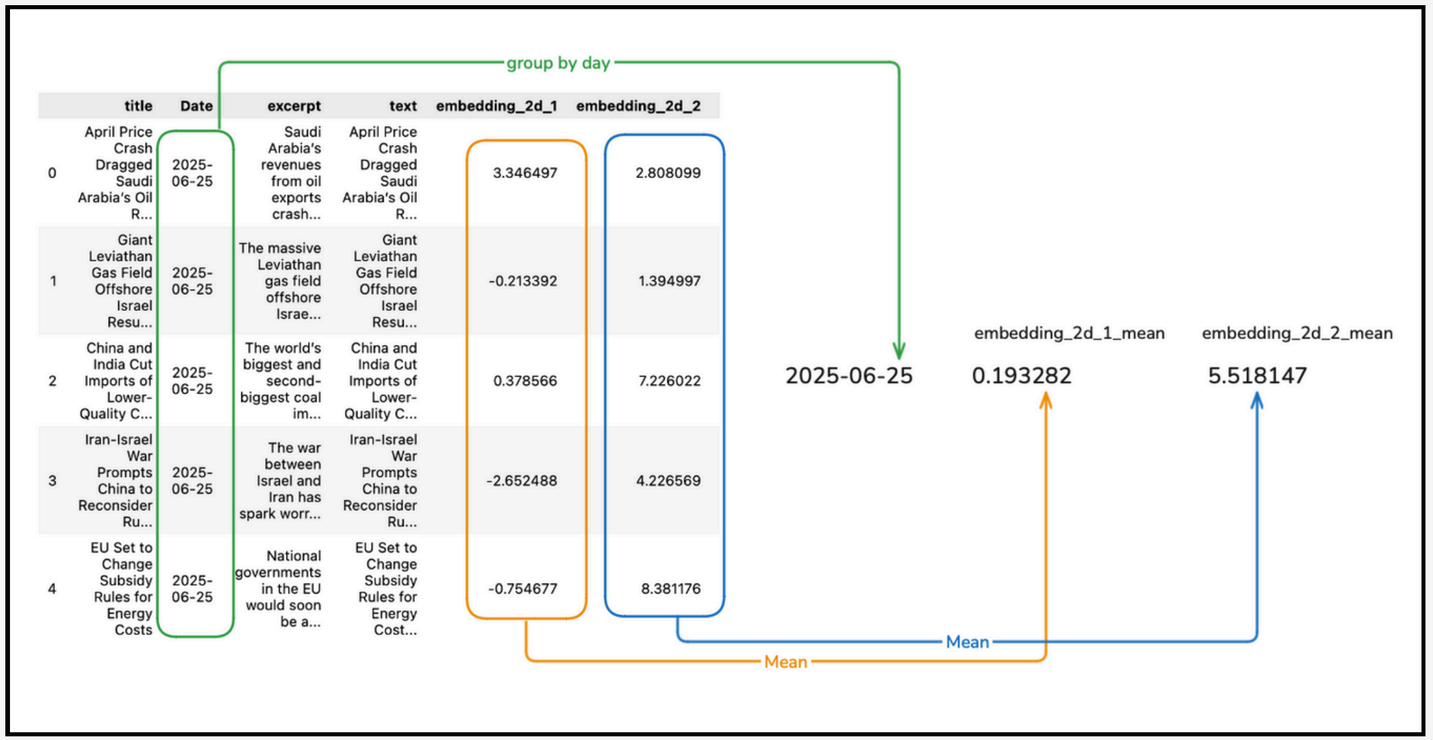
Cluster 6 displays a dramatic increase in 2022, correlating with the geopolitical fallout of Russia's invasion of Ukraine. As Western nations imposed extensive sanctions on Russian energy exports, media coverage on pipeline shifts, LNG redirection, and the weaponization of energy soared. This surge marks a clear pivot in narrative, from isolated regional restrictions to a system-level reordering of global energy alliances.

Cluster 3 – Crude Export Metrics and Trade Flows

Structural Growth: Post-2016, peaking in 2024

Cluster 3 shows a consistent upward trend beginning in 2016, reflecting intensified interest in OPEC decisions, Chinese and Indian import volumes, and global oil demand patterns. By 2024, the cluster reaches peak prominence, likely fueled by rising Asian demand, shifting supply chains, and the reevaluation of trade flows in light of geopolitical fragmentation and the energy transition.

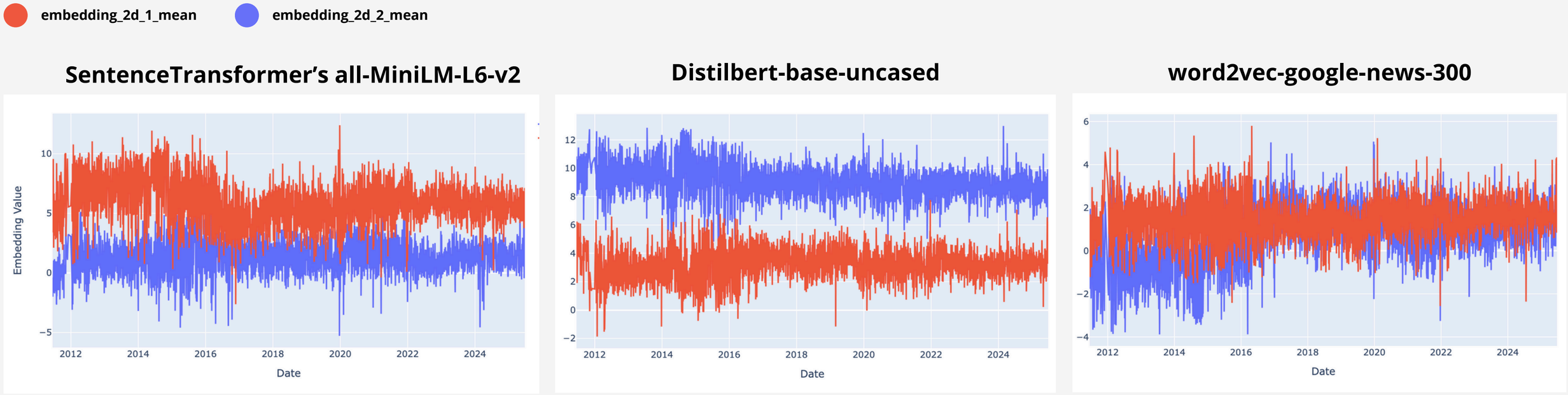
Feature Creation: Daily Mean



Feature Cereation

- **Semantic aggregation:** Averaging provides a compact summary of the collective thematic content of all articles published on a given day, capturing shared or dominant topics while smoothing out individual variability.
- **Temporal alignment:** This aggregation ensures consistency with the daily resolution of the target variable (Brent crude oil price), allowing for the construction of input features that match the granularity required for forecasting.

Feature Creation: Daily Mean

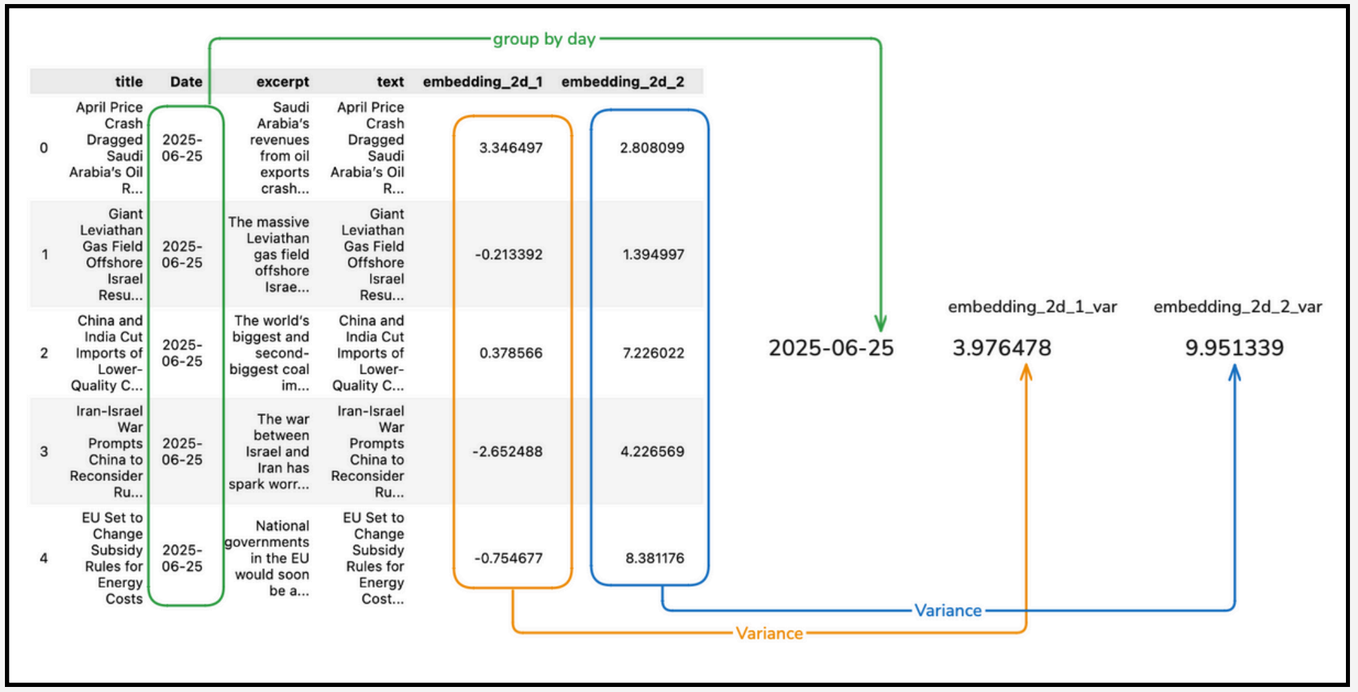


- Displays the largest temporal variation, particularly in the first component, indicating strong responsiveness to changes in daily news semantics.
- As a transformer model fine-tuned for semantic similarity, it effectively captures contextual nuances and emerging topical shifts.

- Exhibits a more stable temporal profile, with reduced variance and a consistently elevated second component, possibly indicating directional bias.
- Being not fine-tuned for sentence-level tasks, it provides robust but less reactive embeddings, suitable for applications requiring smoother dynamics.

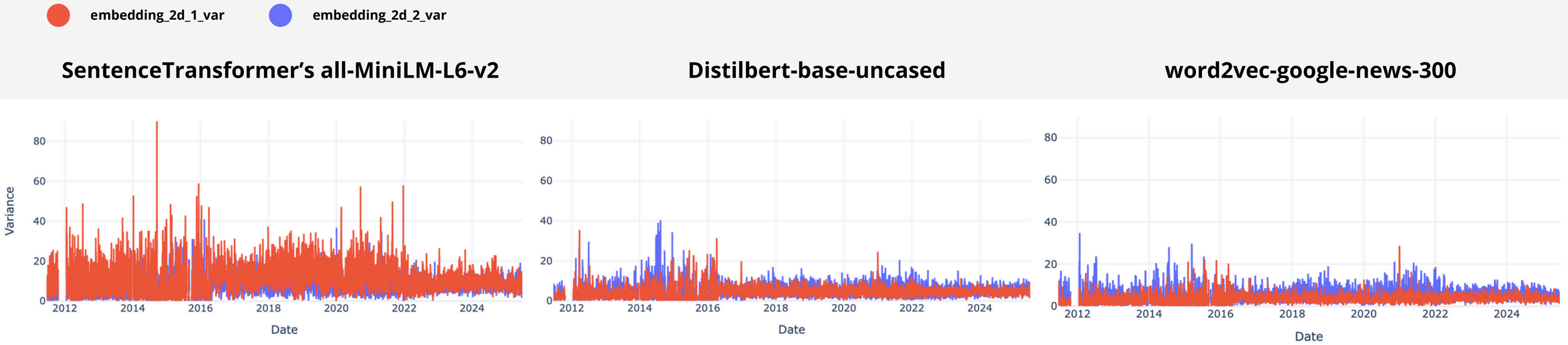
- Shows the lowest variance and narrowest value range, with embeddings appearing noisy and semantically flattened.
- As a static word embedding model, it lacks contextual sensitivity, making it less suitable for detecting daily thematic shifts in the news corpus.

Feature Creation: Daily Dispersion



- The daily embedding variance measures how semantically diverse the news articles are on a given day: higher variance indicates greater heterogeneity in content.
- This feature captures the thematic fragmentation of daily discourse and can signal complex or uncertain news cycles, making it useful for analyzing informational contexts that may affect markets.

Feature Creation: Daily Dispersion



- Exhibits the highest and most persistent daily variance, especially in the first component, indicating strong responsiveness to semantic fragmentation within the news corpus
- Its fine-tuning for semantic similarity allows it to capture subtle contextual differences, making it well-suited for detecting days with diverse and complex topic structures

- Shows moderate and less volatile dispersion, with occasional spikes, particularly in the second component, reflecting a more compressed semantic representation
- Due to the absence of sentence-level fine-tuning, the model yields stable but less context-sensitive embeddings, favoring generalization over granularity

- Records the lowest daily variance across both components, suggesting limited capacity to represent thematic heterogeneity within daily news cycles
- As a static embedding model, Word2Vec lacks contextual adaptation, resulting in flattened semantic representations and reduced expressiveness

Feature Creation: Cosine Similarity

This metric serves as a baseline indicator of semantic stability over time.

embedding_mean (original not reduced)

et = array([-1.89991733e-02, 3.86615365e-02, 4.53784511e-02, 1.30004865e-02,
3.10601213e-02, -1.52258242e-02, -3.10302773e-02, -4.61249069e-03,
25/06/2025 -5.19865827e-02, -8.33604943e-03, -3.24329236e-02, -6.43527135e-03,
-3.52412832e-02, 4.12157820e-02, -4.71333313e-04, 1.30559780e-02,
-2.45269388e-04, -1.26744513e-02, -5.69443107e-02, -3.24946823e-02,
2.38695038e-02, -1.37616716e-02, 3.55496382e-02, -5.59044583e-03,
2.70149750e-03, 1.16264826e-02, -1.39570471e-02, len(e1) = 384

et_1=array([-4.07353371e-02, -2.32926295e-02, 6.49505155e-02, 4.22269152e-02,
1.92683885e-02, -1.43993075e-02, -5.71766602e-02, 1.79065559e-02,
-2.92232453e-02, -1.65051063e-02, -4.36173615e-02, 1.92352732e-02,
24/06/2025 -4.09183169e-02, -9.78721956e-03, -1.68768062e-02, 1.35257378e-02,
-1.34937732e-03, -2.26622190e-02, -2.55685042e-02, -2.95962981e-02,
1.93371243e-02, 6.48071390e-03, 6.19906335e-03, -1.56951187e-02,
3.93466118e-02, 2.64027605e-02, -2.19921782e-02, len(e1) = 384

$$\text{cosine\_sim}_t = \frac{\mathbf{e}_t \cdot \mathbf{e}_{t-1}}{\|\mathbf{e}_t\| \|\mathbf{e}_{t-1}\|} = 0.730947$$

Values close to 1 indicate strong similarity and thematic continuity
Values close to 0 denote orthogonal or highly divergent discourse content

Semantic Change Metrics Based on Cosine Similarity

Semantic Drift

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

High value (≈ 1):

Significant semantic change from the previous day

Low value (≈ 0):

Minimal semantic change; themes are persistent

Drift Velocity

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

High value (≈ 1):

Semantic content has been changing consistently over several days

Low value (≈ 0):

→ Stable semantic environment across recent days.
→ Reflects narrative consistency.

Drift Velocity Difference

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

High value (≈ 1):

→ The speed of semantic change is increasing
→ Discourse is becoming more volatile or dynamic

Low value (≈ 0):

The rate of change is slowing down

Drift Acceleration

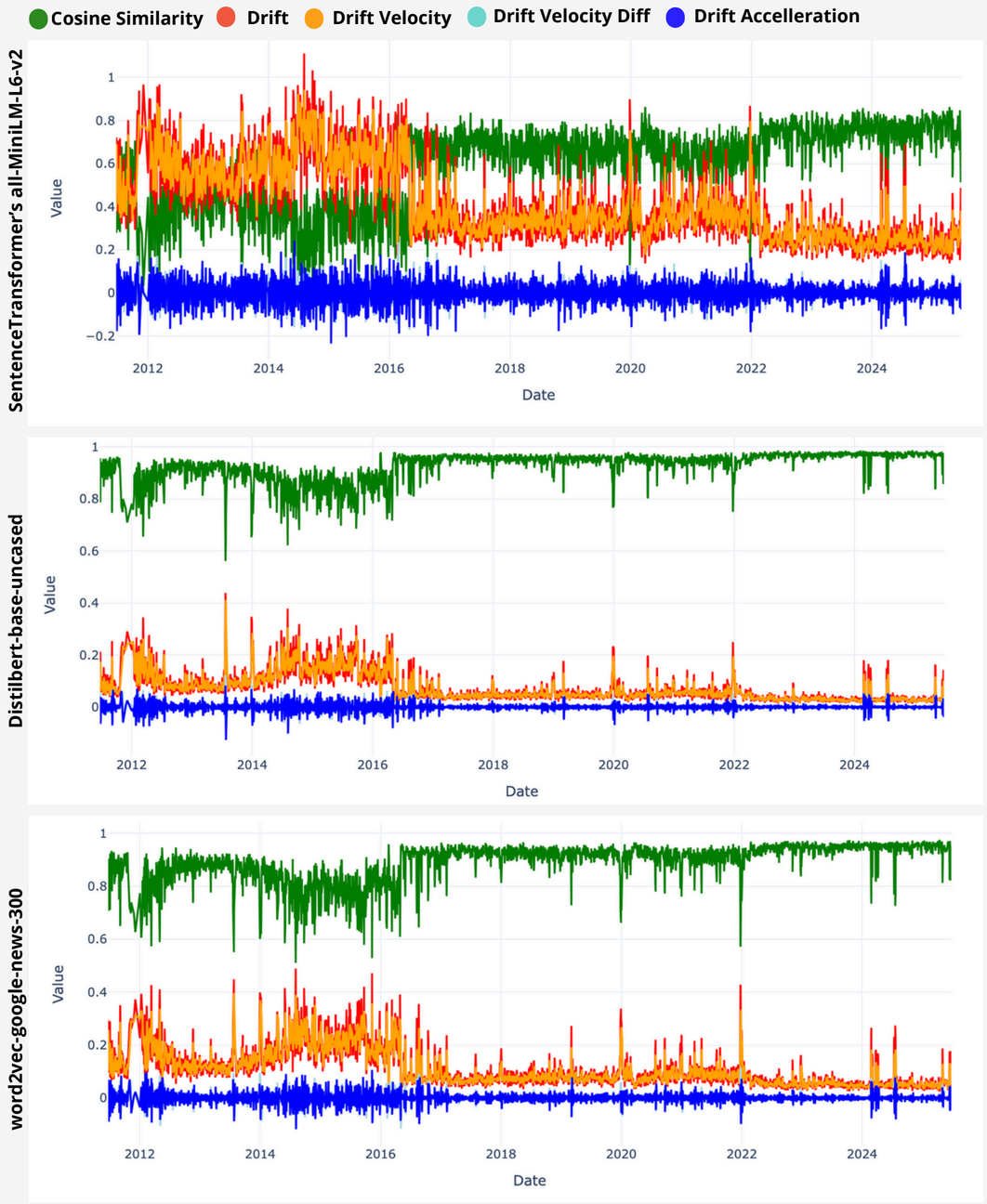
$$a_t = v_t - 2v_{t-1} + v_{t-2}$$

High value (≈ 1):

→ Sudden onset of rapid semantic change
→ May capture shocks, disruptions, or breaking news

Low value (≈ 0):

Fast deceleration in semantic change



Semantic Change Metrics Based on Cosine Similarity

- **MiniLM (SentenceTransformer's all-MiniLM-L6-v2)**
  - Shows high drift and low cosine similarity especially before 2018, indicating strong responsiveness to daily semantic changes.
  - Displays large oscillations in drift velocity and acceleration, capturing rich dynamics and abrupt shifts in discourse.
  - Best suited for detecting subtle topic shifts and fragmentation over time.
- **DistilBERT-base-uncased**
  - Exhibits moderate drift and consistently high cosine similarity (above 0.8 post-2016), reflecting semantic stability with some sensitivity.
  - Drift-related metrics are smoother than MiniLM but still informative.
  - Balances semantic coherence and change detection reasonably well.
- **word2vec-google-news-300**
  - Maintains very high cosine similarity throughout, often exceeding 0.9, and low drift, suggesting minimal variation in semantic representation across days.
  - Drift velocity and acceleration are low and stable, indicating limited capacity to reflect evolving discourse.
  - Performs poorly for detecting semantic novelty or fragmentation due to its static and averaged nature.

Feature Creation: Sentiment Analysis of Oil News Titles

FinBert

financial sentiment analysis

Task-specific Fine-tuning

FinBERT is fine-tuned on a labeled dataset for sentiment classification: Financial Phrase Bank , which contains thousands of financial sentences labeled as positive, neutral, or negative

This step involves:
- Adding a classification head (typically a softmax layer for 3-class classification)
- Training the model to minimize categorical cross-entropy loss

Domain Adaptation

FinBERT is further pre-trained on financial texts:

- Thomson Reuters TRC2 Corpus, a collection of financial news articles and documents

This step enables the model to learn financial terminology, phrasing, and syntactic patterns

General Pre-training

Inherited from BERT:
- Trained on English Wikipedia and BookCorpus
- Provides general language understanding

CrudeBert

Fine-Tuning Procedure

- Model architecture: BERT-base with a classification head (softmax output)
- Input: tokenized headlines (avg. 10.4 tokens per example)
- Loss function: weighted cross-entropy to account for class imbalance
- Train/val/test split: 60/20/20, stratified by label

Although the number of neutral examples was small, they were kept in the dataset to expose the model to non-informative cases and reduce overfitting to binary extremes.

The result was a dataset of approximately 30,000 labeled headlines: ~15,000 positive ~14,000 negative ~500 neutral

Automatic Labeling Based on Economic Logic

Positive (+1) if the event was expected to cause higher prices (supply disruptions, increasing demand, rising imports)

Negative (-1) if the event was expected to cause lower prices (increased production, export growth, demand declines)

Neutral (0) if the event was not expected to impact prices significantly

Keyword-based topic matching
Polarity detection using directional terms
A logic layer encoding Adam Smith's price theory:
↓ supply or ↑ demand → price ↑ → positive
↑ supply or ↓ demand → price ↓ → negative

Creation of a Silver-Standard Dataset

Collected approximately 46,000 crude oil-related news headlines (2000-2021) from the RavenPack Realtime News Discovery platform

Identifyied frequent topics across these headlines, such as:

Accidents Oil discoveries Import/export changes Supply/demand fluctuations Pipeline disruptions

apply to titles of oil news dataset

title	Date
April Price Crash Dragged Saudi Arabia's Oil R...	2025-06-25
Giant Leviathan Gas Field Offshore Israel Resu...	2025-06-25
China and India Cut Imports of Lower-Quality C...	2025-06-25
Iran-Israel War Prompts China to Reconsider Ru...	2025-06-25
EU Set to Change Subsidy Rules for Energy Costs	2025-06-25

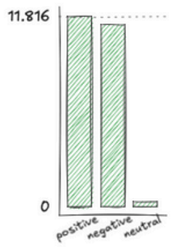
output

sentiment	
positive	1
negative	-1
negative	-1
positive	1
negative	-1

mapping to scalar

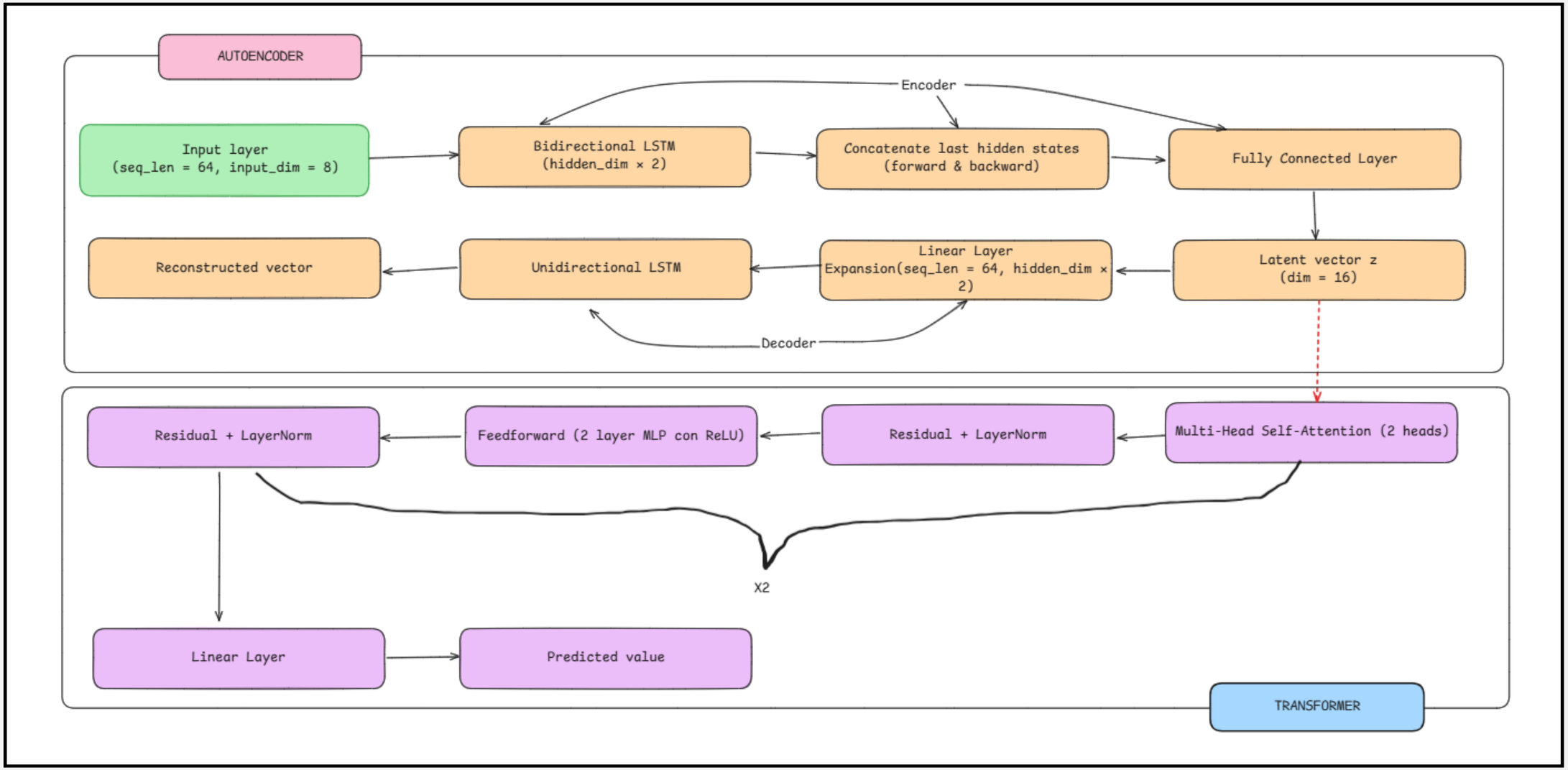
avg daily - 0,2

The average sentiment score on 25/06/2025 was -0.2, reflecting a slightly bearish tone in news coverage. Such sentiment is typically associated with expectations of a modest decline in crude oil prices

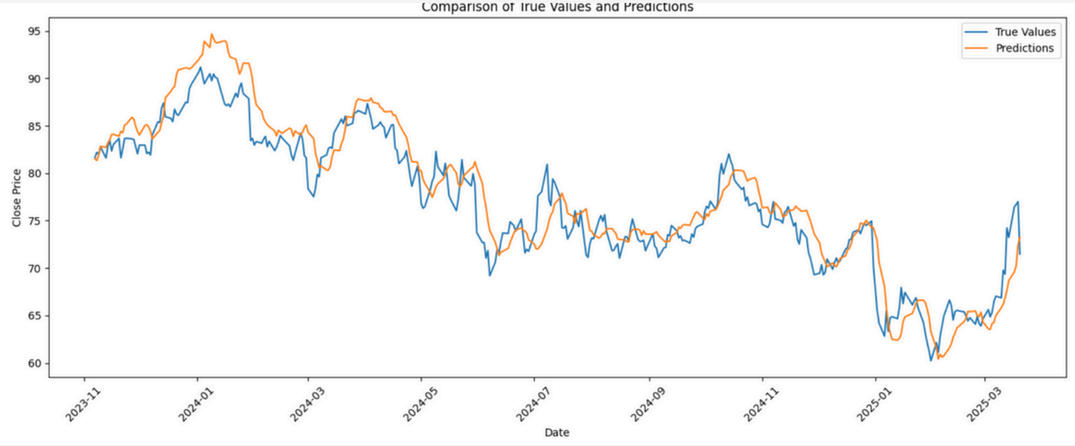


Count of positive, negative or neutral articles in oil news dataset

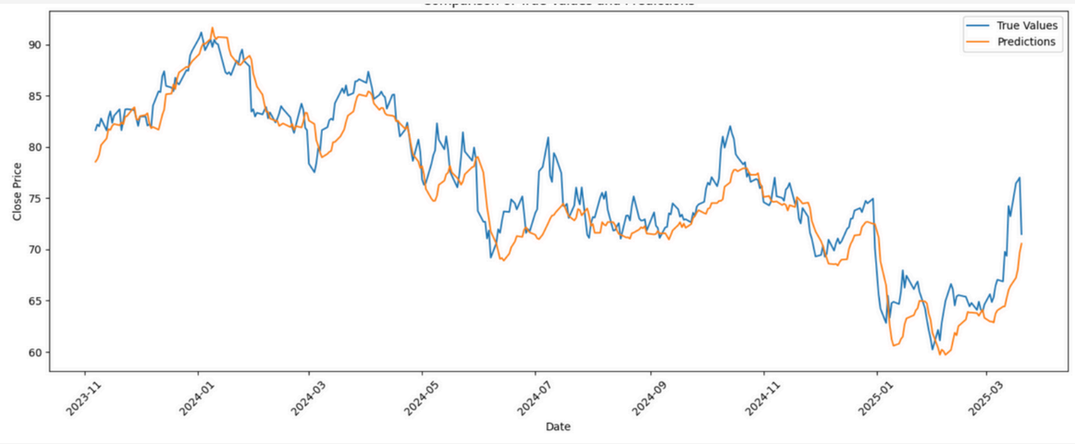
Transformer-based Stock Price Prediction with Autoencoder Denoising



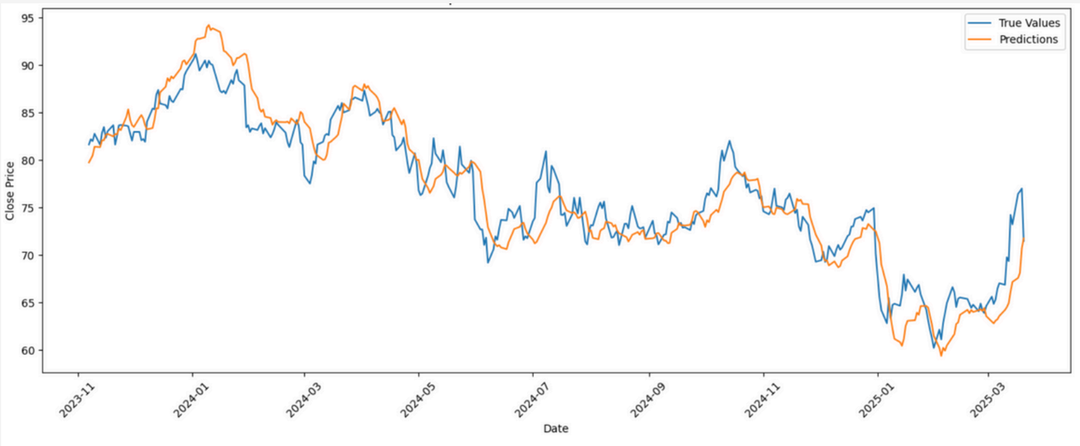
SentenceTransformer's all-MiniLM-L6-v2



Distilbert-base-uncased



word2vec-google-news-300

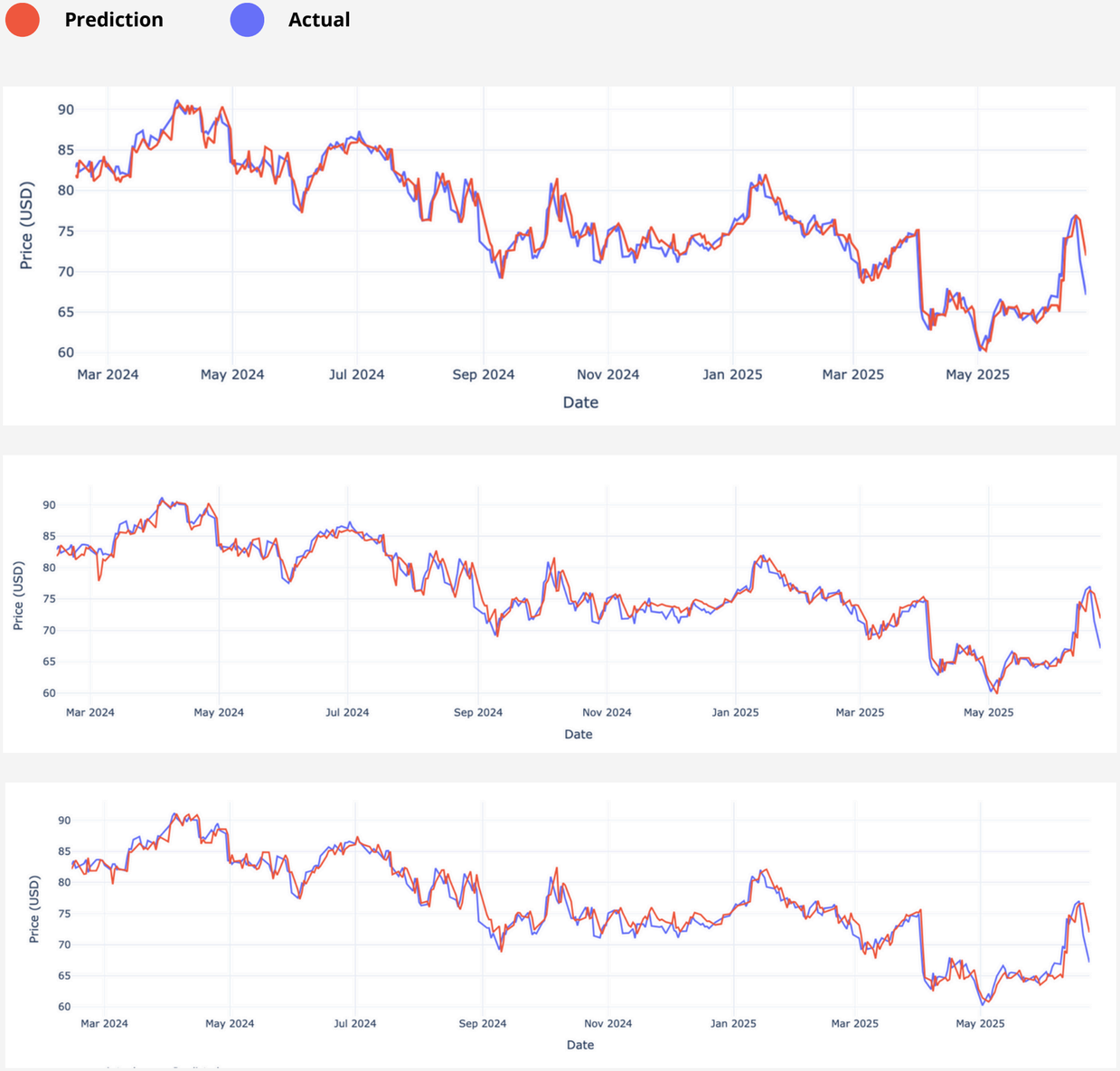


Model evaluation

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

- MiniLM embeddings out-performed both Word2Vec and DistilBERT across most evaluation metrics

Random Forest




```
Feature Used = [
    'Price','SENT_1','SENT_2','cosine_sim','drift',
    'drift_velocity','drift_velocity_diff','drift_acceleration',
    'var_2d_1','var_2d_2','sentiment'
]
```

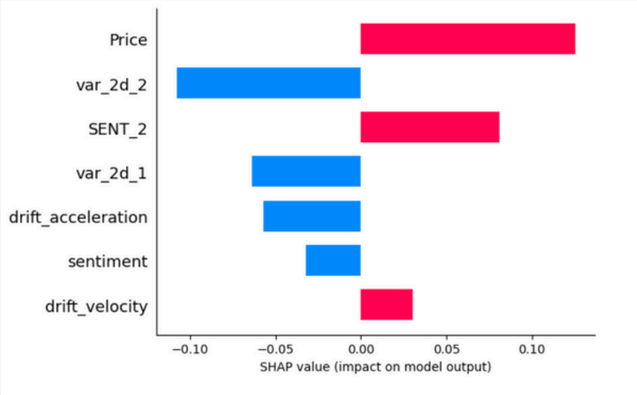
Model Evaluation Metrics

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

Understanding Why: Shap values

The SHAP values represent the contribution of each feature from June 17 in predicting the Brent crude oil price for June 18

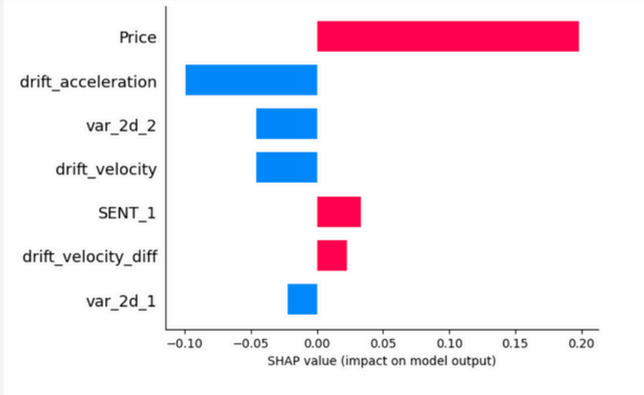
SentenceTransformer's all-MiniLM-L6-v2



Price BZ=F 76.449997
Target 76.699997
Predicted 76.930700

Most prevalent Cluster is **13**: Global oil price trends including gasoline demand, OPEC production forecasts, EIA and IEA reports, and supply-demand analysis

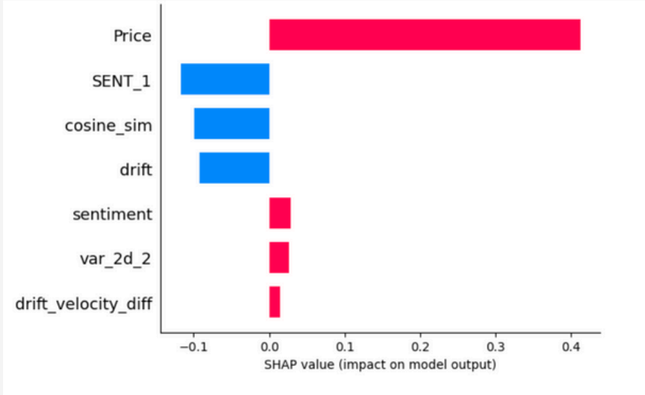
Distilbert-base-uncased



Price BZ=F 76.449997
Target 76.699997
Predicted 76.381650

Most prevalent Cluster is 9: Attacks on oil infrastructure (Libya tankers, pipelines, Houthi/ISIS threats, Saudi/Nigerian fields)

word2vec-google-news-300



Price BZ=F 76.449997
Target 76.699997
Predicted 76.520800

Most prevalent Cluster is 7: Oil price dynamics and demand analysis (gasoline trends, supply/demand balance, OPEC signals)

XGBoost

- 90% train, 10% test
- Number of estimators (trees): 200
- Maximum tree depth: 6
- Learning rate (η): 0.07
- Subsample ratio (for rows): 0.90
- Column sample ratio per tree (colsample_bytree): 0.90
- Tree construction method: histogram-based (tree_method = "hist")

```
Feature Used = [
    'Price','SENT_1','SENT_2','cosine_sim','drift',
    'drift_velocity','drift_velocity_diff','drift_acceleration',
    'var_2d_1','var_2d_2','sentiment'
]
```

Model Evaluation Metrics

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

Conclusion

Semantic features improve prediction

Sentiment, drift, and variance from news capture signals missed by historical prices, especially during shocks or geopolitical events.

News and data work better together

Embeddings turn qualitative narratives into predictive features that enhance model foresight

Future Developments

Use more News Sources

The embedding pipeline could be expanded to incorporate cross-lingual news sources, enabling a global perspective on oil market sentiment

Multimodal data inclusion

The inclusion of multimodal data, such as macroeconomic indicators, energy inventory reports, could further improve forecasting performance