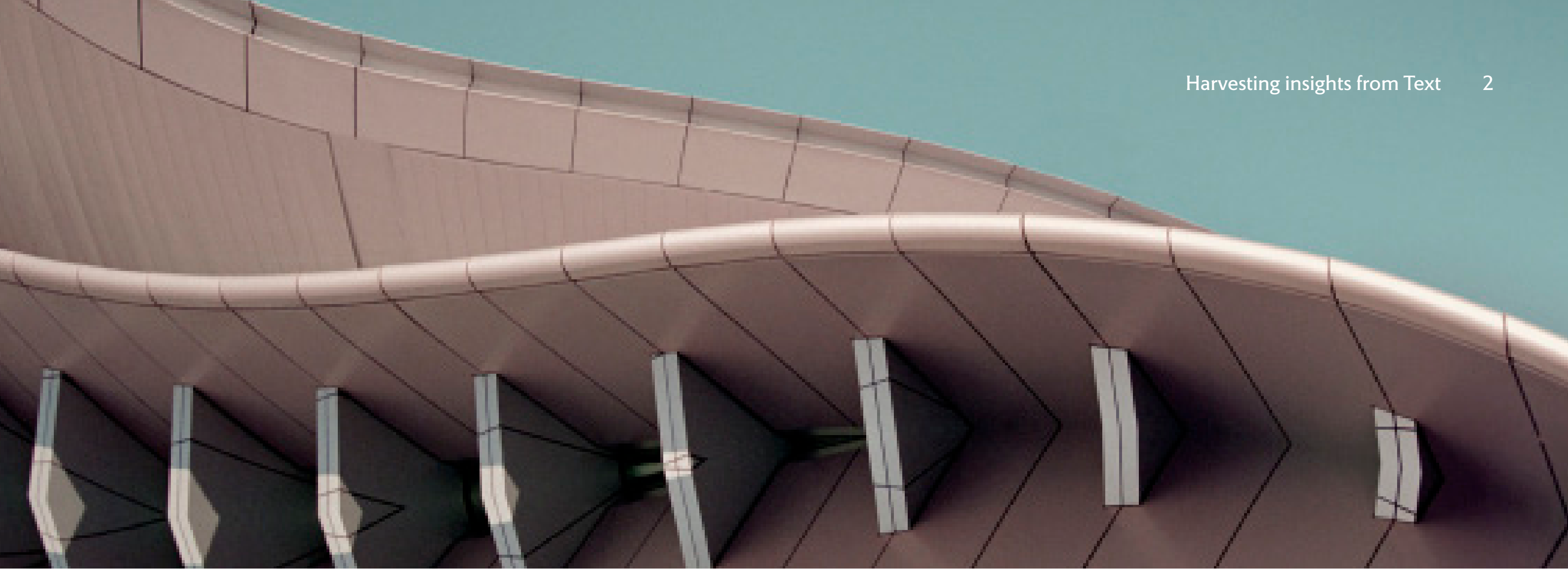
Harvesting insights from Text

Deriving greater value from unstructured text and analytics for quants and data scientists

LSEG DATA & ANALYTICS



Our text and analytics offerings for AI and data science purposes

In today's data-driven world, unstructured text represents a vast and valuable resource, encompassing everything from news articles and research papers to transcripts and financial filings. Harnessing this wealth of information requires sophisticated analytics to transform raw text into actionable insights. LSEG offers an unmatched breadth and depth of unstructured financial content, produced to meet the diverse needs of our customers.

Beyond this, our data science and quantitative teams have developed advanced analytics tools. As a result, customers are able to extract meaningful signals from unstructured text, further equipping them to derive meaning from the unstructured text available, as well as supporting them with predicting future outcomes. In addition to our proprietary Machine Readable News feeds, News Analytics and StarMine models, we also deliver tools from our close partners, Exponential Technology and MarketPsych, to create synergies with our other data available.

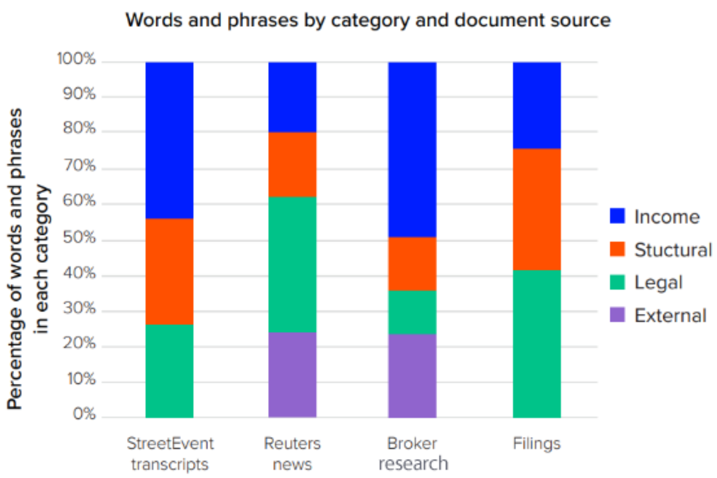
These tools enable teams to leverage the expanding number of use cases within text analytics, and provide critical insights to unlock competitive advantage. Common uses range from event-based trading ahead of or on the back of economic releases and exclusive M&A newsbreaks, to quantitative investment decision-making using sentiment and buzz metrics derived from this vast textual corpus.

Whether you need the raw data or the refined signals our teams create, our comprehensive solutions are tailored to help you navigate and capitalise on the complexities of unstructured text. Our offerings can also support LLM usage, subject to terms.

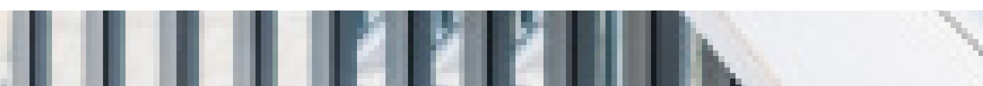
Contents

1. Unstructured Text – Harnessing raw text

Today, it is estimated that at least 80% of data and content available is unstructured. This has led to an increased need for financial services firms around the world to use artificial intelligence (AI), machine learning (ML) and natural language processing (NLP) to parse textual data hundreds of thousands of times faster and more accurately than humans for financial analysis and insight generation. There is greater demand than ever to show the value of unstructured data by making insights consumable from News, Research, Transcripts and Filings content. When various unstructured text sources are used in combination the value of these content sets is further enhanced.



Source: Starmine Credit Risk Model – Chart demonstrating the value of different text sources – Transcripts, News, Research and Filings – to provide unique insights into companies' financial health. A comparison of dictionary word overlap among these sources and a breakdown of key financial language categories (Income, Structural, Legal and External) is shown across different document types. As a result, using multiple sources enhances the depth of financial analysis.



News

Overview: Real-time News includes feed and archive access to exclusive Reuters news and third-party newswires. Both streaming feeds and deep historical archives can be delivered directly to applications, enabling algorithms to capitalise on the power of news. Plug-and-play search and symbology APIs strip out the noise and help you get to what you need fast. These are available in a range of latencies (low latency, hourly, intra-day, end-of day), machine-ready formats and delivery channels.

Content:
- LSEG's programmatic news feed solution is the only one in the industry that's powered by award-winning, trusted Reuters News
- Renowned for breaking news and market-moving scoops, Reuters provides timely and trusted national and international news to financial professionals reporting from 200 locations, with 2,600 journalists
- Broad asset class coverage including equities, commodities, foreign exchange, fixed income, futures and macroeconomic news
- Coverage of 50,000 publicly listed companies in English across the globe
- 2.6 million+ stories per year by Reuters (~7,000 stories per day) in 12 languages
- 130+ third-party newswires and exchanges (including corporate and regulatory announcements) resulting in a total of approx. 30,000 unique stories per day on average
- 29+ year News Archive point-in-time recording of streamed News

Metadata:
- Vast array of metadata applied, that is well-suited to train machine learning models
- Each story is scored for author sentiment, relevance, significance and novelty
- Textual news includes both topic and entity tagging leveraging RICs and PermIDs
- 2,600+ topic codes and other metadata used
- Topic tagging leverages machine learning using 2,000+ trained, automated topic models that are run against each story to accurately apply appropriate topic tags
- Industry classification runs 5 levels deep and includes 1,100+ tags

Research

Overview: Real-time Research API supports a single integrated API for quantitative research and investment management customers into their entitled research collection, providing the ability to search and filter across brokers using common metadata from over 1,200 contributors.

Content:
Extensive and unique research history, dating as far back as 1982.
- Breadth of contributors: 500+ premium brokers available upon approval, 700+ active open access research contributors.
- LSEG Real-Time Research API provides live and historical data across our corpus of 1200+ contributors using a single access point and data model.
- Standardised delivery, data model and metadata: provides game-changing cost and technical scale to allow data science teams to go from ideation to profit quickly.

Metadata:
- Normalised data model across contributors
- PermIDs for analysts, industries, companies, and contributors
- StarMine author ratings
- 20 report styles
- 100+ report reasons identifying drivers/themes for publication
- Unified data model for historical and live data
- Spans research contributors, including metadata and company identifiers
- Single API for real-time updates
- API and Bulk delivery options to simplify build-and-maintain technology
- Extensive historical data for backtesting
- 50+ metadata fields for filtering and search, including:
  - Organisational, country, and industry PermIDs
  - StarMine analyst ratings
  - Research themes (e.g., target price changes, market share, accounting issues) created by an advanced ML tagging system
  - Report type classifiers for equities, credit, economic, fixed income, and ESG
- Normalised timestamps across all research providers, in UTC

Transcripts	Filings
Overview: LSEG's Event Transcripts and Briefs provide verbatim reports and unbiased summaries of corporate and institutional events. Collectively, they represent the deepest available archive of this content in the industry. Coverage history for both transcripts and briefs goes back to 2001. **Content:** Verbatim transcripts and summaries of over 10,000 global companies: – 15 event types covered including earnings calls, M&A calls, sales calls and analyst meetings. Also includes corporate conference presentations for conferences hosted by the top global investment banks – Data can be embargoed – Over 36,000 transcripts and summaries produced every year – 1,500 companies covered live, streaming event presentations in near real-time (desktop only) – Integrated state-of-the-art speech-to-text technology generating highly accurate transcriptions ready to be manually finalised **Metadata:** – Corporate, institutional and corporate event types – Event title – Document type – Company name and identifier – Sentiment scores (desktop only) – Topic and entity tagging (desktop only) – DealIDTags (desktop only) – Speakers' identification with PermIDs (desktop only) – MP3 and presentation filings – English content	**Overview:** Largest and most consistent source for global public company corporate disclosures. With the largest historical filings library in the market, Filings provides unrivalled depth and access to unique US historical documents, including annual reports, insider reports and pricing supplements. – Global network of local experts results in unique coverage and unrivalled timeliness, with near real-time availability for document feeds and news dissemination channels and same-day availability for manually captured documents – 65+ years of history and 67 million documents in 40+ languages **Content:** – Coverage includes all US and global exchanges, including partial coverage of private and ESG companies, US municipal bonds and fixed income-related documents – Hundreds of document types, including annual and quarterly reports, equity and fixed income prospectuses, municipal bonds, M&As, ownership and other regulatory filings – 65,000 global companies; 14,000 SEC companies (active); 7,500 private companies and growing – Almost 4 million documents sourced in 2023; over 10,000 filings collected on average every day – 20,000+ documents captured in a given peak day **Metadata:** – Feed Filings are available on products seconds after being released. Semi-automated markets are sourced within 10 minutes. Other documents are made available on day of publication – Standardised Electronic Data Gathering, Analysis and Retrieval (EDGAR) tags for easy navigation, including filer status and document title – Display of SEC Delta reports for 10-Q and 10-K filings – Translated filings from Portuguese to English and English to Japanese – Different file formats in Repository: PDF, OCR, TXT, XML, HTML, XHTML, XBRL, EDGAR – Entity and topic tags (desktop only)

	Unstructured Text			
	News	**Research**	**Transcripts**	**Filings**
Top use cases	News, Research, Transcripts and Filings are complementary to each other, enhancing the value of these content sets when various unstructured text sources are used together. Top use cases include: – Quantitative trading – Market Surveillance – Research and backtesting – Research analysis (buy- and sell-side)			
Languages	Reuters: 12 Third-party news: 20+	30	English	40
Packages	– Reuters company news by region – Reuters political/ general/economic – All Reuters News – Global press wires – 130+ third-party sources	– Per pre-approval contributor – Open access contributor list	– Global	– Global
History	Reuters News: from 1996 Other news sources: from 2003	Archives available from 1982, varying by contributor	Archives available going back to 2001	From 1968
Delivery frequency	Real-time	Near real-time	Near real-time	Near real-time
Delivery methods and format	**Feed:** JSON – Real-time broadcast – API – Multicast via NY4 for ultra-low latency **Archive:** – SFTP delivered historical and end-of-day JSON files	**Feed:** PDF/text with JSON metadata **Archive:** Text with JSON metadata – PDF upcoming – API – Secure large file delivery	**Feed:** PDF/XML – Secure large file delivery – SFTP	Documents available in multiple formats including XBRL, HTML and PDF

2. Analytics – Delivering greater insights

Turn unstructured data into actionable insight and manage risk with our comprehensive solutions. Leverage News Analytics and MarketPsych Analytics to capture what's happening, and employ StarMine and Global Macro Forecasts delivered in partnership with Exponential Technology for predictive analytics. From leading news analytics with ultra-low latency delivery, to deep historical news archives and feeds, our solutions can optimise your programmatic workflows.

Analytics to support understanding of what has taken place

Today, vast amounts of news and social media data is unstructured. Our Machine Readable News Analytics tools can turn such inputs into actionable insights to maximise returns and spot event risk. News flow and sentiment are important sources of signals in quantitative stock selection and systematic trading. They also support understanding of market perception with regards to the relevance and confidence of news.

In addition, MarketPsych Analytics tools digitise themes and sentiments into machine-readable values and signals, covering all major countries, commodities, currencies, cryptocurrencies, equity sectors, and public and private companies. MarketPsych sentiment indicators can be leveraged for alpha generation, volatility forecasting, risk management, event monitoring, macroeconomic nowcasting and research.

News Analytics	MarketPsych Analytics
Overview: Using cutting-edge Natural Language Processing, News Analytics provides measurement of company sentiment, relevance and novelty of text, as well as other valuable metadata. Supporting use cases such as trade signal detection for quant investment strategies, backtesting and market surveillance, and with a particular focus on Equities and Commodities & Energy, News Analytics covers historical data dating back to 2003 and 40,000+ companies. So that when it comes to making sense of the news, and what it means for your business, News Analytics does the hard work for you. **Content:** Analytics on Reuters News plus around 50 third-party news sources: – Delivers headlines and scoring on news story body – Over 32,000 companies in English and 6,000 companies in Japanese – 42 Commodity and Energy (C&E) topics **Metadata:** Leveraging natural language processing (NLP) based calculations, metadata incorporates information including: – Sentiment: positive, neutral, negative tone at entity and document level – Relevance: importance of the entity to the news item – Novelty: similarity to previous news – Volume: number of recent mentions – Headline classification and headline analysis	**Overview:** MarketPsych Analytics from LSEG analyses news and social media in real-time to convert the volume and variety of professional news and the internet into manageable information flows that drive sharper decisions. The analytics are delivered as real-time data series that can be easily incorporated into your investment and trading decision processes – quantitative or qualitative. **Content:** Analytics from 3,000+ curated news sources and social media outlets published by the minute, hourly and daily aggregations for various entities: – 252 countries and territories – 100,000+ global companies, sectors and ETFs – 62 stock indexes and sovereign bonds – 44 currencies – 53 commodities – 500+ cryptocurrencies **Metadata:** – Buzz metrics that indicate how frequently a topic is being referenced in the media – Sentiment scores (positive or negative) – Emotional indicators such as fear, joy and trust – Fundamental and macro themes including earnings forecasts, layoffs, innovation and inflation

MarketPsych ESG Analytics	MarketPsych Transcript Analytics
Overview: LSEG's MarketPsych ESG Analytics analyse both financial and corporate sustainability-related news and social media in near real-time. Powered by MarketPsych data's AI-based natural language processing engine, the analytics are derived from millions of daily articles across thousands of global news and social media outlets. They provide numerical environmental, social and governance (ESG) insights on companies and countries to drive better investment decisions. LSEG's MarketPsych ESG Analytics of ESG data is designed to drive better investment decisions. **Content:** Analytics from 6,000+ curated news sources and social media outlets published by the minute, hourly and daily aggregations for various entities: – 252 countries and territories – 100,000+ global companies Two packages are available: – The Core package provides 17 daily scores for every company and country. Updated daily and based on a 365-day exponential average – The Advanced package provides access to Core plus 300 real-time analytics for companies and 456 analytics for countries. Minute, hourly and daily updates are available **Metadata:** – Core ESG scores ranging from 1 to 100 – Advanced bipolar ESG scores ranging from -1 to 1 – Advanced unipolar controversy scores ranging from 0 to 1 Scores cover 100+ environmental, social and governance themes including workplace safety, social unrest and emissions controversies.	**Overview:** A refined data feed that leverages cutting-edge Natural Language Processing (NLP) to decipher corporate discourse quickly and systematically. This solution delves into the nuances of every spoken sentence, offering insights into sentiment, relevant companies and topics being mentioned, and the intricate web of emotions speakers convey. Designed for the discerning eyes of institutional and quantitative firms, LSEG's MarketPsych Transcript Analytics provides a detailed, data-driven view on the undercurrents of corporate calls. **Content:** Analytics, powered by MarketPsych's proprietary NLP model, is applied to LSEG's machine-readable corporate Transcripts. A user interface allows interaction with the data via drop-down menus, charting tools and a generative AI agent. Two types of feeds are provided: – Standard consists of aggregated sentiments for each speaker, section and document – Premium provides detailed metadata down to the sentence level including verb tense, part of speech and the metadata as outlined below **Metadata:** – The output feeds include metadata derived from each sentence in the transcripts using several fine-tuned RoBERTa-based classifiers – Entities: Millions tagged from 20+ entity types, including companies, countries, locations, currencies, commodities, and more – Topics: 1,000+ topics and 4,000 events relevant to business, finance, legal and more – Sentiment: Numerically classifies the polarity of each sentence according to its financial and ESG context – Emotional profiling: 13 dimensions of emotional tones classified in each sentence, including fear, annoyance and optimism

	Analytics for capturing what's happening			
	News Analytics	**MarketPsych Analytics**	**MarketPsych ESG Analytics**	**MarketPsych Transcript Analytics**
Top use cases	– Quantitative trading – Market surveillance – Research and backtesting	– Quantitative trading – Market sentiment visualisation – Asset research: investigate market-moving events – Risk management – Monitor global macroeconomics and nowcast indicators – Cryptocurrency trading and monitoring	– Quantitative investors to provide uncorrelated alpha signals – Discretionary investors to allocate assets and manage risks – Corporate clients to monitor market perceptions – Regulators to precisely direct investigations – Analysts and researchers to identify greenwashing	– Alpha generation: Create custom data feeds for quantitative investment – Document search for unique trends, phrases and sentiments – Improve portfolio risk management
Languages	– English – Japanese: processed in native language, not translated	13: English, Arabic, Chinese, Japanese, Portuguese, Dutch, French, German, Indonesian, Italian, Korean, Spanish and Russian	13: English, Arabic, Chinese, Japanese, Portuguese, Dutch, French, German, Indonesian, Italian, Korean, Spanish and Russian	English
Packages	– Companies by region and language – C&E topics – Macroeconomic news and events	– Companies (including sectors, ETFs and indexes) – Countries – Country-level equity, rates, fixed income – Currencies – C&E – Cryptocurrencies	– The Core package provides 17 daily scores for every company and country – The Advanced package provides access to Core plus 300 real-time analytics for companies and 456 analytics for countries	Transcripts of earnings conference calls, conference presentations, guidance calls, sales presentations, and more since 2002 for 12,000+ global companies
History	From 2003	– From 1998 for most packages – From 2009 for cryptocurrencies	From 1998 (depending on languages)	From 2002
Delivery frequency	Near real-time	By the minute Hourly Daily	By the minute Hourly Daily	Event-driven
Delivery methods and format	– Real-time broadcast: JSON – SFTP Archive: JSON or tabular	– SFTP: live feed and archive – API: push, pull, and bulk file delivery available – Website data file download – QAD – Feed: Tabular & JSON – Archive: Tabular	– SFTP: live feed and archive – API: push, pull, and bulk file delivery available – Website data file download – LSEG Datastream – Feed: Tabular & JSON – Archive: Tabular	– API (JSON and CSV outputs) – API provides flexibility to generate custom data requests for historical testing and live deployment via a UI App

Analytics to help predict future impact

Grounded in sound economic intuition and backed by rigorous analysis, the StarMine suite covers critical areas such as value, momentum, ownership, risk and quality. We deliver the tools, content, text and insights needed to reliably predict market movements, identify gaps and pinpoint important trends. The StarMine suite also helps when introducing new angles to investment strategies, testing, validating and benchmarking proprietary quantitative methods.

Global Macro Forecasts are advanced economic indicators powered by Exponential Technology. These leverage cutting-edge indicators to provide actionable insights into global economic trends weeks in advance enabling customers to defensively pre-position portfolios for potential macro surprises or offensively take active risk if the consensus fails to incorporate economic predictions.

StarMine M&A Target Model

Overview: With the StarMine M&A Target Model you get access to the first commercially available model that uses a Large Language Model (LLM) to rank potential Merger and Acquisition (M&A) targets within the next 12 months for over 38,000 public companies.
The StarMine M&A Target Model uses a LLM on Reuters news textual content, machine learning on fundamental point-in-time datasets and credit profiles. The model has two components that separately analyse industry-leading structured and unstructured data to deliver a best-of-breed ranking.

Content: The model provides a measurement of the likelihood of a public company to be acquired by analysing textual and fundamental content about the company. The model consists of two components – a text component and a fundamental component. The text component follows a bottom-up approach – starting with Reuters news documents, calculating document scores using the BERT-RNA model and aggregating document level scores to a company-level score. The fundamental component uses structured data from fundamentals, deals, shareholder activism, pricing and corporate actions to generate a company score. The model combines outputs of both components to generate a final M&A Target Score.

- Over 38,000 public companies covered globally. Some regions have more M&A activity than others, and the model accounts for this by training the model separately on different regions.
- Model identifies and incorporates sectors having waves of M&A activity and companies that have been takeover targets and/or involved in proxy fight in the past.
- Model leverages the breadth and depth of LSEG's point-in-time data, and Reuters as the trusted source of breaking news.
- Point-in-time daily history from January 2000.

Metadata:
In addition to model final scores, the model provides component details:

- Score: percentile ranking based on the relative likelihood of acquisition
- Components: Text, fundamental, credit, size, valuation, prior target, proxy fight, sector wave, dividend
- Document details used by model

StarMine Text Mining Credit Risk Model

Overview: StarMine Text Mining Credit Risk Model (TMCR) assesses the risk in publicly traded companies by systematically evaluating the language in Reuters news, events conference call transcripts, corporate filings (10-K, 10-Q and 8-K), and select broker research reports to predict which firms are likely to come under financial distress and which are likely to thrive. It is a percentile ranking (1-100) of stocks, with 100 corresponding to the healthiest companies.

Content:
This very unique component of StarMine Credit Risk applies sophisticated text mining algorithms to Events earnings conference call transcripts, financial statements and other regulatory filings, Reuters news, and select broker research reports to identify language that is predictive of credit risk. The model uses custom dictionaries for each type of document to accurately assess the unique diction and style in each one. The model allows analysts to quickly identify the most important documents for a company out of the potentially hundreds they may be responsible for, and it gives quantitative managers a powerful, new quantitative signal by systematically analysing a large body of previously untapped qualitative data.

- Over 39,000 public companies covered globally. Model maps default probability to letter ratings like the agency ratings
- Model identifies key language from multiple text sources to turn raw textual data into credit scores
- Model analyses each document source independently and then combines to create an overall probability of default
- Point-in-time daily history from 1998

Metadata:
- In addition to model final scores, model provides component details:
  - Score: predicts which firms are likely to come under financial distress and which are likely to thrive
  - Components: Transcripts, News, Filings, Research
  - Document details used by model

StarMine MarketPsych Media Sentiment Model

Overview: The StarMine MarketPsych Media Sentiment Model (MMS) marks the first StarMine equity returns model based on news and social media sentiment. The model is a stock ranking system that provides a 1 to 100 daily percentile ranking for over 16,000 global stocks. The MMS scores are designed to forecast the next month's relative share price returns, with higher ranked stocks outperforming lower. The model is derived from LSEG's MarketPsych Analytics, a market leader in financial media sentiment data.

Content:
The StarMine MarketPsych Media Sentiment model (MMS) is designed to capture both under- and over-reaction patterns by optimally weighing these pertinent themes and sentiments. Aside from considering the impacts themselves, MMS also accounts for the persistence of their effects on price, varying by news and social media. The model construction process is carefully chosen to:

- Determine which of LSEG's MarketPsych sentiment indices perform consistently across regimes
- Investigate the duration of prediction of each of LSEG's MarketPsych indices
- Develop optimal handling of sparse and noisy data
- Create a logical combination of LSEG's MarketPsych indices that allows for further study

MMS model construction uses sentiment indices that fall into three categories. Each of equity, business and management groups express a unique aspect of media sentiment about the company.

- Over 16,000 public companies covered globally
- Point-in-time daily history from 1998

Metadata
- In addition to model final scores, the model provides component details:
  - Score: The model condenses the predictive power of media topics, themes, and sentiments into a simple 1-100 ranking
  - Components: Equity, business and management scores

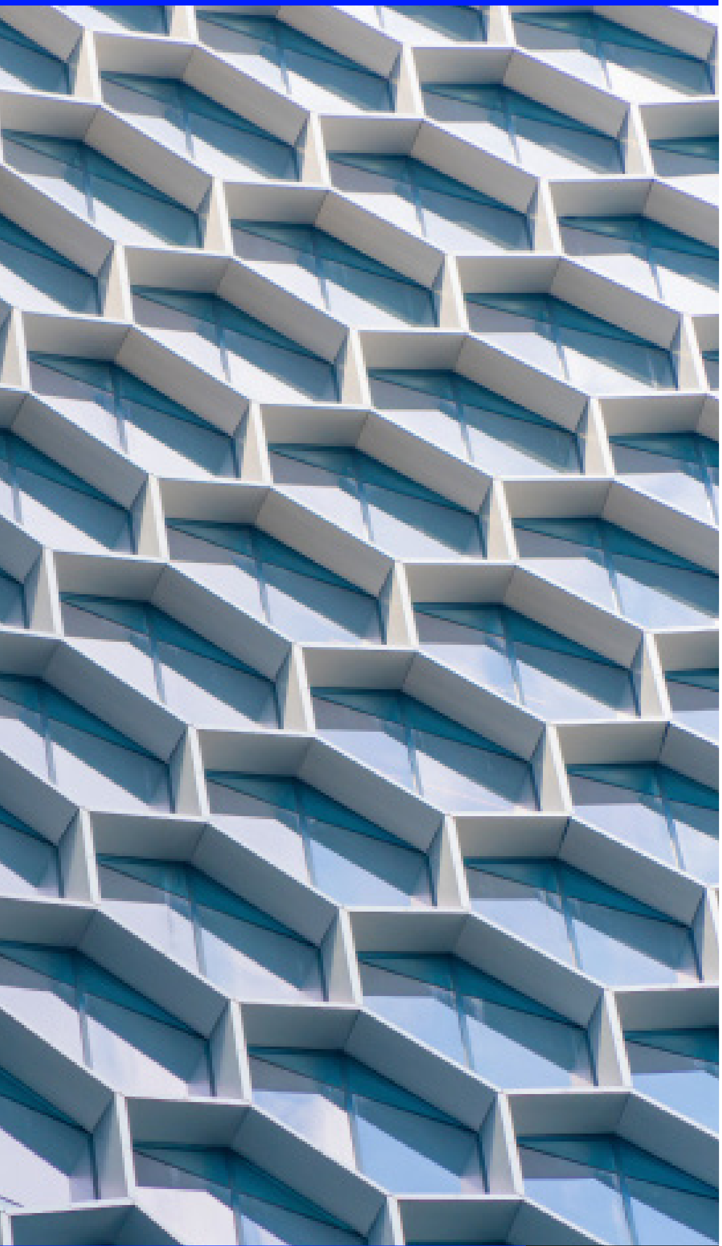
Global Macro Forecasts

Overview: LSEG's Global Macro Forecasts leverage cutting-edge indicators to provide actionable insights into global economic trends. They use fundamental modeling techniques for each macroeconomic release, including novel, real-time business activity and survey data, machine learning algorithms and advanced statistical methods to accurately forecast economic indicators weeks in advance.

Content:
- Forecasts include the US Consumer Price Index (CPI), US Retail Sales, Conference Board Consumer Confidence Index and Michigan Consumer Sentiment Index
- Individual CPI components including food, shelter, gasoline and transportation are broken out in addition to topline CPI
- The indicators span timeframes ranging from one month to two days ahead of economic data releases
- Month-over-month and year-over-year percentage changes, as well as the absolute figures, accompany accuracy performance stats over one, two and five-year horizons

Metadata:
- Forecasts are powered by LSEG data including Reuters news, ECI and Polls, together with Tick History data, for a robust, powerful solution

	Predictive analytics			
	StarMine M&A Target Model	**StarMine Text Mining Credit Risk (TMCR) Model**	**StarMine MarketPsych Media Sentiment (MMS) Model**	**Global Macro Forecasts**
Top use cases	– Quantitative trading – Research and backtesting – As a factor in multi-factor portfolio selection – Mitigate portfolio risk – Idea generation	– Analysis of corporate counterparty risk – Research and backtesting – Screening and portfolio selection – Mitigate portfolio risk – Simplified analytics and insights for advisors	– Quantitative investment: MMS is an orthogonal factor for stock price prediction – Stock ranking and recommendation systems – Risk reduction in portfolio management	– Enhance proprietary analytics – Improve rebalance timing to improve performance and support client retention – Better anticipate macro events and understand market dynamics – Forecast macroeconomic surprises ahead of announcements
Languages	– English	– English	13	English
Packages	– Companies by region – Global, AMERS, EMEA and APAC regions	– Companies by region – Global, AMERS, EMEA and APAC regions	Overall, business, management and equity scores	– US Consumer Price Index (CPI) including individual CPI categories – US Michigan Consumer Sentiment Index – Conference Board Consumer Confidence Index – US Retail Sales – National Retail Federation (NRF) – Other indicators to follow
History	From 2000	From 1998	From 1998	From 2017
Delivery frequency	Daily	Daily	Daily	From one month to two days ahead of economic data releases
Delivery methods and format	**Feed:** Data Platform – JSON, CSV QA – SQL Deployed/ Onsite **Desktop:** API, UI, Excel	**Feed:** SFTP – Text, CSV QA – SQL Deployed/ Onsite **Desktop:** API, UI, Excel	CSV SQL Text Deployed/Onsite Servers FTP SFTP API QAC (cloud instance)	**Feed:** JSON for real-time; JSON CSV or Parquet for historical **Desktop:** REST API, visualisation tools



Find out more

Machine Readable News

Text Analytics

StarMine Analytics and Models

MarketPsych Analytics and Models

For more information please contact us | LSEG

Visit **lseg.com** | @LSEGplc LSEG

LSEG DATA & ANALYTICS