EVENT-DRIVEN FEEDS

QUANTITATIVE RESEARCH

Newsfeed Demystified

Prepared by: Lei Huang Ph.D., Product Manager, Bloomberg L.P.

Bloomberg
FOR ENTERPRISE

Contents

Introduction.......... 3

Machine-Readable Newsfeeds.......... 4

Sample Data.......... 5

Topic Codes.......... 7

Application 1: When to Find Important News.......... 11

Application 2: Where to Find Important News.......... 14

Application 3: Predicting Which News Will Move the Market.......... 17

Final Remarks.......... 24

Introduction

The modern investing landscape is extraordinarily complex. Success is no longer about who you know and what they know. Over the past decade, data has become the most important component of the financial system. News, social media, economic releases, corporate actions and other unstructured datasets flow through machine-readable feeds bit by bit and are processed by computers that can make instantaneous trading decisions.

It is virtually impossible for even the most experienced and knowledgeable investment professionals to compete directly with such real-time information processing. Humans spend most of their time concentrating on understanding the data itself, designing ways to use it algorithmically and thus enhancing their investment models' performances.

During a typical data onboarding process, major challenges need to be properly addressed before the full potential of the data can be unlocked. Because of their varying business needs, different firms usually have very specific application scenarios; accordingly, pinpointing relevant pieces of information within the haystack of available data can be extremely difficult. To be effective, researchers must have an in-depth understanding of the processes involved in data origination, curation and delivery—domain nuances that can often seem mysterious to those other than data vendors. This lack of transparancy leaves researchers with little idea of what the possibilities are and what they could realistically expect.

This paper is designed to help demystify the process of consuming a newsfeed. We will examine the relationship between the effect of news on humans and on machines using Bloomberg's Event-Driven Feeds. The datasets provide unique views into aggregate news consumption activities as well as the short-term market impact on related stocks.

We will also address two key questions:

- Do the number of financial professionals reading a news story influence market movement as much as newsfeeds that inform algos?
- What are some of the most impactful topic codes assigned to real-time news items?

Machine-Readable Newsfeeds

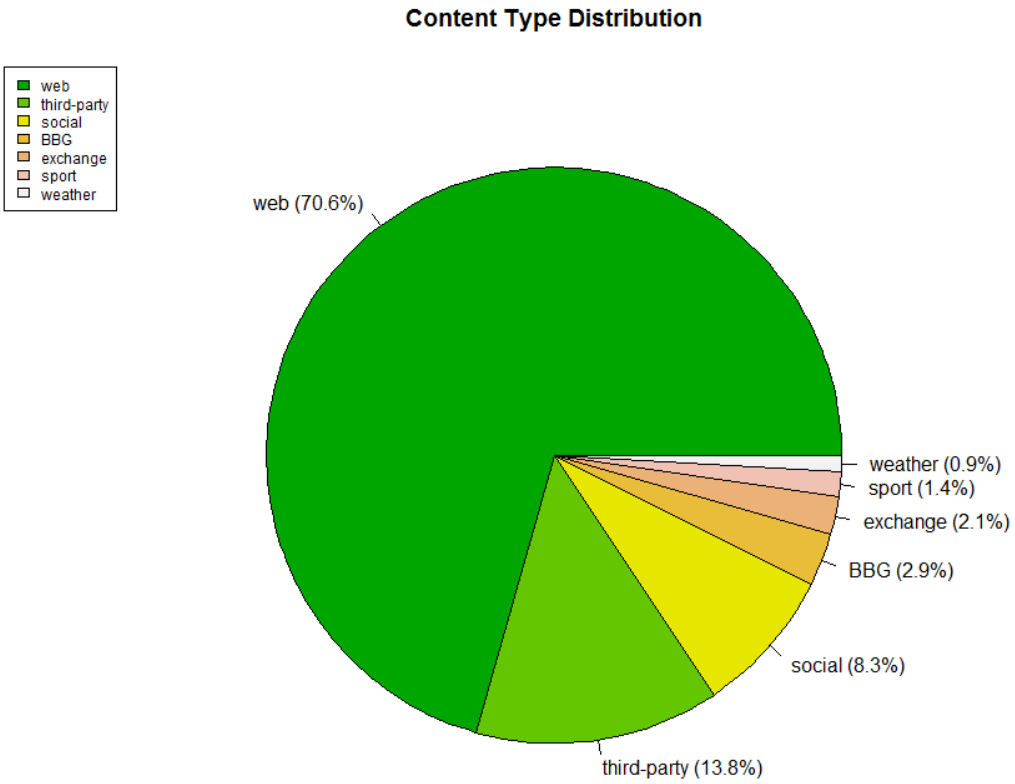
Machine-readable newsfeeds provide real-time delivery of textual news in machine-readable formats, including XML and JSON. They are designed for systematic applications that leverage news information to meet individual business demands. Established use includes low-latency trading of headlines, news detection for defensive market making, sentiment-driven longer-term strategies, news surveillance for compliance and risk management, etc.

A wealth of information is available in a typical commercial offering. For example, Bloomberg's Textual News Feed is offered by Bloomberg Enterprise Solutions as part of its broader Event-Driven Feeds category. The feed aggregates global breaking news from 151 Bloomberg bureaus worldwide as well as from valuable third-party sources. Textual news information is paired with metadata that encodes related companies, topics and people. Each message is timestamped point-in-time with millisecond precision.

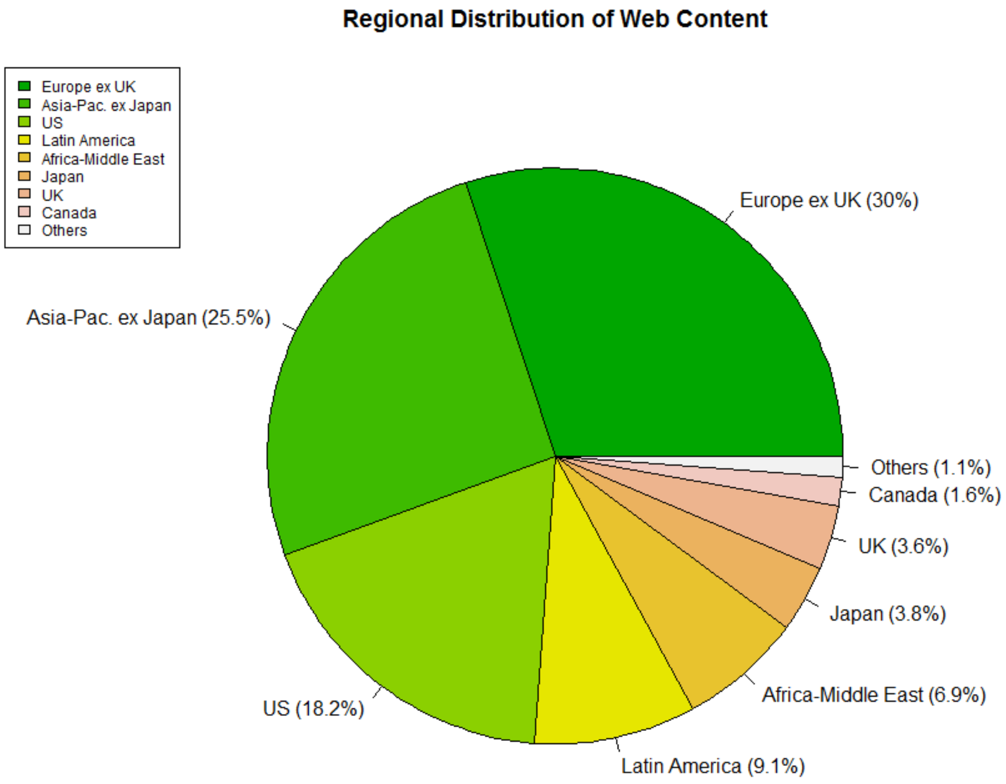
Sample Data

Live recorded messages in the Bloomberg Headline and Classification Data feed (EID 54418) from July 1, 2016 to Dec. 31, 2016, were used for this study. This machine-readable textual feed contains both real-time headlines and metadata for topic/company/people classifications. The six-month archive contains 83 million unique headlines sourced from 399 different wires.

The following chart shows the breakdown of the headlines by types of source wires. The majority of the headlines were scraped from public websites, followed by those published by third-party wires and social media platforms. Bloomberg's newsroom (BBG) generated approximately 3% of all headlines that appeared in the feed.



The following chart shows the breakdown of the web-scraped headlines by geographic location of the source websites. Europe excluding the UK, Asia-Pacific excluding Japan, and the U.S. are the top three regions, contributing approximately 74% of the global web-scraped content.



Topic Codes

Machine-readable topic codes, which are a unique feature of the real-time newsfeed, bring important structure to otherwise unstructured content flow. Bloomberg curates a large volume of topic codes that are frequently reviewed and updated, with 6,248 unique topic codes appearing in the six-month archive.

Topic codes can be assigned to headlines via four different mechanisms at different stages:

- They can be discretionarily assigned by reporters and editors upon initial news release
- They can be automatically derived from originally assigned codes by a rules-based classification engine
- They can be subsequently discretionarily assigned by editors as special tags for content promotion
- They can be automatically assigned whenever preset news impact objectives are hit

The last mechanism is especially interesting. These automated, retrospective tagging processes are triggered by objective rules. Thus news stories with realized impact can be determined unambiguously, albeit at a significant delay after initial publication. In practice, such "after-the-fact" codes cannot be used proactively to identify noteworthy content from incoming messages of a live feed. However, they are very useful filters for historical analysis to build the foundation on which real-time applications can be developed.

Two kinds of retrospective codes will be analyzed in this study. They both measure the realized impact of a given headline but from different perspectives.

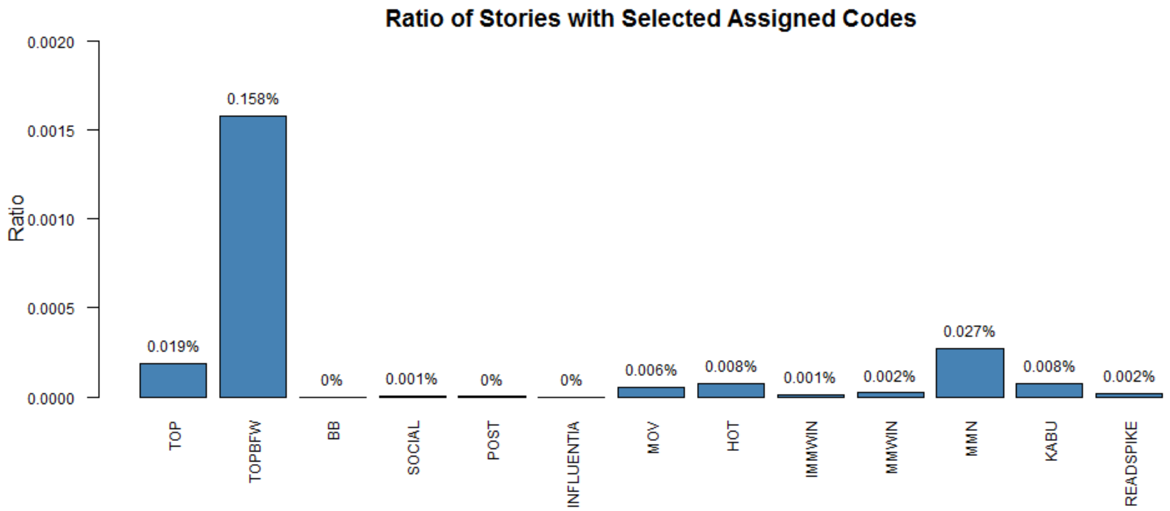
The first is readership codes, which measure the headline consumption from the Bloomberg Terminal user community. On the Terminal screen, news headlines automatically scroll to keep users informed on latest developments. Each headline usually stays on the screen for a very brief time; the physical display size also limits its character length. For additional information, such as the story body or a link to the original news article, users have to hit the headline link to be redirected to another page.

Readership codes are assigned to headlines when their aggregate count of Terminal users meets various preset thresholds. For example, a headline will get a READ25 code after it is read by more than 25 unique Terminal users within an eight-hour rolling window. The following chart shows the ratio of all headlines with readership codes using 15 different thresholds. Overall, only a very small portion of the headlines received significant user hits.

For instance, around 0.027% of headlines were hit by more than 500 unique readers (READ500)—this corresponds to 22,400 headlines in the six-month archive.

The second kind is the MMN (Market-Moving News) code, which measures the short-term market impact of the headlines. Bloomberg systematically tracks intraday price actions and news publication activities for a large universe of stock tickers in real time. Whenever a headline is identified as driving a short-term price anomaly, such as an immediate price jump or strong trend formation, it will automatically be assigned a MMN code. On average, about 0.027% of headlines receive the MMN code. This corresponds to 22,598 headlines in the six-month archive.

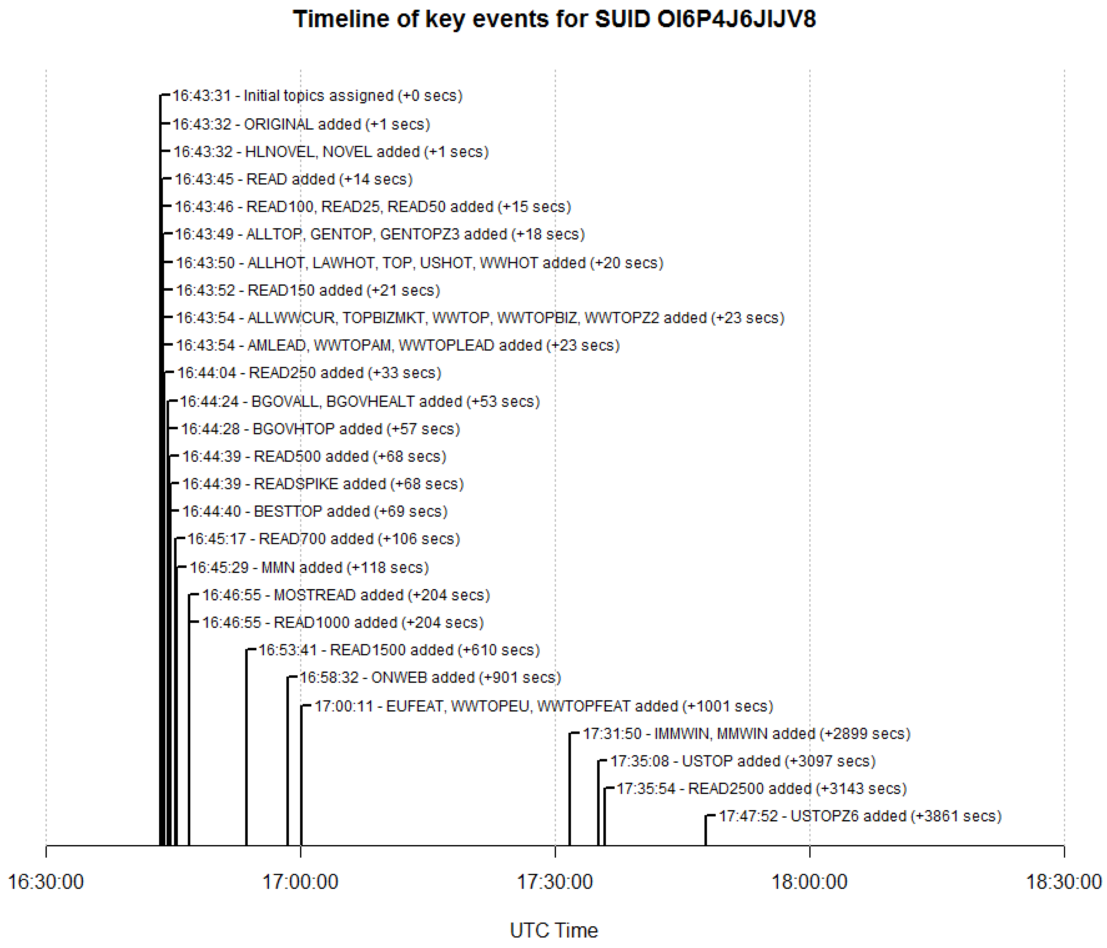
The following chart shows the ratio of all headlines that contain the MMN code, as well as a variety of other codes that are commonly used to indicate headline importance.



The following table gives definitions of the topic codes shown in the chart above.

Topic Code	Definition	Assignment method
TOP	Top Stories	Judgment-based
TOPBFW	First Word Top Stories	Judgment-based
BB	Most Important News	Judgment-based
SOCIAL	Most Important Social Media	Judgment-based
POST	Most Important Weibo Posts	Judgment-based
INFLUENTIA	BN Most Influential News	Judgment-based
MOV	Equity Movers	Judgment-based
HOT	Hot Headlines	Judgment-based
IMMWIN	Influential Market Mover	Judgment-based
MMWIN	Market Mover	Judgment-based
MMN	Market-Moving News	Rule-based
KABU	Japan Equity Market-Moving News	Judgment-based
READSPIKE	Readership Spikes	Rule-based

The multi-stage code assignment process can result in a single story making multiple rounds of appearances in the feed, known as “story passes.” In the Bloomberg system, passes of the story can be identified by the same SUID—Story Unique Identifier. For example, on Dec. 14, 2016, Bloomberg News reported that the U.S. had filed first charges in a generic drug price-fixing probe. The following chart shows the timeline of major event updates associated with that story (SUID OI6P4J6JIJV8).

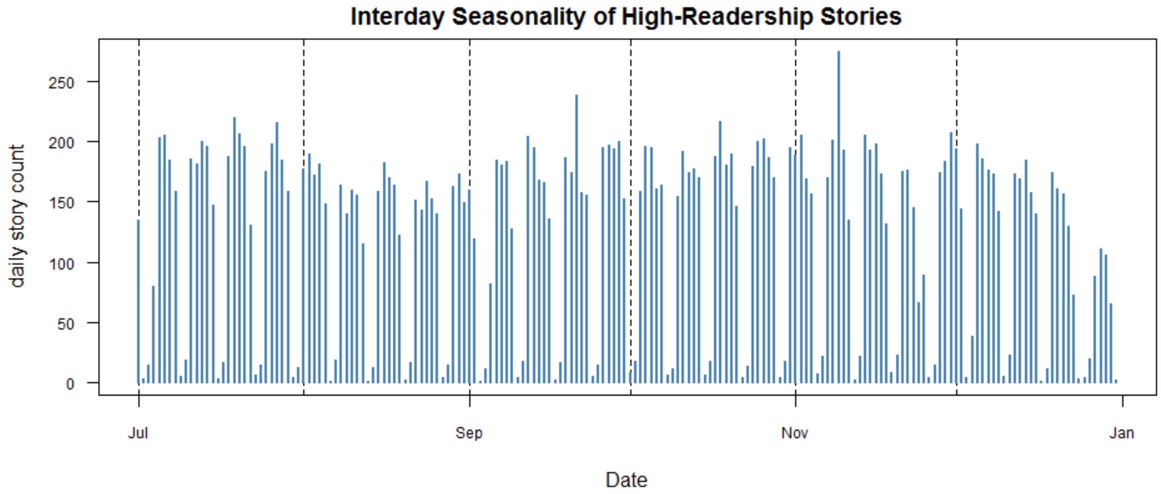


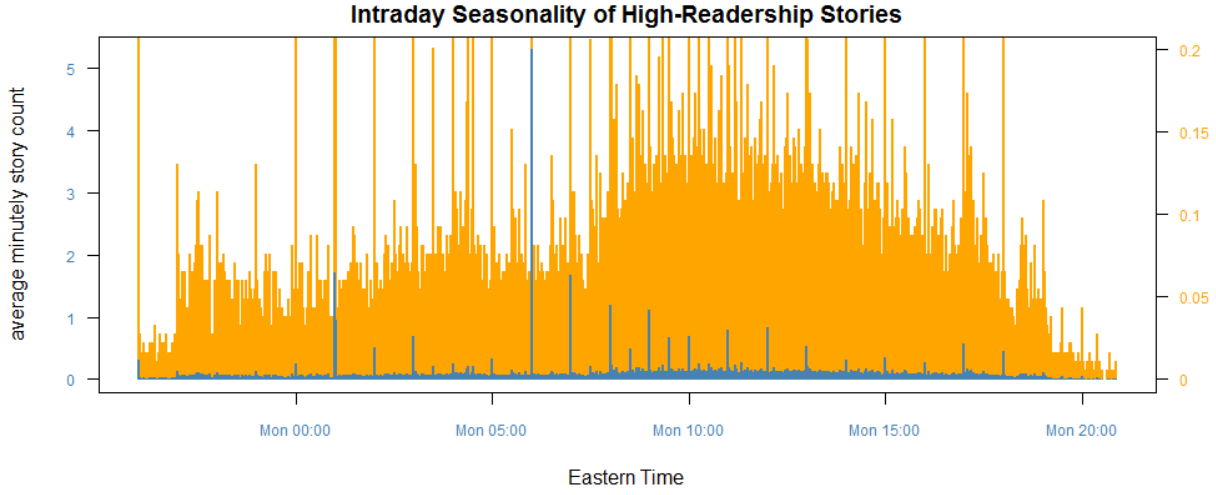
In the following sections we demonstrate three simple applications that can leverage topic codes to extract valuable insights from the newsfeed archive.

Application 1: When to Find Important News

Each day, hundreds of thousands of news headlines flash on trading screens worldwide. The sheer volume of this information flow makes it impossible for any human reader to digest the news in its entirety. A common question: How often do news headlines arrive throughout the day? Are there any time windows in which important news headlines are more likely to appear?

To answer these questions, we tracked the daily and intraday arrival patterns of important headlines by leveraging the two retrospective topic codes: READ500 and MMN.

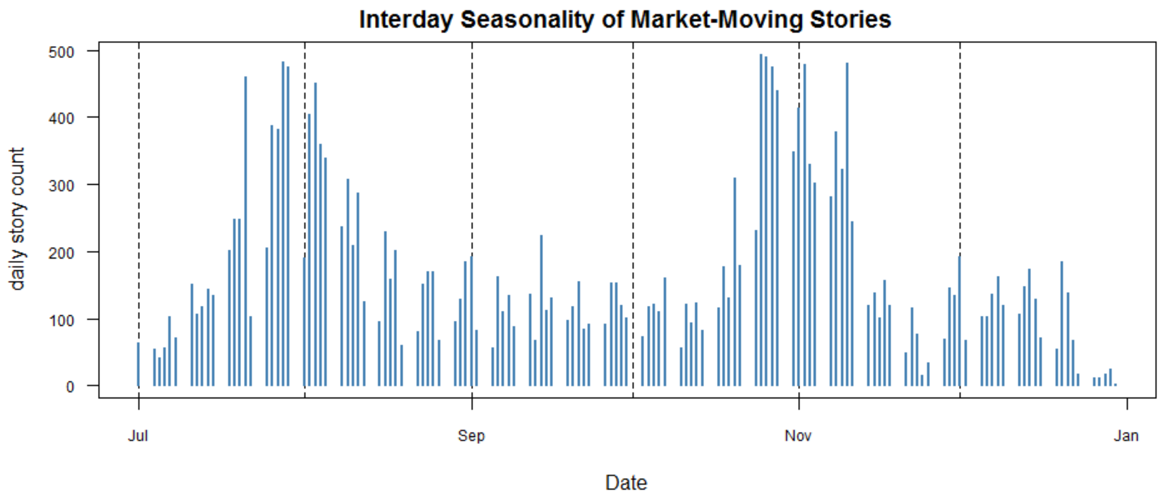


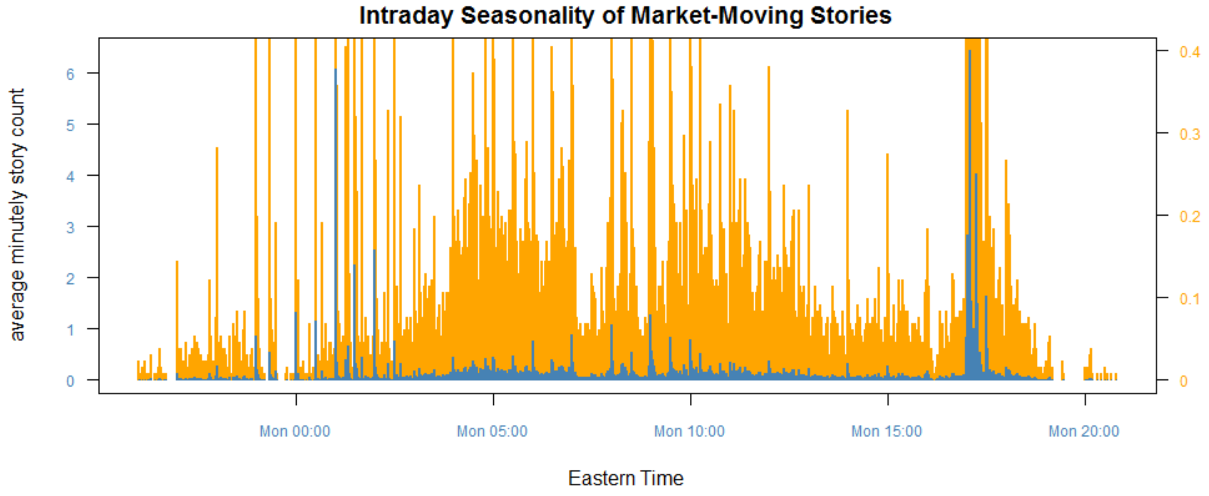


The charts immediately above show the seasonality of stories with the retrospective readership code READ500.

The top panel shows the daily statistics during the second half of 2016. The daily publication traffic is generally stable, fluctuating around 150-200 headlines each day during weekdays, of which Friday is usually less eventful than Monday to Thursday. Weekends see much-reduced traffic, with virtually no headlines on Saturdays and fewer than 10 headlines on Sundays. The overall traffic tapers off significantly during holiday seasons, such as at the end of November (Christmas season) and final weeks of December (end-of-year holiday season).

The bottom panel shows the intraday statistics in Eastern time. The blue bars are the full-scale plottings of the minute-by-minute publication counts. The orange bars are the rescaled plottings of the same time series (according to the right-side axis and labels) to help visualize the baseline seasonality. The intraday headline arrival pattern is notable for isolated peaks on top of a modulated baseline. The isolated peaks are in exact coincidence with the whole or half hours—directly related to news embargoes. The top five times (Eastern time) with the highest minute-by-minute counts are 5am (~5.3 headlines), 6am (~1.7 headlines), midnight (~1.7 headlines), 7am (~1.2 headlines) and 8am (~1.1 headlines). The modulated baseline shows that high-readership stories arrive at a stable rate from midnight to around 6am, after which the rate picks up significantly and peaks from 9am to noon. The arrival rate starts to weaken during afternoon hours. The weakest activity is observed from 6pm to 8pm, after which the arrival rate recovers to a level similar to the early morning hours.


Interday Seasonality of Market-Moving Stories


Intraday Seasonality of Market-Moving Stories

The charts above show the seasonality of stories with the retrospective market impact code MMN. The same methodologies are used as in the previous analysis.

The top panel shows the daily statistics during the second half of 2016. The daily publication traffic reflects very prominent quarterly seasonality, which is directly related to earnings reporting schedules. No traffic is seen on the weekends—no market data is available for intraday impact calculation because the stock market is closed. Practically, the market impact for weekend market-moving news will be reflected by the price gaps between Friday close and Monday open. However, such measurements are not supported by current market monitors.

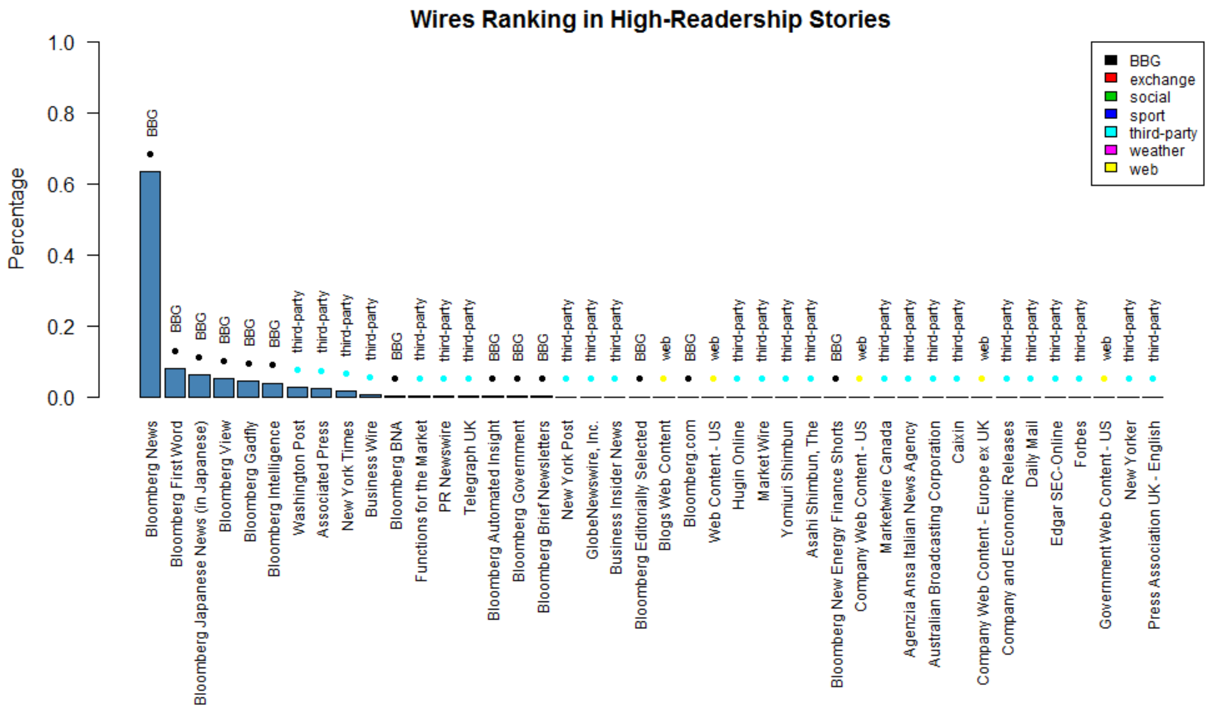
The bottom panel shows the intraday statistics in Eastern time. The intraday patten of market-moving stories shows a number of notable differences compared with that of high-readership stories (as shown previously). First, the preferred times of market-moving stories have a very different distribution. The top five times (Eastern time) with the highest minute-by-minute counts are 4:05pm (~6.4 headlines), midnight (~6.1 headlines), 4:15pm (~4 headlines), 4:01pm (~2.8 headlines) and 1am (~2.5 headlines). Three of the top spots are found in the half-hour window starting at 4pm—a time when public companies often release quarterly earnings results. Second, the activity levels from 3am to 6am are comparable to those during U.S. morning trading hours (9am to 12 noon). This is the result of headlines affecting stock prices in the other major global exchanges.

The results obtained using two different types of subsets give us direct evidence that the news stories that receive wide readership attention are not exactly the same as those that actually move the stock market. The discrepancy could be due to many reasons. One plausible explanation is that news stories gain in popularity if they are related to macro-level political and economic events, but are more likely to move the stock market if they contain corporate-level, fact-based information that can materially change a company's valuation. This explanation will be further substantiated in the next two sections.

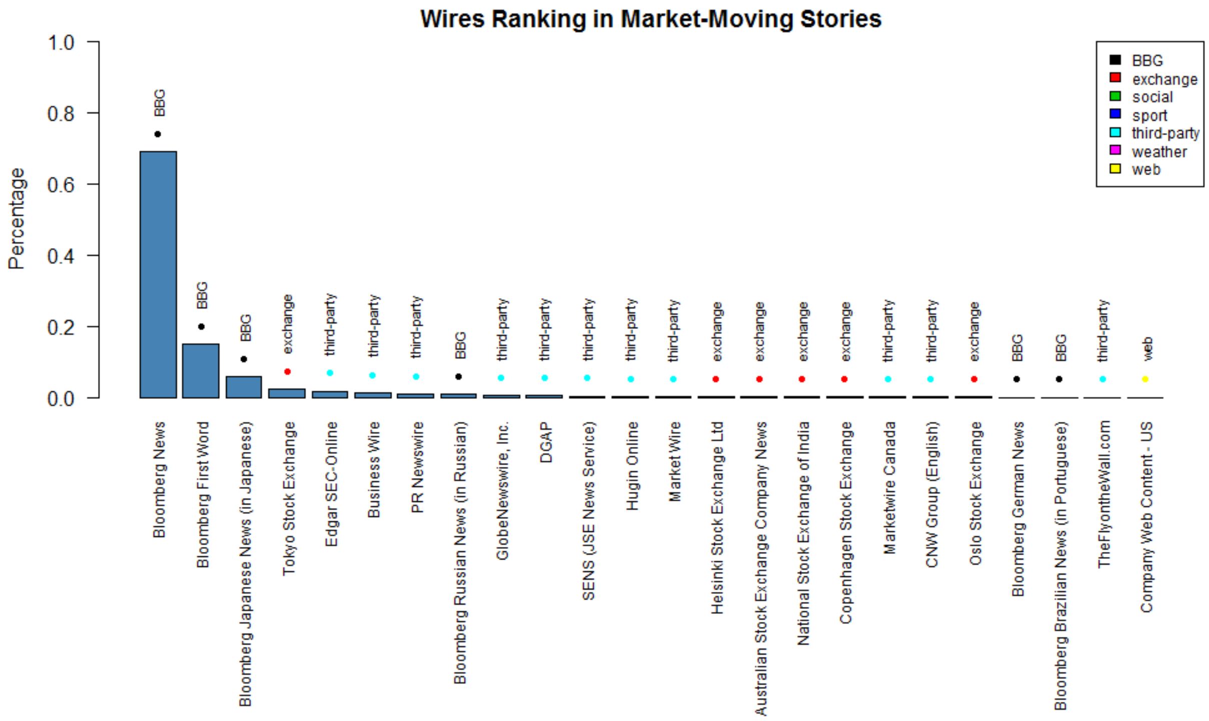
Application 2: Where to Find Important News

Another common question of news consumers: How do stories published by different newswires differ from one another? Do certain wires publish more impactful news content than others and, if so, which ones?

One caveat: not all wires are available to all Terminal users. Subscription to certain wires, known as "premium content," needs to be specially entitled, with associated fees and contractual arrangements. Our analysis uses EID 54418 of Bloomberg's Event-Driven Feeds—a standard subscription that contains all non-premium sources. Therefore, the following results and comparisons apply only to those non-premium wires.



The chart above shows the wire sources of stories with the retrospective readership code READ500. The wires are categorized into seven different groups. Each bar shows the percentage of stories published by a given wire. Although a total of 399 different wires are in the feed, only 42 wires actually published headlines with significant readership. Of those 42 wires, 13 are Bloomberg wires, 24 are third-party wires and 5 are web-scraping wires. The top three wires (Bloomberg News, Bloomberg First Word and Bloomberg—Japanese News) account for 78% of the high-readership headlines. Other top-ranking non-Bloomberg wires are well-known, including those of the *Washington Post*, The Associated Press and the *New York Times*.



The chart above shows the wire sources of stories with the retrospective market impact code MMN. Compared with the previous chart, this shows even fewer wires—24 in total—that published headlines with immediate market impact. Of those 24 wires, 6 are Bloomberg wires, 6 are direct exchange feeds, 11 are third-party wires and 1 is a web-scraping wire. The top three wires (Bloomberg News, Bloomberg First Word and Bloomberg—Japanese News) are the same as those measured by readership; in total, they account for 90% of market-moving headlines. Other top-ranking non-Bloomberg wires are mostly exchange wires (Tokyo Stock Exchange), regulatory wires (Edgar SEC Online) or press release distributors (Business Wire, PR Newswire, etc.).

The wire source comparison confirms our original hypothesis. People are more willing to spend time reading news from well-known, reputable publishers. These skillfully written news articles usually cover hot and popular topics to engage readers' attention. Stories that move stock prices, however, are usually those containing important, previously unknown material information on individual companies. Such information can be offered as prepared press releases (from press release distributors or stock exchanges) or as regulatory filings (from regulatory wires). While most of these market-moving stories are relatively dull for human readers, they are actually easily parsable by applications that employ natural language process (NLP) techniques.

Application 3: Predicting Which News Will Move the Market

The previous sections provide insight into the aggregate distribution of noteworthy news content across time and origin—however, that knowledge is not sufficient to determine the potential impact of individual headlines in real time. While retrospective topic codes triggered by realized impact can be a powerful tool for historical analysis, by definition, they are not present in the initial headline release. As previously noted, more than 6,000 unique topic codes are present in Bloomberg's six-month archive. A question that newsfeed users often ask: Which initial topic codes are most relevant for high-impact stories?

To determine the code relevancy, we define the conditional probability as follows:

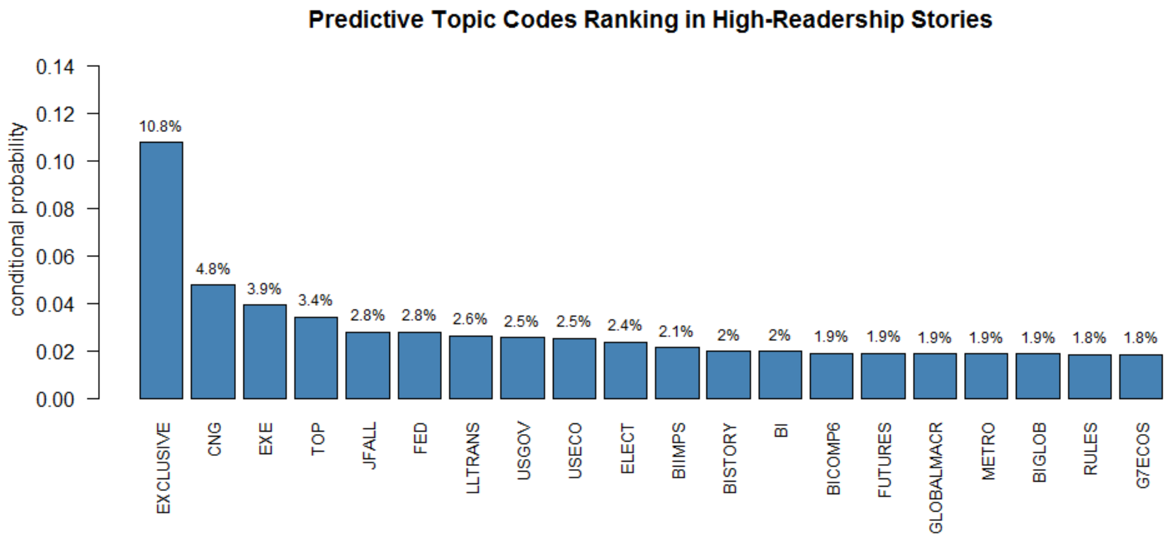
$$P(retrospective\ target\ code|initial\ code) = \frac{P(initial\ code|retrospective\ target\ code) \cdot P(retrospective\ target\ code)}{P(initial\ code)}$$

In the equation, "retrospective target code" represents the type of impact measurement (readership or market impact); "initial code" represents individual codes under test that are available in the first appearance of each story.

The $P(retrospective\ target\ code|initial\ code)$ is the conditional probability of seeing target codes (in subsequent updates) given the appearance of initial code (in the first message). The $P(initial\ code|retrospective\ target\ code)$ is the conditional probability of seeing the initial code (in the first message) given the appearance of target code (in subsequent updates). The $P(retrospective\ target\ code)$ and the $P(initial\ code)$ are unconditional probabilities within the entire sample set.

To further reduce the noise, we remove initial codes that have very low unconditional probabilities. Such codes are usually idiosyncratic to specific stories and therefore carry little information for more-generic applications. Specifically, we remove initial codes if:

$$P(initial\ code) < P(retrospective\ target\ code)$$

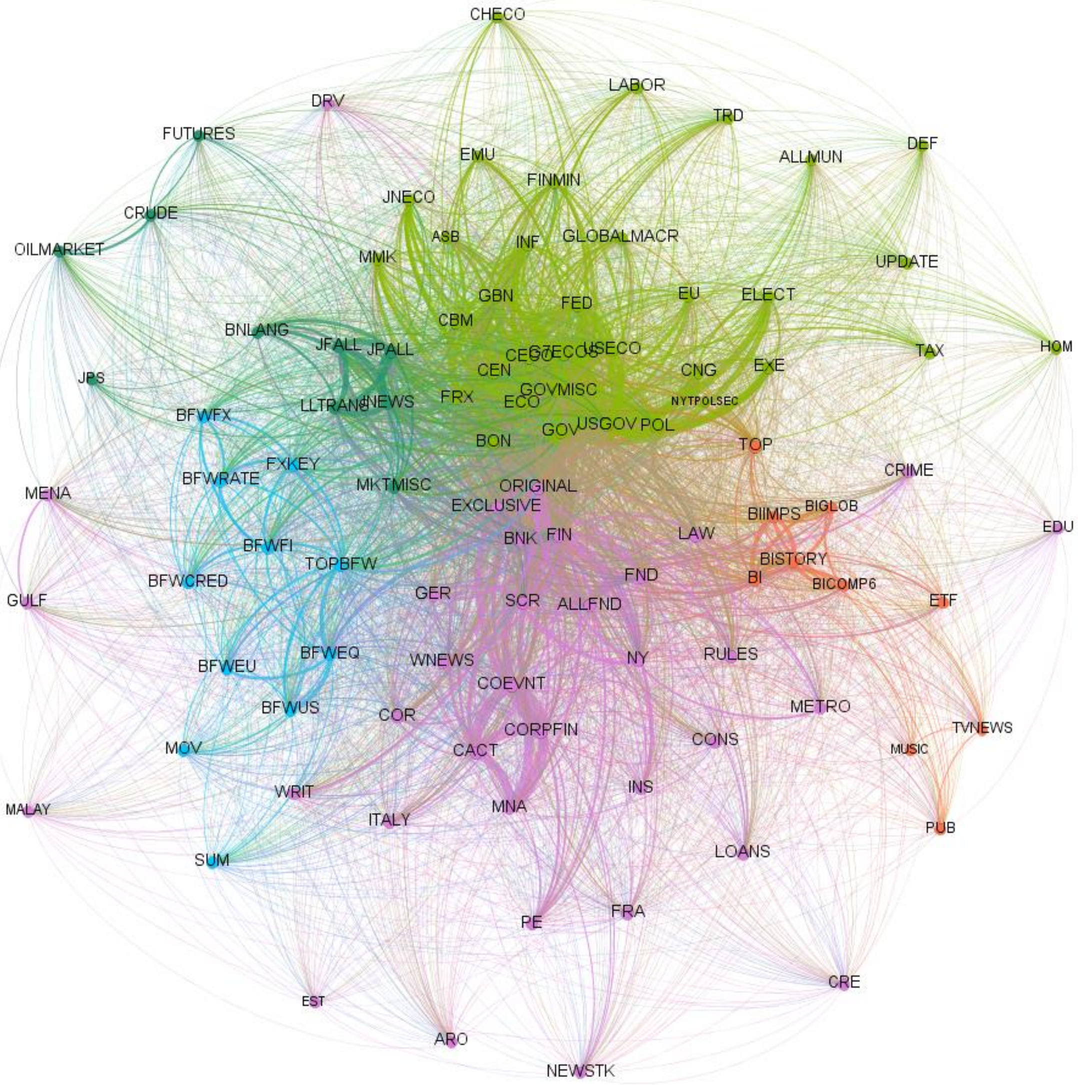


The graph above shows the ranking of initial topic codes for stories with the retrospective readership code READ500. It shows the top 20 codes ranked by their conditional probabilities. Almost all codes are related to macro-level politics and the economy.

The following table gives code definitions.

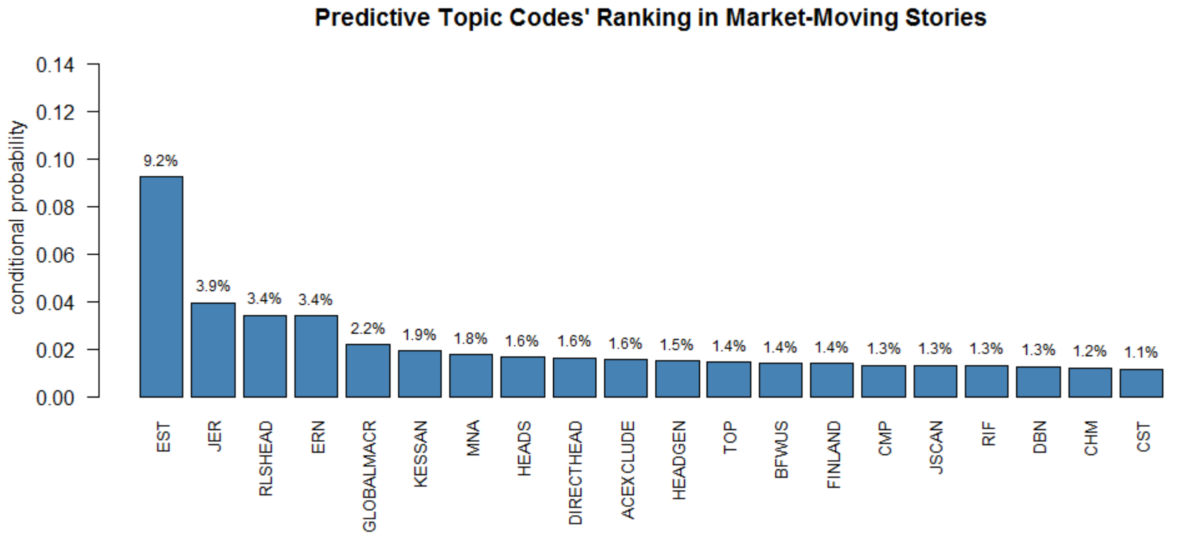
Topic Code	Definition
EXCLUSIVE	Bloomberg News Exclusives
CNG	Congress, U.S.
EXE	The White House
TOP	Top Stories
JFALL	Overseas News (Japanese)
FED	Federal Reserve
LLTRANS	Local Language Translations
USGOV	U.S. Government
USECO	U.S. Economy
ELECT	Elections
BIIMPS	BI Impact Statement
BISTORY	BI Storyboard
BI	BI Analysis
BICOMP6	BI Competitor Analysis 6
FUTURES	Futures
GLOBALMACR	Global Macro
METRO	Metropolitan-area News
BIGLOB	BI Global
RULES	Rules and Regulations
G7ECOS	Group of Seven Economics

Because of intrinsic logical dependencies, topic codes are also interconnected. When a group of codes is strongly associated with each other, the codes tend to coexist for given stories. This can quickly diminish the additional predictive power of new codes if their closely related peers are already being used—a behavior similar to the classic colinearity effect in multi-variable regression problems.



To help understand the interconnectivity of high-frequency codes, we directly measure coexistance behavior at first appearances of stories. The chart above shows the network graph of the top 100 high-frequency initial topic codes with high-readership stories (with the retrospective code READ500). Individual codes are plotted as nodes; coexistances between a pair of codes are plotted as undirected edges. The size of the node is proportional to its degree of connectivity; the width of edge is proportional to the coexistance frequency. Using modularity analysis, the codes are found to be partitioned into five sub-communities, as shown by different colors in the chart. Communities appear to form for different reasons. Such formation could be attributable to association with similar wire groups, such as codes with BF (Bloomberg First Word) prefixes and BI (Bloomberg Intelligence) prefixes, or association with specific asset classes (such as OILMARKET—FUTURES—CRUDE triplets) or languages (such as the LLTRANS—JFALL—JPALL—JNEWS quartet).

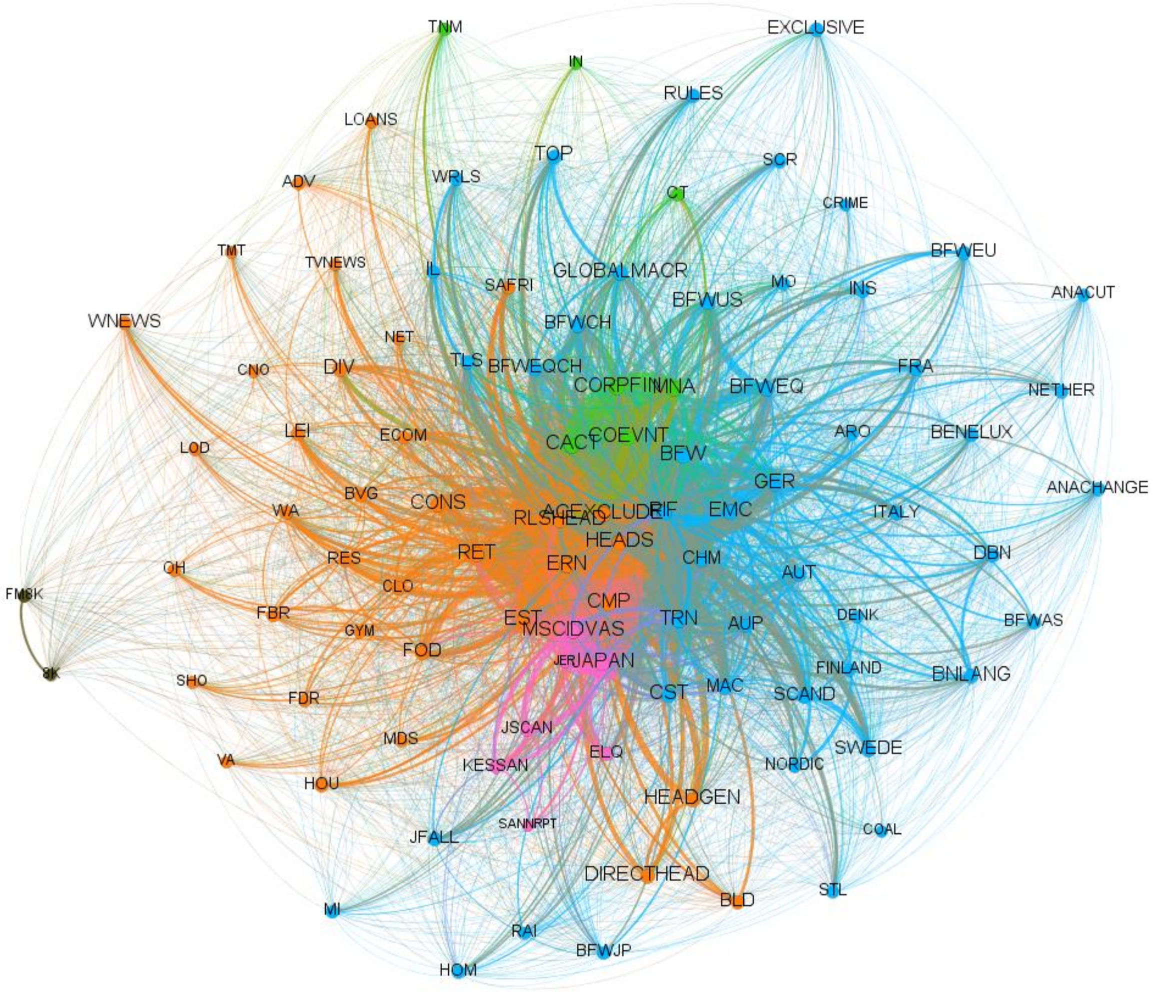
Similar analysis can also be applied to topic codes associated with market-moving stories.



The chart above shows the ranking of initial topic codes for stories with the retrospective market impact code MMN. It shows the top 20 codes ranked by their conditional probabilities. Most of the codes are related to either corporate fundamentals (earnings, M&A, etc.) or timeliness of the news (headline-related).

The following table gives topic code definitions.

Topic Code	Definition
EST	Earnings Estimates
JER	Japan Earnings
RLSHEAD	Headlines on Press Releases
ERN	Earnings
GLOBALMACR	Global Macro
KESSAN	Company Earnings (Japanese)
MNA	Mergers & Acquisitions
HEADS	Headlines
DIRECTHEAD	As-is Headlines Press Releases
ACEXCLUDE	Excluded from Alert Catcher
HEADGEN	Auto-generated Headlines
TOP	Top Stories
BFWUS	First Word U.S. Equities News
FINLAND	Finland
CMP	Computer-generated
JSCAN	Japanese Scanned Material
RIF	RIF Headlines
DBN	Bloomberg News (German)
CHM	Chemicals
CST	Engineering and Construction Services



The chart above shows the network graph of the top 100 high-frequency initial topic codes with market-moving stories with the retrospective code MMN. The codes are found to be partitioned into five sub-communities. Three communities are easy to explain: one related to 8K filings, one related to corporate events and one related to Japanese news content.The remaining two communities are less clear to separate; additional analysis is needed to further disentangle the network.

Final Remarks

This report only scratches the surface of the potential in mining machine-readable newsfeed data. The three examples, while each addressing slightly different questions, combine to paint a coherent picture: The divergence is striking between what people read and what actually moves the market.

News-driven trading strategies have been incredibly efficient in pricing individual breaking headlines into the stock market, with response times shortening from hundreds of milliseconds only a few years ago to less than a millisecond today. This has evolved into a speed game where infrastructure plays a bigger role than research. However, a large area of uncharted territory remains where significant new opportunities can be found by intelligence players. By quantitatively tracking the sequence of related news items, one can construct the news-derived fundamental context critical in the determination of longer-term price movements.

With the continuing commoditization of powerful processers, storage and data science toolkits, reacting instantaneously to huge volumes of real-time news programatically is no longer a pipe dream. In time, machine-readable newsfeeds will become an indispensable part of every financial firm's toolkit.