


How Web Data Powers Predictive Analytics in Finance

 SESAMm **web**hose.io

INTRODUCTION.....	1
Section 1: Beating the Market with Alternative Web Data	2
Why Investment Managers are Leveraging Alternative Web Data	3
How Alternative Web Data Gives Investment Managers an Edge	5
Section 2: The Challenges of Collecting Alternative Web Data	7
Using Pattern Matching and Heuristics to Structure the Web	8
Is Web Scraping Legal?	8
But What About Copyright?	9
Section 3: Case Studies	10
Pfizer – General Analysis.....	10
Pfizer – ESG Analysis	12
Pfizer – Products Analysis	13
A Look Ahead at Alternative Web Data	14

INTRODUCTION

In the past, investment management institutions relied mostly on traditional data to gain an edge in investing. Traditional data ranges from SEC filings to earnings reports and pricing information — any type of data produced by the company itself. The rise of the digital age, however, has opened up new sources of data for investors beyond the scope of traditional data. The seemingly infinite scope of alternative data includes data produced from credit cards, satellites, social media and perhaps most importantly — the web.

With the additional integration of alternative data, investment management institutions and hedge funds in particular that once relied only on traditional data now have an edge in predicting the rise and fall of the markets. As increasing numbers of financial institutions jump on the bandwagon of alternative data, spending on alternative data by trading and asset management firms is set to exceed \$7 billion by 2020.¹

What was only a few years ago a question of when institutions should start using data has shifted to the question of how they can organize and structure these mostly unstructured datasets. And with 4 billion webpages and 1.2 million terabytes of data on the internet estimated to be generated globally by 2025, there is no shortage of web data to sort through. As increasing numbers of investment management institutions incorporate alternative web data into their predictive algorithms, it will change the face of investment as we know it.

This white paper is intended to be a guide for investment management (IMs) institutions to better understand how alternative web data is quickly becoming an essential component for generating alpha and mitigating investment risk. In addition, it explores different models of web data crawlers and what IMs need to look for as they incorporate alternative web data into their predictive analytics models.

¹ Alternative data for investment decisions: Today's innovation could be tomorrow's requirement. Deloitte Center for Financial Services. 2017.

SECTION 1: BEATING THE MARKET WITH ALTERNATIVE WEB DATA

“Your company’s biggest database isn’t your transaction, CRM, ERP or other internal database. Rather it’s the Web itself... Treat the Internet itself as your organization’s largest data source.” -- [Gartner](#)

As previously mentioned, alternative data includes any type of data that is beyond the scope of traditional data: satellite imagery, social media data, and web data (which includes news sites, blogs, discussions and forums) along with credit card data. Alternative web data, which falls under the broader category of big data, is typically unstructured and demands a process for structuring it in order to deliver insights.

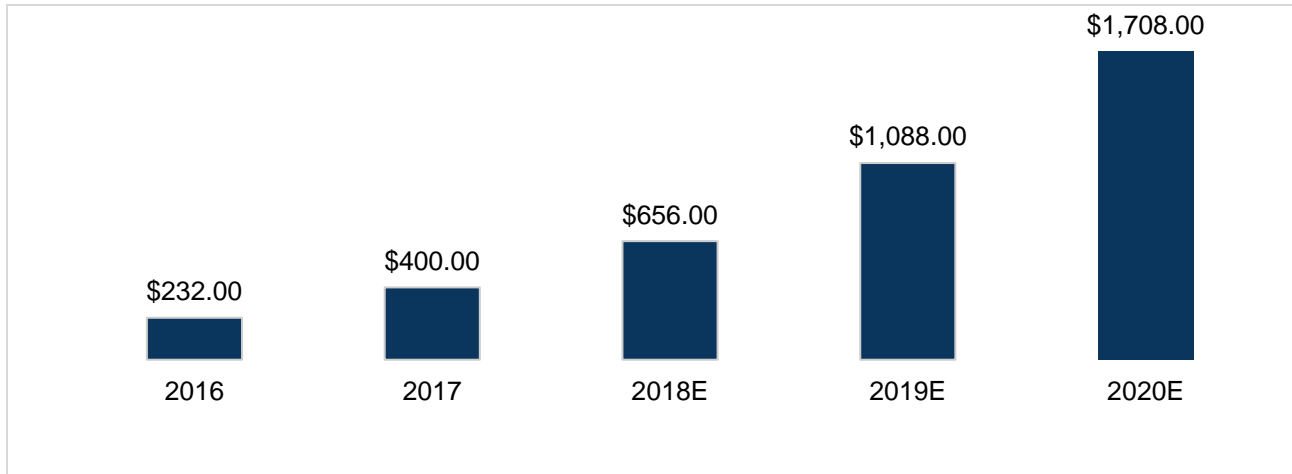
The Most Popular Types of Alternative Data		
Type	Example	Structured or Unstructured
Satellite imagery	Counting the number of oil-storage tanks to calculate inventory, weather data	Unstructured
Social network	<i>Naver.com</i> <i>Baidu.com</i> <i>vk.com</i>	Unstructured
Web data (news sites, blogs, discussions, forums)	<i>Marketwatch.com, Fool.com,</i> <i>Seekingalpha.com</i> <i>finance.yahoo.com</i>	Unstructured
Credit card data	Credit card numbers, dates, financial amounts, phone numbers, addresses, product names, etc.	Structured

WHY INVESTMENT MANAGERS ARE LEVERAGING ALTERNATIVE WEB DATA

Web data, which includes posts from news articles, blogs, discussions and forums offers advantages over other types of alternative data. The scale and diversity of web data is vast enough to offer highly personalized and relevant datasets for specific industries and use cases. Web monitoring services such as Webhose that also offer archived data can be particularly valuable in producing such datasets. In addition, web data such as news and blogs are constantly updated and offer companies the ability to continually keep up with their industry as well as the competition in near real-time, as we will see in a case study later in this document.

Due to the above characteristics, alternative web data can be especially valuable for investment managers in its ability to enhance signal. The first in the financial industry to take advantage of this type of alternative data a few years ago as a whole were hedge funds, but it has since gained traction in the remaining buy-side institutions as well. A [WBR survey](#) in the third quarter of 2018 found that 79% of investment institutions use alternative data. Of the non-users surveyed, 82% of them plan to incorporate alternative data into their trading strategies within the next year. The question is no longer when an investment management firm will start to buy alternative data, but what type of alternative data will help them in their investment strategy.

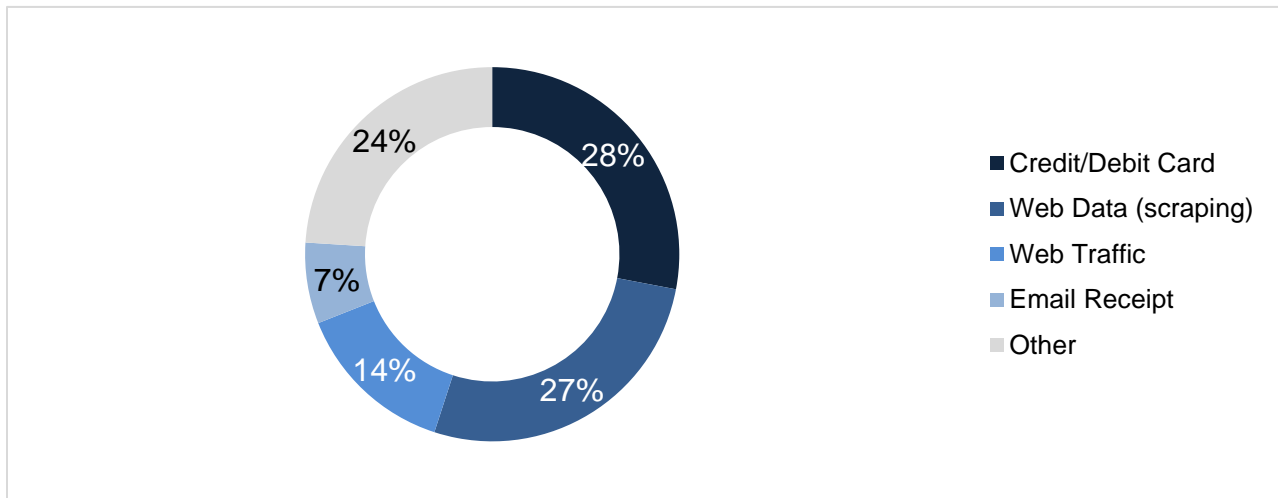
Total Buy-Side Spend on Alternative Data (\$m)



Source: Alternativedata.org

While current figures of buy-side institutions using alternate web data are unavailable, we do know that alternative web data is one of the more accurate datasets for these investment institutions, along with credit card data.

Most Accurate/Insightful Datasets

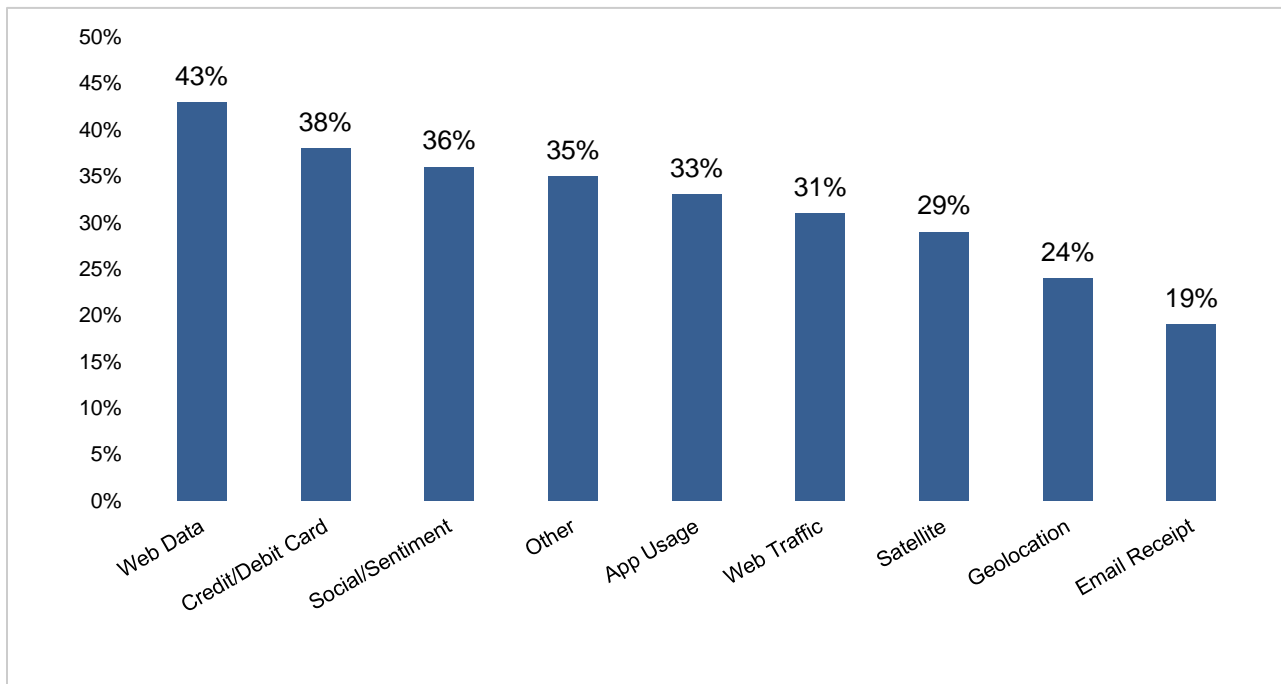


Source: Alternativedata.org

HOW ALTERNATIVE WEB DATA GIVES INVESTMENT MANAGERS AN EDGE

A recent [Greenwich Associates study](#) announced that 72% of investment management institutions reported that alternative data of all types has enhanced their signal. A fifth of respondents reported that alternative data was responsible for them receiving more than 20% of their alpha. And alternative web data is the first type of data these institutions invest in, right next to credit card data.

Funds Using Dataset | % of Funds Using Dataset



Source: Alternativedata.org

The accuracy of alternative web data and its potential for highly personalized datasets is what delivers investment managers that edge.

Here are a few examples of how alternative web data specifically has helped investors to both generate greater alpha while at the same time mitigating risk:

- Almost a week before Gilead Sciences bought Kite Pharma, an immunotherapy cancer treatment company, for \$11.9 billion, it was predicted by AI technology, allowing investors to take advantage of the predicted market movement. Kite Pharma rose 28% that same week of the buyout.²
- During the US government shutdown of 2019, important reports from the US Department of Agriculture ceased to be published, making it difficult for farmers and traders to make important decisions about what crops to grow, trade and sell. With the help of publicly available alternative web data from different government and non-governmental sources in addition to various satellite data, private agencies were able to publish their own crop supply and demand forecasts, which were critical for both farmers and commodities investors.³
- After the Tōhoku earthquake hit Japan, the web monitoring of news stories was able to identify the link between the disaster and the price of the iPad2 by finding an additional news story about the destruction of a major manufacturing plant that produces the NAND flash, a critical component of the iPad2.⁴

² Ram, Aliya and Wigglesworth, Robin. "When Silicon Valley came to Wall Street." Financial Times. October 26, 2017.

³ Meyer, Gregory and Terazono, Emiko. "New Crop Data Advisors Cash in on US Shutdown." March 8, 2019. Financial Times.

⁴ Williams, Janaya. "Solving mysteries using predictive analytics." April 23, 2014.

SECTION 2: THE CHALLENGES OF COLLECTING ALTERNATIVE WEB DATA

“Getting information off the internet is like taking a drink from a firehose.”

– Mitchell Kapor, founder of Lotus Development Corporation and the designer of Lotus 1-2-3, co-founder of the [Electronic Frontier Foundation](#)

Before investors can apply alternative web data for purposes of alpha and risk mitigation, however, they must first collect the data, which presents a range of different challenges.

First, the crawling service used by investment management institutions must be solid and reliable. That means that it must be:

- *Able to easily structure and organize the data.* Most web data are unstructured and difficult to access, since it’s not usually tagged or labeled. Even when organizations have access to unstructured data, they have no way of indexing and structuring it in a manner in which they can deliver insights to their customers. In the financial realm, accessing and organizing data means identifying data that has a low signal to noise ratio. In other words, it means filtering out information that isn’t important to investors (noise).
- *Comprehensive yet accurate.* A superior crawler will collect unstructured data and structure it without missing key details (such as the dates and full titles of a blog post). This is crucial, since inaccurate details plugged into trading algorithms can impact the signal generated, costing investors millions.
- *Open and transparent.* Crawling, or scraping the web involves a number of privacy and safety issues. A data monitoring service that is legal and compliant with a site’s Terms of Service (TOS) is essential—which may or may not allow web crawling or scraping.

USING PATTERN MATCHING AND HEURISTICS TO STRUCTURE THE WEB

The ability of web crawlers to structure web content is critical. Before a system can analyze content, it must first know where the content is. It must be able to map fields and their values. Fields like title, post text, comments, dates, author names etc. must be extracted and tagged.

It's easy to write specific crawlers to crawl a small number of sites that will extract those pre-defined fields from these small number of sources. But when you need data from millions of sources you haven't previously crawled, you need an advanced crawler. Webhose's crawlers use sophisticated pattern matching heuristics to match patterns on newly discovered websites. It leverages knowledge about the structure of previously crawled sites onto sites it has never crawled before. This ability enables structuring the web on scale.

IS WEB SCRAPING LEGAL?

Yes, unless you use it unethically. At Webhose we make sure to crawl only publicly available content. The crawler is very efficient and tries to minimize the resources it takes from any site it crawls.

Keep in mind that an advanced crawler crawl millions of sites. Since it's impossible to contact each site owner and ask for permission, the crawler needs to make it easy to be identified and to give the sites the choice to block the crawler in case they don't wish it to access their content.

These are the steps we take to do this:

- The crawler automatically lets the website it crawls identify it by using a user agent: `omgili/0.5 +https://omgili.com`
- Fixed IPs are used for easier identification on server logs
- The crawler follows standard crawling directive like robots.txt and HTML meta tags

By allowing the crawler to scrape a website, the site owner benefits by:

- Being able to be connected to hundreds of apps, services and marketplaces, which can then link back, potentially sending thousands of relevant visitors to the web property.
- If the site runs advertisements, being noticed and linked to by these services can increase the site's attractiveness to advertisers and the revenue the site generates.
- Instead of being crawled by hundreds of inefficient crawlers downloading the same data repeatedly, companies tap into an already crawled repository to download the data in a machine-readable format.

BUT WHAT ABOUT COPYRIGHT?

The content we crawl is being converted into a M2M (machine to machine) data derivative in a machine-readable format (JSON or XML). It is not presented to humans, and it is used as a technological method to transfer content from point A to point B, just like you would if you'd write your own crawler. It is of course forbidden to use the data crawled and present it as your own, on your own website or resource.

SECTION 3: CASE STUDIES

Let's take a deep dive into some examples of how Webhose's web data can be leveraged with a software such as [TextReveal](#), SESAMm's dedicated Natural Language Processing (NLP) solution for investing.

The goal of this NLP technology is to draw meaningful insights of people from several types of documents (social media, news, forums, etc.) on a daily basis and correlate these insights with market movements. The software aggregates and manipulates textual contents in various forms in order to capture comprehensive information. With new insights and deep information, IMs can enhance their investment decision process.

By mixing several different approaches of one company analysis, we will show how web data and NLP together can provide predictive insights for IMs.

PFIZER – GENERAL ANALYSIS

In the general analysis below, the goal is to forecast a global price trend on the Pfizer stock. We do this by gathering data about the Pfizer company to take a look at global sentiment. This includes articles not only about Pfizer company, but also the CEO and names of Pfizer specific drugs and related underlying trends. In this case, the analysis was performed on 75 million articles from blogs, news articles, discussions mostly in English, but also German, Spanish, French and Dutch.

Using advanced NLP technology, the resulting graph reveals interesting correlations between market data and volumes leading to market movements. The most significant event we see in the general analysis is a global negative sentiment trend in July of 2019 (shown in the blue line on the top) followed by a drop in the market (shown in the red and green bar charts on the bottom). The market later turned bullish, meaning that it is trading upwards.



Figure 1: Pfizer volume, market prices and sentiment data from SESAMm's Markets API dashboard

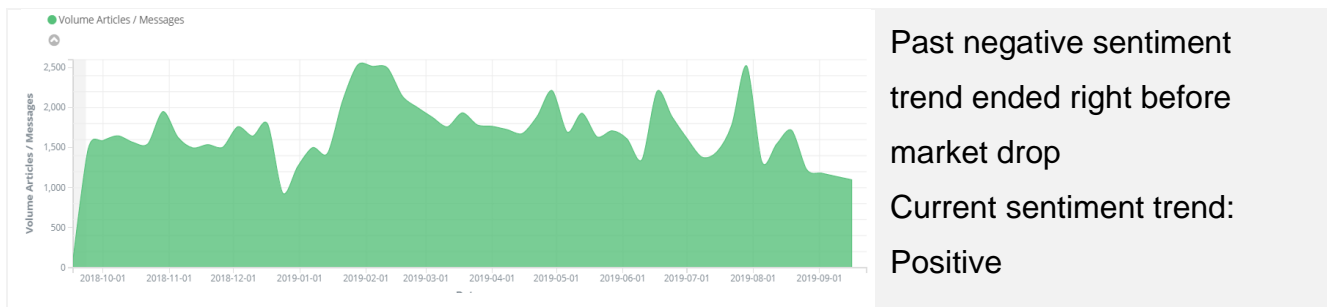


Figure 2: Volume of articles & messages identified about Pfizer based on a small data sample

Most of the additional spikes in the market can be traced to events related to the company, including market events or massive lawsuits or Environmental, Social and Governance (ESG) events.

PFIZER – ESG ANALYSIS

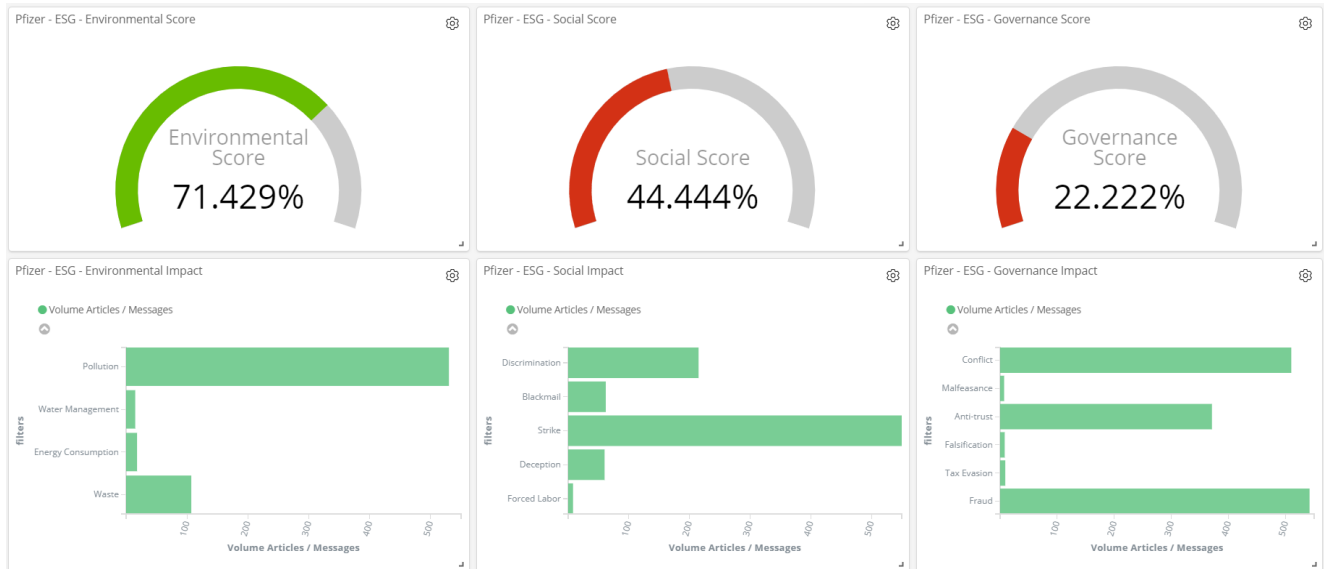


Figure 3: ESG scores based on data sample with specific filters from SESAMm’s Markets API dashboard

Let’s take a closer examination into the Environment, Social and Governance (ESG) risk of the company, which examines the sustainability and social impact of a company for investors.

The NLP technology also automatically detects concepts related to pollution, water management, energy consumption and waste to associate them with high volumes of mention as well as sentiment to create a global Environmental score of the company. After this analysis, we see that Pfizer has a pretty good environmental reputation; it is not mentioned in problems related to the environment. Investors interested in creating a portfolio with a positive environmental impact or low environmental risk should choose Pfizer.

Social and Governance scores, on the other hand, are a different story. The Social score detected significant numbers of references of the company related to discrimination, strikes and blackmail. The Governance score found references to the company and negative topics in the volume of articles and messages, such as conflict, anti-trust and fraud.

As a result of this analysis, investors who wish to distance themselves from stocks with a bad reputation for legal action or treatment of employees should not invest in Pfizer stocks. Let's look at one additional type of predictive insight: Product Analysis.

PFIZER – PRODUCTS ANALYSIS

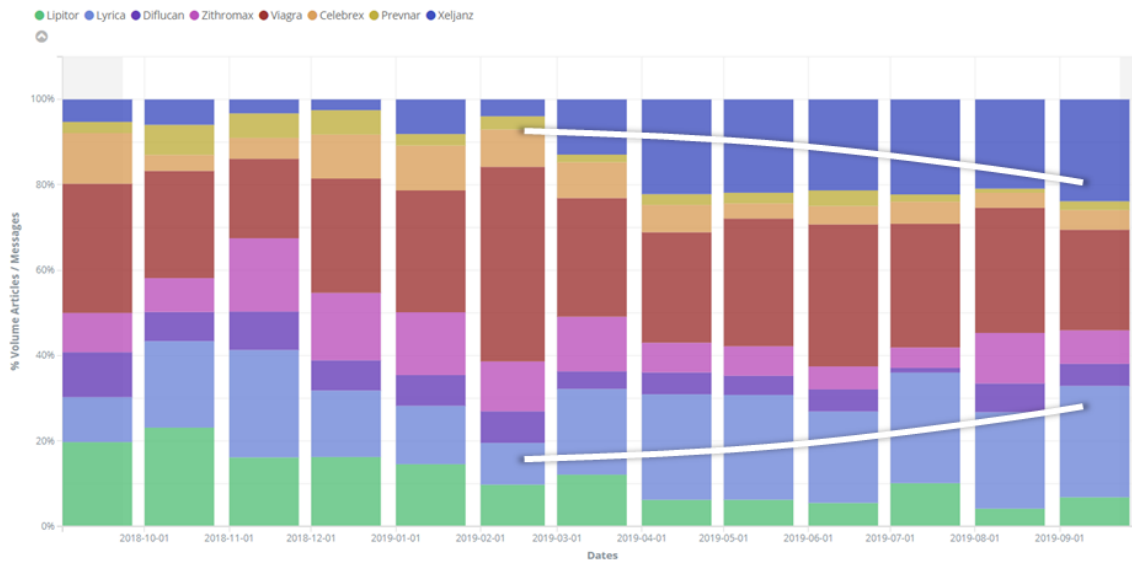


Figure 4: ESG scores based on data sample with specific filters SESAMm's Markets API dashboard

A Product Analysis entails the automatic detection of Pfizer and its products. Let's look at the analysis of a number of its latest products, including Lyrica and Xeljanz. This includes the number of mentions and its correlation with either positive or negative sentiment of these products.

For instance, data can be collected from American consumer blogs to identify whether or not there are side effects of these drugs. For instance, many mentions of bad side effects bode poorly for the drug. We can see that both Lyrica and Xeljanz have a high volume in the growth of mentions of the products, and there is a strong likelihood that these mentions are correlated with high growth. These three analyses of Pfizer, General, ESG and Product are all done by feeding Webhose's high-quality data into SESAMm's advanced NLP TextReveal.

A LOOK AHEAD AT ALTERNATIVE WEB DATA

Although it is being rapidly adopted all along all types of investment management institutions, alternative data is still an emerging trend, with many new applications that will be revealed in the future. The sheer amount of web data presents exciting new types of analysis for investment management institutions such as product, sentiment and ESG analysis. At the same time, however, it presents specific challenges. Along with the biggest challenge of organizing and identifying important data from the massive amount of data present on the web, collecting and crawling alternative web data must also be legal and safe. Inaccurate data can directly influence investment models or trading signals generated, resulting in losses of billions of dollars. Unscrupulous crawling policies can destroy a brand's reputation in an instant.

Along with being able to verify the accuracy and relevancy of data, crawling data on the web must ensure compliance with the Terms of Service (TOS) of the different websites, especially the corporate giants such as Twitter and Facebook. A web data service that offers comprehensive and safe coverage, the ability to scale and full transparency will be able to withstand the inevitable changes in a fast-paced digital world that is radically transforming the field of investment.

Power Investments with Web Data

Want to learn more about how to crawl the web for alternative data?

[Talk to an Expert](#)

About Webhose

[Webhose](#) is the leading data collection provider turning unstructured web content into machine-readable data feeds. It delivers comprehensive, up-to-the-minute coverage of the open web that includes millions of news articles and blog posts in addition to vast coverage of online discussions, forums and review sites in all languages. Webhose also offers a dark web monitoring and data breach detection service that provides coverage of the dark networks and includes millions of sites, files, marketplaces and messaging platforms crawled daily.

About SESAMm

[SESAMm](#) is an innovative fintech company specializing in big data and artificial intelligence for investment. Its team builds analytics and investment signals by analyzing billions of web articles and messages using natural language processing and machine learning. With its NLP platform and quantitative data science platform, SESAMm addresses the entire value chain of alpha research. SESAMm's 40 people team in Paris, New York, Metz, and Luxembourg works with major hedge funds, banks and asset management clients around the world for both fundamental and quantitative use cases.