webhose.io

How DataRobot is Replacing Clickbait with a Viral Score

We've all done it at one point: You click on an article online with a catchy title, only to realize the content of the article wasn't related to the title at all. Unfortunately, many online publishers looking to monetize their content have inadvertently subscribed to a model which encourages clickbait by paying their journalists based on the number of clicks their articles are able to generate. The result, many experts say, is a dumbing down of content across all but the most elite media sites.

What Every Publisher Needs to Ask Before Going Live

How can publishers continue to monetize their content while at the same time encourage sophisticated high-quality articles that tackle more complex issues?

How can advertisers ensure that people who click on content are spending time reading and viewing their advertisements, and not just clicking away immediately?

How can publishers encourage new content that people truly want to read and share with their friends and families?
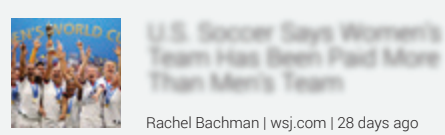
Collecting and Filtering Data for Predictive Modeling

DataRobot helps transform organizations into AI-driven enterprises by implementing machine learning models that solve complex business problems and drive real ROI. They sought an alternative method for publishers to monetize their content, while providing true value to readers. The new approach they chose involved building a highly accurate predictive model which would know which newly published articles will go viral.

But first, they needed data to train their predictive models.

Webhose delivered them the data, providing articles that can be searched according to title, text, publisher, author, data published, sentiment, external links and most importantly for DataRobot, a viral score.


```
performance_score: 5,
    domain_rank: 387,
    social:
        uuid:"5efaa091386d6bfc502742bc058cc6ff9ce1546b",
        url:"http://omgili.com/ri/.wHSUbtEfZRq9N.uhgW_KXYK
        D80roPvSJdmBmpPF6TLQPClkEYNLJ6kDXyejxudVNTZDyP0bOt
        ord_in_thread:0,
    author:"Rachel Bachman",
    published:"2019-07-30T02:48:00.000+03:00",
    title:"U.S. Soccer Says Womens Team Has Been Paid
    text:"Letter from U.S. Soccer president Carlos Cor
        womens teams pay-equity lawsuit The U.S. Soccer Fe
        president Monday saying it has paid the U.S. Women
```

U.S. Soccer Says Women's Team Has Been Paid More Than Men's Team

Rachel Bachman | wsj.com | 28 days ago

The U.S. Soccer Federation released a letter from its president Monday saying it has paid the U.S. Women's National Team more than the men's team in recent years, a move that comes ahead of mediation in the women's team's pay-equity lawsuit against its employer. The pay analysis - which U.S. Soccer president Carlos Cordeiro said was conducted...

For example, one of the datasets collected from Webhose included stories that only had titles mentioning USA, Canada or Mexico that were written in English with a viral performance score greater than zero.

As DataRobot sought to apply Webhose's data towards its predictive model, it discovered correlations between Webhose's score and specific keywords, the title, external links, and a new predictive viral score.

For instance, it found that the text, title of the article and size of the publisher had the greatest influence on the predictive viral score. DataRobot was also able to improve predictive accuracy by using data fields created by Webhose, such as the list of people in the article and categorizing those people according to the sentiment expressed.

Most Important Features in a News Article Linked to Virality

DataRobot also found that certain phrases or words were more likely to drive viral news than others. For instance, words like donald, interview, women and public (marked in red in the image below) were more likely to be found in articles that went viral. In contrast, words like Washington, second, billion, business, and minister were far less likely to be found in articles that went viral. In addition to text, however, the section where the article was listed also influenced the viral score.
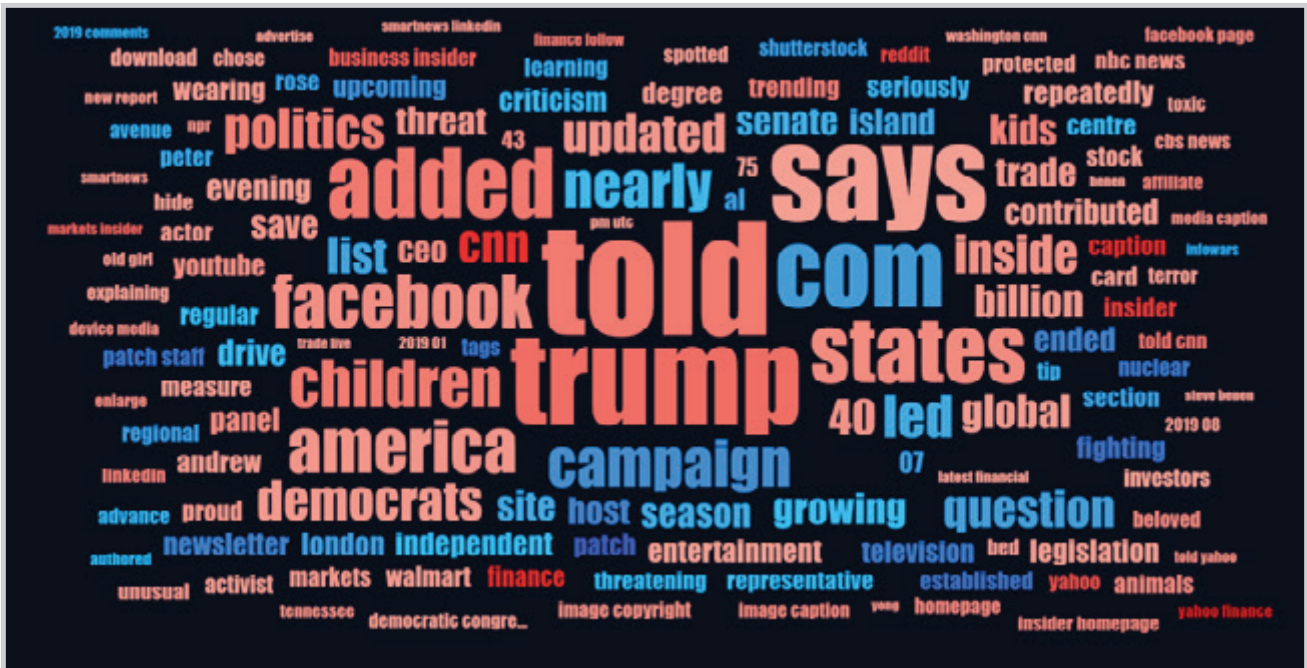
Words and Phrases Linked to Virality

In addition, DataRobot found a correlation between the number of external links and the predicted viral score - up to a dozen external links increased the likelihood of it going viral. (This was the peak; after that, the likelihood of it going viral flattened off).
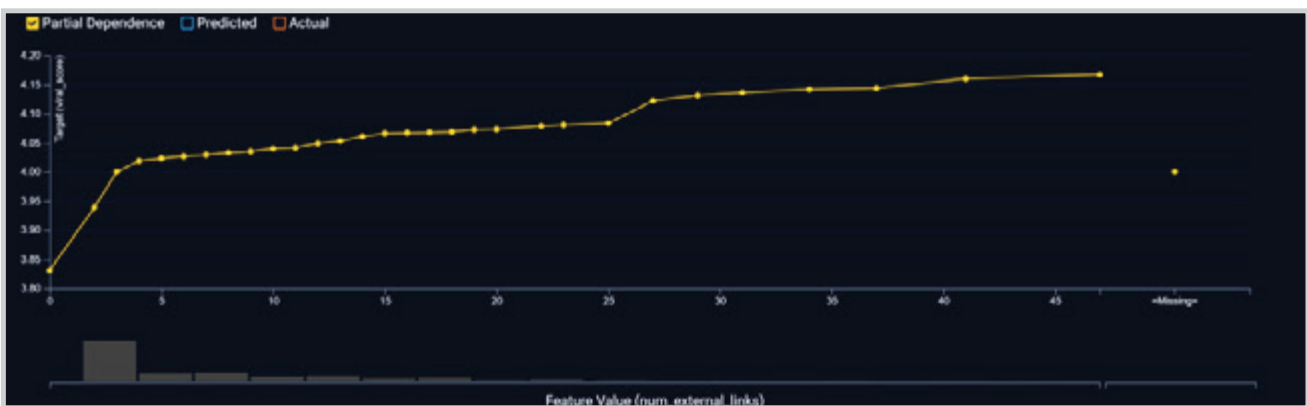
Correlation Between the Number of External Links and the Predicted Viral Score

webhose.io

With Webhose's granular filtering capabilities, including its ability to search according to viral scores based on past history, DataRobot was able to discover the correlations above and build a model that could predict the virality of an article in the future. The result was that DataRobot found that the top 5% of articles for future viral potential were those that it had given an average viral score of more than 9 out of 10. Using this predictive model, publishers can better prioritize which stories should go on the front page. In addition, this alternative monetization model encourages publishers and journalists to focus more on writing and publishing high-quality content rather than on generating clickbait titles for monetary gain.

With the help of Webhose's data, DataRobot was able to develop over 90 different machine learning models and select the one that delivered the most accurate viral score. Relying on manual work to gather the data and develop these different models could take months. But with automated machine learning applications, it took just a few hours.

How to Calculate an Accurate Predictive Viral Score

Step 1

Use Webhose to search for news articles according to advanced search capabilities that include title, text, publisher/author, data published, sentiment, external links, and viral score.

Step 2
DataRobot takes the data and discovers correlations and patterns in stories that go viral and those that do not.

Step 3
DataRobot produces a future viral score based on a function of the story, the way it's written, the name of the publisher, the date it was published, etc, factoring in the correlations it found from Webhose's data.

Step 4
DataRobot is able to automatically generate over 90 different machine-learning models that product a viral score. By comparing the accuracy of each model, it is able to find the one that is the best.

With Webhose's datasets that include clean and structured data from news article, DataRobot can now offer publishers with an alternative to the clickbait advertising model. The future viral score it offers based on its predictive modeling can also pave the way for publishers to offer higher-quality content that tackles more complex issues.

About **web**hose.io

Webhose.io, the brainchild of Ran Geva and Guy Mor, two entrepreneurs with extensive experience in technology, data mining and product development provides on-demand access to web data feeds. Webhose empowers you to build, launch, and scale data-driven operations as you grow. Every web data feed is optimized to deliver up-to-the-minute coverage of a specific content domain including news, blogs, online discussions and forums, and more.

About **DataRobot**

DataRobot is the category creator and leading provider of automated machine learning. Organizations worldwide use DataRobot to empower the teams they already have in place to rapidly build and deploy machine learning models and create advanced AI applications. With a library of hundreds of the most powerful open source machine learning algorithms, the DataRobot platform encapsulates every best practice and safeguard to accelerate and scale data science capabilities while maximizing transparency, accuracy, and collaboration. By making data scientists more productive and enabling the democratization of data science, DataRobot helps organizations transform into AI-driven enterprises.