NewsQ Algorithm Design

Andrew Ouyang, Daniar Tabys, Matthew Gebara, and Neil Patel

What is news?:

News is information that builds upon preexisting knowledge. News should be reliable, relevant, unbiased, and goal-oriented. It is also important to note that the definition of news depends on the audience's background. For example, news in one country may not be relevant for citizens in another country.

Some of the most relevant problems with modern journalism include the focus on sensational headlines and the spread of misinformation. To apply our algorithm to address a specific problem, we decided to choose articles from the United Kingdom as the subject for testing our algorithm. The United Kingdom is notorious for having a sensational tabloid culture that ranks entertainment value above informational content. Our algorithm aims to penalize the use of outlandish headlines and make recommendations based on the content of the article. For the benefits of democracy to be fully realized, the average person must be properly informed, and thus should not have to waste time wading through pages and pages of biased opinions and rumors. In this project we worked to push journalism in the direction of facts, and to keep the world informed with the best information out there.

YOU DONE!
ETON MESS
100 DAYS THAT WILL MAKE OR BREAK HIM
DAILY EXPRESS
HANG ON TO YOUR HATS...HERE COMES BORIS!
The Daily Telegraph
I'm the dude*
THE TIMES
Johnson goes to work
Like a slumbering giant, we're going to rise and ping off the guy ropes of self-doubt and negativity... and unite this amazing country
The Sun
NAA NAA NAA NA-NA-NA-NAA
HEY DUDE! DON'T MAKE IT BAD
The Guardian
An ambition is fulfilled. But what next for Britain?
Daily Mirror
BORIS JOHNSON, PRIME MINISTER
IT'S REALLY NOT FUNNY ANY MORE
Daily Mail
NOW BRING US SUNSHINE!
in crucial first 100 days of power

Our project:

Our response to the challenge to design an algorithm that will rank ~50 news articles based on a variety of factors and weighing scales is as follows. We will be using our algorithm to rank news articles from the United Kingdom based primarily on their reliability, relevancy, impartiality, and central objective. Specifically, we will be using the following criteria to rank the news articles. Each factor is weighted according to the rank of importance, with high rank of importance corresponding to higher weighting.

Criteria (high to low importance):

1. **Sentiment analysis ("Vibe Check") -** This will be the metric that measures the bias that the writer adds to an article. The overall sentiment of the article will be compared to the general sentiment of the topic. If there is a large mismatch between the two, points will be deducted.
2. **Headline relevance -** This is a measure of the relevance of article content in relation to the article headline. If the keywords in the main body of the article are represented in the article title, the headline is a good representation of the article contents and would receive a higher rating.
3. **Relevance of topic -** This measures the relevance of article content in relation to the search topic. A stronger relationship between search topic and article content will result in a higher rating.
4. **Date relevance -** This is a measure of an article's relevance to current events. A more recent article is more likely to have relevant information. A publish date that is closer to the specified latest date would result in a higher rating.
5. **Quotation frequency -** This is a measure of the reliability and credibility of the writer's claims. Generally, citing sources often will result in a higher rating. Although this would not be completely accurate in promoting the most credible articles, the weighing takes this into account.

Sentiment analysis

Using an API from DeepAI, we assigned each sentence in a news article to a category of “Very Negative”, “Negative”, “Neutral”, “Positive”, or “Very Positive”. To transform and normalize this data set, we assigned each category to a number in the range 1-5 respectively. The mean of the data set for each news article is compared to the general sentiment of the topic, where a smaller difference would result in a smaller penalty and a larger difference would result in a larger penalty.

To find the general sentiment of the topic which provides a baseline of comparison, we took the mean sentiment of the entire set of news articles. Instead of using a “Neutral” sentiment of 3 for the baseline, we used the mean of the entire data set to account for skewed sentiments for specific topics.

Headline relevance

Using keywords from the article contents provided by the Media Cloud API, we compared the relevancy of the headline in relation to the article. The keywords are identified through a representative sample of all articles specified by the topic, location, and time period. Our goal is to filter out headlines that are unrelated to the keywords that denote the user’s interests. This also serves to discourage sensational headlines (seen in the U.K.’s tabloid culture) that have little relation to the article’s contents. We expect headlines to serve as a summary of article content rather than as just marketing.

Relevance of topic

Looking through the text, we found the number of times the topic appears in the document. We expect that an article that is more relevant to the topic of the search would mention the topic more often. For longer articles, we expect more instances of the topic, but this has been normalized to prevent larger documents from receiving a more favorable rating.

Date relevance

An outdated article is likely less reliable and relevant than a current up-to-date article. Keeping this in mind, recent articles are prioritized by penalizing those farther from the specified latest date. This factor, however, has been weighted less as part of the algorithm since the other factors are more significant and the fact that some articles can be considered reliable and relevant irregardless of the publish date.

Quotation frequency

In general, we expect that instances of quotations in an article denote where the author cites a source, either formally or informally. By using our algorithm to return a count of quotation mark usage normalized to the total word count, we hope to place greater priority for recommendation on articles that have a higher frequency of quotations. To prevent excessive penalizing of articles that have few quotations, we put in place a system that rewards the use of quotations rather than penalizes a lack of quotations. This system also sets a maximum reward for heavy quotation usage to prevent inflating the rating of particular news articles.

Next Steps:

To improve upon our current project, we would fine-tune the algorithm and add more factors to make the article ranking more accurate. For example, improvements could be done by checking the amount of ad content on a news site or by utilizing a U.K. fact checker. We could also potentially utilize machine learning by creating a training set and applying attributes known to make an article more or less reliable. To make the algorithm more useful for a wider audience, we can create a front end interactive user-friendly interface to allow users to browse unbiased and trustworthy articles of their interest.