B | Africa Growth Initiative
at BROOKINGS

Exploring the media sentiment around Africa: Data and methodology

Data

We compiled all articles mentioning the word “Africa” in the headline or the first paragraph for three newspapers: *The Wall Street Journal*, *The New York Times* and *The Economist*. This data was accessed via Lexis Nexis for the years:

	Years	Total Articles
WSJ	1980-2015	4593
NYT	1985-2015	45934
Economist	1975-2015	10557

The three newspapers were chosen for their broad readership, *The Wall Street Journal* and *The New York Times* are two of the three most circulated newspapers in North America.[1] *The Economist* was also chosen for its close to 2.5 million print readers in North America, which is roughly the same circulation as the NYT and WSJ.[2] Additionally, the WSJ and *The Economist* are marketed as periodicals focused on investment and economics; *The New York Times* is not marketed as such. We wanted to see if there was any difference in coverage focus as a result.

The time span chosen reflects the availability of data through the LexisNexis system. The search criteria (“Africa” in the headline or lead paragraph) is intended to be as broad as possible while maintaining relevance to the main research question at hand. The headline and abstract serve as the reader’s introduction to the article; even those who do not finish the article feel their sentiment and tone. By restricting our search to the word “Africa” in these two areas, we measure sentiment towards Africa as opposed to that of individual countries. Further research on individual countries would be a natural next step.

Methodology

Using the LexisNexis database, we retrieved the headlines and lead paragraphs of all Economist, *The New York Times* and *The Wall Street Journal* articles from 1990 to 2015 that mentioned the term “Africa” in the headline or first paragraph.

In the first phase, we wanted to identify trends in how often different countries were mentioned.

In the second stage, we formulated a measure of sentiment by comparing the headlines and lead paragraphs of articles to the Harvard IV-4 semantic dictionary. This process measures the frequency of positive and negative content in the text. The Harvard IV-4 dictionary is designed for use across a variety of contexts; in order to acquire a more domain-specific measure of sentiment, we also ran our articles against the Loughran and McDonald financial dictionary.[3]

[1] http://www.wsj.com/articles/SB10001424052702304178104579535822452265610

[2] http://economistgroupmedia-1530222749.us-east-1.elb.amazonaws.com/sites/default/files/TEWA_JJ15_V2_0.pdf

[3] wjh.harvard.edu/~inquirer

Finance research on textual analysis has gained prominence in recent years. Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Laughran and McDonald (2011) have all used the Harvard IV-4 General Inquirer Dictionary (H4D) to examine the effects of word sentiment on financial results.

Laughran and McDonald (2011) also introduced a dictionary of their own, intended for use exclusively on financial documents. Their research indicates that the Laughran and McDonald dictionary (LMD) has a better correlation with financial metrics.

In order to determine sentiment, we used a classic text classification approach, as used in the papers above.

Both the H4D and the LMD have lists of words classified as positive and negative. We iterate through each article's headline and lead paragraph and count the number of H4D-positive words and LMD-positive words, along with the corresponding negative words.

We then define a sentiment index metric as follows, where x represents the total number of positive words that appear in each article, and y the corresponding negative amount:

$$s.i. = \frac{x-y}{x+y}$$

By definition, $s.i. \in (-1,1)$. Other authors, like Tetlock et. al. (2008), have used only the number of negative words divided by the total number of words. Our approach was chosen to ensure that positive words are explicitly taken into account in the analysis.

In the second phase, our research focused on the difference in sentiment when including both the headline and lead paragraph and when only analyzing the headline. We organized articles by year to get a macroscopic view of sentiment. We find evidence across all three periodicals that headlines are significantly more negative than headlines and lead paragraphs taken together.