Criteria, Scoring and Collection Methodology

Article Criteria:

The dataset contains 1,000 articles. Articles were only included in the dataset if they satisfied the following criteria:

- Written in English language
- Published between January 1, 2021 and December 31, 2021
- No word limit
- Contains substantively vaccine-related content:
    - Vaccine-related content satisfies at least one of the following: 1). mentions vaccines in the context of how they work, their effectiveness, what activities can be done before/during/after vaccines, who can take vaccines, what counter indications exist, whether vaccines are safe; 2). mentions vaccines in the context of vaccine policy, 3). mentions vaccines in the context of thoughts, behaviors and attitudes.

Parameters of dataset:

- Article URL: Link to full text of article
- Score: Interval between 1-5
- Confidence: Percentage between 0-100
- Source: Database/List used to surface article
- Local: Binary indicator for whether the article came from a local (1) or non-local source (0)
- US: Binary indicator for whether the article came from a US (1) or non-US source (0)

Overtone Scoring Definitions:

This dataset labels each article with Overtone's "Score" and "Confidence" measures:

Scores (integervalue, 1–5): The informational quality score is a measure of the value that has been added to an article by its author to inform their readers. The text is rated by the Overtone algorithm according to the presence or lack of journalistic signals. These signals include the following:

- Original reporting: Does the article provide new information or is it taken from elsewhere? Is the article vague or does it provide evidence to support its statements?
- Good sourcing: Does the article cite interviews, data or documents? Does it simply republish information that is published online as a release or social media post? Does it include multiple different sources or does it rely on only one?
- Exploration of ideas: Does the article focus solely on one story angle or does it provide context to inform the reader of the related issues? Are statements and evidence analyzed and explained in a nuanced manner?

Scoring Categories:
1 - Aggregation, Status Update, Non-News
These stories convey news but with limited explanation of the meaning, such as a press release or sourcing reliant on aggregation from a company, other outlets or social media posts. Articles that score 1 may also be non-news, such as purely personal blog posts or recipes.

2- Bare Bones, One Source
These stories provide some context but with limited original reporting or value-add, such as a report that is largely aggregated or based on one source, such as local police. These stories convey one angle to the news without consideration of the larger picture.

3- Daily Story
These stories provide at least some value-add, such as citing an expert to comment on news broken elsewhere. They may provide context on the news with some original analysis but do not delve into multiple angles. Event-based reporting may often fall into this category.

4 - Added Value, Original Reporting
These stories delve deeper than one angle, juxtaposing the ideas of one expert to others or additional facts. Beyond reporting they may show reporters work looking into data or facts and analyzing them for the reader. Reported feature stories are often in this category.

5 - Enterprise, In-Depth, Investigative
These stories show the presence of most or all of the quality journalistic signals. This includes information from solid sources such as interviews or documents, original reporting and exploration of ideas. These stories are likely to be bringing new information to light or deeply exploring a known issue rather than retreading previously reported material.

Confidence (percentage): The confidence accompanying an Overtone score is a measure of how confident the Overtone English model is in its output. This is measured from 0 to 100, with lower confidence scores indicating either an unfamiliarity with the type of text provided, or the possibility that the text could have also received another score.

Dataset Curation Methodology

Step 1: Identification of sources for articles

In order to ensure that a variety of vaccine-article types and sources were included in the dataset, the curation team derived articles from three main resources: 1) The Vaccine Safety Net (VSN), 2) A list of sources from Wikimedia DC, the regional outreach organization of Wikimedia Foundation and ARTT project partner, and 3) The Overtone.ai database. These three resources helped ensure that there was a variety of article types in the dataset.

The first resource was a list of sources provided by the VSN. The VSN is a global network of websites established by the World Health Organization, that provides reliable information on vaccine safety. A second resource was a list of sources provided by project partners from within the Wikipedia community. As one purpose of this dataset was to provide a range of article quality, the sources provided by these partners were considered both reliable and unreliable, as according to Wikipedia's Reliable Sources guidelines. A final resource was the Overtone.ai database. The team drew 4,000 pieces of content from the Overtone database, which is a large repository of English-language content used by Overtone for media monitoring.

Step 2: Initial Collection

Once the main sources were determined, the team started a structured search for articles. The website Buzzsumo was utilized for surfacing relevant articles from the VSN and Wikimedia sources. Search used keywords "vaccine" AND "source URL," and were filtered by: language (English), content by journalists, content by all publishers, and date. Results were sorted by total engagement ranking, and up to 10 of the top articles were included for each source. Articles from the Overtone database that received various scores (note: for more information on scoring, see section below) between January 1 and December 31, 2021 and mentioned the keyword "vaccines" were also added to the dataset.

Step 3: Scoring

After the initial collection, the preliminary list of approximately 1000 articles was scored with Overtone.ai software. The collection was then checked to ensure there was a representational amount of content in the following three categories: *Score* (At least 150 1s, 250 2s, 250 3s, 250 4s, and 100 5s); *Outlet type* (At least 100 articles local outlets); *Region* (At least 100 articles from non-US sources)

Step 4: Refinement

The collection of articles was then further refined. First, the team added in additional articles, taken from the Overtone database. These articles were mainly scored as 2s and 3s. Second, to reach the maximum article limit of 5000, articles were sourced from Event Registry, a news search and analytical platform. The search utilized the Event Registry "concept" of "Covid-19 vaccine", and was filtered to include only articles that were from the year 2021, English language articles from the U.S. and U.K.. The results of this search were randomized, and the top 5,000 were collected. Finally, the team manually checked each article to ensure each was composed of "sufficiently vaccine-related" content.

Step 5: Secondary Scoring (repeat of Step 3)

Step 6: Creation of Final Dataset

In order to create the final list for the dataset, the team filtered the spreadsheet based on the individual scored articles that were marked "Yes" for sufficiently vaccine related. An HTTP check was done to remove bad links (i.e. 404s), and the remaining articles were moved to the final list. If there were excess articles (i.e. above 250 2s), the articles were randomized and the articles below the cut-offs were removed and stored in a separate file.