The authors bear all responsibility in case of violation rights. Upon acceptance, the dataset will be publicly released on GitHub under the CC-BY 4.0 license, and will be updated if videos are removed from their respective sources by uploaders.

Dataset and code are included in the supplementary materials. These will be uploaded to a public GitHub repository upon acceptance.

**Datasheet (in the Gebru et al. format[1])**

**Motivation**

**For what purpose was the dataset created?** The dataset was created for the purpose of creating AI systems that can interpret the wide range of news content available online, in a variety of languages.
**Who created the dataset, and on behalf of which entity?** Omitted.
**Who funded the creation of the dataset?** Omitted.

**Composition**

**What do the instances that comprise the dataset represent?** Each instance is a video and text description describing a current event.
**How many instances are there in total?** 2,396 videos.
**Does the dataset contain all possible instances or is it a sample of instances from a larger set?** The dataset does not contain all possible instances – it is a small sample of the possible instances in the domain.
**What data does each instance consist of?** Each instance consists of a video and text description (denoted as a URL and description label) and their corresponding language, current event name, article URL(s) (in English and the video language if it is not English), and English event article excerpts.
**Is there a label or target associated with each instance?** Yes, videos are labeled with their current event, language, whether their text description is the video description or video title, and event article links. Videos are not cleaned or preprocessed for the dataset.
**Is any information missing from individual instances?** No.
**Are relationships between individual instances made explicit**? Videos/text belonging to the same current event are labeled as such.
**Are there recommended data splits?** No.
**Are there any errors, sources of noise, or redundancies in the dataset?** Yes, in Section 3 we explain that identical videos are removed but similar video content may exist. Rare annotation errors may exist and the dataset will be updated on GitHub if any errors are found.
**Is the dataset self-contained, or does it link to or otherwise rely on external resources?** The dataset links to videos and external text articles.
**Does the dataset contain data that might be considered confidential?** No.
**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** Yes, the dataset includes videos of disasters, political protests, etcetera. The videos have not been checked for possible offensive or harmful content.

---

[1] Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. "Datasheets for datasets." Communications of the ACM 64, no. 12 (2021): 86-92.

**Does the dataset identify any subpopulations?** No.
**Is it possible to identify individuals, either directly or indirectly from the dataset**? It may be possible. Videos may identify individual people, but the owners of the videos chose to upload this content publicly and have the choice to take it down at any time (consequently removing it from the dataset as well). Individuals may also be identified via the supporting text articles, which are linked in the dataset.
**Does the data contain data that might be considered sensitive in any way?** Yes, see above.

**Collection process**
**How was the data associated with each instance acquired?** Data collection details are provided in Section 3 of the paper.
**What mechanisms or procedures were used to collect the data?** See Section 3.
**If the dataset is a sample from a larger set, what was the sampling strategy?** See Section 3. Videos were selected manually based on relevance.
**Who was involved in the data collection process and how were they compensated?** Data collection was done by the authors.
**Over what timeframe was the data collected?** The data was collected between November 2022 and May 2023.
**Were any ethical review processes conducted?** No.
**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources?** Data was collected via video sharing sites.
**Were the individuals in question notified about the data collection?** N/A.
**Did the individuals in question consent to the collection and use of their data?** N/A.
**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** N/A.
**Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?** No.

**Preprocessing/cleaning/labeling**
**Was any preprocessing/cleaning/labeling of the data done?** Videos are labeled with their current event, language, whether their text description is the video description or video title, and event article links. Videos are not cleaned or preprocessed for the dataset.
**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?** The video URLs and article URLs to the source material are included in the dataset. English article excerpts used as event queries are saved in the dataset.
**Is the software that was used to preprocess/clean/label the data available?** No, but necessary steps are detailed in the paper.

**Uses**
**Has the dataset been used for any tasks already?** Yes, it has been used for the experiments detailed in the paper.
**Is there a repository that links to any or all papers or systems that use the dataset?** No.
**What (other) tasks could the dataset be used for?** The dataset could be used for a variety of tasks including video retrieval, report generation, question-answer pair generation, etc.

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** Possibly – this varies with the application it is used for.
**Are there tasks for which the dataset should not be used?** Judgment should be used for tasks that may directly affect real people.

**Distribution**

**Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** Yes, it will be publicly available.
**How will the dataset be distributed?** Via GitHub.
**When will the dataset be distributed?** The data is released on GitHub and will be finalized before the camera-ready version is submitted.
**Will the dataset be distributed under a copyright or other intellectual property license, and/or under applicable terms of use?** The dataset will be licensed under CC-BY 4.0.
**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** The videos belong under their respective sources' licenses.
**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

**Maintenance**

**Who will be supporting/hosting/maintaining the dataset?** The first author will maintain the dataset and it will be hosted on GitHub.
**How can the owner/curator/manager of the dataset be contacted?** They can be contacted via email.
**Is there an erratum?** An erratum will be included in the repository if necessary.
**Will the dataset be updated?** The dataset will be updated to correct any errors that are found.
**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** Yes, all the visual data will be retained as long as the individual creators of the videos used keep their content publicly available. If an individual creator removes their content from the third-party site, it will no longer be accessible for the dataset.
**Will older version of the dataset continue to be supported/hosted/maintained?** No.
**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Yes, they are free to fork the GitHub repository.