Russia-Ukraine News Reports Dataset

Nityam Pareek
210150021

April 2024
GitHub

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset is created to enable explorers and data scientists to analyze the progression of political ties between Russia and Ukraine from various angles. In pressing issues which affect the world, it is important to have a dataset which observes the events from many different sources and time frames.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Web scraping techniques have been used to compile the dataset from the Global Database of Events, Language, and Tone (GDELT) - the world's largest, most comprehensive, and open database of human society, continuously monitoring global news media in over 100 languages.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This dataset has been created as a mini-project for the course 'Multimodal Data Processing and Learning - II' conducted by the Mehta Family School of Data Science and Artificial Intelligence at the Indian Institute of Technology Guwahati. There is no monetary grant for the same.

Any other comments? No further comments.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances that comprise the dataset primarily consist of data about news articles related to Russia-Ukraine relations including the URLs, source country, language, and date between January 2017- December 2023. The dataset also includes the tone and number of published articles according to day.

How many instances are there in total (of each type, if appropriate)?

The dataset contains a total of 16.8k news articles between the years 2017-2023. Average tones and article count data contains 2551 rows.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset does not contain all possible instances, it is a sample from the vast amounts of news articles published on the topic. The sampling is not random as articles are sampled according to an 'influence' metric by the API. However, we can safely say that the sample is a representative of the larger set because it contains data from 121 unique countries. However, it is biased towards countries like Turkey, Ukraine, Russia, and USA which publish news articles more frequently and have larger reach than other countries' media.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The dataset consists of the following three csv files:

1. **news_articles.csv** : This file contains the news article (URL), date of entry into the database, title, image (URL), domain, language and source country of the news article.

2. **average_tone.csv** : This contains information about the average tone of news articles published about the topic on a particular day on a scale of +5 (very positive) to -5 (very negative).

3. **num_articles.csv** : This contains the information about the total number of published articles, along with the number of articles related to out topic on a particular day.

Is there a label or target associated with each instance? If so, please provide a description.

There is no explicit label or target associated with each instance. The dataset has been created in an open ended fashion to encourage explorers to engage in a wide variety of NLP tasks.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Some individual instances may contain missing or redacted images, because all news articles do not publish images. Some urls may also be broken because the articles were redacted.

Are relationships between individual instances made explicit (e.g.,

users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

All instances are related to the Russia-Ukraine relations. They can be grouped by source country, language, or date.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are no such recommended splits for this dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A source of redundancies in this dataset may be the inclusion of those articles which are not exactly relevant to our topic. Some media houses may publish articles in the name of 'Russie-Ukraine' simply for a larger reach, without much contribution. Some articles may also be conveying the same information, but in a different language leading to redundancies.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset heavily relies on external URLs for articles and corresponding images. As is the case with news articles, there is no guarantee that they will exist or remain constant in the future. No archival versions have been created. There are no known restrictions in accessing the URLs at the time of creation of the dataset. However in future it is upto media houses whether or not they put those articles behind a paywall.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

Since all data is in the public domain, there are no confidential sources.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Since the data is on Russia-Ukraine relations, it may contain articles and photographs which talk about war and suffering. This might cause anxiety to some people.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the dataset does not target individuals or subsections of a population.

Does the dataset identify any subpopulations (e.g., by age, gender)?

If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Not applicable.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Not applicable.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

Not applicable.

Any other comments? No further comments.

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The official API was used to query the GDELT database.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sampling strategy was automated by the API, which used an 'influence' metric in the backend to query the database.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The GDELT project is supported by Google Jigsaw. Exact compensation details are unknown.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The API allows access to data post 1 January 2017. This matches the timeframe of the instances in the dataset. Old news articles have been deliberately included to provide explorers with a time series like data.

Were any ethical review processes conducted (e.g., by an institutional

review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.
Information about such a review process is unknown.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No, the dataset does not target individuals or subsections of a population.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
Not applicable.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
Not applicable.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
Not applicable.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
Not applicable.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
Not applicable.

Any other comments?
No further comments.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
No data preprocessing was required at the time of creation of this dataset. Mechanisms to update the dataset for use at a future data are made available on GitHub.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Not applicable.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Not applicable.

Any other comments?
No further comments

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
Any present usage of this dataset is not known.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
Not applicable.

What (other) tasks could the dataset be used for?
The dataset can be used in a variety of NLP tasks like generating articles given time and country, predicting the average tone of a set of articles, predicting the domain of an article, comparing similar documents across languages, 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? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset should not be used in any situation where we are seeking to establish a notion of 'ground truth'. As is the case with media, the biases of humans will be reflected in the dataset along with any exaggeration done to make the article more 'interesting'.

Any other comments? No further comments.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset will be publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset will be posted on GitHub. It does not have a DOI as of the date on this report.

When will the dataset be distributed?
The dataset is hosted on GitHub as of the date on this report

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
The dataset is published under the MIT license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No such restrictions have been imposed as of the date on this report.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No such restrictions are known to apply as of the date on this report.

Any other comments?

No further comments.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will be hosted on GitHub. While I am not sure how often I will be able to maintain it, I will make the code publicly available for whoever wants to create an updated version of the dataset for themselves.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

I can be contacted at the E-Mail address ndnityam@gmail.com.

Is there an erratum? If so, please provide a link or other access point.

There is no erratum as of the date on this report.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

The dataset will not be updated from my side unless requested.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, the dataset does not target individuals or subsections of a population.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

This dataset contains news reports in a time series format, which will not lose their relevance with time. Any future time series maintenance is not guaranteed.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Mechanisms to fetch and clean new data have been made publicly available on Github.

Any other comments?

No further comments.