EveOut: Reproducible Event Dataset for Studying and Analyzing the Complex Event-Outlet Relationship

Swati
swati@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Tomaž Erjavec
tomaz.erjavec@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

ABSTRACT

We present a dataset consisting of 77,545 news events collected between January 2019 and May 2020. We selected the top five news outlets based on Alexa Global Rankings and retrieved all the events reported in English by these outlets using the *Event Registry API*. Our dataset can be used as a resource to analyze and learn the relationship between events and their selection by the outlets. It is primarily intended to be used by researchers studying bias in event selection. However, it may also be used to study the geographical, temporal, categorical and several other aspects of the events. We demonstrate the value of the resource in developing novel applications in the digital humanities with motivating use cases. Website with additional details is available at *http://cleopatra.ijs.si/EveOut/*.

KEYWORDS

Dataset, News Event Analysis, Event selection bias, News coverage

1 INTRODUCTION

News outlets are constantly faced with the task of selecting events they will report on, dependent on the perceived interest of the event to their readership. This can be driven by various factors, such as the geographical origin of the event, involvement of well-known persons, etc. Such selection requires monitoring of current affairs to determine their news value for the outlet.

Machine learning tools may help outlets to deal with the large numbers of events, help them explore strategies for selecting publishable events, and build dedicated decision support systems for this task. The effectiveness of these systems depends on the availability of news event collections complemented by relevant event details such as date, category, country of occurrence, brief description, etc.

In this paper we introduce EveOut, the first large publicly available data set of 77,545 English news events with a variety of features collected between January 2019 and May 2020. It includes events in eight different categories of news, i.e. business, politics, technology, environment, health, science, sports, and arts-and-entertainment. We hope that EveOut will encourage publishers and others involved in the news production process to develop tools to enhance digital journalism. The data set would also allow researchers from digital humanities to study and analyze the relationship and impact of different features on the selection of events by the outlets.

1.1 Contributions

The paper makes the following three contributions to science:

- The dataset generation scripts, which provide a structured reproducible approach to building a publicly available dataset of news events with varied features. This will not only speed up the development of future versions of EveOut, but will also help to create custom datasets with the desired outlets and features.
- The compilation of EveOut, a novel dataset with a rich range of event features and spanning multiple news categories.
- Identification of possible use cases intended to facilitate the creation of tools to improve digital journalism and to help researchers study the complex relationship between events and news outlets.

2 DATASET

Several news outlets may cover a single world event as a story in a variety of different ways. A collection of one or more stories, all of which describe the same world event, is referred to as an *'event'* in the entire paper. In the following subsections, we define our data generation process and provide statistics on the resulting dataset.

2.1 Data Source

We use **Event Registry**[1] [4] as the data source which monitors, collects, and provides news articles from news outlets around the world in over 30 languages. It also identifies the major incidents reported in the articles and aggregates them into clusters known as events. For example, *"missiles launched by Iran at US forces in Iraq"* is an event reported across the globe in over 3,200 news articles.

To construct an event, Event Registry follows a series of steps. News aggregation is the first step in which RSS feeds are constantly monitored for new articles. The next major step is the semantic event information extraction, which retrieves information from the articles in a structured way to be used in subsequent steps. Clustering algorithms are then used to group articles that describe the same event. In the last step, the article clusters are marked as events and are annotated with rich metadata such as a unique id to track the event coverage, categories to which it may belong, geographical location, sentiment, etc. As a result, its extensive temporal coverage can be used effectively to study the complex correlation between events and news outlets.

[1] https://eventregistry.org

Information Society 2020, 5–9 October 2020, Ljubljana, Slovenia

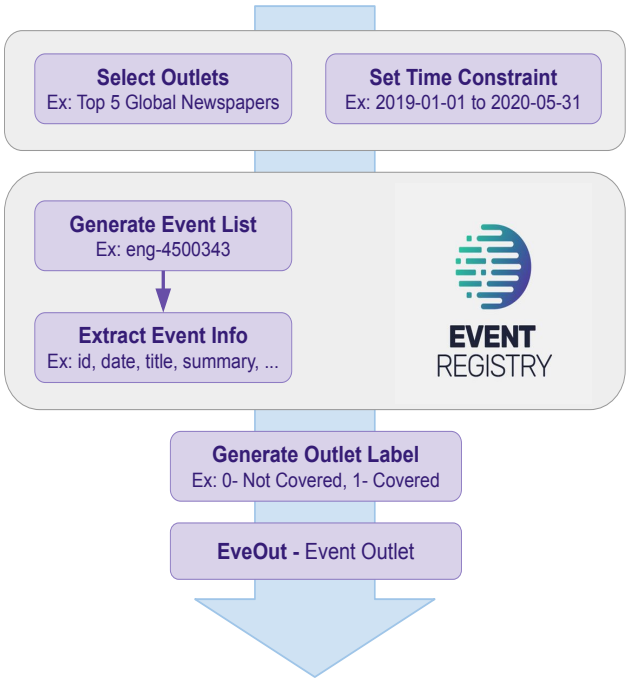

Figure 1: EveOut dataset generation process.

Table 1: Description of the dataset attributes.

Attribute	Description
uri	a unique event identifier
title	title of the event in English
event_date	date in yyyy-mm-dd format
sentiment	event sentiment
categories	event categories
loc_country	country where the event occurred
loc_continent	continent where the event occurred
total_article_count	total number of articles published
article_count	total number of articles published in English
summary	summary of the event
outlet_list	list of outlets that reported the event

2.2 Data Generation Process

To generate the dataset we adopted an automated approach which is depicted in Figure 1. We use Event Registry API to collect event related information mentioned in Table 1. The script is designed to simplify the release of future versions and to be able to replicate the process of generating custom datasets. The outlined process is the result of the resource's core requirement to best address the potential use-cases referred to in Section 4.

For data generation, we first selected the top five news outlets based on Alexa Global Rankings [2]. We then used an explicit temporal query (Q_t) to retrieve all events in all news categories from the Event Registry API. $Q_t = \{Q_{text}, Q_{time}\}$ consists of the text component Q_{text} and the time component Q_{time}. Next, we set the time limit $Q_{time} = [Q_{sd}, Q_{ed}]$ for extracting events that occurred within the specified time where, Q_{sd} = '2019-01-01' and Q_{ed} = '2020-05-31' signify the event's start date and end date. Since the outlet's event selection policy may change over time, we selected this time frame as recent data tends to be more reliable in predicting event coverage patterns. We then set $Q_{text} = \{Q_{out}, Q_{lang}, Q_{cat}\}$ where, Q_{out} = {'*nytimes*', '*indiatimes*', '*washingtonpost*', '*usatoday*','*chinadaily*'}, Q_{lang} = {'*eng*'}, and Q_{cat} = {'*politics*', '*business*', '*sports*', '*arts and entertainment*', '*science*', '*technology*', '*health*', '*environment*'} represent the outlets, languages and news categories respectively.

From the extracted event list, we first excluded events that were not covered by any of the selected outlets. We then extracted individual outlets from the event's outlet list and created a column in the dataset to represent each of them. We use a binary scalar value to indicate whether the outlets covered the event or not. The event coverage by the outlets is not uniform, which can be visualized in Figure 2.

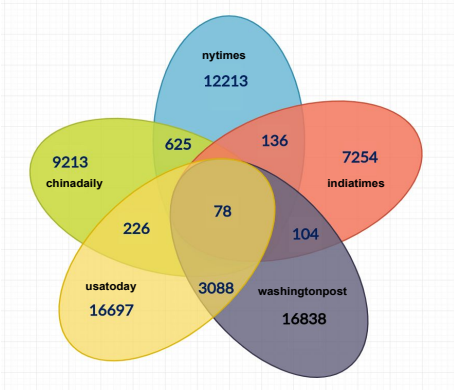

Figure 2: Distribution of event coverage by the outlets.

3 AVAILABILITY

The GitHub repository containing the scripts is available at ***https://github.com/Swati17293/EveOut***. To facilitate discoverability and preservation, the full data set is archived as an online resource at ***https://doi.org/10.5281/zenodo.3953878***. EveOut is available in three common formats (JSON, XML, and CSV) for direct download and use. The documentation meets the requirements of the *FAIR Data principles*[3] with all necessary metadata defined. Under the *Creative Commons Attribution 4.0 International license*, it is freely available to make it reusable for almost any purpose. A separate web page with detailed statistics and illustrations can be found at ***http://cleopatra.ijs.si/EveOut/*** for in-depth analysis.

3.1 Reusability

The resource is currently being used for individual projects and as a contribution to the project's deliverables of the Marie Skłodowska-Curie CLEOPATRA Innovative Training Network[4]. A major part of this project aims to provide a temporal, cross-lingual analysis of concepts around different events, exploring how language impacts the mediatic narratives built by the media. It also aims to analyse news reporting bias and multiple media

[2]https://www.alexa.com/topsites/category/Top/News/Newspapers

[3]http://www.nature.com/articles/sdata201618/

[4]http://cleopatra-project.eu/

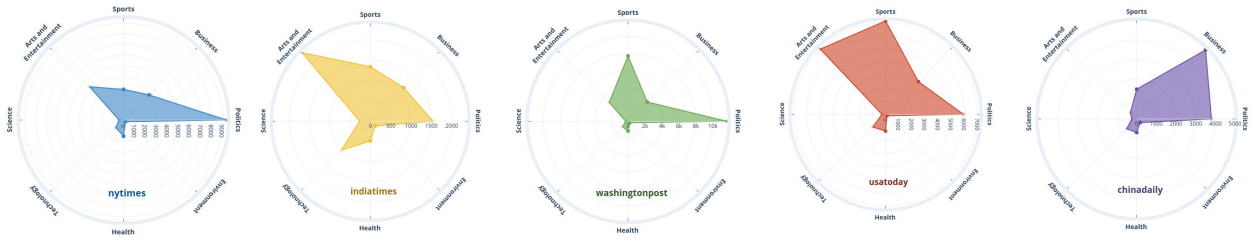

Figure 3: Overview of the category-wise event coverage by the outlets.

narratives which would enable to filter out appropriate information which then will be used to build information representation tools. Since EveOut serves as the basis for the study and analysis of events and their attributes, it is ideally suited to the project needs.

4 POTENTIAL USE CASES

4.1 Examine Event-Selection Bias

It is important for a journalist to know which event is worthy enough to be published. Even readers would be interested to know the factors that affect this selection. An automated solution can be devised using EveOut to provide an overview of the event and to visualize differences in coverage.

4.2 Outlet Prediction

EveOut is designed to predict the likelihood of an event being covered by the outlet. It would enable the publishers of the outlets to assess the significance of the event. In addition, it may also be used by independent editors who prefer to report on events covered by mainstream outlets.

5 STATISTICS AND ANALYSIS

In this section we provide further information about the data contained in EveOut, focusing explicitly on the distribution of events between the outlets.

With regard to the distribution of event categories covered by the outlets, as shown in Figure 3, *'politics'* is the most common category, while *'environment'* is the least common category. It is also worth noting that each outlet focuses on the different categories of events aside from *'politics'*. For instance, *'indiatimes'* focuses more on events related to *'arts and entertainment'*, whereas *'chinadaily'* tends to cover more *'business'* related events.

As far as the coverage of the event over time is concerned, it is also inconsistent as depicted in Figure 6. Furthermore, the event-coverage of *'usatoday'* and *'washingtonpost'* is slightly inconsistent. It is also interesting to note the sharp decline in coverage by *'usatoday'* in *'Aug* 2019' and by *'washingtonpost'* in *'May* 2020'.

The drop in the graph for washingtonpost in *'May* 2020 is due to its event preference. It is evident from washingtonpost's radial graph in Figure 3 that its coverage is biased towards politics and sports. These two categories alone represent around 50% of events in the dataset. However, this percentage dropped to 40% in *'May* 2020 and, as a result, the coverage of washingtonpost dropped significantly. Increase of event coverage in *'Mar* 2019 is also attributed to the fact that about 56% of events were from these two categories. In nutshell, if the outlet favors a certain category of events and, in a specific time frame, and events of that category are high/low than usual, it will be reflected in the outlet's coverage pattern.

Figure 4 reveals that instead of favoring events with neutral sentiment, outlets tend to favor events with positive sentiment. In addition, event coverage by *'usatoday'* and *'washingtonpost'* is quite diverse with respect to sentiments.

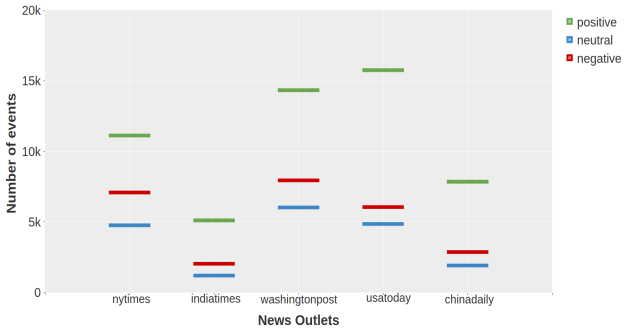

Figure 4: Distribution of event coverage by the outlets with respect to sentiments.

In terms of the sentiments used in each category as plotted in Figure 5, it is worth noting that *'technology'* and *'sports'* events are mostly positive.

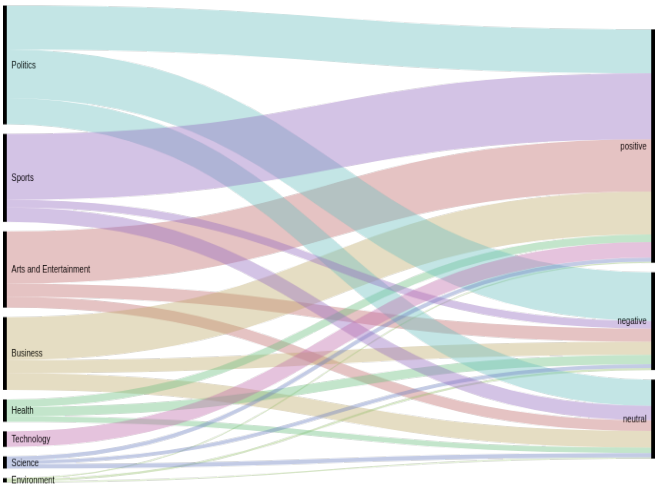

Figure 5: Distribution of category over sentiments.

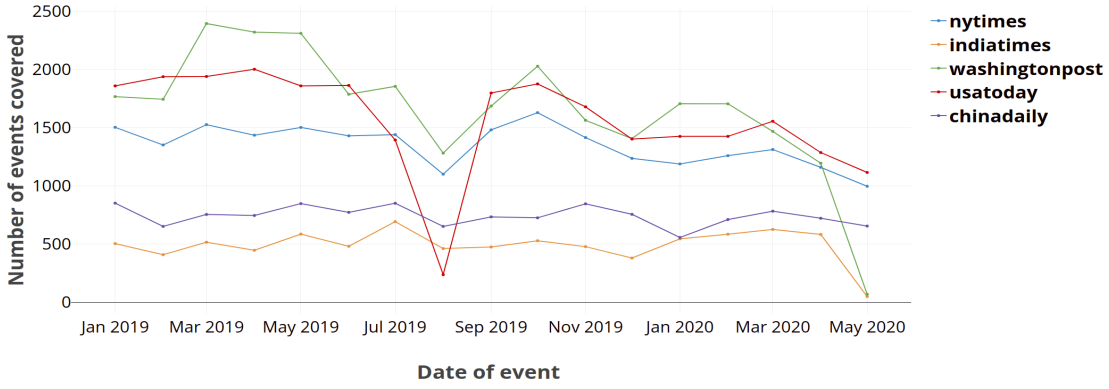

Figure 6: Distribution of the event coverage by the outlets over time.

6 RELATED WORK

There are a number of datasets that focus on news articles [7]. As far as the availability of event-centric datasets is concerned, there is a scarcity of publicly available datasets. There are few related research on the event data [3, 1], but the extracted/generated datasets for the experiments is also not publicly accessible.

GDELT [5] is the most popular, very large and publicly available event-oriented news dataset. It contains data in multiple languages from a wide range of online publications. It's collection of world events is centered on location, network and temporal attributes. There is no attribute defining the outlet list for the event in the dataset. As a result, there is a lack of knowledge essential to the analysis of the event-outlet relationship that is the foundation of our dataset.

In addition, the existing event datasets [6, 2] are category-dependent (*politics/healthcare/disaster etc.*) which renders them useful for specific research purposes only. Therefore, by providing a generalized event-centric news dataset, EveOut addresses the stated dataset bottleneck.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the EveOut dataset, which covers events reported by the top five global news outlets for over 17 months. We have ensured that the dataset complies with the *FAIR principles*. In conjunction with the data set, we provide the source code for reproducing the dataset with varied features. For instance, it is possible to generate a reduced version of EveOut, focused on just one category, say *'politics'*. Specific outlets, dates, and languages can also be specified in accordance with the requirements. We illustrate potential use cases to show how the dataset could be used to study the pattern of event coverage of an individual outlet and to predict whether or not the outlet will cover a specific event. Researchers from digital humanities can also use it for an in-depth analysis of complex event-outlet relationships. In the future , we intend to extend the dataset to include events described in different languages.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 812997.

REFERENCES

[1] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. 2018. Selection bias in news coverage: learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, 535–543.

[2] Cindy Cheng, Joan Barceló, Allison Spencer Hartnett, Robert Kubinec, and Luca Messerschmidt. 2020. Covid-19 government response event dataset (coronanet v. 1.0). *Nature Human Behaviour*, 1–13.

[3] Felix Hamborg, Norman Meuschke, and Bela Gipp. 2018. Bias-aware news analysis using matrix-based news aggregation. *International Journal on Digital Libraries*, 1–19.

[4] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[5] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: global data on events, location, and tone, 1979–2012. In *ISA annual convention*. Volume 2, 1–49.

[6] Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: an armed conflict location and event dataset: special data feature. *Journal of peace research*, 47, 651–660.

[7] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, X. Xie, Jianfeng Gao, Winnie Wu, and M. Zhou. 2020. Mind: a large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3597–3606. DOI: 10.18653/v1/2020.acl-main.331. https://www.aclweb.org/anthology/2020.acl-main.331.