EveOut: an Event-centric News Dataset to Analyze an Outlet's Event Selection Patterns

Swati, Dunja Mladenić and Tomaž Erjavec
Jožef Stefan Institute and Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
E-mail: swati@ijs.si, dunja.mladenic@ijs.si, tomaz.erjavec@ijs.si

Keywords: dataset, news event analysis, event selection bias, news coverage, gatekeeping bias, outlet prediction, digital humanities

Received: January 10, 2021

Automation of computational models to study the structure of events and their value to news outlets is an effective way to understand event-outlet relationships. However, the scarcity of publicly available, comprehensive event-centric news datasets restricts the implementation of such models. To overcome this bottleneck, we collected seventeen months of event data using Event Registry to generate EveOut, a publicly available event-centric news dataset. To conduct statistical analysis, we first select five English-language and three in Slovenian-language news outlets. We then retrieved all the events covered by them and used it to document the prevalence of geographical, temporal, categorical, and several other aspects of the event selection bias by these outlets. We illustrate the significance of our dataset in the field of digital humanities by identifying a motivating use case. The dataset is publicly available from the dedicated website `http://cleopatra.ijs.si/EveOut/`*, which provides a detailed description of the fields, usage information, and a link to the GitHub repository.*

Povzetek: V prispevku predstavljamo EveOut, javno dostopno množico podatkov, zgrajeno na osnovi dogodkov, o katerih poročajo mediji. EveOut je zasnovana za pomoč pri analizi in razumevanju zapletenega odnosa medijev do poročanja o posameznem dogodku. Zgrajeno množico podatkov smo tudi uporabili za raziskovanje geografskih, časovnih, vsebinskih in drugih vidikov pristranskosti nekaj izbranih poročevalcev pri izbiri dogodkov, o katerih poročajo.

1 Introduction

News outlets are constantly confronted with the task of selecting events to be reported on. This selection is based on the newsworthiness of an event which can be defined by the presence or absence of several news values such as the inclusion of the power elites, the relevance, and popularity of the topic, etc. Determining the news value for an outlet may result in a selection bias, also known as gatekeeping bias [7]. A journalist, for instance, is more likely to report on an event that includes fresh data on an existing and trending event.

Gatekeeping bias can be significantly reduced by studying and analyzing the correlation and impact of different features on the selection of events by the outlets and then using the knowledge to automate the event selection process. Computational models for the study of complex event-outlet relationships can help explore the strategies for selecting publishable events and automating the event selection process. However, to stimulate the development of these models, the availability of data on news events and their relevant details is necessary.

In this paper, we present EveOut, the first, large, publicly available event dataset generated by leveraging the events collected using EventRegistry [4]. The resulting dataset, called "EveOut", consists of $81,562$ news events in English and Slovenian language, with a varied range of features retrieved for the period between January 2019 and May 2020.

We hope that EveOut will serve as a benchmark dataset to study the event-outlet correlation and help mitigate the impact of implicit bias present in the production and reporting process. We also hoped that it will encourage the publishers and others involved in the news production process to develop tools to enhance digital journalism and facilitate research in this field.

The contributions of this paper are as follows:

- We present EveOut, a publicly available novel event-centric news dataset generated with a wide range of features using the EventRegistry platform.
- We provide flexible dataset generation scripts that facilitate the generation of custom versions of EveOut with the required features.
- We identify a potential use-case 'outlet prediction task'. For the task, we then illustrate how conditional probabilistic models can be used to estimate the correlation between outlets.
- We present a detailed statistical analysis with respect to multiple event features to compare, contrast, and infer the coverage pattern of the selected news outlets publishing in English and Slovenian languages.

2 Related work

A number of datasets are based on news articles but, to the best of our knowledge, only a few datasets that explicitly focus on event-centric data have been proposed. GDELT (Global Data on Events, Location, and Tone) [5] is a CAMEO-coded [6], open, and large-scale news dataset that tracks the news media around the world in multiple languages. The articles are then compiled into a list of events, for rich and insightful event analytics. However, there is no attribute in the dataset that would specify the outlet(s) for the event. As a result, there is a lack of information in GDELT that is crucial to the study of the event-outlet relationship that forms the foundation of our dataset.

In terms of availability, publicly available event datasets are scarce. There is some related research on event data [3, 1], but the datasets extracted/generated for the experiments are not publicly available. Besides, the majority of the existing event datasets [2] are category-dependent (*politics, healthcare, disaster, etc.*) which renders them useful for specific research purposes only. EveOut addresses these bottlenecks by introducing a generalized publicly available event-centric news dataset.

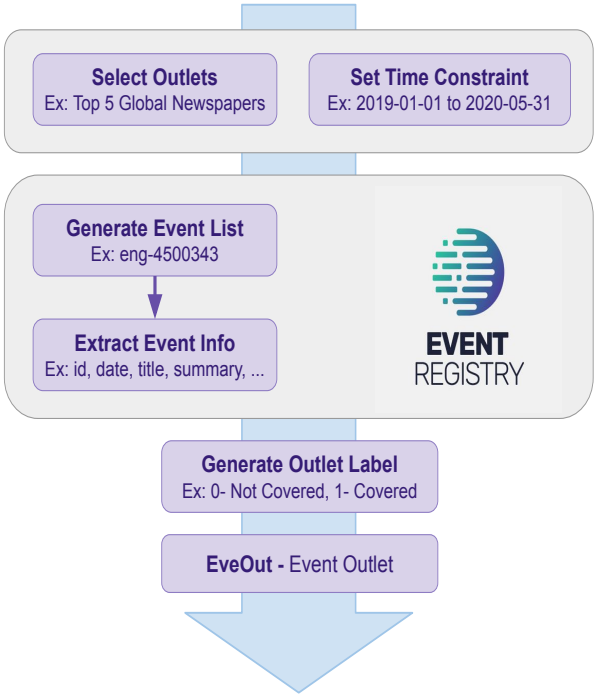
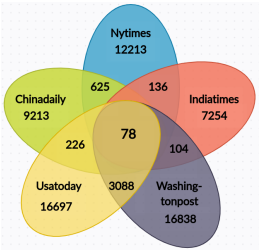
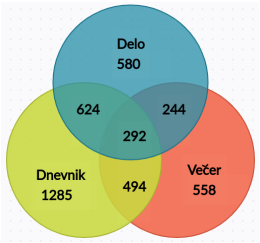


Figure 1: EveOut generation process, composed of user selection of the outlets and the time period, automatic extraction of the data from Event Registry and labeling of the extracted data.



(a) English news outlets.



(b) Slovenian news outlets.

Figure 2: Distribution of event coverage by the outlets.

3 Data description

3.1 Raw data source

Event Registry[1] [4] monitors, collects, and delivers news articles from news sources around the world in more than 30 languages. It extracts semantic information from the articles and if the same event is described in multiple articles, it aggregates them into clusters using several clustering algorithms. These article clusters are referred to as events. For instance, *"Trump threatens to shut down social media firms"* is an event recorded internationally in more than 1,220 news articles. Each event is then annotated with various metadata, such as a unique id to track the coverage of the event, topic, categories to which it may belong, geographical location, sentiments, etc. As a result, its large-scale temporal coverage can be used effectively to study the event-outlet relation.

3.2 Data generation process

For the data generation process, as depicted in Figure 1, we first selected five English and three Slovenian news outlets (for the sake of simplicity, we refer news outlets publishing in English/Slovenian language as English/Slovenian news outlets throughout the paper). We selected these outlets following the work in [8] which is based on Alexa Global Rankings of top news outlets.

We then used an explicit temporal query (Q_t) to retrieve all events in all news categories using 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}[2], Q_{cat}[3], and Q_{lang}[4] represents the list of out-

[1] https://eventregistry.org
[2] https://eventregistry.org/documentation?tab=suggSources
[3] https://eventregistry.org/documentation?tab=suggCategories
[4] https://github.com/EventRegistry/

Attribute	Description
uri	a unique event identifier
title	event title in the specified language
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 the specified language
summary	summary of the event
outlet_list	list of outlets that reported the event

Table 1: Description of the dataset attributes.

P(O_1|O_2)	Nytimes	Indiatimes	Washingtonpost	Usatoday	Chinadaily
Nytimes	1.00	0.09	0.28	0.24	0.19
Indiatimes	0.03	1.00	0.03	0.03	0.01
Washingtonpost	0.33	0.09	1.00	0.26	0.19
Usatoday	0.27	0.09	0.25	1.00	0.13
Chinadaily	0.10	0.01	0.08	0.06	1.00

(a) English news outlets.

P(O_1|O_2)	Delo	Dnevnik	Večer
Delo	1.00	0.33	0.33
Dnevnik	0.51	1.00	0.49
Večer	0.30	0.29	1.00

(b) Slovenian news outlets.

Table 2: Conditional probability of an event to be covered by an outlet (in rows), provided it is covered by another outlet (in columns).

lets, categories, and languages respectively.

For English news outlets, we set $Q_{out} = \{$'$nytimes.com$', '$indiatimes.com$', '$washingtonpost.com$','$usatoday.com$','$chinadaily.com.cn$'$\}$ and $Q_{lang} = \{$'eng'$\}$ and we set $Q_{out} = \{$'$delo.si$', '$dnevnik.si$', '$vecer.com$'$\}$ and $Q_{lang} = \{$'slv'$\}$ for Slovenian news outlets. We fixed $Q_{cat} = \{$'$news/Politics$', '$news/Business$', '$news/Sports$', '$news/Arts\ and\ Entertainment$', '$news/Science$', '$news/Technology$', '$news/Health$', '$news/Environment$'$\}$ to represent the news categories. If an event falls into more than one category, it is labeled with multiple categories.

We first excluded events from the extracted event list that weren't covered by any of the selected outlets. We then extracted individual outlets from the event's outlet list and generated a column in each dataset to denote individual outlets. We used a binary scalar value to indicate whether the outlets covered the event or not. Table 1 describes the attributes of the generated dataset. From Figure 2, it is apparent that the event coverage by the outlets is not uniform.

4 Availability and reusability

For ease of discovery and preservation, EveOut is archived as an online resource at ***https://doi.org/10.5281/zenodo.3953878***. It is well documented in accordance with the requirements of the *FAIR Data principles*[5] and is freely accessible under the *Creative Commons Attribution 4.0 International license* to make it reusable for nearly any purpose. For dataset regeneration, the GitHub repository at ***https://github.com/Swati17293/EveOut*** gives the source code of the collection process. For an in-depth analysis, a separate web page with detailed statistics and illustrations can be found at ***http://cleopatra.ijs.si/EveOut/***.

The resource is currently being used in several studies within a larger research project[6]. 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. Since EveOut serves as the basis for the study and analysis of events and their attributes, it is ideally suited to the project needs.

5 Potential use case - outlet prediction

Outlet Prediction is the task of estimating the probability that an event will be covered by an outlet. In addition to allowing the publishers of the outlets to evaluate the significance of the event, this task is intended to benefit independent editors who prefer to report on events covered by mainstream outlets. It can be best assessed by calculating the conditional probability P of an event covered by an outlet O_1 given that it is already covered by another outlet O_2 using the following equations.

$$P(O_1|O_2) = \frac{P(O_1 \cap O_2)}{P(O_2)}, \text{ if } P(O_2) > 0 \quad (1)$$

$$= 0, \text{ if } P(O_2) = 0 \quad (2)$$

Table 2a shows that apart from *'Indiatimes'* and *'Chinadaily'*, rest of the outlets tends to overlap each other in terms of event coverage. It is also interesting to note from Table 2, that among the listed outlets, the likelihood of any outlet to cover an event, given that it is already covered by any other outlet is higher (higher P values) for Slovenian outlets.

Unlike the selected English outlets which are supposedly global, the selected Slovenian outlets are the major outlets in Slovenia which is a small country. This difference influences the coverage pattern of the outlets, which reveals how regional priorities affect the event selection process. For instance, $P(Dnevnik|Delo) = 0.51$ and $P(Dnevnik|Večer) = 0.49$ which is quite high as compared to the others which indicates that if an event is covered by either *'Delo'* or *'Večer'* it is highly probable that it will be covered by *'Dnevnik'*.

6 Statistics and analysis

The statistical analysis of our dataset with regard to the distribution of events between the outlets is summarized and visualized in this section.

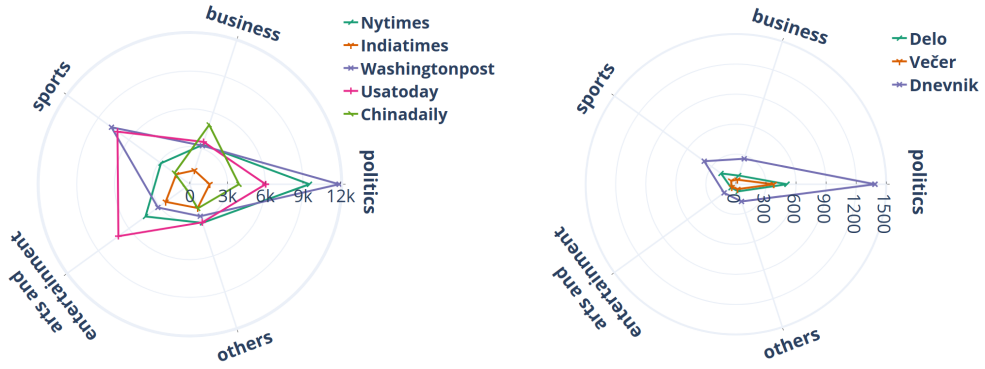
Figure 3a and 3b represents the distribution of event categories covered by the English and the Slovenian news outlets. It is evident from the distribution that each English news outlet focuses on a different event category other than *'Politics'*. For instance, *'Indiatimes'* focuses more on events related to *'Arts and Entertainment'*, whereas *'Chinadaily'* tends to cover more *'Business'* related events. In contrast to English outlets, the event coverage by Slovenian outlets is similar in addition to *'Politics'* focusing on *'Sports'* and to some extent on *'Business'*.

By plotting the proportion of event coverage over time, as shown in Figure 4, the pattern of event coverage by the outlets can be better visualized. In particular, *'May* 2020' contrasts the percentage of event coverage by the English and Slovenian news outlets. Moreover, unlike other news outlets, coverage of events by *'Usatoday'*, *'Washingtonpost'*, and *'Večer'* is somewhat inconsistent. A substantial decline in the coverage of *'Washingtonpost'* in *'May* 2020' is also noteworthy in the graph. It is due to its event preference which is evident from its radial graph in Figure 3a. Its coverage is skewed towards *'Politics'* and *'Sports'* which alone represents around 50% of events in the dataset. However, this percentage dropped to 40% in *'May* 2020', and as

event-registry-python/wiki/Supported-languages

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

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



(a) English news outlets.

(b) Slovenian news outlets.

Figure 3: Category-wise distribution of event coverage by the outlets. (Category 'others' includes: environment, health, science, and technology)

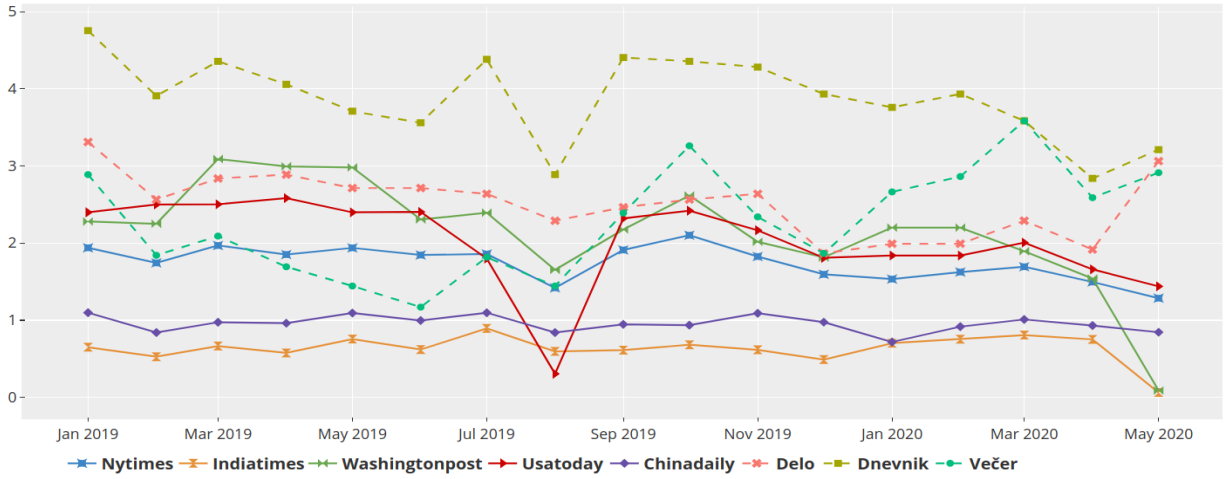


Figure 4: Distribution of the percentage of event coverage by the news outlets over time.

a result, its coverage declined substantially. In a nutshell, if the outlet favors a certain category of events and, in a specific time frame, events of that category are higher/lower than usual, it would be reflected in the outlet's coverage pattern.

Figure 5 reflects the inclination of the news outlets towards geographical bias which indicates that they prefer to cover events relevant to the geographical area in which they are based.

7 Conclusions and future work

In this paper, we presented a novel event-centric dataset EveOut for the study and analysis of complex event-outlet relationships. We also provide flexible data generation scripts, to speed up the development of future versions of EveOut. We also mentioned a potential use case to illustrate how the dataset could be used to study the event coverage patterns of the outlet and to estimate the correlation between the outlets using conditional probabilistic models.

We also conducted a statistical study to compare and contrast five English and three Slovenian outlets to examine their event selection patterns. We found that *'Politics'* is the most popular category, while *'Environment'* is the least popular category covered by the outlets. We also identified that news outlets, as expected, tend to cover geographically relevant events. In particular, we discovered that if the outlet favors a certain category of events and, if in a specific time frame, events collected of that category are higher/lower than usual, then this is reflected in the outlet's coverage pattern.

Although several features, such as event description, have not been analyzed in our study, it is expected that these features will also help to identify the inherent bias present in the event selection process. We hope that our dataset will not only help to discover and interpret event selection bias but will also help researchers to develop tools to enhance digital journalism.

Different news outlets may have different policies for selecting events. For example, some news outlets may want to publish only the top events of the day, while others may want to include exclusive global events. As part of our future work, an automated solution could be developed using

(a) English news outlets.

(b) Slovenian news outlets.

Figure 5: Distribution of country-wise coverage of events by the outlets. Notice the higher coverage density in (a) The USA, India, and China (b) Slovenia.

EveOut to provide an overview of the event and to visualize the differences in coverage, as it is important for journalists to know which event is worthy of publication and which factors influence the selection process.

In the future, it would also be interesting to have a distribution of articles with positive and negative sentiment for specific events and outlets. This would reveal not only the outlet's political orientation but also the editorial's overall attitude.

Acknowledgement

This work was supported by the Slovenian Research Agency under the project J2-1736 Causalify and co-financed by the Republic of Slovenia and the European Union's H2020 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*. International World Wide Web Conferences Steering Committee, 535–543. https://doi.org/10.1145/3184558.3188724.

[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. https://doi.org/10.1038/s41562-020-0909-7.

[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. https://doi.org/10.1007/s00799-018-0239-9.

[4] Gregor Leban, Blaž 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. https://doi.org/10.1145/2567948.2577024.

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

[6] Philip A Schrodt, Omür Yilmaz, Deborah J Gerner, and Dennis Hermreck. 2008. The cameo (conflict and mediation event observations) actor coding framework. In *2008 Annual Meeting of the International Studies Association*.

[7] Stuart N Soroka. 2016. Gatekeeping and the negativity bias. In *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.43.

[8] Swati, Erjavec Tomaž, and Mladenić Dunja. 2020. Eveout: reproducible event dataset for studying and analyzing the complex event-outlet relationship.