UNIVERSITÀ DEGLI STUDI
DI SALERNO

# GOOD OR BAD

## INFORMATION SYSTEMS FOR BIG DATA

**Alessandro Falco - Peter Citro - Matteo Toriello**

“Good news or Bad news first?”

# Table of Contents

# INTRODUCTION

The “Good or Bad” project described in this report aims to create an interactive web page that provides users with a clear and concise overview of the global situation by collecting real-time news data from the GDELT database. This platform continuously monitors news worldwide, offering a heterogeneous and complex data stream that requires robust data mining process to extract valuable insights. The core of the project consists of three main pages designed to facilitate access to the most relevant information:

- **Trending News**: Shows the most cited news, allowing users to identify dominant global topics over a given period;
- **Good News**: Highlights the best news, providing insight into positive international events;
- **Bad News**: Presents the worst news, enabling the monitoring of significant negative situation.

A central challenge of the project was managing the significant amount of real-time data and optimizing analyses to ensure high performance and quick user responses. The system combines various technologies and tools to organize, filter, and analyse data, ensuring clear and fast visualization for the user.

# GDELT DATA PRESENTATION

The GDELT dataset (Global Database of Events, Language and Tone) contains multiple fields documenting global events, encoded for structured analyses. It comprises various tables, including the Event Table and the Mentions Table.

For this project, the focus was on the Event Table, as it contains all the necessary fields for analyses presented on the web page. In GDELT version 2, this table includes about 62 variables, some of which are redundant.

## Event Table

1. **Data & Identification**
    - **GlobalEventID**: Unique identifier for each event.
    - **Day, MonthYear, Year, FractionDate**: Various date formats to facilitate integration with different software;
2. **Actor's fields**
    - **Actor1 e Actor2**: Each event involves two main actors with fields describing:
        - **ActorCode**: A unique code for the actor;
        - **ActorName**: Name or classification (es. "KURD" o "UNITED STATES").
        - **CountryCode, KnownGroupCode, EthnicCode, ReligionCode, TypeCode**: Codes identifying geographical, religious, ethnic affiliations, or social roles (e.g. NGO, government).
3. **Event & Actions**
    - **EventCode, EventBaseCode, EventRootCode**: CAMEO codes describing event actions in a hierarchical structure. For instance, a detailed code "0251" can be simplified to "025" or "02" for aggregated analysis;
    - **QuadClass**: High-Level classification (e.g., 1=Verbal Cooperation, 2=Material Cooperation…);
    - **GoldsteinScale**: A scale from -10 to +10 estimating the theorical impact of the event on stability;
    - **NumMentions, NumSources, NumArticles**: Measures of event relevance based on mentions, sources and articles.
4. **Event locations**
    - **ActorGeo e ActionGeo**: Fields geolocating actors and actions, including latitude, longitude and specific country or city codes.
5. **Tone & Data Management**
    - **AvgTone**: Score ranging from -100 (very negative) to +100 (very positive) estimating the "tone" of articles;
    - **DATEADDED, SOURCEURL**: Date of event addition to the database and original source URL.

# ARCHITECTURE

The main challenge during the design phase was to develop an architecture capable of integrating various technologies cohesively and practically. The main objective was to ensure rapid visualization of requested results while keeping the system simple to manage.

As illustrated in the following figure, the proposed architecture is not overly complex. While there is room for improvement, this structure is considered adequate to support the system.

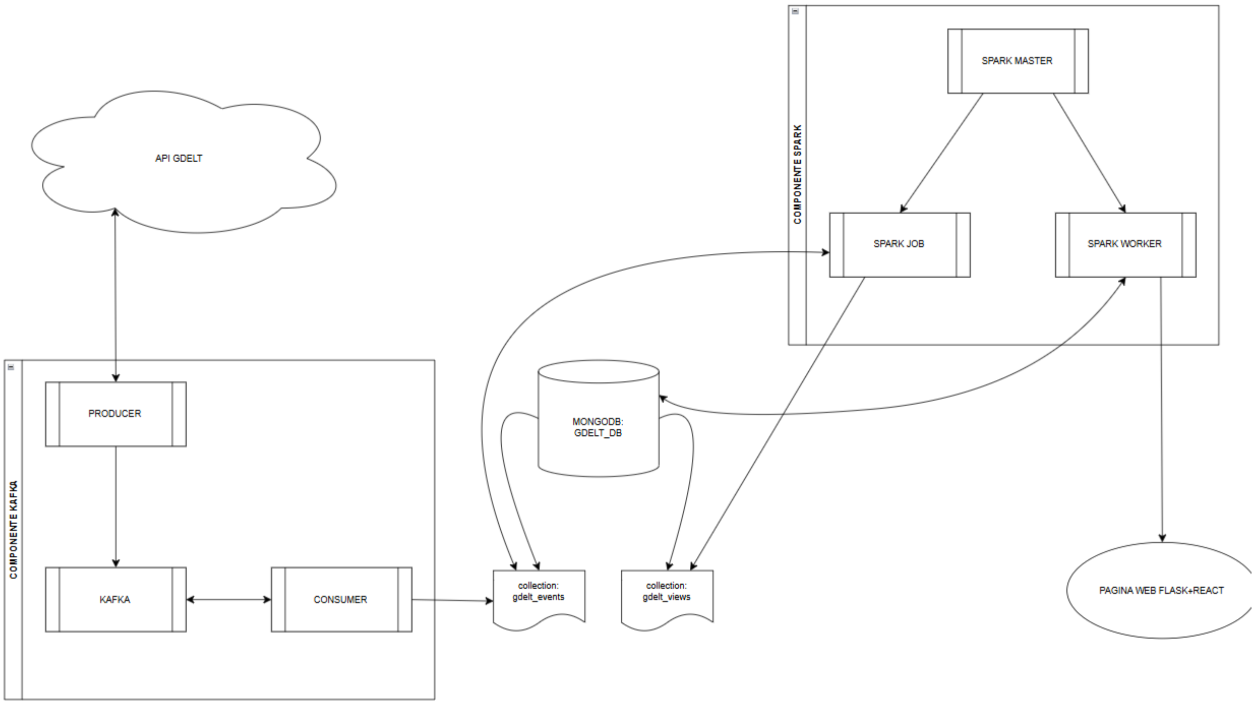
The system's operation can be summarized as follows:

1. **Kafka Producer**: Reads data from GDELT API and sends it to Kafka, which acts as a buffer;
2. **Kafka Consumer**: Retrieves data from Kafka and stores it in the NoSQL database GDELT_DB, in the *gdelt_events* collection.
3. **Spark Components**:
    a. **Spark Job**: Simulates a streaming process (repeated every 15 minutes). Reads data from the *gdelt_events* collection in MongoDB and generates three pre-computed views to speed up GET requests on the web page. The views are:
        i. *General;*
        ii. *Best_events_view;*
        iii. *Worst_event_view;*

    These filter data, selecting only those pertaining to the current and previous days, optimizing initial visualizations.

    b. **Spark Worker**: Manages POST requests from the web page. For instance, it executes queries to retrieve specific data from the entire *gdelt_events* collection based on user-defined filters.
4. **Flask-App**: Represents the web page used to interact with the data. The backend logic is implemented in Flask, while the frontend is developed with HTML, CSS and JavaScript. Additionally, a React Component, developed by our professor G. Fenza from University of Salerno, was integrated to filter news based on CameoCode, offering a cleaner and more targeted visualization.

Further details about individual component are provided in the "Technologies Used" section of this report.


API GDELT
COMPONENTE KAFKA
PRODUCER
KAFKA
CONSUMER
collection: gdelt_events
MONGODB: GDELT_DB
collection: gdelt_views
COMPONENTE SPARK
SPARK MASTER
SPARK JOB
SPARK WORKER
PAGINA WEB FLASK+REACT

# TECHNOLOGIES USED

This section provides detailed analysis of the technologies used in the project, explaining the process enabling the system's functionality and implementation.

***Kafka Producer***

The first component analysed is the Kafka Producer, which is fundamental for collecting and sending GDELT data. As described earlier, this component's primary task is to fetch data from the GDELT API and send it to Kafka. However, its role goes beyond simple message transmission: it is also responsible for initial data processing, ensuring the raw data is recoverable in the original GDELT db.

During its first execution, the Kafka Producer performs an initial check on the MongoDB collection and executes one of the following scenarios depending on the collection's state:

1. Empty Collection: Retrieves data from the GDELT API starting from 01/01/2024 to the execution date to populate MongoDB for the first time;
2. Non-Empty Collection: Identifies the most recent date in the saved data within MongoDB and retrieves missing data from the API up to the current date. This process is crucial to prevent dataset discontinuities, even during prolonged program interruptions.

After this, the Kafka Producer begins its natural cycle: every 15 minutes, it sends requests to the GDELT API for the latest 15-minute update and transmits the data to Kafka.
Before sending data to Kafka, the producer performs several checks and processes on the raw data to ensure proper visualization in subsequent system components. These checks, implemented after studying the dataset and conducting EDA, include:

- Managing Missing Data: For example, if a nation "is affected" by an event such as an attack, and the Actor1Name field is empty, the value is replaced with Actor2Name to ensure correct visualization in charts and tables;
- Geolocation Recovery: If ActionGeo_CountryCode is empty, the system attempts to retrieve the country code using latitude and longitude with the Nominatim object from the gropy.geocoders library;
- Fallback Assignments: If Actor2Name is empty, it assigns the value of ActionGeo_CountryCode, and if Actor1Name is empty it assigns the value of Actor2Name.

***Kafka Consumer***

The Kafka Consumer retrieves data from the Kafka buffer in batches of 100 records and then uploads them to MongoDB. Before doing so, it performs a duplication check, ensuring only new data is added verifying the GlobalEventID of each record against existing entries in the MongoDB collection.

To ensure system stability, error management mechanism is implemented to prevent interruptions in the flow between GDELT, the Producer, Kafka and the Consumer. This approach ensures operational continuity even in the presence of temporary anomalies.

***MongoDB***

For Data storage, the NoSQL MongoDB technology was chosen to effectively manage data volume and simplify parallel query execution performed by Spark during processing.

In MongoDB, several collections were created, each designed for a specific purpose in system execution. Key collections include:

1. *gdelt_events*. The primary collection where all data from the Kafka Consumer is stored. It contains the entire dataset provided by the GDELT API, forming the basis for all subsequent analyses;
2. *most_mentioned_view*. Manages GET requests on the “Most Mentioned News” page. This collection stores data for the current and previous days, selecting only the necessary columns to optimize backend performance and reduce complexity;
3. *best_events_view*. Manages GET requests for the “Good News” page. This collection stores positive events from the current and previous days. Data must have AvgTone>0 and GoldsteinScale>=5 to be included;
4. *worst_events_views*. Supports the GET requests for the “Bad News” page. This collection stores negative events from the current and previous days. Data must have AvgTone<0 and Goldsteinscale<=-5.

These MongoDB collections enable targeted and fast data management, efficiently supporting the system’s core functionalities.

***Spark-Job***

The implementation of the Spark Job arose from the need to keep batch vies used by GET requests on the website up-to-date. This approach was adopted to reduce computational load during initial requests, ensuring quick response times when launching various site page.

The Spark Job operates with a dedicated configuration of 2GB of RAM and 4 cores, periodically updating data every 15 minutes. If data for the requested days (current and previous) is unavailable, the job’s re-execution time is reduced to one minute, accelerating the recovery of missing information.

This component reads raw data from the *gdelt_events* collection in MongoDB and, based on the criteria described earlier, applies filters to create batch views: *most_mentioned_view, best_events_view* and *worst_events_view*. These views are written in overwrite mode, replacing existing views (MongoDB configured in this system does not support append mode).

Thanks to this component, the system alleviates computational load by providing pre-computed views for initial GET requests, significantly improving response times.

***Spark***

In the context of the “Good or Bad” project, Spark plays a fundamental role in managing analyses and processing the data requested by the web page. Its flexibility and ability to process large amounts of data in a distributed manner make it the natural choice for our system.

To ensure parallel execution of this task and the Spark Job, the resources for this process were limited to 8 cores and 4GB of RAM.

Spark is responsible for executing all operations requested by the web page, including:

1. Data Loading: Retrieves data from the *gdelt_events* collection in MongoDB, using it as the basis for subsequent processing;
2. Statistics Computation: Generates statistics useful for creating tables and charts displayed on the web page. These summarize relevant information for the user, such as the count of major events, the day’s average tone and more;

3. Executing Custom Queries: Performs data queries based on user-applied filters on the web page, allowing personalized results by selecting only the information of interest from the entire dataset.

Thanks to Spark, the system efficiently manages data volumes, providing quick responses to user requests.

***Flask-App***

The backend of the website was developed in Python using the Flask library, which, thanks to integration with Jinja2, allows dynamic embedding of Python objects directly into HTML pages.

This approach made it possible to manage the main functionalities of the website, rendering the interface intuitive and functional without the need to use other tools like Grafana, which would have constrained the visualization aspect.

The three main pages of the site (*Trending News, Good News and Bad News)* share a common structure and logic to ensure consistency in the user experience.

Common aspects include:

1. Header: Fixed banner at the top of the page containing the title and navigation buttons;
2. Sidebar: Side bar with filters to customize data visualization. The bar is fixed to remain visible while scrolling. Available filters include:
    a. CameoCode Filter: Allows selection of specific codes (or group of codes) for the news. This section was developed in React by prof. G. Fenza and integrated after minor modifications to ensure proper communication with the backend and visual consistency;
    b. Date Filter: Enables the selection of a time range. Naturally, the broader the range, the longer the response times;
    c. State Filter: Filters news by country code, allowing analysis of a single country situation;
3. Style: The site uses a single CSS file, primarily generated using ChatGPT;
4. JavaScript: The same JavaScript functions were reused on each page. These are simple and manage basic functionalities, such as ensuring the start date does not exceed the end date when selecting filters;
5. Utils: A folder dedicated to support files. Within it is the *Funzioni_Spark.py* file, containing functions to generate charts and statistics displayed on web pages. This approach keeps the backend code clean.

***Index: Notizie in Voga***

The main page of the site provides a general overview of news during the selected period (initially the current day and the previous one). Key sections of page include:

- Map Section: An interactive map highlighting the most trending news (those with NumMentions above the average) with clickable points. Clicking on the points redirects users to the news site; however, this functionality is strongly discouraged as the news collected from GDELT is not filtered and may redirect to fake or malicious sites. Additionally, there is a list of the five most frequent CameoCodes with the number of associated events. Clicking on a CameoCode highlights only the corresponding point on the map.
- Trends Section: This section analyses the average tone (AvgTone) of the news. It includes:
    - Histogram, displays the distribution of news tone during the period;
    - Dynamic Bar shows the tone average. A negative value indicates a generally negative period, while a positive value suggests a favourable period.

***GoodVibes/BadVibes***

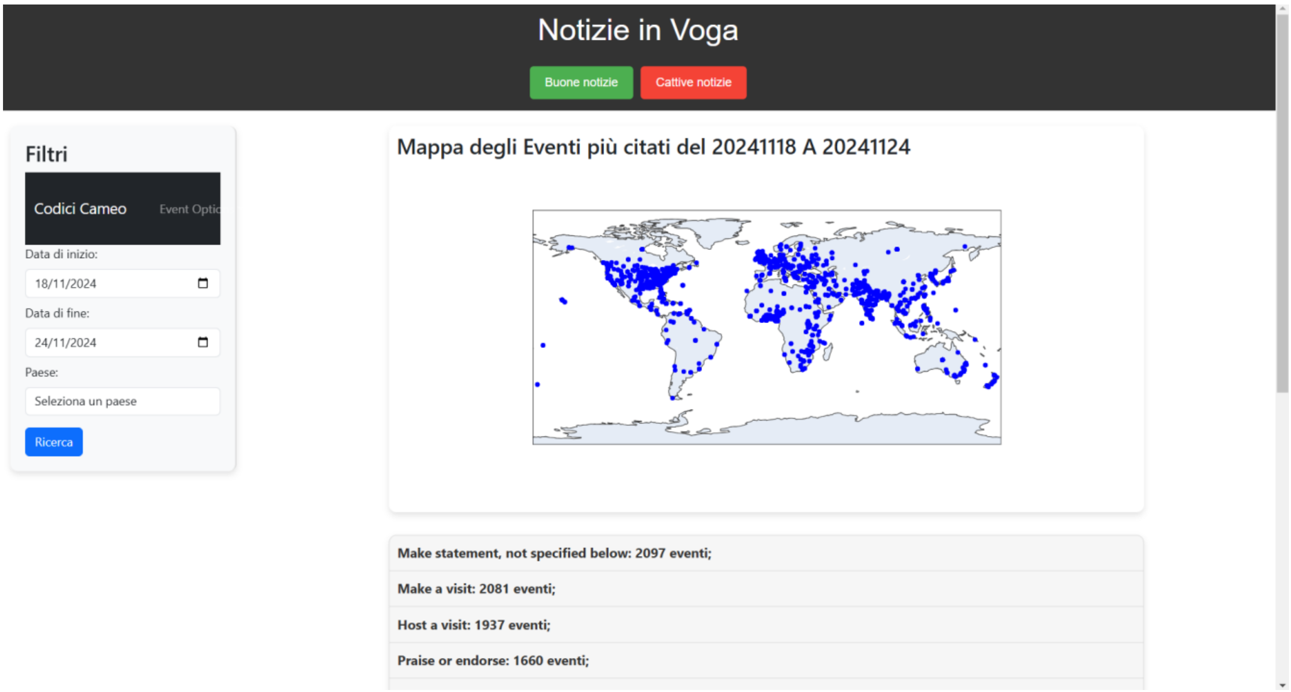
The Good News and Bad News pages offer identical analyses for positive and negative news, respectively. For Good News, the selection criteria are AvgTone>0 and GoldsteinScale>=5, while for Bad News, they are AvgTone<0 and GoldsteinScale<=-5.

Both pages include:

- Map Section: Has the same functionalities as the map section of the main page. The only difference is the color of the points displayed on the map.
- Pie Chart Section: Includes two pie charts showing the distribution of CameoCodes relative to GoldsteinScale and AvgTone. The charts are interactive: specific codes can be deselected or additional details viewed by hovering over the “slices”. The interactive bar showing the AvgTone average is also included in this section.
- Bar Chart Section: Displays a bar chart showing the countries with the highest number of positive (or negative) news during the selected period. A list of the top five countries by event count is also included;
- Table Section: Two tables provide a detailed overview of the news:
    - Top 5 Best (or Worst) News of the Period: lists the most relevant news (sorted by GoldsteinScale and AvgTone) with details about involved actors, tone, GoldsteinScale, mention count, CameoCode and the news link;
    - Top 5 Most Mentioned Best (or Worst) News of the Period: Like the first table but sorted by NumMentions.

# USAGE EXAMPLE

***Le notizie più in voga della settimana***



Notizie in Voga

Filtri

Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

Mappa degli Eventi più citati del 20241118 A 20241124

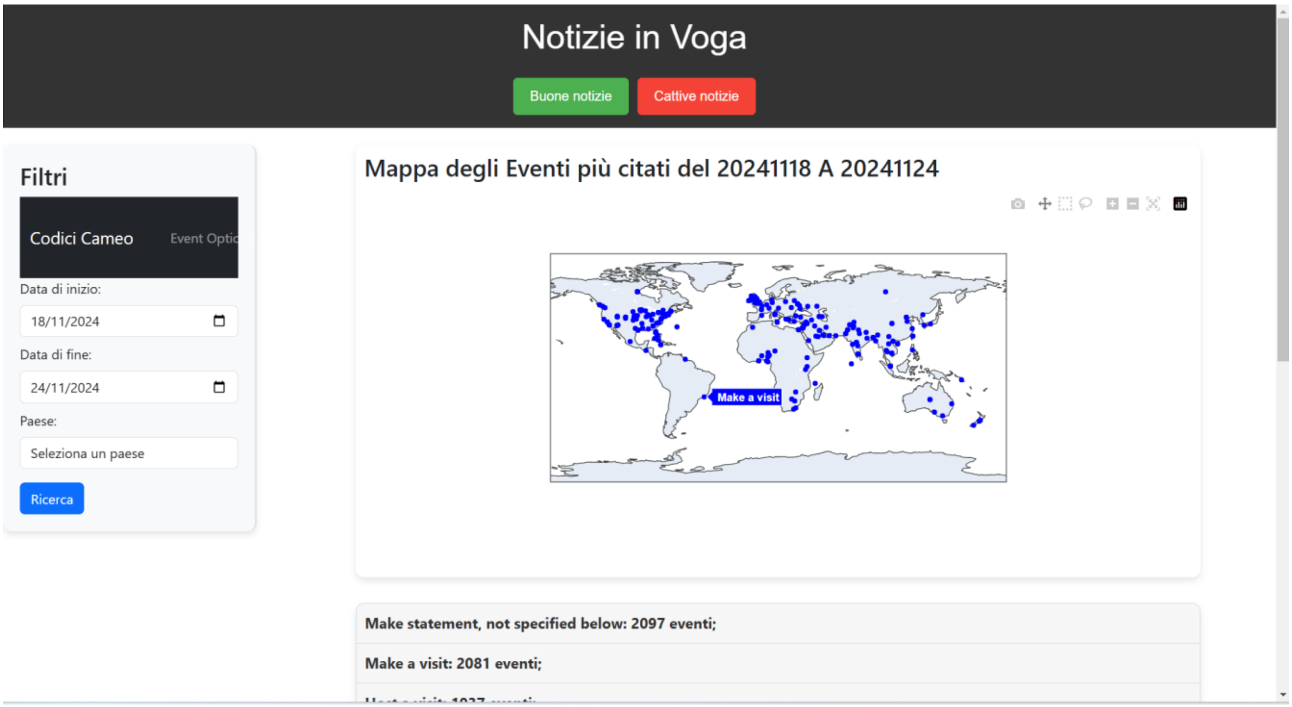
**Make statement, not specified below: 2097 eventi;**

**Make a visit: 2081 eventi;**

**Host a visit: 1937 eventi;**

**Praise or endorse: 1660 eventi;**

***Page after click on "Host a visit"***



Notizie in Voga

Filtri

Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

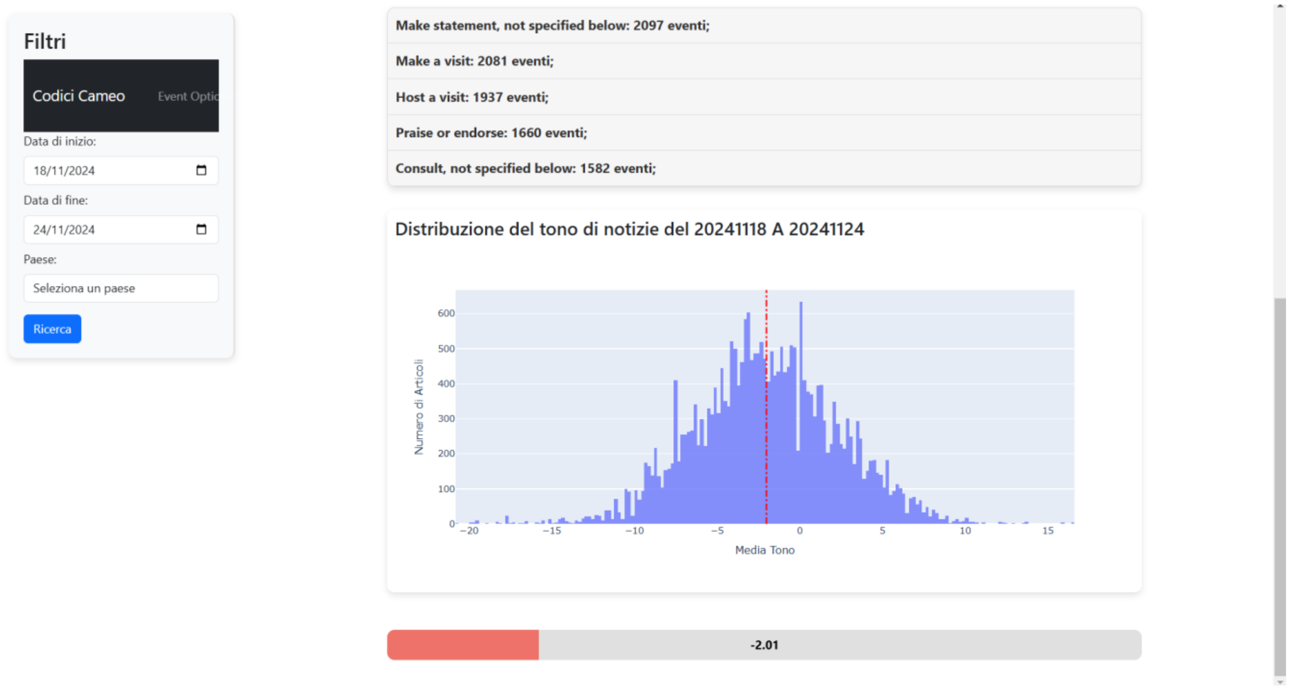
Mappa degli Eventi più citati del 20241118 A 20241124

**Make statement, not specified below: 2097 eventi;**

**Make a visit: 2081 eventi;**

**Host a visit: 1937 eventi;**

### *Average Tone of news from 18/11/2024 to 24/11/2024*



Filtri

Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

**Make statement, not specified below: 2097 eventi;**

**Make a visit: 2081 eventi;**

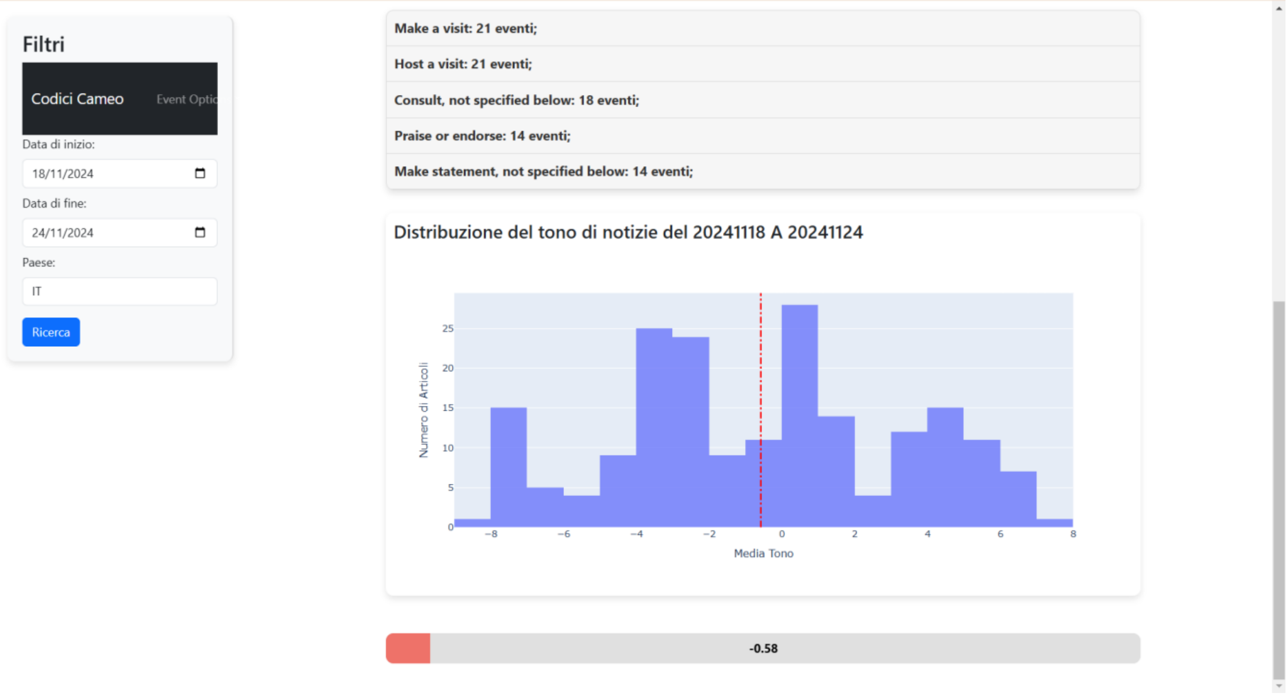
**Host a visit: 1937 eventi;**

**Praise or endorse: 1660 eventi;**

**Consult, not specified below: 1582 eventi;**

Distribuzione del tono di notizie del 20241118 A 20241124

### *Distribution of trending news in Italy*



Filtri

Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: IT

**Make a visit: 21 eventi;**

**Host a visit: 21 eventi;**

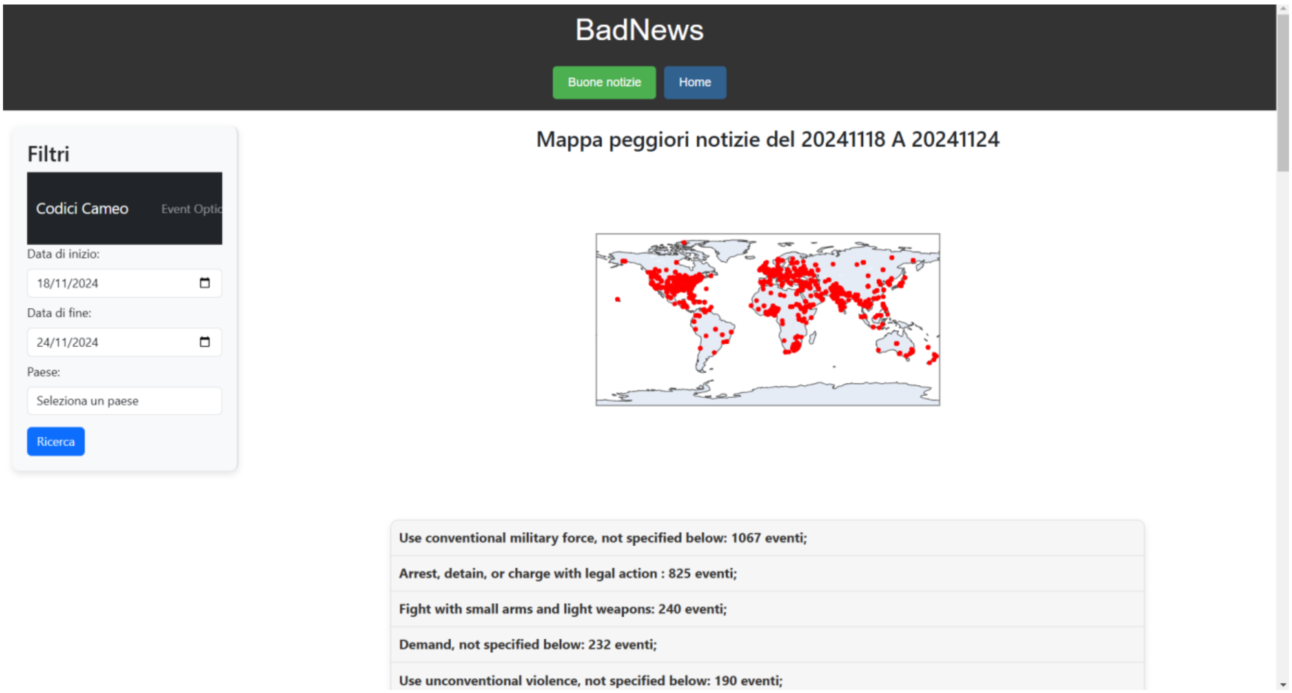
**Consult, not specified below: 18 eventi;**

**Praise or endorse: 14 eventi;**

**Make statement, not specified below: 14 eventi;**

Distribuzione del tono di notizie del 20241118 A 20241124

***worse week news***



Mappa peggiori notizie del 20241118 A 20241124

**Use conventional military force, not specified below: 1067 eventi;**

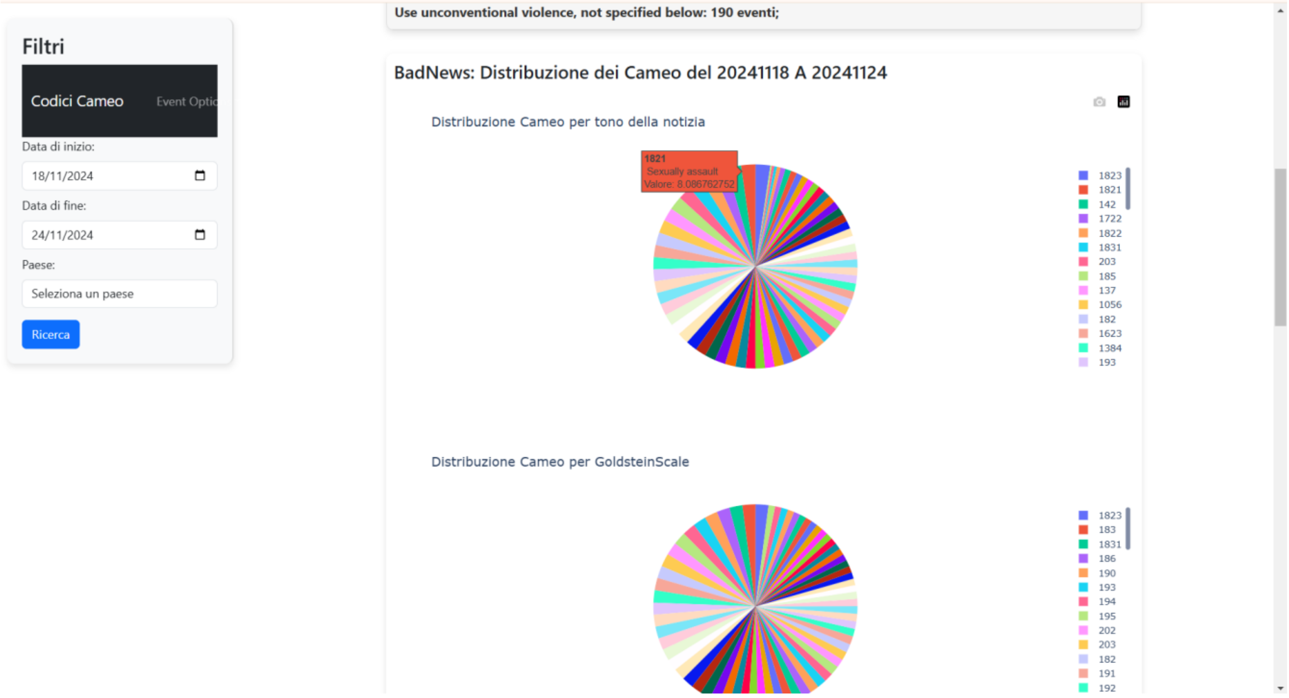
**Arrest, detain, or charge with legal action : 825 eventi;**

**Fight with small arms and light weapons: 240 eventi;**

**Demand, not specified below: 232 eventi;**

**Use unconventional violence, not specified below: 190 eventi;**

***Cameo distribution for the worse news***



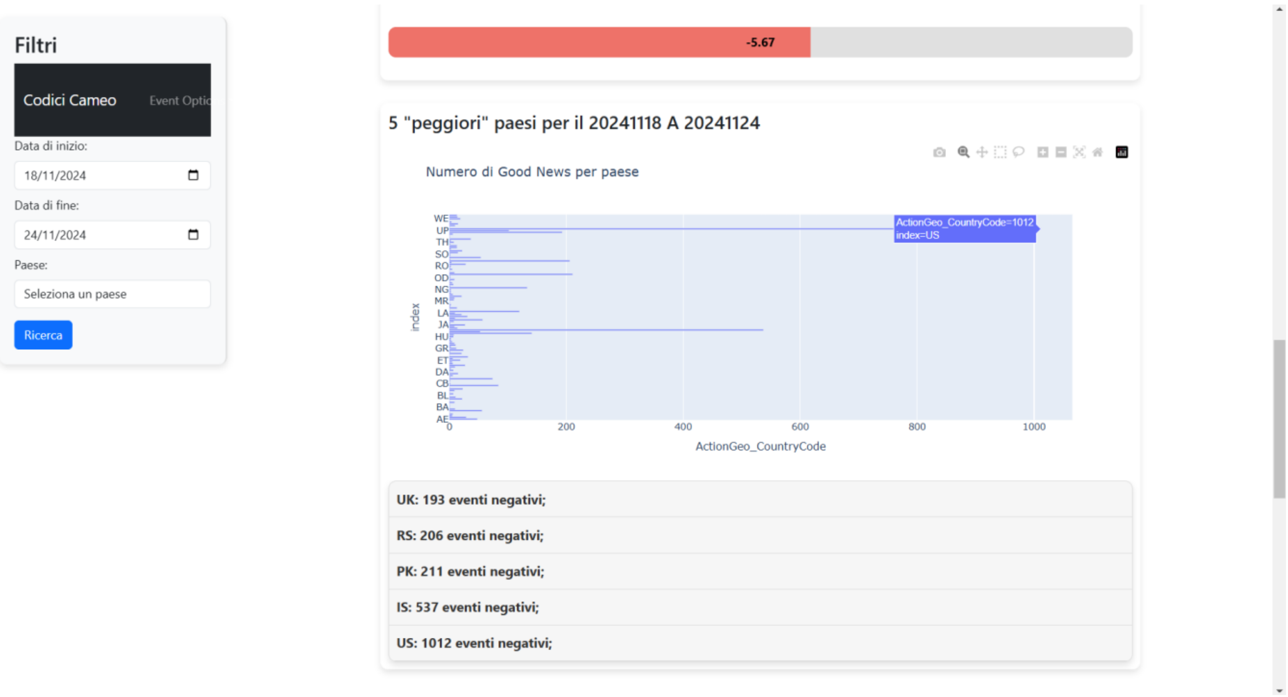
**Use unconventional violence, not specified below: 190 eventi;**

BadNews: Distribuzione dei Cameo del 20241118 A 20241124

Distribuzione Cameo per tono della notizia

Distribuzione Cameo per GoldsteinScale

### *5 Countries with the highest number of worse events*



Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

5 "peggiori" paesi per il 20241118 A 20241124

**UK: 193 eventi negativi;**

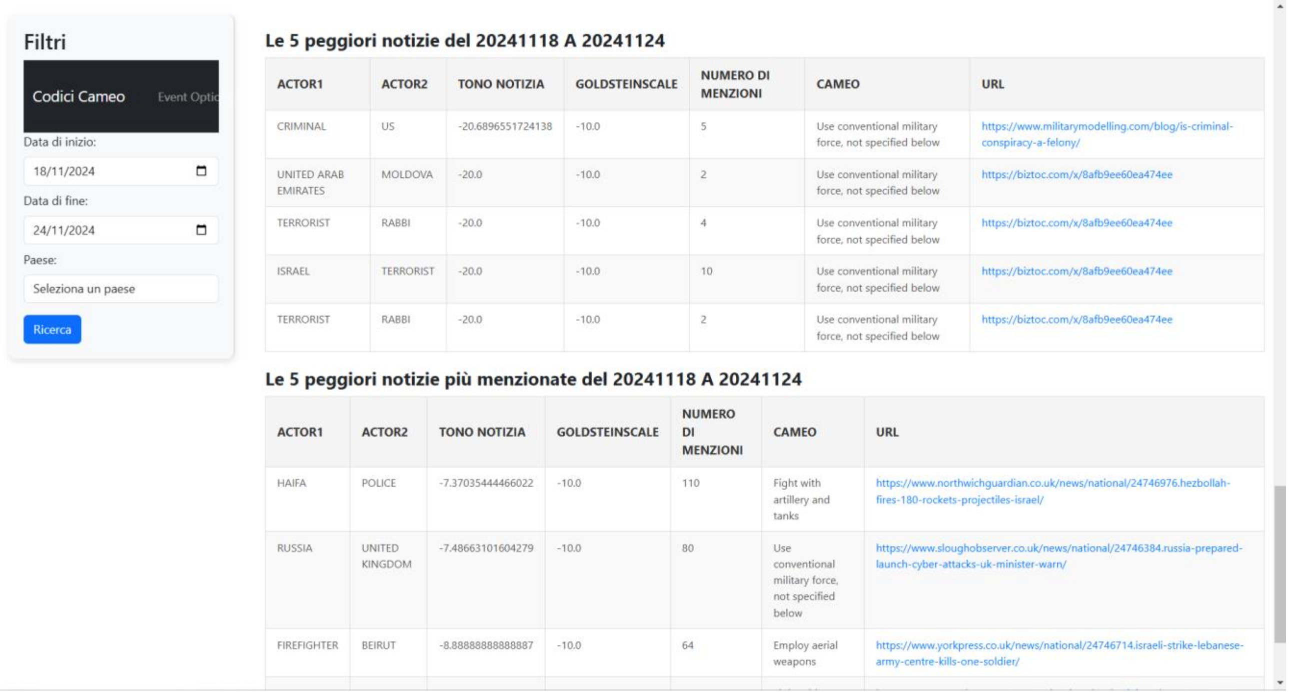
**RS: 206 eventi negativi;**

**PK: 211 eventi negativi;**

**IS: 537 eventi negativi;**

**US: 1012 eventi negativi;**

### *5 worse news*



Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

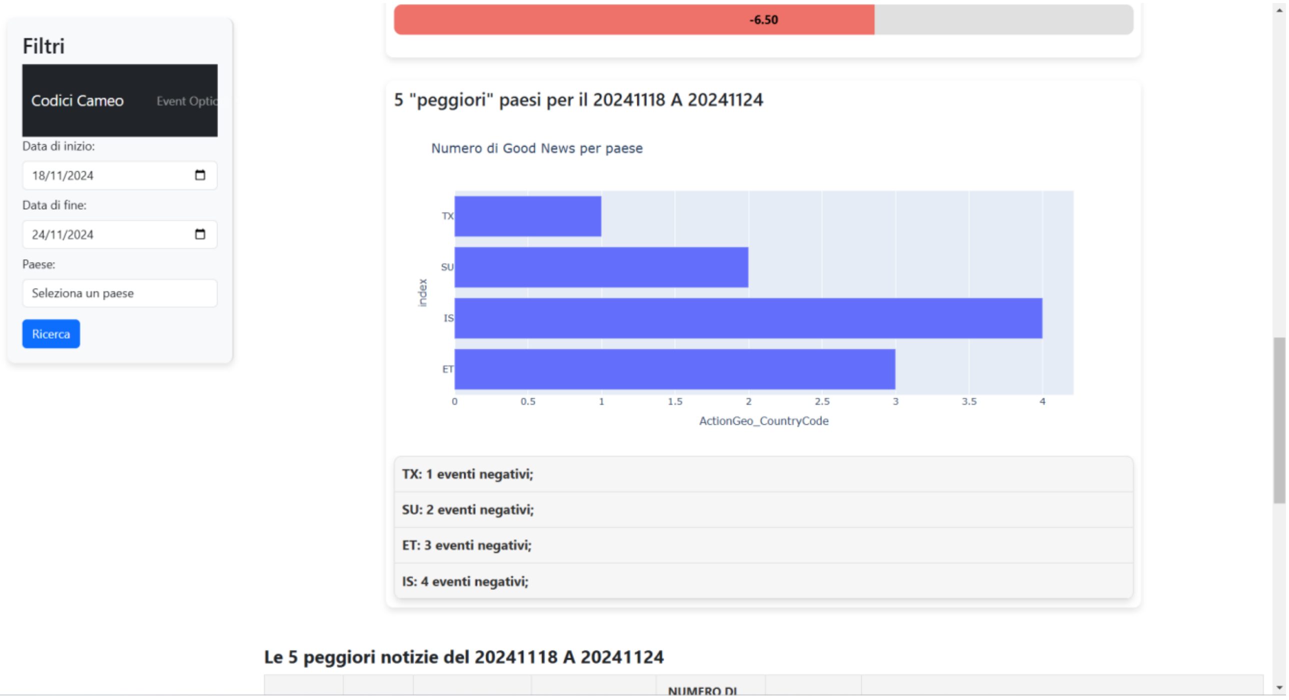
**Le 5 peggiori notizie del 20241118 A 20241124**

| ACTOR1 | ACTOR2 | TONO NOTIZIA | GOLDSTEINSCALE | NUMERO DI MENZIONI | CAMEO | URL |
|---|---|---|---|---|---|---|
| CRIMINAL | US | -20.6896551724138 | -10.0 | 5 | Use conventional military force, not specified below | https://www.militarymodelling.com/blog/is-criminal-conspiracy-a-felony/ |
| UNITED ARAB EMIRATES | MOLDOVA | -20.0 | -10.0 | 2 | Use conventional military force, not specified below | https://biztoc.com/x/8afb9ee60ea474ee |
| TERRORIST | RABBI | -20.0 | -10.0 | 4 | Use conventional military force, not specified below | https://biztoc.com/x/8afb9ee60ea474ee |
| ISRAEL | TERRORIST | -20.0 | -10.0 | 10 | Use conventional military force, not specified below | https://biztoc.com/x/8afb9ee60ea474ee |
| TERRORIST | RABBI | -20.0 | -10.0 | 2 | Use conventional military force, not specified below | https://biztoc.com/x/8afb9ee60ea474ee |

**Le 5 peggiori notizie più menzionate del 20241118 A 20241124**

| ACTOR1 | ACTOR2 | TONO NOTIZIA | GOLDSTEINSCALE | NUMERO DI MENZIONI | CAMEO | URL |
|---|---|---|---|---|---|---|
| HAIFA | POLICE | -7.37035444466022 | -10.0 | 110 | Fight with artillery and tanks | https://www.northwichguardian.co.uk/news/national/24746976.hezbollah-fires-180-rockets-projectiles-israel/ |
| RUSSIA | UNITED KINGDOM | -7.48663101604279 | -10.0 | 80 | Use conventional military force, not specified below | https://www.sloughobserver.co.uk/news/national/24746384.russia-prepared-launch-cyber-attacks-uk-minister-warn/ |
| FIREFIGHTER | BEIRUT | -8.88888888888887 | -10.0 | 64 | Employ aerial weapons | https://www.yorkpress.co.uk/news/national/24746714.israeli-strike-lebanese-army-centre-kills-one-soldier/ |

### *Worst news for Cameo 20: “Use unconventional Mass Violence”*



Data di inizio: 18/11/2024

Data di fine: 24/11/2024

Paese: Seleziona un paese

5 "peggiori" paesi per il 20241118 A 20241124

**TX: 1 eventi negativi;**

**SU: 2 eventi negativi;**

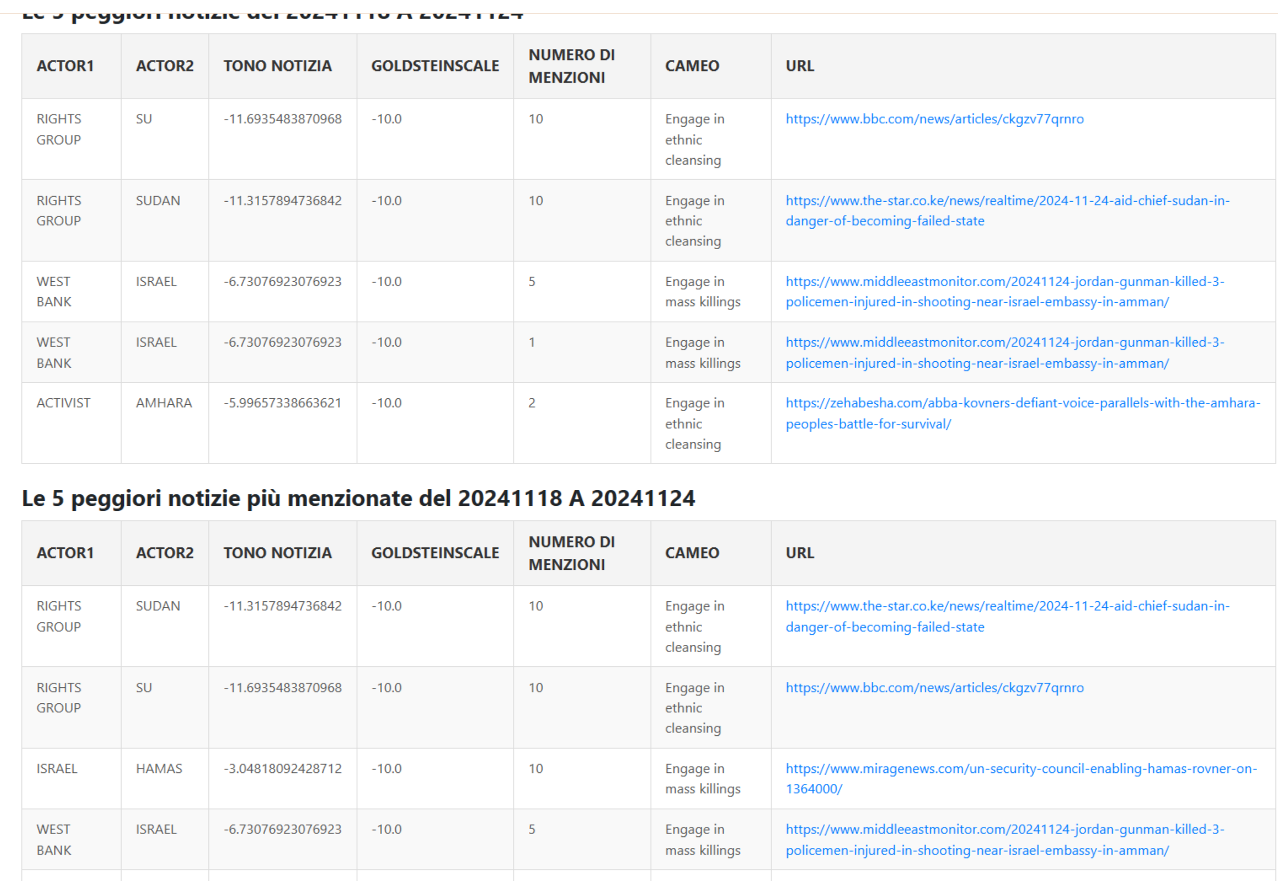
**ET: 3 eventi negativi;**

**IS: 4 eventi negativi;**

**Le 5 peggiori notizie del 20241118 A 20241124**

NUMERO DI

### *5 worse news for Cameo 20*



**Le 5 peggiori notizie del 20241118 A 20241124**

| ACTOR1 | ACTOR2 | TONO NOTIZIA | GOLDSTEINSCALE | NUMERO DI MENZIONI | CAMEO | URL |
|---|---|---|---|---|---|---|
| RIGHTS GROUP | SU | -11.6935483870968 | -10.0 | 10 | Engage in ethnic cleansing | https://www.bbc.com/news/articles/ckgzv77qrnro |
| RIGHTS GROUP | SUDAN | -11.3157894736842 | -10.0 | 10 | Engage in ethnic cleansing | https://www.the-star.co.ke/news/realtime/2024-11-24-aid-chief-sudan-in-danger-of-becoming-failed-state |
| WEST BANK | ISRAEL | -6.73076923076923 | -10.0 | 5 | Engage in mass killings | https://www.middleeastmonitor.com/20241124-jordan-gunman-killed-3-policemen-injured-in-shooting-near-israel-embassy-in-amman/ |
| WEST BANK | ISRAEL | -6.73076923076923 | -10.0 | 1 | Engage in mass killings | https://www.middleeastmonitor.com/20241124-jordan-gunman-killed-3-policemen-injured-in-shooting-near-israel-embassy-in-amman/ |
| ACTIVIST | AMHARA | -5.99657338663621 | -10.0 | 2 | Engage in ethnic cleansing | https://zehabesha.com/abba-kovners-defiant-voice-parallels-with-the-amhara-peoples-battle-for-survival/ |

**Le 5 peggiori notizie più menzionate del 20241118 A 20241124**

| ACTOR1 | ACTOR2 | TONO NOTIZIA | GOLDSTEINSCALE | NUMERO DI MENZIONI | CAMEO | URL |
|---|---|---|---|---|---|---|
| RIGHTS GROUP | SUDAN | -11.3157894736842 | -10.0 | 10 | Engage in ethnic cleansing | https://www.the-star.co.ke/news/realtime/2024-11-24-aid-chief-sudan-in-danger-of-becoming-failed-state |
| RIGHTS GROUP | SU | -11.6935483870968 | -10.0 | 10 | Engage in ethnic cleansing | https://www.bbc.com/news/articles/ckgzv77qrnro |
| ISRAEL | HAMAS | -3.04818092428712 | -10.0 | 10 | Engage in mass killings | https://www.miragenews.com/un-security-council-enabling-hamas-rovner-on-1364000/ |
| WEST BANK | ISRAEL | -6.73076923076923 | -10.0 | 5 | Engage in mass killings | https://www.middleeastmonitor.com/20241124-jordan-gunman-killed-3-policemen-injured-in-shooting-near-israel-embassy-in-amman/ |

# CONCLUSIONS

The Good or Bad project, in our opinion, has successfully demonstrated the ability to integrate different technologies to address the complexity of analysing and managing large volumes of real-time data. Through tools like Kafka, Spark, Flask and other, adopted technologies, the system achieves the goal of offering an interactive and intuitive platform capable of synthesizing global information in a clear and immediate way.

The features presented in the *Trending News, Good News and Bad News* sections can provide valuable support for users in their daily lives. For example, an interested reader could access the platform to obtain an up-to-date overview of global events or the situation in their own country, using the available filters to customize the view.

Additionally, the project offers significant room for improvement, including:

1. Expansions of analytical features: Incorporating new metrics and graphical visualization to deepen the insights already available;
2. Improvement of data quality: Adding advanced filters to reduce the risk of displaying content from unreliable sources such as malicious websites or fake news;
3. Optimization of architecture: Exploring the adoption of additional technologies to enhance system scalability and robustness;
4. Extension of the user interface: Developing advanced functionalities like data export or content sharing to make the platform even more useful.

In conclusion, in our opinion, *Good or Bad* represents an excellent starting point for the creation of an open-source solution that is useful to the web community. On one hand, it provides end users with an innovative, intuitive and simple platform that enables them to monitor global news. On the other, it servers as an example for developers who might leverage the code produces for this project to create their own solutions for GDELT data management.