How the Pandemic Emphasized the Importance of Real-World Data

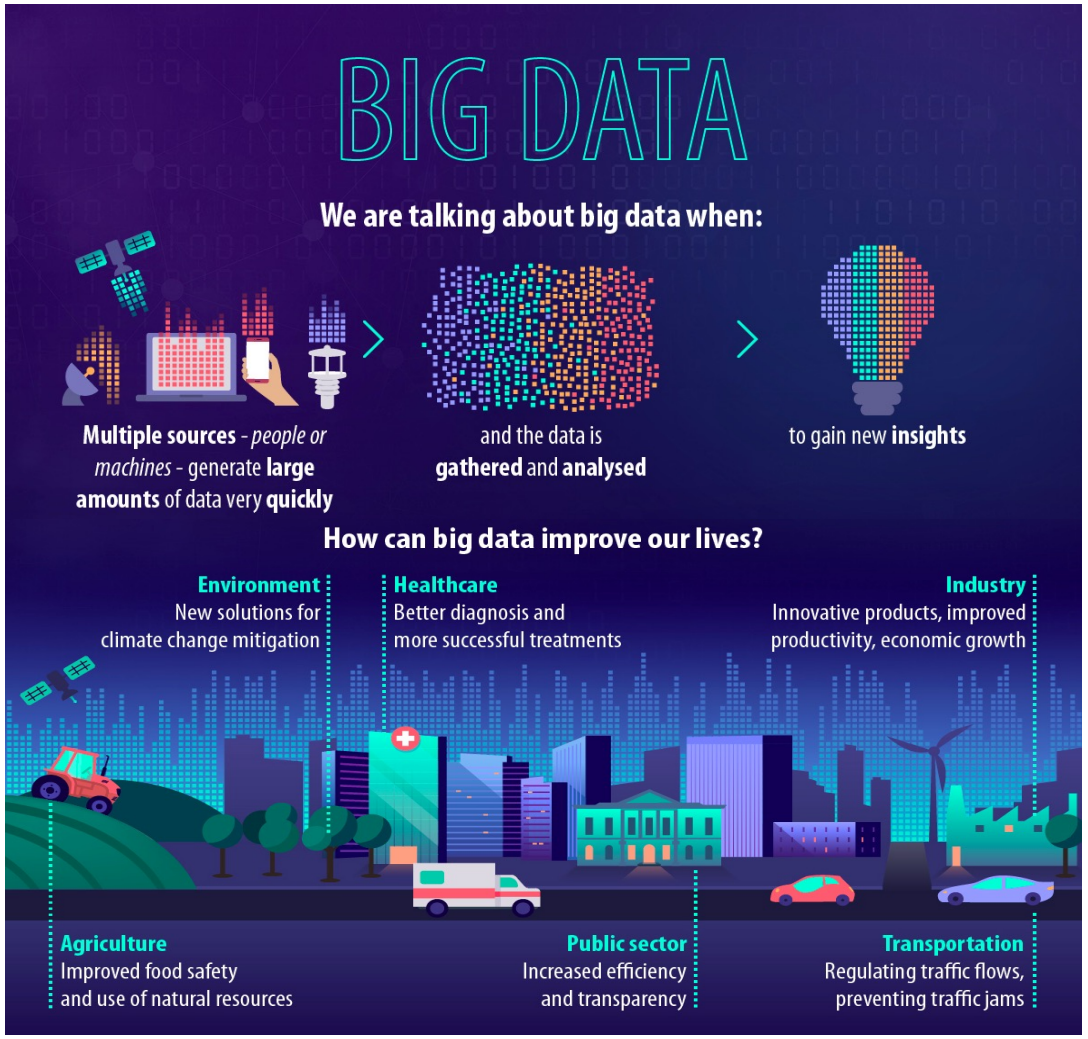
Francesco Branda

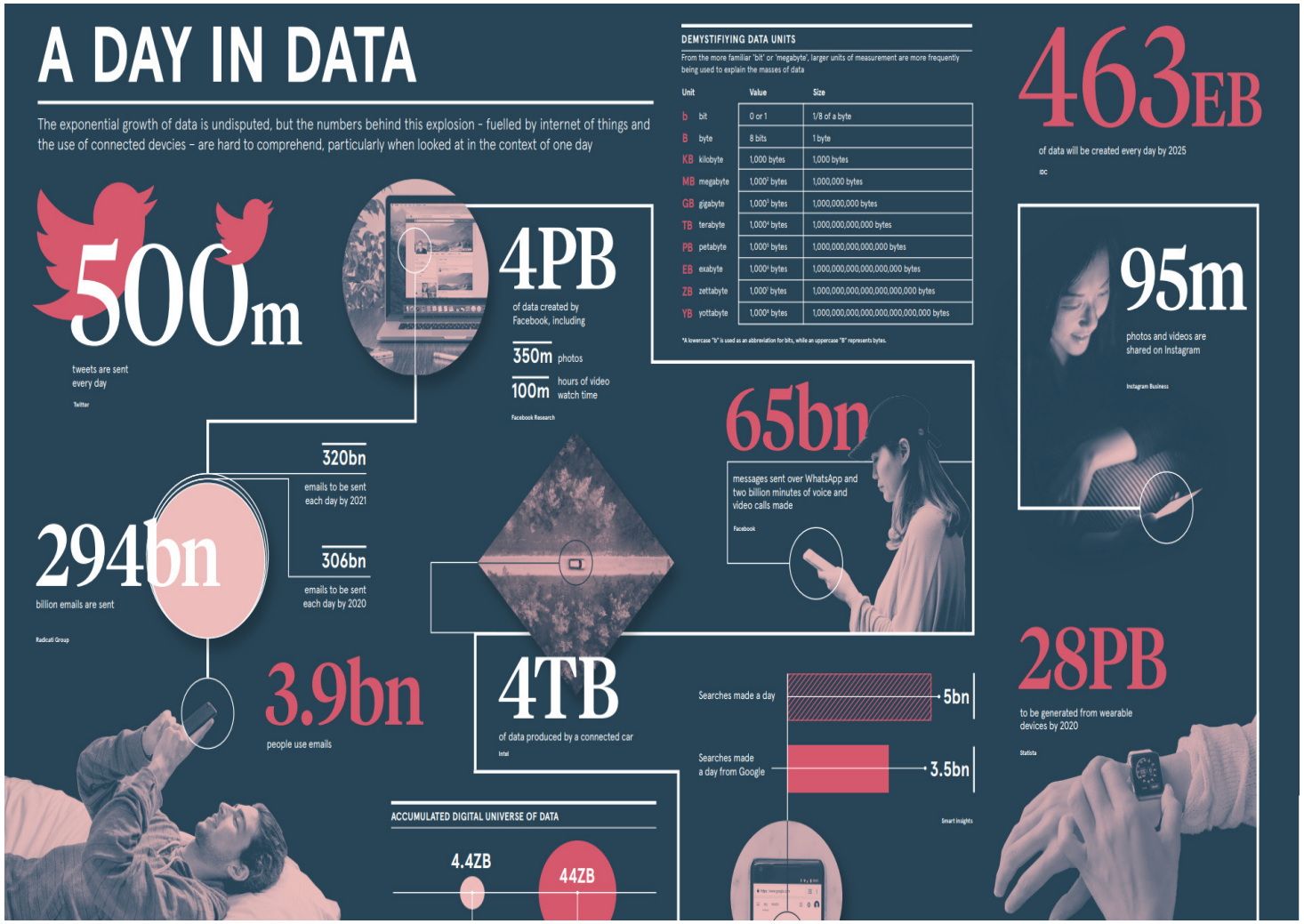
University of Calabria & DtoK Lab

E-mail: francesco.branda@unical.it

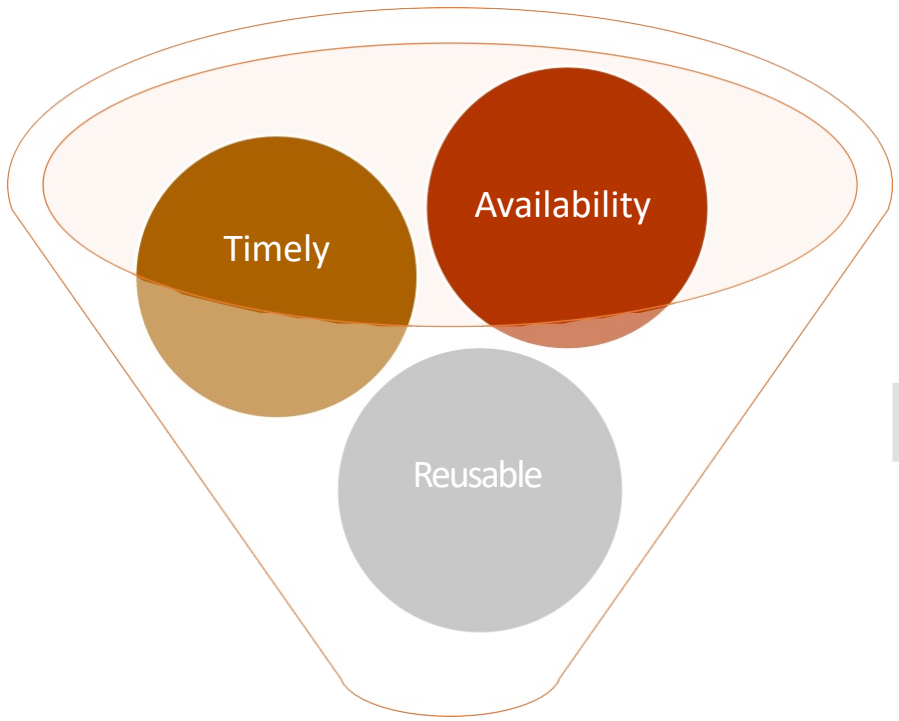
Web: https://francescobranda.netlify.app/

Overview


BIG DATA
We are talking about big data when:
Multiple sources - people or machines - generate large amounts of data very quickly
and the data is gathered and analysed
to gain new insights
How can big data improve our lives?
Environment
New solutions for climate change mitigation
Healthcare
Better diagnosis and more successful treatments
Industry
Innovative products, improved productivity, economic growth
Agriculture
Improved food safety and use of natural resources
Public sector
Increased efficiency and transparency
Transportation
Regulating traffic flows, preventing traffic jams

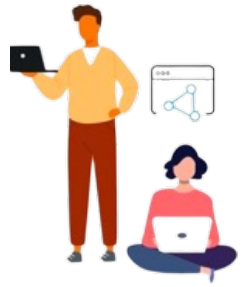

A DAY IN DATA
The exponential growth of data is undisputed, but the numbers behind this explosion - fuelled by internet of things and the use of connected devcies – are hard to comprehend, particularly when looked at in the context of one day
500m
tweets are sent every day
Twitter
4PB
of data created by Facebook, including
350m photos
100m hours of video watch time
Facebook Research
320bn
emails to be sent each day by 2021
306bn
emails to be sent each day by 2020
294bn
billion emails are sent
Radicati Group
3.9bn
people use emails
4TB
of data produced by a connected car
Intel
65bn
messages sent over WhatsApp and two billion minutes of voice and video calls made
Facebook
DEMYSTIFIYING DATA UNITS
From the more familiar "bit" or "megabyte", larger units of measurement are more frequently being used to explain the masses of data
Unit | Value | Size
b bit | 0 or 1 | 1/8 of a byte
B byte | 8 bits | 1 byte
KB kilobyte | 1,000 bytes | 1,000 bytes
MB megabyte | 1,000² bytes | 1,000,000 bytes
GB gigabyte | 1,000³ bytes | 1,000,000,000 bytes
TB terabyte | 1,000⁴ bytes | 1,000,000,000,000 bytes
PB petabyte | 1,000⁵ bytes | 1,000,000,000,000,000 bytes
EB exabyte | 1,000⁶ bytes | 1,000,000,000,000,000,000 bytes
ZB zettabyte | 1,000⁷ bytes | 1,000,000,000,000,000,000,000 bytes
YB yottabyte | 1,000⁸ bytes | 1,000,000,000,000,000,000,000,000 bytes
*A lowercase "b" is used as an abbreviation for bits, while an uppercase "B" represents bytes.
ACCUMULATED DIGITAL UNIVERSE OF DATA
4.4ZB
44ZB
Searches made a day
5bn
Searches made a day from Google
3.5bn
Smart Insights
463EB
of data will be created every day by 2025
IDC
95m
photos and videos are shared on Instagram
Instagram Business
28PB
to be generated from wearable devices by 2020
Statista

Research problem



Data access

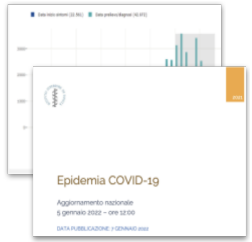
Metadata supply and quality



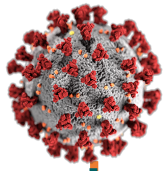
Data reuse

Data literacy

Data-driven surveillance system



Epidemiological surveillance



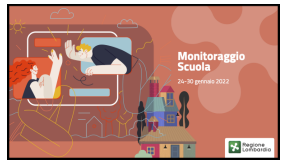
All SARS-CoV-2 sequences

VoC/VoI

- Integrated surveillance (case-based data sent to TESSy)
- Monthly prevalence surveys (cross-sectoral)
- National vaccination repository
- Genomic surveillance and analysis platform
- School monitoring

Ministry of Education

National Institute of Health (ISS)



Regional weekly monitoring

Personal contribution

https://www.covid19dataportal.it/it/data_types/forms/

COVID-19 Data Portal ITALY

Chi siamo Portale Europeo Supporto & Feedback Cerca en | it

Genomica & Trascrittomica Dati sulle Proteine Dati di Imaging Dati Sanitari Ricerca Eventi

Ricerca Italiana

Home / Tipologia di Dati / Ricerca Italiana

Progetti di Ricerca

Visualizzazione 1-3 di 3 elementi.

Ricerca: Argomento :

Descrizione	Leader del progetto	Contatto del Repository	Istituzione	Argomento
From the infection report to the vaccines: all the data on the Covid emergency in Calabria on a single platform	Francesco Branda	Francesco Branda francesco.branda@unical.it	Università della Calabria Rende	Health data
Descrizione completa	The COVIDA project is dedicated to the collection and visualization of data related to the COVID-19 emergency in Calabria and makes available to the scientific community all the necessary information, such as the total number of infections recorded on the territory, with a series of detailed indications (hospitalized, cured, deceased, number of swabs performed) to monitor and classify the epidemic risk, and the number of subjects vaccinated with the first dose, those vaccinated with a full cycle, the progress of vaccinations by category and age group, to evaluate the progress of the vaccination campaign. The platform can be reached at the link https://covida.tk/			
Full-genome sequencing of SARS-CoV-2 strains from the province of Parma	Enrico Silini	Riccardo Percudani riccardo.percudani@unipr.it	Università di Parma Parma	Genomics/Trascrittomics
Synthetic Antibodies neutralize SARS-CoV-2 infection of mammalian cells	Giuseppe Novelli	Giuseppe Novelli novelli@med.uniroma2.it	Università degli Studi di Roma Tor Vergata Roma	Health data Other

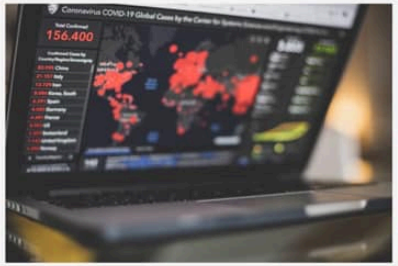
Visualizza 5 elementi per pagina

« 1 »

Portale COVID-19

COVIDA

Seguici su:



Andamento vaccinazione | Andamento epidemiologico | Bollettino giornaliero

COVID-19 Regione Calabria

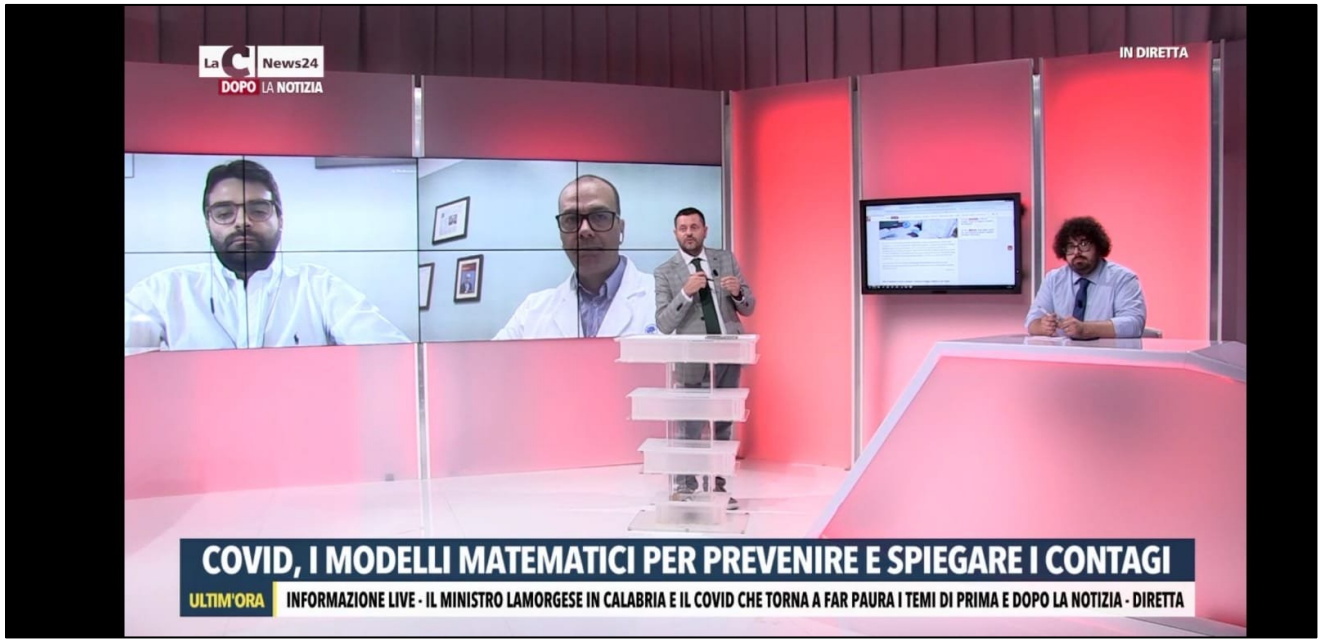
Ultimo aggiornamento: 21-07-2022

+3.020*	+4	+1.503	72.904
Nuovi casi	Deceduti	Dimessi / Guariti	Casi attivi

315 (-8)	18 (-1)
Ricoverati	Terapia intensiva

*Numero tamponi: 12.208 (+3.870 rispetto a ieri)

Personal contribution



diseases

MDPI

Article

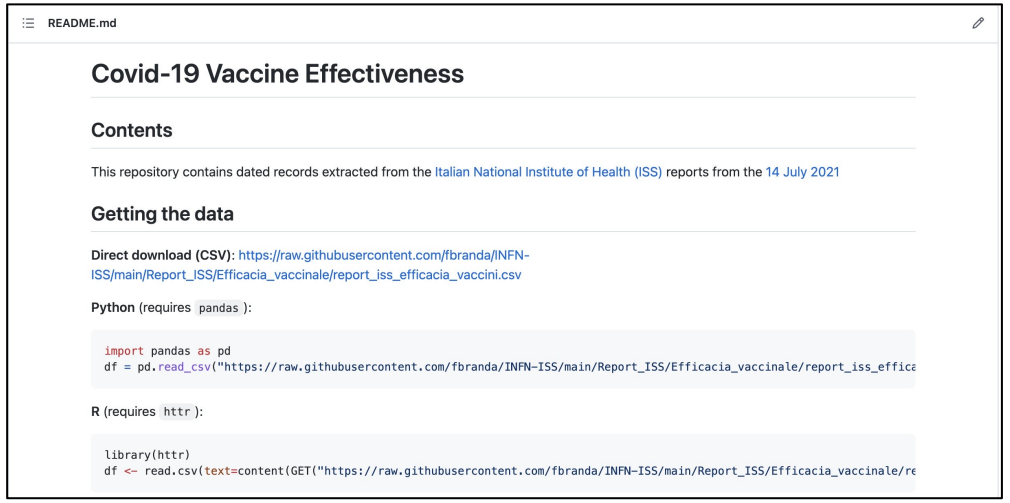
Predicting the Spread of SARS-CoV-2 in Italian Regions: The Calabria Case Study, February 2020–March 2022

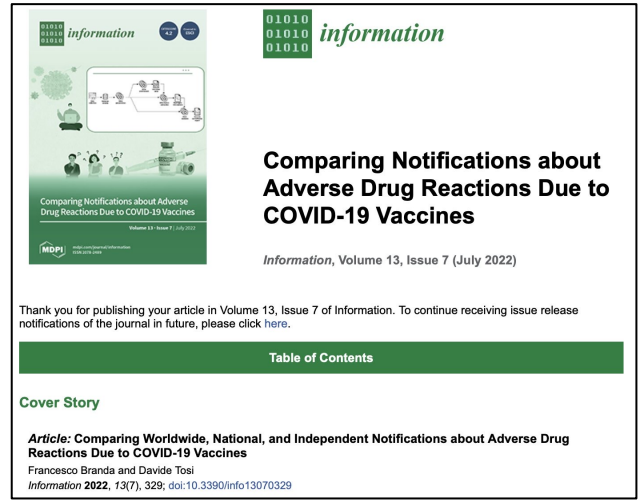
Francesco Branda [1,†], Ludovico Abenavoli [2,*,†], Massimo Pierini [3,4,†] and Sandra Mazzoli [4,†]

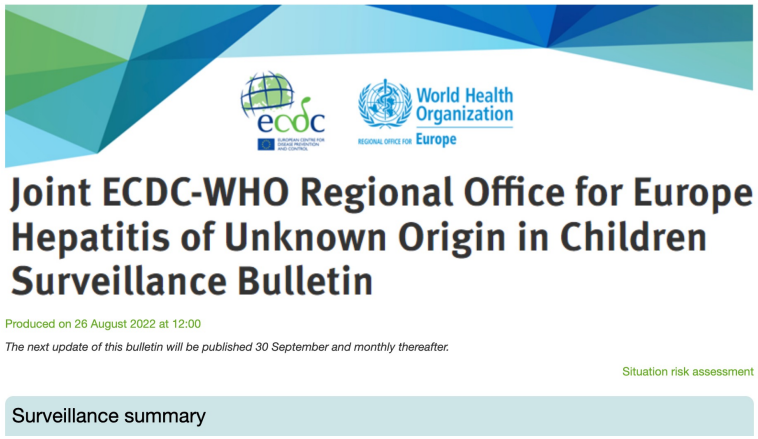
[1] Department of Computer Science, Modeling, Electronics and Systems Engineering (DIMES), University of Calabria, 87036 Rende, Italy; francesco.branda@unical.it
[2] Department of Health Sciences, University Magna Graecia, 88100 Catanzaro, Italy
[3] Guglielmo Marconi University, 00193 Rome, Italy; info@epidata.it
[4] SITO WEB del Gruppo Epidemiologico, EpiData.it, 24121 Bergamo, Italy; sandra.mazzoli50@gmail.com
* Correspondence: l.abenavoli@unicz.it
† These authors contributed equally to this work.

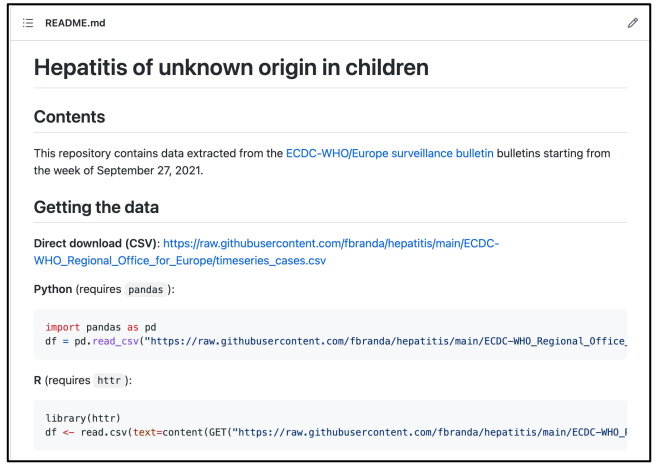
Abstract: Despite the stunning speed with which highly effective and safe vaccines have been developed, the emergence of new variants of SARS-CoV-2 causes high rates of (re)infection, a major impact on health care services, and a slowdown to the socio-economic system. For COVID-19, accurate and timely forecasts are therefore essential to provide the opportunity to rapidly identify risk areas affected by the pandemic, reallocate the use of health resources, design countermeasures, and increase public awareness. This paper presents the design and implementation of an approach based on autoregressive models to reliably forecast the spread of COVID-19 in Italian regions. Starting from the database of the Italian Civil Protection Department (DPC), the experimental evaluation was performed on real-world data collected from February 2020 to March 2022, focusing on Calabria, a region of Southern Italy. This evaluation shows that the proposed approach achieves a good predictive power for out-of-sample predictions within one week (R-squared > 0.9 at 1 day, R-squared > 0.7 at 7 days), although it decreases with increasing forecasted days (R-squared > 0.5 at 14 days).

Personal contribution

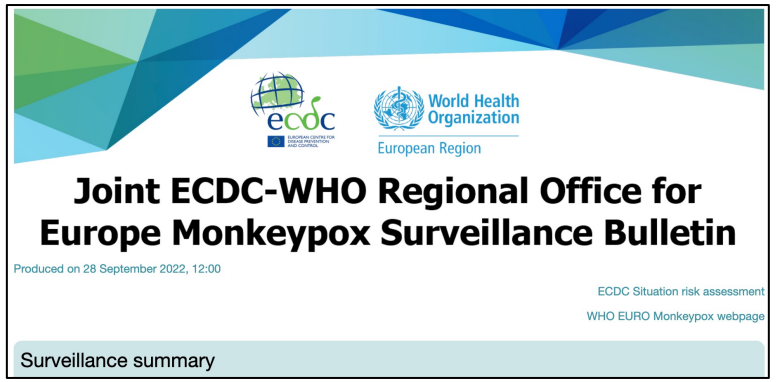


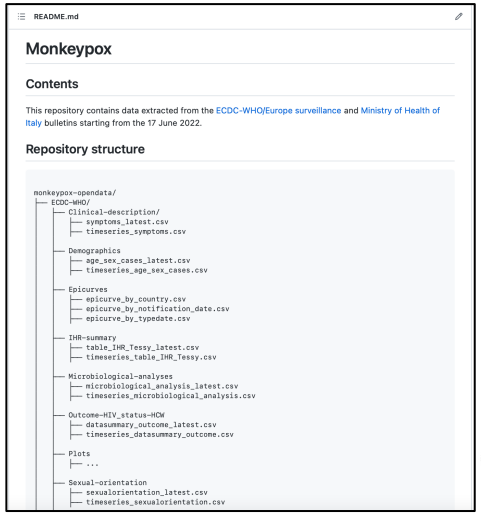


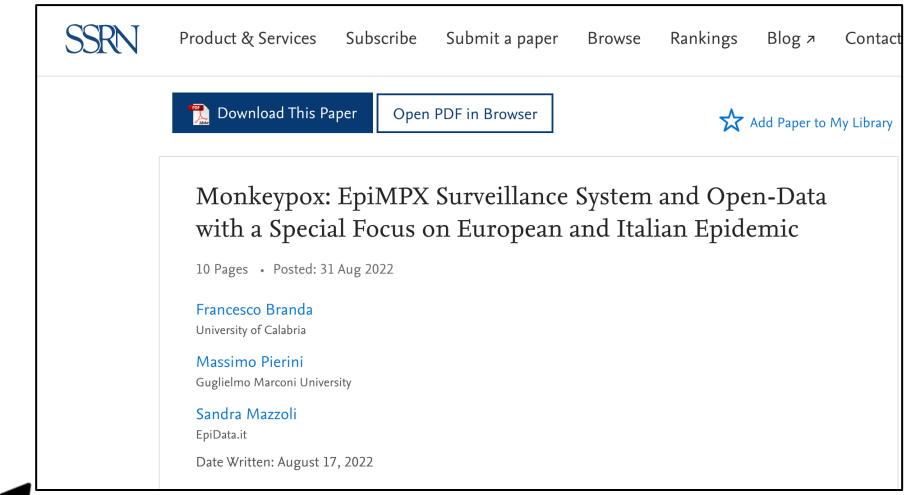


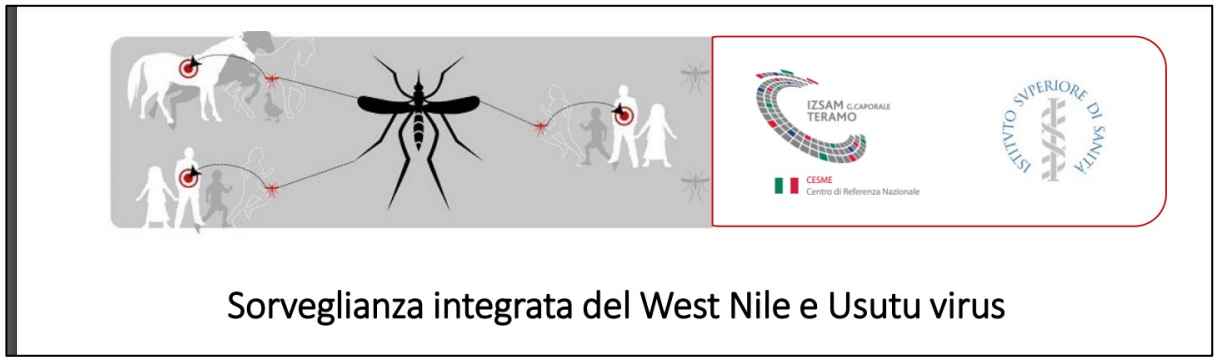


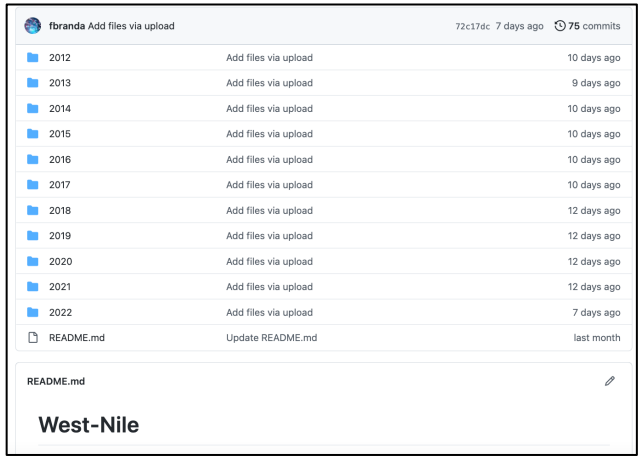
Personal contribution



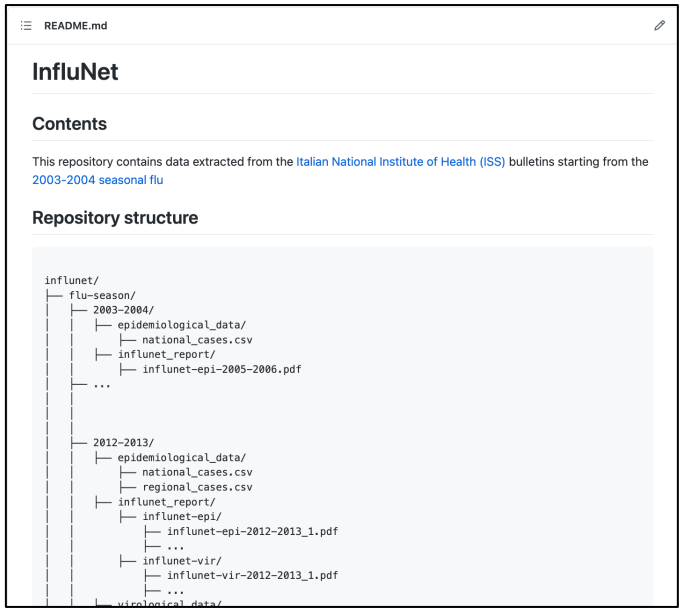


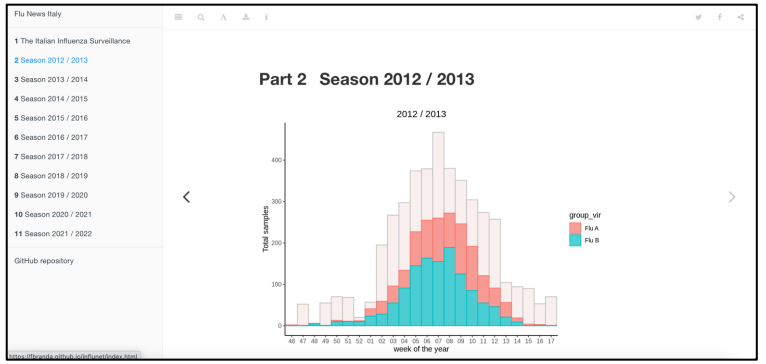




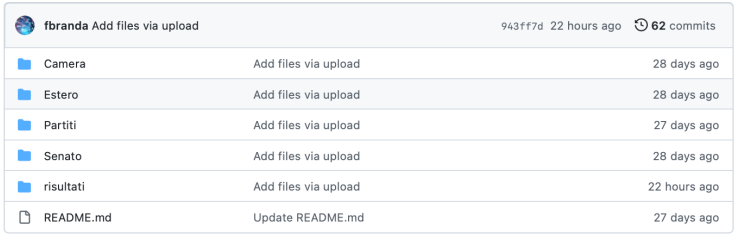


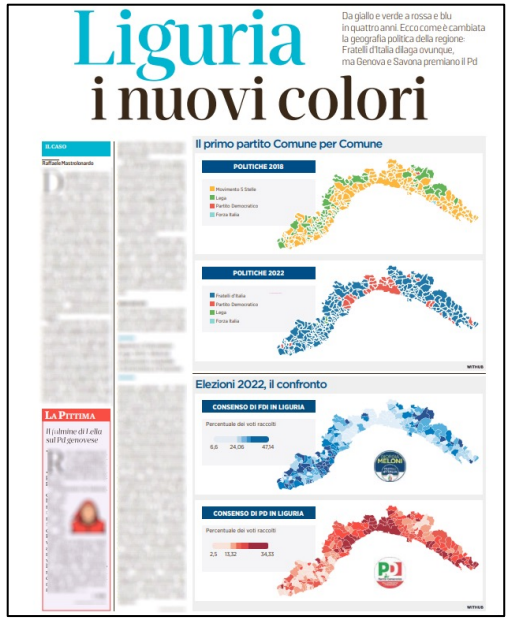
Personal contribution





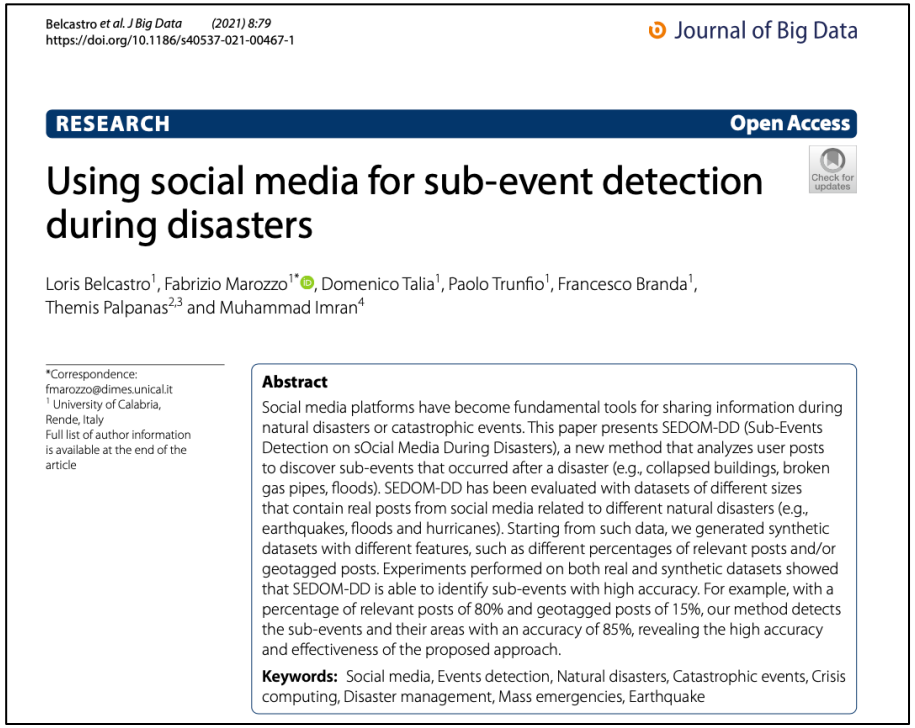
Personal contribution





Use case: SEDOM-DD

- Social media platforms have become an **important source** of information that can be exploited to understand **human dynamics** and **behaviors**.
- In the context of natural disasters, the very large use of social media platforms has enabled eyewitnesses and other disaster-affected people to share information about their damages, risks and emergencies in real time.
- The use of social media posts to help rescue and intervention activities remains an open challenge as users often publish posts containing inaccurate information, slang or abbreviated words, or without using geolocalization.
- The proposed **methodology**, called **SEDOM-DD** (**Sub-Events Detection on sOcial Media During Disasters**), aimed at detecting sub-events during disasters from social media data.



Belcastro *et al. J Big Data* (2021) 8:79
https://doi.org/10.1186/s40537-021-00467-1

Journal of Big Data

RESEARCH Open Access

Using social media for sub-event detection during disasters

Loris Belcastro[1], Fabrizio Marozzo[1*], Domenico Talia[1], Paolo Trunfio[1], Francesco Branda[1], Themis Palpanas[2,3] and Muhammad Imran[4]

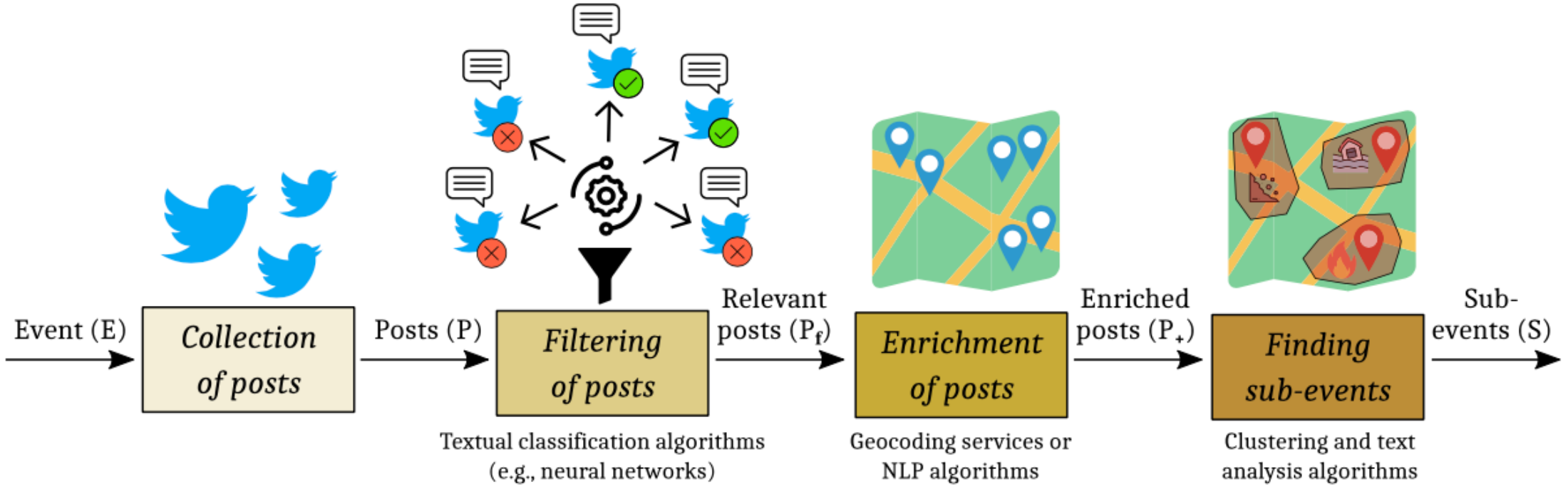
*Correspondence:
fmarozzo@dimes.unical.it
[1] University of Calabria, Rende, Italy
Full list of author information is available at the end of the article

Abstract

Social media platforms have become fundamental tools for sharing information during natural disasters or catastrophic events. This paper presents SEDOM-DD (Sub-Events Detection on sOcial Media During Disasters), a new method that analyzes user posts to discover sub-events that occurred after a disaster (e.g., collapsed buildings, broken gas pipes, floods). SEDOM-DD has been evaluated with datasets of different sizes that contain real posts from social media related to different natural disasters (e.g., earthquakes, floods and hurricanes). Starting from such data, we generated synthetic datasets with different features, such as different percentages of relevant posts and/or geotagged posts. Experiments performed on both real and synthetic datasets showed that SEDOM-DD is able to identify sub-events with high accuracy. For example, with a percentage of relevant posts of 80% and geotagged posts of 15%, our method detects the sub-events and their areas with an accuracy of 85%, revealing the high accuracy and effectiveness of the proposed approach.

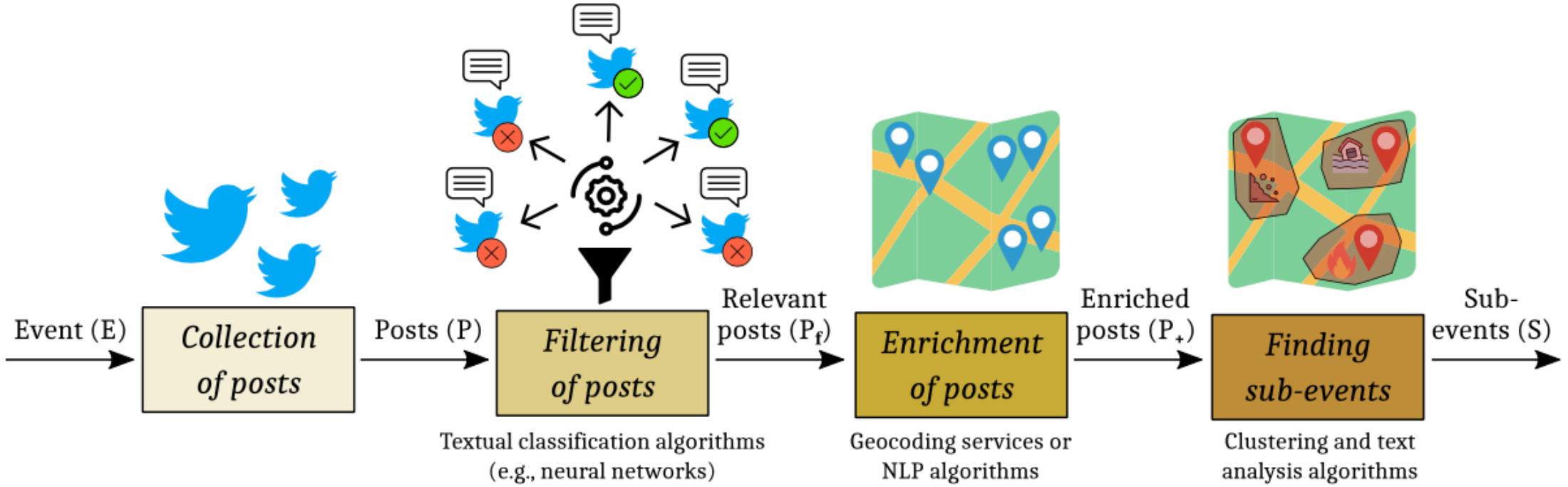
Keywords: Social media, Events detection, Natural disasters, Catastrophic events, Crisis computing, Disaster management, Mass emergencies, Earthquake

Execution flow of SEDOM-DD



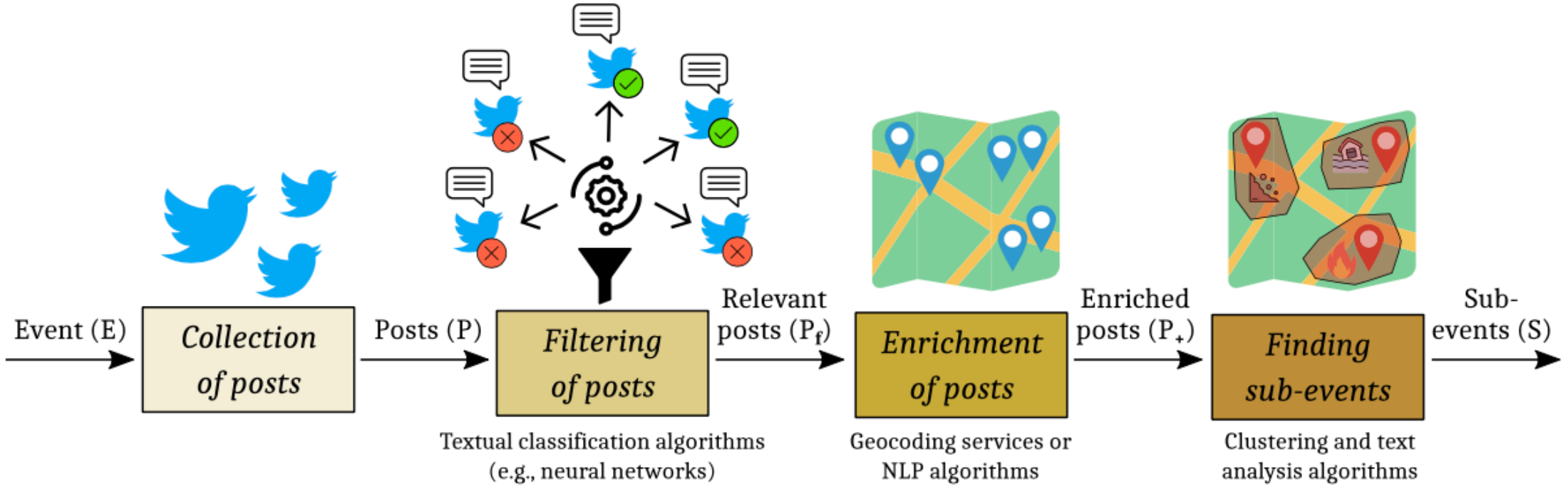
1) *Data collection:* given a disaster event and its impact areas, all the posts generated in the event's area are collected. These posts can be collected from social media platforms (e.g., Twitter) through queries based on keywords or locations.

Execution flow of SEDOM-DD



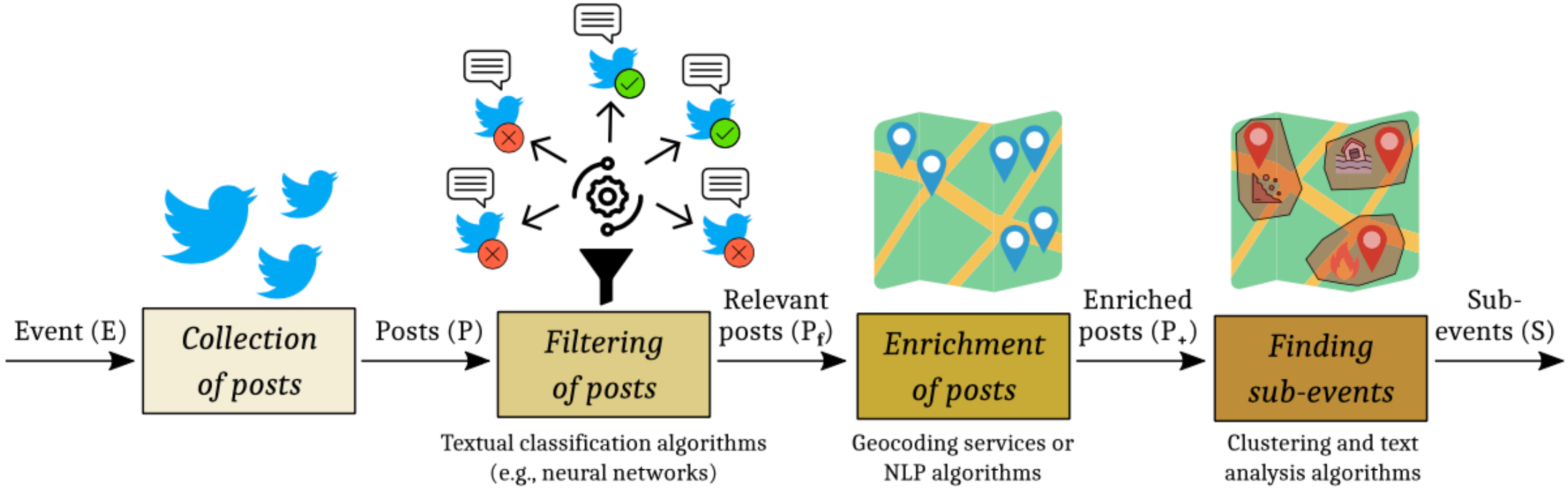
2) *Filtering of posts:* we use supervised machine learning techniques to identify relevant posts. Posts that refer to the disaster and that come from users who live in the affected area are relevant for analysis, and thus are maintained.

Execution flow of SEDOM-DD



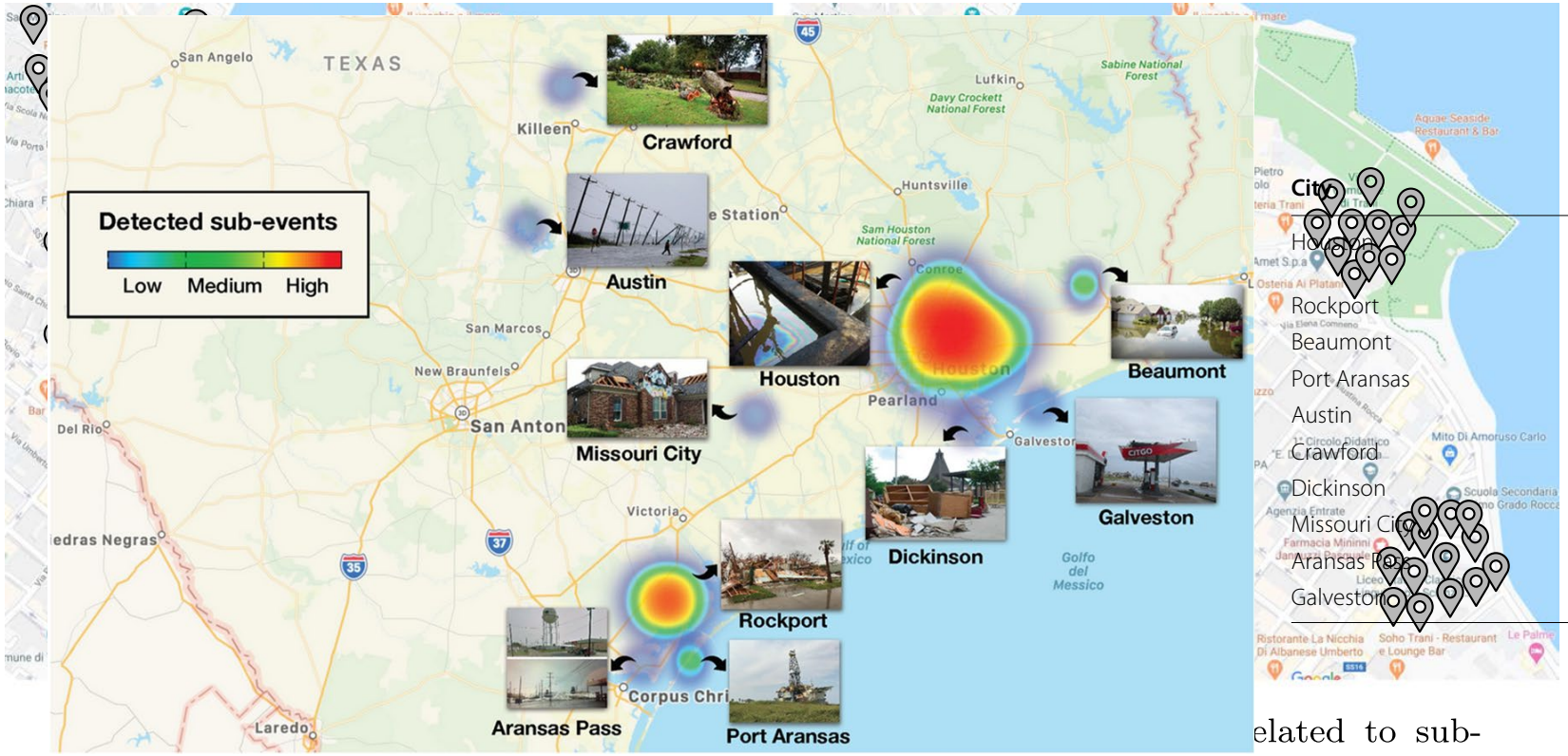
3) *Enrichment of posts:* since many posts are relevant for analysis but are not geotagged, the information contained in the text is used to estimate the coordinates of the location where such posts were created.

Execution flow of SEDOM-DD

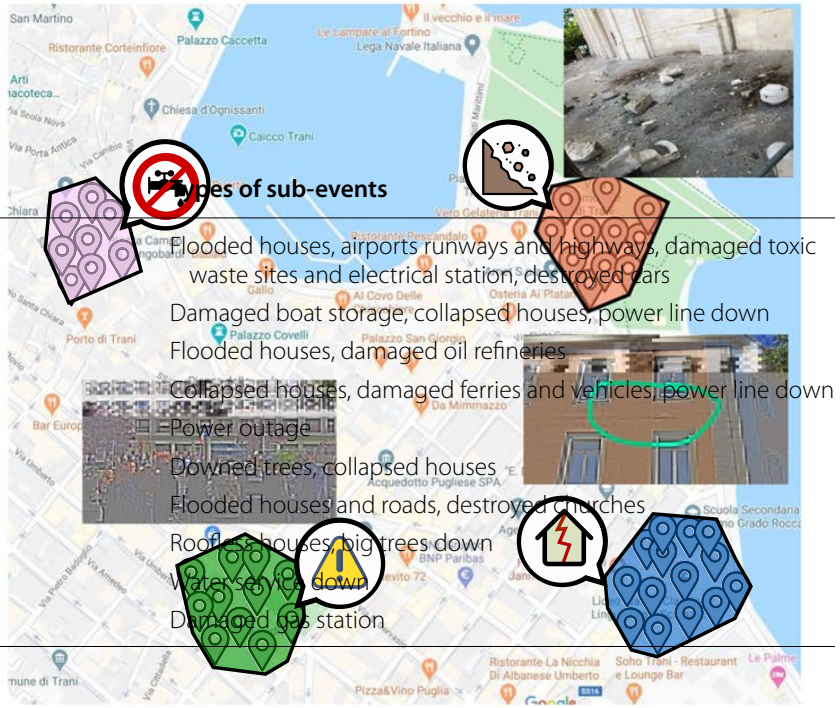


4) *Finding sub-events:* geotagged posts are analyzed and aggregated for finding clusters of posts mentioning a common problem (i.e., a specific sub-event that occurred in a certain area).

An example of using SEDOM-DD



elated to sub-events.

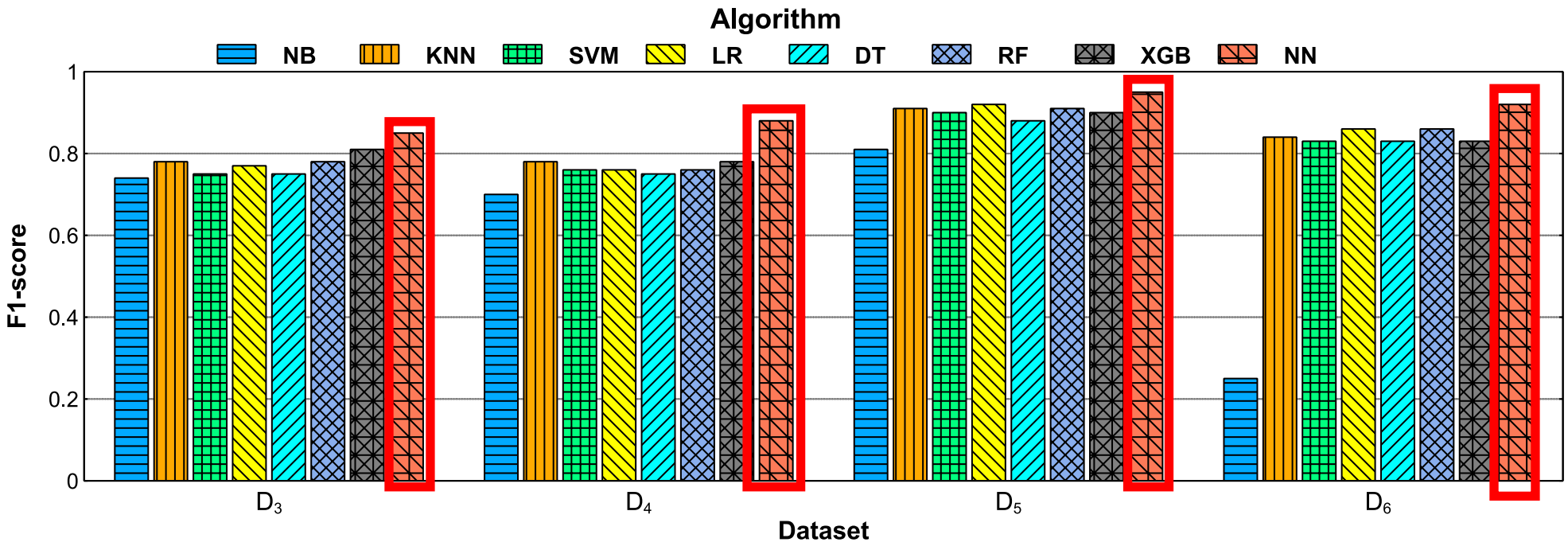


(c) Sub-events detection.

Experimental evaluation

- The algorithm based on **neural networks** was the most accurate with an accuracy of 83%, followed by the algorithms XGBoost (81%) and Random Forest (80%).

Algorithms	Acc	Prec	Rec	F1
Naïve Bayes	0.753	0.735	0.753	0.739
KNN	0.807	0.803	0.807	0.781
SVM	0.776	0.765	0.776	0.751
Logistic Regr.	0.790	0.773	0.790	0.766
Decision Tree	0.744	0.755	0.744	0.753
Random For.	0.795	0.794	0.790	0.783
XGBoost	0.815	0.812	0.815	0.809
Neural Net.	**0.830**	**0.826**	**0.864**	**0.845**



Use case: IOM-NN with sentiment analysis

- In recent years, the increasing use of social media also allows for the analysis of collective sentiment and the dynamics of public opinion.
- The proposed **methodology**, called **IOM-NN** (**Iterative Opinion Mining using Neural Networks**), aimed at discovering the political polarization of social media users during election campaigns characterized by the competition of political factions.
- Experimental results show the great effectiveness of IOM-NN, which was able to correctly identify the winning candidate in 10 out of 11 swing states, compared to the average of latest opinion polls before the election, which identified the winner in 9 out of 11 states.

Social Network Analysis and Mining (2022) 12:83
https://doi.org/10.1007/s13278-022-00913-9

ORIGINAL ARTICLE

Check for updates

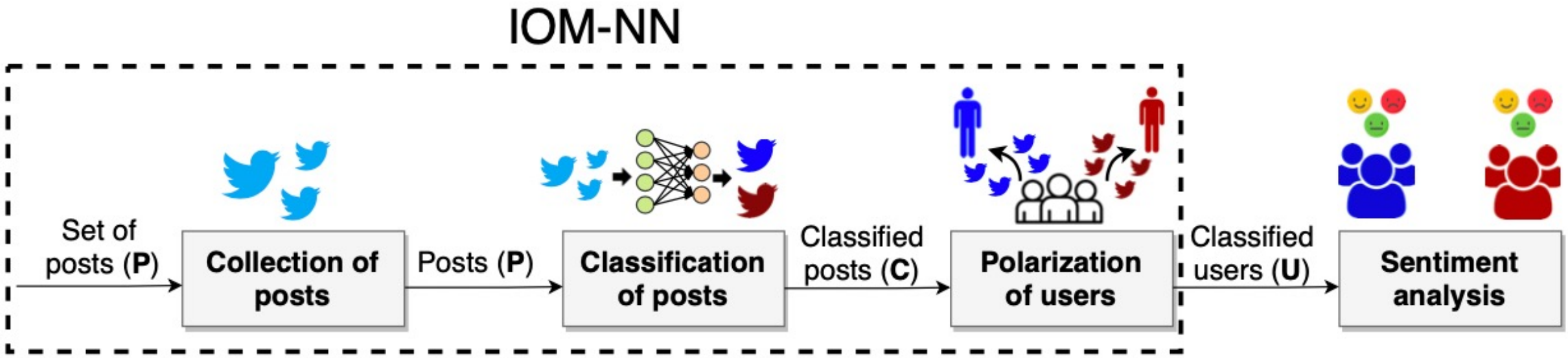
Analyzing voter behavior on social media during the 2020 US presidential election campaign

Loris Belcastro[1] · Francesco Branda[1] · Riccardo Cantini[1] · Fabrizio Marozzo[1] · Domenico Talia[1] · Paolo Trunfio[1]

Received: 20 January 2022 / Revised: 11 May 2022 / Accepted: 23 June 2022

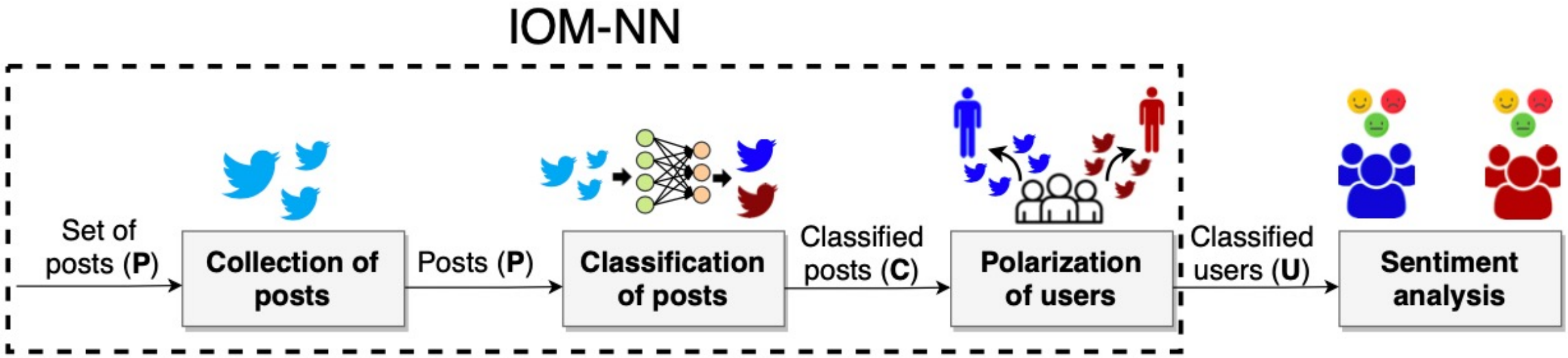
Abstract
Every day millions of people use social media platforms by generating a very large amount of opinion-rich data, which can be exploited to extract valuable information about human dynamics and behaviors. In this context, the present manuscript provides a precise view of the 2020 US presidential election by jointly applying topic discovery, opinion mining, and emotion analysis techniques on social media data. In particular, we exploited a clustering-based technique for extracting the main discussion topics and monitoring their weekly impact on social media conversation. Afterward, we leveraged a neural-based opinion mining technique for determining the political orientation of social media users by analyzing the posts they published. In this way, we were able to determine in the weeks preceding the Election Day which candidate or party public opinion is most in favor of. We also investigated the temporal dynamics of the online discussions, by studying how users' publishing behavior is related to their political alignment. Finally, we combined sentiment analysis and text mining techniques to discover the relationship between the user polarity and sentiment expressed referring to the different candidates, thus modeling political support of social media users from an emotional viewpoint.

Execution flow of the proposed methodology



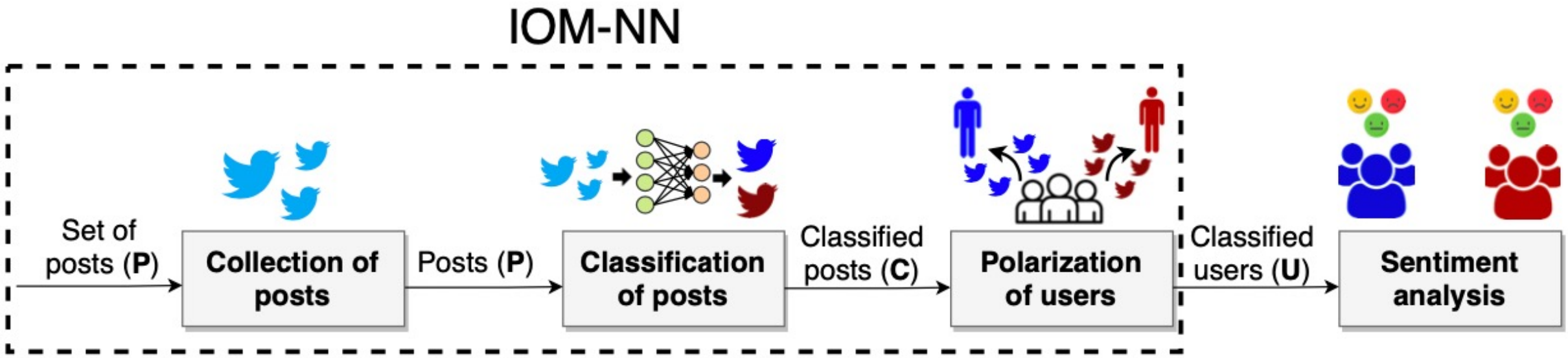
1) *Collection of posts:* data are gathered from social media by using a set of keywords related to the selected political event.

Execution flow of the proposed methodology



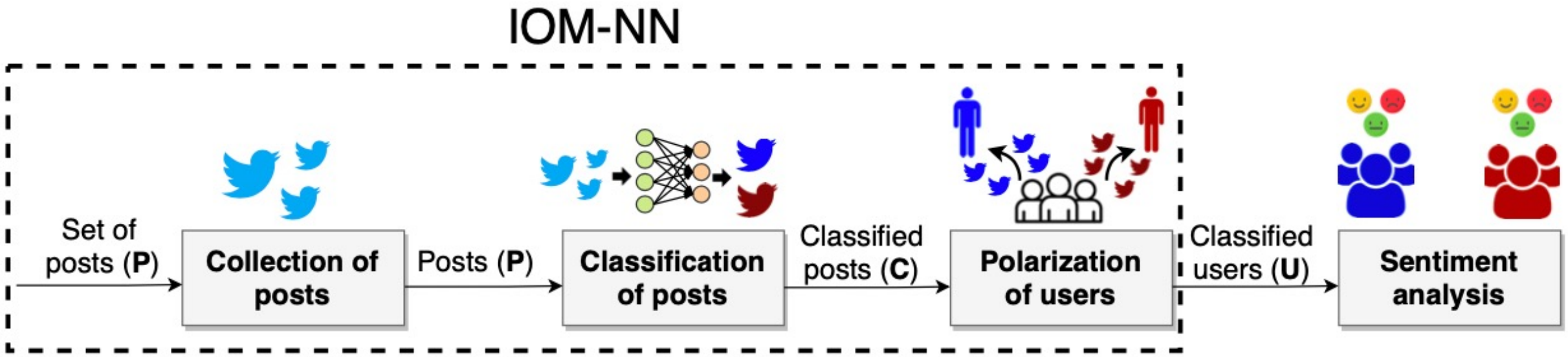
2) *Classification of posts:* the collected posts are classified in favor of a faction according to the detected political support.

Execution flow of the proposed methodology



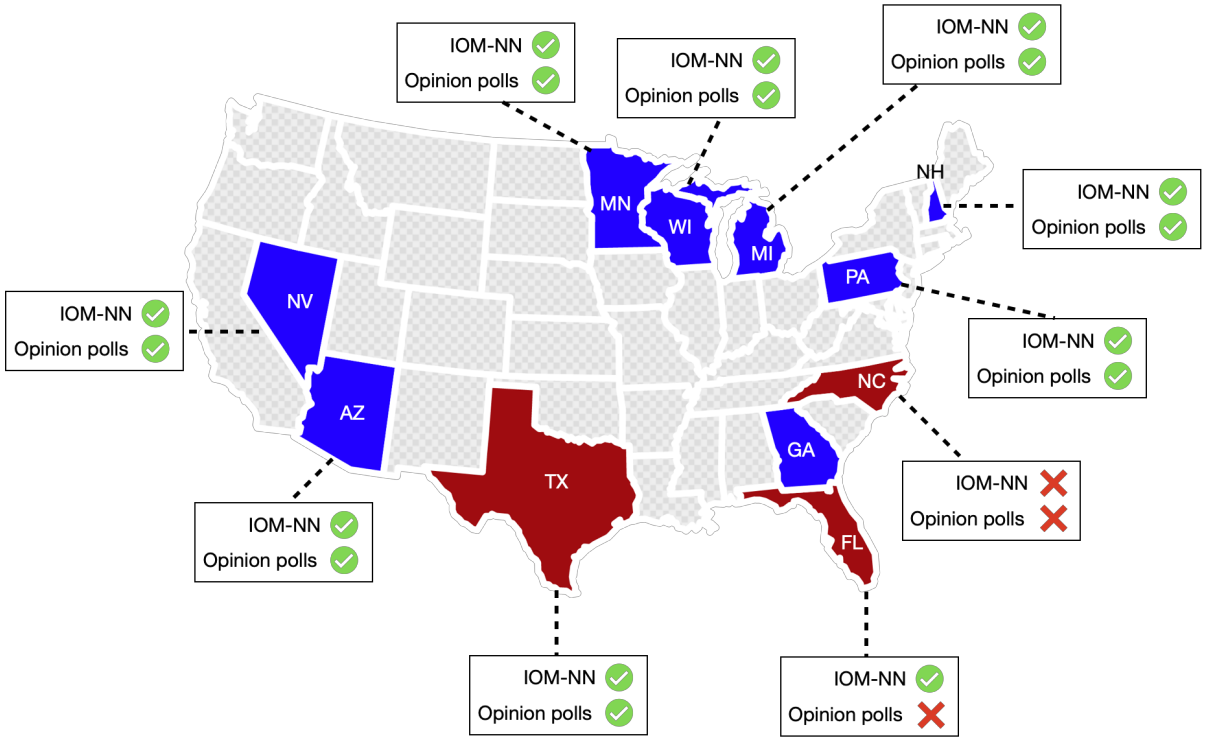
3) *Polarization of users*: the classified posts are analyzed for determining the polarization of users towards a faction.

Execution flow of the proposed methodology



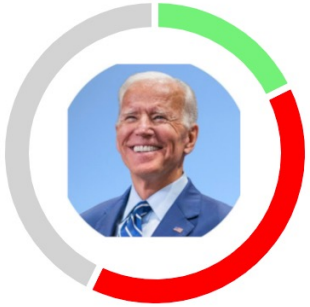
4) *Sentiment analysis:* the polarized posts are exploited for investigating the relationship between the political orientation of users and the sentiment they expressed in referring to the different candidates.

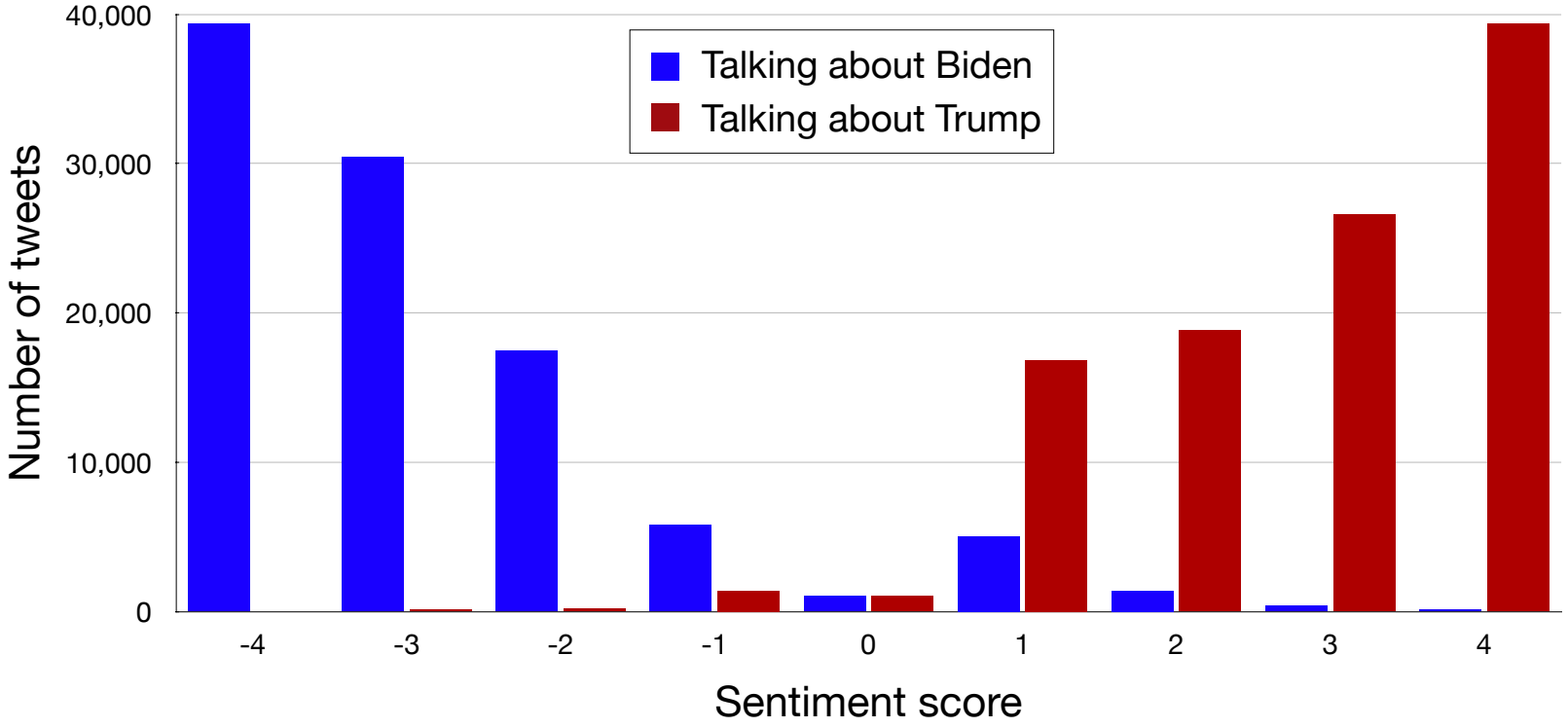
Experimental evaluation

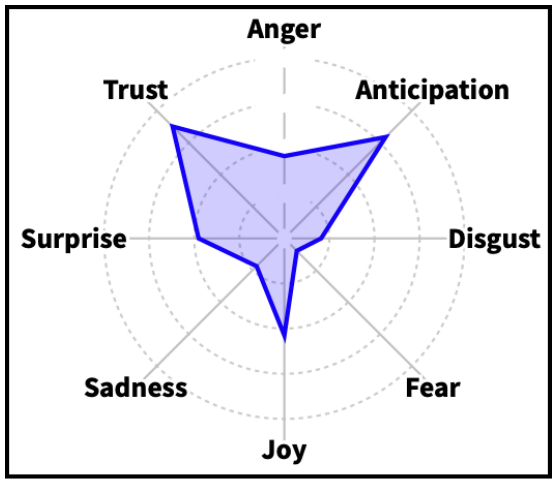


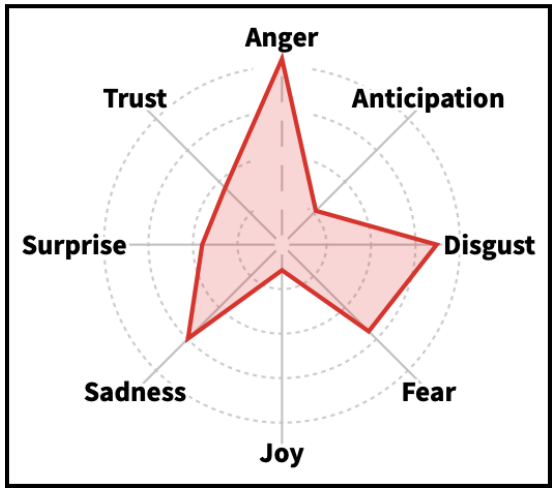
State	Real percentages		Opinion polls		IOM-NN	
	B	*T*	*B*	*T*	*B*	*T*
Arizona	49.4	49.1	**48.0**	45.8	**50.2**	48.3
Florida	47.9	51.2	48.7	46.0	48.0	**51.1**
Georgia	49.5	49.2	**47.6**	47.4	**52.7**	46.0
Michigan	50.6	47.8	**49.9**	44.4	**55.4**	43.0
Minnesota	52.4	45.3	**51.6**	41.8	**55.1**	42.6
Nevada	50.1	47.7	**49.4**	44.4	**49.8**	48.0
New Hampshire	52.7	45.4	**53.4**	42.4	**50.9**	47.3
North Carolina	48.6	49.9	47.8	47.5	56.6	41.9
Pennsylvania	50.0	48.8	**49.4**	45.7	**55.7**	43.1
Texas	46.5	52.1	47.5	**48.8**	46.1	**52.5**
Wisconsin	49.4	48.8	**52.0**	42.8	**56.3**	41.9
Correctly classified	-		**9/11**		**10/11**	
Tweets	-		-		670,451	
Users	-		$\approx$ 11,000		57,116	
Avg. Acc	-		0.82		0.91	

Experimental evaluation

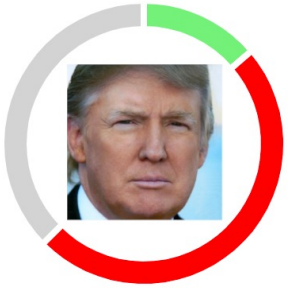


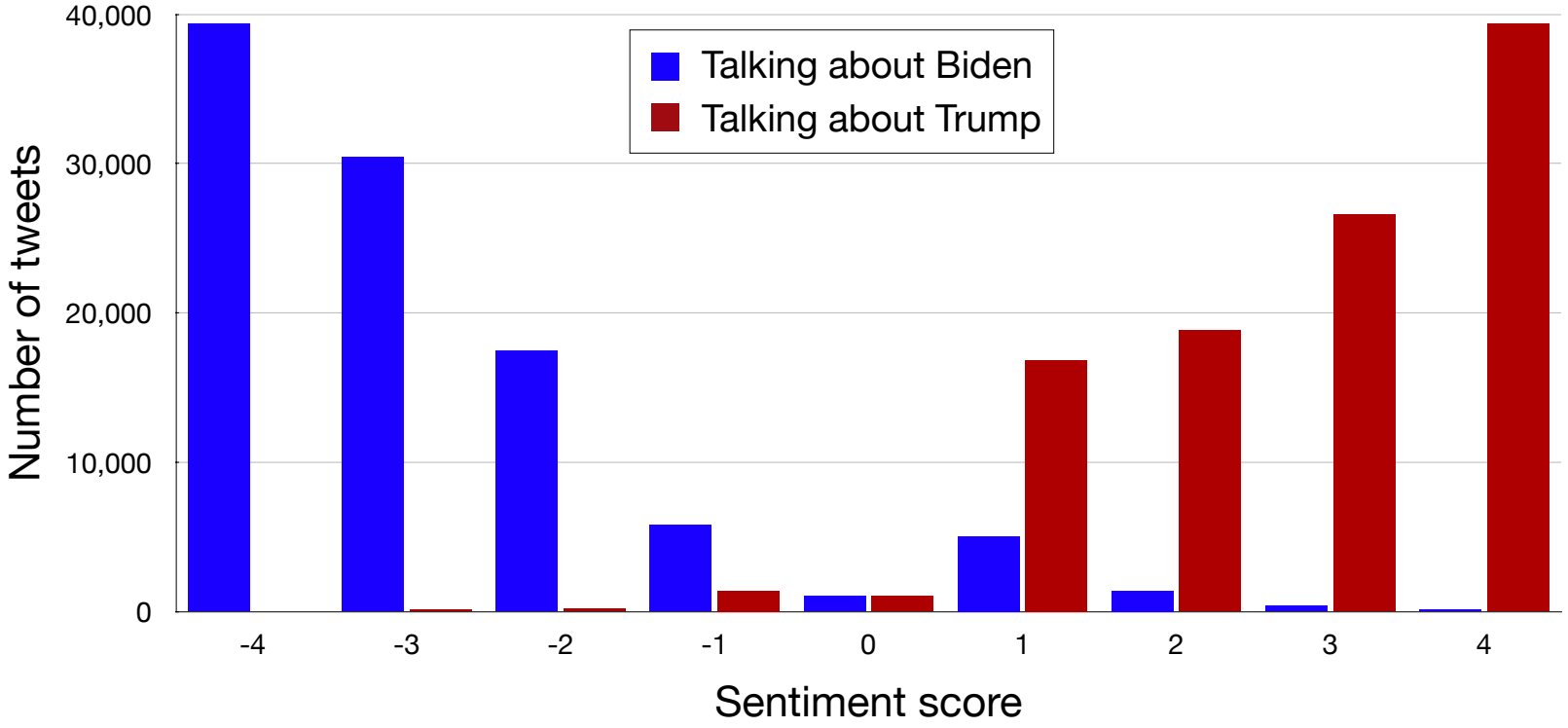


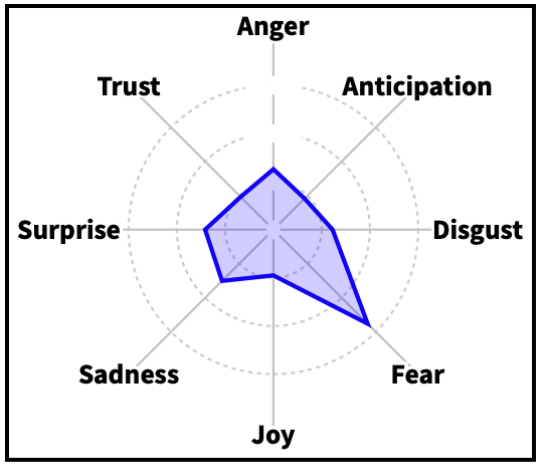


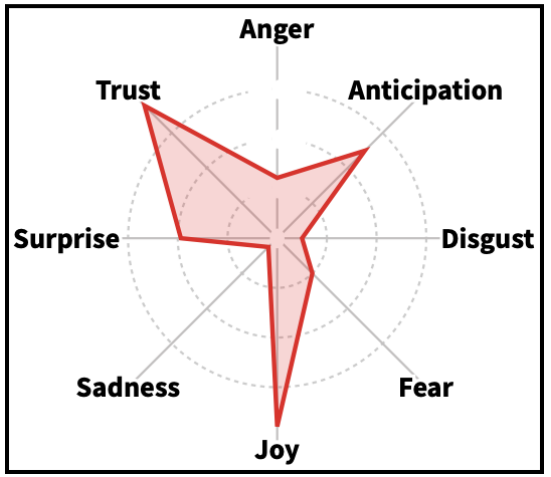


Experimental evaluation









Key messages

- Open Data are fundamental for i) conducting real-time situation analysis; ii) facilitating coordination and collaboration between national and local governments; iii) securing public trust in government through better transparency and improved communications; iv) countering misinformation.

- Governments at all levels need to build up their capacities to overcome data silos and skill gaps to address diverse dimensions of data governance. These range from ensuring the consistency of data collection to enhancing government accountability in sharing data and strengthening data quality and data security for a timely and proper response.

- It is important to adopt a holistic and whole-of-government approach to data governance with the engagement of all stakeholders and partners across sectors. Building data partnerships with all stakeholders can help leverage digital solutions driven by the private sector, promote publication of data produced by civil society organizations on open government data portals or open government data on non-government data portals, and support data sharing among all stakeholders.

Questions