



GLOBAL DATABASE OF EVENT, LANGUAGE AND TONE

INTRODUCTION AND DEDUPLICATION

[HTTPS://WWW.GDELTPROJECT.ORG/](https://www.gdelproject.org/)

Dr. Deepti Joshi

Department of Cyber and Computer Sciences

The Citadel, The Military College of South Carolina

GDELT PROJECT OVERVIEW

- Monitors the world's broadcast, print, and web news
- Reports in at least 65 languages are translated into English and processed in near-real-time
- World-wide coverage
- Dating back to January 1, 1978
- Updates every 15 minutes



WHY DO WE CARE ABOUT GDELT?

- **Our goal** – *anticipating social unrest* by understanding local contexts and discovering the **fuel** and **triggers** for unrest within any given spatial context (country, state, county/district..)
- GDELT is a source for **ground truth** for historic data
- In order to determine the long-term fuel factors that lead to unrest, the **number of unrest events** reported by GDELT becomes the **dependent variable** for social scientists

GDELT PROJECT- TYPES OF DATA PRODUCED

- 3 streams of data being produced:
 - one codifying **physical activities** around the world in over 300 categories
 - one recording the **people, places, organizations**, millions of **themes** and thousands of **emotions** underlying those events and their interconnections
 - one codifying the **visual narratives** of the world's news imagery

GDELT EVENT DATABASE

- The GDELT Event Database records
 - over 300 categories of physical activities around the world using the CAMEO (Conflict and Mediation Event Observations) codebook (<http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf>)
 - Up to 59 attributes captured for each event including approximate location of the event and the actors involved

CAMEO: PRIMARY EVENT CATEGORIES

2.1	MAKE PUBLIC STATEMENT
2.2	APPEAL
2.3	EXPRESS INTENT TO COOPERATE
2.4	CONSULT
2.5	ENGAGE IN DIPLOMATIC COOPERATION
2.6	ENGAGE IN MATERIAL COOPERATION
2.7	PROVIDE AID
2.8	YIELD
2.9	INVESTIGATE
2.10	DEMAND
2.11	DISAPPROVE
2.12	REJECT
2.13	THREATEN
2.14	PROTEST
2.15	EXHIBIT MILITARY POSTURE
2.16	REDUCE RELATIONS
2.17	COERCE
2.18	ASSAULT
2.19	FIGHT
2.20	ENGAGE IN UNCONVENTIONAL MASS VIOLENCE

GDELT EVENT DATABASE

GOAL

- Take a sentence like "*The United States criticized Russia yesterday for deploying its troops in Crimea, in which a recent clash with its soldiers left 10 civilians injured*" and
- transform into three structured database entries, recording
 - *US CRITICIZES RUSSIA,*
 - *RUSSIA TROOP-DEPLOY UKRAINE (CRIMEA),* and
 - *RUSSIA MATERIAL-CONFLICT CIVILIANS (CRIMEA)*

GDELT EVENT DATABASE

RECORD ATTRIBUTES

GOBALEVENTID	SQLDATE	MonthYear	Year
FractionDate	Actor1Code	Actor1Name	Actor1CountryCode
Actor1KnownGroupCode	Actor1EthnicCode	Actor1Religion1Code	Actor1Religion2Code
Actor1Type1Code	Actor1Type2Code	Actor1Type3Code	Actor2Code
Actor2Name	Actor2CountryCode	Actor2KnownGroupCode	Actor2EthnicCode
Actor2Religion1Code	Actor2Religion2Code	Actor2Type1Code	Actor2Type2Code
Actor2Type3Code	IsRootEvent	EventCode	EventBaseCode
EventRootCode	QuadClass	GoldsteinScale	NumMentions
NumSources	NumArticles	AvgTone	Actor1Geo_Type
Actor1Geo_FullName	Actor1Geo_CountryCode	Actor1Geo_ADM1Code	Actor1Geo_Lat
Actor1Geo_Long	Actor1Geo_FeatureID	Actor2Geo_Type	Actor2Geo_FullName
Actor2Geo_CountryCode	Actor2Geo_ADM1Code	Actor2Geo_Lat	Actor2Geo_Long
Actor2Geo_FeatureID	ActionGeo_Type	ActionGeo_FullName	ActionGeo_CountryCode
ActionGeo_ADM1Code	ActionGeo_Lat	ActionGeo_Long	ActionGeo_FeatureID
DATEADDED	SOURCEURL		

GDELT EVENT DATABASE

SAMPLE DATA - 1

GBALEVENTID	SQLDATE	MonthYear	Year	FractionDate	Actor1Code	Actor1Name	Actor1CountryCode
962238012	20210101	202101	2021	2021.0027			
962239581	20210101	202101	2021	2021.0027	IND	SRINAGAR	IND
962242626	20210101	202101	2021	2021.0027	AGR	FARMER	

Actor1Type1Code	Actor1Type2Code	Actor1Type3Code	Actor2Code	Actor2Name	Actor2CountryCode
			JUD	MAGISTRATE	
			COP	POLICE	
AGR			IND	DELHI	IND

GDELT EVENT DATABASE

SAMPLE DATA - 2

Actor2Type1Code	IsRootEvent	EventCode	EventBaseCode	EventRootCode	QuadClass	GoldsteinScale	NumMentions	NumSources
JUD	1	140	140	14	3	-6.5	10	1
COP	1	140	140	14	3	-6.5	20	1
	1	140	140	14	3	-6.5	19	4

NumArticles	AvgTone	Actor1Geo_Type	Actor1Geo_FullName	Actor1Geo_Country Code	Actor1Geo_ADM1 Code	Actor1Geo_Lat	Actor1Geo_Long
10	-3.82513661202	0					
20	-7.71276595744	4	New Delhi, Delhi, India	IN	IN07	28.6	77.2
19	-3.07924278361	4	Delhi, Delhi, India	IN	IN07	28.6667	77.2167

GDELT EVENT DATABASE

SAMPLE DATA - 3

Actor1Geo_FeatureID	Actor2Geo_Type	Actor2Geo_FullName	Actor2Geo_CountryCode	Actor2Geo_ADM1Code	Actor2Geo_Lat	Actor2Geo_Long	Actor2Geo_FeatureID
	4	Chaparia, Jharkhand, India	IN	IN38	24.05	86.9	-2092901
-2106102	4	New Delhi, Delhi, India	IN	IN07	28.6	77.2	-2106102
-2094230	4	Delhi, Delhi, India	IN	IN07	28.6667	77.2167	-2094230

ActionGeo_Type	ActionGeo_FullName	ActionGeo_CountryCode	ActionGeo_ADM1Code	ActionGeo_Lat	ActionGeo_Long	ActionGeo_FeatureID
4	Chaparia, Jharkhand, India	IN	IN38	24.05	86.9	-2092901
4	New Delhi, Delhi, India	IN	IN07	28.6	77.2	-2106102
4	Delhi, Delhi, India	IN	IN07	28.6667	77.2167	-2094230

DATEADDED	SOURCEURL
20210101	https://timesofindia.indiatimes.com/city/varanasi/3-diagnostic-centres-sealed-licences-cancelled-for-laxity/articleshow/80053054.cms
20210101	https://www.dawn.com/news/1598998/kashmiri-families-dispute-indian-claims-slain-youths-were-fighters
20210101	https://www.chinadaily.com.cn/a/202101/01/WS5fee7a7ba31024ad0ba9ff1f.html

PROBLEMS WITH GDELT (WITHIN OUR CONTEXT)

- **Our goal** – get the number of unrest (protest) events within a country at the district (county) level and a given time frame from GDELT
- **Issues in achieving our goal:**
 - Large amount of **duplication** was discovered within the dataset for India for 2016 protest events
 - GDELT reported 59,422 protest events in India in 2016, while other popular datasets reported 9,600 (ACLED) and 8,500 (ICEWS)
- Thus, in order to make GDELT more usable for our purposes, we need a strategy to **de-duplicate GDELT**

GDELT DE-DUPLICATION - 1

- **Preliminary Observation**

- The same URL appears multiple times within the data

Article mentioned 186 times

https://www.kptv.com/news/university-of-oregon-to-require-students-staff-covid-19-vaccinations-for-fall/article_12318c40-b1dd-11eb-866b-e3f9afd5b0fc.html

- Really short
- Only has 4 paragraphs
- No pictures separating text

Article mentioned 184 times

<https://www.dailymail.co.uk/news/article-9583213/White-House-warn-Israel-media-not-harmed-organizations-demand-explanation-airstrike.html>

- Has a lot of pictures from different locations, with explanations attached
- Pictures are depicting protest in multiple places

Article mentioned 1 time

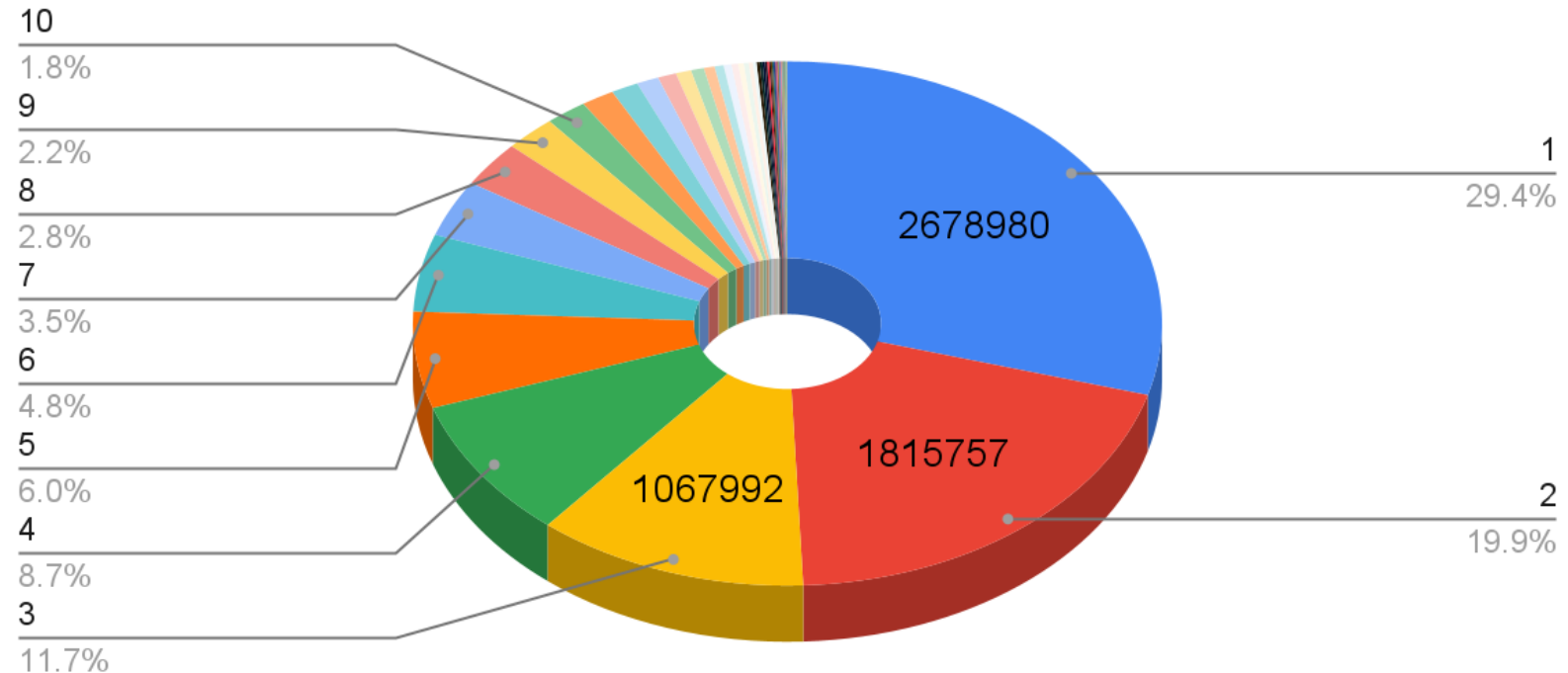
<https://www.nytimes.com/live/2021/05/15/world/israel-gaza-updates#us-envoy-israel-gaza>

- Has Multiple articles following it
- Contains a lot of paragraphs
- Has a plot of pictures separating different articles

GDELT DE-DUPLICATION - 2

Number Of Times a Source URL Appears in a day

Whole World for One Year June 14th, 2020 to June 13th, 2021



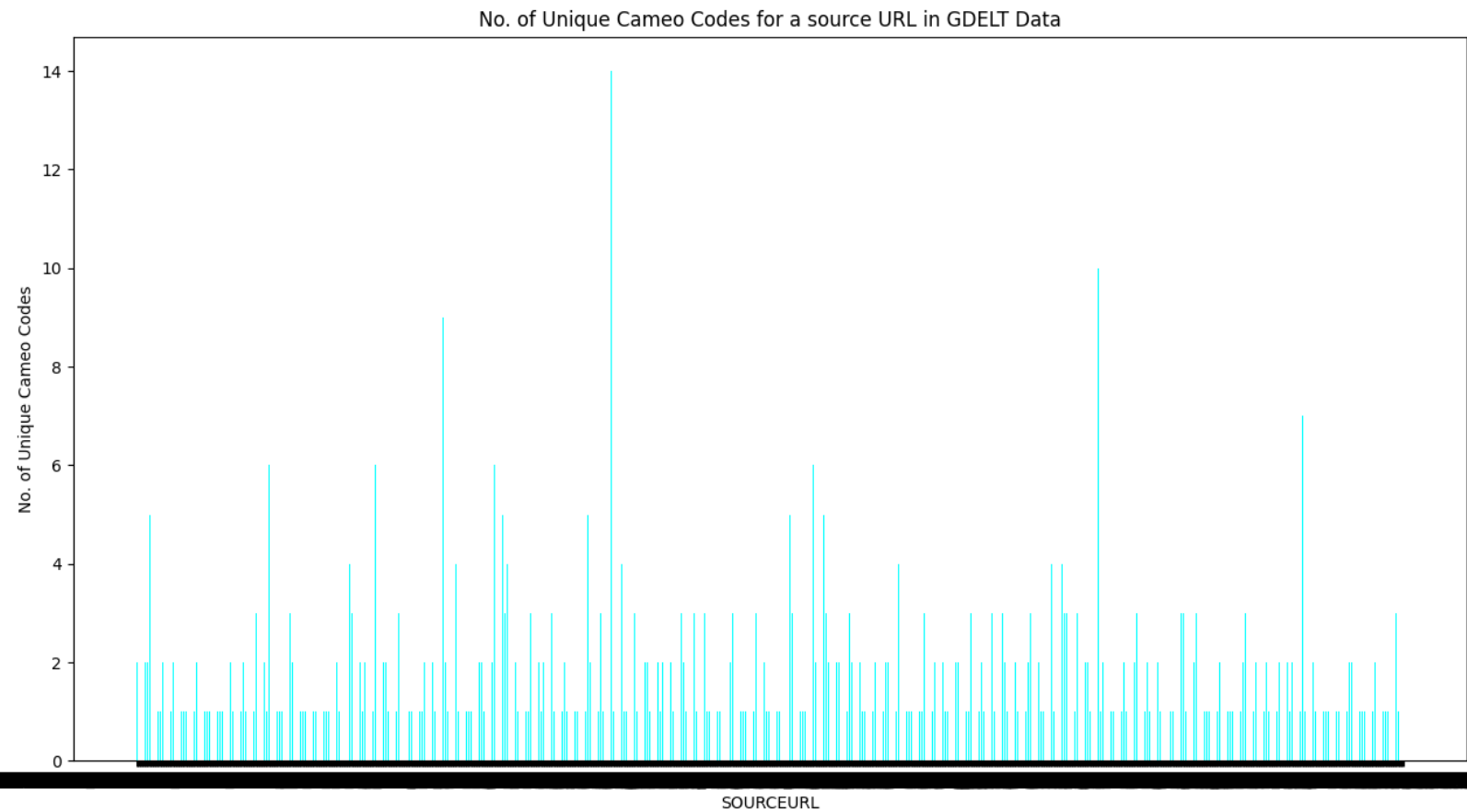
GDELT DE-DUPLICATION - 3

Why does the same URL appear more than once?

- Ignoring the Global event ID no 2 rows are exactly the same.
- Several cameo codes are related in that both can easily apply to the same article. Example: Russia demands aid and Germany refuses to provide aid are two different cameo codes and could cause another row to appear in GDELT
- Longer articles with multiple sentences/paragraphs reporting on related details to the event have multiple cameo codes.

GDELT DE-DUPLICATION - 4

India Events from May 26th, 2021



GDELT DE-DUPLICATION - 5

- Article with 14 Cameo Codes and 29 rows in GDELT
- URL: <http://www.kashmirtimes.com/newsdet.aspx?q=110070>

Home / Columnist

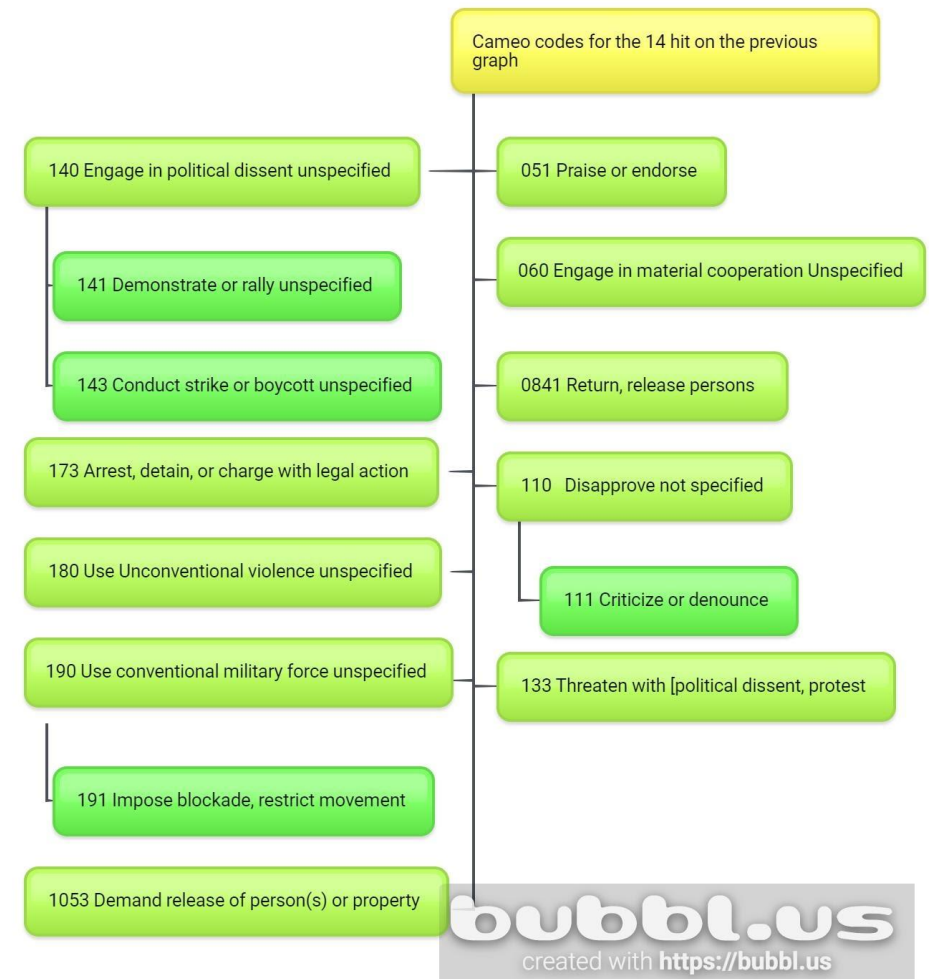
HOW PROTESTERS FORCED 'STUBBORN' MODI GOVERNMENT, REFUSING CRITICISM, TO THE TABLE

Kashmir Times. Dated: 5/26/2021 11:17:10 AM

"India has in recent years been home both to mass protests and heavy state repression. In 2019, the focus was on the struggles of people against discriminatory new citizenship laws that targeted India's sizeable Muslim minority, and the repression of the formerly semi-autonomous region of Jammu and Kashmir."

"For many farmers, however, the new laws seemed to offer a direct threat to their way of life. India's 'Green Revolution', introduced in the late 1960s with the northern state of Punjab at its core, saw the introduction of higher-yielding crops, the intensive and subsidised use of irrigation, fertilisers and pesticides....."

"Images of brutality went round the world, showing police in riot gear hitting defenceless farmers, many of them of advanced age, with batons. The government mobilised a swift backlash to this humiliation. Heavy charges of sedition and terrorism were brought against protesters, activists and journalists."



GDELT DE-DUPLICATION - 6

- De-duplication Strategy 1 (DDS1)

- Merge rows with the same URL
 - Leads to lots of different types of events being represented as a single event
 - Leads to the 29 events from the previous slide to be collapsed into one event
 - Not a valid strategy

- De-duplication Strategy 2 (DDS2)

- Merge rows with the same URL, Date, Location (Latitude and Longitude) and Event Root Code
 - Leads to the 29 events from the previous slide to be collapsed into 10 events

GDELT DE-DUPLICATION - 7

Date	Original Count	DDS1	DDS2
20210517	118507	25044	65352
20210518	129875	28715	72904
20210519	143205	30316	79499
20210520	140502	29925	77881
20210521	130338	27600	71884
20210522	79159	15680	42819
20210523	67373	13603	37041
20210524	116818	24373	64848

GDELT DE-DUPLICATION - 8

- De-duplication Strategy 3 (DDS3)

- Merge rows with the same URL, Location (Latitude and Longitude) and Event Root Code within a year
 - The logic here is that sometimes the article maybe updated across multiple days and thus the same event may be reported multiple times

- De-duplication Strategy 4 (DDS4)

- First apply DDS3
- Next, merge rows with the same Actor1, Actor 2, Date, Location (Latitude and Longitude) and Event Root Code

GDELT DE-DUPLICATION - 9

- De-duplication Strategy 5 (DDS5)
 - First apply DDS3
 - Next, merge rows with the same Actor1, Actor 2, Location (Latitude and Longitude) and Event Root Code

GDELT DE-DUPLICATION - 10

Protest Events (Cameo Root Code 14) reported in GDELT for the year 2020

	Original count	DDS2	DDS3	DDS4	DDS5	% of events remaining
India	48693	31140	30898	17920	12860	26.41
Afghanistan	1988	1329	1320	851	617	31.04
Mexico	2817	1899	1882	1204	857	30.42
Syria	2590	1714	1704	1188	863	33.32
Ukraine	2079	1377	1371	962	671	32.28
United States	218947	148861	148020	84825	46153	21.08
Pakistan	12486	7950	7906	5050	3700	29.63

GDELT DE-DUPLICATION - 11


- DDS4 and DDS5 allow us to group together articles that report about the same event but include different details.
- For example: The following two URLs now belong to the same event
 - <https://economictimes.indiatimes.com/news/politics-and-nation/intelligence-failure-to-blame-for-red-fort-violence-congress-adhir-chowdhury/articleshow/80513222.cms>
 - <https://www.news18.com/news/india/actor-deep-sidhu-accused-of-instigating-protesters-to-storm-red-fort-named-in-fir-3352541.html>

Intelligence failure to blame for Red Fort violence: Congress' Adhir Chowdhury

PTI • Last Updated: Jan 28, 2021, 08:35 PM IST


SHARE FONT SIZE SAVE PRINT

Synopsis
Thousands of farmers, protesting against the new agricultural laws, had clashed with the police on January 26. Many of them, driving tractors, reached the Red Fort and scaled the walls of the monument.



Kolkata: Leader of the Congress in Lok Sabha Adhir Chowdhury claimed on Thursday that intelligence failure on the part of the Delhi Police was to blame for the chaos and violence at Red Fort on January 26.

Actor Deep Sidhu, Accused of Instigating Protesters to Storm Red Fort, Named in FIR



The national capital on Tuesday witnessed clashes between protesters and police during the tractor parade by farmers to press their demand of repealing three new agri laws.

● NEWS18.COM
● LAST UPDATED: JANUARY 28, 2021, 07:58 IST
● FOLLOW US ON: [Facebook](#) [Twitter](#) [Instagram](#)
[Telegram](#) [Google News](#)

Deep Sidhu has been named as an accused in the Republic Day violence.

The Delhi Police has named actor Deep Sidhu and gangster-turned-social activist Lakha Sidhana in an FIR

NEXT STEPS - 1

1. Validation

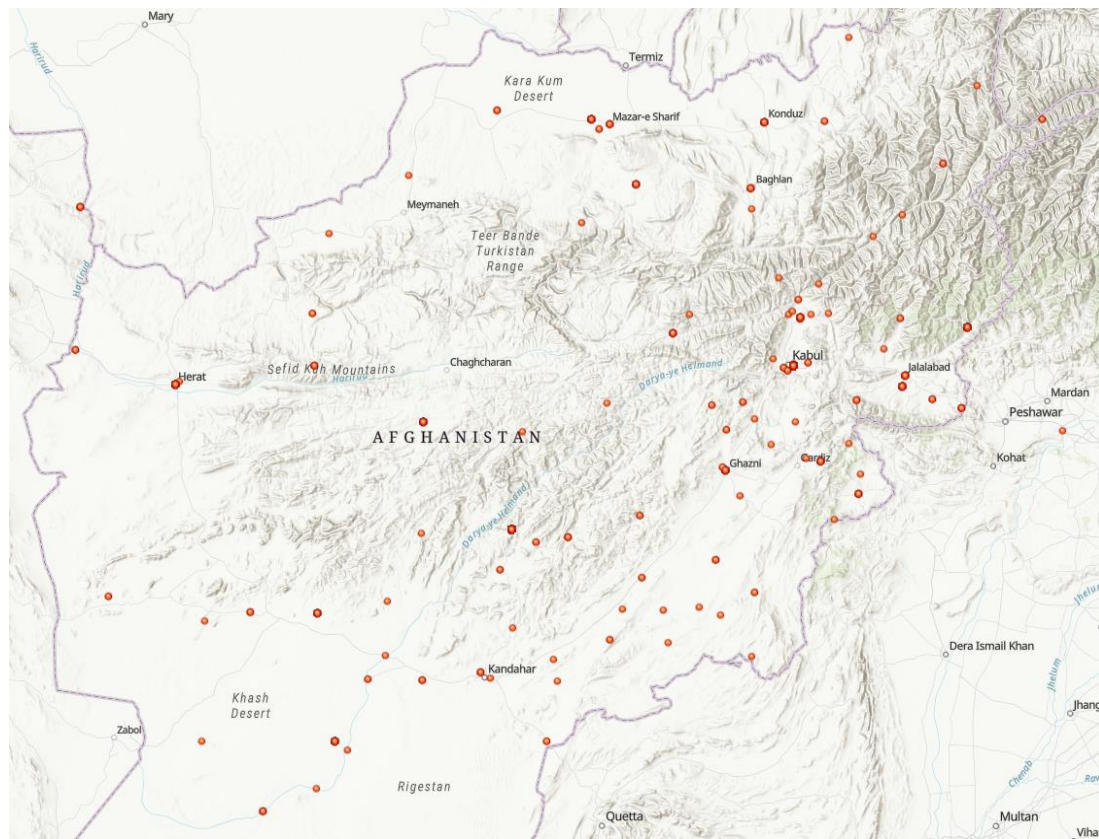
- Validate the de-duplication process by analyzing the results obtained manually
- Validate the de-duplication process by comparing the number of events with those reported in other datasets – ACLED and ICEWS

2. Content Analysis

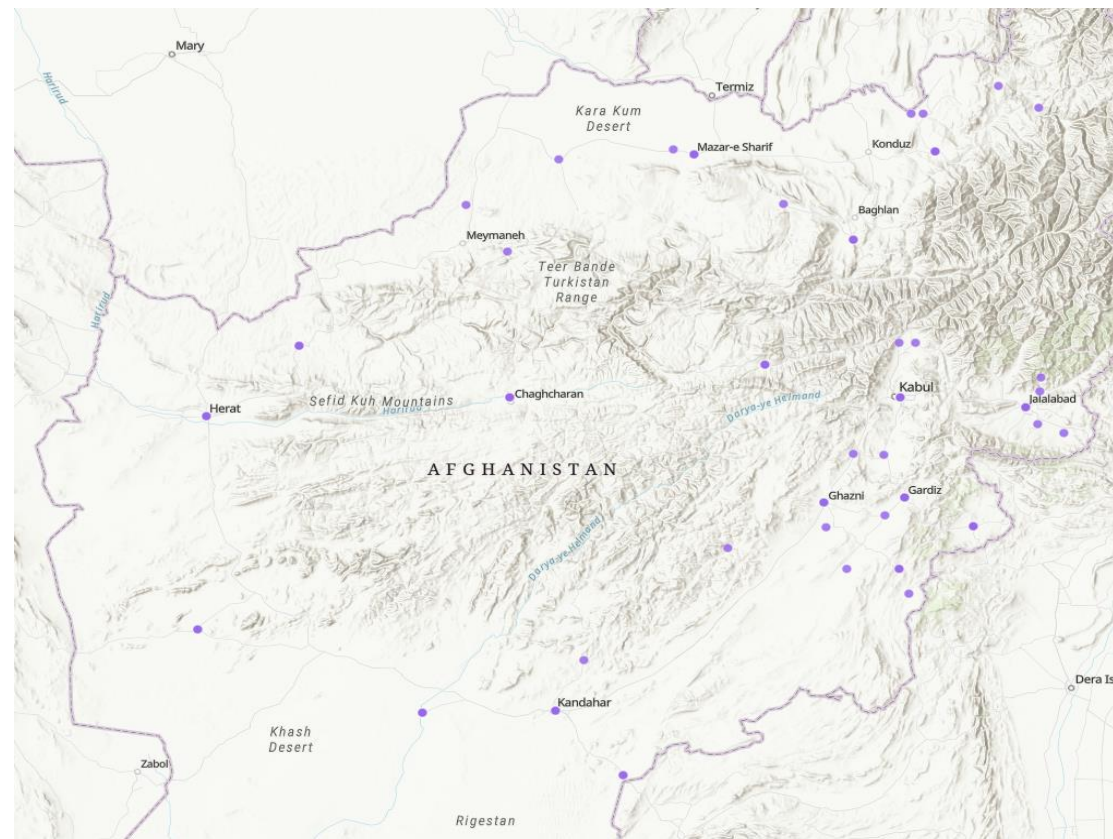
- Compare the de-duplicated events from GDELT with those reported in ACLED and ICEWS – Are the same events reported in both?

NEXT STEPS - 2

COMPARING GDELT WITH ACLED



GDELT



ACLED