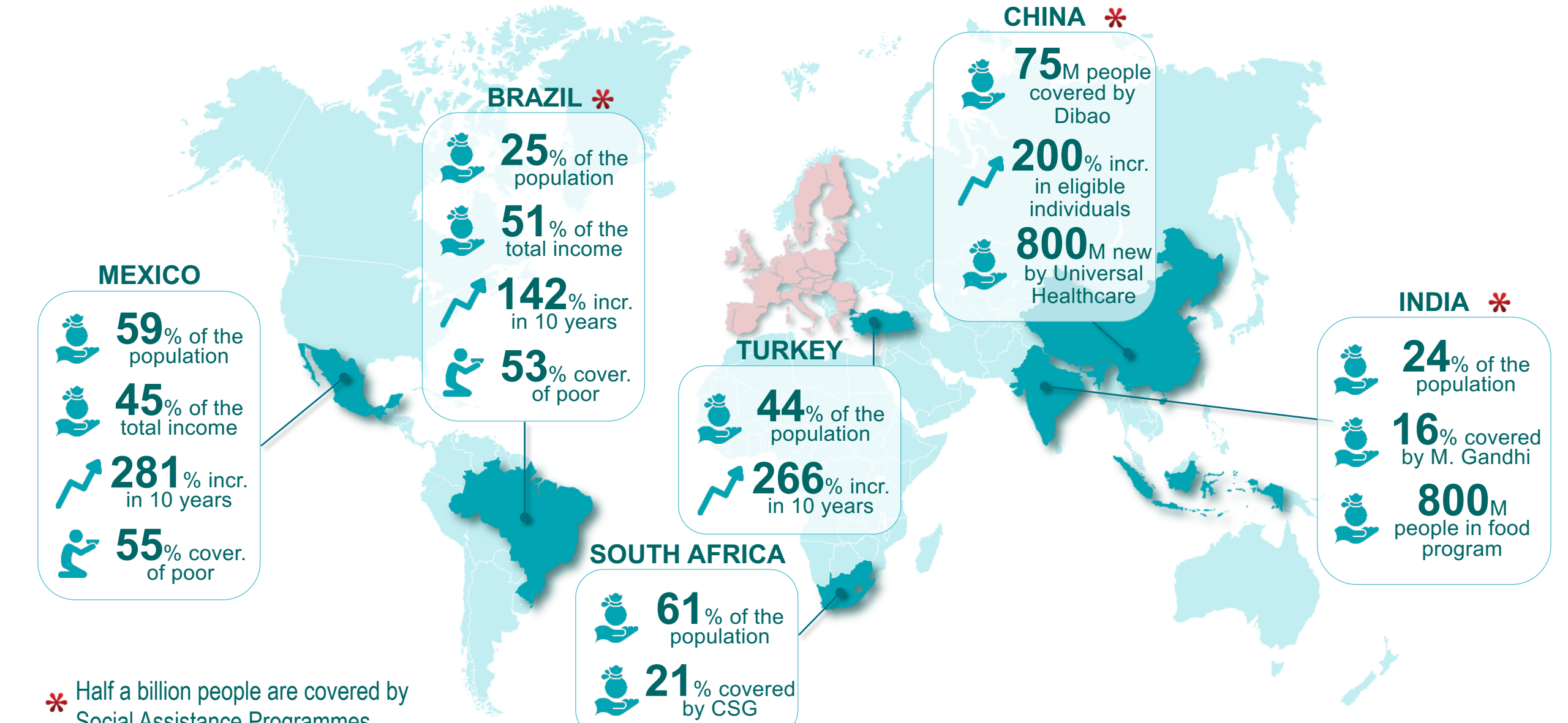
# Cross-context News Corpus for Protest Events related Knowledge Base Construction

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Osman Mutlu, Çağrı Yoltar, and Fırat Duruşan, Burak Gürel
Koc University, Istanbul, Turkey

Automatic Knowledge Base Construction Conference
June 22-24, 2020

# EMERGING WELFARE

## Welfare State Expansion in Emerging Markets since the 1990s



### BRAZIL *

- **25**% of the population
- **51**% of the total income
- **142**% incr. in 10 years
- **53**% cover. of poor

### MEXICO

- **59**% of the population
- **45**% of the total income
- **281**% incr. in 10 years
- **55**% cover. of poor

### TURKEY

- **44**% of the population
- **266**% incr. in 10 years

### SOUTH AFRICA

- **61**% of the population
- **21**% covered by CSG

### CHINA *

- **75**M people covered by Dibao
- **200**% incr. in eligible individuals
- **800**M new by Universal Healthcare

### INDIA *

- **24**% of the population
- **16**% covered by M. Gandhi
- **800**M people in food program

* Half a billion people are covered by Social Assistance Programmes

# Relation between Welfare Expansion and Protests

- What is the cause of the welfare state expansion in emerging markets?
  - Do protests play any role?
- **Protest (Contentious Politics)**: Any form of grassroots political action, and actions of political and non-governmental organizations that are aimed at mobilizing the public in the name of political demands and grievances.
  - Street demonstrations
  - Industrial actions
  - Group clashes
  - Actions targeting officials or civilians
  - The time and place of an event should be understandable from the text
  - Plans, threats, etc. do not count!
- Capturing cross-context variability of the protest events requires us to study them in detail across contexts.
- For instance, the terms "**bandh**" and "**idol immersion**" are event types that are specific to India and not covered by any general spurpose protest key terms list.

# Protests

- The **rioting** crowd broke windows and overturned cars.
- The union began its **strike** on Monday.
- Shah 's supporters also had gone on the **rampage** outside the principal 's office on tuesday .

- CPI(M) stages **protest** rally in Bhavnagar. the bhavnagar unit of communist party of india cpi(m) on friday staged a **demonstration** opposite the local post office here...

# CREATION OF PROTEST DATASETS - HISTORY

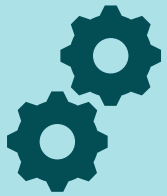
**Recent datasets:**

GDELT (Global Database of Events, Language and Tone), ICEWS (Integrated Crisis Early Warning System), NAVCO (Nonviolent and Violent Campaigns and outcomes), MMAD (Mass Mobilization in Autocracies Database),

**MANUAL**

- Pros: high quality, full-control
- Cons: Slow, non-replicable in case of updates to coding manual or detection of coding errors, constant effort requirement for expansion, limited scope, relatively hard to catch errors

**AUTOMATIC**



- Pros: fast, replicable, sizeable, "**transparent**"
- Cons: needs high quality corpus and/language resources, requires special attention to minority or new cases

# Reliability and Validity

## RELIABILITY

**Protest databases should have comparable results**

- Variable nature of contentious politics in different countries and time periods
- Using key terms to make variability more manageable but sacrifice recall performance.
- ML models trained using data from a country does not work on data from another country

## VALIDITY

**Results of event-coding projects should reflect unique real-world events**

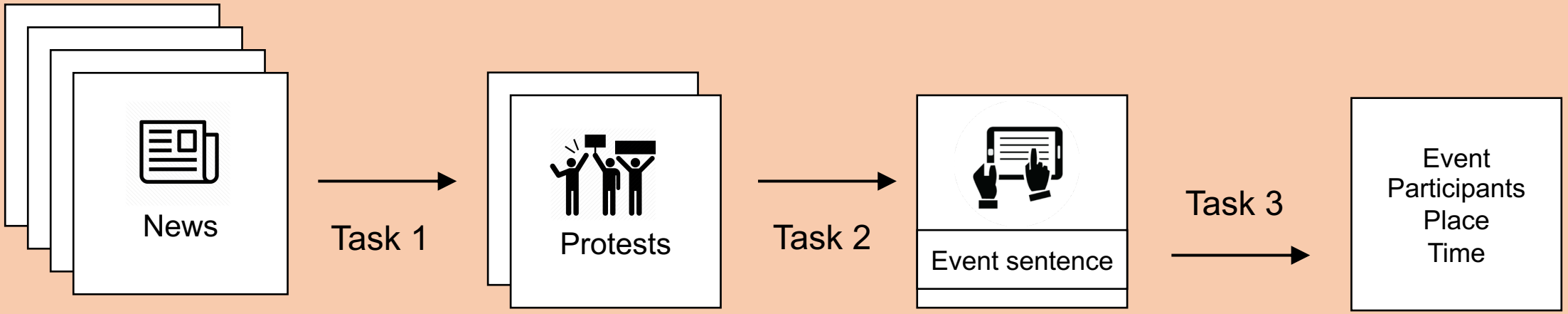
- Event extraction remains unresolved, low recall and precision (instead of real world events)
- Corpus flaws (no gold standard corpus)
- Global source dominated, instead of local

**Generalizability:** The principle which guides our overall methodology and task design is that both the protest event ontology and the automated tools must be so dynamic as to be able to accommodate source variability.

# CREATION OF A CROSS-CONTEXT GOLD STANDARD ANNOTATED CORPUS

- We use local news sources primarily (international only if there is censorship such as in China) from China, India, and South Africa
  - All are in English language for now (Portuguese and Spanish are being prepared)
- Random and active learning based sampling of news articles from multiple local sources (no key term selection)
- Annotation follows the pipeline structure: Document, Sentence, Token
  - The Krippendorf's alfa scores for these levels are above .75, .65, and between .34 and .60 respectively
- Double-annotation with agreement monitoring and correction steps
- All disagreements, %10 of the agreements, and ML-model errors were checked manually by the annotation supervisor

# CREATION OF PROTEST DATASET – AUTOMATED APPROACH TASK STRUCTURE



# Document level annotations

| | ES | INT | IEX | NIEX | PD | RCV1 | SCMP1 | SCMP2 | TH | ToI |
|---|---|---|---|---|---|---|---|---|---|---|
| Protest | 151 | 262 | 296 | 71 | 69 | 802 | 17 | 19 | 264 | 481 |
| Non-Protest | 149 | 738 | 265 | 630 | 732 | 367 | 985 | 483 | 782 | 1985 |
| Sampling | K | AL | AL | R | R | AL | R | R | AL | R&AL |

- K, R, and AL stand for key term based, random, and active learning respectively
- ES, INT, IEX, NIEX, PD, RCV1, SCMP, TH, and ToI refer to EventStatus corpus, Guardian, Indian Express, New Indian Express, People's Daily, South China Morning Post, The Hindu, and Times of India respectively.

# Sentence level annotations

| | INT2 | SCMP3 | NIEX2 |
|---|---|---|---|
| Protest | 1,658 | 511 | 1,299 |
| Non-protest | 9,045 | 2,847 | 7,083 |

- Sentence count per source is provided

# Token level annotations

| Tag name | Time | Trigger | Place | Facility | Participant | Organizer | Target |
|---|---|---|---|---|---|---|---|
| India | 822 | 1,378 | 645 | 392 | 2,283 | 1,260 | 1,453 |
| China | 144 | 142 | 82 | 52 | 272 | 88 | 109 |
| IAA | 60.07 | 50.02 | 41.82 | 39.10 | 39.50 | 47.44 | 34.38 |

- The count of the annotations per information type is provided.
- Inter-annotator agreement (IAA) is reported in terms of Krippendorf's alfa
- These annotations are checked and improved in terms of spotchecks and machine learning model error analysis.

# Cross-context performance

| | India | China | Int-China | South Africa |
|---|---|---|---|---|
| Document | .89 | .82 | .83 | .85 |
| Sentence | .85 | .79 | .83 | .85 |
| Token | .74 | .67 | N/A | N/A |

- We fine-tune the pretrained BERT-Base with the data from India. 512 tokens for document and token levels and 128 for sentence level.
- The F1 scores are reported on a held-out data from India, the whole data from China and South Africa.
- The cross-context setting causes the scores to drop on data from the other context.

# Detailed token level performance

| | Trigger | Time | Place | Facility | Participant | Organizer | Target |
|---|---|---|---|---|---|---|---|
| Precision | 0.756 | 0.663 | 0.724 | 0.436 | 0.649 | 0.568 | 0.497 |
| Recall | 0.691 | 0.704 | 0.646 | 0.436 | 0.564 | 0.619 | 0.485 |
| F1 | 0.722 | 0.683 | 0.683 | 0.436 | 0.604 | 0.593 | 0.491 |

- The performance of the BERT model is provided per information type.
- Flair NER model yielded significantly better results, which are .780, .697, and .652 for the place, participant, and organizer types.
- A BERT based event extraction model that is trained on ACE event extraction data yielded .543 F1 for trigger detection on the CONFLICT part of its test data. But this model yielded .479 F1 on our test data.

# Combination of document, sentence, and token models

| | Tok | Sent+Tok | Doc+Tok | Doc+Sent+Tok |
|---|---|---|---|---|
| Precision | .624 | .696 | .660 | **.701** |
| Recall | **.663** | .561 | .647 | .547 |
| F1 | .643 | .621 | **.653** | .614 |

- We analyzed predictions of all components on 100 positively and 100 negatively predicted documents by the document classifier.

# Corpus release

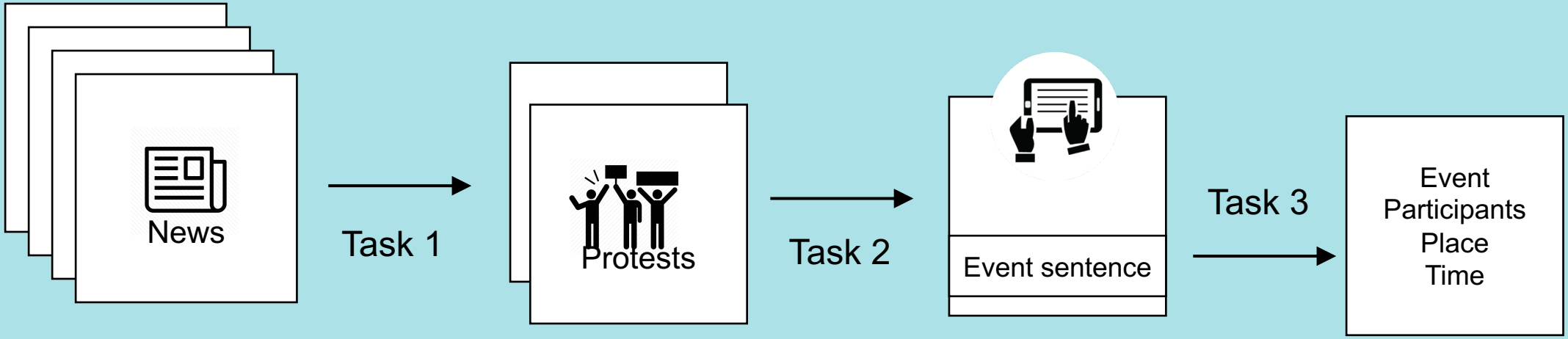
- The data is shared in a way that does not violate copyright of the news sources
  - Configuration of a Docker image that allow reproducing the datasets was shared with participants
  - And/Or Only the relevant part of a news article
- For more details contact us or check https://github.com/emerging-welfare/glocongold

# Next steps

- Tackling the following tasks
  - Event separation in multi-event news articles
  - Event, participant, and organizer semantic category identification
  - Cross-lingual generalization (between English, Spanish and Portuguese)
  - Creating a global database of protests
  - Comparing the quality of the international and local sources
- Seeking attention of computational linguistics and machine learning. Communities to collaborate on these tasks

# CLEF 2019 Lab ProtestNews

- Conference and Labs of the Evaluation Forum (CLEF) 2019 Lab ProtestNews: Extracting Protests from News
  - Train on data from India evaluate on data from China
  - Document classification, sentence extraction, and event extraction tasks




- **Website**: https://emw.ku.edu.tr/clef-protestnews-2019/

# LREC 2020 Workshop - Automated Extraction of Socio-political Events from News

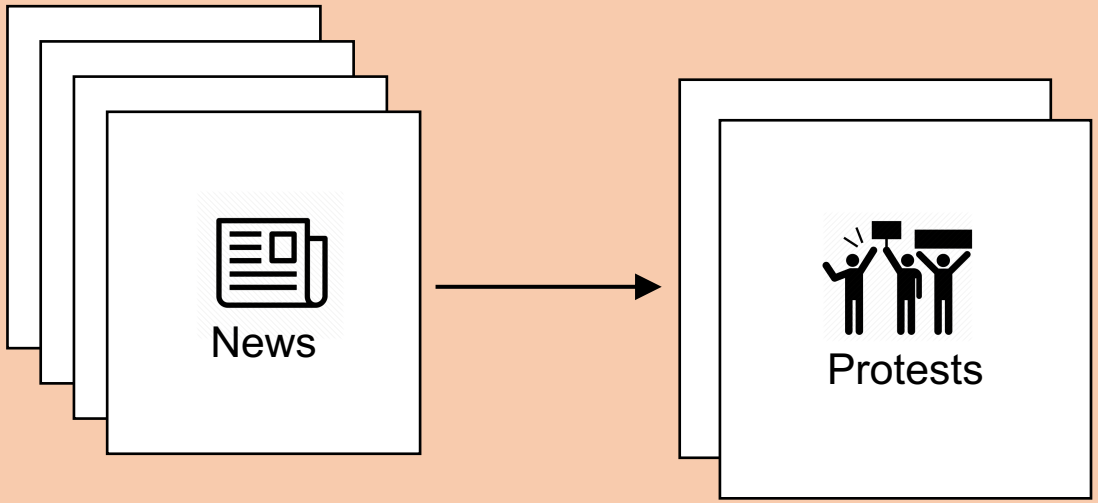
- Automated Extraction of Socio-political Events from News (AESPEN)
- **Shared task**: Event Sentence Coreference Identification (ESCI)
  - Which sentences are about the same event in a document?
    - Singh had recently blamed Advani for coming to Gujarat Chief Minister Narendra Modi ' s rescue and ensured that he was not sacked , in the wake of the riots .
    - On Kandahar plane hijack issue , Singh said Advani was not speaking the truth .
    - Elaborating on the three issues , Singhvi said , " The BJP gave sermons on Raj Dharma and turned a Nelson ' s eye to the communal carnage , which became a big blot on the fair name of the country .
- **Website**: https://emw.ku.edu.tr/aespen-2020/

Thanks for your attention!

Please contact ahurriyetoglu@ku.edu.tr for any comment, question, or remark?

Project home page: https://emw.ku.edu.tr/

Please consider joining our email list: automated-political-event-collection@googlegroups.com


News
Protests