Sustaining Modern Infrastructure For Political And Social Event Data

Patrick Brandt[1], Latifur Khan[1], Vito D'Orazio[2], Javier Osorio[3]

[1]The University of Texas at Dallas, [2]West Virginia University, [3]The University of Arizona

Big Picture

The main goal of this project is to integrate and expand our end-to-end cyberinfrastructure for robust creation, validation, access, and analysis of political event data. Event data in this context refers to a machine-coded description of an entity (e.g., a political actor) doing something to another entity, as extracted from news reports. We focus on political and social events about conflict and cooperation between governments, individuals, non-governmental organizations, rebel groups, and others. We rely on natural language processing tools to code event data by annotating the kinds of political events that are of interest to political scientists, international relations scholars, sociologists, and the national security community.

Our system scrapes contemporaneous news reports in English, Spanish, and other languages,and automatically encodes relevant political events for data analysts.

Data access and usability:

- Our data along with other open event data are available through our API and R interface.
- https://github.com/eventdata/UTDEventData
- https://eventdata.utdallas.edu/

Event data extraction:

- ConfliBERT: domain-specific pre-trained language model for conflict and political violence [1].
- Available at github.com/eventdata/ConfliBERT
- CoMe-KE: Transformers-based method for discovering political actors and extracting relevant role and location information from text sources [4].

Real world application:

- Created a new model to forecast political violence using event data and other input features [2].
- Entered the Violence Early Warning System competition to forecast subnational conflict events in Africa [3].

New challenges in event data research:

- New methods to use ConfliBERT to extract political events, and CoMe-KE for political entities.
- Expand ConfliBERT for multiple languages.
- Broader infrastructure of support for the research community to contribute to event data research.

This CSSI involves multiple institutions (University of Texas at Dallas, University of Arizona, West Virginia University) and disciplines (Computer Science, Political Science, Data Science). It has supported over 15 research assistants in Computer Science and Political Science graduate programs and has provided undergraduate research experience for students at the University of Texas at Dallas and the University of Arizona.

ConfliBERT: A Pretrained Language Model for Political Conflict and Violence[1]

Find our model at: github.com/eventdata/ConfliBERT

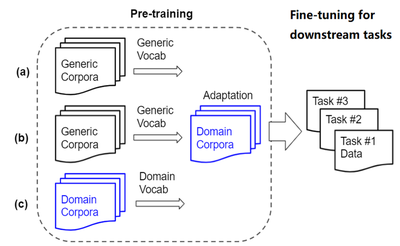
- Analyzing conflicts and violence with text methods are challenging due to the specialized language and lack of labeled data in that domain.
- Pre-trained transformers (e.g., BERT) using large-scale unlabeled text can significantly alleviate the bottleneck of annotation using transfer learning.
- Pre-trained, domain-specific transformers can out-perform generic language models.

ConfliBERT Methods

We compare the original BERT and two implementations of the domain-specific pretrained model, ConfliBERT:

(a) Continual training (Cont):
Continual training of original BERT with its original vocabulary on specific domains.

(b) Pretraining from scratch (SCR):
Pretraining BERT from scratch using new domain vocabulary on specific domains.



ConfliBERT Evaluation

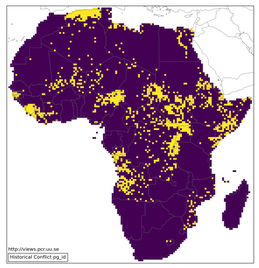
Dataset	Domain	Tasks	Train/dev/test	Metrics	BERT uncased	BERT cased	Confli.-Cont uncased	Confli.-Cont cased	Confli.-SCR uncased	Confli.-SCR cased
BBC	General	BC	1588/315/322	F1	97.24	96.38	97.9	96.95	98.08	**98.13**
20 News.	General	BC	9044/2270/7532	F1	80.30	79.58	80.4	80.51	**81.05**	80.37
Gun V.	Violence	BC	3387/423/423	macro F1	84.30	85.24	90.02	**90.27**	86.35	86.13
GLOCON	Protest	Sent BC	1549/193/193	macro F1	84.53	84.92	85.60	85.72	**86.57**	82.20
		Doc BC	782/130/130	macro F1	88.97	84.61	89.76	89.97	**91.13**	88.27
GTD	Terrorism	MCC	2825/471/471	macro F1	83.55	82.05	81.97	83.23	**83.82**	83.16
SATP	Terrorism	BC	5956/744/745	F1	87.78	87.10	87.51	87.49	88.12	**88.72**
		Rel MLC	1085/232/232	example F1	87.81	88.36	88.14	88.37	**89.08**	88.64
		All MLC	4794/1192/1489	example F1	63.36	63.32	64.14	63.72	64.47	**64.53**
Insight C.	Crime	MLC	1002/332/319	example F1	68.57	67.83	69.09	69.15	68.68	**69.47**
India P.	Violence	Sent MLC	14943/3172/3276	example F1	64.89	64.54	63.03	63.40	**67.27**	66.22
		Doc MLC	905/165/187	example F1	66.80	63.41	67.09	67.38	**69.97**	66.71
Event S.	Protest	TS MCC	1818/226/227	macro F1	70.65	67.15	73.32	**75.03**	72.55	70.94
		BC	4010/500/501	F1	91.72	90.67	**92.42**	91.85	92.10	92.40
CAMEO	Politics	PC MCC	1348/224/225	macro F1	86.44	85.85	**87.88**	86.12	87.64	87.83
		ST NER	1153/224/225	macro F1	72.29	72.25	74.00	**74.45**	74.35	72.87
MUC-4	Terrorism	NER	1300/200/200	macro F1	62.96	60.33	60.29	60.90	**63.97**	60.31
Re3d	Defence	NER	574/191/200	macro F1	63.44	62.46	64.40	66.20	**66.40**	64.23

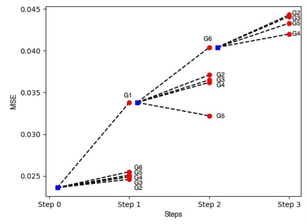
Table 4: The datasets, tasks and summary results of our evaluation.

Our models consistently report the best result (in bold), although performance varies (e.g., by case and task).

Conflict Forecasting with Event Data and Spatio-Temporal Graph Convolutional Networks[2,3]

- Leverages spatial and temporal dependencies and event data to forecast armed conflict in Africa.
- Inspired by the mechanism of message passing in graphs, graph convolutional networks (GCNs) are used to aggregate up to n-hop spatial neighbor nodes with each location in the data.
- Propose to use GCN along with LSTM to account for both spatial and temporal nature of the data.
- Used this method to forecast subnational armed conflict in Africa (Violence Early Warning System)
- We outperformed the competition baseline and showed the value of event data for forecasting.





CoMe-KE: A Transformers-Based Approach for Knowledge Extraction on Conflict and Mediation Domain[4]

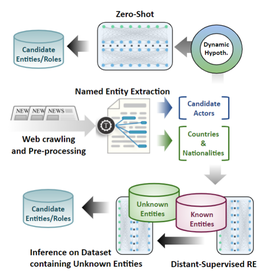
Problem Definition

- **Given:** A natural language corpus as input.
- **Desired:** Extract new ***political entity*** to extend the CAMEO repository.

Step 1: Political Entity Retrieval: Extract **political entity** and **locations**;

Step 2: Relation Extraction: Given (1), learn semantic relations for each pair.

We design and test two approaches for this step **(ZS and DS)**.



APIs and Usability

- API access to our event data and to other open event data resources.
  - eventdata.utdallas.edu/UTDEventData/Data/
- Access our API and query the data via the UTDEventdata R package.
  - https://github.com/KateHyoung/UTDEventData
- TwoRavens for Event Data user interface for event data download and data manipulations.
  - tworavens.com

Future Work

- Extending ConfliBERT to support multiple languages.
- Prompt-based zero-/few-shot learning approach for event coding.
- Expand on use of transformer models for end-to-end event data extraction .

References and Resources

1. Y. Hu, M. Hosseini, E. S. Parolin, J. Osorio, L. Khan, P. Brandt, V. D'Orazio, "ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence." Proceedings of The North American Chapter of ACL (NAACL) 2022.
2. Y. Li, B. Dong, L. Khan, P. Brandt, V. D'Orazio. "Data-Driven Time Series Forecasting for Social Studies Using Spatio-Temporal Graph Neural Networks." in Proceedings of ACM Conference on Information Technology for Social Good (GoodIT'21), pp. 61-66.
3. P. Brandt, V. D'Orazio, L. Khan, Y. Li, J. Osorio, M. Sianan, "Conflict Forecasting with Event Data and Spatio-Temporal Graph Convolutional Networks." *International Interactions*. 2022.
4. E. S. Parolin, Y. Hu, L. Khan, J. Osorio, P. Brandt, V. D'Orazio, "CoMe-KE: A New Transformers-Based Approach for Knowledge Extraction in Conflict and Mediation Domain," in Proceedings of IEEE International Conference on Big Data, 2021, pp. 1449-1459

Project website: eventdata.utdallas.edu
Project github: github.com/eventdata

Funding statement: This material is based upon work supported by the National Science Foundation under Grant Numbers DMS 1737978 and OAC 1931541. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.