# A Task Set Proposal for Automatic Protest Information Collection across Multiple Countries

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

Koc University, Rumelifener yolu, Sarıyer, İstanbul, Turkey
{ahurriyetoglu,eryoruk,dyuret,cyoltar,bgurel,fdurusan,omutlu}@ku.edu.tr

**Abstract.** We propose a coherent set of tasks for protest information collection in the context of generalizable natural language processing. The tasks are news article classification, event sentence detection, and event extraction. Having tools for collecting event information from data produced in multiple countries enables comparative sociology and politics studies. We have annotated news articles in English from a source and a target country in order to be able to measure the performance of the tools developed using data from one country on data from a different country. Our preliminary experiments have shown that the performance of the tools developed using English texts from India drops to a level that are not usable when they are applied on English texts from China. We think our setting addresses the challenge of building generalizable NLP tools that perform well independent of the source of the text and will accelerate progress in line of developing generalizable NLP systems.

**Keywords:** natural language processing · information retrieval · machine learning · text classification · information extraction · event extraction · domain adaptation · transfer learning · computational social science · contentious politics · protest information

## 1 Introduction

Comparative social studies on social protest requires collecting protest event data from multiple countries. The utility of these collections increases with the number of countries covered, the length of the time span and the weight of the information gathered from local sources. The performance of natural language processing (NLP) tools, those of text classification and information extraction in our setting, has not been satisfactory against the requirements of longer time coverage and working on data from multiple countries [16, 8]. In this study, we introduce a set of tasks, supported with the relevant data, for facilitating the creation of protest event databases that are better equipped to handle variations in country settings through both space and time. The setting we propose facilitates testing and improving state-of-the-art methods for text classification and information extraction on English news article texts from India and China. The direction of our work is towards developing generalizable information systems that perform comparatively well on texts from multiple countries.

The need for collecting protest or conflict data has been satisfied by utilizing manual [4, 18], semi-automatic [9], and automatic [9, 2, 10, 11, 13] methods -each of which presents a different set of challenges that limit the utility of that method. The methods that rely on manual and semi-automated coding, though reliable, require a tremendous amount of effort to replicate on new data as they depend intensely on high quality human effort. On the other hand, text classification and information extraction systems that rely on automated methods yield less reliable results as they tend to perform poorly on texts different from the ones they were developed and validated on [6, 12]. The huge amount of news articles that are required to be analyzed and the constant need of repeating the same analyses on new data force us to push limits of automated protest information collection yet again. Furthermore, addressing and remedying performance issues when faced with difficulties presented by variations across datasets requires the tools to be as generalizable as possible.

Much of the difficulty presented by automated methods of data collection on contentious politics events[1] stems from the fact that contentious politics take slightly different forms in different countries and time periods in line with spatial and temporal variation of sociopolitical phenomena. The automated tools run the risk of being biased towards the country and/or time period of the cases that they are trained upon and the need to adapt them to different cases leads developers to either redesign tools from scratch for each individual case or take certain shortcuts which somehow makes variety more manageable. A common such recourse which imposes a level of uniformity to data universe is key term based filtering -a method which relies on an a priori set of keywords related to protest events to filter irrelevant cases out of the training dataset. It is our conviction that this method is arbitrary and possibly cripples the reliability of data collection from the outset by leaving out potentially relevant protest events. Moreover, there is no inbuilt way to determine if or to what extent such unwanted exclusion occurs as the filtering is external to the training-evaluation cycle.

Rather than developing case specific classifiers for every single country or limiting the raw data via key term filters, we strive to develop generalizable information systems that perform comparatively well on multiple country settings and can be applied to any set of random selection of news articles. In order to accommodate the geographical and historical variability of sociopolitical contexts, the chief aspect of our task design takes the tools that are developed on the basis of the data from a certain country and evaluates them on data from a different country. Thus, the evaluation feedback forms a novel basis on which the tools are further developed to accommodate even more variation in the future. This rolling training-evaluation cycles is expected to create a virtuous circle of feedback loop which will be more generally applicable with every new country case that is introduced.

[1] The term used when referring to these events in collective is "repertoires of contention" [7, 14]. We will use "protest events" from here on for the sake of brevity simplicity.

This paper describes how we will realize the proposed setting within the lab ProtestNews in the 2019 edition of the Conference and Labs of the Evaluation Forum (CLEF).[2,3,4] We introduce the methodology we apply to create the corpus, and the task set we propose, in 2 and 3 respectively. We report our preliminary results in Section 4 and conclude our report by pointing to future directions of our work in Section 5.

## 2 Data

We collect online English news articles from a source and a target country, India and China respectively.[5] We first download the freely accessible part of an online news archive and create a random sample of these articles from each source in order to have a representative sample for labelling and annotation for each task.

We apply the same labelling and annotation manuals on data collected from different countries. This approach enables obtaining comparable measures of automatic system performance. Our data preparation process applies state-of-the-art annotation methodology in terms of being based on an annotation manual, sampling the news articles from various sources and periods, and continuously monitoring the annotations to achieve a high inter-annotator agreement.

Annotators that are master students or PhD candidates in social or political sciences work in pairs. In each pair, both annotators annotate the same document, sentence, or token depending on the task.[6] The annotation start by labelling articles in a sample of news articles as containing a protest or not. Sentences of these positively labelled documents are then labelled as containing protest information or not. These sentences should contain either an event trigger or a reference to an event trigger in order to be labelled as positive. Finally, the protest-related sentences are annotated at token level for the information they denote.[7] The supervisor, who is a social scientist and responsible of maintaining the annotation manuals as well, resolves the disagreements between the annotators.

We analyze the annotator agreements as well. To prevent cases where annotators may agree on wrong labelling, we applied the following means of improving the corpus. First, we regularly apply a spot check in which the expert double checks a small sample of labels and annotations the annotators agree on the attached label. Second, any erroneous annotation in the positive cases may be

[2] http://www.clef-initiative.eu, accessed January 19, 2019.

[3] http://clef2019.clef-initiative.eu, accessed January 19, 2019.

[4] https://emw.ku.edu.tr/clef-protestnews-2019, accessed January 19, 2019.

[5] Using available corpora that are already being allowed to be distributed freely is not an option for our setting due to the requirement of having a representative sample from the source and target countries. Also, the dataset should contain data created in more than one country in order to be useful in our setting.

[6] The overlap ratio is 100%.

[7] We mainly annotate the event trigger, place, time, participant, organizer, and target of the protest.

captured in the following step where the annotators do a more detailed annotation for the following task. Third, we semi-automatically check the documents labelled as non-protest, by training a classification model on 80% of the all labelled and adjudicated documents or sentences and testing on the remaining 20%. The cases that are predicted as protest by the classifier but labelled as non-protest by the annotators are double checked manually to verify they are indeed non-protest. Finally, in order to eliminate risk of wrong labelling due to lack of knowledge about a country, a domain expert instructs the annotators before they start to do annotation.

We distribute the data in a way that does not violate copyright of the news sources. This involves only sharing information that is needed to reproduce the corpus from the source in cases it is not allowed to distribute the news articles.

## 3 Tasks

We designed the tasks as depicted in Figure 1. The analysis should start by predicting whether a random news article mentions a protest. Then, the sentence(s) that contain protest information should be identified. Finally, protest information such as participants, place, and time should be detected in the protest related sentences. This order of tasks provides a controlled setting that enables error analysis and optimization possibility during annotation and tool development efforts.

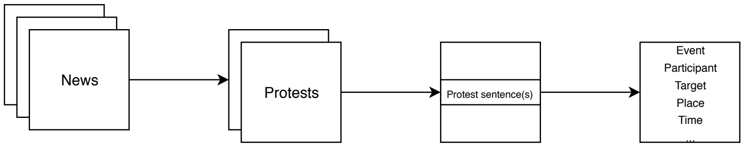

**Fig. 1.** The lab consists of a) *Task 1*: News article classification as protest vs. non-protest, b) *Task 2*: Protest sentence detection, and c) *Task 3*: Event extraction. Tasks 2 and 3 will be based on news articles labeled for task 1. Participants can choose to participate in one or more of these tasks independent of each other.

The set of tools that will tackle these tasks should be implemented and validated for data originated from a country and tested on data collected from a different country, which are India and China. There will be two level of evaluation, which we refer as Test 1 and Test 2, on data that is not accessible to the lab participants. The first level, which is Test 1, is on test data from the country used for training and developing the methods. The second evaluation, which is Test 2, will be on data from the target country. The primary score for ranking the submissions will be the one on the target country.

We use macro averaged F1 for evaluating the Task 1 and Task 2. The event extraction task, which is Task 3, will be evaluated on F1 score that will be based on the ratio of the match between the prediction and the annotations in the test sets.

Although the annotation effort is continuing to increase amount of news articles for each task, we would like to report the recent approximate number of news articles we have labelled and annotated for each task in Table 1. The training and development columns illustrates the number of documents that will be accessible to the lab participants. These documents are from the source country. The document count for Test 1 and Test 2 columns are from the source and the target countries respectively.

**Table 1.** Number of annotated news articles for each task

| | Training | Development | Test 1 | Test 2 |
|---|---|---|---|---|
| Task 1 | 8,000 | 1,000 | 1,000 | 4,000 |
| Task 2 | 600 | 100 | 100 | 200 |
| Task 3 | 300 | 50 | 50 | 20 |

## 4 Preliminary Results

We performed various analyses and experiments on the corpus we created in order to further shed light on characteristics of the dataset and the tasks we propose. First, we filtered our corpus with the key terms that were used by Wang et al. (2016) [15], Lorenzini et al. (2016) [10], and Weidman and Rød (2019) [17]. The Table 2 shows the protest coverage of these key terms in our corpus. The low recall demonstrates the difference between the coverage of a random sample and a key term filtered sample.[8] We assume that our random sampling method ensures complete recall.

**Table 2.** Coverage of the key terms used by recent studies in our corpus

| | Precision | Recall | F1-score |
|---|---|---|---|
| Wang et al. (2016) | .57 | .75 | .65 |
| Lorenzini et al. (2016) | .42 | .88 | .57 |
| Weidman and Rød (2019) | .60 | .58 | .59 |

We have performed automatic classification experiments by training binary machine learning models for task 1 and task 2. For task 1, a support vector machine (SVM) and a deep neural network (DNN) classifiers were trained using the training data by being optimized on the development data. The SVM model

[8] The difference between our and these projects' annotation manuals potentially affects the precision and recall as well.

has yielded .85 and .25 F1 score on Test 1 and Test 2 respectively. The pretrained BERT model's [5] performance is .90 and .64 in the same setting. For task 2, three binary sentence classifiers, which are random forest, decision tree, and SVM, were created using the training and development data. The F1 scores of these classifiers are .47, .52, and .56 on Test 1 data. Finally, our experiments for task 3 yielded around .30 lower F1 score than 't is reported in publications of these tools on test 1 data [1, 3].[9]

## 5 Conclusion and Future Work

Comparative social science studies deploy concepts, and work on variables that must be applicable across multiple different countries and time periods. As the particular cultural, political and linguistic characteristics of each different geographical and historical context reflect on the news articles, the NLP tools that are utilized to construct news databases used by these studies must have generalized applicability. The preliminary analysis and tool performance results show that the difference in news content and performance differences on data from different countries are significant, which presents a challenge for the text processing systems aiming at such generalizability. The task design we propose in this paper is expected to fulfill such requirements, and will certainly be enriched and moved closer to perfection through contributions in this shared task.

As to the future development path of our line of research, we envision the following improvements to the dataset in line of our broader goal of developing tools for creating a high-quality global protest database with general applicability: (i) the corpus should be extended with English data from additional countries; (ii) data in languages other than English should be included; (iii) instead of labeling only as protest or non-protest, categorization of protest events into types such as demonstration, industrial action, group clash, and armed militancy should be integrated into the task set; and (iv) the problem of distinguishing expressions of events that have not taken place, such as threats and plans of protest events, from events that have taken place must be addressed. Tasks which label planned/threatened events separately from events and non-events promises to tackle this challenge.

## Acknowledgments

This work is funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare.[10] We are grateful to our steering committee members for the CLEF 2019 lab Sophia Ananiadou, Antal van den Bosch, Kemal Oflazer, Arzucan Özgür, Aline Villavicencio, and Hristo Tanev. Finally, we thank to Theresa Gessler and Peter Makarov for their contribution in organizing the CLEF lab by reviewing the annotation manuals and sharing their work with us respectively.

[9] https://github.com/emerging-welfare/ie-tools-test-on-India-b1, accessed January 19

[10] https://emw.ku.edu.tr, accessed January 19

## References

1. Akdemir, A., Hürriyetoğlu, A., Yörük, E., Gürel, B., Yoltar, c., Yüret, D.: Towards Generalizable Place Name Recognition Systems: Analysis and Enhancement of NER Systems on English News from India. In: Proceedings of the 12th Workshop on Geographic Information Retrieval. pp. 8:1–8:10. GIR'18, ACM, New York, NY, USA (2018). https://doi.org/10.1145/3281354.3281363, http://doi.acm.org/10.1145/3281354.3281363
2. Boschee, E., Natarajan, P., Weischedel, R.: Automatic Extraction of Events from Open Source Text for Predictive Forecasting. In: Subrahmanian, V. (ed.) Handbook of Computational Approaches to Counterterrorism, pp. 51–67. Springer New York, New York, NY (2013). https://doi.org/10.1007/978-1-4614-5311-6_3, https://doi.org/10.1007/978-1-4614-5311-6_3
3. Büyüköz, B., Hürriyetoğlu, A., Yörük, E., Yüret, D.: Examining Existing Information Extraction Tools on Manually-Annotated Protest Events in Indian News. In: Proceedings of Computational Linguistics in Netherlands (CLIN). CLIN29 (2019)
4. Chenoweth, E., Lewis, O.A.: Unpacking nonviolent campaigns: Introducing the NAVCO 2.0 dataset. Journal of Peace Research **50**(3), 415–423 (2013). https://doi.org/10.1177/0022343312471551, https://doi.org/10.1177/0022343312471551
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Ettinger, A., Rao, S., Daumé III, H., Bender, E.M.: Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task. In: Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems. pp. 1–10. Association for Computational Linguistics (2017), http://aclweb.org/anthology/W17-5401
7. Giugni, M.G.: Was It Worth the Effort? The Outcomes and Consequences of Social Movements. Annual Review of Sociology **24**, 371–393 (1998), http://www.jstor.org/stable/223486
8. Hammond, J., Weidmann, N.B.: Using machine-coded event data for the micro-level study of political violence. Research & Politics **1**(2), 2053168014539924 (2014). https://doi.org/10.1177/2053168014539924, https://doi.org/10.1177/2053168014539924
9. Leetaru, K., Schrodt, P.A.: Gdelt: Global data on events, location, and tone, 1979–2012. In: ISA annual convention. vol. 2, pp. 1–49. Citeseer (2013)
10. Lorenzini, J., Makarov, P., Kriesi, H., Wueest, B.: Towards a Dataset of Automatically Coded Protest Events from English-language Newswire Documents. In: Paper presented at the Amsterdam Text Analysis Conference (2016)
11. Schrodt, P.A., Beieler, J., Idris, M.: Three'sa charm?: Open event data coding with el: Diablo, Petrarch, and the open event data alliance. In: ISA Annual Convention (2014)
12. Soboroff, I., Ferro, N., Fuhr, N.: Report on GLARE 2018: 1st Workshop on Generalization in Information Retrieval: Can We Predict Performance in New Domains? SIGIR Forum **52**(2), 132–137 (2018), http://sigir.org/wp-content/uploads/2019/01/p132.pdf
13. Sönmez, Ç., Özgür, A., Yörük, E.: Towards building a political protest database to explain changes in the welfare state. In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and

Humanities. pp. 106–110. Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/W16-2113, http://www.aclweb.org/anthology/W16-2113

14. Tarrow, S.: Power in Movement: Social Movements, Collective Action and Politics. Cambridge Studies in Comparative Politics, Cambridge University Press (1994), https://books.google.com.tr/books?id=hN5nQgAACAAJ
15. Wang, W.: Event Detection and Extraction from News Articles. Ph.D. thesis, Virginia Tech (2018)
16. Wang, W., Kennedy, R., Lazer, D., Ramakrishnan, N.: Growing pains for global monitoring of societal events. Science **353**(6307), 1502–1503 (2016). https://doi.org/10.1126/science.aaf6758, http://science.sciencemag.org/content/353/6307/1502
17. Weidmann, N.B., Rød, E.G.: The Internet and Political Protest in Autocracies, chap. Coding Protest Events in Autocracies. Oxford University Press, Oxford (2019)
18. Yoruk, E.: The politics of the Turkish welfare system transformation in the neoliberal era: Welfare as mobilization and containment. The Johns Hopkins University (2012)