Overview of CLEF 2019 Lab ProtestNews: Extracting Protests from News in a Cross-context Setting

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

Koc University, Istanbul 34450, Turkey
{ahurriyetoglu,eryoruk,dyuret,cyoltar,bgurel,fdurusan,omutlu,
aakdemir}@ku.edu.tr
http://www.ku.edu.tr

Abstract. We present an overview of the CLEF-2019 Lab ProtestNews on Extracting Protests from News in the context of generalizable natural language processing. The lab consists of document, sentence, and token level information classification and extraction tasks that were referred as task 1, task 2, and task 3 respectively in the scope of this lab. The tasks required the participants to identify protest relevant information from English local news at one or more aforementioned levels in a cross-context setting, which is cross-country in the scope of this lab. The training and development data were collected from India and test data was collected from India and China. The lab attracted 58 teams to participate in the lab. 12 and 9 of these teams submitted results and working notes respectively. We have observed neural networks yield the best results and the performance drops significantly for majority of the submissions in the cross-country setting, which is China.

Keywords: natural language processing · information retrieval · machine learning · text classification · information extraction · event extraction · computational social science · generalizability

1 Introduction

We describe a realization of our task set proposal [4] in the scope of CLEF-2019 Lab ProtestNews.[1,2] The task set aims at facilitating development of generalizable natural language processing (NLP) tools that are robust in a cross-context setting, which is cross-country in this lab. Since the performance of NLP tools

---

[1] http://clef2019.clef-initiative.eu/

[2] https://emw.ku.edu.tr/clef-protestnews-2019/

significantly drop in a context different from the one they are created and validated [1, 2, 6], measuring and improving state-of-the-art NLP tool development methodology is the primary aim of our efforts.

Comparative social and political science studies facilitate protest information to analyze cross-country similarities, differences, and effect of these actions. Therefore our lab focuses on classifying and extracting protest event information in English local news articles from India and China. We believe our efforts will contribute to enhance the methodologies applied to collect data for these studies. This need was motivated based on the recent results that shows NLP tools, those of text classification and information extraction, have not been satisfactory against the requirements of longer time coverage and working on data from multiple countries [7, 3].

This first iteration of our lab attracted 58 teams from all around the world. 12 of these teams submitted their results to one or more tasks on the CodaLab page of the lab.[3] 9 teams described their approach in terms of a working note.

We introduce the task set we tackle, the corpus we have been creating, and the evaluation methodology in Sections 2, 3, and 4 respectively. We report the results in Section 5 and conclude our report in Section 6.

2 Task Set

The lab consists of the tasks document classification, event sentence detection and event extraction, which are referred as task 1, task 2, and task 3 respectively, as demonstrated in Figure 1. The document classification task, which is task 1, requires predicting whether a news article report at least one protest event that has happened or is happening. It is a binary classification task that require to predict whether a news article label should be 1 (positive) or 0 (negative). The sentences that contain any event trigger should be identified in task 2, which is event sentence detection task. Sentence labels are 0 and 1 as well. This task could be handled either as classification or extraction task as we provide order of the sentences in their respective articles. Finally, the event triggers and event information, which are place, facility, time, organizer, participant, and target, should be extracted in task 3. This order of tasks provides a controlled setting that enables error analysis and optimization possibility during annotation and tool development efforts. Moreover, this design enable analyzing steps of the analysis that contributes to explainability of the automated tool predictions.

3 Data

We provide the number of instances for each task in Table 1 in terms of training, development, test 1, and test 2 data. The training and development data was collected from online local English news from India. Test 1 and test 2 data refer to data from India and China respectively.
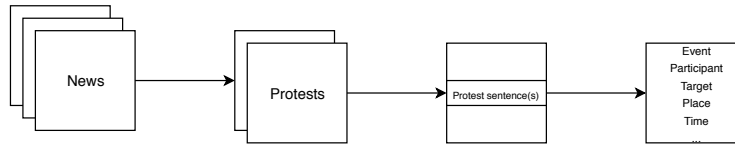
---

[3] https://competitions.codalab.org/competitions/22349

Fig. 1. The lab consists of a) *Task 1*: Document classification, b) *Task 2*: Event sentence detection, and c) *Task 3*: Event extraction.

Table 1. Number of instances for each task

	Training	Development	Test 1 (India)	Test 2 (China)
Task 1	3,429	456	686	1,800
Task 2	5,884	662	1,106	1,234
Task 3	250	36	80	39

A sample from task 1 contains the news article's text, its URL and its label that is assigned by the annotation team. For task 2: sentences, their labels, their order in the article's text they belong, and the URL of their article are available in samples. The release format of the data for task 2 enable participants to treat this task either as classification of individual sentences or extracting event relevant sentences from a document.

The data for task 3 consists of snippets that contain one or more event trigger that refer to the same event. Multiple sentences may occur in a snippet in case these sentences refer to the same event.[4] The tokens in these snippets are annotated using IOB, inside, outside, beginning, scheme. The examples of data is provided in Figure 2.

There is not any overlap of news articles across tasks. This separation was required in order to avoid any misuse of data from one task to infer the labels for another task without any effort.

3.1 Distribution

We distributed the data set 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 for task 1 and task 2 and only relevant snippets for task 3. We released a Docker image that contains the toolbox[5] required to reproduce the news articles on the computer of a participant. The toolbox generates a log of the process that reproduce the data set and we have requested these log files from the participants. The toolbox is a pipeline that scrapes HTMLs, converts HTMLs to text and finally performs specific filling operation for each of task 1 and task 2. To the best of our knowledge, the toolbox succeeded in enabling participants create the data set on their computers. Only one participant from

---

[4] Snippets we share contain information about only a single event.
[5] https://github.com/emerging-welfare/ProtestNews-2019

```
{"text": "... Police suspect that the panchayat members, including the Salwa
Judum leader, were abducted and killed by Maoist rebels, who had left the
bodies near the village. Meanwhile, security forces and Naxalites had an
encounter near village Belgaon. ...", "label":1}
```

A sample from data set for task 1.

```
{"sentence": "Police suspect that the panchayat members, including the
Salwa Judum leader, were abducted and killed by Maoist rebels, who had
left the bodies near the village.", "sentence_number":3, "label": 1},
{"sentence": "Meanwhile, security forces and Naxalites had an encounter
near village Belgaon.", "sentence_number":4, "label": 1}
```

Corresponding sentence samples for the article above.

```
            ⋮
including   O
the         O
Salwa       B-target
Judum       I-target
leader      I-target
,           O
were        O
abducted    B-trigger
and         I-trigger
killed      I-trigger
by          O
Maoist      B-participant
rebels      I-participant
            ⋮
```

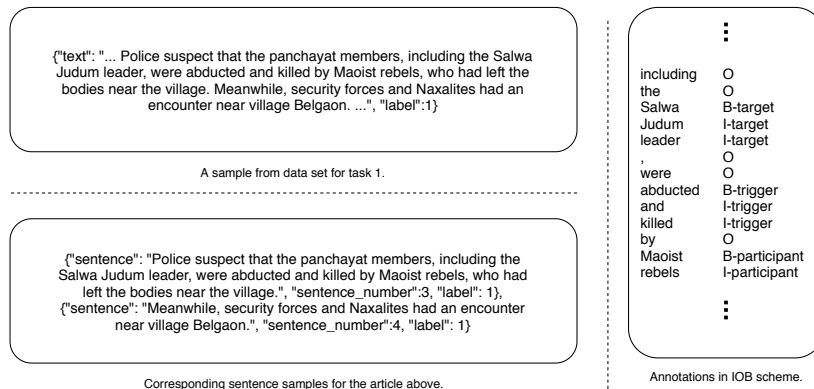Annotations in IOB scheme.

Fig. 2. Data samples for task 1, task 2, and task 3

Iran was not able to download the news articles due to restrictions to access online content that are specific to his geolocation.

4 Evaluation Setting

We use macro averaged F1 due to class imbalance present in our data for evaluating the task 1 and task 2. The event extraction task, which is task 3, was evaluated on the average F1 score of all information types that was based on the ratio of the full match between the prediction and the annotations in the test sets, using a python implementation[6] of CoNLL 2003 shared task [5] evaluation script.

We performed two levels of evaluation that were on data from the source country (Test 1) and from target country (Test 2). The participants were informed only about labels of the training and development data from the source country. They did not see labels of any test set. The number of allowed submissions and the period the participants can submit their predictions for Test 1 and Test 2 was determined in a way that restrict the possibility of over-fitting on test data. We limited the number of submission and the submission period in order to make sure the participants do not overfit to the test data based.

We applied three cycles of evaluation. Participants could submit unlimited number of results without being able to see their score. Their last submitted results' scores were announced at the end of each evaluation cycle. First and second cycles aimed at providing feedback to the participants. The third and final cycle was the deadline for submitting results.

Finally, we provided a baseline submission for task 1 and task 2 in order to guide the participants. This baseline was based on predictions of the the best scoring machine learning model among Support Vector Machines, Naive

---

[6] https://github.com/sighsmile/conlleval

Bayes, Rocchio classifier, Ridge Classifier, Perceptron, Passive Aggressive Classifier, Random Forest, K Nearest Neighbors, and Elastic Net on development set. The best scoring model was a linear support vector machines classifier that was trained using stochastic gradient descent.

5   Results

We facilitated CodaLab platform for managing the submissions and maintain a leaderboard.[7] The leaderboard for task 1 and task 2 is presented in Table 2. The column names have the following format Test <task number>-<test number>, e.g. Test 1-1 stands for the Test 1 of task 1. The results for ProtestLab_Baseline is the aforementioned baseline that was submitted by us.

Table 2. Results that are ranked based on average F1 scores of Test 1 and Test 2 for task 1 and task 2

	Test 1-1	Test 1-2	Test 2-1	Test 2-2	Avg
ASafaya	.81	.63	**.70**	.60	.69
PrettyCrocodile	.79	.60	.65	**.64**	.67
LevelUp Research	**.83**	**.65**	.66	.45	.65
Provos_RUG	.80	.59	.63	.55	.64
GS	.78	.56	.64	.58	.64
Be-LISI	.76	.50	.58	.30	.54
ProtestLab_Baseline	.83	.49	.58	.20	.52
CIC-NLP	.59	.50	.52	.34	.49
SSNCSE1	.38	.15	.56	.35	.36
iAmirSoltani	.69	.36	-	-	.26
Sayeed Salam	.55	.28	-	-	.20
SEDA lab	.58	-	.15	.02	.19

The summary of the approaches and results of each team that participated in task 1 and or task 2 are provided below.[8]

ASafaya (Sakarya University) submitted the best results for task 2 and for average of task 1 and task 2 using Bidirectional Gated Recurrent Unit (GRU) based model. Although this model perform the best on average, the performance of this model drops across the context significantly.

PrettyCrocodile (National Research University HSE) submitted the second best average results that were predicted using Embeddings from Language Models (ELMo). The performance of the model is comparable in the cross-context setting for task 2.

---

[7] https://competitions.codalab.org/competitions/22349#results

[8] We have not received details of the submissions from CIC-NLP, iAmirSoltani, and Sayeed Salam. The details of other approaches can be found in the respective working notes that were published in proceedings of CLEF 2019 Lab ProtestNews.

LevelUp Research (University of North Carolina at Charlotte) has applied multi-task learning based on LSTM units using word embeddings from a pre-trained FastText model. This method yielded the best results for task 1.

Provos_RUG (University of Groningen) has implemented a feature based stacked ensemble model based on FastText embeddings and a set of different basic Logistic Regression classifiers that enabled their predictions to rank fourth among the participating teams.

GS (University of Bremen) has stacked the word embeddings such as GloVe and FastText together with the contextualized embeddings generated from Flair language models (LM). This approach was ranked fourth in general and third for task 2.

Be-LISI (Universit de Carthage) combined the logistic regression with linguistic processing and expansion of the text with related terms using word embedding similarity. This approach marked a significant difference in terms of overall performance, which is the drop from .64 to .54, in comparison to higher ranked submissions.

SSNCSE1 (Sri Sivasubramaniya College of Engineering) reported results of their bi-directional LSTM that applies Bahdanau, Normed-Bahdanau, Luong, and Scaled-Luong attentions. The submission that uses Bahdanau attention yielded the results reported in Table 2.

SEDA lab (University of Exeter) applied support vector machines and XGBoost classifiers that are combined with various word embedding approaches. Results of this submission showed promising performance in terms of precision on both document and sentence classification tasks.

We analyze task 3 results separate from task 1 and task 2 as it differs from them. The F1 scores for task 3 are presented in Table 2.

Table 3. Results that are ranked based on average score of Test 1 and Test 2 for Task 3

	Test 3-1	Test 3-2	Avg
GS	.604	.532	.568
DeepNEAT	.601	.513	.557
Provos_RUG	.600	.456	.528
PrettyCrocodile	.524	.425	.474
LevelUp Research	.516	.394	.455

GS (University of Bremen) submitted the best results for task 3 using a BiLSTM-CRF model incorporating pooled contextualized flair embeddings, and their model was the best in generalizing.

DeepNEAT (FloodTags & Radboud University) compares the submitted ELMO+BiLSTM model to a traditional CRF and shows that the former is better and more generalizable.

Provos_RUG (University of Groningen) divides the task 3 into two as event trigger detection task and event argument detection task using BiLSTM-CRF model with word embeddings, POS embeddings, and character-level embeddings for both subtasks. He further extends the features for latter subtask with learned embeddings for dependency relations and event triggers.

PrettyCrocodile (National Research University HSE) makes use of ELMO embeddings with different architectures, achieving her best score for task 3 using a BiLSTM.

LevelUp Research (University of North Carolina at Charlotte) implemented a multi-task neural model that require a time-ordered sequence of word vectors representative of a document or sentence. The LSTM layer has been replaced by a layer of bidirectional gated recurrent units (GRU).

6   Conclusion

The results show how text classification and information extraction tool performances drops between two contexts. The scores on data from the target country are significantly lower than on data from the source country. Only the PrettyCrocodile team performed comparatively well across contexts for task 2. Although it is not the best scoring system for neither task 1 nor task 2, Pretty-Crocodile team's approach show some promise toward tackling the generalizability of NLP tools.

The generalization of automated tools is an issue that has recently attracted much attention.[9] However, as we have determined in our lab, generalizability is still a challenge for state-of-the-art methodology. Consequently, we will continue our efforts by repeating this practice and extending the data and will be adding data from new countries and languages to our setting. The next iteration will run in the scope of the Workshop on Challenges and Opportunities in Automated Coding of COntentious Political Events (Cope 2019) at European Symposium Series on Societal Challenges in Computational Social Science (Euro CSS 2019).[10,11,12]

Acknowledgments

---

[9] https://sites.google.com/view/icml2019-generalization/cfp

[10] https://competitions.codalab.org/competitions/22842

[11] https://emw.ku.edu.tr/?event=challenges-and-opportunities-in-automated-coding-of-contentious-political-events&event_date=2019-09-02

[12] http://symposium.computationalsocialscience.eu/2019/

for their contribution in organizing the CLEF lab by reviewing the annotation manuals and sharing their work with us respectively.

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. 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
3. 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
4. Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., Mutlu, O.: A task set proposal for automatic protest information collection across multiple countries. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 316–323. Springer International Publishing, Cham (2019)
5. Sang, E.F., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050 (2003)
6. 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
7. 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