# CLEF ProtestNews Lab 2019: Contextualized Word Embeddings for Event Sentence Detection and Event Extraction

Gabriella Skitalinskaya, Jonas Klaff, and Maximilian Spliethöver

University of Bremen, 28359 Bremen, Germany
{gabski,joklaff,mspl}@uni-bremen.de

**Abstract.** In this work we describe our results achieved in the ProtestNews Lab at CLEF 2019. To tackle the problems of event sentence detection and event extraction we decided to use contextualized string embeddings. The models were trained on a data corpus collected from Indian news sources, but evaluated on data obtained from news sources from other countries as well, such as China. Our models have obtained competitive results and have scored 3rd in the event sentence detection task and 1st in the event extraction task based on average F1-scores for different test datasets.

**Keywords:** Contextualized String Embeddings · Classification · Named Entity Recognition.



## 1 Introduction

Automated protest news mining can play a great role in analyzing and understanding protests and their media coverage, especially on a global scale. Such research may be able to support different research domains by capturing the protest's evolution over time and identifying the origins of riots and social movements. Additionally, by analyzing news sources from a wide range of countries we can get a better understanding of the worldwide media coverage of protest events.

The CLEF-2019 ProtestNews! Lab [19] tries to tackle this problem and has introduced three shared tasks aimed at identifying and extracting event information from news articles across multiple countries. The aim of the shared tasks is the development of a generalizeable text classification and information extraction tool that could be applied to datasets from different countries without additional training.

The first task can be described as a binary classification task aimed at discriminating between news articles related to protest events and any other news

articles. In the second task, the tool should be able to determine whether a sentence is an event sentence, i.e. contains an event trigger or a mention of it. Finally, the third task is a named entity recognition (NER) task focused on extracting various types of information from a given event sentence such as location, time and participants of an event. In this paper we will only cover the second and third tasks.

For every task a set of news articles from one country (India) was provided with a predefined training and development split. The resulting models were then evaluated on news articles from the same country as in the training set and on an additional set containing data from another country(China).

The rest of this paper is organized as follows. We discuss relevant literature in Section 2. Section 3 gives details on the training dataset and the description of the proposed approaches. Section 4 provides experimental evaluation, and important insights gained during our work. We conclude in Section 5, outlining our contributions and directions for future research.

## 2 Related work

Various Natural Language Processing (NLP) techniques have been utilized in research to automatically analyze and extract information about events from free-form texts focusing on different types of events and pursuing different goals. We would like to give a few examples of typical approaches used to address the problems of our interest. The authors of [7] use a Conditional Random Fields (CRF) model to evaluate social media posts of a big event in order to gather information about smaller sub-events, that are collocated to the more popular one. The authors do not focus on political events, but try to find a more generalizable approach, applicable to multiple types of events.

In contrast to that, [8] focuses on the analysis of activism. Their main approach is to extract event information from natural language text and visualize it afterwards. Instead of social media posts, as in [7], [8] utilize news articles from media outlets.

In [8] the authors use a simple count of certain linguistic features to determine if a sentence is relevant. Similarly, [9] creates simple hand-crafted rules for different NLP tags (like part-of-speech and named-entity tags) to classify sequences into protest relevant or not.

The authors of [10] present an actual use case by using the gathered information to create a protest/demonstration forecast system that is able to predict the occurrences of planned protests by analyzing "open-source documents that appear to indicate civil unrest event planning". They apply simple statistical models to do phrase filtering and Probabilistic Soft Logic to identify geographical information.

It can be seen that most of the considered approaches use simple term representations, that do not take into account the context of the term or only do so for the train set, which means that terms that never occurred in the training set will always have a zero-vector. To achieve a better approximation for such

words, the authors of [14] train their model to generate representations for parts of words. The idea is to better incorporate subword information and to be able to generate encodings for out-of-vocabulary terms by combining the encoding of different parts of the word.

Distributed term representations (word embeddings) and a Bi-LSTM have often been used in recent research to solve the task of sequence tagging and classification (see for example [13, 17]). Due to the success of this combination and improvements to the contextualized word embeddings in the recent years, we have chosen to use the flair embeddings and their language model for all tasks in consideration. The chosen approaches will be described in more detail in the following sections.

## 3 Methodology

### 3.1 Dataset

The provided datasets are individual for each task and consist of newspaper articles, taken from Indian and Chinese online-newspapers. The training data consists of news articles from Indian sources whereas the test data is represented by two sets, one containing Indian news sources (test) and the other Chinese news sources (test_china). All datasets are in the English language.

In the case of Task 2, the provided training dataset is imbalanced, 988 sentences have been tagged as protest-related and 4897 as not. Detailed descriptive statistics for each dataset can be found in Table 1.

For Task 3 the provided train dataset consists of 21623 labeled tokens. The tokens were labeled according to the BIO labeling scheme were a (B) label indicates the first token of an entity, an (I) label all following tokens and an (O) label all tokens, outside of entities. Entities of interest included: 'participant', 'trigger', 'loc', 'place', 'etime', 'fname', 'target'. Full explanations on the what is considered in each entity type can be found in [19].

**Table 1.** Dataset description per task

| | **Task2** | | | | **Task3** |
|---|---|---|---|---|---|
| | ***# of sent*** | ***min length*** | ***max length*** | ***average length*** | ***# of tokens*** |
| *train* | 5882 | 3 | 988 | 146 | 21623 |
| *dev* | 662 | 15 | 539 | 153 | 3188 |
| *test* | 1107 | 2 | 499 | 147 | 6506 |
| *test_china* | 1235 | 6 | 664 | 141 | 4345 |

### 3.2 Approach

In the framework of the ProtestNews Labs we wanted to evaluate how well contextual string embeddings perform in sequence labeling and classifying protest-

related news, and whether the models trained on data from one country can be applied to data from other countries.

**Task 2.** In this task the goal was to classify whether the above mentioned sentences contain an event-trigger or not. To solve this task we chose an approach using contextualized distributed term representations [13] to represent the input text. For this purpose, we stacked different traditional word embeddings such as GloVe [12] and FastText [14] together with the contextualized embeddings generated from Flair language models (LM), as suggested by [13]. Flair LMs are character-level Bi-LSTMs, pre-trained on the task of predicting the most probable next character in a sequence of characters. Therefore, they encode in their representation all previous and all following words from the given input sequence. In our work, we used the Flair-LMs pre-trained on a news corpus (news-forward-fast) and the corresponding inverted corpus(news-backward-fast). We choose the described approach for its ability to capture the context of the input, which proved to be useful for this specific task.

In the next step, we used the generated representations for every word in the given input sequence to derive a vector representation for the whole input sequence by using LSTM-based document embeddings [6]. In contrast to pooled document embeddings which (by default) represent the document by averaging all word representations, LSTM-based document embeddings take the word vectors as input features and are fine-tuned to the specific downstream task to extract the resulting document embedding from the last hidden state after the fine-tuning [6].

The resulting document embeddings were used to classify every input sequence as containing an event-trigger (1) or not (0). The classification itself was performed by a linear transformation, using a single linear layer with the dimension of the resulting document embeddings.

We tested both the original as well as preprocessed input sequences, where the preprocessing steps included stopword removal, removal of named entities, such as date and time and removal of short words. Since the preprocessed sequences lead to a significant performance drop, we fine-tuned the Flair-LM on the original texts.

**Task 3.**The aim of this task is to develop a generalized model to extract event related information, such as location, time and participants, from given sentences. The sentences are, as above mentioned, taken from online newspapers. The task is framed as a named entity recognition (NER) task and therefore a token labeling problem.

For task 3 we chose pooled contextualized embeddings [15] for their better performance compared to the non-pooled version. While the standard version of the contextualized flair embeddings does only account for the context of a token per sentence, the pooled embeddings combine the contexts of all usages of the term in the input and concatenates the resulting vector with the contextualized vector for the sequence of interest. The process of generating the embeddings is described in detail in [15].

In both approaches models were trained using the provided *train* set and validated on the *dev* data.

## 4 Results

### 4.1 Task 2. Event sentence detection

In the framework of the sentence event detection task we have submitted two runs experimenting with different minibatch sizes. In both runs we have used LSTM-based document embeddings built with stacked Glove and contextualized string embeddings. In our experiments, we used the hyperparameter settings proposed by the authors in [13]. The only difference between Run 1 and Run 2 is in the minibatch size, which has been set to 16 and 8 respectively.

The results obtained by our runs for each dataset as well as the best results in the track and baseline provided by the organizers are presented in Table 2. According to [19], a Linear Support Vector Classification with a stochastic gradient descent learning model was selected as a baseline model. For the official testing phase the average of F-scores obtained for each task was used as the performance measure. The second submission was used as our final submission, positioning us in third place.

When comparing the results achieved by the first and second run, a decrease in quality of classification with the increase in minibatch size can be observed. This may be explained by the following. In cases where models tend to overfit, the gradients calculated with a small batch size are much more noisy than gradients calculated with large batch size, so it is easier for the model to escape from sharp minimizers, and thus leads to a better generalization [16].

One of the goals of the ProtestNews Labs track is to build a model able to generalize outside of the country domain used for training, making it possible to use the the same model to detect event sentences in news sources from other countries. Thus, it is interesting to see not only how well the model performs on data of the two considered countries, but also how big the difference between the achieved results is. In Table 2 it can be seen that in Run 2, the gap between the India test score and China test score is the lowest, which can indicate a higher cross-country generalization ability of the proposed model.

**Table 2.** Evaluation of the results obtained by different runs on the test datasets

| | **Set 1 (China)** | **Set 2 (India)** | **Average F-score** |
|---|---|---|---|
| | ***F-score*** | ***F-score*** | |
| *Baseline* | 0.200 | 0.582 | 0.391 |
| *Best results* | 0.604 | 0.706 | 0.655 |
| *Run 1* | 0.523 | 0.617 | 0.570 |
| *Run 2* | 0.583 | 0.648 | 0.615 |

### 4.2 Task 3. Event tagging

We have submitted 2 runs experimenting with different standard word embeddings, which were stacked with pooled contextualized string embeddings, as recommended in [15]. Using these embeddings we trained our own Sequence Tagging Models. In the first run we have used FastText embeddings [14], whereas in the second run we tried the Glove embeddings [12]. In our experiments, we use the hyperparameter settings recommended by the authors in [15] as they achieved state-of-the-art performance in other natural language processing tasks.

The results obtained by our runs for each dataset are presented in Table 3. During the testing phase the average of F-scores obtained for each dataset was used as the performance measure. The second submission was used as our final submission and landed us the first place. It can be seen that there is a considerable difference in the results obtained for different countries.

**Table 3.** Evaluation of the results obtained by different runs on the test datasets

| | **Set 1 (China)** | | | **Set 2 (India)** | | | **Macro average** | **Macro average** | **Macro average** |
|---|---|---|---|---|---|---|---|---|---|
| | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| *Run 1* | 62.72 | 41.00 | 49.59 | 61.44 | 47.98 | 53.88 | 62.08 | 44.49 | 51.73 |
| *Run 2* | 62.65 | 46.24 | 53.21 | 66.20 | 55.67 | 60.48 | 64.43 | 50.96 | 56.85 |

## 5 Conclusion and Future Work

In this paper we tackled the problem of event sentence detection and event tagging in protest-related news articles at the CLEF ProtestNews Lab. The proposed solutions were based on using contextualized string embeddings. We achieved the best F-score in extracting relevant information from event-related sentences, and the third-best F-score in classifying sentences from news articles.

The improvement of the generalization ability of the approach will be the main focus of our future work. We will try other embeddings such as bert[18] to further investigate if an attention based incorporation of the context improves the performance.

## References


1. BeautifulSoup - bs4, https://www.crummy.com/software/BeautifulSoup/. Last accessed 21 May 2019
2. Natural Language Toolkit, https://www.nltk.org/. Last accessed 21 May 2019
3. Explosion AI - spaCy, https://www.spacy.io. Last accessed 21 May 2019
4. Annotation Specifications - Named Entities, https://spacy.io/api/annotation#named-entities, Last accessed 21 May 2019

5. TextBlob: Simplified Text Processing, https://textblob.readthedocs.io/en/dev/, Last accessed 21 May 2019
6. Zalando Research - flair GitHub Repository, https://github.com/zalandoresearch/flair, Last accessed 21 May 2019
7. Khurdiya, Arpit, Dey, Lipika, Mahajan, Diwakar, Verma, Ishan: Extraction and Compilation of Events and Sub-events from Twitter. In: Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01, pp. 504-508, IEEE Computer Society (2012).
8. Ploeger, T., M. Kruijt, L. M. Aroyo, F. G. A. de Bakker, I. Hellsten, A. S. Fokkens-Zwirello, J. E. Hoeksema, S. ter Braake: Extractivism: Extracting Activist Events from News Articles Using Existing NLP Tools and Services. In: CEUR Workshop Proceedings, pp. 3041, CEUR Workshop Proceedings (2013).
9. Huang, Bert, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, Naren Ramakrishnan: Planned Protest Modeling in News and Social Media. In: Proceedings of the Twenty-Seventh Innovative Applications of Artificial Intelligence Conference, AAAI Press (2015).
10. Papanikolaou, Konstantina, Haris Papageorgiou, Nikos Papasarantopoulos, Theoni Stathopoulou, George Papastefanatos: Just the Facts with PALOMAR: Detecting Protest Events in Media Outlets and Twitter. In: The Workshops of the Tenth International AAAI Conference on Web and Social Media Social Media in the Newroom: Technical Report WS-16-19, 13542, pp. 135-142, AAAI Press (2016).
11. Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space. ArXiv:1301.3781, arXiv (2013).
12. Pennington, Jeffrey, Richard Socher, Christopher Manning: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 15321543, Association for Computational Linguistics (2014).
13. Akbik, Alan, Duncan Blythe, Roland Vollgraf: Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 16381649 Association for Computational Linguistics (2018).
14. Bojanowski, Piotr, Edouard Grave, Armand Joulin, Tomas Mikolov: Enriching Word Vectors with Subword Information. ArXiv:1607.04606, arXiv (2016).
15. Akbik, Alan, Tanja Bergmann, and Roland Vollgraf: Pooled Contextualized Embeddings for Named Entity Recognition. (2019).
16. Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P. :On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016).
17. Peters, Matthew E. and Neumann, Mark and Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke, Deep contextualized word representations, Proc. of NAACL (2018).
18. Devlin, J., Chang, M., Lee. K., Toutanova, K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv preprint arXiv:1810.04805 (2018).
19. Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruş an, F., Mutlu, O., A Task Set Proposal for Automatic Protest Information Collection Across Multiple Countries. In European Conference on Information Retrieval, pp. 316-323, Springer, Cham (2019) https://link.springer.com/chapter/10.1007/978-3-030-15719-7_42