CASE 2021 Task 2: Socio-political Fine-grained Event Classification using Fine-tuned RoBERTa Document Embeddings

Samantha Kent
Fraunhofer FKIE
Fraunhoferstraße 20
53343 Wachtberg
samantha.kent@
fkie.fraunhofer.de

Theresa Krumbiegel
Fraunhofer FKIE
Fraunhoferstraße 20
53343 Wachtberg
theresa.krumbiegel@
fkie.fraunhofer.de

Abstract

We present our submission to Task 2 of the Socio-political and Crisis Events Detection Shared Task at the CASE @ ACL-IJCNLP 2021 workshop. The task at hand aims at the fine-grained classification of socio-political events. Our best model was a fine-tuned RoBERTa transformer model using document embeddings. The corpus consisted of a balanced selection of sub-events extracted from the ACLED event dataset. We achieved a macro F-score of 0.923 and a micro F-score of 0.932 during our preliminary experiments on a held-out test set. The same model also performed best on the shared task test data (weighted F-score = 0.83). To analyze the results we calculated the topic compactness of the commonly misclassified events and conducted an error analysis.

1 Introduction

Event detection and classification as Natural Language Processing (NLP) tasks can be used to analyze data gathered in the information space. The findings of this analysis can then be connected to events in the physical world and contribute to situational awareness, particularly when they are related to socio-political events. The sheer amount of data that is generated and stored in the information space every day, means that strategies need to be developed to be able to efficiently and effectively process this data. Given the large amounts of data, deep learning strategies are often preferred. However, time and computational constraints may play a role in deciding how to extract and analyze data.

Task 2 in the Socio-political and Crisis Events Detection Shared Task at the CASE @ ACL-IJCNLP 2021 workshop aims at the fine-grained classification of events (Haneczok et al., 2021). The task is based on data extracted from the Armed Conflict Location & Event Data (ACLED) database (Raleigh et al., 2010). It consist of socio-political events that have been annotated based on the ACLED event taxonomy, and includes 6 event types and 25 event subtypes. The aim of this task is to label event snippets using a model trained on data from the ACLED dataset, in order to see how robust models are when presented with data that is not directly covered by ACLED or contains unseen event classes. The results presented in this paper pertain only to subtask 1, where the task is the classification of 25 different event subtypes with ACLED-compliant labels. In other words, all the classes are seen classes from the ACLED dataset. The second and third subtasks are zero-short learning tasks that contain unseen classes.

This paper proceeds by first describing the data collection process in section 3. Section 4 contains the system description and the following section contains the experimental results. Section 6 provides an overview of the results based on the test data provided by the organizers. Finally, in section 7 the error analysis provides an insight into the system results and the data.

2 Related Work

Previous research in event detection and classification shows that there are numerous approaches to solve the problem of detecting events in texts. Xiang and Wang (2019) give a coherent overview of suitable strategies, starting with earlier approaches like pattern matching, and describing methods of machine learning as well as deep learning. There have been a number of shared tasks that have taken place in previous years that contribute to research conducted in this area. Specifically, the shared tasks CLEF 2019 Protest News (Hürriyetoğlu et al., 2019), AESPEN 2020 (Hürriyetoğlu et al., 2020), and CASE 2021 (task 1) (Hürriyetoğlu et al., 2021) focus on event detection at both the sentence and

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 208–217

document level, as well as event co-reference resolution.

Currently, not much research has been conducted that further analyzes event data once the events have been identified. There are a handful of studies across different domains. Peng et al. (2019) achieve state of the art results detecting and classifying social event data with a Pairwise Popularity Graph Convolutional Network (PP-GCN) with an external knowledge base. Nugent et al. (2017) compare different supervised classification methods for detecting a range of different events, and achieve good results with Support Vector Machines (SVM) and Convolutional Neural Networks (CNN). A benchmark corpus for fine-grained political event classification was created by the organizers of this task and an initial exploration and classification of the data is reported on in Piskorski et al. (2020) and Piskorski and Jacquet (2020). The findings reported that BERT transformer models achieved a micro F1 of (0.943-0.949) and a macro F1 of (0.860-0.889). More simple TF-IDF-weighted character n-gram models also achieved good results. A large dataset of 600,000 annotated ACLED event snippets was used as training data.

3 Data collection

Due to copyright reasons, the data used in this paper was collected directly from the ACLED website.[1] To create the corpus, all data from each available region was downloaded and then filtered using the following steps.

Firstly, all events with less than 25 tokens and more than 1000 tokens were removed. The next step was to balance the corpus based on the 25 different fine-grained event classes. Originally, the largest class in the corpus consisted of 36.69% of the events, compared to the smallest with 0.001%. To create a more balanced version of the corpus, we extracted a sample of events per class, with the smallest classes being fully represented and extracting only a percentage of the largest classes. Note that it was not possible to fully balance all of the classes as there was only a very small amount of data for classes such as CHEM_WEAP. A random sample of this balanced corpus was then split into a train (n=94000), development (n=9000), and test (n=2500) corpus, which also all contain the balanced class distribution. We observed that randomizing the order of events was crucial, to avoid introducing a bias based on the different ACLED regions. Figure 2 illustrates the distribution of the corpus. A more detailed table can be found in appendix A.

In a further step, we created three different versions of the original corpus. The first version, referred to as ACLED_N, contains the original text from the ACLED download, an example of which can be found below.

```
{text: CPI(M) activists attacked
    a BJP rally in Hrishyamukh on
    18 January 2018.,
 subtype: FORCE_AGAINST_PROTEST}
```

Based on the results presented by Piskorski et al. (2020), where the BERT transformer model performed slightly better on the corpus with less preprocessing, we decided to include a version with little to no pre-processing. In ACLED_L, we replaced all locations from the text using the Flair Named Entity Recognition (NER) tagger (Akbik et al., 2018) with the generic token 'LOC'. The third version, ACLED_T, contains a pre-processed version of the original text, but without any time stamps. All dates and times were removed from the text and replaced with 'TIME'. These two alternative versions of the corpus were created to analyse whether or not the information specific to one particular event or set of events would be transferable to the classification of other events.

4 System Description

We submitted five system runs for evaluation. The systems differ slightly from each other, either in the model or the way the used data was pre-processed. The general approach for all submitted systems was to use fine-tuned pre-trained transformer document embeddings. All experiments were conducted using the Flair framework (Akbik et al., 2019).

4.1 System 1 - RoBERTa ACLED_L

For system 1, we fine-tuned the RoBERTa base model (Liu et al., 2019), and trained the embeddings using a learning rate of 3e-5, a batch size of 16. Based on our experiments, we trained the model for 2 epochs, because we found that the model overfits if we trained for more than 2 epochs. After each epoch the training data was shuffled and this was also done in the subsequent systems. Additionally, we assigned weights to the different event classes. This was done to smooth out any

[1] https://acleddata.com/data-export-tool/

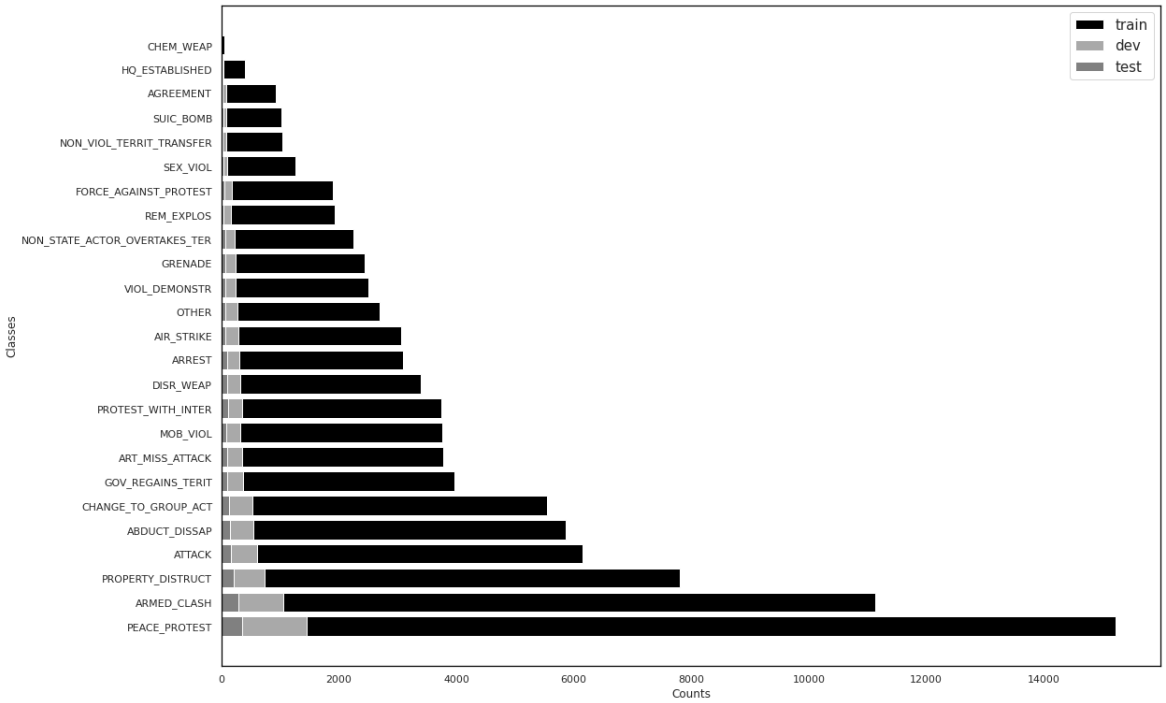

Figure 1: Class distribution

remaining differences in class sizes. We used the ACLED_L version of the corpus as text input.

4.2 System 2 - RoBERTa ACLED_N

System 2 again uses the RoBERTa base model (Liu et al., 2019) and the previously mentioned parameters for learning rate (3e-5), batch size (16) and number of epochs (2). The difference to system 1 is, that the text that was used during the fine-tuning of the model was not pre-processed. This means that the text snippets that were obtained from the ACLED (Donnay et al., 2019) database were fed into the system in their original state and, therefore, all information included in the text was kept.

4.3 System 3 - BERT ACLED_L

For system 3, we used the pre-trained BERT base-cased model (Devlin et al., 2019) along with a learning rate of 3e-5, a batch size of 16 and 2 epochs for training. As in system 1, we used the ACLED_L corpus.

4.4 System 4 - BERT ACLED_N

System 4 used the same settings as system 3, meaning, the pre-trained BERT base-cased model (Devlin et al., 2019), a learning rate of 3e-5, a batch size of 16 and 2 epochs for training. The input data for system 4 consisted of the original text from ACLED_N.

4.5 System 5 - BERT ACLED_T

Our last system, system 5, made use of the pre-trained BERT base-cased model (Devlin et al., 2019). The learning rate was set to 3e-5, the batch size to 32. It was trained for 2 epochs. For the text input we used the text from ACLED_T, where all time and date stamps have been removed.

5 Preliminary experiments

Preliminary model evaluations on 10 held-out test sets show that each of the systems performed comparatively well. The RoBERTa model with the normal ACLED text as input performed slightly better than the other systems. Table 1 below shows the range of Macro and Micro F1 scores across the 10 test sets. Model performance increased or decreased slightly, depending on the samples in the individual test sets. The results also illustrate that the removal of the location or time mentions in the event snippets, does not greatly influence system performance. Rather, the preliminary tests indicate that the fine-tuned RoBERTa embeddings benefit from the inclusion of the more detailed ACLED specific information.

An analysis of the results of the individual classes, shows that each of the 25 subtypes achieve f1-scores of over 0.800. The two lowest scoring classes are HQ_ESTABLISHED and

	Macro F1	Micro F1	Weighted F1
RoBERTa ACLED_L	0.887 - 0.919	0.917 - 0.929	0.917 - 0.929
RoBERTa ACLED_N	0.894 - **0.923**	0.916 - **0.932**	0.918 - **0.931**
BERT ACLED_L	0.868 - 0.911	0.913 - 0.928	0.913 - 0.928
BERT ACLED_N	0.869 - 0.900	0.907 - 0.925	0.907 - 0.925
BERT ACLED_T	0.889 - 0.918	0.913 - 0.929	0.913 - 0.929

Table 1: Preliminary Evaluation Results

VIOL_DEMONSTR, with an F-score of 0.814 and 0.819 respectively. The highest scoring class is CHEM_WEAP (F-score = 1), however there are only two instances present in this test set. PEACE_PROTEST and ABDUCT_DISSAP also score highly, achieving an F-score of 0.978 and 0.975 respectively. A table containing a detailed overview of each class can be found in appendix A.

6 Results

Table 7 shows the results of our five system submissions. The systems were tested on a test set provided by the organizers, consisting of 829 samples for subtask 1. We find that System 2, using the RoBERTa base model (Liu et al., 2019) and ACLED_N as input, performs best with an average weighted F-score of 0.83, average macro F-score of 0.794 and average micro F-score of 0.829. A more detailed overview can be found in appendix A.

Additionally, the second model that uses original ACLED text, System 4, achieves the second best result. As was the case in our preliminary experiments, we see that the inclusion of specific location and timestamps in the training data, does not greatly influence the ability of the system to predict the different classes correctly or incorrectly.

	Macro F1	Micro F1	Weighted F1
RoBERTa ACLED_L	0.797	0.770	0.799
RoBERTa ACLED_N	**0.829**	**0.794**	**0.830**
BERT ACLED_L	0.808	0.768	0.808
BERT ACLED_N	0.802	0.774	0.812
BERT ACLED_T	0.793	0.766	0.793

Table 2: System Results

7 Error Analysis

To get a better insight into the workings of our systems, we conducted an error analysis on the test data provided by the organizers for all five submissions. In order to investigate misclassifications made by the models, we decided to look at the performance of the system with regard to the individual classes.

7.1 Analysis of Word Frequencies

As can be seen in Table 3, all models score low F-scores for either the class OTHER or the class PROPERTY_DISTRUCT, or both. The results obtained for these classes substantially lower the overall average F-scores of the models.

	Worst Class	F1
RoBERTa ACLED_L	OTHER	0.42
RoBERTa ACLED_L	PROPERTY_DISTRUCT	0.46
RoBERTa ACLED_N	PROPERTY_DISTRUCT	0.35
RoBERTa ACLED_N	OTHER	0.40
BERT ACLED_L	PROPERTY_DISTRUCT	0.30
BERT ACLED_L	OTHER	0.34
BERT ACLED_N	OTHER	0.28
BERT ACLED_N	MOB_VIOL	0.49
BERT ACLED_T	PROPERTY_DISTRUCT	0.41
BERT ACLED_T	NON_STATE_ACTOR _OVERTAKES_TER	0.49

Table 3: Event Type Error Analysis

All models achieve the highest scores for the classes SUIC_BOMB, GRENADE and CHEM_WEAP. We looked at the word distribution these classes have in our training data as can be seen in figure 2 and 3.

Considering these distributions, it can be stated that a specific vocabulary, as can be found in the class SUIC_BOMB, is advantageous for a correct classification, while a heterogeneous vocabulary, as can be found in the class OTHER, is disadvantageous. One can tell that while the by far most frequently occurring word in texts regarding the event type SUIC_BOMB, namely ”suicide”, is clearly indicative for the given class, the most frequently used words in connection with the event type OTHER, namely ”activity”, ”violent”, ”area” and ”force” are rather generic. Furthermore, they can also be found frequently in a number of texts connected

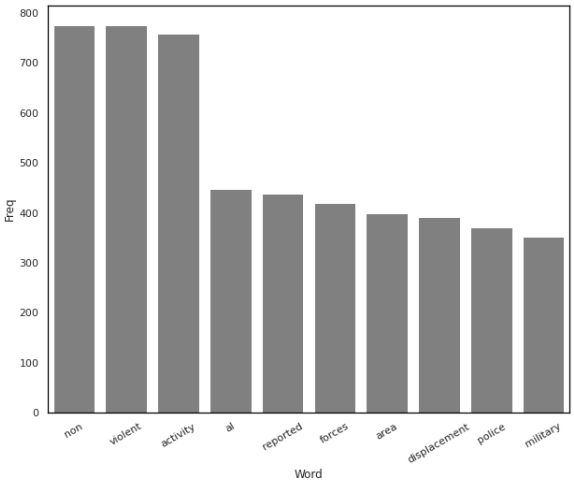

Figure 2: Top 10 words in the class OTHER

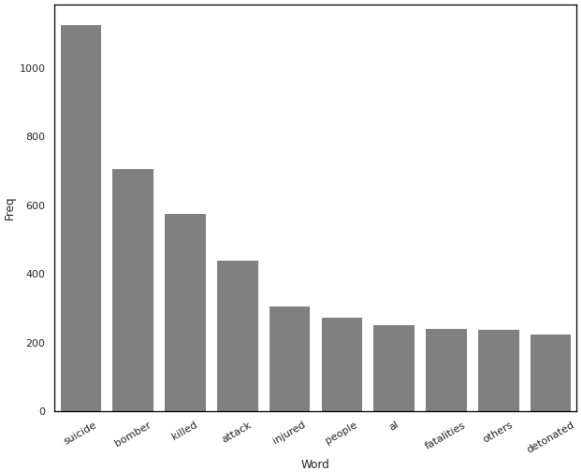

Figure 3: Top 10 words in the class SUIC_BOMB

to other classes (e.g., ”area”: AIR_STRIKE, CHANGE_TO_GROUP_ACT, ”force”: NON_STATE _ACTOR_OVERTAKES_TER, NON_VIOL_TERRIT_TRANSFER). This does not hold true for the word ”suicide”.

7.2 Frequent Errors

Looking further at the errors, we see that 65 samples of the test data were classified incorrectly by all five models. This makes up between 37% and 45% of errors for the respective systems. It is noticeable that all five models frequently predict the class MOB_VIOL for sentences that are gold labeled as PROPERTY_DISTRUCT (between 5 and 9 times for the respective systems). No other two classes are confused this often, and to investigate further we analysed these two classes with regard to their topic compactness. We calculated the topic distances of the sentences in comparison to the topic centroids per class in the training data. Figures 4 and 5 show the results of the topic compactness analysis.

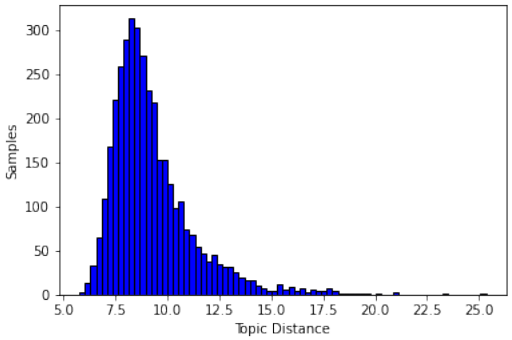

Figure 4: Distribution of document vectors to topic centroid in class MOB_VIOL

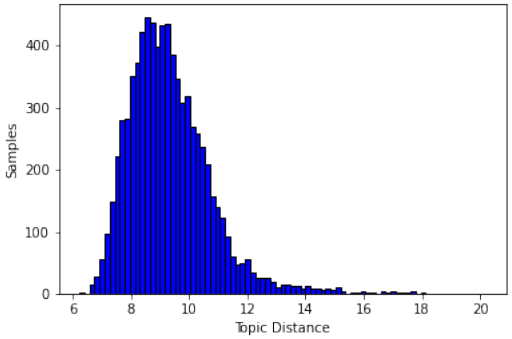

Figure 5: Distribution of document vectors to topic centroid in class PROPERTY_DISTRUCT

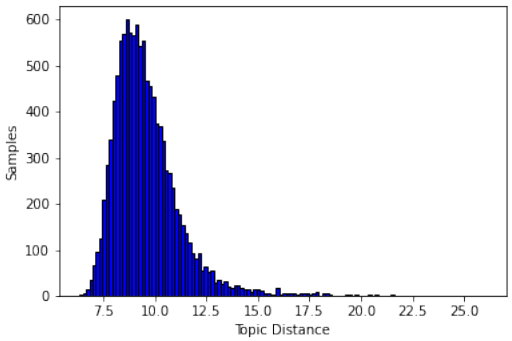

Figure 6: Distribution of document vectors to topic centroid in the combined class PROPERTY_DISTRUCT and MOB_VIOL

We see that both classes, MOB_VIOL and PROPERTY_DISTRUCT, are quite compact. There are

some outliers, but most of the document vectors are clustered close to each other and the topic centroid. However, if we combine the classes into one topic and again analyse the distribution of document vectors to the topic centroid, we find that there are also very few outliers, as can be seen in figure 6. This means that the examples for MOB_VIOL and PROPERTY_DISTRUCT in our training data are similar to each other, which may explain why our models consistently confuse these two classes with regard to the test data provided by the organizers.

Looking at the test samples, we further find that due to the large number of classes, it is also difficult for human annotators to distinguish between the different classes in some cases. An example for this is the following:

```
{text: Police said two groups
   from different communities in
   Chhabra town of Rajasthan's
   Baran district pelted stones
   on each other and torched
   vehicles parked around after
   putting six shops afire,
 guess: MOB_VIOL,
 gold: PROPERTY_DISTRUCT}
```

All our models consistently predict the event class MOB_VIOL for this example, the gold standard annotation is, however, PROPERTY_DISTRUCT. It can be argued that the given example actually includes both event classes, with the first part of the sentence, "Police said two groups from different communities in Chhabra town of Rajasthan's Baran district pelted stones on each other" being an instance of MOB_VIOL, while the second part, "and torched vehicles parked around after putting six shops afire", belongs to the class PROPERTY_DISTRUCT. Test instances like this pose a challenge for the models.

8 Conclusion

In this study we proposed the use of fine-tuned RoBERTa transformer document embeddings for the fine-grained classification of socio-political events. We balanced the corpus to ensure that the 25 subtypes were represented as equally as possible. Compared to the results that were achieved during the preliminary experiments, we observed a drop in performance on the test set provided by the organizers. However, compared to the baseline figures provided by the organizers in (Piskorski et al., 2020), we achieved very similar results with less training data. This suggests that balancing the training data data had a positive effect on model performance.

Our analysis of the results of both different test sets, the set created for preliminary experiments and the set provided by the organizers for system evaluation, show that there is definitely a difference in performance in the various classes. It also highlighted the issue of events that could be classed as more than one different subtype, and the challenge that these events pose for fine-grained classification. Depending on the given use case, parts of our system could already be implemented in a real world setting in order to analyze the flow of data in the information space and achieve situational awareness in the physical world, as clear cut classes like CHEM_WEAP and GRENADE are identified reliably. In a military setting, for example, these classes are far more relevant than occurrences of PROPERTY_DISTRUCT.

In future work, it would be interesting to evaluate if the use of more training data, while still trying to obtain a more even distribution of classes, would further increase performance. Particularly, it raises the question if more training data would increase performance for the classes that currently do not perform as well. A more thorough class analysis, which would contribute to understanding why there seem to be systematic errors in specific classes, could provide insight into answering this question.

Acknowledgments

This research has been supported through funding from Philip Morris Impact as part of the Fraud Information Fusion Intelligence Project.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Karsten Donnay, Eric T. Dunford, Erin C. McGrath, David Backer, and David E. Cunningham. 2019. Integrating conflict event data. *Journal of Conflict Resolution*, 63(5):1337–1364.

J Haneczok, G Jacquet, J Piskorski, and N Stefanovitch. 2021. Fine-grained event classification in news-like text snippets shared task 2, case 2021. In *Proceedings of the Workshop on Challenges and Applications of Automated Text Extraction of Socio-Political Event from Text (CASE 2021), co-located with the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, case 2021. In *Proceedings of the Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Erdem Yörük, Deniz Yüret, Çağrı Yoltar, Burak Gürel, Fırat Duruşan, Osman Mutlu, and Arda Akdemir. 2019. Overview of clef 2019 lab protestnews: Extracting protests from news in a cross-context setting. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 425–432, Cham. Springer International Publishing.

Ali Hürriyetoğlu, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu. 2020. Automated extraction of socio-political events from news (AESPEN): Workshop and shared task report. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 1–6, Marseille, France. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Tim Nugent, Fabio Petroni, Natraj Raman, Lucas Carstens, and Jochen L. Leidner. 2017. A comparison of classification models for natural disaster and critical event detection from news. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3750–3759.

Hao Peng, Jianxin Li, Qiran Gong, Yangqiu Song, Yuanxin Ning, Kunfeng Lai, and Philip S. Yu. 2019. Fine-grained event categorization with heterogeneous graph convolutional networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3238–3245. International Joint Conferences on Artificial Intelligence Organization.

Jakub Piskorski, Jacek Haneczok, and Guillaume Jacquet. 2020. New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6663–6678, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jakub Piskorski and Guillaume Jacquet. 2020. TF-IDF character N-grams versus word embedding-based models for fine-grained event classification: A preliminary study. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 26–34, Marseille, France. European Language Resources Association (ELRA).

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47(5):651–660.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

A Class Distribution and Results

Class	Train Nr. %	Dev Nr. %	Test Nr. %
PEACE_PROTEST	15229 (16.04)	1452 (16.13)	362 (14.48)
ARMED_CLASH	11132 (11.72)	1055 (11.73)	291 (11.64)
PROPERTY_DISTRUCT	7802 (8.22)	732 (8.13)	216 (8.64)
ATTACK	6153 (6.48)	610 (6.78)	166 (6.64)
ABDUCT_DISSAP	5871 (6.18)	550 (6.11)	154 (6.16)
CHANGE_TO_GROUP_ACT	5548 (5.84)	529 (5.88)	135 (5.40)
GOV_REGAINS_TERIT	3974 (4.18)	366 (4.07)	104 (4.16)
ART_MISS_ATTACK	3770 (3.97)	360 (4.00)	103 (4.12)
MOB_VIOL	3755 (3.95)	320 (3.56)	92 (3.68)
PROTEST_WITH_INTER	3740 (3.94)	354 (3.94)	109 (4.36)
DISR_WEAP	3388 (3.57)	330 (3.67)	99 (3.97)
ARREST	3098 (3.26)	311 (3.46)	97 (3.89)
AIR_STRIKE	3061 (3.22)	299 (3.32)	73 (2.92)
OTHER	2701 (2.84)	276 (3.07)	63 (2.52)
VIOL_DEMONSTR	2507 (2.64)	246 (2.73)	75 (3.00)
GRENADE	2439 (2.57)	238 (2.64)	70 (2.80)
NON_STATE_ACTOR_OVERTAKES_TER	2252 (2.37)	220 (2.44)	67 (2.68)
REM_EXPLOS	1938 (2.04)	170 (1.89)	44 (1.76)
FORCE_AGAINST_PROTEST	1900 (2.00)	178 (1.98)	56 (2.24)
SEX_VIOL	1260 (1.33)	106 (1.18)	30 (1.20)
NON_VIOL_TERRIT_TRANSFER	1033 (1.09)	90 (1.00)	27 (1.08)
SUIC_BOMB	1023 (1.08)	82 (0.91)	30 (1.20)
AGREEMENT	927 (0.98)	78 (0.86)	22 (0.89)
HQ_ESTABLISHED	406 (0.43)	40 (0.44)	13 (0.52)
CHEM_WEAP	57 (0.06)	8 (0.08)	2 (0.08)
Total	94964	9000	2500

Table 4: Corpus class distribution of the 25 event subtypes.

System	Avg Type	Avg Prec	Avg Recall	Avg F-score
System 1	micro avg	0.797	0.797	0.797
System 1	macro avg	0.790	0.778	0.770
System 1	weighted avg	0.824	0.797	0.799
System 2	micro avg	0.829	0.829	**0.829**
System 2	macro avg	0.807	0.808	**0.794**
System 2	weighted avg	0.851	0.829	**0.830**
System 3	micro avg	0.808	0.808	0.808
System 3	macro avg	0.787	0.779	0.768
System 3	weighted avg	0.828	0.808	0.808
System 4	micro avg	0.802	0.802	0.802
System 4	macro avg	0.788	0.789	0.774
System 4	weighted avg	0.841	0.802	0.812
System 5	micro avg	0.793	0.793	0.793
System 5	macro avg	0.780	0.780	0.766
System 5	weighted avg	0.817	0.893	0.793

Table 5: Detailed System Results

Class	Precision	Recall	F1-score	Support
ABDUCT_DISSAP	0.9787	0.9718	0.9753	142
AGREEMENT	0.8974	1.0000	0.9459	35
AIR_STRIKE	1.0000	0.9333	0.9655	75
ARMED_CLASH	0.9429	0.8771	0.9088	301
ARREST	0.9500	0.9500	0.9500	80
ART_MISS_ATTACK	0.9515	0.9608	0.9561	102
ATTACK	0.8361	0.9217	0.8768	166
CHANGE_TO_GROUP_ACT	0.9716	0.9648	0.9682	142
CHEM_WEAP	1.0000	1.0000	1.0000	2
DISR_WEAP	0.9444	0.9659	0.9551	88
FORCE_AGAINST_PROTEST	0.8302	0.9167	0.8713	48
GOV_REGAINS_TERIT	0.8900	0.8900	0.8900	100
GRENADE	0.9744	0.9870	0.9806	77
HQ_ESTABLISHED	0.6875	1.0000	0.8148	11
MOB_VIOL	0.8598	0.8846	0.8720	104
NON_STATE_ACTOR_OVERTAKES_TER	0.8889	0.9014	0.8951	71
NON_VIOL_TERRIT_TRANSFER	0.8966	0.8387	0.8667	31
OTHER	0.9531	0.8971	0.9242	68
PEACE_PROTEST	0.9863	0.9703	0.9782	370
PROPERTY_DISTRUCT	0.9700	0.9417	0.9557	206
PROTEST_WITH_INTER	0.9082	0.8990	0.9036	99
REM_EXPLOS	0.9107	0.9273	0.9189	55
SEX_VIOL	0.9630	0.8966	0.9286	29
SUIC_BOMB	0.9643	0.9643	0.9643	28
VIOL_DEMONSTR	0.7722	0.8714	0.8188	70
weighted avg	0.9338	0.9312	0.9318	2500

Table 6: RoBERTa ACLED_N Detailed Class Evaluation - Prelim. Test Data

Class	Precision	Recall	F1-score	Support
DISR_WEAP	0.915	0.931	0.923	58
ABDUCT_DISSAP	0.792	0.950	0.864	20
AGREEMENT	1.000	0.774	0.873	31
AIR_STRIKE	1.000	0.833	0.909	36
ARMED_CLASH	0.817	0.742	0.778	66
ART_MISS_ATTACK	0.838	0.861	0.849	36
ATTACK	0.806	0.926	0.862	27
CHANGE_TO_GROUP_ACT	0.731	0.633	0.679	30
CHEM_WEAP	1.000	0.865	0.928	37
ARREST	1.000	0.676	0.807	34
FORCE_AGAINST_PROTEST	0.692	0.783	0.735	23
GOV_REGAINS_TERIT	0.822	0.974	0.892	38
GRENADE	0.958	0.958	0.958	48
HQ_ESTABLISHED	0.724	0.955	0.824	22
MOB_VIOL	0.414	0.706	0.522	17
NON_STATE_ACTOR_OVERTAKES_TER	0.625	0.833	0.714	24
NON_VIOL_TERRIT_TRANSFER	0.800	0.762	0.780	21
OTHER	0.333	0.500	0.400	8
PEACE_PROTEST	0.864	0.895	0.879	57
PROPERTY_DISTRUCT	0.714	0.238	0.357	21
PROTEST_WITH_INTER	0.514	0.864	0.644	22
REM_EXPLOS	1.000	0.917	0.957	36
SEX_VIOL	0.957	0.957	0.957	23
SUIC_BOMB	0.976	0.976	0.976	41
VIOL_DEMONSTR	0.881	0.698	0.779	53
weighted avg	0.851	0.829	0.830	829

Table 7: RoBERTa ACLED_N Detailed Class Evaluation - Task Test Data