Fine-grained Event Classification in News-like Text Snippets Shared Task 2, CASE 2021

Jacek Haneczok
Erste Digital
Vienna, Austria
jacek.haneczok@gmail.com

Guillaume Jacquet
Joint Research Centre
European Commission
Isrpa, Italy
guillaume.jacquet@ec.europa.eu

Jakub Piskorski
Linguistic Engineering Group
Institute for Computer Science
Polish Academy of Sciences
Warsaw, Poland
jpiskorski@gmail.com

Nicolas Stefanovitch
Joint Research Centre
European Commission
Isrpa, Italy
nicolas.stefanovitch@ec.europa.eu

Abstract

This paper describes the Shared Task on Fine-grained Event Classification in News-like Text Snippets. The Shared Task is divided into three subtasks: (a) classification of text snippets reporting socio-political events (25 classes) for which vast amount of training data exists, although exhibiting different structure and style vis-a-vis test data, (b) enhancement to a generalized zero-shot learning problem, where 3 additional event types were introduced in advance, but without any training data ('unseen' classes), and (c) further extension, which introduced 2 additional event types, announced shortly prior to the evaluation phase. The reported Shared Task focuses on classification of events in English texts and is organized as part of the Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), co-located with the ACL-IJCNLP 2021 Conference. Four teams participated in the task. Best performing systems for the three aforementioned subtasks achieved 83.9%, 79.7% and 77.1% weighted F_1 scores respectively.

1 Introduction

The task of event classification is to assign to a text snippet an event type using a domain specific taxonomy. It constitutes an important step in the process of event extraction from free texts (Appelt, 1999; Piskorski and Yangarber, 2013) which has been researched since mid 90's and gained a lot of attention in the context of development of real-world applications (King and Lowe, 2003; Yangarber et al., 2008; Atkinson et al., 2011; Leetaru and Schrodt, 2013; Ward et al., 2013; Pastor-Galindo et al., 2020). While vast amount of challenges on automated event extraction, including event classification, has been organised in the past, relatively little efforts have been reported on approaches and shared tasks focusing specifically on fine-grained event classification.

This paper describes the Shared Task on Fine-grained Event Classification in News-like Text Snippets. The task is divided into three subtasks: (a) classification of text snippets reporting socio-political events (25 classes) for which vast amount of training data exists, although exhibiting slightly different structure and style vis-a-vis test data, (b) enhancement to a generalized zero-shot learning problem (Chao et al., 2016), where 3 additional event types were introduced in advance, but without any training data ('unseen' classes), and (c) further extension, which introduced 2 additional event types, announced shortly prior to the evaluation phase. The reported Shared Task focuses on classification of events in English texts and is organized as part of the Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021) (Hürriyetoğlu et al., 2021), co-located with the ACL-IJCNLP 2021 Conference. Four teams actively participated in the

The views expressed in this article are those of the authors and not necessarily those of Erste Digital.

Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) , pages 179–192

task.

The main rationale behind organising this Shared Task is not only to foster research on fine-grained event classification, a relatively understudied area, but to specifically explore robust and flexible solutions that are of paramount importance in the context of real-world applications. For instance, often available training data is slightly different from the data on which event classification might be applied (data drift). Furthermore, in real-world scenarios one is interested in quickly tailoring an existing solution to frequent extensions of the underlying event taxonomy.

The paper is organized as follows. Section 2 reviews prior work. Section 3 describes the Shared Task in more detail. Section 4 describes the training and test datasets. Next, the evaluation methodology is introduced in Section 5. Baseline and participant systems are described in Section 6. Subsequently, Section 7 presents the results obtained by these systems, whereas Section 8 discusses the main findings of the Shared Task. We present the conclusions in Section 9.

2 Prior Work

The research on event detection and classification in free-text documents was initially triggered by the Message Understanding Contests (Sundheim, 1991; Chinchor, 1998) and the Automatic Content Extraction Challenges (ACE) (Doddington et al., 2004; LDC, 2008). The event annotated corpora produced in the context of the aforementioned challenges fostered research on various techniques of event classification, which encompass purely knowledge-based approaches (Stickel and Tyson, 1997), shallow (Liao and Grishman, 2010; Hong et al., 2011) and deep machine learning approaches (Nguyen and Grishman, 2015; Nguyen et al., 2016).

Multi-lingual Event Detection and Co-reference challenge was introduced more recently in the Text Analysis Conference (TAC) in 2016[1] and 2017[2]. In particular, it included an Event Nugget Detection subtask, which focused on detection and fine-grained classification of intra-document event mentions, covering events from various domains (e.g., finances and jurisdiction).

One could observe in the last decade an ever growing interest in research on fine-grained event classification. Lefever and Hoste (2016) compared SVM-based models against word-vector-based LSTMs for classification of 10 types of company-specific economic events from news texts, whereas Nugent et al. (2017) studied the performance of various models, including ones that exploit word embeddings as features, for detection and classification of natural disaster and crisis events in news articles. Jacobs and Hoste (2020) reports on experiments of exploiting BERT embedding-based models for fine-grained event extraction for the financial domain.

Although most of the reported work in this area focuses on processing English texts, and in particular, news-like texts as presented in Piskorski et al. (2020), some efforts on event classification for non-English language were reported too. For instance, Sahoo et al. (2020) introduced a benchmark corpus for fine-grained classification of natural and man-made disasters (28 types) for Hindi, accompanied with evaluation of deep learning baseline models for this task. Furthermore, an example of fine-grained classification of cyberbullying events (7 classes) in social media posts was presented in Van Hee et al. (2015).

Work on classification of socio-political events and the related shared tasks, although not focusing on fine-grained classification, but covering event types which are in the scope of our task, was presented in Hürriyetoğlu et al. (2021) and Hürriyetoglu et al. (2019).

3 Task Description

The overall objective of this Shared Task is to evaluate the 'flexibility' of fine-grained event classifiers. Firstly, we are interested in the robustness vis-a-vis the input text structure, i.e., how classifiers trained on short texts from a curated database perform on news data taken from diverse sources where this structure is somewhat different. This corresponds to Subtask 1, which can be considered as a regular classification task. Secondly, we wanted to study how classifiers can be made flexible regarding the taxonomy used, with the aim of easily tailoring them for specific needs. This corresponds to Subtask 2 and 3, which were framed as generalized zero-shot learning problems: the label set for Subtask 2 was announced in advance, while the label set for Subtask 3 was announced on the day of the

[1] https://tac.nist.gov//2016/KBP/Event/index.html

[2] https://tac.nist.gov/2017/KBP/Event/index.html

competition.

The aforementioned objectives arise from the practical constraints of working with real data, being exposed to data drift and having different users being interested in different facets of the same events.

In order to train a fine-grained event classifier, we proposed to use ACLED (Raleigh et al., 2010) event database and the corresponding taxonomy described in the ACLED Codebook[3], which has 25 subtypes of events related to socio-political events and violent conflicts. ACLED created a large dataset of events over several years which are manually curated with a common pattern in the way of reporting events and uses a complex event taxonomy: The boundary between the definition of similar classes can be highly intricate, and can seem at point quite arbitrary. Nevertheless, ACLED presented itself as the best possible training material for the specific objectives of this Shared Task.

More precisely, the formal definitions of the different subtasks are as follows:

- **Subtask 1:**
  Classification of text snippets that are assigned to ACLED types only,
- **Subtask 2 (generalized zero-shot):**
  Classification of text snippets that are assigned to all ACLED types plus three unseen (non-ACLED) types, namely: `Organized Crime`, `Natural Disaster` and `Man-made Disaster`, these new types were announced in advance, but no training data was provided,
- **Subtask 3 (generalized zero-shot):**
  Classification of text snippets that are assigned to two additional unseen event types (`Diplomatic Event` and `Attribution of Responsibility`) on top of the ones of Subtask 2, these new types were not announced in advance.

The participating teams had the possibility to submit solutions to any number of subtasks without condition, whereas per subtask up to 5 system responses could be submitted for evaluation. More information on the event types for this Shared Task is provided in Appendix A.

[3] https://acleddata.com/acleddatanew/wpcontent/uploads/dlm_uploads/2019/01/ACLED_Codebook_2019FINAL.docx.pdf

4 Data

4.1 Training Data

For the training purposes the participants were allowed to either exploit any freely available existing event-annotated textual corpora and/or to exploit the short text snippets reporting events which are part of the large event database created by ACLED and which can be obtained from ACLED data portal[4] for research and academic purposes. Furthermore, the participants were also recommended to exploit as an inspiration the techniques for text normalization and cleaning of ACLED data, and some baseline classification models trained using ACLED data described in Piskorski et al. (2020).

4.2 Test Data

For the purpose of evaluating the predictive performance of the competing systems a dedicated test set was created based on news-like text snippets. To this end we sourced the web to collect short texts reporting on events either in the form of online news or of a similar style. We posed simple queries with label-specific keywords using conventional search engines to collect relevant text snippets. The most frequent keywords from ACLED datasets have been used a basis to form these queries. The collected set of snippets was cleaned by removing duplicates and further enhanced by adding both manually as well as automatically perturbed short news-like texts. More specifically, for selected snippets the most characteristic keywords were manually replaced by either less common or more vague expressions, so that the event type from the ACLED taxonomy can be still predicted, albeit making it more difficult. Also the reported figures, methods or outcomes of the event were subject to changes. Furthermore, about 15% of the text snippets were automatically perturbed[5] by: (a) replacing all day and month names mentions with another randomly chosen day and month resp., and (b) replacing each occurrence of a toponym referring to a populated place with randomly chosen toponym selected from GEONAMES gazetteer[6] of about 200K populated cities, whose population is at least 500. The perturbed snippets were additionally inspected in order to make sure that the changes allow for guessing the

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

[5] The choice of 15% was motivated by the willingness to add some (but not too much) additional complexity to the task.

[6] https://www.geonames.org/

event type vis-a-vis ACLED taxonomy. Only the perturbed version of the original text snippet were included in the test dataset, the original ones were discarded. An example of original text and the automatically perturbed version thereof is provided in Figure 1.

A Catalan pro-independence demonstrator throws a fence into a fire during a protest against police action in Barcelona, Spain, October 26, 2019

A Madukkarai pro-independence demonstrator throws a fence into a fire during a protest against police action in Podosinovets, Hohenmölsen, June 26, 2019

Figure 1: Sample text snippet reporting a violent demonstration event (top) and the perturbed version thereof (bottom).

The distribution of the counts by event type is shown in Figure 3, whereas the distributions of the sequence length by event type is shown in Figure 4. The created test set consists in total of 1019 text snippets, 190 of which were annotated with labels corresponding to the zero-shot classes. An example of text snippet reporting a `Government regains territory` event is provided in Figure 2.

Syrian government forces have captured a central town and adjacent villages, boosting security in nearby areas loyal to President Bashar Assad, and marched deeper into a rebel-held neighborhood of Damascus, Syrian state media and an opposition monitoring group said Sunday.

Figure 2: Sample text snippet reporting an event.

The annotation was performed by two pairs of independent annotators, cross-validating the annotated snippets. The initial disagreement rate was observed to be roughly 10-15%. Most unclear text snippets, for which there were comparably strong arguments for assigning two or more labels, were removed from the test dataset. For text snippets reporting on multiple events, the more recent event was considered to be the main event (and given the priority for determining the type), whereas the remaining events were considered only as background information. Some ambiguities were solved by aligning on common assumptions, e.g. if there is

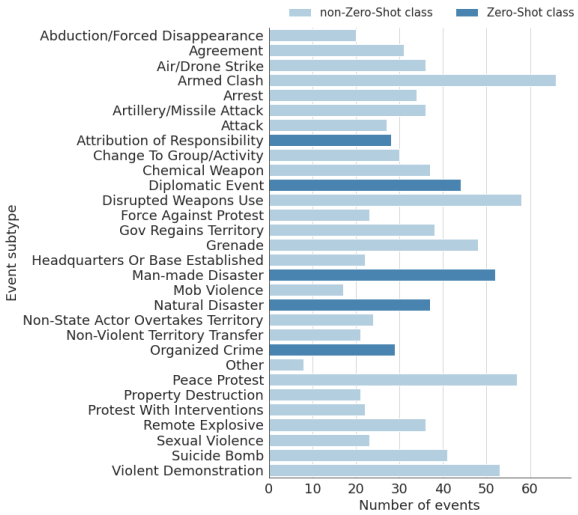

Figure 3: Event type count distribution in the test dataset.

no explicit mention of violence, a protest reported in the snippet was considered to be a peaceful one.

Figure 4: Distribution of the length of the text snippets by event type in the test dataset.

It is important to emphasize that the created test dataset for the Shared Task reported in this paper contains text snippets reporting events, which were prepared solely for the purpose of evaluating solutions for automated fine-grained classification of events reported in short texts.[7]

[7]**Disclaimer:** A significant fraction of the text snippets in the test dataset has no link to any real-world event whatsoever and, in particular, the locations mentioned therein were selected completely at random. As such, even though some of

5 Evaluation methodology

For measuring the event classification performance we used $precision$, $recall$, and the $micro$, $macro$ and $weighted$ F_1 metric. While the micro version calculates the performance from the classification of individual instances vis-a-vis the all-class model, in macro-averaging, one computes the performance of each individual class separately, and then an average of the obtained scores is computed. The $weighted$ F_1 is similar to the $macro$ version, but computes the average considering the proportion for each class in the dataset.

6 Systems

6.1 Baseline Systems

We provide two baseline systems: a simple character n-gram based L2-regularized logistic regression model and a system based on two Transformer-based deep neural representation models.

6.1.1 L2-regularized Logistic Regression on character n-grams ($L2LR_{baseline}$)

For Subtask 1 we have trained a L2-regularized Logistic Regression-based model with log-scaled TF-IDF values of 3 to 5 character ngrams found in the text snippets as features[8] (non-optimized, with $C = 1.0$ and $\epsilon = 0.01$) using LIBLINEAR library[9]. In particular, a more balanced subset of ca. 129K event snippets from ACLED-III (Piskorski et al., 2020) was used, i.e., all high-populated classes were under-sampled with a maximum of 10K instances per class.

6.1.2 Combined deep Transformers BERT and BART ($BB_{baseline}$)

As our main baseline model for Subtasks 1-3 we use a combination of two Transformer-based unsupervised language representation models: a multi-layer bidirectional Transformer encoder BERT (Devlin et al., 2019) and a sequence-to-sequence autoencoder BART (Lewis et al., 2019). As a base classifier we employ the BERT-BASE model, pre-trained using two unsupervised tasks: masked language model and next sentence prediction on lower-cased English text of the BooksCorpus (800M words) and English Wikipedia (2,500M words) and fine-tuned for supervised classification using ACLED-III data as described in Piskorski et al. (2020). For Subtasks 2-3 involving a zero-shot learning problem our baseline system relies on the following further steps. The test set observations (text snippets) for which the predicted logits (outputs before the $softmax$ normalization) obtained using fine-tuned BERT fall below the threshold $l = 7$, or for which the predicted label corresponds to the `Other` class, are passed to the second stage of processing using BART. In the second stage with the objective to tackle the zero-shot learning problem we use BART-LARGE-MNLI, pre-trained on the Multi-Genre Natural Language Inference (MNLI) corpus of 433k sentence pairs annotated with textual entailment information (Williams et al., 2018). In this stage, the classification task is reformulated as the natural language inference (NLI) task of determining whether a *hypothesis* is true (entailment) or false (contradiction), given a *premise*. We follow the approach proposed in Yin et al. (2019) and take the text snippet as the *premise* and the descriptive forms of candidate labels as alternative *hypotheses*. The final label is assigned in this stage based on the largest probability of entailment obtained using BART. For each text snippet being processed in this stage the set of candidate labels is defined as consisting of the label predicted in the first stage by the BERT model and all labels of the zero-shot (unseen) classes relevant for the respective subtask.

6.2 Participant Systems

Eight teams registered for the task, whereas four teams submitted their system responses: **ICIP** (Institute of Software Chinese Academy of Sciences), **FKIE-ITF** (Fraunhofer Institute for Communication, Information Processing and Ergonomics), **IBM-MNLP** (IBM Multilingual Natural Language Processing), **UNCC** (University of North Carolina Charlotte). All participants took part in all 3 subtasks, with the exception of FKIE-ITF which took part only in Subtask 1. We provide short overview of these systems.

For Subtask 1 all teams used a fine-tuned ROBERTA as their base classification model. For Subtask 2, most of the teams used a hybrid solution, using a diversity of classifiers, one team did use few shot learning (therefore diverging from the zero shot problem statement). For Subtask 3, where a

the text snippets in the test dataset might have a link to some real-world events the information contained in the snippets may contradict factual information. Consequently, this dataset should not be used as a database of events for the analysis of real-world socio-political developments and conflict events.

[8] An n-gram is considered as a feature only if it appears at least 15 times in the training data.

[9] https://www.csie.ntu.edu.tw/~cjlin/liblinear

zero-shot classifier was mandatory, all participants based their system on a Transformer-based model trained on an NLI task, with some variations.

Despite using the same base approaches, each team focused in its submission on different ways to improve it: ICIP tried different attention mechanisms; FKIE-ITF (Kent and Krumbiegel, 2021) explored different text pre-processing techniques and used sub-sampling; IBM-MNLP (Barker et al., 2021) tried re-ranking different combination of few-shot, zero-shot and regular classifiers; UNCC (Radford, 2021) focused on using a single NLI learning approach for all tasks and used a specific subsampling.

7 Evaluation Results

The results for all submitted system responses for all 3 subtasks in terms of precision, recall and F_1 weighted average scores are provided in Table 1, 2 and 3 respectively, detailed results are given in Appendix B. Each team had the possibility to submit a maximum of 5 configurations per subtask, all of which are reported in the table, and identified by a numerical extension. As an overview of the obtained results, the best performing systems for the three subtasks are 83.9%, 79.7% and 77.1% weighted F_1 scores respectively.

The two teams that reported using undersampling due to lack of sufficient computational resources, are also the ones having the overall lowest score on Subtask 1.

In Table 2, all submissions of team IBM-MNLP are few-shots excepts for their last submission: IBM-MNLP 2.4. Both of their few-shot and zero-shot configurations perform better then systems of any other team for Subtask 2. In Table 3, their first and third submissions are zero shot for the 5 new types, while their two other submissions are zero-shot only for the 2 new types.

For Subtask 3, the best weighted F_1 score for zero-shot classifier restricted to the 5 new classes only are the following: 65.1% for ICIC, 52.9% for IBM-MNLP and 26.2% for UNCC, c.f. Table 7 for details.

System	Prec.	Rec.	F_1
$L2LR_{baseline}$	*0.728*	*0.668*	*0.678*
$BB_{baseline}$	*0.861*	*0.837*	*0.838*
FKIE-ITF 1.1	0.824	0.797	0.799
FKIE-ITF 1.2	0.851	0.829	0.830
FKIE-ITF 1.3	0.828	0.808	0.808
FKIE-ITF 1.4	0.841	0.802	0.812
FKIE-ITF 1.5	0.817	0.793	0.793
IBM-MNLP 1.1	0.851	0.830	0.828
IBM-MNLP 1.2	0.856	0.834	0.835
IBM-MNLP 1.3	**0.861**	**0.838**	**0.839**
ICIP 1.1	0.857	0.826	0.829
ICIP 1.2	0.855	0.829	0.831
ICIP 1.3	0.834	0.789	0.796
ICIP 1.4	0.858	0.828	0.832
ICIP 1.5	0.857	0.825	0.829
UNCC 1.1	0.798	0.739	0.736

Table 1: Overall performance overview Subtask 1: weighted average scores.

System	Sys. type	Prec.	Rec.	F_1
$BB_{baseline}$	*Zero-S.*	*0.811*	*0.787*	*0.788*
IBM-MNLP 2.1	Few-S.	0.824	0.782	0.779
IBM-MNLP 2.2	Few-S.	0.817	**0.797**	**0.797**
IBM-MNLP 2.3	Few-S.	0.824	0.794	0.790
IBM-MNLP 2.4	Zero-S.	0.809	0.786	0.785
ICIP 2.1	Zero-S.	0.798	0.744	0.742
ICIP 2.2	Zero-S.	0.823	0.781	0.776
ICIP 2.3	Zero-S.	0.820	0.775	0.769
ICIP 2.4	Zero-S.	0.827	0.781	0.779
ICIP 2.5	Zero-S.	**0.829**	0.784	0.782
UNCC 2.1	Zero-S.	0.670	0.658	0.635
UNCC 2.2	Zero-S.	0.670	0.658	0.635

Table 2: Overall performance overview Subtask 2: weighted average scores.

System	Sys. type	Prec.	Rec.	F_1
$BB_{baseline}$	*Zero-S.*	*0.803*	*0.745*	*0.753*
IBM-MNLP 3.1	Zero-S.	0.793	0.744	0.746
IBM-MNLP 3.2	Few-S.	0.787	0.755	0.756
IBM-MNLP 3.3	Zero-S.	0.793	0.744	0.746
IBM-MNLP 3.4	Few-S.	0.787	0.755	0.756
ICIP 3.1	Zero-S.	0.790	0.741	0.733
ICIP 3.2	Zero-S.	0.818	0.775	0.765
ICIP 3.3	Zero-S.	0.810	0.768	0.757
ICIP 3.4	Zero-S.	0.818	0.775	0.767
ICIP 3.5	Zero-S.	**0.821**	**0.778**	**0.771**
UNCC 3.1	Zero-S.	0.643	0.625	0.602
UNCC 3.2	Zero-S.	0.644	0.629	0.605

Table 3: Overall performance overview Subtask 3: weighted average scores.

8 Discussion

8.1 Overall Results

The results of all three subtasks provide interesting insights on fine-grained event classification in the context of real-world applications, where practical constraints can lead to a setup with a drift between the data on which the models were trained and for which predictions are generated, and where unseen classes can naturally pose a zero-shot learning problem. Firstly, we conclude that in Subtask 1 the Transformer-based BERT and ROBERTA were observed to lead to virtually the same level of per-

formance in terms of all considered metrics. This observation is interesting, as e.g. on the GLUE benchmark (Wang et al., 2018) ROBERTA is shown to outperform BERT. Secondly, after enhancing the classification task to a generalized zero-shot learning problems in Subtask 2 and 3, the submitted results suggest that the best solutions are, very similar to our baseline $BB_{baseline}$ described in Section 6.1.2, based on the two-stage approach employing a supervised, fine-tuned Transformer-based classifier and another Transformer-based model instance trained on the MNLI data for tackling the zero-shot classification as the sentence-entailment problem. Interestingly, only one team (UNCC) submitted a single-stage model, trained on the entailment-like reformulation of the classification problem. We hypothesize that compared to the single-stage entailment-like setup, the two-stage approaches might more effectively utilize the information provided in the available training data. The significant differences in performance values between these two paradigms in all three subtasks (73.6% vs. 83.9% in Subtask 1, 63.5% vs. 79.7% in Subtask 2 and 60.5% vs. 77.1% in Subtask 3) might seem to confirm this hypothesis. However, it should be stressed that the submissions following the single-stage entailment-like setup were made with a disclaimer on computational limitations.

In order to provide some flavour of most typical errors and difficulties of automatically labelling event snippets using ACLED taxonomy Figure 5 provides the confusion matrix, normalized over the true conditions (rows), for the $BB_{baseline}$ approach applied to solve Subtask 1.

The most significant type of error is the misclassification of `Force Against Protest` as `Protest With Interventions` (39%), `Property Destruction` as `Mob Violence` (29%) and as `Violent Demonstration` (24%) and `Artillery/Missile Attack` as `Armed Clash` (19%). Given a fine line between these types, the above error rates are not surprising. More generally, one can observe that distinguishing between the sub-types belonging to the same main type (see the ACLED taxonomy in Appendix A), is typically more challenging. Also, it is not surprising that the `Other` class has also a relatively low recall of 50%.

As regards models robustness, in Piskorski et al. (2020), the reported F_1 score of the BERT-based ACLED-trained classifier when evaluated on ACLED data yield about 94.4%. In Subtask 1, using similar Transformer-based classifier lead to a maximal score of 83.9%: we observe approx. 10 percentage point drop in performance. It is important to mention herethat the former model used 80% of the ACLED data for training, whereas the latter used the entire ACLED dataset reported in Piskorski et al. (2020).

Class-wise performance comparison of both classifiers are reported in Table 8.

Such a performance drop can be explained in part by the fact that text snippets in the ACLED follow a pattern that is different than news-like reporting, and as such the classifier struggles to generalize to the real-world news-like reporting style, despite the standard regularization techniques.

The performance drop is not equally distributed over the classes. Actually, when applying to news data, roughly half of the classes have better scores, and half have worse scores.

One possible reason for this performance drop seems to be the three most populated classes in the ACLED dataset (`Armed Clash`, `Attack`, `Artillery/Missile Attack`) which on average lost 18 points when compared with the results of the baseline model $BB_{baseline}$.

8.2 ACLED taxonomy

Having used ACLED taxonomy in the context of this Shared Task have resulted in some reflections, both in terms of experience of using it to annotate text snippets reporting events and its practicality for a real-world application for automatically labelling news-like texts.

As regards the annotation of news-like text snippets great care has been taken to follow strictly the ACLED Codebook. This turned to be a harder task than initially expected, in part due to shortcomings of the Codebook, and, in part due to the nature of how events are reported in the news.

News texts often assume a known global context and do not provide enough information to allow to clearly assign an ACLED event subtype. This is due the high specificity of ACLED subtypes that make it hard, for instance, to classify a text describing a demonstration, if it can not be understood from the text whether the event was violent, and if this was the case, which side started the violence, i.e., the demonstrators or the authority tasked to thwart the demonstration. All such information

is needed to select the proper ACLED event class. Having said this, it is worthwhile to mention here that sometimes the nuances between the definitions of the event types are very small and we also found certain inconsistencies between the entries in the ACLED event database itself, e.g. for the `Protest with Intervention` and `Excessive force against the protesters` categories the corresponding text descriptions did not differ much, and at times using certain instrument to intervene was mentioned in the case of both events. Clearly, when encoding an event using ACLED taxonomy based on HUMINT and without considering any source text the human knows the event type upfront, and hence, the resulting text describing the event might not fully reflect/mirror the specific of the particular event type. This poses a certain limitation to what extent the textual descriptions of events in ACLED can be useful for training models to be applied on news-like data, but to have a better picture a full-fledged study of the aforementioned inconsistencies should be carried out, which is out of scope of the Shared Task.

The high specificity of the ACLED taxonomy is also at times problematic as it was not designed for multi-label classification tasks. As such, an attack on a civilian with a suicide bomber can not be classified as suicide bombing event according to ACLED taxonomy if any other interaction took place and is reported, for instance, if the text mentions also assailants attack with firearms first before detonating the bomb or if the police tries to stop them. In such a case the `Armed Clash` event type has to be used. On the other hand, intuitively, it would make sense that the text is tagged with at least two labels: `Attack` (attack on civilian) and `Suicide bombing`, or potentially also a tag that represents an authority intervention. ACLED taxonomy imposes a complex and incomplete set of priorities in order to enforce an event to be labelled using a mono-dimensional classification.

Another issue encountered when using this taxonomy is related to the fact that definitions of some event classes are unclear and not intuitive per-se. For instance, the class `Arrest` which accounts for either mass arrests or arrest of VIPs, but not for arrests of "one or few" people, which fall under a different type. Furthermore, problematic is also the fact that some classes are actually determined not only by what actually happened but also by who was the main actor involved. For instance, the class `Government retakes territory` and `Non-state actor captures territory` are almost indistinguishable when the named entities are shuffled. What is more, the taxonomy does not specify how to handle certain cases, e.g., when a non-government actor is acting on behalf of or is supported by the government in regaining/overtaking territory.

Lastly, disregarding the strictly mono-dimensional nature of ACLED taxonomy, most news text snippets (even single sentences) report on more than one event, and determining which one is the salient one is not always straightforward even to human annotators. One of our observations is that for labelling news reporting on events a multi-class labelling approach would be more intuitive and logical.

9 Conclusions

This paper reported on the outcome of the Shared Task on Fine-grained Event Classification in News-like Text Snippets that has been organized as part of the Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021), co-located with the ACL-IJCNLP 2021 Conference.

8 teams registered to participate in the task, while 4 of them submitted system responses for 3 subtasks, two of which were generalized zero-shot learning tasks. Given the specific set up of the shared task, i.e., the training data being somewhat different from the test data and inclusion of 5 unseen classes the top results obtained can be considered good, however, there is definitely place for improvement. Furthermore, we intend to carry out comparative error analysis across systems, which might reveal some additional insights into the complexity of the task.

Further documentation and related material on the reported Shared Task can be found at `https://github.com/emerging-welfare/case-2021-shared-task/tree/main/task2`, whereas the test dataset alone is also available at: `http://piskorski.waw.pl/resources/case2021/data.zip` for research purposes.

We believe that the reported results, findings and the annotated test dataset will contribute to stimulating further research on fine-grained event classification.

References

Douglas E. Appelt. 1999. Introduction to information extraction. *AI Commun.*, 12(3):161–172.

Martin Atkinson, Jakub Piskorski, Roman Yangarber, and Erik van der Goot. 2011. Multilingual Real-Time Event Extraction for Border Security Intelligence Gathering. In *Open Source Intelligence and Counter-terrorism*. Springer, LNCS, Vol. 2.

Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. 2021. IBM MNLP IE at CASE 2021 Task 2: NLI Reranking for Zero-Shot Text Classification. In *Proceedings of the 4^{th} Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics (ACL).

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer.

Nancy A. Chinchor. 1998. Overview of MUC-7. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

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.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using Cross-Entity Inference to Improve Event Extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Jakub Piskorski, Reyyan Yeniterzi, and Erdem Yörük. 2021. Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021): Workshop and Shared Task Report. In *Proceedings of the 4^{th} Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics (ACL).

Ali Hürriyetoglu, Erdem Yörük, Deniz Yuret, Çagri Yoltar, Burak Gürel, Firat Durusan, 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 - 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9-12, 2019, Proceedings*, volume 11696 of *Lecture Notes in Computer Science*, pages 425–432. Springer.

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, pages 1–28.

Gilles Jacobs and Veronique Hoste. 2020. Extracting fine-grained economic events from business news. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 235–245, Barcelona, Spain (Online). COLING.

Samantha Kent and Theresa Krumbiegel. 2021. CASE 2021 Task 2: Socio-political Fine-grained Event Classification using Fine-tuned RoBERTa Document Embeddings. In *Proceedings of the 4^{th} Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics (ACL).

Gary King and Will Lowe. 2003. An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders. *International Organization*, 57:617–642.

LDC. 2008. Annotation Tasks and Specification. ONLINE: https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2.

Els Lefever and Véronique Hoste. 2016. A Classification-based Approach to Economic Event Detection in Dutch News Text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 330–335, Portorož, Slovenia. European Language Resources Association (ELRA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Shasha Liao and Ralph Grishman. 2010. Using Document Level Cross-Event Inference to Improve Event Extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797, Uppsala, Sweden. Association for Computational Linguistics.

Thien Huu Nguyen, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi. 2016. Joint Learning of Local and Global Features for Entity Linking via Neural Networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2310–2320, Osaka, Japan. The COLING 2016 Organizing Committee.

Thien Huu Nguyen and Ralph Grishman. 2015. Event Detection and Domain Adaptation with Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.

Timothy 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, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 3750–3759.

J. Pastor-Galindo, P. Nespoli, F. Gómez Mármol, and G. Martínez Pérez. 2020. The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. *IEEE Access*, 8:10282–10304.

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 Roman Yangarber. 2013. Information extraction: Past, present and future. In Thierry Poibeau, Horacio Saggion, Jakub Piskorski, and Roman Yangarber, editors, *Multi-source, Multilingual Information Extraction and Summarization*, Theory and Applications of Natural Language Processing, pages 23–49. Springer Berlin Heidelberg.

Benjamin Radford. 2021. CASE 2021 Task 2: Zero-Shot Classification of Fine-Grained Sociopolitical Events with Transformer Models. In *Proceedings of the 4^{th} Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics (ACL).

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.

Sovan Kumar Sahoo, Saumajit Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A Platform for Event Extraction in Hindi. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2241–2250, Marseille, France. European Language Resources Association.

Mark Stickel and Mabry Tyson. 1997. Fastus: A cascaded finite-state transducer for extracting information from natural-language text. In *Finite-State Language Processing*, pages 383–406. MIT Press.

Beth M. Sundheim. 1991. Overview of the Third Message Understanding Evaluation and Conference. In *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and Fine-Grained Classification of Cyberbullying Events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Michael D Ward, Andreas Beger, Josh Cutler, Matt Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing GDELT and ICEWS event data. *Analysis*, 21:267–297.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Roman Yangarber, Peter Von Etter, and Ralf Steinberger. 2008. Content Collection and Analysis in the Domain of Epidemiology. In *Proceedings of DrMED 2008: International Workshop on Describing Medical Web Resources at MIE 2008: the 21^{st} International Congress of the European Federation for Medical Informatics 2008*, Goeteborg, Sweden.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Appendices

A Event Types

The ACLED event taxonomy comprises of six main event types which are further subdivided into 25 sub-event types as follows:

```
BATTLES
 - Armed clash
 - Government regains territory
 - Non-state actor overtakes territory
EXPLOSION AND REMOTE VIOLENCE
 - Chemical weapon
 - Air/drone strike
 - Suicide bomb
 - Shelling/artillery/missile attack
 - Remote explosive/landmine/IED
 - Grenade
VIOLENCE AGAINST CIVILIANS
 - Sexual violence
 - Attack
 - Abduction/forced disappearance
PROTESTS
 - Peaceful protest
 - Protest with intervention
 - Excessive force against protesters
RIOTS
 - Violent demonstration
 - Mob violence
STRATEGIC DEVELOPMENTS
 - Agreement
 - Arrests
 - Change to group/activity
 - Disrupted weapons use
 - Headquarters or base established
 - Looting/property destruction
 - Non-violent transfer of territory
 - Other
```

For further details on ACLED event taxonomy please refer to the ACLED codebook.

We provide here the description of the 5 new types used in the Shared Task. The first three new types cover contextually important security- and safety-related events and developments that are not related to political violence and not considered to contribute to political dynamics within and across multiple states. The last two new types cover events directly related to security situation, and as such fall under the `Strategic Development` main event type of ACLED, however, they are mainly related to announcements instead of concrete deeds. The 5 additional new types are as follows:

- `Organized crime`: This event type covers incidents related to activities of criminal groups, excluding conflict between such groups: smuggling, human trafficking, counterfeit products, property crime, cyber crime, assassination (for criminal purposes), corruption, etc.
- `Natural Disaster`: This event type covers any kind of natural disasters and hazards where there is a direct or potential harm, including: earthquakes, tsunami, floods, storms, fires, volcano eruptions, landslides, avalanches, infectious disease outbreaks, pandemics, climate related, etc.
- `Man-made Disaster`: This event type covers any kind of disasters caused by humans where there is a direct or potential harm, such as: industrial accidents, traffic incidents, infrastructure failure, foodchain contamination, etc.
- `Diplomatic Event`: This event type covers any kind of diplomatic action or announcement that have a potential impact on the security situation or denoting the attitude of a country towards a conflict. As such this type covers diplomatic measures declaration (e.g. sanctions or closure of embassies), threats, call for actions, praises and condemnations.
- `Attribution of Responsibility`: This event type covers announcements related to the responsibility of attacks and hostile operations. In particular, this event type covers group claiming their own responsibility, accusation of responsibility and denial of responsibility.

B Complete Evaluation Tables

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
$L2LR_{baseline}$	0.668	0.668	0.668	0.702	0.647	0.650	0.728	0.668	0.678
$BB_{baseline}$	0.837	0.837	0.837	0.837	0.804	0.807	0.861	0.837	0.838
FKIE-ITF 1.1	0.797	0.797	0.797	0.790	0.778	0.770	0.824	0.797	0.799
FKIE-ITF 1.2	0.829	0.829	0.829	0.807	0.808	0.794	0.851	0.829	0.830
FKIE-ITF 1.3	0.808	0.808	0.808	0.787	0.779	0.768	0.828	0.808	0.808
FKIE-ITF 1.4	0.802	0.802	0.802	0.788	0.789	0.774	0.841	0.802	0.812
FKIE-ITF 1.5	0.793	0.793	0.793	0.780	0.780	0.766	0.817	0.793	0.793
IBM-MNLP 1.1	0.830	0.830	0.830	0.828	0.787	0.792	0.851	0.830	0.828
IBM-MNLP 1.2	0.834	0.834	0.834	0.849	0.793	0.810	0.856	0.834	0.835
IBM-MNLP 1.3	0.838	0.838	0.838	0.854	0.800	0.814	0.861	0.838	0.839
ICIP 1.1	0.826	0.826	0.826	0.827	0.800	0.796	0.857	0.826	0.829
ICIP 1.2	0.829	0.829	0.829	0.824	0.802	0.798	0.855	0.829	0.831
ICIP 1.3	0.789	0.789	0.789	0.805	0.766	0.765	0.834	0.789	0.796
ICIP 1.4	0.828	0.828	0.828	0.827	0.803	0.799	0.858	0.828	0.832
ICIP 1.5	0.825	0.825	0.825	0.825	0.799	0.795	0.857	0.825	0.829
UNCC 1.1	0.739	0.739	0.739	0.770	0.697	0.698	0.798	0.739	0.736

Table 4: Overall performance overview for Subtask 1.

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
$L2LR_{baseline}$	0.668	0.585	0.624	0.627	0.578	0.581	0.638	0.585	0.593
$BB_{baseline}$	0.79	0.79	0.79	0.797	0.763	0.767	0.811	0.787	0.788
IBM-MNLP 2.1	0.782	0.782	0.782	0.809	0.753	0.752	0.824	0.782	0.779
IBM-MNLP 2.2	0.797	0.797	0.797	0.807	0.761	0.773	0.817	0.797	0.797
IBM-MNLP 2.3	0.794	0.794	0.794	0.811	0.759	0.764	0.824	0.794	0.790
IBM-MNLP 2.4	0.786	0.786	0.786	0.790	0.750	0.758	0.809	0.786	0.785
ICIP 2.1	0.744	0.744	0.744	0.767	0.733	0.718	0.798	0.744	0.742
ICIP 2.2	0.781	0.781	0.781	0.788	0.767	0.750	0.823	0.781	0.776
ICIP 2.3	0.775	0.775	0.775	0.786	0.760	0.743	0.820	0.775	0.769
ICIP 2.4	0.781	0.781	0.781	0.793	0.767	0.752	0.827	0.781	0.779
ICIP 2.5	0.784	0.784	0.784	0.795	0.769	0.755	0.829	0.784	0.782
UNCC 2.1	0.658	0.658	0.658	0.648	0.632	0.613	0.670	0.658	0.635
UNCC 2.2	0.658	0.658	0.658	0.648	0.632	0.613	0.670	0.658	0.635

Table 5: Overall performance overview for Subtask 2.

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
$L2LR_{baseline}$	0.668	0.544	0.600	0.585	0.539	0.542	0.593	0.544	0.551
$BB_{baseline}$	0.745	0.745	0.745	0.771	0.709	0.720	0.803	0.745	0.753
IBM-MNLP 3.1	0.744	0.744	0.744	0.769	0.714	0.720	0.793	0.744	0.746
IBM-MNLP 3.2	0.755	0.755	0.755	0.764	0.723	0.728	0.787	0.755	0.756
IBM-MNLP 3.3	0.744	0.744	0.744	0.769	0.714	0.720	0.793	0.744	0.746
IBM-MNLP 3.4	0.755	0.755	0.755	0.764	0.723	0.728	0.787	0.755	0.756
ICIP 3.1	0.741	0.741	0.741	0.762	0.725	0.708	0.790	0.741	0.733
ICIP 3.2	0.775	0.775	0.775	0.788	0.757	0.738	0.818	0.775	0.765
ICIP 3.3	0.768	0.768	0.768	0.779	0.749	0.729	0.810	0.768	0.757
ICIP 3.4	0.775	0.775	0.775	0.788	0.757	0.741	0.818	0.775	0.767
ICIP 3.5	0.778	0.778	0.778	0.791	0.760	0.744	0.821	0.778	0.771
UNCC 3.1	0.625	0.625	0.625	0.620	0.599	0.580	0.643	0.625	0.602
UNCC 3.2	0.629	0.629	0.629	0.621	0.602	0.582	0.644	0.629	0.605

Table 6: Overall performance overview for Subtask 3.

System	Sys. type	Prec.	Rec.	F_1
IBM-MNLP 3.1	Zero-S.	0.915	0.389	0.529
IBM-MNLP 3.2	Few-S.	0.896	0.553	**0.668**
IBM-MNLP 3.3	Zero-S.	0.915	0.389	0.529
IBM-MNLP 3.4	Few-S.	0.896	0.553	**0.668**
ICIP 3.1	Zero-S.	0.917	0.532	0.599
ICIP 3.2	Zero-S.	**0.941**	0.547	0.621
ICIP 3.3	Zero-S.	0.916	0.521	0.589
ICIP 3.4	Zero-S.	0.928	0.563	0.635
ICIP 3.5	Zero-S.	0.929	**0.579**	0.651
UNCC 3.1	Zero-S.	0.562	0.179	0.244
UNCC 3.2	Zero-S.	0.571	0.200	0.262

Table 7: Performance overview Subtask 3: weighted average scores on the 5 unknown types.

Figure 5: Confusion matrix for $BB_{baseline}$ applied to Subtask 1.

ACLED class	F_1 on ACLED	F_1 on News-like data	ΔF_1
Abduction/forced disappearance	0.903	0.865	-0.04
Agreement	0.831	0.842	0.01
Air/drone strike	0.987	0.971	-0.02
Armed clash	0.956	0.736	-0.22
Arrests	0.89	0.938	0.05
Attack	0.915	0.727	-0.19
Change to group/activity	0.838	0.844	0.01
Chemical weapon	0.829	0.943	0.11
Disrupted weapons use	0.891	0.938	0.05
Excessive force against protesters	0.692	0.650	-0.04
Government regains territory	0.839	0.904	0.07
Grenade	0.893	0.909	0.02
Headquarters or base established	0.758	0.905	0.15
Looting/property destruction	0.808	0.370	-0.44
Mob violence	0.851	0.595	-0.26
Non-state actor overtakes territory	0.784	0.776	-0.01
Non-violent transfer of territory	0.73	0.781	0.05
Other	0.64	0.400	-0.24
Peaceful protest	0.984	0.902	-0.08
Protest with intervention	0.813	0.755	-0.06
Remote explosive/landmine/IED	0.97	0.959	-0.01
Sexual violence	0.93	0.955	0.02
Shelling/artillery/missile attack	0.978	0.841	-0.14
Suicide bomb	0.933	0.907	-0.03
Violent demonstration	0.862	0.772	-0.09

Table 8: Comparison of $BB_{baseline}$ performances when applied on ACLED data vs. news-like data: weighted average scores