New Benchmark Corpus and Models for Fine-grained Event Classification: To BERT or not to BERT?

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

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

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

Abstract

We introduce a new set of benchmark datasets derived from ACLED data for fine-grained event classification and compare the performance of various state-of-the-art machine learning models on these datasets, including SVM based on TF-IDF character *n*-grams and neural context-free embeddings (GLOVE and FASTTEXT) as well as deep learning-based BERT with its contextual embeddings. The best results in terms of micro (94.3-94.9%) and macro F_1 (86.0-88.9%) were obtained using BERT transformer, with simpler TF-IDF character *n*-gram based SVM being an interesting alternative. Further, we discuss the pros and cons of the considered benchmark models in terms of their robustness and the dependence of the classification performance on the size of training data.

1 Introduction

Since an ever-growing amount of information on events of any type is transmitted via web in the form of free texts (e.g. online news) one has witnessed in the last decades an emergence of research on development of methods and tools for automated detection and extraction of structured information on events from textual sources (King and Lowe, 2003; Yangarber et al., 2008; Atkinson et al., 2011; Piskorski et al., 2011; Leetaru and Schrodt, 2013; Ward et al., 2013; Pastor-Galindo et al., 2020). One particular step in the event extraction process is event classification, i.e., assigning to a text snippet including event trigger an event type using a domain specific taxonomy, which is the main focus of this paper.

While vast amount of tasks and challenges on automated event extraction, including event classification, has been organised over the years, relatively little work has been reported on approaches for fine-grained event classification. Furthermore, the existing freely available datasets used for training and evaluation purposes are rather of tiny size, ranging usually up to 5-10K events. Due to the emergence of deep learning-based approaches for the entire range of NLP tasks, there is a particular need to have larger event classification corpora in order to gain better insights into the performance of such methods and their comparison with shallow learning approaches, e.g., in terms of training data sizes required to obtain 'acceptable' performance, types of embeddings and model robustness vis-a-vis different data characteristics. The rise of deep learning-based approaches allowing for model pre-training in an unsupervised manner using only plain text and then utilizing transfer learning (via re-using the pre-trained model and only fine-tuning it in a supervised manner), poses additional questions with regard to the required data sizes

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

Proceedings of the 28th International Conference on Computational Linguistics, pages 6663–6678
Barcelona, Spain (Online), December 8-13, 2020

and the robustness of transfer learning against various data characteristics. One of the most groundbreaking moments in the field of NLP was the release of BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019) in October 2018, as the first deeply bidirectional, unsupervised language representation model, leading to a significant uplift in multiple performance benchmarks with limited task-specific fine-tuning. Hence, among the benchmark machine learning models included in the work reported in this article we study in particular the performance of BERT on the fine-grained event classification tasks derived from ACLED.

The main contributions of the work reported in this paper can be summarized as follows:

- we introduce a set of relatively large benchmark datasets derived from ACLED[1] data - a manually curated event repository - each consisting of circa 600K short event descriptions for the evaluation of fine-grained event classification, which covers 25 event types (all types are related to political violence events, crisis situations, protest and unrest events),

- we compare the performance of various state-of-the-art benchmark models on these datasets, spanning SVM and NN-based classifiers that exploit for its feature representations: (a) TF-IDF character *n*-grams, (b) off-the-shelf pre-trained non-contextual GLOVE and FASTTEXT embeddings, (c) contextual pre-trained BERT embeddings, and (d) contextual fine-tuned BERT embeddings. Further, we discuss the pros and cons of using these models in terms of their robustness and practical application in a real-world set-up.

The exploitation of the ACLED data for evaluation of fine-grained event classification models has, in particular, the following two major advantages: (a) the ACLED event descriptions resemble very much texts that can be found in online news reporting on events, and (b) ACLED data is to a certain extent noisy in terms of grammatical correctness, which provides an excellent material to test models robustness vis-a-vis lower quality data.

To our best knowledge, no similar corpora in terms of size for the task at hand exist, and given the specific nature of the dataset (i.e. text snippets resembling news reporting), we believe that measuring the event classification performance of a given method on these dataset might constitute a good approximation of the to-be-expected performance when applying the same method on real news articles.

The rest of the paper is structured as follows. First, an overview of related work is provided in Section 2. Subsequently, Section 3 describes the corpora derived from ACLED data. Next, Section 4 introduces the benchmark models for the event classification task, whereas Section 5 presents the results of the performance of these models on the ACLED corpora and basic error analysis. The main findings and practical implications thereof are summarized in Section 6. Finally, Section 7 gives conclusions and an outlook on future work.

2 Related Work

The early research on event detection and classification in textual documents was driven by the Message Understanding Contests (Sundheim, 1991; Chinchor, 1998) and the Automatic Content Extraction (ACE) Challenges (Doddington et al., 2004; LDC, 2008). Many approaches to event detection and classification have been reported and evaluated on the event corpora (ca. 6000 event mentions in ca. 500 documents) developed in the context of the aforementioned ACE Challenges, which range from shallow (Liao and Grishman, 2010; Hong et al., 2011) to deep machine learning approaches (Nguyen and Grishman, 2015; Nguyen et al., 2016). The more recently introduced Multi-lingual Event Detection and Co-reference challenge in the context of the Text Analysis Conference (TAC) in 2016[2] and 2017[3], included an Event Nugget Detection subtask, which focused on detection and fine-grained classification of intra-document event mentions (9 types and 38 subtypes), covering events from various domains (e.g., finances and jurisdiction). The evaluation datasets used in the context of TAC are rather tiny though (<10K events).

[1] https://www.acleddata.com
[2] https://tac.nist.gov//2016/KBP/Event/index.html
[3] https://tac.nist.gov/2017/KBP/Event/index.html

In the last decade other efforts on more fine-grained event classification that cover various domains were reported too. For instance, (Lefever and Hoste, 2016) compared SVM-based models against word-vector-based LSTMs for classification of 10 types of company-specific economic events from online news, 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 (7 types) in news articles. While most of the work in this area focused on English language and processing news texts in particular, some efforts on event classification for non-English language and other domains were reported too. A benchmark corpus for fine-grained classification of man-made and natural disasters (28 types) for Hindi, accompanied with evaluation of deep learning baseline models for this task, has been presented in (Sahoo et al., 2020). Furthermore, an example of fine-grained classification of cyber-bullying events (7 classes) in social media posts was reported in (Van Hee et al., 2015). This paper reports on the creation of benchmark corpora for fine-grained event classification of political violence, conflict situation and protest events from short text snippets, where the main difference vis-a-vis the benchmark corpora reported elsewhere is the size of the corpora, significantly bigger (ca.. 600K events) versus other known event classification corpora (usually of the size in the range of 2-10K). The work reported in this paper builds on the preliminary study of the ACLED data for event classification presented in (Piskorski and Jacquet, 2020) and extends it in various dimensions.

3 Event Classification corpora derived from ACLED data

The Event Classification corpus was derived from event data collected in the context of the Armed Conflict Location & Event Data Project (ACLED)[4]. ACLED (Raleigh et al., 2010) gathers human-moderated records on most important facts about political violence and protest events across various continents with a specific focus on Africa, Asia, the Middle East, and Southeastern and Eastern Europe. The collected event records contain information on the date of the event, location, the key actors involved, type of violence and number and description of fatalities. For the sake of creating corpora for the Event Classification task we have extracted from circa 615K manually-curated event records available on the ACLED web page[5] three elements, including: (a) event snippets, being free-text descriptions of the events, which mention basic information on all key information on the event, (b) event type, and (c) and event subtype. The ACLED event type ontology has 6 main even types (battles, explosion and remote violence, violence against civilians, protests, riots, strategic developments), which are further subdivided into 25 fine-grained subtypes. The detailed definitions of the various event types and subtypes in ACLED are reported in the so called ACLED Codebook (ACLED, 2019). Some examples of event descriptions for violent demonstration, peaceful protest, and armed clash events resp. are given below.

1. *Several people were injured when demonstrations erupted at Sangam following the death of a local militant in a gun fight with government forces the day before. The forces resorted to lathicharge followed by bursting of teargas shells to disperse the stone-pelting demonstrators.*

2. *Striking members of the Punjab State Ministerial Staff Union staged a protest in Bathinda on Friday against the state government's alleged 'anti-employee policies'.*

3. *On 12-March-2013, the Myanmar army fired machine guns at a KIA post in Mu Bum [could not find; geocode for Momauk where the Myanmar army's LIB 437 is based]. No fatalities noted.*

While the texts in the first two examples resemble texts that could as well appear in news articles, the third example contains some comments in brackets provided by the human experts.

From the raw data extracted from ACLED event records three event corpora were created, each being result of cleaning and normalising the original free-text event descriptions. The main drive behind creating three corpora was to move from textual data that contain some "noise" and some not fully grammatically correct constructions to a corpus containing grammatically correct sentences and constructions which are very close to texts appearing in the news reporting on events.

[4] https://www.acleddata.com

[5] https://www.acleddata.com/curated-data-files/

ACLED-I was created through carrying out most basic cleaning of the texts, including: (a) removing from the event descriptions quotation and similar non-content relevant characters, (b) removing too obvious markers that would artificially hint a classifier to guess the correct event (sub)type, e.g., initial phrases like "*Attack:*" corresponding directly to the definition of the event type, and (c) filtering out event records, whose event descriptions consist of less than 20 characters, which were deemed as non informative.

ACLED-II was created via applying the following treatments on ACLED-I: (a) removing outliers, i.e., events, whose description is longer than 650 characters, (b) removing from event descriptions circa 100 unique phrases (provided that they appear at the beginning of the event description and are followed by a colon) that might indicate the event type[6] that are, however, not identical with the event type definitions (e.g. *Detonation:*), (c) removing references to urls, (d) removing comments in brackets introduced by human experts in texts, e.g., *[size=thousands]*, *[codes as 10 ....]*, (e) removing non-Latin based characters, (f) correcting errors related to missing whitespaces at the end of sentences.

ACLED-III resulted from further cleaning and normalisation of ACLED-II, which included: (a) removing events whose description is shorter than 60 characters, (b) removing additional non-sentence-like structures corresponding to comments introduced by human experts encoding the events (i.e, constructions in brackets like in ACLED-II d) above, but significantly longer), (c) normalisation of various non-alphanumeric symbols, (d) removing numeric encoding of locations (coordinates), and (f) removing all event descriptions that contain at least one sentence, which could not be parsed by Stanford PCFG Parser (Klein and Manning, 2003) and resulting in a tree with a root labelled as `"S"`, `"FRAG"` or `"NP"`. In this context we made an assumption that parse trees with roots labelled with tags other than the ones mentioned before constitute potential indicators of non grammatically correct sentences/utterances/nominal phrases. For instance, the parse tree for the event description '*Reports that the CSNPD attacked a truck near Gore killing two people and wounding six*' was labelled with `X`, and thus eliminated (subject missing).

For testing robustness of the benchmark models in the event classification task, an additional version of ACLED-III was generated, in which two main type of modifications were carried out on the ACLED-III corpus: (a) all day and month names were replaced with randomly selected days and months, and (b) each occurrence of a toponym referring to a populated place was replaced with randomly chosen toponym selected from a GEONAMES gazetteer [7] of circa 200K populated cities, whose population is at least 500. The main drive behind these modifications was to simulate data drift that can be expected in the domain of the data at hand. This alternate version of ACLED-III will be referred to with **ACLED-III-Δ**.

Corpus	Number of events	Average event description length (in characters)	Average number of sentences/phrases	Number of unique words	Fraction of alphabetic chars
ACLED-I	611678	188.8	1.70	216297	80.47%
ACLED-II	610107	184.9	1.69	214249	80.53%
ACLED-III	588940	186.1	1.48	211561	80.56%
ACLED-III-Δ	588940	193.3	1.48	323880	80.87%

Table 1: Basic ACLED datasets statistics.

The basic statistics for all three ACLED datasets are provided in Table 1. The event subtype distribution diagram for all three ACLED corpora is presented in Figure 1. From the diagram one can observe that there are 8 event subtypes for which more than 20K instances exist. On the other hand, there are two event subtypes for which there are only few hundred instances (`Chemical weapon`, `Headquarters or base established`). The cleaning of the data did not result in any significant changes in the event subtype distribution for ACLED-II and ACLED-III resp. (see Figure 1). The full list of event types and their corresponding subtypes, accompanied by more-detailed statistics is provided in Table 5 in Appendix A.

[6] Not removed previously while creating ACLED-I

[7] `http:\www.geonames.org`

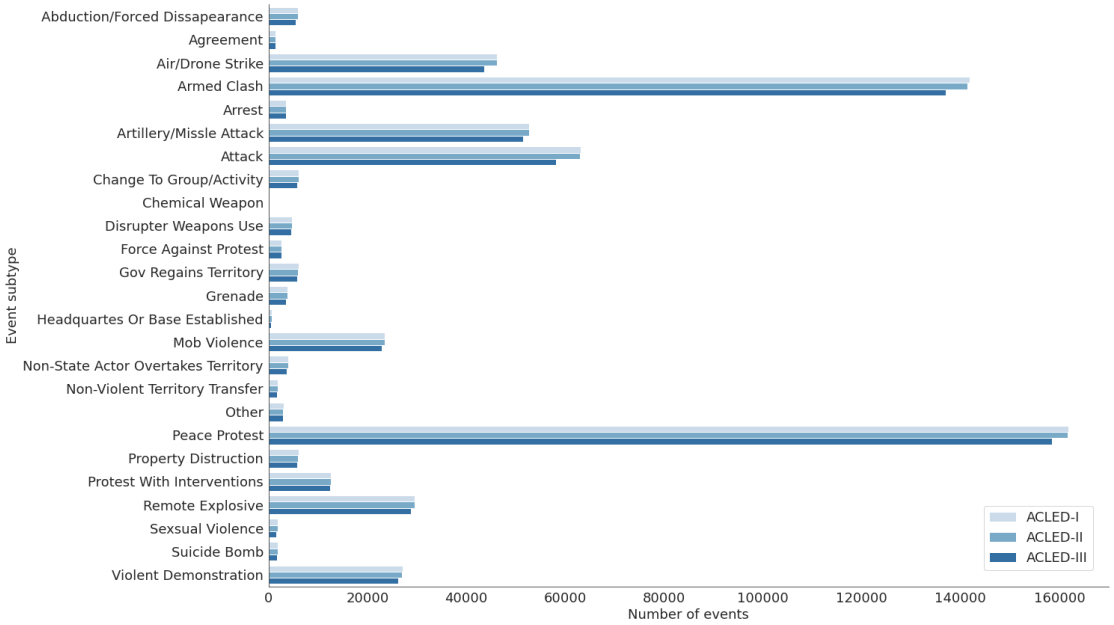

Figure 1: Event subtype distribution for ACLED-I, ACLED-II and ACLED-III.

The distribution of the length of event descriptions for all three ACLED datasets is shown in Figure 2. For the vast majority the length is between 30 and 400 characters, which corresponds to the length of a title and 1-2 leading sentences in a news article reporting on an event. ACLED-I corpus contains all the outliers, i.e., events with description of more than 1000 characters.

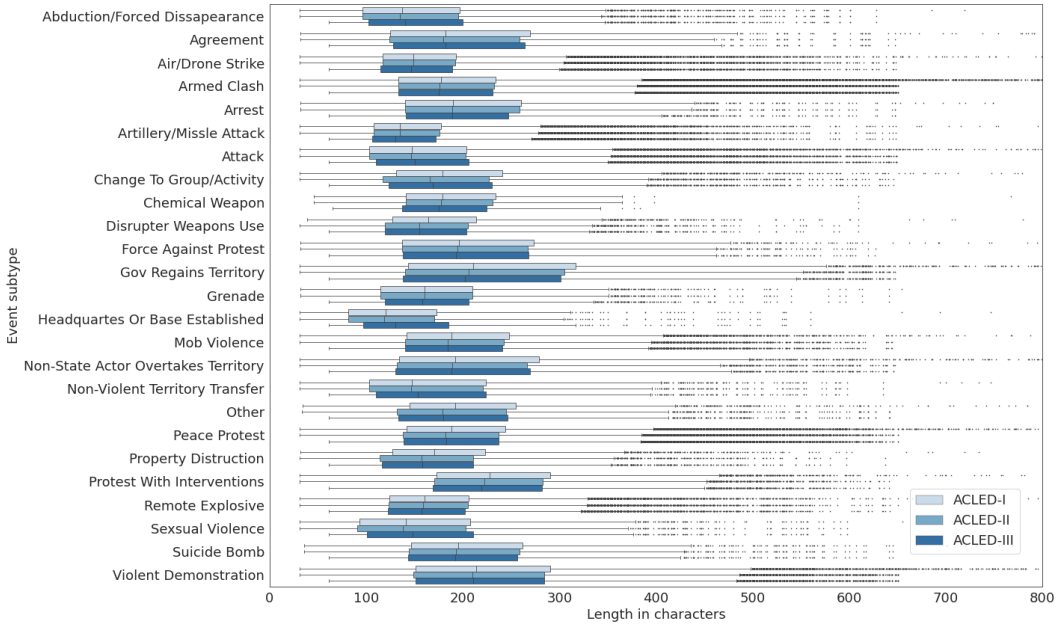

Figure 2: Event description length distribution for ACLED-I, ACLED-II and ACLED-III datasets.

4 Benchmark models

4.1 SVM with TF-IDF char *n*-grams

For the first benchmark model we follow a bag-of-words (BoW) model for extracting TF-IDF features from the character *n*-grams contained within each event description and train a linear SVM model as the classifier. We use an *n*-gram range between 3 and 5-grams (this turned to be the best setting based

on empirical observations). We exclude the *n*-grams occurring in less than 5 event descriptions. We observed during our experiments that these parameters could be slightly modified without important impact on the classification results. The vectorisation is implemented with L2 normalisation, in order to normalise for the number of expressions in each class, and sublinear TF calculations (which log-scales the TF counts).

The dimensionality of the TF-IDF vectors varies depending on the training set size, and each event description is represented by a large sparse vector instead of the short full vector used in the word embedding representation. For the ACLED-III corpus, the TF-IDF vectors vary from 41 054, when using 1% of the training set, to 364 023 when using the full training set.

Regarding non-linear kernels, due to the fact that with the given size of the data the standard non-linear SVMs do not scale well, we have run some initial experiments following the common approach of using kernel map approximations and applying linear SVM on the top of it. Using Nystroem method (Williams and Seeger, 2001) for approximating RBF kernel as well as using Monte-Carlo sampling from the Fourier transformation of the RBF (Rahimi and Recht, 2008) and chi-squared kernels (Vedaldi and Zisserman, 2012) resulted in either worse or similar performance than plain linear SVM. Although only several alternatives for non-linear kernels have been examined and all are subject to sampling errors inherent in the applied approximations, we hypothesise that these results are an indication that with the underlying BoW feature space the problem is either linearly separable or close to linearly separable.

4.2 SVM with non-contextual word embeddings

In the second benchmark model we explored a SVM trained on non-contextual word embeddings. A word embedding is a function $Words \rightarrow \mathbb{R}^d$ that maps words to real-valued vectors of a fixed dimension (Bengio et al., 2003). Recently, various studies reported that word embeddings perform surprisingly well for text classification tasks (Reimers and Gurevych, 2019), in particular in the context of machine learning models that rely on vector representation as input to enjoy richer representations of text input while alleviating high-dimensionality issues. We experimented with two popular non-contextual word embeddings, namely, GLOVE and FASTTEXT embeddings.

GLOVE (Pennington et al., 2014) word embeddings are obtained through exploitation of aggregated global word-word co-occurrence statistics from a large corpus. We used the pre-trained GLOVE 300-dimensional vectors trained on WIKIPEDIA and the English Gigaword corpus[8]. For computing GLOVE vector for an event description the single GLOVE embeddings of all words contained in the event description were averaged (unknown words were discarded in this process).

FASTTEXT embeddings (Mikolov et al., 2018) are based on a model, in which each word is represented as a bag of character *n*-grams, and the vector representing the word is constructed as the sum of the vectors for the character *n*-grams it consists of. We exploited the pre-trained 300-dimensional FASTTEXT vectors, trained on Common Crawl[9] and Wikipedia (Grave et al., 2018) using CBOW with position-weights with character *n*-grams of length 5, and a window of size 5.

4.3 SVM with contextual word embeddings

Our third benchmark model is SVM trained on contextual word embeddings. In particular, we used in our experiments the embeddings based on BERT model, designed to pre-train deep bidirectional representations from unlabeled text data. Such pre-trained BERT model can be then fine-tuned with an additional output layer for classification. The main difference vis-a-vis the classical word embeddings like WORD2VEC is the fact that BERT produces word representations that are dynamically informed by the words around them. Further details on the BERT model are provided in the next section and here we just provide a brief description of the two explored strategies for extracting word embeddings from BERT: based on non-fine-tuned and fine-tuned models. The sequence embeddings based on the non-fine-tuned model are taken as the average of all sub-word embeddings from the given text sequence, extracted from the second-to-last hidden layer. On the other hand, the sequence embeddings based on

[8] https://catalog.ldc.upenn.edu/LDC2011T07

[9] https://commoncrawl.org/

the fine-tuned model are taken as the final hidden vectors of the special [CLS] tokens (which are fed into the output layer for classification).

4.4 Fine-tuned BERT

As introduced in Section 4.3, BERT is a deep bidirectional language representation model which can be pre-trained in an unsupervised manner and then fine-tuned for the specific downstream task, in our case classification. BERT's architecture is a multi-layer bidirectional Transformer encoder based on the original implementation described in (Vaswani et al., 2017). In our work we used the so called BERT-BASE version based on 12 Transformer blocks (layers), 738-dimensional hidden vectors, 12 self-attention heads and in total 110M paremeters. The model is pre-trained using two unsupervised tasks: masked language model (task of predicting some masked tokens) and next sentence prediction. The pre-training corpus was lower-cased English text of the BooksCorpus (800M words) and English Wikipedia (2,500M words). For fine-tuning all model parameters are initialized with the values from the pre-trained model, an additional output layer for classification is used and all the parameters are updated based on the labeled data. For further details see (Devlin et al., 2019) and the references included therein.

5 Experiments

We have evaluated five SVM-based classifiers (two of which are feeded with BERT embeddings) and one end-to-end deep NN-based BERT classifier. More specifically, the following models are included in our experimental setup: (a) SVM with TF-IDF char *n*-grams (SVM-CHAR), (b) SVM with FASTTEXT embeddings (SVM-FAST), (c) SVM with GLOVE embeddings (SVM-GLOVE), (d) SVM with non fine-tuned BERT embeddings (SVM-BERT), (e) SVM with fine-tuned BERT embeddings (SVM-F-BERT), and finally (f) the deep bidirectional transformer encoder BERT (BERT). All models were used for running experiments on all four ACLED datasets.

5.1 Experiment settings

For implementing the SVM models, we use scikit-learn (Pedregosa et al., 2011). The SVM pairwise classification is implemented using scikit-learn's LinearSVC SVM classifier with the One-Versus-One wrapper (Pedregosa et al., 2011). For the experiments with BERT we have used the Pytorch-Transformers library by HuggingFace (Wolf et al., 2019). We have fine-tuned the pre-trained BERT model for 3 epochs with a learning rate of 3e-5 and a batch size of 32. Padding and truncation of the input text sequences have been performed with the maximum sequence length of 64 (extending the maximum sequence length to 128 led to only marginal performance gain).

We use a shuffle-split 80% training (for BERT, 75% training, 5% development), 20% testing. When testing different portions of the training set, 1%, 5%, 10%, 50% and 100%, the test set remain the same and the portions are created using stratification split to make sure that the heterogeneous class distribution is maintained. For the 1% portion configuration, due to the high variability of such small training set, we use a 10-fold shuffle-split cross-validation configuration and the F_1-scores reported for this 1% portion configuration correspond to the average obtained on the 10 folds.

5.2 Evaluation Metrics

For measuring the event classification performance we used 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 25-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.

5.3 Results

First, in Table 2 the comparison of micro, macro and weighted F_1 score for all 6 benchmark models trained on 100% of the training data for all ACLED corpora is provided (with the exception of SVM-BERT and SVM-F-BERT which were evaluated only on ACLED-III). One can observe that BERT consistently

outperforms other models on all three corpora, which is followed by SVM-F-BERT in the case of ACLED-III. Somewhat surprisingly, SMV-CHAR model constitutes an extremely well-performing runner-up to the fine-tuned BERT-based approaches, and due to its simplicity makes it attractive from the application point of view. Furthermore, like shown in other studies (Reimers and Gurevych, 2019), fine-tuning BERT transformer results in performance boost.

Corpus	SVM-CHAR	SVM-FAST	SVM-GLOVE	SVM-BERT	SVM-F-BERT	BERT
			micro F_1			
ACLED-I	92.4	82.6	85.4	-	-	**94.9**
ACLED-II	91.8	82.4	85.3	-	-	**94.4**
ACLED-III	91.8	82.3	85.0	87.3	93.8	**94.3**
			macro F_1			
ACLED-I	83.8	61.7	70.9	-	-	**88.9**
ACLED-II	80.7	60.2	69.1	-	-	**87.0**
ACLED-III	80.6	59.4	68.5	72.7	85.0	**86.0**
			weighted F_1			
ACLED-I	92.3	81.7	84.9	-	-	**94.8**
ACLED-II	91.7	81.4	84.8	-	-	**94.4**
ACLED-III	91.6	81.3	84.5	87.0	93.8	**94.2**

Table 2: Comparison of micro, macro and weighted F_1 scores on ACLED-I, ACLED-II and ACLED-III datasets using 100% of the training data.

Interestingly, the best results obtained by all models were actually on ACLED-I. Without speculating whether the differences between the results on ACLED-I versus the two other are statistically significant we can hypothesize that better results on ACLED-I might be due to: (a) some discriminatory power of the "noise" that was removed from ACLED-I while creating the other corpora, e.g., the specific comments added by the humans (in ACLED-I) might have been associated with specific type of events, and (b) presence of some initial phrases in the event descriptions in ACLED-I which might have constituted good indicators of the event type (see Section 3), which are absent in the two other datasets.

In Figure 3 we provide the learning curves for micro and macro F_1 score for ACLED-III dataset and four main benchmark models (four for simplicity reasons), using different portions ($1\%, 5\%, 10\%, 50\%$ and 100%) of the training data of ACLED-III. One can observe that already with 1% of the training data (approx 4.7K events) all models obtain micro F_1 above 70%, whereas reaching 70% macro F_1 requires circa 10% (approx 50K events), and only BERT achieves this result actually. However, with smaller amount of data (i.e., less than 2-3%) BERT might not constitute the best choice in terms of macro F_1 as one can infer from the diagram in Figure 3. According to Table 2 and Figure 3, BERT outperforms its competitors in all configurations and for all datasets, except the case in which the training set is relatively tiny. With 1% of the training data (approx. 5K events), BERT macro F_1 drops under SVM-CHAR and SVM-GLOVE. This drop is specifically visible when observing the obtained results per class, which are provided in Table 4 in Appendix A that compares macro F_1 scores per class for BERT and SVM-CHAR with 1% and 100% of the training data available. BERT macro F_1 is equal or close to zero for 9 most poorly populated classes, whereas out of the 12 most poorly populated classes, SVM-CHAR obtains better macro F_1 for 11 of them. This illustrates how unstable BERT can be when the training set is tiny.

For the two best performing models, namely, SVM-CHAR and BERT, Figure 4 provides the comparison of the learning curves for weighted F_1 score across different ACLED corpora, from which we can observe again that although ACLED-II and ACLED-III were supposed to contain less "noise" (which turned to have some discriminatory power) they are actually "harder" than ACLED-I.

Finally, in order to give an insight into the models robustness in the context of data drift, Table 3 provides the comparison of four benchmark models[10] in terms of micro and macro F_1 on ACLED-III and ACLED-III-Δ datasets using full training data. Although there is significant lexical variation between the two corpora (see 1), one could not observe dramatic loss in the performance of any of the models.

[10]Similar behaviour was observed for the other two models, therefore we did not include them here.

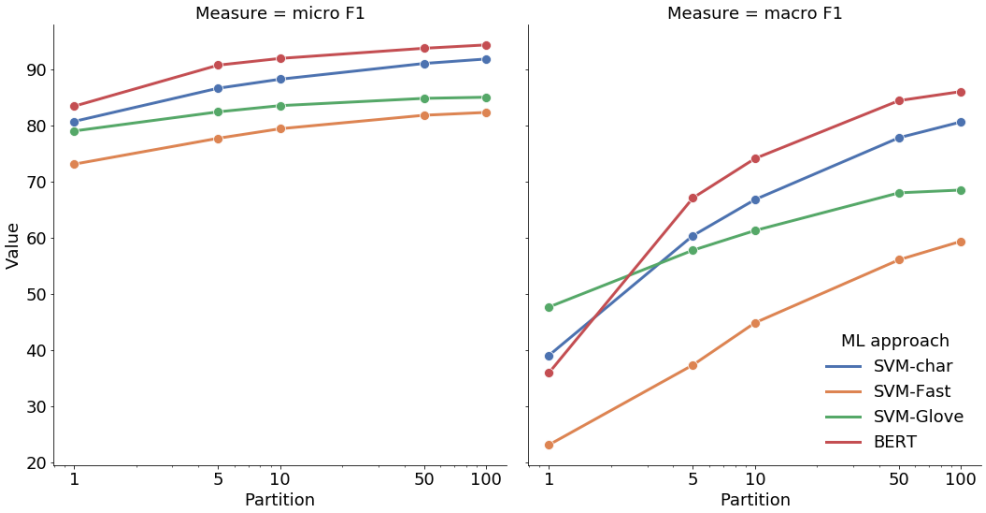

Figure 3: Comparison of micro and macro F_1 measure results on ACLED-III dataset for all benchmark models using different portion of the training data.

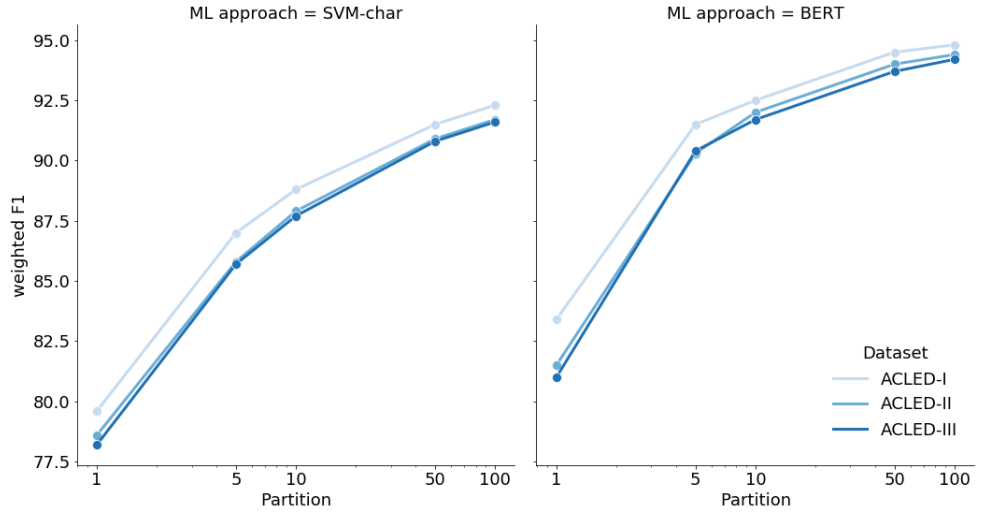

Figure 4: Comparison of weighted F_1 scores on ACLED-I, ACLED-II and ACLED-III datasets for SVM-CHAR and BERT models. The values on the x-axis indicate the percentage of the training dataset exploited for training of the respective models.

Corpus	SVM-CHAR	SVM-FAST	SVM-GLOVE	BERT	SVM-CHAR	SVM-FAST	SVM-GLOVE	BERT
		micro F_1				macro F_1		
ACLED-III	91.8	82.3	85.0	94.3	80.6	59.4	68.5	86.0
ACLED-III-Δ	90.9	79.1	83.7	93.7	78.3	54.1	64.8	84.8

Table 3: Comparison of micro/macro F_1 on ACLED-III/ACLED-III-Δ using 100% of the training data.

5.4 Error Analysis

In order to carry out a basic error analysis we have focused on the models trained on the full training data of ACLED-III and computed confusion matrices on the common test set used for the evaluation of all models. Based on the confusion matrices normalized for predicted conditions (columns) calculated with SVM-GLOVE, SVM-FAST, SVM-CHAR and BERT models, and shown in Figures 5 and 6 in Appendix A, we make the following observations. The most prevalent type of error is the misclassification

of many types of events as Armed Clash or Attack events, which applies mainly to SVM-FAST and SVM-GLOVE, and to a much lesser extent to SVM-CHAR and BERT. We hypothesize that this type of mismatches is mainly due to the fact that both armed clashes and attacks are mentioned in the text reporting other events, e.g., events on regaining and overtaking territory. Certain significant fraction of mismatches results most likely from small nuances in the definition of the respective event types, e.g., in the context of events related to protests: Peaceful protest, Protest with intervention and Force against protesters, again, more prominent phenomenon in the case of SVM-FAST and SVM-GLOVE. Further, we observe one specific error outlier that applies only to SVM-FAST, whereas it remains insignificant in the context of other models, namely, a mismatch of Chemical weapon events as Air/drone strike (38%) or Artillery/Missile attack (29%) events. For the best performing models, namely, SVM-CHAR and BERT, somewhat unsurprisingly, the Other event type is misclassified most, i.e., in 59% and 41% of cases respectively. Lastly, we note that the highest percentage of misclassifications between two event types for SVM-GLOVE, SVM-FAST, SVM-CHAR and BERT is 67%, 52%, 32% and 14% respectively.

6 Discussion

Based on the results presented in Section 5.3 we can draw some general conclusions regarding real-world usability of the benchmark models evaluated. All SVM-based and fine-tuned BERT-based transformer model happened to be resistant to somewhat more noisy data, where in the case of the latter model we were expecting to see some more visible performance loss given that the ACLED event descriptions appeared to be different from the corpora on which the pre-trained BERT embeddings were learned. As regards the classification performance BERT appears to be the winner among the compared models and evaluation metrics, however, when time complexity is a concern (BERT being known to be of magnitudes slower), the runner-up, namely TF-IDF character *n*-gram based SVM performed surprisingly well, and is not lagging far behind, would be definitely the better choice. We have also observed that dropping all *n*-grams with TF-IDF below 0.001 reduces the dimensionality of the feature space by the factor of at least x10 with only marginal impact on classification performance, which makes SVM-CHAR even more attractive. In addition, in the case of selecting a solution for obtaining the best results in terms of macro F_1 and having available only a tiny fraction of the training data (i.e. less than ca. 20K events) with some very sparsely populated event classes like in ACLED, using BERT would not be convenient vis-a-vis other models that managed better to tackle the data sparseness problem. Furthermore, all models did not turn to dramatically suffer from significant data drift based on some rudimentary robustness tests carried out.

7 Conclusions

In this paper we presented large datasets for evaluation of fine-grained event classification (25 classes), which were derived from ACLED data, a human-created event repository. We compared the performance of 6 state-of-the-art benchmark models, spanning SVM and NN-based classifiers that exploit TF-IDF character *n*-grams and off-the-shelf pre-trained non-contextual and contextual word embeddings as features. The best results in terms of micro (94.3-94.9%) and macro F_1 (86.0-88.9%) were obtained using the popular BERT transformer, however the significantly simpler TF-IDF character *n*-gram based SVM constitutes an interesting alternative.

There are various avenues to explore in the future, including, i.a., (a) evaluation of models that exploit the hierarchy of the event types, (b) carrying out more advanced robustness tests (Jin et al., 2020), (c) exploration of other transformer-based approaches (Sanh et al., 2019; Adhikari et al., 2019), (d) more in-depth cross-model error analysis, and (e) alleviating the problem with uneven event type distribution, and creating somewhat more "balanced" version of the datasets. All raw ACLED-derived corpora can be downloaded at: http://piskorski.waw.pl/resources/acled/ACLED-DATASETS.zip, whereas the corresponding versions with partitions into training and test data are accessible at: http://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/2020_annotated_event_dataset/Folds/. Please note that for the purpose of carrying out the evaluations reported in this paper the test data from Fold 1 was used.

References

ACLED. 2019. Armed Conflict Location & Event Data Project (ACLED) Codebook. Technical report. Accessed at: https://www.acleddata.com/resources/general-guides/.

Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. 2019. DocBERT: BERT for Document Classification. *CoRR*, abs/1904.08398.

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

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3:1137–1155.

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, June. 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, May. European Language Resources Association (ELRA).

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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, June. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In *Proceedings of AAAI 2020*.

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.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, page 423–430, USA. Association for Computational Linguistics.

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).

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, July. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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, July. 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, December. The COLING 2016 Organizing Committee.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David, Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

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, May. European Language Resources Association (ELRA).

Jakub Piskorski, Hristo Tanev, Martin Atkinson, Eric van der Goot, and Vanni Zavarella. 2011. Online News Event Extraction for Global Crisis Surveillance. In Ngoc Thanh Nguyen, editor, *Transactions on Computational Collective Intelligence*, pages 182–212. Springer Berlin Heidelberg, Berlin, Heidelberg.

Ali Rahimi and Benjamin Recht. 2008. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

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, September. INCOMA Ltd. Shoumen, BULGARIA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Andrea Vedaldi and Andrew Zisserman. 2012. Efficient additive kernels via explicit feature maps. *IEEE transactions on pattern analysis and machine intelligence*, 34(3):480–492.

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.

Christopher KI Williams and Matthias Seeger. 2001. Using the Nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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.

Appendices

A Supporting statistics and information

	1% training BERT	SVM-CHAR	100% training BERT	SVM-CHAR	support (test)
Peaceful protest	0.968	0.931	0.984	0.976	31511
Armed clash	0.882	0.823	0.956	0.929	27506
Attack	0.767	0.709	0.915	0.869	11560
Shelling/artillery/missile attack	0.939	0.904	0.978	0.968	10440
Air/drone strike	0.951	0.941	0.987	0.979	8698
Remote explosive/landmine/IED	0.862	0.858	0.970	0.952	5770
Violent demonstration	0.697	0.628	0.862	0.817	5179
Mob violence	0.488	0.581	0.851	0.804	4646
Protest with intervention	0.630	0.456	0.813	0.756	2455
Looting/property destruction	0.100	0.162	0.808	0.764	1193
Government regains territory	0.464	0.375	0.839	0.758	1174
Change to group/activity	0.442	0.232	0.838	0.784	1148
Abduction/forced disappearance	0.554	0.523	0.903	0.845	1065
Disrupted weapons use	0.087	0.309	0.891	0.836	877
Non-state actor overtakes territory	0.000	0.204	0.784	0.645	753
Grenade	0.732	0.634	0.893	0.867	692
Arrests	0.000	0.135	0.890	0.815	688
Other	0.000	0.086	0.640	0.518	553
Excessive force against protesters	0.000	0.142	0.692	0.599	512
Suicide bomb	0.000	0.256	0.933	0.858	369
Non-violent transfer of territory	0.006	0.073	0.730	0.661	341
Sexual violence	0.000	0.034	0.930	0.893	292
Agreement	0.008	0.023	0.831	0.768	260
Headquarters or base established	0.000	0.000	0.758	0.750	88
Chemical weapon	0.000	0.000	0.829	0.743	18
macro avg	0.383	0.401	0.860	0.806	117788
weighted avg	0.819	0.785	0.942	0.916	117788

Table 4: Macro F_1 scores per class for BERT and SVM-CHAR on ACLED-III dataset.

Event Type	Event Subtype	ACLED-I	ACLED-II	ACLED-III
BATTLES		151955	151193	146441
	Armed clash	141871	141331	136944
	Government regains territory	6119	5975	5809
	Non-state actor overtakes territory	3965	3887	3688
EXPLOSION AND REMOTE VIOLENCE		134153	134052	129273
	Chemical weapon	106	105	103
	Air/drone strike	46222	46177	43617
	Suicide bomb	1775	1760	1738
	Shelling/artillery/missile attack	52716	52692	51484
	Remote explosive/landmine/IED	29514	29501	28804
	Grenade	3820	3817	3527
VIOLENCE AGAINST CIVILIANS		70844	70733	65100
	Sexual violence	1770	1759	1544
	Attack	63121	63027	58124
	Abduction/forced disappearance	5953	5947	5432
PROTESTS		177082	176916	173443
	Peaceful protest	161829	161701	158500
	Protest with intervention	12636	12611	12414
	Excessive force against protesters	2617	2604	2529
RIOTS		50545	50341	48964
	Violent demonstration	27092	26919	26147
	Mob violence	23453	23422	22817
STRATEGIC DEVELOPMENTS		27099	26872	25719
	Agreement	1415	1394	1340
	Arrests	3518	3505	3432
	Change to group/activity	6112	6025	5737
	Disrupted weapons use	4641	4629	4507
	Headquarters or base established	589	584	468
	Looting/property destruction	6008	5973	5719
	Non-violent transfer of territory	1821	1814	1674
	Other	2995	2948	2842
TOTAL		611678	610107	588940

Table 5: ACLED-I, ACLED-II and ACLED-III event corpus statistics: Number of events.

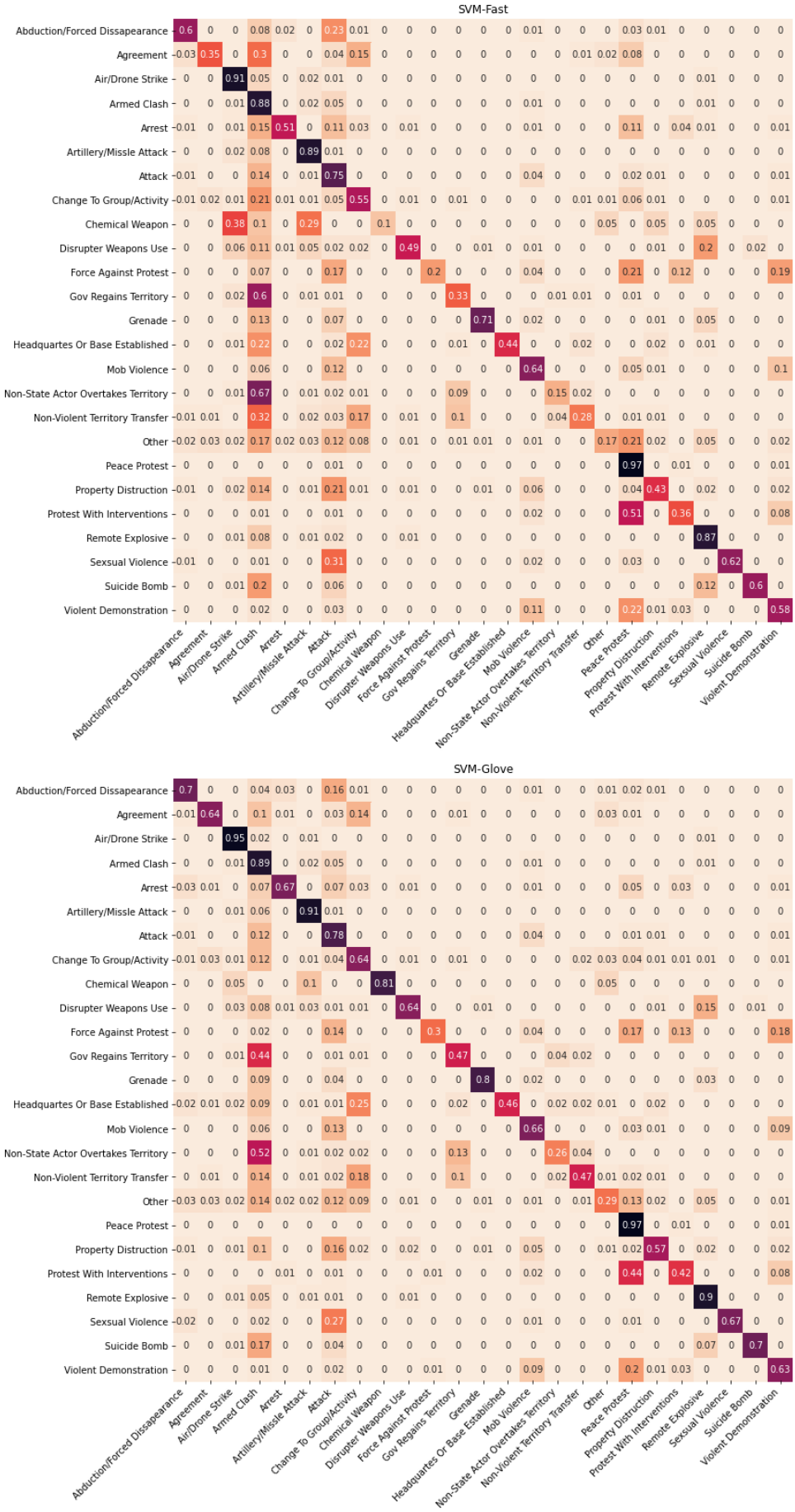

Figure 5: Confusion matrices for SVM-FAST (top) and SVM-GLOVE (bottom) evaluated on ACLED-III dataset.

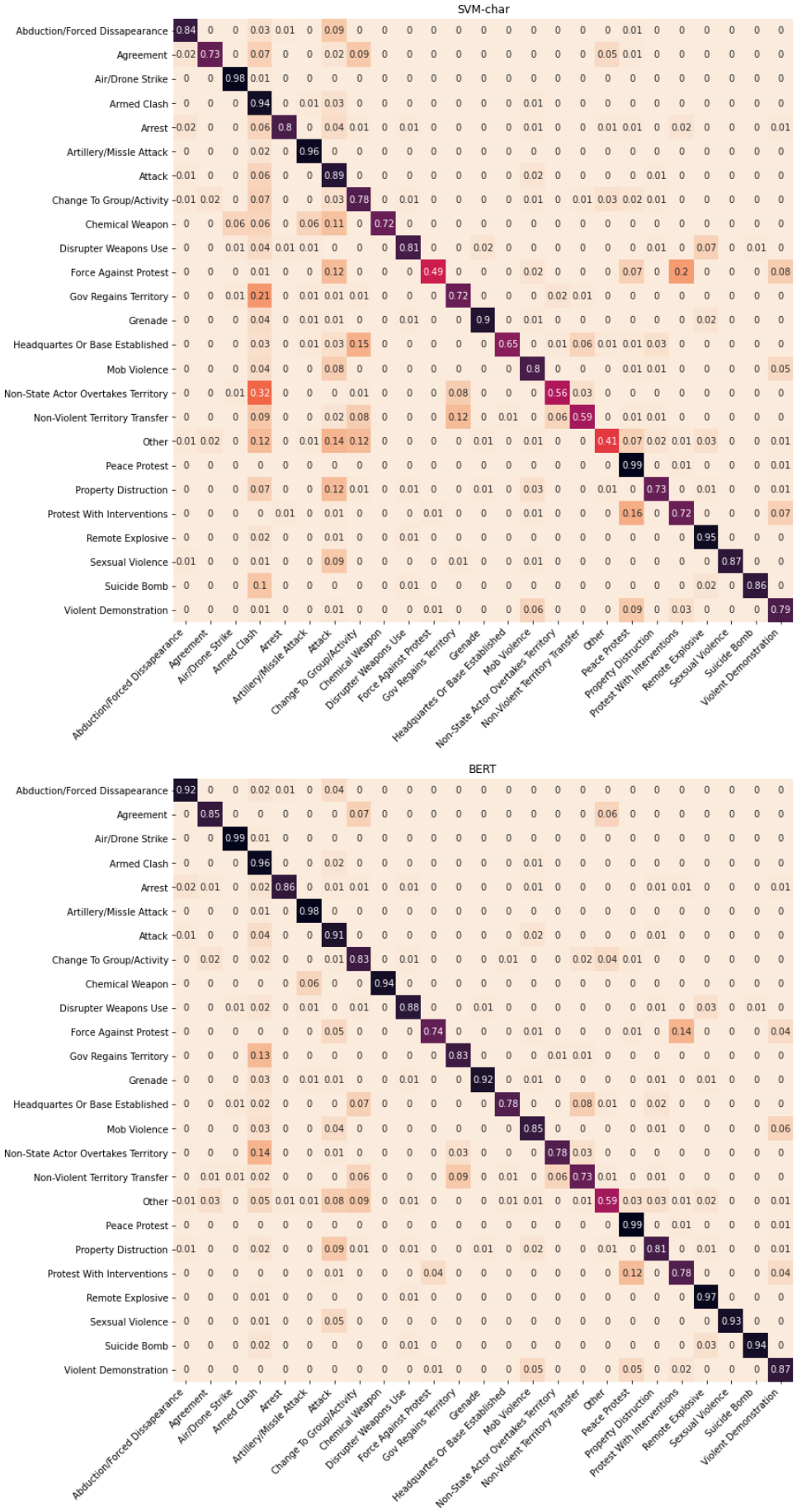

Figure 6: Confusion matrices for SVM-CHAR (top) and BERT (bottom) evaluated on ACLED-III dataset.