Proceedings of AESPEN 2020, pages 26–34
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

TF-IDF Character *N*-grams versus Word Embedding-based Models for Fine-grained Event Classification: A Preliminary study

Jakub Piskorski, Guillaume Jacquet
Polish Academy of Sciences, Joint Research Centre - European Commission
jpiskorski@gmail.com, guillaume.jacquet@ec.europa.eu

Abstract

Automating the detection of event mentions in online texts and their classification vis-a-vis domain-specific event type taxonomies has been acknowledged by many organisations worldwide to be of paramount importance in order to facilitate the process of intelligence gathering. This paper reports on some preliminary experiments of comparing various linguistically-lightweight approaches for fine-grained event classification based on short text snippets reporting on events. In particular, we compare the performance of a TF-IDF-weighted character *n*-gram SVM-based model with SVMs trained on various off-the-shelf pre-trained word embeddings (GLOVE, BERT, FASTTEXT) as features. We exploit a relatively large event corpus consisting of circa 610K short text event descriptions classified using 25-event categories that cover political violence and protest events. The best results, i.e., 83.5% macro and 92.4% micro F_1 score, were obtained using the TF-IDF-weighted character *n*-gram model.

Keywords: event classification, machine learning, word embeddings, subword models

1. Introduction

Recently various organisations around the world have acknowledged the paramount importance of exploiting the ever-growing amount of information published on the web on various types of events for early detection of threats, carrying out risk analysis and predicting future developments (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). Since a clear majority of information on relevant events is provided in the form of free text (e.g. news articles), an important task is to automatically detect mentions of events of interest in such texts and to classify them using domain specific taxonomies.

This paper reports on a preliminary study of exploiting linguistically-lightweight approaches for fine-grained event classification for short texts reporting on events. In particular, we compare the performance of various SVM-based classifiers, including a TF-IDF-weighted character *n*-gram model with various models that exploit off-the-shelf pre-trained word embeddings (GLOVE, BERT and FASTTEXT) as features.

Our research has two aims. Firstly, we are interested to develop a robust, fine-grained, event classifier that can be: (a) easily ported across languages, (b) quickly adapted to new domains/event taxonomies, and (c) applied to classify events based solely on short text snippets. The decision to focus on classification from short texts come from the type of incomplete event data that is often at hand, e.g. historical news event information stored in so called event templates, that apart from automatically extracted meta-data include only the title and 1-2 initial sentences from a news article from which the event information was extracted (Atkinson et al., 2017). Secondly, we are interested to gain a better understanding of the amount of data that is required to obtain 'acceptable' classification performance in order to better estimate the effort required for new classification-scheme development cycles.

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

- we make available a clean and tuned version of a large corpus of circa 600K short text snippets tagged with fine-grained event category labels (mainly covering political violence and protest events) for event classification experiments, which was derived from a manually-curated event repository created by human experts in the context of the ACLED[1]
- we report on the comparison of the performance of various SVM-based classifiers, including a TF.IDF-weighted character *n*-gram model and models that exploit various pre-trained word embeddings as features, evaluated on the aforementioned event corpus.

We are not aware of any similar study on automated event classification in terms of the size of the underlying training-test dataset and fine-grained event categories. Furthermore, given the specific nature of the dataset exploited, i.e. text snippets resembling news headlines and initial sentences in news articles, we believe that the reported results constitute a good approximation for the to-be-expected performance when applying the same methods on real news-article corpora.

The rest of the paper is structured as follows. First, Section 2. provides an overview of related work. Next, Section 3. describes the dataset used for carrying out the experiments. Subsequently, Section 4. presents the results of the performance of the various classification models explored. Finally, Section 5. gives conclusions and an outlook on future work.

2. Related Work

The research and progress on the task of identifying event mentions in text documents and classification of these

[1] `https://www.acleddata.com` initiative, and

events was initially driven by the Message Understanding Contests (Sundheim, 1991; Chinchor, 1998) and the Automatic Content Extraction (ACE) Challenges (Doddington et al., 2004; LDC, 2008). In particular, 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).

Recently, the Multi-lingual Event Detection and Coreference task has been introduced as part of the Text Analysis Conference (TAC) in 2016[2] and 2017[3], which included an Event Nugget Detection subtask, focusing on detection and classification of intra-document event mention types and subtypes with 9 and 38 categories respectively, that cover events from various domains (e.g., finances and jurisdiction). The related evaluation datasets are rather tiny, i.e., ca. 500 documents with less than 10K labelled event mentions.

Furthermore, a CLEF ProtestNews Track was organized recently (Hürriyetoğlu et al., 2019) with three shared tasks aimed at identifying and extracting event information from news articles across multiple countries, where one of the tasks explicitly focused on classification of the news articles into "protests" versus "non protests" depending on whether the article reports on protests, and a more fine-grained binary classification task that focused on labelling sentences that refer to reporting on protest events. Similarly to the TAC tasks, the evaluation datasets are rather small (4K news articles, and 6K labelled sentences). In particular, approaches exploiting word embeddings to tackle these tasks have been reported (Ollagnier and Williams, 2019).

The work most similar to ours on event classification has been presented in (Nugent et al., 2017). This paper studies the performance of various models, including ones that exploit word embeddings as features, for detection and classification of natural disaster and critical socio-political events in news articles, based on analysing their initial sentences. However, the underlying event type taxonomy is relatively coarse-grained (7 types) and the size of the evaluation dataset is relatively small (ca. 2.5K documents).

In the work reported in this paper we only focus on the task of event classification, and given the specific dataset (in particular, its size) exploited for carrying out our, it is difficult to make direct comparisons with the shared tasks and evaluation campaigns mentioned above.

3. Datasets

For carrying out our research, we exploited the data gathered in the context of the Armed Conflict Location & Event Data Project (ACLED)[4]. ACLED (Raleigh et al., 2010) collects human-moderated records on the dates, actors, types of violence, locations, and fatalities of all reported political violence and protest events across Africa, some regions of Asia, the Middle East, and Southeastern and Eastern Europe and the Balkans. In particular, we exploited the manually curated data provided on the ACLED web page[5] and extracted from them event records consisting of: event type, event subtype and textual description, which mentions basic information on the event. ACLED uses an event ontology consisting of 6 main event types, which are subdivided into 25 more fine-grained subtypes, listed in Table 1. Two examples of event descriptions for `Abduction/forced disappearance` and `Peaceful protest` events resp. are given below.

```
[1] A girl was kidnapped in Ain El
    Turk, Oran by unidentified
    individuals. Police managed to
    free the girl 3 days later.

[2] On 20 February, a group of 30
    anarchists protested in front
    of the Russian consulate in
    north Athens unfurling banners
    in support of Russian anarchists
    and scattering fliers.
```

The detailed definition of the ACLED event hierarchy is presented in (ACLED, 2019). We were able to extract from ACLED curated resources 614107 event triples, consisting of the type, subtype and short event description. We will refer to this corpus as ACLED-O (ACLED Original). This corpus was subsequently cleaned, through: (a) removing from the event descriptions quotation and similar non-content relevant characters, (b) removing too obvious markers that would artificially help the classifier such as initial phrases in the event descriptions indicating the specific event type or subtype, e.g. "*Arrest:*", and (c) filtering out event triples that contain event descriptions consisting of less than 20 characters (considered as non informative). We will refer to the resulting corpus as ACLED-C (ACLED Clean). Finally, we created a third version of the corpus to check if the mention of geographical names in an event description could impact the results of the classifier. We replaced in ACLED-C the occurrences of geographical names with a generic location tag, using the GEONAMES[6] gazetteer. The resulting dataset will be referred to as ACLED-CG. The specific event type/subtypes and related statistics of the ACLED-C datasets are listed in Table 1.

The distribution of the length of event descriptions for the ACLED-C dataset is shown in Figure 1. We can observe that the length of the vast majority of the event descriptions 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. We have, however, observed some outliers with a length of more than 1000 characters.

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

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

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

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

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

Event Type	Event Subtype	Number	Percent.
BATTLES		151955	24.84%
	Armed clash	141871	23.19%
	Government regains territory	6119	1.00%
	Non-state actor overtakes territory	3965	0.65%
EXPLOSION AND REMOTE VIOLENCE		134153	21.93%
	Chemical weapon	106	0.02%
	Air/drone strike	46222	7.56%
	Suicide bomb	1775	0.29%
	Shelling/artillery/missile attack	52716	8.62%
	Remote explosive/landmine/IED	29514	4.83%
	Grenade	3820	0.62%
VIOLENCE AGAINST CIVILIANS		70844	11.58%
	Sexual violence	1770	0.29%
	Attack	63121	10.32%
	Abduction/forced disappearance	5953	0.97%
PROTESTS		177082	28.95%
	Peaceful protest	161829	26.46%
	Protest with intervention	12636	2.07%
	Excessive force against protesters	2617	0.43%
RIOTS		50545	8.26%
	Violent demonstration	27092	4.43%
	Mob violence	23453	3.83%
STRATEGIC DEVELOPMENTS		27099	4.43%
	Agreement	1415	0.23%
	Arrests	3518	0.58%
	Change to group/activity	6112	1.00%
	Disrupted weapons use	4641	0.76%
	Headquarters or base established	589	0.10%
	Looting/property destruction	6008	0.98%
	Non-violent transfer of territory	1821	0.30%
	Other	2995	0.49%
TOTAL		611678	

Table 1: ACLED-C event corpus statistics: Number and percentage of event types and subtypes.

4. Experiments

4.1. Classification Tasks

In our research we were primarily interested in the fine-grained event classification vis-a-vis the subtypes enumerated in Table 1, which we call **Event Subtype Classification**. For the sake of completeness, and given the availability of the corpora introduced in the previous Section we also evaluated the performance of coarse-grained event classification, which will be referred to as **Event Type Classification**, in line with the terminology introduced in the ACLED corpus. In particular, we compared the results obtained on all three versions of this corpus, i.e., (ACLED-O, ACLED-C and ACLED-CG).

4.2. Approaches

We compare two main approaches to the Event Subtype/Type Classification, both using Support Vector Machine (SVM) model, where one is based on TF-IDF character *n*-grams features, and the other on exploiting various word embeddings as features for training the models. The SVM classification is 'pairwise' (One-Versus-One; OVO), meaning that a binary classifier is trained for each pair of classes and the class which receives most votes (highest count) is selected. This method of multi-class classification was favoured over One-Versus-Rest classification due to overall better results obtained. We chose an SVM classification approach following its widely-acknowledged strong performance on text classification tasks (Joachims, 1998; Yang and Liu, 1999; Qin and Wang, 2009; Ye et al., 2009; Chesney et al., 2017).

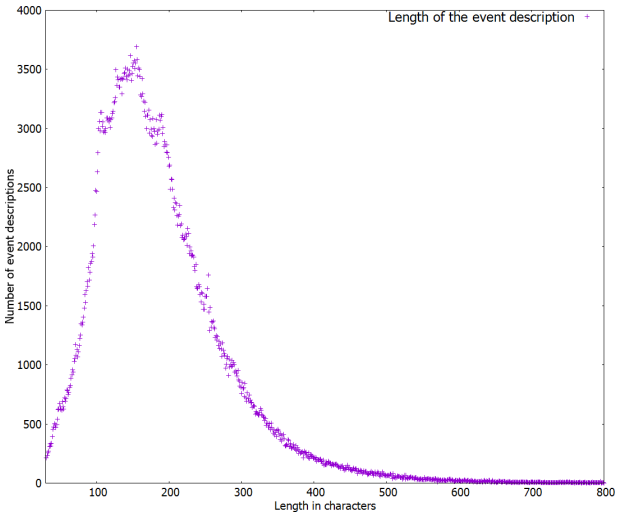

Figure 1: The distribution of the length of event descriptions for ACLED-C dataset.

4.2.1. TF-IDF character *n*-gram approach

We follow a bag-of-words (BoW) model for extracting TF-IDF features from the character *n*-grams contained within each event description. We use an *n*-gram range between 3 and 5-grams. 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).

In contrast to the word embedding approaches described in the next section, here 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. In the presented experiments, the TF-IDF vectors varied from 26 705, when using 0.5% of the training set, to 365 175 when using the full training set.

4.2.2. Word embedding-based approach

Word embeddings have proved to be an efficient method for solving various natural language processing tasks in recent years, enabling, in particular, various machine learning models that rely on vector representation as input to enjoy richer representations of text input while alleviating high-dimensionality issues. Formally, 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). Many authors have reported that word embeddings perform surprisingly well for text classification tasks (Reimers and Gurevych, 2019). In our initial experiments we used the popular GLOVE, BERT 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. For our experiments we used the pre-trained GLOVE 300-dimensional vectors trained on WIKIPEDIA and the English Gigaword corpus[7]. To compute a GLOVE embedding for an event description we averaged the single GLOVE embeddings of all words contained in the event description.

BERT (Devlin et al., 2019) is a pre-trained transformer network (Vaswani et al., 2017), which can be used to extract word and sentence embedding vectors for various NLP tasks. 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. For our experiments we exploited the pre-trained BERT multilingual (104 languages) cased model[8] that produces 768-dimensional vectors. As with GLOVE, we averaged the single BERT embeddings for all words in each event description. We have chosen the averaging of the BERT vectors based on the relatively good results reported on 7 different classification tasks in (Reimers and Gurevych, 2019), which yielded on average almost identical results vis-a-vis exploiting the [CLS] special token output from a BERT transformer.

FASTTEXT embeddings (Mikolov et al., 2018) are based on a model where 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. In our experiments, 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. Experiment settings

For implementing the SVM models, we use Scikit-learn (Pedregosa et al., 2011), the machine learning library for Python. The SVM pairwise classification is implemented using Scikit-learn's LinearSVC SVM classifier with the One-Versus-One wrapper (Pedregosa et al., 2011).

We use 10-fold shuffle-split cross-validation, split 75% training and 25% testing for all experiments. The general approach was as follows: the corpus is randomly shuffled (with a constant random state initialisation value for reproducibility) 10 times, and each shuffled version is then separated for training and testing. With this method, it is not guaranteed that each fold will be different, but it is likely with sizeable data sets; nonetheless, we favour this technique over k-fold cross-validation as it maximises the training data available, even for the smallest event subtypes.

4.4. Evaluation Methodology

For the sake of evaluating the event classification performance we used the classical precision, recall, and F_1 metrics, which are formally defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

[7] https://catalog.ldc.upenn.edu/LDC2011T07
[8] https://github.com/google-research/bert
[9] https://commoncrawl.org/

$$F_1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \quad (3)$$

where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively. To obtain a fine-grained picture, we evaluate both $micro$ and $macro$ versions of the introduced metrics and denote them with P_{mic}, P_{mac}, R_{mic}, R_{mac}, $F_{1_{mic}}$, $F_{1_{mac}}$ resp. While the micro versions calculate the performance from the individual true positives, true negatives, false positives, and false negatives of the 25-class model, in macro-averaging, one computes the performance of each individual class separately, and then an average of the obtained scores. In other words, micro versions of the metrics are indicators of the performance quality at the individual event level (biased by event type frequency), whereas the macro versions are indicators of the performance quality at the event type level disregarding the event type distribution.

4.5. Results

First, we evaluated the performance of the TF-IDF character *n*-gram based SVM on each of the three corpora, namely, ACLED-O, ACLED-C and ACLED-CG. The performance of the respective models is presented in Table 2. Given that there were no observable differences in performance on the three corpora, in particular between ACLED-C and ACLED-CG, all further experiments were carried out only on ACLED-C corpus.

Dataset	P_{mic}	R_{mic}	$F_{1_{mic}}$	P_{mac}	R_{mac}	$F_{1_{mac}}$
ACLED-O	0.924	0.924	0.924	0.884	0.816	0.845
ACLED-C	0.924	0.924	0.924	0.871	0.807	0.835
ACLED-CG	0.921	0.921	0.921	0.872	0.805	0.834

Table 2: Character *n*-gram-based SVM results on 75% of the ACLED-O, ACLED-C and ACLED-CG datasets

The micro and macro F_1 scores for the **Event Subtype Classification** task (fine grained classification) using different portions of the ACLED-C corpus for training and testing purposes ($0.5\%, 1\%, 5\%, 10\%, 50\%$ and 100%) are presented in Figure 2. The corresponding macro precision and recall figures are compared in Figure 3. We can observe that:

- overall, the TF-IDF character *n*-gram based model performs better than word embedding-based models, except the case when less than ca. 3% of data (ca. 20K events) is available for training, in whose context GLOVE-based approach works better with respect to the macro F_1 score,
- in particular, with the full dataset available (600K events) the TF-IDF character *n*-gram based model (reaching max. of 83.5% macro and 92.4% micro F_1 score) clearly outperforms ($> 10\%$) the word embedding-based approaches,
- already with a very small portion of the data, i.e., 0.5% (ca. 3K events) one obtained fairly good micro F_1 scores, ranging from 71.8% to 77.4%, whereas obtaining macro F_1 scores above 60% requires at least 10 to 50% of the data (60-300K events) for the various word embedding-based models,
- in general, out of the three word embedding-based approaches, GLOVE appears to work best, although with availability of more data the differences between F_1 scores for all three word embedding-based approaches become smaller and converge.

The micro and macro F_1 scores for the **Event Type classification** task (coarse grained) using different portions of the ACLED-C corpus for training and testing purposes ($0.5\%, 1\%, 5\%, 10\%, 50\%$ and 100%) are presented in Figure 4. In general, we can observe the same patterns as in the case of fine-grained event classification, i.e., TF-IDF character *n*-gram based model performs better (reaching max. 94.6% micro and 92.5% macro F_1 scores when using the entire corpus), GLOVE outperforming the other word embedding-based models with smaller amount of training data, etc. However, not surprisingly though, the main difference in this context are significantly higher micro and macro F_1 scores ranging from 78 to 85% and 68 to 77% resp. when training the models on a tiny portion of the data (i.e., 0.5% of the data, which corresponds to ca. 3K events). Similar results were obtained in the work reported in (Nugent et al., 2017) that compared the performance of similar-in-nature models trained and evaluated on comparable corpora in terms of its size.

4.6. Error Analysis

To get a better insight into the most frequent errors for the event subtype classification task we computed confusion matrices for the different approaches evaluated and concluded that the types of errors were similar across the different settings. Therefore, for the sake of completeness, we only present here the confusion matrix for the GLOVE-based SVM classifier, which is depicted in Figure 5.
We can observe from the confusion matrix that:

- classification of event subtypes within the `Explosions and Remote Violence` event type works best, i.e., true positive rate ranging from 82% to 95%,
- classification of event subtypes within the `Strategic Developments` and `Riots` main event types yields worst results on average, i.e., true positive rate ranging from 0.60% to 0.79%,
- most of the errors within the `Battles`, `Riots` and `Protests` main event types are due to mislabelling the event subtype with another subtype within the same main event type, which appears to be a logical consequence of small nuances of the definitions of the specific event subtypes and resulting high overlap of the respective vocabulary used in the event descriptions, e.g., `Government regains territory` versus `Non-state actor overtakes territory`

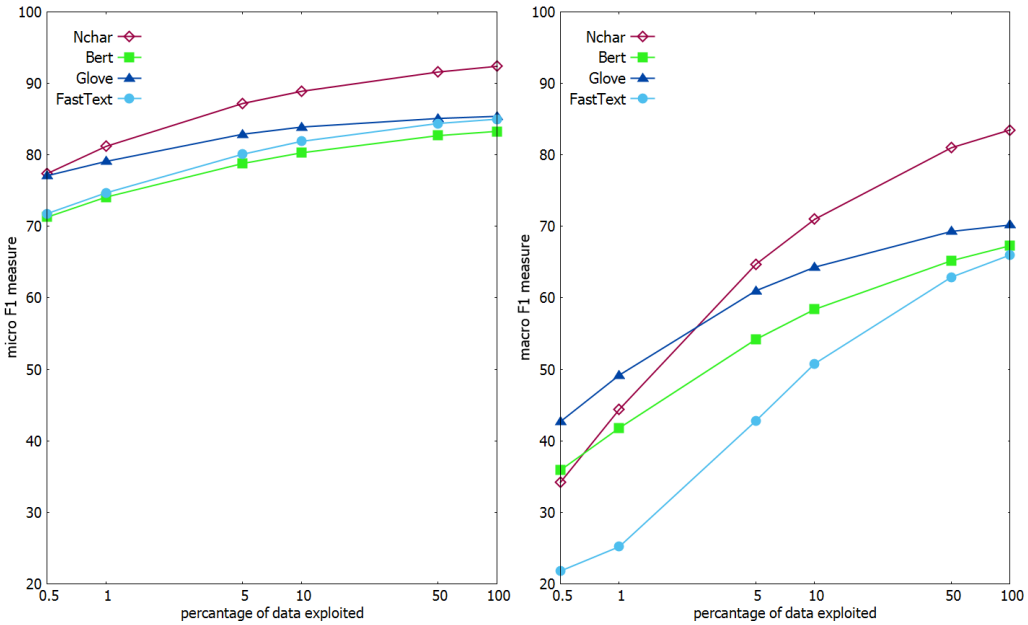

Figure 2: Micro and Macro F_1 measure results for Event Subtype Classification on the different subsets of ACLED-C dataset.

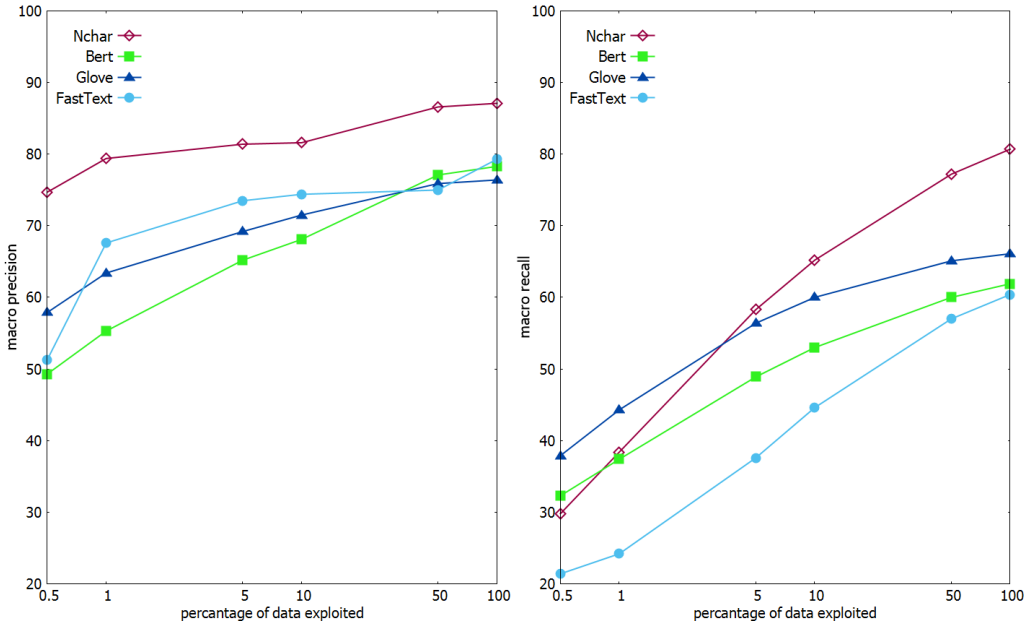

Figure 3: Macro Precision and Recall results for Event Subtype Classification on the different subsets of ACLED-C dataset.

or `Peaceful protest` event subtype versus `Protest with intervention`,

- vast majority of errors in general is due to wrongly classifying the event subtype as `Armed clash` (in the first row of the matrix one can observe clashes for 23 subtypes with the aforementioned event subtype), followed by errors resulting from misclassification of the subtype as `Attack`, which is most likely due to the fact that armed clashes and attacks constitute ca. 23% and 10% of all events resp., and
- finally, some more prominent observable clashes between event subtype misclassifications that go beyond the same main event and are worth mentioning are the ones that potentially result from similar vocabulary used (small nuances in the definition), e.g., the two following event descriptions were mis-classified by all approaches. [1] was supposed to be of type `Non-state actor overtakes territory` but has been classified as `Gov. regains terr`. Instead, [2] was supposed to be of type `Gov. regains terr` but has been classified as `Non-state actor overtakes territory`.

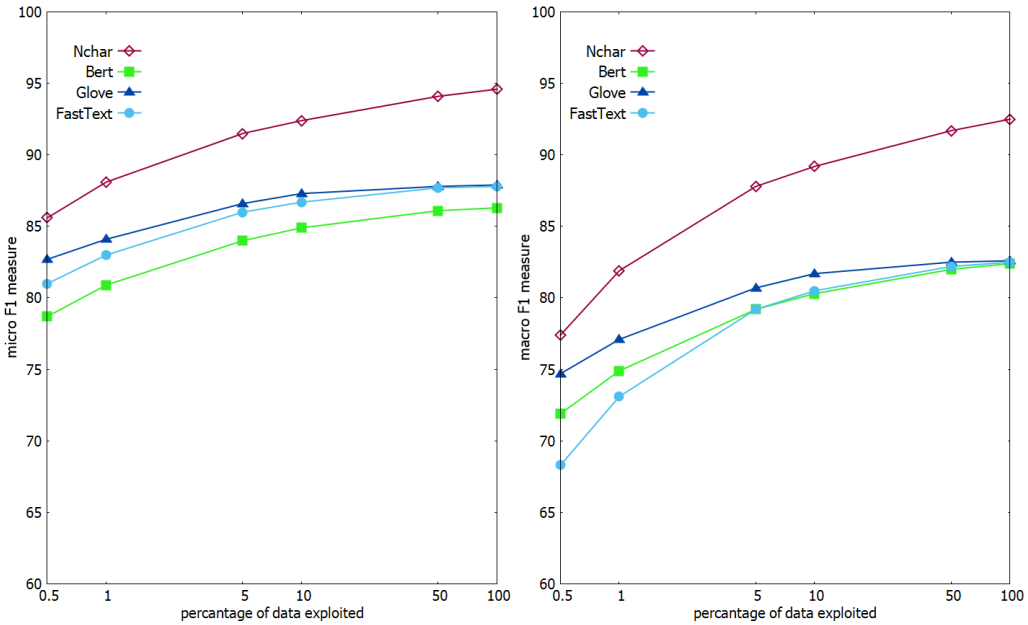

Figure 4: Micro and Macro F_1 measure results for Event Type Classification on the different subsets of ACLED-C dataset.

	Armed clash	Government regains territory	Non-state actor overtakes territory	Chemical weapon	Air/drone strike	Suicide bomb	Shelling/artillery/missile attack	Remote explosive/landmine/IED	Grenade	Sexual violence	Attack	Abduction/forced disappearance	Peaceful protest	Protest with intervention	Excessive force against protesters	Violent demonstration	Mob violence	Agreement	Arrests	Change to group/activity	Disrupted weapons use	Headquarters or base established	Looting/property destruction	Non-violent transfer of territory	Other
Armed clash	0.84	0.15	0.14	0.06	0.02	0.12	0.06	0.03	0.05	0.02	0.10	0.05	0	0	0.01	0.01	0.07	0.06	0.04	0.05	0.06	0.04	0.05	0.07	0.05
Government regains territory	0.02	0.63	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0.03	0	0.10	0
Non-state actor overtakes territory	0.01	0.12	0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.06	0.01
Chemical weapon	0	0	0	0.94	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Air/drone strike	0.01	0	0	0	0.95	0.01	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0.01	0	0.01
Suicide bomb	0	0	0	0	0	0.82	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Shelling/artillery/missile attack	0.02	0.01	0	0	0.01	0	0.92	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0.01	0.01	0	0
Remote explosive/landmine/IED	0.01	0	0	0	0	0	0	0.90	0.03	0	0.01	0	0	0	0	0	0	0	0	0	0.06	0	0.01	0.01	0.03
Grenade	0	0	0	0	0	0	0	0	0.85	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0
Sexual violence	0	0	0	0	0	0	0	0	0	0.88	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Attack	0.05	0.01	0.02	0	0	0.01	0	0.01	0.02	0.05	0.76	0.10	0	0.02	0.08	0.02	0.10	0.02	0.06	0.02	0	0.02	0.10	0.02	0.05
Abduction/forced disappearance	0	0	0	0	0	0	0	0	0	0	0.02	0.78	0	0	0	0	0	0.01	0.05	0.01	0	0.01	0.01	0.01	0.02
Peaceful protest	0	0	0	0	0	0	0	0	0	0.02	0.01	0.01	0.92	0.13	0.05	0.08	0.02	0.01	0.02	0.03	0	0	0.02	0.02	0.04
Protest with intervention	0	0	0	0	0	0	0	0	0	0	0	0	0.03	0.67	0.13	0.05	0.01	0	0.03	0	0	0	0	0	0.01
Excessive force against protesters	0	0	0	0	0	0	0	0	0	0	0	0	0	0.04	0.59	0.02	0	0	0	0	0	0	0	0	0
Violent demonstration	0	0	0	0	0	0	0	0	0	0.01	0.01	0	0.03	0.11	0.13	0.73	0.10	0	0.01	0.01	0	0	0.03	0	0.02
Mob violence	0.01	0	0	0	0	0.01	0	0	0.01	0.03	0.04	0	0.01	0.01	0.01	0.08	0.69	0	0.01	0.01	0	0	0.04	0.01	0.01
Agreement	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0.69	0	0.02	0	0.01	0	0.02	0.03
Arrests	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0.02	0	0	0	0.01	0.70	0.01	0.01	0	0	0	0
Change to group/activity	0	0.01	0.01	0	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0.09	0.02	0.75	0.01	0.06	0.01	0.06	0.06
Disrupted weapons use	0	0	0	0	0	0.03	0	0.02	0.01	0	0	0	0	0	0	0	0	0	0.02	0.01	0.79	0	0.01	0	0.01
Headquarters or base established	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.77	0	0.02	0.01
Looting/property destruction	0	0	0	0	0	0	0	0	0.01	0	0.01	0.01	0	0	0	0	0.01	0.01	0	0.01	0.02	0	0.70	0.01	0.02
Non-violent transfer of territory	0	0.05	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.02	0	0.02	0	0.06	0	0.60	0.01
Other	0	0	0	0	0	0	0	0.01	0.01	0	0	0.01	0	0	0	0	0	0.06	0.01	0.03	0	0	0.01	0.02	0.61

Figure 5: The confusion matrix for Event Subtype classification using GLOVE word embeddings.

[1] on 25 january the ajdabiya shura council claimed to have retaken the 18 gate from the lna 21 border guards the gate southeast of the city was secured by the lna on 10 jan.

[2] rebels attacked the town of tin hama being held by gatia progovernment troops and briefly took over before the malian military intervened and chased the rebels out.

5. Conclusions

In this paper we reported on some preliminary experiments comparing various linguistically-lightweight approaches for fine-grained event classification based on short text snippets reporting on events. The results of our tests on a relatively large event corpus revealed that a TF-IDF-weighted character *n*-gram SVM-based model outperforms (reaching 83.5% macro and 92.4% micro F_1 score) SVM models that exploit various of-the-shelf pre-trained word embeddings as features.

While the results reported in this paper are promising and the event description in the ACLED corpus used for the evaluation strongly resemble the headlines and leading sentences of news articles reporting on events, one can only hypothesize that similar results could be obtained on real news data. Also, there are other more complex ways of exploiting word embeddings using neural architectures that were not explored in this work. Therefore, in order to get a more in-depth insight and more complete picture we intend to explore the performance of other shallow learners, including non-linear SVM models, decision trees and deployment of other type of word embedding-based approaches too, e.g. Sentence-BERT embeddings (Reimers and Gurevych, 2019) and tuning thereof for the particular task at hand. Furthermore, future work might also encompass: (a) exploring ways to combine the TF-IDF character *n*-gram and word embedding-based approaches to boost the performance, and (b) studying the impact of the length of the event descriptions on the overall performance.

Furthermore, we intend to create two additional corpora: (a) one consisting of real news article snippets reporting on events in order to study whether one can obtain similar performance to the one reported in this paper, and (b) a multilingual version of the ACLED corpus in order to study the portability of the approaches across languages, benefiting in particular from the existence of pre-trained multilingual word embeddings, such as the ones we experimented with in this paper.

The ACLED-C dataset and the corresponding word embedding vectors that were computed and used for carrying out the experiments reported in this paper are accessible at `https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/LANGUAGE-TECHNOLOGY/2020_annotated_event_dataset/ACLED-G_dataset/`. The The ACLED-C dataset is also available as one file[10] from `http://piskorski.waw.pl/resources/acled/ALL.zip`

6. Bibliographical References

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

Atkinson, M., Piskorski, J., Yangarber, R., and van der Goot, E. (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.

Atkinson, M., Piskorski, J., Tanev, H., and Zavarella, V. (2017). On the Creation of a Security-Related Event Corpus. In *Proceedings of the Events and Stories in the News Workshop 2017*, pages 59–65, Vancouver, Canada. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155.

Chesney, S., Jacquet, G., Steinberger, R., and Piskorski, J. (2017). Multi-word entity classification in a highly multilingual environment. *Proceedings of EACL 2017 Multi-Word Expressions Workshop*.

Chinchor, N. A. (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*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (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).

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., and Zhu, Q. (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.

Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., and Mutlu, O. (2019). A task set proposal for automatic protest information collection across multiple countries. In Leif Azzopardi, et al., editors, *Advances in Information Retrieval*, pages 316–323, Cham. Springer International Publishing.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer.

King, G. and Lowe, W. (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.

[10] Each line contains the event description followed by (tab-separated) event type and subtype.

ONLINE: https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications.

Leetaru, K. and Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2. Citeseer.

Liao, S. and Grishman, R. (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.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Nguyen, T. H. and Grishman, R. (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.

Nguyen, T. H., Fauceglia, N., Rodriguez Muro, M., Hassanzadeh, O., Massimiliano Gliozzo, A., and Sadoghi, M. (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.

Nugent, T., Petroni, F., Raman, N., Carstens, L., and Leidner, J. L. (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.

Ollagnier, A. and Williams, H. (2019). Classification and event identification using word embedding. In *Proceedings of CLEF*.

Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., and Martínez Pérez, G. (2020). The not yet exploited goldmine of osint: Opportunities, open challenges and future trends. *IEEE Access*, 8:10282–10304.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., David, Cournapeau, Brucher, M., Perrot, M., and Édouard Duchesnay. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., and Zavarella, V. (2011). In Ngoc Thanh Nguyen, editor, *Transactions on Computational Collective Intelligence*, pages 182–212. Springer Berlin Heidelberg, Berlin, Heidelberg.

Qin, Y.-P. and Wang, X.-K. (2009). Study on multi-label text classification based on svm. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, volume 1, pages 300–304. IEEE.

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

Reimers, N. and Gurevych, I. (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.

Sundheim, B. M. (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*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (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.

Ward, M. D., Beger, A., Cutler, J., Dickenson, M., Dorff, C., and Radford, B. (2013). Comparing GDELT and ICEWS event data. *Analysis*, 21:267–297.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM.

Yangarber, R., Etter, P. V., and Steinberger, R. (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.

Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3):6527–6535.