Discovering Black Lives Matter Events in the United States: Shared Task 3, CASE 2021

Salvatore Giorgi[1,2], **Vanni Zavarella**[3], **Hristo Tanev**[3], **Nicolas Stefanovitch**[3], **Sy Hwang**[1], **Hansi Hettiarachchi**[4], **Tharindu Ranasinghe**[5], **Vivek Kalyan**[6], **Paul Tan**[6], **Shaun Tan**[6], **Martin Andrews**[6], **Tiancheng Hu**[7], **Niklas Stoehr**[7], **Francesco Ignazio Re**[7], **Daniel Vegh**[7], **Dennis Atzenhofer**[7], **Brenda Curtis**[2], **Ali Hürriyetoğlu**[8]

[1]University of Pennsylvania, [2]National Institute on Drug Abuse,
[3]European Commission, [4]Birmingham City University, [5]University of Wolverhampton,
[6]Handshakes, [7]ETH Zurich, [8]Koc University

sgiorgi@sas.upenn.edu, ahurriyetoglu@ku.edu.tr

Abstract

Evaluating state-of-the-art event detection systems for determining the spatio-temporal distribution of the events on the ground is infrequently performed. Despite this, the ability to both (1) extract events "in the wild" from text and (2) properly evaluate event detection systems has potential to support a wide variety of tasks such as monitoring the activity of socio-political movements, examining media coverage and public support of these movements, and informing policy decisions. Given the global response to the murder of George Floyd, an unarmed Black man, at the hands of police officers, we study performance of event detection systems on detecting Black Lives Matter (BLM) events from tweets and news articles. This shared task asks participants to identify BLM related events from large unstructured data sources, using systems pretrained to extract socio-political events from text. We evaluate several metrics, assessing each system's ability to monitor the evolution of protest events both temporally and spatially. Results show that identifying daily protest counts is an easier task than classifying spatial and temporal protest trends simultaneously, with maximum performance of 0.745 (Spearman) and 0.210 (Pearson r), respectively. Additionally, all baselines and participant systems suffered from low recall (max 5.08), confirming the high impact of media sourcing in the modelling of protest movements.

1 Introduction

Typically, performance evaluations of automated event coding engines are carried out with respect to benchmarks made of annotated linguistic units (e.g. clause, sentence or document). While this is crucial in order to factorize the individual, linguistic subtasks composing the event extraction process, it does not estimate the overall usability of machine-coded event data sets for micro-level modelling of social processes, particularly in the domain of socio-political and armed conflict, where spatial analysis has become standard.

The complex dynamics of the Black Lives Matter movement and its varied media coverage by news outlets and social media make it a particularly relevant use case for assessing the capability of automated, Event Extraction systems to model socio-political processes. The "Discovering Black Lives Matter Events" task[1] organized in the context of the Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE) 2021 workshop aims at doing so by challenging Event Extraction (EE) engines to extract a collection of protest events from two heterogeneous text collections (i.e., news and social media). The task's evaluation is done by measuring a number of spatio-temporal correlation coefficients against a curated Gold Standard data set of protest incidents from the BLM movement.

During May and June of 2020, protests occurred across the globe in response to the murder of George Floyd, an unarmed Black man, by Derek Chauvin, a white police officer. In the U.S., the number of locations holding demonstrations related to this murder outnumbered any other demonstration in U.S. history (Putnam et al., 2020). These events were more often than not associated with the Black Lives Matter (BLM) movement, either (1) directly through organizing or (2) indirectly through the slogan "Black Lives Matter" or shared political agendas such as police abolition and protests against police violence towards Black communities. Since its inception in 2013, the Black Lives Matter movement, a loose network of affiliated organizations, has organized demonstrations around a large number of police shootings and killings and sought to raise awareness of systematic vio-

[1] https://github.com/emerging-welfare/case-2021-shared-task

lence against Black communities. While support for Black Lives Matter has varied over its lifetime (Horowitz, 2020), the work done over the past years laid the foundation for the global response seen in the wake of George Floyd's murder.

This task is the third in a series of tasks at CASE 2021 workshop (Hürriyetoğlu et al., 2021b). The first task is concerned with protest news detection at multiple text resolutions (e.g., the document and sentence level) and in multiple languages: English, Hindi, Portuguese, and Spanish (Hürriyetoğlu et al., 2021a). Teams which participated in Task 1 were invited to participate in this third task: "Discovering Black Lives Matter Events in the United States". This task is an evaluation only task, where all models are (1) trained on the data supplied in Task 1, (2) applied to the news and social media data (i.e, New York Times and Twitter data), and (3) evaluated on a manually curated, Gold Standard BLM protest event list. Each team's system is compared to simple baselines in order to properly evaluate their accuracy.

2 Related Work

Summary measures such as precision, recall, and F1 are limited in their capacity to inform about the quality of the predictions of an automated system (Derczynski, 2016; Yacouby and Axman, 2020). Moreover, evaluating capabilities of a system on detecting socio-political events from text requires additional metrics such as spatio-temporal correlation of the system output and the actual distribution of the events (Wang et al., 2016; Althaus et al., 2021).

Several studies focused on assessing the correlation of machine-coded event data sets with Gold Standards based on disaggregated event counts, for example Ward et al. (2013) and Schrodt and Analytics (2015). Hammond and Weidmann (2014) applied disaggregation of events incidents across PRIO-GRID geographical cells (Tollefsen et al., 2012) to assess the Global Database of Events, Language and Tone (GDELT) data approximation of the spatio-temporal pattern of conflicts. Zavarella et al. (2020) adapted this method to administrative units for measuring the impact of event de-duplication on increasing correlation with the Armed Conflict Location and Event Data (ACLED) data sets for a number of conflicts in Africa. In this report we report on an evaluation task, which we refer as Task 3, we provide a detailed analysis of the capabilities of the best performing systems on Task 1 (Hürriyetoğlu et al., 2021a) in this respect. We believe this effort will shed light on system performances beyond precision, recall, and F1.

3 Data

The goal of this task is to evaluate the performance of automatic event detection systems on modeling the spatial and temporal pattern of a social protest movement. We evaluate the capability of participant systems to reproduce a manually curated BLM-related protest event data set, by detecting BLM event reports, enriched with location and date attributes, from a news corpus collection, a Twitter collection, and from the union of the two.

3.1 Training Data

As a usability analysis, no training data were provided for this Task. Namely, the event definition applied for coding the reference event data set is the same as the one adopted for Shared Task 1 (Hürriyetoğlu et al., 2021a) and any data utilized for Task 1 and Task 2, such as the one from Hürriyetoğlu et al. (2021), or any additional data could be used to build a system/model run on the input data.

3.2 Input Data

We provide two types of input data. The first is a generic, not topic filtered collection of all news items (Title and Lead Paragraph) from the New York Times for the target time range May 25th - June 30th. The second is a collection of Black Lives Matter related tweets (Giorgi et al., 2020).

New York Times The New York Times (NYT) data sets consists of 5,347 articles published between May, 25 and June 30, 2020. The data associated with each article includes published date, print headline, lead paragraph, web URL, authors, and an abstract, among other meta-data. This is a general set of NYT articles (i.e., articles may or may not be related to BLM), unlike the Twitter data set which only contains tweets related to BLM or counter protests (e.g., All Lives Matter and Blue Lives Matter).

Twitter We used an open source data set of tweets containing keywords related to Black Lives Matter and the counter protests: All Lives Matter and Blue Lives Matter. While this data set contains tweets dating back to the origins of the Black

Lives Matter movement, the tweets used in this task are limited to the date range: May 25, 2020 (the date of George Floyd's murder) to June 30, 2020. These tweets were pulled in real time using the Twitter API's keyword matching with the following three keywords: *BlackLivesMatter*, *AllLivesMatter*, and *BlueLivesMatter*. This data set consists of 30,160,837 tweets. Participants were given full access to each tweet's meta-data (including the tweet's text), which could include URLs, location information, and dates.

3.3 Gold Standard Data

For the Gold Standard data (i.e., the BLM events list we wish to automatically detect) we considered two online sources of Black Lives Matter protest events: Creosote Maps [2] and Race and Policing[3]. Starting with these two data sets, we first checked if the source URL link was still active. If not, we referenced other data sets for the event in question: Wikipedia (a list of George Floyd protests in and outside of the U.S.) and the New York Times. If a valid article was not found matching this protest date and location, then we performed a Google search for the specific event. If still nothing was found, then the event was removed from the data set. If at any point, we discovered a valid URL for the event, we ran a validation check. This check asked: (1) is the source a tweet or Facebook post; (2) does the source describe an upcoming event; (3) is the source irrelevant to the protest at the location; (4) does the source have enough information; and (5) is the source not accessible because of a paywall. If the source passed this check, we then scraped the source for the publication date and days of the week in the article text. If the publication date and the day of the week *do not* match, we then inferred the date of the protest by the mention of the day of the week closest to the publication date. Finally, we manually checked the scraped or inferred dates and record this as the event date.

In the end, this produced 3,463 distinct U.S. events between May 25 and June 30, 2020 with date, city, and state information. Of these events, only 537 (approximately 15% of the events) occurred after the first week of June. To compensate for the lack of coverage across all of June, we used the open source data set from the The Crowd Counting Consortium (CCC)[4]. From our original data set of 3,463 events, 754 events also occurred in the CCC data, matching on (1) URL or (2) both date and city. We then combined the two data sets (i.e., the CCC events with our original list) and removed duplicates. This resulted in 7,976 protest events in our final Gold Standard data. The U.S. map in Figure 1 shows the spatial distribution of these events (yellow dots).

4 Evaluation

System performance is evaluated by computing correlation coefficients on event counts aggregated on cell-days, using uniform grid cells of approximately 55 kilometers sides from the PRIO-GRID data set (Tollefsen et al., 2012). We use these analytical measures as a proxy to the spatio-temporal pattern of the BLM protest movement.

4.1 Data Normalization

In order to be joined with PRIO-GRID shapefiles, string-like location information of system output data had to be normalized to coordinate pairs. To do this we used the OpenStreetMap Nominatim search API[5]. For structured location name representations (i.e., *city*, *state*, *country*) we used a parametric search, otherwise we used free-form query strings. We note that geographical coordinate conversion from Nominatim places the event at the geographical centroid of the polygon of the assigned administrative unit. In our evaluation, we discarded the system output event records with no source location information or whose string-like location attribute returned null results in Nominatim API.

4.2 Metrics

We use the cell-days counts for two different analysis: the correlation with the total daily "protest cell" counts (i.e., time trends alone) and the event counts for each cell-day (i.e., spatial and temporal trends together).

Temporal Trends The first analysis only considers the total number of "activated" cells (i.e., for which at least one Protest event was recorded), in the system output and Gold Standard data set. This time series analysis is sufficient to estimate how well the automatic systems capture the time trends of the protest movement. However, it does not

[2] https://www.creosotemaps.com/

[3] http://raceandpolicing.com/

[4] https://sites.google.com/view/crowdcountingconsortium/home

[5] https://nominatim.org/release-docs/develop/api/Search/#parameters

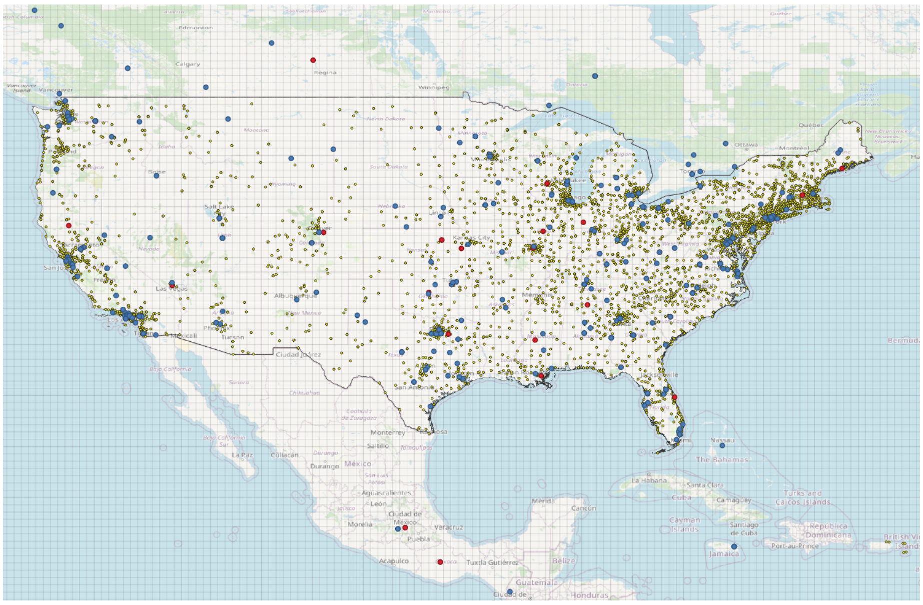

Figure 1: The geo-referenced BLM protest event records from Gold Standard (small yellow dots) overlaid with the PRIO-GRID cells over the US. The larger red and blue dots represent events recognized by the Baseline system from NYT and Twitter, respectively.

compute accuracy of system data in estimating the spatial variation of the target process.

Spatial and Temporal Trends To this purpose, we also measure the correlation coefficients on the absolute event counts with respect to Gold Standard, over each single cell-day.

For both analyses, we use two types of correlation coefficients to assess variable's relationship: Pearson coefficient r and Spearman's rank correlation coefficient ρ. Moreover, we used Root Mean Squared Error (RMSE) to measure the absolute value of the error on estimating cell/event counts from the Gold Standard.

4.3 Baseline

As a baseline, we used the output from NEXUS, a state-of-the-art engine for events detection from news (Tanev et al., 2008) that has been used in the area of security and disaster management[6]. We denote this system as *Baseline* throughout. Nexus is based on a blend of rule-based cascaded grammars for detection event slots (i.e. perpetrator, various types of affected people, infrastructure and vehicle targets and weapons used), and a combination of keyword-based and statistical classifiers for detection of event classes. The dictionaries underlying the extraction grammars of the system have been learned using weakly supervised lexical learning on generic news corpora (Tanev and Zavarella, 2014; Zavarella et al., 2014). No learning was performed on domain corpora in protest movements or related themes. Details on Nexus full taxonomy of event categories can be found in Atkinson et al. (2017). For this task, we filter the events belonging to the following type set: Disorder/Protest/Mutiny, Boycott/Strike, Public Demonstration, Riot/Turmoil, Sabotage/Impede, Mutiny. NEXUS performs event geocoding by (1) matching populated place names from the GeoNames gazetteer[7] in the news item; (2) resolving them into unique location entities via disambiguation heuristics (Pouliquen et al., 2006); and (3) selecting a single main event location based on the text proximity with the matched event components (see the slots above) in the news article. In order to mitigate the lack of geographical context in the tweet body, when processing the Twitter data, we ran Nexus on an enriched text, which included the String value of the *full_name* field in the *Place* child object of the tweet, whenever that was available[8]. This resulted in a small fraction of 32,085 tweets with geographical information (out of the roughly 30 million tweets originally sam-

[6] A spin-off of the NEXUS system is the Medical NEXUS, an event detection system for disease outbreaks and food poisoning (Linge et al., 2012)

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

[8] `https://developer.twitter.com/en/docs/twitter-api/data-dictionary/object-model/tweet`

pled). For the sake of comparison, we shared with participants this subset of tweets, together with the assigned location.

4.4 Nexus Deduplication

This system, developed by the Task organizers and denoted *NexusDdpl*, is an extension of the Baseline system, where an event deduplication has been integrated as a post-processing module. The algorithm uses two metrics based on geographical distance between two event points and semantic distance, respectively. The semantic distance is computed using the cosine between the projections of the sentence embeddings of the texts of the events records. The LASER embeddings (Schwenk and Douze, 2017) were used for that purpose. Twitter data has been cleaned of hashtags, URLs, and accounts names, as these have a negative impact on the semantic similarity measure. In order to be considered duplicate two events must have both distance measures under a fixed threshold, which were set to 2km for spatial distance, 0.20 for semantic distance on NYT data, 0.30 for semantic distance on Twitter data. The reason of these different threshold depending on the data sets is that Twitter data are noisier than NYT data, with higher variations in text size and style when describing a single event. As such looser threshold was required. When applying on the combination of both data sets, we use a compromise threshold of 0.35 was used.

4.5 Team Systems

Four teams participated in this event: *DaDeFrNi*, *EventMiner*, *Handshakes*, and *NoConflict*. We briefly describe the systems below and ask the reader to refer to their systems papers for additional details.

DaDeFrNi This team considered two slightly different procedures for this task. For the NYT data set, they first extracted geo-entities from each article using the Python library geography, which was used to classify each entity in one of the three categories "city", "country", and "region". For the cases where an article contained the name of a city but did not provide any region or country reference, *DaDeFrNi* retrieved the necessary information by checking the city name against a worldwide cities database. When the name of a city was associated with several locations, we filtered the city with the highest population, along with its corresponding "region" and "country". For the Twitter data set, given the large size of the data, the above procedure was computationally expensive. Thus, the Python library spaCy (Honnibal et al., 2020) for retrieving NER/GPE entities, given its much smaller computational cost. The complete system details can be found in Ignazio Re et al. (2021).

EventMiner Team *EventMiner*'s approach for Task 3 is mainly based on transformer models (Hettiarachchi et al., 2021). This approach involved three steps: (1) event document identification, (2) location detail extraction, (3) and event filtering to identify the spatial and temporal pattern of the targeted social protest movement. Event documents are identified using the winning solution submitted to CASE 2021 Task 1-Subtask 1: event document classification (Hettiarachchi et al., 2021). Next, the location details in event described tweets are extracted. Since this team only focused on the Twitter corpus, they used tweet metadata to extract location details. However, since the majority of the tweets are not geotagged and to extract the location details mentioned in the text, they used a NER approach too. For NER, a transformer model is fine-tuned for token classification using the data set released with the WNUT 2017 Shared Task on Novel and Emerging Entity Recognition (Derczynski et al., 2017). The BERTweet model is used since it is pretrained on Tweets (Nguyen et al., 2020). To convert the location details into an unique format and fill the missing details (e.g. region, country), locations are geocoded using the GeoPy library[9]. For the final step, event tweets with location details are grouped based on their created dates and locations and removed the groups with fewer tweets assuming that important events generate a high number of tweets. Three systems were submitted. For the first system, denoted by †, only the new events are included (i.e., events with locations which are identified in the previous day are removed). The second system ††, includes all the extracted events (i.e., no filtering as in †). Finally, the third system ††† further filters the events from † to include U.S. events only. Please see Hettiarachchi et al. (2021) for more details

Handshakes This model is a pretrained XLM-RoBERTa model, fine-tuned on the multi-language article data from Task 1 Subtask 1 and sentence data from Subtask 2, with a classification head that predicts if the input text is a protest or not. We make use of the provided location data in the

[9] https://geopy.readthedocs.io

data sets, where available. Please see Kalyan et al. (2021) for further details.

NoConflict Team *NoConflict* used their model of protest event sentence classification from the winning submission of the English version of Task 1 Subtask 2. Their model is based on a RoBERTa (Liu et al., 2019) backbone with a second pretraining (Gururangan et al., 2020) stage done on the POLUSA (Gebhard and Hamborg, 2020) data set before finetuned on Subtask 2 data. For the NYT data set, they first filtered the articles based on the section name. They then ran their model on the abstract of each article to identify ones containing protest events. For each remaining article, they run a transformer-based (Vaswani et al., 2017) named entity recognition from spaCy (Honnibal et al., 2020) to identify the location and date of the events. They covert the location to absolute location using the Geocoder library and convert the date of the event to the absolute date based on the article's publication date. If the relative location or date is unavailable, they default to those included in the metadata. The event sentence classification system details can be found in Hu and Stoehr (2021). Three systems were submitted for the NYT data, denoted $\diamond$, $\diamond\diamond$, and $\diamond \diamond \diamond$. Each system used a set of manually curated keywords applied to different parts of each data point. Theses rules are included in the Appendix. For the Twitter data set, Team *NoConflict* ran their model on the full text of each tweet to identify protest events. For each potential event tweet, they identify the location and time based on the metadata of the tweet itself and the main tweet if it is a retweet.

5 Results

Table 1 shows the Pearson *r*, Spearman correlation coefficient ρ, and Root Mean Squared Error (RMSE) for the total daily protest cell counts of the Baseline and participant systems, over the 35 days target time range. When a run for both source types exists for a system, we also evaluate the union of the two event sets (noted as "Merged" in Tables). Here, the correlations are between the total number of cells per day where the system found an event vs. the number of cells where event happened according to the Gold Standard (i.e., temporal patterns and not spatial patterns). These correlation measures are tolerant to errors in geocoding (as far as the events are located in U.S.) and evaluate the capability of the system to detect protest events in the news and social media, independent of their location. We see the following: (1) *NoConflict* surpasses the *Baseline* with the NYT, Twitter, and Merged data in both Pearson r and Spearman ρ, and (2) *EventMiner* and *HandShakes* surpasses *Baseline* with Twitter data in Pearson r (both systems have lower Spearman ρ than *Baseline*). Additionally, *NoConflict* surpasses the *NexusDdpl* system (using NYT, Twitter, and Merged data), and the *HandShakes* system surpasses the *NexusDdpl* system using Twitter data.

	Data	r	ρ	RMSE
Baseline	NYT	0.646	0.626	301.98
	Twitter	0.337	0.367	291.01
	Merged	0.353	0.334	288.04
NexusDdpl	NYT	0.646	0.626	301.98
	Twitter	0.337	0.367	291.01
	Merged	0.357	0.334	287.85
DaDeFrNi	NYT	-0.366	-0.264	287.04
	Twitter	-0.202	-0.280	306.77
	Merged	-0.408	-0.365	287.26
EventMiner	Twitter†	0.451	0.327	300.15
	Twitter††	0.427	0.312	299.59
	Twitter†††	0.453	0.343	300.83
HandShakes	Twitter	0.424	0.254	**276.13**
NoConflict	NYT$^{\diamond}$	0.725	0.669	302.14
	NYT$^{\diamond\diamond}$	**0.745**	**0.762**	302.96
	NYT$^{\diamond\diamond\diamond}$	0.601	0.658	303.407
	Twitter	0.534	0.524	287.88
	Merged	0.522	0.537	286.59

Table 1: Correlation coefficients and error rates for daily protest cell counts: r represents Pearson correlation coefficient, ρ is Spearman's rank correlation coefficient, and RMSE is the Root Mean Squared Error computed on day-cell units. Superscripts refer to the various systems submitted by *EventMiner* and *NoConflict*, as described in Section 4.5.

Table 2 reports Pearson *r*, Spearman correlation coefficient ρ, and Root Mean Squared Error (RMSE) over cell-day event counts of the Baseline and participant systems with respect to Gold Standard, for the 35 days time range. Here the variables range over the whole set of PRIO-GRID cells included in the US territory and, thus, shows the correlation of event numbers across geo-cells, thus evaluating the system's geolocation capabilities. *NoConflict* (NYT$^{\diamond}$) had the highest Pearson r and lowest RMSE across all systems, as well as the highest Spearman ρ (with the Merged data). Using Twitter data alone, the *Baseline* and *NexusDdpl* systems outperformed all others in terms of Pearson r, however *NexusDdpl* had a higher Spearman ρ. However, when looking at both correlation metrics simultaneously, no system is above the *NexusDdpl* baseline.

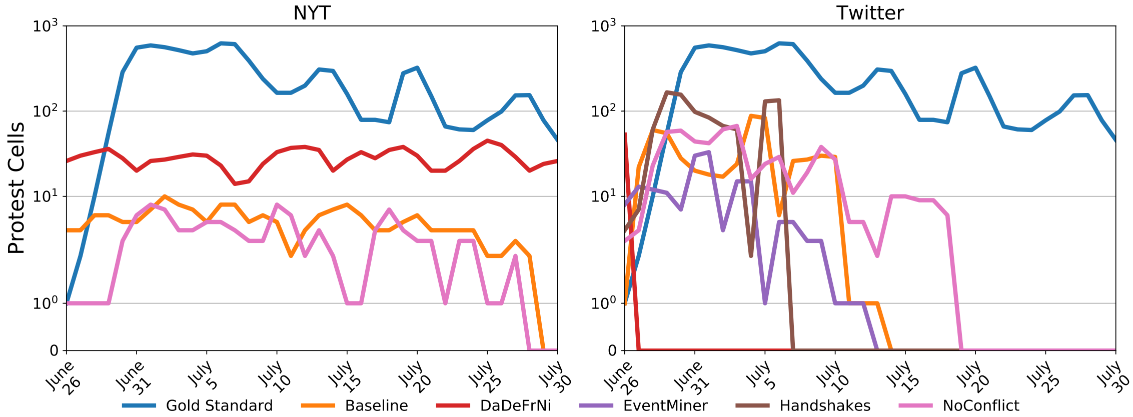

Figure 2: Time series of total daily protest cells from the Gold Standard (in blue), against system runs on New York Times (left) and Twitter (right) input data. Protest cell counts are on a log scale. *Baseline* and *NexusDdpl* systems produce the same cell count numbers (see Table 2), so the *NexusDdpl* system was omitted.

In Figure 2 we plot the time series of total daily protest cells for the best performing instance of each system on New York Times (left) and Twitter (right) data, respectively. We see the systems evaluated on the NYT data failing to pick up both variation in the temporal patterns (i.e., a large number of protests early in late May and early June, which gradually declines with weekly spikes) and the magnitude of the events (i.e, most systems pick up less than 100 events per day). Systems evaluated on Twitter data pick up more events in late May and early June, but still fail to pick up the magnitude of the events.

A more lenient representation of the agreement with Gold Standard is shown in Table 3. Here we report the confusion matrix between grid cells that Gold Standard and system runs code as experiencing at least a protest event. It can be observed that only few of the cells classified as Protest by Gold Standard are detected by the automatic systems, which on the other hand incorrectly classified as Protest several additional cells.

	Data	r	ρ	RMSE
Baseline	NYT	0.096	0.089	0.732
	Twitter	0.171	0.127	0.785
	Merged	0.181	0.132	0.724
NexusDdpl	NYT	0.100	0.088	0.725
	Twitter	0.193	0.124	0.777
	Merged	0.192	0.129	0.715
DaDeFrNi	NYT	0.165	0.136	0.711
	Twitter	0.002	-0.004	69.171
	Merged	0.003	0.122	87.422
EventMiner	Twitter†	0.155	0.077	0.715
	Twitter††	0.147	0.077	0.715
	Twitter†††	0.157	0.076	0.715
HandShakes	Twitter	0.109	0.105	0.783
NoConflict	NYT$^{\diamond}$	**0.210**	0.095	**0.712**
	NYT$^{\diamond\diamond}$	0.196	0.086	0.714
	NYT$^{\diamond\diamond\diamond}$	0.184	0.082	0.715
	Twitter	0.020	0.138	148.18
	Merged	0.018	**0.145**	148.20

Table 2: Correlation coefficients and error rates for *cell-day* event counts of the Baseline and participant systems with respect to Gold Standard. Superscripts refer to the various systems submitted by *EventMiner* and *NoConflict*, as described in Section 4.5.

6 Conclusions

The goal of the "Discovering Black Lives Matter Events" Shared Task was to explore novel performance evaluations of pretrained event detection systems. These systems were applied to large noisy, heterogeneous text data sets (i.e., news articles and social media data) related to a specific protest movement, namely, Black Lives Matter. Thus, the systems are being evaluated out-of-domain in terms of both data type (i.e., the systems are trained on news data and evaluated on both news and social media) and protest movement context (i.e., the training data are not necessarily related to BLM). Systems are evaluated in their ability to identify both events across time as well as their distribution across space. This evaluation scenario proved difficult for all systems participating in the shared task. A major problem, as shown on Table 3, is the system's low recall. No system was able to outperform the *NexusDdpl* baseline both in precision and recall together. The only system which outperformed the baseline in either recall or F1 is the *DaDeFrNi* (Ignazio Re et al., 2021), with a recall of 5.08 and F1 of 8.86. On the other hand, two systems surpass the baseline in precision: *EventMiner* (Hettiarachchi et al., 2021) and *NoConflict* (Hu and Stoehr, 2021),

		Gold Standard		Precision	Recall	F1
		true	false			
Baseline	true	330	341	49.2	3.87	7.20
	false	8163	195790			
NexusDdpl	true	326	353	48.0	3.84	7.11
	false	8167	195778			
DaDeFrNi	true	431	802	35.0	**5.08**	**8.86**
	false	8062	195329			
EventMiner†††	true	94	74	56.0	1.11	2.17
	false	8399	196057			
Handshakes	true	328	631	34.2	3.86	6.94
	false	8165	195500			
NoConflict$^{\diamond\diamond\diamond}$	true	81	29	**73.6**	0.95	1.88
	false	8412	196102			

Table 3: Confusion matrix of grid cells experiencing at least one Protest event (true) versus inactive cells (false), for the Gold Standard, Baseline and participant systems. Unless denoted by a superscript, all systems use the "merged" version (i.e., both NYT and Twitter data sets) except for *HandShakes* system which uses only Twitter data.

with precisions of 56.0 and 73.6, respectively.

The low recall at this years shared task may well be due to the low coverage of protest events of the highly diffused BLM movement both in the NYT and Twitter corpus, so the upper bound of the recall may turn out not to be much higher than the system performance. One possible explanation for this is that a significant part of the BLM events in the Gold standard are located in small towns, for which NYT has a limited coverage and also they were not in the focus of social media, due to their small scale. *NexusDdpl* turned out to be quite high both in terms of event detection accuracy, as well as geo-coding correlation. While no single system outperformed all others in tracking both temporal and spatial trends, *NoConflict* had a clear advantage (i.e., the highest scoring system in 2 out of 3 metrics) in terms of tracking daily events.

Acknowledgments

The author from Koc University was funded by the European Research Council (ERC) Starting Grant 714868 awarded to Dr. Erdem Yörük for his project Emerging Welfare. The authors from the National Institute on Drug Abuse were supported in part by the Intramural Research Program of the NIH, National Institute on Drug Abuse (NIDA).

References

Scott Althaus, Buddy Peyton, and Dan Shalmon. 2021. A total error approach for validating event data. *American Behavioral Scientist*, 3(2).

Martin Atkinson, Jakub Piskorski, Hristo Tanev, and Vanni Zavarella. 2017. On the creation of a security-related event corpus. In *Proceedings of the Events and Stories in the News Workshop*, pages 59–65.

Leon Derczynski. 2016. Complementarity, F-score, and NLP evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 261–266, Portorož, Slovenia. European Language Resources Association (ELRA).

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Lukas Gebhard and Felix Hamborg. 2020. The polusa dataset: 0.9 m political news articles balanced by time and outlet popularity. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468.

Salvatore Giorgi, Sharath Chandra Guntuku, Muhammad Rahman, McKenzie Himelein-Wachowiak, Amy Kwarteng, and Brenda Curtis. 2020. Twitter corpus of the #blacklivesmatter movement and counter protests: 2013 to 2020. *arXiv preprint arXiv:2009.00596*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Jesse Hammond and Nils B Weidmann. 2014. Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2):2053168014539924.

Hansi Hettiarachchi, Mariam Adedoyin-Olowe, Jagdev Bhogal, and Mohamed Medhat Gaber. 2021. DAAI at CASE 2021 task 1: Transformer-based multilingual socio-political and crisis event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Juliana Horowitz. 2020. *Amid protests, majorities across racial and ethnic groups express support for the Black Lives Matter movement*. Pew Research Center.

Tiancheng Hu and Niklas Stoehr. 2021. Team noconflict at case 2021 task 1: Pretraining for sentence-level protest event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Ali Hürriyetoğlu, Osman Mutlu, Farhana Ferdousi Liza, Erdem Yörük, Ritesh Kumar, and Shyam Ratan. 2021a. Multilingual protest news detection - Shared Task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

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

Ali Hürriyetoğlu, Erdem Yörük, Osman Mutlu, Fırat Duruşşan, Çağrı Yoltar, Deniz Yüret, and Burak Gürel. 2021. Cross-Context News Corpus for Protest Event-Related Knowledge Base Construction. *Data Intelligence*, 3(2):308–335.

Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, and Niklas Stoehr. 2021. Team dadefrni at case 2021 task 1: Document and sentence classification for protest event detection. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Vivek Kalyan, Paul Tan, Shaun Tan, and Martin Andrews. 2021. Handshakes ai research at case 2021 task 1: Exploring different approaches for multilingual tasks. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, online. Association for Computational Linguistics (ACL).

Jens P Linge, Marco Verile, Hristo Tanev, Vanni Zavarella, Flavio Fuart, and Erik van der Goot. 2012. Media monitoring of public health threats with medisys. *C. WILLIAM, CWR. WEB-STER, D. BALAHUR, et al*, pages 17–31.

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

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, et al. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *arXiv preprint cs/0609065*.

Lara Putnam, Erica Chenoweth, and Jeremy Pressman. 2020. The floyd protests are the broadest in us history—and are spreading to white, small-town america. *Washington Post*, 6.

Philip A Schrodt and Parus Analytics. 2015. Comparing methods for generating large scale political event data sets. In *Text as Data meetings, New York University, 16–17, 2015*, pages 1–32.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *arXiv preprint arXiv:1704.04154*.

Hristo Tanev, Jakub Piskorski, and Martin Atkinson. 2008. Real-time news event extraction for global crisis monitoring. In *Natural Language and Information Systems*, pages 207–218, Berlin, Heidelberg. Springer Berlin Heidelberg.

Hristo Tanev and Vanni Zavarella. 2014. Multilingual lexicalisation and population of event ontologies: A case study for social media. In *Towards the Multilingual Semantic Web*, pages 259–274. Springer.

Andreas Forø Tollefsen, Håvard Strand, and Halvard Buhaug. 2012. Prio-grid: A unified spatial data structure. *Journal of Peace Research*, 49(2):363–374.

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 *Advances in Neural Information Processing Systems*.

Wei Wang, Ryan Kennedy, David Lazer, and Naren Ramakrishnan. 2016. Growing pains for global monitoring of societal events. *Science*, 353(6307):1502–1503.

Michael D Ward, Andreas Beger, Josh Cutler, Matthew Dickenson, Cassy Dorff, and Ben Radford. 2013. Comparing gdelt and icews event data. *Event Data Analysis*, 21(1):267–297.

Reda Yacouby and Dustin Axman. 2020. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 79–91, Online. Association for Computational Linguistics.

Vanni Zavarella, Jakub Piskorski, Camelia Ignat, Hristo Tanev, and Martin Atkinson. 2020. Mastering the media hype: Methods for deduplication of conflict events from news reports. In *Proceedings of AI4Narratives — Workshop on Artificial Intelligence for Narratives*.

Vanni Zavarella, Hristo Tanev, Ralf Steinberger, and Erik Van der Goot. 2014. An ontology-based approach to social media mining for crisis management. In *SSA-SMILE@ ESWC*, pages 55–66. Citeseer.

A Additional System Details

The *NoConflict* team produced three separate rule-based systems for the NYT data. NYT$^{\diamond}$: include keywords “Police Brutality, Misconduct and Shootings”, “Attacks on Police”, “George Floyd Protests (2020)”, “Demonstrations, Protests and Riots”, “Black Lives Matter Movement”; excluded keywords: “Hong Kong Protests (2019)”; include section name: “U.S.”, “Politics”, “New York”, “World”; exclude News Desk: “Arts & Leisure”, “Gender”, “Investigative”, “Special Sections”, “Sports”, “Science”, “Magazine”, “Video”, “Podcast”, “News Desk”; exclude if present in abstract or lead paragraph: “Hong Kong”. NYT$^{\diamond\diamond}$: include keywords: “Police Brutality, Misconduct and Shootings”, “Attacks on Police”, “George Floyd Protests (2020)”, “Demonstrations, Protests and Riots”, “Black Lives Matter Movement”; exclude keywords: “Hong Kong Protests (2019)”; include section name: “U.S.”, “Politics”, “New York”, “World”; exclude News Desk: “Arts & Leisure”, “Gender”, “Investigative”, “Special Sections”, “Sports”, “Science”, “Magazine”, “Video”, “Podcast”, “News Desk”, “Washington”, “Politics”; exclude if present in abstract or lead paragraph: “Hong Kong”. NYT$^{\diamond\diamond\diamond}$: include keywords: “Police Brutality, Misconduct and Shootings”, “Attacks on Police”, “George Floyd Protests (2020)”, “Demonstrations, Protests and Riots”, “Black Lives Matter Movement”; exclude keywords: “Coronavirus (2019-nCoV)”, “Quarantines”, “Hong Kong Protests (2019)”; include section name: “U.S.”, “Politics”, “New York”; exclude News Desk: “Arts & Leisure”, “Gender”, “Investigative”, “Special Sections”, “Sports”, “Science”, “Magazine”, “Video”, “Podcast”, “News Desk”, “Washington”, “Politics”, “Foreign”; exclude if present in abstract or lead paragraph: “Hong Kong”.