# A Multi-instance Learning Approach to
# Civil Unrest Event Detection using Twitter

**Alexandra DeLucia**[*], **Mark Dredze**[*], **Anna L Buczak**[†]

[*]Center for Language and Speech Processing, Johns Hopkins University
[†]Johns Hopkins University Applied Physics Laboratory
{aadelucia, mdredze}@jhu.edu, Anna.Buczak@jhuapl.edu

## Abstract

Social media has become an established platform for people to organize and take offline actions, often in the form of *civil unrest*. Understanding these events can help support pro-democratic movements. The primary method to detect these events on Twitter relies on aggregating many tweets, but this includes many that are not relevant to the task. We propose a multi-instance learning (MIL) approach, which jointly identifies relevant tweets and detects civil unrest events. We demonstrate that MIL improves civil unrest detection over methods based on simple aggregation. Our best model achieves a 0.73 F1 on the Global Civil Unrest on Twitter (G-CUT) dataset.

https://github.com/AADeLucia/
MIL-civil-unrest

## 1 Introduction

Social media has become an established platform for people around the world to share opinions and react to socio-political events. Platforms enable communication between like-minded individuals and can facilitate offline action. These actions can take the form of democratic expression, such as protests or other types of *civil unrest*. In the right situation, these events can lead to pro-democracy actions and lead to political changes towards more free and open governments and societies. Researchers who study these political movements often turn to social media platforms, especially the globally used Twitter,[1] to facilitate an understanding of how these events develop (Smidi and Shahin, 2017; Soengas-Pérez, 2013; Steinert-Threlkeld, 2017), which in turn may be used to study pro-democratic movements.

As part of that research program, different efforts have considered how to detect or forecast the start of the spread of these movements, including answering what will happen when. Detection and forecasting models identify civil unrest either at the macro- (global or country) (Muthiah et al., 2015; Islam et al., 2020) or micro- (city) level (Alsaedi et al., 2017; Giorgi et al., 2021) using one or multiple data sources, like social media, news, or economic indicators. Others study event extraction, whose goal is to extract information about an ongoing or recent event, such as where it happened and who was involved.

One of the challenges of developing models for detecting civil unrest events – is there an ongoing event? – is developing models responsive to rapid on the ground changes. Social media provides a mechanism for rapid detection; messages can be collected and analyzed as the event unfolds. Several studies have examined how Twitter can be utilized in a civil unrest detection model (Chinta et al., 2021; Islam et al., 2020; Muthiah et al., 2015). While there are a wealth of tweets from all over the world at any given time, not all tweets from a given location are relevant to an event. Filtering and identifying relevant tweets remains a challenging problem (Sech et al., 2020; Mishler et al., 2017; Rogers et al., 2019; Zhang and Pan, 2019).[2] Given the goal of detecting *any* event, it is important for a detection method to work even without knowing which tweets are necessarily relevant.

We follow Wang et al. (2016) and propose a multi-instance learning (MIL) approach to detecting civil unrest events at the country-level using Twitter data. In MIL, examples are grouped and labeled as a group instead of individually (i.e., weak supervision). Instead of aggregating all tweets from a given country within a specified time period, we utilize the MIL formulation where at least one tweet is relevant while most are not to predict

---

[1]We discuss recent access changes to Twitter API in Section 8.

[2]Rogers et al. (2019) used data from a Russian social media site (VKontakte) and Zhang and Pan (2019) used data from a Chinese site (Sina Weibo).

an event. We learn a tweet-level representation using a BERT-style model and then group these representations (i.e., *instances*) into a single *bag* (country/time period, in our case a single day).

We apply this method to the Global Civil Unrest on Twitter (G-CUT) dataset (Chinta et al., 2021),[3] which contains 200 million English tweets from 2014–2019 from 42 countries in Africa, the Middle East, and Southeast Asia. We focus on English tweets to take advantage of this large dataset and for easier analysis without the need for translation. Following Chinta et al. (2021), we use the Riots and Protests labels at the day level for a country from the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010) as ground truth, where a bag is positive if at least one event occurred in that country on that day.

We show that providing the model with all tweets (i.e., not filtering for civil unrest) and allowing it to choose relevant information leads to improved performance on detecting that an event occurred on a specific day in a country, as measured by F1. We support these results with an analysis showing the key tweets identified by the model during prediction.

In summary, we contribute the following:

- A trained MIL model for civil unrest detection on Twitter that achieves 0.73 F1 on G-CUT.
- Variations of the MIL model with varying levels of bag- and instance-level information.
- Analysis of example tweets identified as important for the model prediction.

## 2   Related Work

Work in civil unrest analysis, also referred to as socio-political event analysis, focuses on event characterization (Scharf et al., 2021), protest detection and forecasting (Hürriyetoğlu, 2021; Hürriyetoğlu et al., 2022), and event extraction (You et al., 2022; Mehta et al., 2022). Prior work on protest detection on the macro level (global or country) is most relevant to our task of country-level civil unrest detection. Existing work differs with regard to source data, event ground truth, and methods.

Most prior work uses news, social media, or economic indicators as features, or sometimes, a combination for a fuller picture. For news data, the goal is often to identify whether an article is

discussing a protest or event, as opposed to identifying whether an event is occurring at a location of interest. Wang et al. (2016) use an MIL framework to predict whether a news article is discussing a protest, with sentences considered as *instances* and news articles as *bags*. They use the sentences that were most informative for the article prediction for further analysis in event extraction. Our approach is inspired by this work; we describe it in Section 3.2. This setup of classifying the overall articles and the sentences within them was a part of the Multilingual Protest News Detection CASE 2021 shared task (Hürriyetoğlu et al., 2021).

Similar to sentence-level event identification, prior work has trained models for social media post-level identification of civil unrest discussion (Sech et al., 2020). While not their final goal, Islam et al. (2020) and Alsaedi et al. (2017) incorporate tweet-level models as filters for whether to include a tweet in future steps. Their tweet labels were gathered from manual annotation. We omit this filtration step since the proposed MIL approach can handle irrelevant tweets. This is an important note since our dataset was not collected with event-specific tweets as in Alsaedi et al. (2017).[4]

Alsaedi et al. (2017) leverage tweets discussing the London Riots (2011) to predict micro (small-scale) events like fires, car accidents, and assaults identified by police records through tweet filtering, clustering, and then automatically selecting a tweet as an event summary. Our model allows for a flexible number of posts to be provided as an explanation for each prediction. They also evaluate their system on larger-scale events across the Middle East in 2015, clustering with a variety of features like hashtags, sentiment, time, and location, if provided. Thus, predictions are on an event-level and not exactly a country-level as ours is (i.e., using only information from a single country).

Zheng and Sun (2019) also extract event-related keywords from clustered tweets, but cluster tweets in an online active-learning MIL setting. This MIL formulation differs from ours because their "bags" are clusters of similar tweets that are hand-labeled, whereas our bags represent a day in a country. They use a strict formulation where a cluster is predicted as positive (i.e., event) if it has at least one positive tweet (i.e., discussing event/unrest), and negative if there are no positive tweets. This assumption

---

[4] The London Riots dataset was collected with a list of event-related hashtags such as #tottenhamshooting and #UKRiots.

does not work for our setting, where we expect unrest-related tweets even on days where no event occurred.

Similar to our goal of detecting unrest on the country-level while maintaining tweets individuality, Islam et al. (2020) learn weights on individual tweets for a location that are updated in a temporally streaming fashion. Their tweet weights are based on a civil unrest dictionary of terms that correspond to different unrest "stages" (observe, agitation, mobilization, organization, occurrence), along with other temporal and spatial information. Our method does not rely on keyword dictionaries and instead, we update embeddings to obtain tweet weights for a given country-day. They also include a spatial component that incorporates information of nearby events. For ground truth they used the Global Data on Events, Location, and Tone (GDELT) database (Leetaru and Schrodt, 2013) and restricted to events with high coverage. In this work we follow other work and use labels from ACLED (Chinta et al., 2021; Zavarella et al., 2022).

Other systems combine multiple streams of information, such as news and Twitter data (Muthiah et al., 2015; Ramakrishnan et al., 2014; Giorgi et al., 2021) and news and macro-socio political indicators like GDELT and Worldwide Governance Indicators (WGI) (Buczak et al., 2022). We focus on Twitter because prior work has shown strong civil unrest indicators in past events such as the Arab Spring (Smidi and Shahin, 2017; Soengas-Pérez, 2013).

There is also work on civil unrest prediction in the fields of social and political science. Goldstone et al. (2010) built a model that incorporates a variety of socioeconomic and political indicators (e.g., infant mortality rate, stability of neighboring countries, and type of government) to predict whether a country in a given year would experience a large-scale event like a regime change or genocide. Similarly, Chenoweth and Ulfelder (2015) use structural condition theories (e.g., political opportunity) to predict large-scale non-violent events in a country-year. While Goldstone et al. (2010) and Chenoweth and Ulfelder (2015) achieved impressive 80% accuracy and 0.75 AUC, respectively, curating their rich political and socioeconomic country profiles requires a large amount of domain knowledge compared to our Twitter-centric approach.

# 3 Multi-Instance Learning for Detecting Civil Unrest

Our task is to identify days on which, in a given country, there is a civil unrest event based on Twitter data. For each country and day, we acquire a large number of tweets potentially relevant to an event. A model examines this data and predicts whether or not an event is taking place. Instead of aggregating the tweets, we propose a multi-instance learning approach that considers whether or not tweets are relevant. Specifically, the model assumes that on a day in which an event occurs, only a subset of the provided tweets are relevant to the event. This framing supports explainable predictions, where the tweets deemed relevant by the model can be examined for further context.

## 3.1 Multi-instance Learning

Multi-instance learning (MIL) is a form of weak supervision wherein individual examples, or *instances*, are grouped in a *bag* and are labeled at the bag level. MIL can be useful for large datasets where labeling individual instances is time-consuming, or for problems where a single label is associated with a set of samples. For example, a task may be to identify if a newspaper contains a fashion section. A newspaper would be represented as a bag, and individual articles as instances. If a newspaper has a fashion section, we assume at least one instance (article) is about fashion; otherwise, no articles are about fashion. Alternatively, in content-based image retrieval, images are segmented and each segment is analyzed individually, and the image is classified based on the contained objects in the instances (segments) (Carbonneau et al., 2018).

In the case of civil unrest detection from tweets, we assume that if an event takes place, then at least one tweet (and likely many) will discuss that event, while many will not. If no event takes place, no tweets discuss an event. In our work, each tweet is an instance, and all tweets from a single country on one day constitute a bag. A positive instance is a tweet that discusses civil unrest (e.g., expressing dissatisfaction or describing a protest in real-time) and a *positive bag* is a day and country where a protest occurred.

There are two different assumptions in MIL. The *standard assumption* assumes that every positive bag contains at least one positive instance and for a negative bag to only contain negative instances.
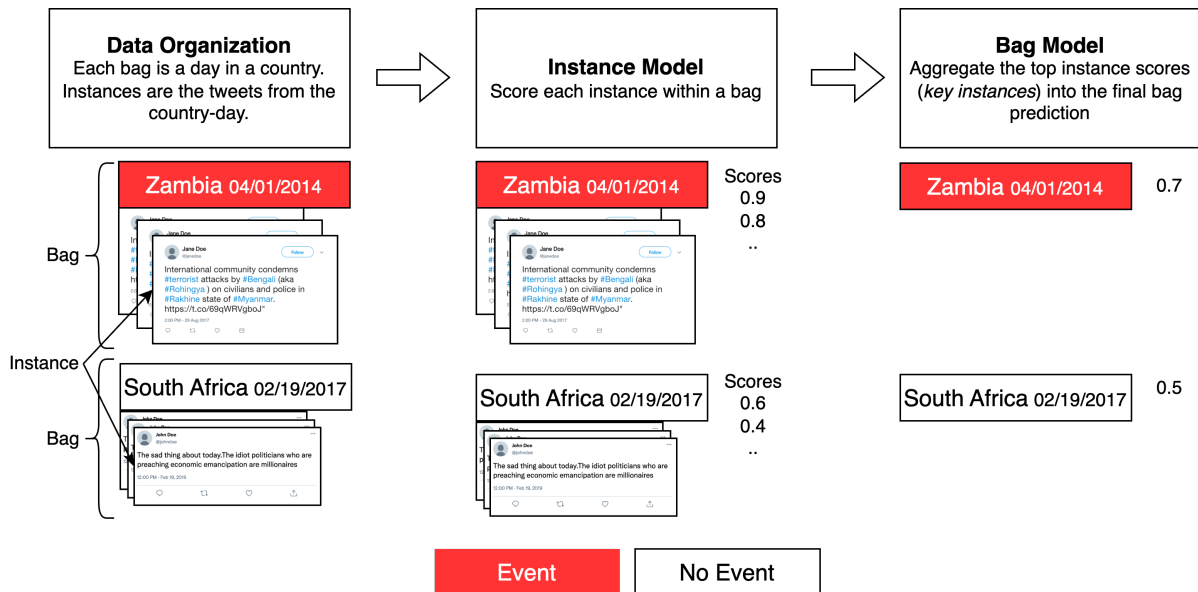
Figure 1: Overview of the proposed multi-instance learning (MIL) approach for civil unrest detection. We follow the country-day groupings and labels from the Global Civil Unrest on Twitter (G-CUT) dataset (Chinta et al., 2021).

This assumption is overly strict for our purposes, so we follow the relaxed *collective assumption*, where bags can contain some level of instances from the other class (Carbonneau et al., 2018). This assumption is a better fit since there can be tweets expressing dissatisfaction or discussing protests on non-event days.

There are multiple ways to combine instance-level features and scores at the bag level, such as averaging the instance-level scores, considering only the top-$k$ instances, or using the max score. In our task, this flexibility can help overcome a weak signal of positive instances in positive bags. Furthermore, MIL identifies instances that were most influential in the final bag prediction, known as "key" instances. While there is a trade-off between optimizing a model for bag classification and instance classification (Vanwinckelen et al., 2016), there is work that uses identified key instances for downstream tasks, such as bag summarization (Wang et al., 2016).

### 3.2 MIL for Twitter

Our MIL-based model for Twitter data is based on Wang et al. (2016), but we replace sentences and news articles with tweets and sets of tweets. Consider a collection of tweets $\mathcal{D} = \{(x_1, y_1), \ldots, (x_i, y_i), \ldots\}$, where a bag $x_i = \{(x_i^1, y_i^1), \ldots, (x_i^j, y_i^j), \ldots\}$ contains all tweets (indexed by $j$) from a single day in a country (country-day for brevity). We aim to find informative key instances that predict the bag-level class of civil unrest (protest for brevity): no protest or protest, $y_i \in \{0, 1\}$. Figure 1 shows an overview of our model.

**Instance Model**  We represent individual tweets with BERTweet (Nguyen et al., 2020), a BERT model pretrained on English tweets. In order to better represent civil-unrest related tweets, we fine-tuned BERTweet on the Civil Unrest on Twitter (CUT) dataset (Sech et al., 2020) with the HuggingFace Trainer. The model has a macro-F1 score of 0.82. More training details are in Appendix B.1. This trained model is the **instance classification model**, where the score (probability) of an instance is $p(y_i^j = 1) = \sigma(\theta x_i^j)$. The instance-level scores are not incorporated in the standard MIL model, but the representations are fine-tuned starting from the instance model representations.

**Bag Model**  The score of bag $x_i$ is determined by aggregating the instance scores. We use the average of the top $K_i$ instances in the bag:

$$p(y_i = 1) = \frac{1}{K_i} \sum_{j < K_i} \sigma(\theta x_i^j) \qquad (1)$$

The number of top instances for bag $x_i$ ($K_i$) is chosen by hyperparameter $0 \leq \eta \leq 1$ so that $K_i \triangleq \max(1, \lfloor |x_i| \times \eta \rfloor)$. This dynamic average was used in Wang et al. (2016) and handles bags with differing volumes of instances. For example, with $\eta = 0.2$, the score of bags with 100 and 87

instances would be based on the top 20 and 17 instances, respectively. We train using the binary cross entropy loss (BCE) for bag-level event prediction. The loss is propagated through to the instance model so that instance representations are adjusted to better predict the overall bag label. We refer to this model as the **MIL** model.

**Instance-level Supervision** Vanwinckelen et al. (2015) showed that a good bag classifier does not imply a good instance classifier. Since we want a model to identify useful key instances for downstream tasks, a model that performs well on both bag and instance classification would be useful. Unlike most MIL tasks, we have instance-level knowledge in the form of tweet-level civil unrest prediction probabilities from our trained instance classification model.[5]

We modify our MIL formulation by incorporating these instance-level scores in addition to the bag labels. Our new loss function is

$$\underbrace{-\frac{1}{|X|} \sum_{x_i \in X} \text{BCELoss}(y_i, p(y_i))}_{\text{bag-level loss}} \qquad (2)$$

$$\underbrace{- \beta \frac{1}{|X|} \sum_{x_i \in X} \frac{1}{|x_i|} \sum_{x_i^j \in x_i} \text{BCELoss}(y_i^j, p(y_i^j))}_{\text{instance-level loss}}$$

see Appendix B.2 for the unabridged loss function and a comparison of our function to the one from Wang et al. (2016).

To encourage correct bag-level classification we used the typical binary cross entropy loss (BCE) for logistic regression. Note that $p(y_i = 1)$ is the same as in eq. (1) and is only calculated using the key instances (controlled by $\eta$). The second portion of the loss function, the instance level loss, is also a BCE loss to minimize the difference between the MIL instance prediction score and the true score from the trained instance model. $\beta$ is a hyperparameter to control the impact of instance-level loss on training. We call this model the MIL with Bag and Instance Supervision (**MIL-BI**).

Observe that the MIL model is a special case of MIL-BI. When $\beta = 0$, no instance-level information is incorporated and it is a standard MIL model with only bag loss. Also, when $\eta = 0$, the top-$k$ average is simply the $\max$ operation, another commonly used MIL aggregation function (albeit rather noisy and prone to false positives).

## 4 Data

We use existing datasets for general civil unrest: the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010), Global Civil Unrest on Twitter (G-CUT) (Chinta et al., 2021), and Civil Unrest on Twitter (CUT) (Sech et al., 2020).

Together G-CUT and ACLED provide the tweets for each day in a country and the label of whether an event occurred on that day. G-CUT contains 200 million English tweets from 2014–2019 covering 42 countries in Africa, the Middle East, and Southeast Asia. Due to the large variety of amount of tweets for each day, we randomly sampled a maximum of 1000 tweets from each country from each day to represent the "bag". We also pruned the dataset further than Chinta et al. (2021) to remove spam-like tweets; see Appendix A for details.

Following Chinta et al. (2021), we use the Riots and Protests labels at the day label for a country from ACLED. We consider a day in a country as "positive" for a civil unrest event if ACLED identified a protest or riot on that day in that country.[6] Even if multiple events are identified on the same day, that still only counts as one positive example. All other days are negative (i.e., no event).

As mentioned in Section 3.2, we used CUT to train our instance model.

## 5 Experiments

We evaluate the proposed MIL models along with baselines for the civil unrest detection task. We follow prior work and evaluate model performance on F1, precision, and recall (Chinta et al., 2021; Alsaedi et al., 2017). The model's prediction is marked as correct if it predicts a civil unrest event occurred on a country-day (i.e., bag) that is also identified by the ACLED ground truth. We use the weighted F1 score due to the class imbalance (roughly 30% positive in the training set).

### 5.1 MIL Models

We evaluate the MIL and MIL-BI models described in Section 3.2 across key instance ratios, $\eta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, and instance supervision, $\beta \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$.

---

[5]This differs from other MIL tasks where no instance labels are available to train an instance classification model.

[6]ACLED includes six main event types: battles, explosions/remote violence, violence against civilians, protests, riots, and strategic development

All models had a batch size of 20, a maximum of 100 instances per bag (depending on the number of tweets per country-day), and were trained for 50 epochs (patience of 20 epochs) with AdamW optimizer and $1 \times 10^{-5}$ learning rate with 100 warmup steps. These models were implemented in PyTorch and trained with the HuggingFace Trainer on 4 NVIDIA A100 80GB GPUs. To ensure the model sees a variety of instances for each bag, the 100 instances were sampled from the maximum of 1000 instances in each bag for each iteration (see Section 4).

## 5.2 Comparison Models

We compare our MIL models against other aggregation-based representations.[7] See Appendix C for training details.

**MIL (max)** We also include MIL with $\eta = 0$ ($|K_i| = 1$) to evaluate max aggregation instead of top-$k$ average.

**MIL-Instance only (MIL-I)** Average only the tweet-level scores and do not include any bag information. This model does not require any training as the instance model is already trained. Instead, the top instance scores are averaged as-is.

**AVG Bag** This model is the most direct comparison to our MIL approach since it tests whether the instance-level representation is useful or if only the bag representation is useful. To represent each day in a country, we use the average instance representation, or average instance model embedding across all tweets for that country for the day. The classifier is a random forest model.

**AVG Bag (BERTweet)** This model is the same as AVG Bag except instead of using the instance representations, we use BERTweet representations. These embeddings were trained on general, non-civil unrest tweets. We include this model to evaluate if the AVG Bag model benefits from the civil unrest-aware embeddings provided by the instance model.

## 5.3 Baselines

The following baselines were re-run from (Chinta et al., 2021). The random baselines were also used in Wu and Gerber (2018) and Qiao and Wang (2015). The train/dev/test split is the same as for the

---

[7]The code from Wang et al. (2016) was not available for reproduction.

MIL models (2014–2016/2017/2018–2019). See Appendix C for details.

**N-gram** A random forest classifier with unigrams from all the tweets for each bag as features. We did not remove location-specific words as in Chinta et al. (2021). We preprocess the tweets with the MIL model tokenizer for a consistent vocabulary (i.e., the BERTweet tokenizer).

**Random baseline** Model that uses the rate of events (i.e., positive class) from the train set to predict whether an event will occur. For example, since the train set has 30% positive examples, this model predicts an event occurs for 30% of the test data. This baseline is included purely for comparison and will not be analyzed in-depth along with the other models.

**Country-Random baseline** This model a country-specific version of the random baseline. It predicts an event for a country based on the rate of events for that country in the training set (2014-2016). For example, the prediction for a bag from Zambia would be based on the positive rate specifically for Zambia in the training set as opposed to the overall positive rate.

| Model | F1 | Precision | Recall |
|---|---|---|---|
| MIL-$max$ | 0.71 | **0.73** | **0.74** |
| MIL ($\eta$=0.4) | **0.73** | 0.73 | **0.74** |
| MIL-BI ($\beta = 1$) | 0.67 | **0.73** | 0.72 |
| MIL-I-$max$ | 0.52 | 0.37 | 0.90 |
| AVG-Bag | 0.48 | 0.33 | 0.88 |
| AVG-Bag BERTweet | 0.38 | 0.58 | 0.29 |
| Ngram | 0.48 | 0.64 | 0.38 |
| Random | 0.31 | 0.33 | 0.28 |
| Country-random | 0.50 | 0.54 | 0.46 |

Table 1: Model performance on civil unrest event detection task. The scores shown are from the test set (years 2018-2019). Reported F1 is weighted-F1.

## 6 Results

The notable findings are as follows:

**The MIL models outperformed all the baselines.** All variations of the MIL models outperformed the other aggregation models and baselines. The strongest baseline was one of the more simple, the Country-Random model, as shown in Table 1. Personalizing the model from the overall positive

| $\eta$ | MIL | | | MIL-I | | |
|---|---|---|---|---|---|---|
| | F1 | Precision | Recall | F1 | Precision | Recall |
| 0.0 | 0.71 | 0.73 | 0.74 | **0.52** | **0.37** | **0.9** |
| 0.1 | 0.73 | 0.73 | 0.74 | 0.34 | 0.43 | 0.29 |
| 0.2 | 0.73 | 0.73 | 0.74 | 0.17 | 0.55 | 0.1 |
| 0.3 | 0.72 | 0.74 | 0.74 | 0.042 | 0.44 | 0.022 |
| 0.4 | **0.73** | **0.73** | **0.74** | 0.0073 | 0.29 | 0.0037 |
| 0.5 | 0.73 | 0.73 | 0.74 | 0.0024 | 0.27 | 0.0012 |
| 0.6 | 0.72 | 0.73 | 0.74 | 0.0016 | 0.32 | 0.00078 |
| 0.7 | 0.72 | 0.74 | 0.74 | 0.00089 | 0.36 | 0.00045 |
| 0.8 | 0.72 | 0.73 | 0.74 | 0.0 | 0.0 | 0.0 |
| 0.9 | 0.72 | 0.73 | 0.74 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.72 | 0.72 | 0.73 | 0.0 | 0.0 | 0.0 |

Table 2: Ablation over the key instance ratio, $\eta$ for the MIL and MIL instance-only (MIL-I) models. The results are on the test set. $\eta = 0$ refers to the *max* aggregation. Reported F1 is weighted-F1.

class rate to the positive class rate for each country helped significantly, with an increase of 0.2 F1 from the Random model. The Ngram model outperformed the Random and AVG-Bag BERTweet models by roughly 0.2 and 0.1 F1, respectively. AVG-Bag BERTweet performed worse than the AVG-Bag model, indicating that the civil unrest pre-training from the instance model was helpful.

**Number of key instances does not have an effect on MIL performance.** The effect of adjusting the key instance ratio for the top-$k$ average had little to no impact on the performance, with all models achieving within $\pm 0.1$ F1 of 0.73 on the test set (Table 2). This low impact might be due to the high variance in the number of tweets per bag (see Appendix Figure 4). However, all models with $\eta > 0$ outperformed $\eta = 0$, or the MIL-$max$ model, indicating an advantage in basing the prediction for a country-day on more than one tweet. While very close, a key instance ratio of $0.4$ had the highest performance and we refer to it as **MIL (best)**.

**Incorporating instance supervision hurts model performance.** We use the best $\eta$ from the MIL sweep $(0.4)$ to experiment with instance-level supervision, $\beta$. Similar to the key instance ratio sweep, the instance loss weight also does not have a large impact on model performance, with only a difference of a few F1 points. Table 3 shows the tested $\beta$ values on the validation set. While the difference is not great, it is still more apparent than with the $\eta$ sweep, indicating incorporating instance loss is more impactful on the model than the number of key instances. As $\beta$ increases, performance

| $\beta$ | F1 | Precision | Recall |
|---|---|---|---|
| 0.0 | **0.73** | 0.73 | 0.74 |
| 0.25 | 0.72 | **0.74** | **0.74** |
| 0.5 | 0.71 | 0.73 | **0.74** |
| 0.75 | 0.70 | 0.73 | 0.73 |
| 1.0 | 0.67 | 0.73 | 0.72 |

Table 3: Instance loss parameter sweep $(\beta)$ for the MIL-BI model. As $\beta$ increases, F1 decreases. All other settings are the same as for the best MIL model. Scores are on the test set.

decreases, confirming the conflict of optimizing for both instance and bag-level classification. While all models with $\beta > 0$ do not perform as well as MIL (best), the best MIL-BI model $(\beta = 0.25)$ achieves an F1 of 0.77 on the validation set and 0.72 on the test set. While $\beta = 0.25$ has the best performance, we analyze $\beta = 1.0$ further in **??** to evaluate whether the decline in civil unrest prediction performance is offset by more informative key instances.

**Bag information is needed alongside instances for accurate bag prediction.** Finally, we do not incorporate the bag labels at all and evaluate the MIL instance-only model. In Table 2 we see the drastic change in model performance, with the lack of training with bag labels leading to a performance worse than MIL and the baselines. The exception is with $\eta = 0$ which outperformed most of the baselines at 0.52 F1, 0.37 precision, and 0.90 recall. However, this high F1 is skewed by the very high

recall as opposed to precision, indicating the model over-predicts positive bags. This use of only the single highest-scoring instance for bag label prediction confirms the presence of positive instances in negative bags, or the *collective assumption* (see Section 3).

**Performance varies across countries** Following prior work that uses tweets from multiple countries, we check our model's performance across all 42 countries in the dataset (Zhang et al., 2022; Chinta et al., 2021). The per-country F1, precision, and recall scores from MIL (best) on the test set are shown in Appendix Figure 5. Roughly half of the countries (22) have an F1 score below the aggregated score, and there is a clear gap in performance between countries with the highest (Pakistan, 1.0 F1) and lowest (Morocco, 0.28 F1) scores. This performance discrepancy can in part be explained by unequal country presence in the training set as well as differing rates of events. As shown in Figure 2, Pakistan (PAK) is not only very prevalent in the data, it also has a very high rate of events, indicating a very simple task if the model associates Pakistani tweets with civil unrest. Countries with either very high or very low levels of civil unrest in the train set generally perform better than those in the middle (40-60% positive events). The relationship is not as clear with Morocco (MAR), which appears to be an outlier, since other countries with similar size and rate of events perform better, such as Thailand (THA).

## 7 Discussion

While performance is important, a strength of the MIL approach is the identification of which tweets contribute the most to the final prediction, i.e. those with the highest probabilities. These top tweets can be used to *explain* the model's prediction. We examine the top tweets for a single event below. Also, we revisit the MIL *collective assumption* in an analysis of civil-unrest related tweet distributions on days with and without events.

### 7.1 Key Instance Analysis: Case Study

In Table 4 we compare top tweets from MIL models of interest: the best-performing MIL-$\eta = 0.4$, MIL-BI-($\beta = 1$), and MIL-$max$. We focus on a single event identified by ACLED, a protest in Sri Lanka on September 5, 2018, shown in Table 4.

The selected event was a large protest concerning a political demonstration demanding the govern-
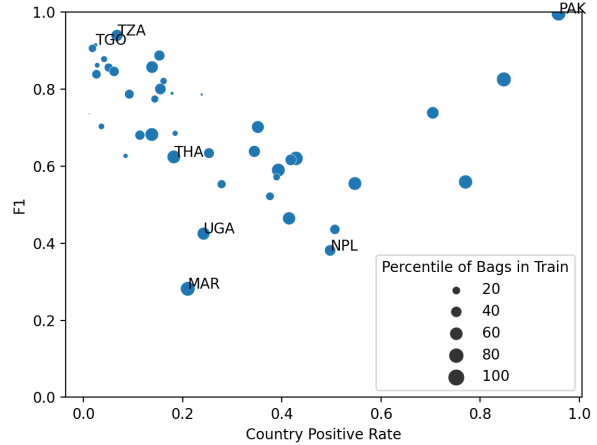


Figure 2: The test F1 scores from the MIL (best) across all countries present in the data. The countries with the highest and lowest F1 scores are annotated. The size of each point is relative to the number of bags each country has in the training data. The model performs best on countries with either very low or very high rates of civil unrest. The included countries are Togo (TGO), Tanzania (TZA), Thailand (THA), Uganda (UGA), Morocco (MAR), Nepal (NPL), and Pakistan (PAK).

ment to step down and was organized by the Joint Opposition, a political alliance. The MIL model predicted a protest with probability $0.53$ and identified informative tweets, with specific mentions of the Joint Opposition as well as police presence for riot control. There is also a noisy, irrelevant tweet directed at the then-president of the US. While the MIL-$max$ model is typically easily skewed by its top tweets, in this specific example it was distracted by an irrelevant tweet about the weather and did not predict that a protest occurred on this day, with a too-low probability of $0.49$. Oddly, while identifying tweets of interest discussing unrest, the MIL-BI model had the lowest prediction of all, with a probability of $0.38$. These tweets are indicative of unrest and one even tags the president of Sri Lanka, but are not as informative as those from the MIL model. From this example, the MIL model identified mostly informative tweets while MIL-$max$ was distracted by irrelevant noise. The MIL and MIL-BI models had low overall prediction confidence due to a skewed positive instance distribution, i.e. while the top tweets had very high scores, most of the tweets were not identified as civil-unrest related and brought down the average.

This is a single qualitative example and more quantitative analyses are needed for evaluating the usefulness of identified key tweets in downstream tasks like event extraction and summarization.

| | | | **Event description:** On 5-6 Sept, in Fort (Colombo, Colombo), thousands gathered at Lake House roundabout in a JO-organized protest demanding the government to step down. Protesters marched from different locations in Colombo city - including Galleface and Kurunduwatta - to Colombo Fort to join a JO-organized protest. Despite peaceful protest, 1 protester died due to cardiac arrest and several hospitalized due to food poisoning, minor injuries, and excessive drinking. |
|---|---|---|---|

| Model | Bag Score | Tweet Score | Tweet |
|---|---|---|---|
| MIL ($\eta = 0.4$) | | 0.99 | @realDonaldTrump What about Saudi attacks ? |
| | 0.53 | 0.99 | The Joint Opposition ( JO) is planning to carry out a huge mass protest called "Jana-balaya Kolabata" against the Government targeting Colombo on the 5th September 2018 from 1400 Hrs. |
| | | 0.98 | Over 5,000 policemen from various units armed with all riot controlling mechanisms will remain standby to face the... |
| MIL-$max$ | 0.49 | 0.49 | current weather in Colombo: scattered clouds, 24°C 88% humidity, wind 3kmh, pressure 1010mb |
| MIL-BI ($\beta = 1.0$) | | 0.99 | Dear Mr. Ranjan, the salve of @RW_UNP, majority is suffering your mismanagement... |
| | 0.38 | 1.0 | @USER @UN Yes UN, world Bank and other international organizations must be responsible for poverty because... |
| | | 0.97 | @vijaytelevision @Vivo_India This cheater; poi Kari deserves ah |

Table 4: Top key tweets identified by MIL models with different parameters for September 5, 2018 in Sri Lanka. The event description is from ACLED. The tweet scores are from the instance model and the bag score is the aggregated score. The bag score differs from the tweet scores for the MIL and MIL-BI models because not all the top $\eta$ tweets are shown.

## 7.2 Distribution of Tweet Scores

An important part of the MIL formulation that we chose was to embrace the noisy Twitter data with the *collective assumption*. In this assumption, a country-day where an event did not occur can still contain civil-unrest related tweets, as identified by the instance model. Appendix Figure 6 shows the distribution of instance scores grouped by country across days with and without an event. The majority of tweets are not unrest-related (i.e., instance score below $0.5$), but there is a long tail toward high-scoring instances. The most important note is that for most countries there is little to no visible difference in civil-unrest related tweets on days with and without events. This is a strong indication of why this task of civil unrest prediction on the country-day level is difficult. We discuss potential model improvements in Section 8. Examples of protest-related tweets on a day without civil unrest are shown in Appendix Table 6.

## 8 Conclusion

Our goal was to evaluate how well a multi-instance learning (MIL) approach to civil unrest detection on Twitter performed to other, aggregated methods. We modeled tweets that occurred on the same day in a country as a *bag* where each tweet is an *instance*. We showed that this formulation worked well, achieving an F1 score of 0.73 on de-

tecting events identified by ACLED. The number of instances that contributed to the final prediction for each bag had little effect, but incorporating instance-level supervision in the form of a loss penalty for misclassifying unrest-related tweets did negatively impact overall event prediction performance.

Since we only evaluated on the civil unrest task, it is unclear if these results are task or data-specific. The remaining challenges are to quantitatively test whether the identified key instances (1) contain event information from the ACLED event(s) (bag labels) and (2) can be used in a full event extraction or summarization pipeline, as in Wang et al. (2016).

We showed that this approach is promising for civil unrest detection, but it can easily be adapted for a new task by substituting new tweets and bag labels, such as the detection of other types of events or even stance classification. The common thread between these problems is that tweets are commonly analyzed in aggregate and all are assumed to be directly related to the topic in question, however since Twitter data is noisy, this is a potentially incorrect assumption.

award N00014-19-1-2316 issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Office of Naval Research.

## Ethical Considerations

The main ethical considerations of this work lie in demographic representation, user privacy, and dual use.

We chose Twitter to gain direct access to the "voice" of the people, but through our filtration of non-English and non-geotagged tweets, the data is not a representative sample of the population. Also, some countries do not have a large number of Twitter users in general. The result is that some countries are vastly over-represented in the dataset than others (e.g. South Africa vs Ethiopia). See the G-CUT paper for details on Twitter coverage of the ACLED-identified civil unrest events.

All of the above culminates in a dataset not representative of the population of a country. In future work we will use a multi-lingual approach to mitigate the bias of using English-only tweets. For the chosen regions of Africa, Middle East, and Southeast Asia, we would incorporate Arabic and French at a minimum.

A way to better represent the population would be to use Twitter Geolocation tools, such as the recently introduced Geo-Seq2seq Carmen (Zhang et al., 2023). This data expansion potentially comes at the expense of user privacy.

And finally, the dual use of a tool which identifies tweets discussing civil unrest is a very real possibility. For the purposes of this work, we focus solely on observing and modeling civil unrest and not instigating or curtailing it.

## Limitations

The limitations in this work are mostly from not fully exploring the model space with respect to training parameters and architecture and the discontinued access to Twitter going forward.

While we ablated over multi-instance learning-specific parameters such as key instance ratio and instance-level loss, there is always more to be done for hyperparameter tuning of the learning rate and other optimizer parameters. Also, to address lim-

itations of commonly used aggregation functions, one could automatically learn an aggregation, such as through an attention layer (Ilse et al., 2018).

Further, the effect of instance model performance on instance model loss inclusion was not included, i.e., as the instance model becomes more accurate, do the bag predictions become more accurate as well? Also, the effects of instance selection was not explored beyond random sampling.

Also of note is the focus on English tweets, which as discussed in Section 8, limits the population represented in the data. If moving to a multilingual setting, we would use Bernice, a multilingual BERT model trained from-scratch on tweets (DeLucia et al., 2022), for the instance representation instead of BERTweet.

We leave these areas to be addressed in future work.

Regarding data access, in March 2023, Twitter changed its API pricing and effectively closed off its public API stream with undetermined plans for an academic pricing tier.[8] This means while no new tweets can be collected, but past tweets already collected can still be modeled. Our MIL approach can be applied to past tweets for historical events, or other social media platforms like Reddit or Facebook.

## References

Nasser Alsaedi, Pete Burnap, and Omer Rana. 2017. Can We Predict a Riot? Disruptive Event Detection Using Twitter. *ACM Transactions on Internet Technology*, 17(2):1–26.

Anna L. Buczak, Benjamin D. Baugher, Adam J. Berlier, Kayla E. Scharfstein, and Christine S. Martin. 2022. Explainable forecasts of disruptive events using recurrent neural networks. In *2022 IEEE International Conference on Assured Autonomy (ICAA)*, pages 64–73.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple in-

---

[8] https://twitter.com/TwitterDev/status/1641222782594990080

stance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353.

Erica Chenoweth and Jay Ulfelder. 2015. Can structural conditions explain the onset of nonviolent uprisings? *Journal of Conflict Resolution*, page 0022002715576574.

Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze, and Anna L. Buczak. 2021. Study of manifestation of civil unrest on Twitter. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 396–409, Online. Association for Computational Linguistics.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Tiancheng Hu, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu. 2021. Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online. Association for Computational Linguistics.

Jack A Goldstone, Robert H Bates, David L Epstein, Ted Robert Gurr, Michael B Lustik, Monty G Marshall, Jay Ulfelder, and Mark Woodward. 2010. A global model for forecasting political instability. *American Journal of Political Science*, 54(1):190–208.

Ali Hürriyetoğlu, editor. 2021. *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*. Association for Computational Linguistics, Online.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. 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)*, pages 79–91, Online. Association for Computational Linguistics.

Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, and Erdem Yörük, editors. 2022. *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. 2018. Attention-based Deep Multiple Instance Learning. *arXiv:1802.04712 [cs, stat]*. ArXiv: 1802.04712.

Kamrul Islam, Manjur Ahmed, Kamal Z. Zamli, and Salman Mehbub. 2020. An online framework for civil unrest prediction using tweet stream based on tweet weight and event diffusion.

Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global Data on Events, Location and Tone,. page 51.

Sneha Mehta, Huzefa Rangwala, and Naren Ramakrishnan. 2022. Improving zero-shot event extraction via sentence simplification. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 32–43, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alan Mishler, Kevin Wonus, Wendy Chambers, and Michael Bloodgood. 2017. Filtering Tweets for Social Unrest. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 17–23.

Sathappan Muthiah, Bert Huang, Jaime Arredondo, David Mares, Lise Getoor, Graham Katz, and Naren Ramakrishnan. 2015. Planned Protest Modeling in News and Social Media. In *Twenty-Seventh IAAI Conference*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. *arXiv:2005.10200 [cs]*. ArXiv: 2005.10200.

Fengcai Qiao and Hui Wang. 2015. Computational Approach to Detecting and Predicting Occupy Protest Events. In *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, pages 94–97.

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*. Publisher: SAGE PublicationsSage UK: London, England.

Naren Ramakrishnan, Gizem Korkmaz, Chris Kuhlman, Achla Marathe, Liang Zhao, Ting Hua, Feng Chen, Chang Tien Lu, Bert Huang, Aravind Srinivasan, Khoa Trinh, Patrick Butler, Lise Getoor, Graham Katz, Andy Doyle, Chris Ackermann, Ilya Zavorin, Jim Ford, Kristen Summers, Youssef Fayed, Jaime Arredondo, Dipak Gupta, Sathappan Muthiah, David Mares, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, and Anil Vullikanti. 2014. 'Beating the news' with EMBERS: forecasting civil unrest using open source indicators. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 1799–1808, New York, New York, USA. ACM Press.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2019. Calls to action on social media: Detection, social impact, and censorship potential. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 36–44, Hong Kong, China. Association for Computational Linguistics.

James Scharf, Arya D. McCarthy, and Giovanna Maria Dora Dore. 2021. Characterizing news portrayal of civil unrest in Hong Kong, 1998–2020. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 43–52, Online. Association for Computational Linguistics.

Justin Sech, Alexandra DeLucia, Anna L. Buczak, and Mark Dredze. 2020. Civil unrest on Twitter (CUT): A dataset of tweets to support research on civil unrest. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 215–221, Online. Association for Computational Linguistics.

Adam Smidi and Saif Shahin. 2017. Social Media and Social Mobilisation in the Middle East: A Survey of Research on the Arab Spring:. *India Quarterly*. Publisher: SAGE PublicationsSage India: New Delhi, India.

Xosé Soengas-Pérez. 2013. The role of the Internet and social networks in the arab uprisings an alternative to official press censorship. *Comunicar*, 21(41):147–155.

Zachary C Steinert-Threlkeld. 2017. Spontaneous collective action: peripheral mobilization during the arab spring. *American Political Science Review*, 111(2):379–403.

Gitte Vanwinckelen, Vinicius Tragante do O, Daan Fierens, and Hendrik Blockeel. 2016. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*, 30(2):313–341.

Gitte Vanwinckelen, Ó ViniciusTragantedo, Daan Fierens, and Hendrik Blockeel. 2015. Instance-level accuracy versus bag-level accuracy in multi-instance learning. *Data Mining and Knowledge Discovery*.

Wei Wang, Yue Ning, Huzefa Rangwala, and Naren Ramakrishnan. 2016. A Multiple Instance Learning Framework for Identifying Key Sentences and Detecting Events. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pages 509–518, Indianapolis, Indiana, USA. Association for Computing Machinery.

Congyu Wu and Matthew S. Gerber. 2018. Forecasting Civil Unrest Using Social Media and Protest Participation Theory. *IEEE Transactions on Computational Social Systems*, 5(1):82–94. Conference Name: IEEE Transactions on Computational Social Systems.

Huiling You, David Samuel, Samia Touileb, and Lilja Øvrelid. 2022. EventGraph: Event extraction as semantic graph parsing. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 7–15, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vanni Zavarella, Hristo Tanev, Ali Hürriyetoğlu, Peratham Wiriyathammabhum, and Bertrand De Longueville. 2022. Tracking COVID-19 protest events in the United States. shared task 2: Event database replication, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 209–216, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Han Zhang and Jennifer Pan. 2019. CASM: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1):1–57.

Jingyu Zhang, Alexandra DeLucia, and Mark Dredze. 2022. Changes in tweet geolocation over time: A study with carmen 2.0. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 1–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jingyu Zhang, Alexandra DeLucia, and Mark Dredze. 2023. Geo-seq2seq: Twitter user geolocation on noisy data through sequence to sequence learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, CA. Association for Computational Linguistics.

Xin Zheng and Aixin Sun. 2019. Collecting event-related tweets from twitter stream. *Journal of the Association for Information Science and Technology*, 70(2):176–186.
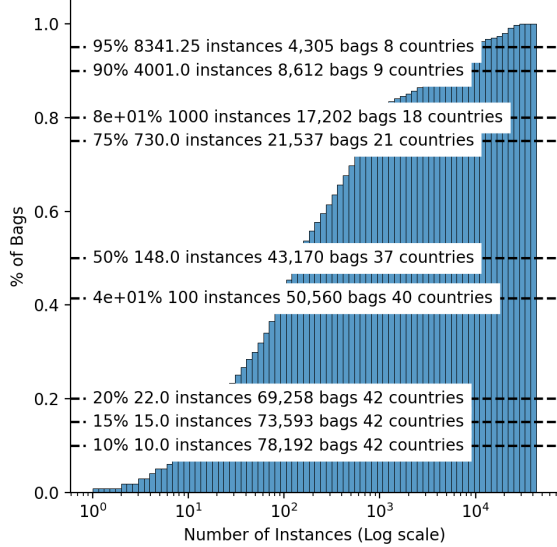
Figure 3: Plot is cumulative and normalized to show the percentage of bags that remain after raising the minimum number of instances threshold. Note the log scale for number of instances.

## A  Data Preparation

As discussed in Section 4, we use the Global Civil Unrest on Twitter (G-CUT) dataset introduced by (Chinta et al., 2021). In this work, we perform further data cleaning:

- Removed retweets and quote tweets

- Removed tweets identified as spam (tweets with more than three hashtags or user mentions or less than three non-URL, hashtag, or user mention tokens)

- Removed exact text duplicates

After cleaning 86,096 bags out of 86,270 (99.80%) remained.

Another change is we removed samples, or bags, that did not have at least 10 tweets, or instances. This threshold was based on dropping the bottom 10% of bags, which still retained 78,192 samples (91% of the original dataset) from all 42 countries (see Figure 3).

## B  Model Training Details

Parameters not detailed in the main paper are discussed here.

### B.1  Civil Unrest Filtration Model

For the Civil Unrest Filtration model, or instance model in the context of multi-instance learning,

| Metric | Validation | Test |
|---|---|---|
| Accuracy | 0.85 | 0.82 |
| Loss | 1.5 | 1.9 |
| F1 (Positive) | 0.65 | 0.6 |
| F1 (Macro) | 0.86 | 0.82 |
| Precision | 0.89 | 0.84 |
| Recall | 0.85 | 0.82 |

Table 5: Results for the validation and test set of the best performing civil unrest filtration model.

we fine-tuned a BERTweet model on the Civil Unrest of Twitter (CUT) dataset (Sech et al., 2020). After removing samples identified as non-English, the dataset consists of 2761 samples, 553 of which discuss general unrest (20%). We split the dataset into train, validation, and test sets of sizes 2235/249/277, respectively.[9] To encourage equal class prevalence we used stratified sampling for the splits. We chose the general unrest label instead of specific protests or events since our overall model aims to predict civil unrest.

The BERTweet model was fine-tuned with a HuggingFace classification head for $100$ epochs, AdamW optimizer, linear schedular warmup of $50$ steps, binary cross-entropy loss, $0.00006815$ learning rate, $0$ weight decay, betas $(0.9, 0.999)$, epsilon $1.000e-8$, $128$ batch size, and early stopping with a patience of $10$. These parameters were chosen after performing a hyperparameter sweep for best positive F1 score on the validation set of 100 trials, selecting randomly from weight decay values $\{1e-10, 1e-09, 1e-08, 1e-07, 1e-06, 1e-05, 0.0001, 0.001, 0.01, 0\}$ and a learning rate uniformly sampled from $[1e-6, 1e-2]$. To address the large class imbalance we included a weight for the positive class calculated as

$$\text{positive weight} = \frac{n_{samples}}{n_{classes} \times n_{positive samples}}$$

where the sample stats are based on the train split. Choosing a model based on best positive F1 score was important because we found that best validation loss did not necessarily correlate with accurate predictions of the positive class, most likely due to the large class imbalance. Performance details of the final model are in Table 5.

---

[9]This is 10% of the dataset set aside for test and then 10% of the remaining data set aside for evaluation.

## B.2 MIL Loss

The full loss function is in Equation (3). While similar to the loss function from Wang et al. (2016), ours is simpler and does not use the instance-ratio control loss or the instance-level manifold propagation. Since the overall model is initialized with meaningful representations from the BERT-based instance model, we omitted the other instance losses. The losses were more necessary in Wang et al. (2016) since the instance/bag representations were learned from averaged word2vec embeddings, which are not as powerful as BERT embeddings. Further, unlike Wang et al. (2016), we have "ground-truth" labels for the instances, allowing us to use a BCE loss instead of their instance hinge loss. A future iteration of this work could incorporate these other losses.

## C   Baseline Training Details

Parameters not detailed in the main paper are discussed here.

### C.1   N-gram Baseline

Inspired by the simple baselines from Chinta et al. (2021), we included a similar baseline in this work.

For a more direct comparison to the MIL and MIL-AVG models we used the BERTweet tokenizer to normalize and tokenize the tweets. We used the random forest classifier implementation from sklearn (Buitinck et al., 2013) with the following settings: 10 estimators, max depth of 32, minimum samples split of 32, and a balanced class weight. These settings were copied from Chinta et al. (2021). We used the same train, validation, and test splits as for the MIL model.

### C.2   AVG-Bag Baseline

The AVG-Bag model is a direct comparison to the standard method of using all tweets to represent each bag as opposed to the MIL approach. With the same instance model discussed in Appendix B.1 and Section 3.2, we represent each bag as the average instance representation ([CLS] token). These representations are then used as features for a random forest model, with the same settings as for the N-gram model. We used the same train, validation, and test splits as for the MIL model.

$$L(x, y; \theta) = -\underbrace{\frac{1}{|X|} \sum_{x_i \in X} y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i))}_{\text{bag-level loss (BCE)}}$$ (3)

$$-\beta \underbrace{\frac{1}{|X|} \sum_{x_i \in X} \frac{1}{|x_i|} \sum_{x_i^j \in x_i} y_i^j \log(p(y_i^j)) + (1 - y_i^j) \log(1 - p(y_i^j))}_{\text{instance-level loss (BCE)}}$$

| Date | Country | Tweet Score | Tweet |
|------|---------|-------------|-------|
| 2017-01-11 | UGA | 1.0 | Somalia's militant Islamist group al-Shabab has shot dead two people it accused of being gay. |
| 2017-02-19 | ZAF | 1.0 | The sad thing about today.The idiot politicians who are preaching economic emancipation are millionaires |
| 2017-04-08 | UGA | 1.0 | Some of issues we need Govt to address:non prioritisation of National Health insurance scheme. #Ugbudget17 @USER @HealthVoice_UG |

Table 6: Example positive tweets (i.e., civil-unrest related) in negative bags (i.e., a country-day with no event) from the validation set (2017). The tweet scores are from the instance classification model (see Section 3.2). UGA and ZAF refer to Uganda and Zambia, respectively.
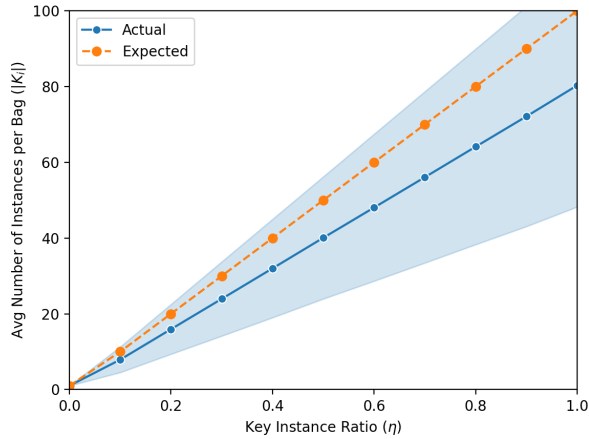


Figure 4: The expected number of key instances for each bag and the average from the train set. The shaded region is the standard deviation.
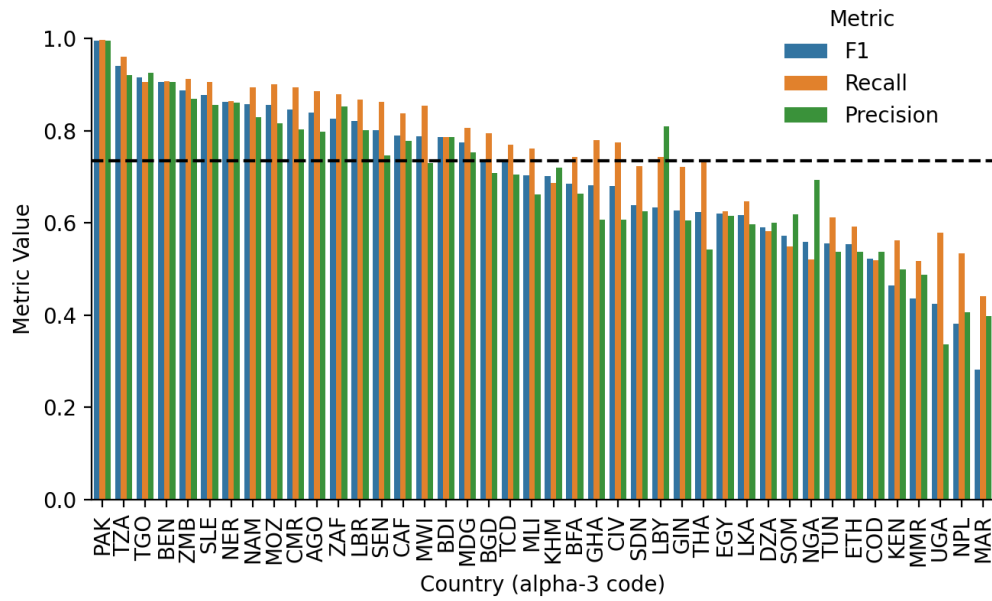
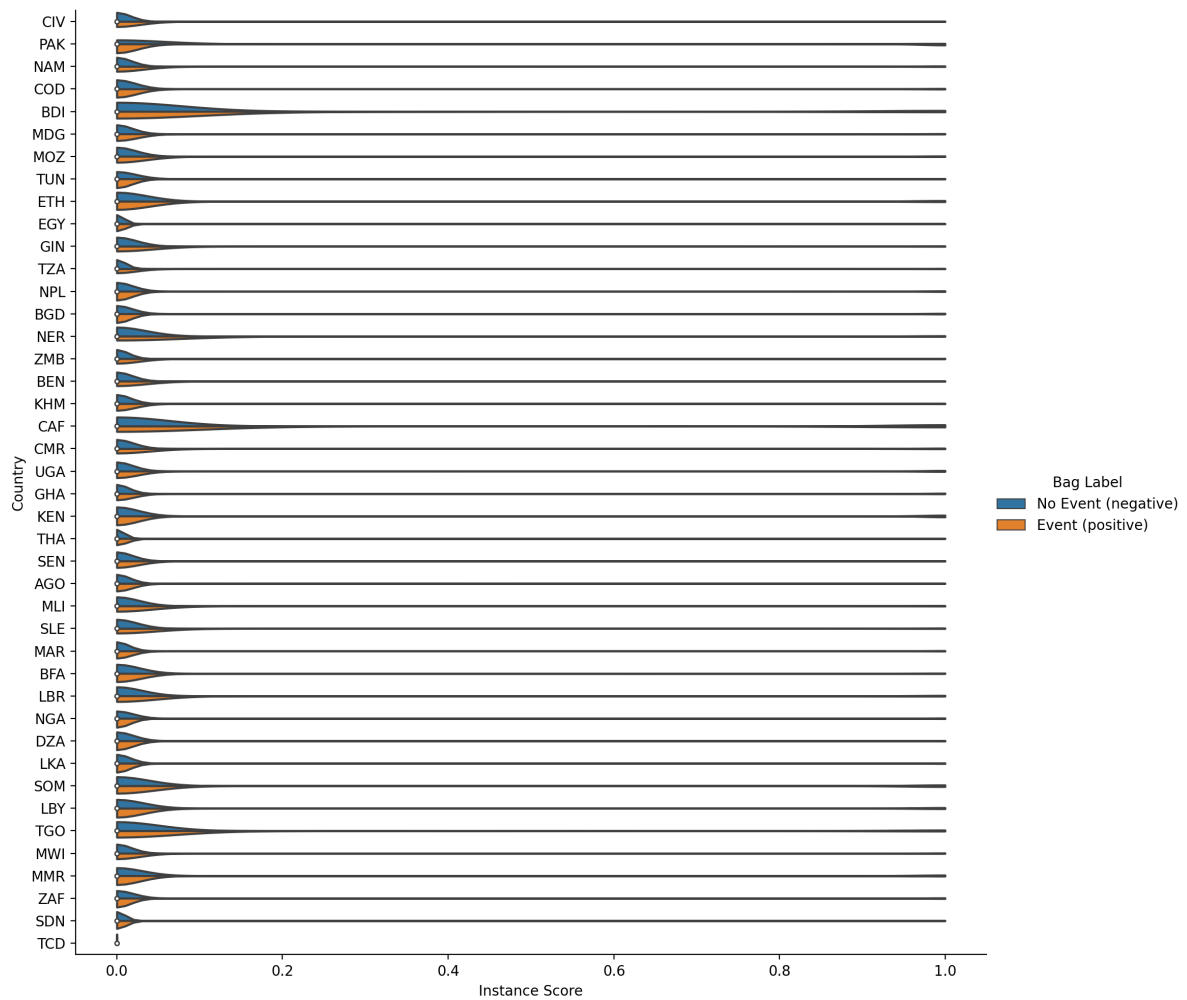Figure 5: Per-country F1 results of top MIL model on the test set.



Figure 6: Distribution of instance scores for each country. Scores are from the instance model (see Section 3.2) on the validation set (year 2017). Countries are identified by their ISO 3166-1 alpha-3 codes.