Studying riots through the lens of social media

Spierenburg, L.J.; van Cranenburgh, S.; Cats, O.

DOI
10.1186/s40537-025-01242-2

Publication date
2025

Document Version
Final published version

Published in
Journal of Big Data

Citation (APA)
Spierenburg, L. J., van Cranenburgh, S., & Cats, O. (2025). Studying riots through the lens of social media. *Journal of Big Data*, *12*(1), Article 182. https://doi.org/10.1186/s40537-025-01242-2

Spierenburg et al. Journal of Big Data (2025) 12:182
https://doi.org/10.1186/s40537-025-01242-2

Journal of Big Data

RESEARCH

Open Access

Studying riots through the lens of social media

Lucas Spierenburg[1*], Sander van Cranenburgh[2] and Oded Cats[1]

*Correspondence:
l.j.spierenburg@tudelft.nl

[1] Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628CN Delft, Netherlands

[2] Transport and Logistics Group, Department of Engineering Systems and Services, Faculty of Technology, Policy, and Management, Delft University of Technology, Jaffalaan 5, 2628BX Delft, Netherlands

Abstract

The emergence of social media offers unprecedented opportunities to map social unrest with high spatiotemporal resolution. This study leverages geolocated social media footage to analyze the spatiotemporal distribution of the 2023 'Nahel Merzouk' riots in France. Using a fine-tuned computer vision model, we detect riot-related content in visual data and validate our approach by comparing the spatiotemporal patterns of detected posts with rioting events reported in the press. Our method yields a spatial resolution of 300 × 300 m, thereby facilitating a detailed analysis of riot distributions at unprecedented scale. By applying density-based clustering, we map riots across seven French cities, revealing their highly localized and bursty dynamics. This study opens pathways for future research on the causes and dynamics of social unrest, enabling a deeper understanding of urban riots and their potential mitigation.

Keywords: Riots, Computer vision, Transfer learning, Social media, Spatiotemporal analysis

Introduction

Riots are instances of agitated social unrest, typically spontaneous and violent, characterized by vandalism, looting, and clashes with law enforcement [25, 52]. Often reflecting latent societal tension, riots result in substantial economic, human, and social costs such as injuries, damage, and mass arrests [12, 24, 43]. Extensive research has been devoted to understanding riots, with two primary objectives: first, identifying the underlying causes to address social tensions and prevent future incidents; and second, analyzing riot dynamics to develop effective crisis management strategies that limit social and material consequences [16, 21, 23, 28, 29, 52]. Such understanding necessitates reliable data that capture the spatiotemporal distribution of such events, highlighting where and when unrest started and how it evolves over time. Data sources commonly used in the literature, such as police records or press reports, suffer from poor spatiotemporal resolution [7, 8, 11, 13]. For instance, the police records used in Bonnasse-Gahot et al. [11] provides the daily number of rioting events at the municipal level which does not allow to track how a single event escalated over time or spread in space.

The emergence of social media opens new opportunities for mapping riots, as they are massively used by urban residents to report incidents and provide data with a much finer spatiotemporal resolution than traditional sources [1, 30, 33–35, 46, 49]. Previous research demonstrates the potential of social media data to complement traditional sources in mapping social unrest [3, 30, 53]. For instance, Alsaedi et al. [3] could trace the London 2011 riots from hashtags[1] of geolocated tweets. These studies showcase the ability of social media data to trace events with substantial agreement with ground-truth datasets [3, 30, 53].

However, research that leverages social media to map riots has been constrained by a significant methodological limitation: the reliance on textual content and structured metadata rather than visual content. Previous studies have focused primarily on analyzing tweets through hashtags, keywords, or post metadata [3, 20, 30, 53]. This approach overlooks rich visual information contained in footage that users extensively share during riot events-content that often provides more direct and objective evidence of unrest than textual descriptions.

Addressing this methodological gap, we develop a computer vision approach capable of identifying riot-related posts from massive unstructured visual datasets. This primary contribution represents a significant departure from prior work by directly analyzing the visual content of social media posts rather than relying on textual or metadata cues. We train a computer vision model to classify photos and videos as riot-related using geolocated footage collected during the 2023 'Nahel Merzouk' riots in France, demonstrating the method's effectiveness through validation against press reports.

As a secondary contribution, our approach enables fine-grained spatiotemporal analysis that demonstrates the analytical potential unlocked by visual social media content. We showcase this capability by animating the evolution of riots across 7 French cities at 300×300 m resolution and hourly intervals, revealing riot dynamics-such as their manifestation as discrete, localized bursts-that traditional data sources and text-based social media approaches cannot capture.

We organize the remainder of the paper as follows. Section "Case study and datasets" introduces the case study and the dataset employed, Sect. "Method" outlines the implemented methodology, Sect. "Results" presents the findings, and Sect. "Discussion and conclusion" concludes the analysis.

Case study and datasets

We study the 'Nahel Merzouk' riots in France which took place in the summer of 2023 (Sect. "Case study"). After collecting geolocated visual data from the social media platform Snapchat during this period, we annotate and organize the dataset to train and evaluate our computer vision model (Sect. "Social media data"). We use an external dataset documenting the spatiotemporal distribution of these riots from press reports to test the validity of model results (Sect. "Press data").

Case study

Our analysis focuses on the 'Nahel Merzouk' riots, which took place across France between June 27th and July 5th, 2023 [43]. These riots were sparked by the fatal police

[1] Short textual tags attached to posts.

shooting of Nahel Merzouk, a 17-year-old teenager of Moroccan and Algerian descent, in the Paris suburb of Nanterre. The incident occurred when Merzouk attempted to drive away from police officers who had stopped him for speeding in a bus lane. The incident, viewed as emblematic of latent discrimination in disadvantaged neighborhoods, quickly ignited widespread outrage [10]. The riots initially broke out in several suburbs of Paris and quickly spread to cities and towns throughout the country. The unrest was marked by intense confrontations with the police, the burning of vehicles and buildings, and widespread acts of vandalism. Public infrastructure, including town halls, schools, and public transport facilities, as well as private properties, were severely damaged. Looting also occurred in many areas, further exacerbating the social and economic impact of the unrest [43]. In response, French authorities deployed tens of thousands of police officers across affected areas and implemented emergency measures, including curfews and restrictions on public gatherings. Social media played a pivotal role in documenting the riots, with users sharing footage about how the events unfolded. Our study demonstrates how these data can be used to draw the spatiotemporal distribution of events that are often difficult to capture with traditional data sources.

Social media data

We collected publicly available footage from the social media platform Snapchat taken in 7 French urban areas; namely Paris, Marseille, Lyon, Toulouse, Lille, Bordeaux, and Grenoble between June 25th and July 16th. We use the OECD definition of the urban core to delineate the study perimeter [15]. Map A in Fig. 1 shows the Paris urban core as delineated by the OECD. A critical constraint in our data collection was Snapchat's ephemeral nature: posts on the platform automatically expire and become unavailable after 1 day to 1 week, depending on user settings and content type. We began systematic data collection on June 29th. Consequently, our dataset provides comprehensive coverage from June 29th onward, while data from earlier dates (June 25th–28th) may be incomplete due to posts that had already expired before our collection began. The data consist of around 107 thousand short videos (approximately 81%) and images (19%) posted publicly by users. A 300×300 m sampling grid was used for data collection. Map B in Fig. 1 shows the spatial distribution of the posts collected within the time period.

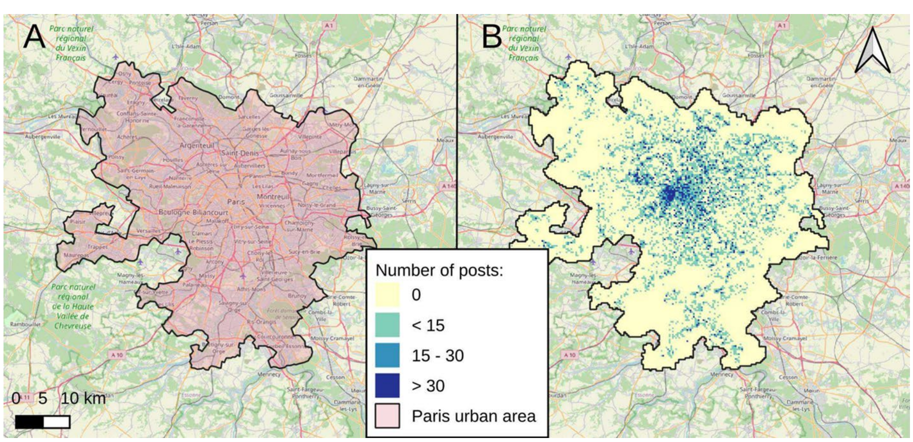

Fig. 1 **A** Paris urban core, as defined by Dijkstra et al. [15]. **B** Number of posts per 300×300 m grid cells

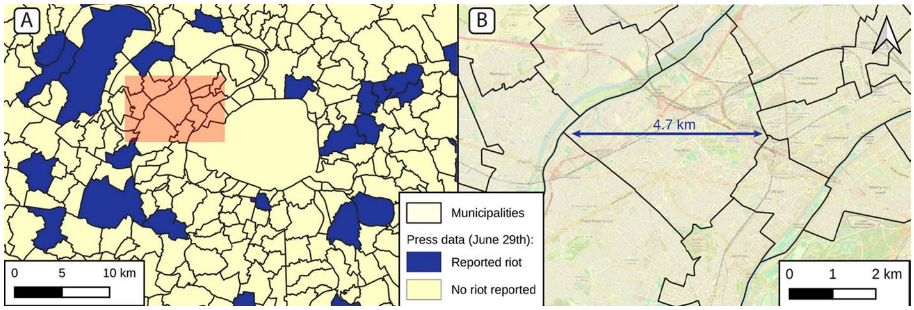

Fig. 2 **A** Rioting events on the night of June 29th were reported in the press dataset. **B** Spatial resolution of the dataset

We label by hand 3917 segments taken from 3234 short videos. A segment consists of a set of successive frames that either (1) represents a rioting event (positive label) or (2) does not represent such an event (negative label). In our annotation process, we specifically focus on violent riot activities including arson, looting, damage to public infrastructure, and confrontations with law enforcement, explicitly excluding peaceful gatherings, demonstrations, or protests without violent elements. We annotate the data conservatively, meaning that ambiguous posts for a human annotator were classified as negative. A video can be composed of several segments with different labels. For instance, if the video scene switches from a car set on fire to the sidewalk, we manually separate the video into two segments, respectively labeled as positive and negative. A total of 257 videos (8%) have at least one video segment relating to a riot in this annotated dataset.

Next, the data are divided into a training set, a validation set, and a test set containing, respectively, 40%, 30%, and 30% of the labeled videos. We ensure that each video is exclusively present in one of the three sets. We tune the weights of the model using the training set. The validation set serves to keep track of the model's performance while training. The test set is a third independent dataset that is used to fairly assess the performance of the model after training on unseen data (Sect. "Validation of the computer vision model").

The implementation of the proposed approach involves major privacy considerations as we collect data from a social media platform. The data collection is non-discriminative, i.e. we cannot filter irrelevant posts prior to collecting data. The data may also contain identifiable information in some cases as the identity of individuals can be revealed if their faces are included in the image or video. Sensitive personal data may therefore be exposed in case of a data breach. We reduce this risk in two stages. We first blur faces upon collection and destroy the raw data to limit identifiable details in the collected data (see Appendix). Then, we destroy the blurred footage after publication and publish publicly the metadata which contains the time, the location of posts, as well as the label predicted by the computer vision model [41].

Press data

We validate the results of our approach by comparing them to a dataset listing all events reported by the local press [5]. This dataset was created by a group of researchers that listed all rioting events reported in the press, day by day, in every municipality. We merge these data with the geographic borders of municipalities, allowing us to map the spatial distribution of riots on a daily basis. For instance, map A in Fig. 2 displays the rioting events reported for the night of June 29th in the city of Paris and neighboring municipalities. The dataset relates 576 events taking place in 408 French municipalities between June 27th and July 3rd. In this dataset, we observe a surge in rioting events between June 28th and June 30th (76% of all registered events) followed a progressive return to normal between July 1st and July 3rd. We test the validity of our approach against this local press dataset by overlapping the spatiotemporal distribution of social media posts representing a riot with a map like map A in Fig. 2. Map B in Fig. 2 shows the typical spatial resolution of the press dataset, using the spatial extent of Nanterre (12 km^2) as an example, which is much coarser than our 300×300 m resolution.

Method

This section describes the training of the computer vision model that detects rioting events from visual social media content (Sect. "Computer vision model") and introduces the tools used for the spatiotemporal analysis of the detected riot-related posts (Sect. "Spatiotemporal analysis").

Computer vision model

This study proposes a model that detects rioting events from geolocated visual content. Due to the unstructured nature of footage data, the model needs to learn what in images pertains to the concept of riots based on a sample of hand-annotated pictures. We address this challenge by fine-tuning a computer vision model that extracts visual patterns relevant to our task and classifies images based on them. Section "Image classification" elaborates on the image classification task and on the model used for that purpose. Section "Training procedure" discusses the training process. Section "Decision threshold tuning" presents the decision threshold tuning for optimal classification performance.

Image classification

The task for the computer vision model is to predict the label of an image, a binary variable with a value of 1 if the image contains a riot and 0 otherwise. Training a deep learning model from scratch requires an extensive labeled dataset. In this work, we reduce drastically the amount of data needed using transfer learning [51]. This approach involves using both the architecture and the weights of an existing model trained for a similar image classification task and retraining part of its layers to specialize it to another use case.

We choose an Efficientnet architecture as the base model for transfer learning that is both more efficient and performs better on benchmark datasets than most image classification models [44]. Vision transformers (ViT) demonstrate higher accuracy but do not justify the increased computational requirements [17]. We implement the smallest architecture 'Efficientnet b0', as our training set is limited and larger architectures did not show significantly better performance, using the weights trained on the ImageNet 2 dataset, publicly available on the Pytorch library [36, 37].

The base model efficientnet b0 is a deep neural network structured as follows: a classifier suggests the class of the image based on the visual features identified by the feature extractor. We replace the base classifier by a custom multilayer perceptron, and optimize its architecture (depth and width) in a hyperparameter tuning step.

The Efficientnet b0 architecture constrains the data resolution. Our 480×852 images are resized to fit the 224×224 image resolution expected by the model [44]. Other alternatives to resizing are cropping and padding, but they resulted in lower performances on the validation set.[2] Finally, we normalize the pixel colors to the same scale as the ImageNet dataset, as it improves the performance.

Training procedure

Our training procedure includes data balancing, hard-example mining, and gradual unfreezing to improve the model's performance. This subsection discusses these three elements. First, the social media dataset is highly imbalanced towards the negative class, which biases the model towards predicting this class more. We address this issue with data balancing. We sample a number of video segments from the negative class equal to the number of video segments from the positive class at each epoch. Second, some scenes from the negative class are harder to differentiate from a riot than others (e.g. a night concert). We implement hard-example mining in the sampling process, where misclassified video segments are more likely to be sampled. Each video segment is associated with difficulty score based on cross-entropy loss, used as a weight in the sampling process. Third, transfer learning is subject to 'catastrophic forgetting', where the model discards patterns observed in the general dataset used for pretraining to capture patterns observed in the training data. This is problematic when the model starts overfitting certain visual details from the training data while discarding more general patterns observed in the pretraining data. Gradual unfreezing addresses this issue by focusing the training on the classifier part of the model (that is unrelated to the pretrained model), while preserving as much as possible the weights from the feature extractor [14]. In an initial phase, we train only the model's classifier and later unfreeze layers from the feature extractor when the performance reaches a plateau on the validation set.

Decision threshold tuning

Once trained, the model produces the probability for an image to represent a riot. For videos in the dataset—composed of a series of still images (frames)—the probability of a video representing a riot is determined by aggregating probabilities across all its frames. We classify a post as representing a riot if the probability provided by the model is greater than a decision threshold. Given the dataset's significant imbalance toward the negative class, the threshold is optimized by maximizing the f1-score rather than accuracy. We then evaluate how the decision threshold influences the f1-score (plot A in Fig. 3). The f1-score reaches its maximum (0.77) for a decision threshold of 0.42, plateaus, and decreases when the threshold exceeds 0.6. The decision threshold yields the best results on the validation set within the range of 0.35 to 0.6.

[2] The reader should note that this result is not generalizable to other case studies as all our images had the same initial dimensions and were therefore distorted in the same way. The model could have learned to compensate for the constant distortion.

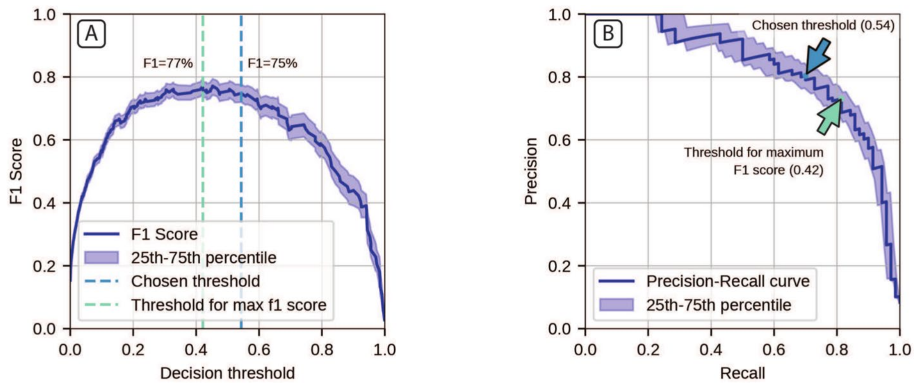

Fig. 3 **A** Effect of the decision threshold on the f1-score. **B** Trade-off between precision and recall in the validation set. Percentiles calculated using bootstrap resampling with 1000 iterations

We conduct a precision-recall analysis to set the threshold to a value within the range 0.35 to 0.6 (plot B in Fig. 3). Within the range of optimal F1-scores, precision is prioritized over recall. In other words, we prioritize being confident in labeling posts as positive at the cost of missing some, rather than aiming to avoid misses at the risk of overshooting. We anticipate significant redundancy in the data, as multiple posts may pertain to the same rioting event. This redundancy reduces the risk of missing riots when mapping their spatiotemporal distribution. Conversely, misclassifying a negative post as representing a riot leads to the erroneous identification of a non-existent event in the spatiotemporal distribution of riots. Based on these considerations, the decision threshold is set to 0.54, which maximizes the precision while maintaining a good f1-score on the validation set (0.75).

Spatiotemporal analysis

Beyond developing the computer vision model, this paper showcases how the method can be applied to reveal riot dynamics at high spatiotemporal resolution. This subsection introduces the analytical tools for this demonstration. We employ two approaches: quantifying clustering behavior using Ripley's spatiotemporal K function to measure burstiness, and applying density-based clustering to group posts into discrete rioting events. Section "Burst analysis" describes the burst analysis methodology. Section "Detection of riots" outlines the clustering approach for riot event detection.

Burst analysis

We quantify the burstiness of riot-related posts using Ripley's spatiotemporal K function, often used in crime epidemiology [4, 19, 22, 39]. This function, expressed in Eq. 1, measures the average number of neighbors each post has within a spatiotemporal neighborhood delimited by spatial neighborhood radius d and temporal neighborhood radius t, quantifying the density of neighboring posts within these radii. S is the spatial extent of the study region, T is the temporal extent, n is the total number of posts, d_{ij} and t_{ij} represent the spatial and temporal distances between posts i and j respectively, and $\mathbf{1}_{condition}$ is the indicator function. The factor w_{ij} corrects for edge effects that arise when spatiotemporal neighborhoods extend beyond the boundaries of the study region, see Gabriel [19] for more details.

$$K(d,t) = \frac{S \cdot T}{n^2} \sum_{i}^{n} \sum_{j,j \neq i}^{n} \frac{1}{w_{ij}} \cdot \mathbf{1}_{d_{ij} \leq d,\, t_{ij} \leq t} \tag{1}$$

The K function is commonly used to qualify a spatiotemporal process as bursty by comparing its value to a theoretical K function assuming random distribution across time and space. Values of $K(d, t)$ that exceed the expected under complete spatiotemporal randomness indicate clustering, while lower values suggest the presence of a regularity pattern. Social media posts are naturally highly clustered in space and time (users may upload several posts in the same location within short time intervals) and a random spatiotemporal process is therefore not a good baseline. We measure the burstiness of riot-related posts by comparing the K function to the K function of a baseline social media activity (non-riot-related posts).

We conduct separate analyses for each of the 7 cities included in our study, where the spatial extent S is defined by the geographical boundaries encompassing all posts within each urban area. The analysis is partitioned temporally on a day-by-day basis to capture daily patterns in riot activity. We measure the K function for riot-related posts during night-time hours (6 p.m. to 6 a.m.) and non-riot-related posts during daytime hours (6 a.m. to 6 p.m.), setting the temporal extent T to 12 h. We then compute the K function across a range of spatial radii d and temporal radii t to identify the characteristic scales at which riot events exhibit the strongest clustering. We then average the K function values across all cities and all days and calculate the ratio between the K function values for every pair of spatial and temporal radii.

Detection of riots

Rioting events cannot be readily derived from positive posts as (1) some posts may be misclassified and (2) several posts may represent the same event. We address these two issues using DBSCAN, a density-based cluster analysis, to group positive posts into rioting events [18]. This approach assigns positive posts that are close in space and time to rioting events while filtering out isolated posts. Several parameters govern the cluster analysis: the distance function, the distance threshold d_{th} under which two posts are considered neighbors, and the minimum number of neighboring posts n_{core} to form a cluster. The distance between two posts is calculated as the straight-line (Euclidean) distance between their spatial locations, plus a weighted measure of their temporal separation (see Eq. 2). In this equation, x_i and y_i represent the spatial coordinates of post i in meters and t_i is the instant at which the post is taken, in seconds. The parameter α relates temporal and spatial proximity. The parameters d_{th}, α, and n_{core} are set based on the results of the burst analysis in "Clustering analysis" Section.

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} + \alpha |t_j - t_i| \tag{2}$$

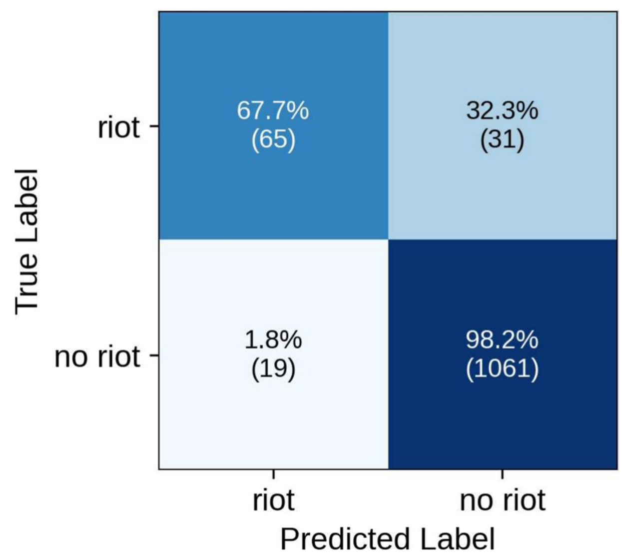

Fig. 4 Confusion matrix for the videos in the test set, normalized by true class

Results

This section validates our computer vision approach for detecting riot-related content in social media footage and demonstrates its application for mapping urban unrest. We first assess the performance of our computer vision model through quantitative analysis on a held-out test set, qualitative examination of misclassified cases, and cross-validation against an external press dataset (Sect. "Validation of the computer vision model"). We then showcase the analytical potential of our validated method by conducting fine-grained spatiotemporal analysis of riot evolution (Sect. "Spatiotemporal analysis of riots"), demonstrating capabilities that traditional data sources cannot provide.

Validation of the computer vision model

In the following, we demonstrate the performance of our method for identifying and mapping riots from user-generated footage. We analyze the performance of our computer vision model in identifying riot-related posts quantitatively in Sect. "Quantitative analysis of model performance" and assess misclassified posts qualitatively in Sect. "Qualitative analysis of misclassified footage". We then validate our approach by comparing the derived spatiotemporal distribution of riots with an external dataset (Sect. "Cross-validation with press data").

Quantitative analysis of model performance

We assess the performance of the computer vision model on the test set, held out of the training procedure (see Sect. "Social media data" for more details). Our assessment combines quantitative analysis using key performance indicators (accuracy, precision, recall) with qualitative analysis of misclassified footage to identify edge cases and understand classification errors. The test set contains 1176 pieces of footage, with 96 riot and 1080 non-riot instances. Given this imbalanced dataset, we focus on precision and recall

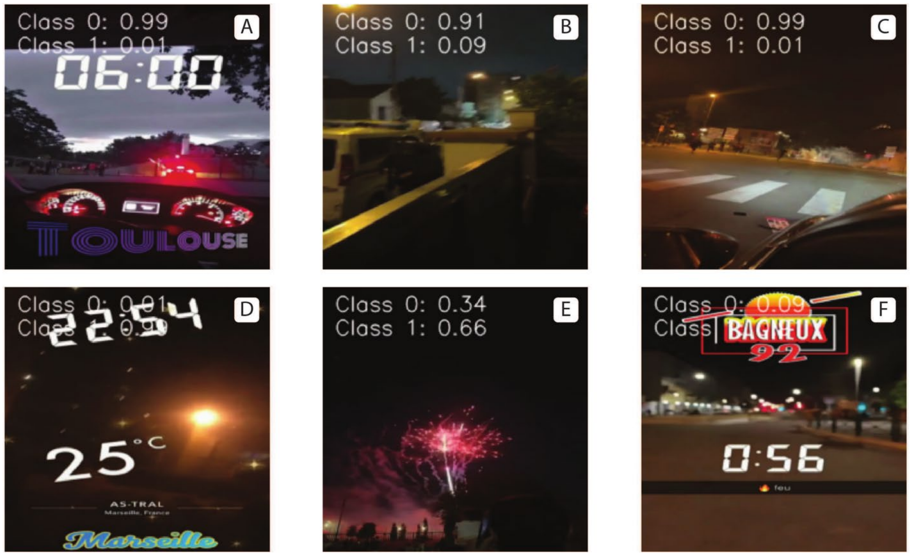

Fig. 5 Examples of ambiguous misclassified posts. **A**–**C** are false negatives, while **D**–**F** are false positives. The top left corner indicates the class probability for a riot (class 1) and "no riot" (class 0). Images have been transformed as they are when supplied to the model (vertical squeeze)

metrics rather than accuracy, as a naive classifier would achieve 92% accuracy by simply predicting the majority class. We separate classified images into true negatives, true positives, false negatives, and false positives (see confusion matrix in Fig. 4). As desired, our model achieves high precision in riot classification, measured at 77% (95% CI 69–86%[3]) on the test set. The recall of 68% (95% CI 58–77%) indicates that while most riot classifications are correct, the model misses approximately one-third of actual riot footage.

Qualitative analysis of misclassified footage

We complement our quantitative analysis of the model's performance with a qualitative analysis of misclassified images. The objective of the qualitative analysis is to examine misclassified images and assess potential impacts on the final results, particularly concerning systematic biases that could affect riot distribution mapping. For instance, if the model consistently identifies nightclub scenes as riots, nightclubs may (incorrectly) appear as hot spots when drawing the spatiotemporal distribution of riots. Of the 31 false negatives, the majority consist of looting scenes (14 instances), destruction events (13 instances), and fire-related incidents (4 instances). These misclassifications can be attributed to two primary factors. First, many videos contain only a few frames clearly depicting rioting activity, and when probabilities are averaged across all frames, the riot-related signal becomes diluted and falls below the classification threshold. Second, destruction and looting scenes are conceptually similar to non-riot activities from a visual perspective—people gathering or carrying objects—making them particularly challenging for the model to distinguish at the individual frame level, especially with limited training examples. The top row of Fig. 5 shows three three of such examples.

[3] Confidence intervals calculated using bootstrap resampling with 1000 iterations.

Similarly, of the 19 false positives, 9 are clearly negative, meaning they would have been correctly labeled by a human annotator. These clearly misclassified posts include four point-of-view driving scenes at night featuring bright halos from streetlights, two television screens displaying riot footage, two general nighttime scenes, and one daytime scene. The remaining 10 posts are ambiguous, as additional context would be needed to confidently classify these scenes as negative. We show three of these ambiguous false positives in the bottom row of Fig. 5. The analysis reveals no concerning patterns of misclassification that would significantly affect riot distribution mapping. Arguably, misclassified posts from the test set do not justify improving the model further, as they do not substantially affect the spatiotemporal distribution of riots. Since we do not observe any systematic bias in the test data, we have no reason to expect false positives to be non-randomly distributed across space; they would add white noise to the distribution. False negatives are less problematic as intense rioting events are likely to be reported by multiple posts.

Cross-validation with press data

After assessing the performance on the test set, we infer the classes of all posts in the full dataset. This section validates the spatial distribution of riots constructed from positive posts by comparing it with the press dataset in the Paris agglomeration (see Sect. "Press data"). We investigate the extent to which both datasets overlap spatially between June 27th and July 2nd in Fig. 6. For a given day, municipalities where riots were reported in the press dataset are highlighted in orange. Then, for each spatial unit, we measure the share of posts classified as riot-related among the unit's total number of posts during

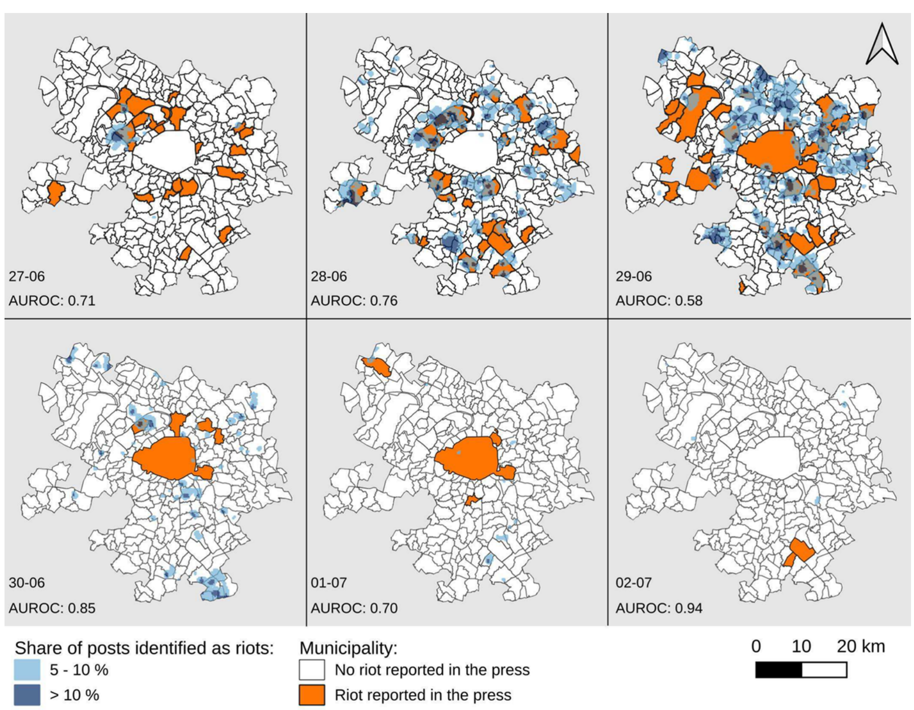

Fig. 6 Spatial overlap between rioting events reported in the press and rioting events reported on social media night after night

that given day. This share is shown in blue in Fig. 6. To provide a quantitative assessment of the correspondence between our approach and the press data, we calculate the Area Under the Receiver Operating Characteristic curve (AUROC) for each day. For each municipality and day, we compute the proportion of riot-related posts among all posts in that municipality as our predictor variable, with the binary outcome indicating whether the press dataset reported a riot event in that municipality on the given day (see Fig. 6).

The quantitative assessment using AUROC reveals variable correspondence between the datasets across different days. The correspondence is moderate to good throughout the full period (AUROC ranging from 0.70 to 0.94) except from June 29th, demonstrating that our approach can effectively discriminate between municipalities with and without press-reported riots on several days. However, correspondence is weaker on June 29th (AUROC = 0.58).

Several factors may explain the observed misalignment across different days. Our social media data collection faced incomplete coverage during early days (June 27th–28th), which may have affected detection accuracy (see Sect. "Social media data"). Additionally, the press dataset involves two distinct interpretation layers that can introduce discrepancies compared to real-time social media documentation: (1) journalist coverage decisions, where reporters make editorial choices about which events to cover based on perceived newsworthiness and (2) subsequent analyst data compilation, where researchers interpret and categorize press reports, potentially introducing additional filtering or classification biases. For example, our approach captures smaller incidents that press outlets might not consider newsworthy enough to report, potentially leading to spatial mismatches. These editorial decisions are further complicated by the temporal evolution of newsworthiness throughout the crisis. The press often prioritizes covering unexpected events during their initial phases, when public interest and the newsworthiness of the situation are at their peak. As a result, even relatively minor incidents may have been reported more extensively on June 27th, contributing to the higher coverage compared to subsequent days. From June 29th, the press reported fewer events than observed using our approach. This might indicate coverage fatigue by press outlets as the crisis progresses into its third day. Finally, temporal discrepancies also occur due to different conventions for dating events that happen after midnight. This temporal discrepancy is evident in the southeastern area of our study region, where municipalities appear as orange (indicating riot activity) on July 2nd in the press dataset, while we report these same events as occurring in the night of July 1st.

Our results demonstrate the ability to identify hot spots in the spatial distribution of riot-related posts that align with press data, indicating that the false negative rate does not hinder the clear identification of rioting events. Moreover, false positives do not appear to generate spurious hot spots. When rioting events are, in fact, more scarce and less intense, towards the end of the crisis, we observe little hot spots in the constructed spatial distribution of riots (bottom right map of Fig. 6). False positives are not concentrated enough to appear as rioting events. We can conclude that the false negative (32%) and the false positive (23%) rates observed on the test set in Sect. "Quantitative analysis of model performance" do not impact substantially the subsequent spatial analysis.

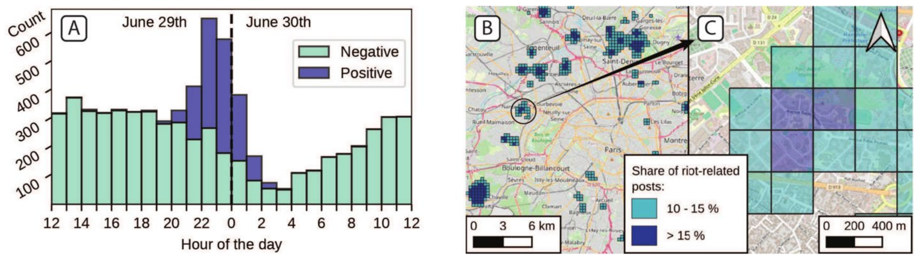

Fig. 7 **A** Temporal distribution of riot-related and non riot-related posts in the Paris agglomeration between June 29th and June 30th. **B** Spatial distribution of riot-related posts in the Paris agglomeration between June 29th and June 30th. **C** Zoom on a neighborhood in Nanterre

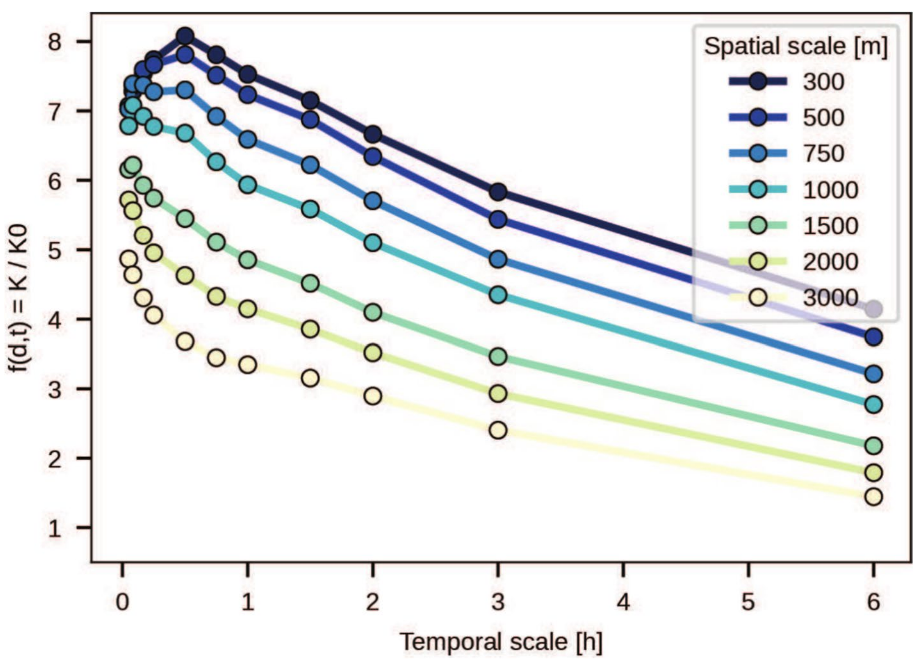

Fig. 8 Ratio *f* between the *K* function for riot-related posts and the *K* function for baseline social media activity (annotated K_0) for several temporal and spatial scale

Spatiotemporal analysis of riots

With our computer vision approach validated, we demonstrate its analytical potential by mapping riot dynamics at 300 × 300 m spatial resolution and hourly temporal intervals. We first conduct descriptive analysis of riot evolution during a high-intensity day (Sect. "Descriptive analysis"), then characterize the burstiness and spatiotemporal scales of riot activity (Sect. "Burst analysis"), and finally regroup posts into discrete rioting events through clustering analysis (Sect. "Clustering analysis").

Descriptive analysis

To demonstrate our approach's effectiveness in mapping riot evolution, we examine the Paris agglomeration during the night of June 29–30, the period with the highest number of recorded rioting events according to public authorities [43]. The histogram

A on Fig. 7 shows the hourly distribution of posts between June 29th midday and June 30th midday for the entire Paris agglomeration. The light green bars indicate the number of negative posts, while the blue bars reflect the number of riot-related posts. The number of negative posts increases in the morning from 5 a.m. to 10 a.m., reaches a peak between 10 a.m. and 7 p.m. and starts decreasing in the evening at 8 p.m., following typical daily activity patterns. In contrast, posts classified as riot-related exhibit a completely different temporal distribution. They occur at night, mostly between 8 p.m. and 3 a.m., peaking at 10 p.m., which is consistent with the rioting activities reported in the press [26, 27]. We also plot the spatial distribution of riot-related posts on the right maps of Fig. 7. With our data, we are able to draw the spatial extent and the intensity of riots across the night. This demonstrates one of the main advantages of our analysis compared to those relying on other data sources. It enables us to describe the evolution of riots hour by hour, at a 300×300 m resolution. Such data are instrumental in understanding the development and dynamics of riots, as they allow us to model the speed and direction in which events spread or compress.

Burst analysis

Throughout the period June 27th-July 2nd and across all 7 cities studied, riot-related posts reported on social media occur in short and localized bursts rather than in a smooth and progressive spatiotemporal distribution. We apply the spatiotemporal K function described in Sect. "Burst analysis" to characterize the clustering behavior of riot-related posts across different temporal and spatial scales. Figure 8 shows the ratio noted f between the K function for riot-related content compared to the baseline social media activity K_0. For all combinations of spatial and temporal radii (d, t), the K function for riot-related posts is consistently larger than the baseline's, implying that riot-related posts are spatially and temporally more concentrated than the benchmark.

Clustering analysis

We cluster social media posts into rioting events using the DBSCAN approach described in Sect. "Detection of riots". We set the α parameter relating the temporal dimension to the spatial dimension in the distance metric to 0.25 based on a sensitivity analysis of the spatiotemporal K function. The maximum of $f = K/K_0$ in Fig. 8 is reached for t_{max}=0.5 h, and d_{max}=300 m., the partial derivative analysis shows that burstiness decays four times stronger with distance than with time[4] ($\frac{\partial f/\partial d}{\partial f/\partial t} = 4$). We want our clustering algorithm to reflect this asymmetry by being more restrictive spatially than temporally. Therefore, we set α to 0.25 (see Eq. 4) to reduce the weight of the temporal dimension in our distance calculation. This allows posts separated by longer time intervals to be grouped into the same riot event while maintaining strict spatial proximity requirements. The distance threshold d_{th} is set to 750, to match the function's maximum (see Eq. 5). The minimum number of neighbors n_{core} to form a cluster is set to 3, which minimizes the number of isolated points (points belonging to no clusters).

[4] When time is expressed in seconds and space in meters.

Table 1 Characteristic spatial and temporal scale of rioting clusters across cities and days with more than 40 riot-related posts

City	Date (dd-mm)	Riot-related posts	Events detected	Median posts per event	Median surface of events (km^2)	Median duration of events (hh:mm:ss)
Paris	27-06	170	19	5	0.141	00:25:07
Paris	28-06	1020	103	5	0.141	00:43:38
Paris	29-06	1370	146	5	0.114	00:38:24
Paris	30-06	491	65	4	0.141	00:16:36
Paris	01-07	203	16	4	0.212	00:42:02
Paris	02-07	52	6	4	0.071	00:05:13
Marseille	30-06	138	10	6	0.212	00:48:07
Marseille	01-07	122	4	6	0.706	02:09:27
Lyon	29-06	123	16	4	0.071	00:41:07
Lyon	30-06	116	17	4	0.071	00:30:21
Lyon	01-07	54	5	4	0.071	00:23:31
Lille	28-06	84	9	4	0.071	01:07:11
Lille	29-06	221	17	5	0.141	01:33:33
Lille	30-06	43	6	6	0.071	00:31:03
Toulouse	28-06	51	7	5	0.212	00:42:28

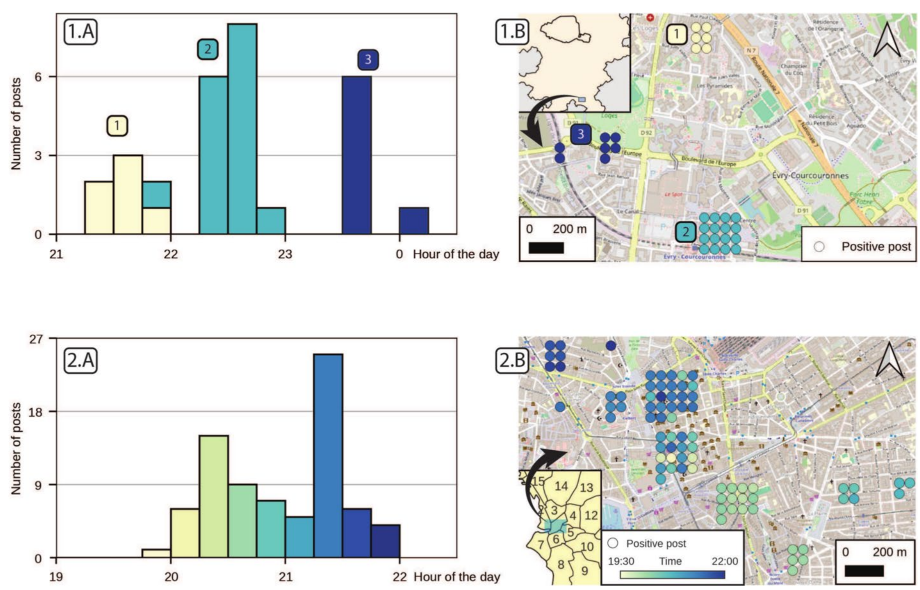

Fig. 9 1. Three bursts identified in the temporal (1.A) and spatial (1.B) distribution of riot-related posts in Évry-Courcouronnes in the night of June 29th. Bursts are labeled by a color and a number. 2. Temporal (2.A) and spatial (2.B) distribution of riot-related posts observed in the Old Port of Marseille in the night of July 1st. The number in map 2.B annotate the districts of Marseille

$$f(d,t) = K(d,t)/K_0(d,t) \tag{3}$$

$$\alpha = \frac{\partial f(d = d_{max}, t = t_{max})/\partial t}{\partial f(d = d_{max}, t = t_{max})/\partial d} = 1/4 \tag{4}$$

$$d_{th} = d_{max} + \alpha \cdot t_{max} \tag{5}$$

Table 1 summarizes the results of the clustering analysis for the cities and days with more than 40 riot-related posts. The average spatial and temporal extent of the clusters confirms the short and localized burst trend observed in the analysis of the spatiotemporal K function. The vast majority (80%) of identified riot clusters last no more than 50 min and remain confined to areas under 0.3 km^2.

One possible explanation for this pattern is the decentralized organization of rioting behavior, where rioters operate in small groups with separate objectives, resulting in localized and short-lived events. Another contributing factor could be police interventions; rapid and targeted actions may disperse a riot, prompting it to re-form at another location. The municipality of Évry-Courcouronnes on the night of June 29th illustrates the typical burst pattern observed throughout our dataset. We identify three intense bursts separated by spatiotemporal gaps of around 800 ms and 1 h (see histogram 1.A and map 1.B in Fig. 9). Yet, we do observe several exceptions such as a riot occurring in the neighborhood of the Old Port of Marseille on July 1st (see histogram 2.A and map 2.B in Fig. 9 and Table 1). The distribution of positive posts is much more homogeneous in space and in time than in the other riots observed in our dataset. Here, rioters may all have shared the same target neighborhood, the Old Port, being central and commercial, overwhelming public authorities.

Discussion and conclusion

This study highlights the potential of social media data for mapping riot-related events at a high spatiotemporal resolution. By employing a computer vision approach, we analyze geolocated visual content at scale, moving beyond the textual data and structured metadata used in prior research. The results demonstrate strong alignment with press-reported data, supporting the validity of our method and reinforcing the value of social media data in studying urban unrest.

In addition to demonstrating significant potential for understanding riot dynamics, our approach also gives rise to concerning possibilities for surveillance overreach and misuse beyond research contexts [32]. Real-time deployment could enable authorities to preemptively disperse peaceful protests before they escalate, potentially infringing on democratic rights to assembly. Beyond immediate deployment concerns, the potential identification of individuals through visual content becomes especially problematic if such frameworks target legitimate protests rather than violent unrest [40]. Additionally, historical riot mapping could justify disproportionate surveillance in specific neighborhoods, perpetuating existing policing inequities through location-based profiling [2]. These dual-use implications highlight the critical need for responsible implementations of such research developments.

Our work presents several limitations due to the use of social media data. First, the data collection is highly dependent on social media platforms and public authorities. Platforms can limit access to their data, provide biased content by blocking or prioritizing certain posts, or even spread fake news [45, 50, 55]. Similarly, public authorities can also temporarily ban or restrict the usage of the platform [6]. In the case study, we successfully collected geolocated social media posts and cross-validated them with press data, though the applicability of this method in future studies will depend on continued data access from social media platforms. Second, user-generated data introduces subjectivity [31]. Even visual posts, though generally more objective than text, may provide an incomplete representation of the scene. Third, platform usage varies geographically and demographically across populations, potentially limiting result generalizability [9]. Fourth, due to the sensitivity of the data used, the raw data cannot be shared openly. This hampers the replicability of the results [42]. Yet, we do share publicly the spatiotemporal distribution of the annotated posts [41].

While presenting notable challenges, our approach opens two promising avenues for further research: one for understanding the dynamics and another for investigating the causes of riots. First, the high-resolution data produced by our approach are instrumental for studies modeling the dynamics of riots [8, 11, 13]. For example, our data can help researchers understand how riots spread, how quickly they escalate, and how localized riots transform into large-scale uprisings, which is instrumental to designing effective crisis management strategies. In particular, agent-based modeling approaches could prove especially valuable in unraveling the microscopic behavioral rules-whether at individual or small—group levels—that aggregate to produce the fragmented spatiotemporal patterns we observe. Second, the data can support studies investigating what motivates participation in riots. Spatial demographic analysis could examine whether rioting intensity correlates with indicators of inequality or urban deprivation at the neighborhood or city level, leveraging our high-resolution data to align with granular socioeconomic datasets [21, 23, 29]. Additionally, investigating the relationship between online mobilization and physical manifestations of unrest could reveal how digital engagement translates into real-world action. This could involve analyzing textual social media data from platforms like X to measure online support for causes and examining how such digital mobilization correlates with the intensity and geographic distribution of actual rioting events captured through our computer vision approach. Such insights are not only valuable for academic understanding but could ultimately inform more effective approaches to addressing and preventing social tensions in urban environments.

Appendix: Anonymization process

To minimize identifiable information in the data, we blur the faces of people appearing on the pictures and videos. We first use an object detection model to delineate boxes containing faces on an image. Then, we apply a Gaussian blur on each of the box. We have considered several methods for detecting faces: Haar cascade classifier, Multi-task Cascaded Convolutional Networks (MTCCN), and “You only look once” (YOLO) [38, 48, 54]. Haar cascade is fast to run but the accuracy is too poor: for half of the videos tested, a face was missed in at least one frame. MTCCN showed acceptable accuracy but the computational time was too long to be used in our case study (around one second per frame). The YOLO model also showed high accuracy and the running time was small enough to be able to scale the approach to the entire dataset. We used the YOLOv8 implementation by Ultralytics [47] finetuned for face detection. After detecting faces, we blur the bounding box of each detected face using a Gaussian blur.

Acknowledgements
This work is supported by the TU Delft AI initiative. Francisco Garrido-Valenzuela and Abel Laval assisted in the data collection.

Author contributions
LS lead the conceptualization, collected and curated the data, developed the methodology, lead the formal analysis, drafted the manuscript. OC and SC participated in the conceptualization, the methodology, and the formal analysis, drafted and reviewed the manuscript, acquired the funding.

Funding
This work is fully funded by the TU Delft AI initiative.

Availability of data and materials
The metadata and classification results generated during the current study are available in the data repository, [41]. This includes timestamps, location data, and classification labels (true/false) for the analyzed social media posts. The original social media content is not included in the repository to protect user privacy, but the provided metadata is sufficient to interpret and build upon the findings reported in the article.

Declarations

Ethics approval and consent to participate
Ethics approval for this study was obtained from the Ethics Committee of Technische Universiteit Delft (id: 3664). The study analyzed publicly available social media data, which was collected and processed in anonymized form.

Consent for publication
Not applicable as this manuscript does not contain any individual person's identifiable data, images, or videos.

Competing interests
The authors declare no competing interests.**Received: 21 January 2025 Accepted: 14 July 2025**
Published online: 26 July 2025

References

1. Ahmed S, Cho J, Jaidka K. Framing social conflicts in news coverage and social media: a multicountry comparative study. Int Commun Gaz. 2019;81(4):346–71.
2. Alikhademi K, Drobina E, Prioleau D, Richardson B, Purves D, Gilbert JE. A review of predictive policing from the perspective of fairness. Artif Intell Law. 2022;30:1–17.
3. Alsaedi N, Burnap P, Rana O. Can we predict a riot? disruptive event detection using Twitter. ACM Trans Internet Technol TOIT. 2017;17(2):1–26.
4. Anselin L, Griffiths E, Tita G. Crime mapping and hot spot analysis. In: Environmental criminology and crime analysis. Willan; 2009. p. 119–38.
5. Aubert R, Dupont S, Laurent S, Pierre S. La cartographie d'une semaine d'émeutes en France. Le Monde, (2023, July). Retrieved from https://www.lemonde.fr/societe/article/2023/07/07/la-cartographied-une-semaine-d-emeutes-en-france_6180894_3224.html (https://www.lemonde.fr/societe/article/2023/07/07/la-cartographie-d-une-semaine-d-emeutes-en-france_6180894_3224.html). Accessed 06 Sept 2024
6. Baker SA. The English riots of 2011: a summer of discontent. Waterside Press; 2012. p. 169–90.
7. Baudains P, Braithwaite A, Johnson SD. Target choice during extreme events: a discrete spatial choice model of the 2011 London riots. Criminology. 2013;51(2):251–85.
8. Baudains P, Johnson SD, Braithwaite AM. Geographic patterns of diffusion in the 2011 London riots. Appl Geogr. 2013;45:211–9.
9. Bazzaz Abkenar S, Haghi Kashani M, Mahdipour E, Jameii SM. Big data analytics meets social media: a systematic review of techniques, open issues, and future directions. Telematics Inform. 2021;57: 101517. https://doi.org/10.1016/j.tele.2020.101517.
10. Beaman J. George Floyd across Europe, or Europe's George Floyds. Eur J Cult Polit Sociol. 2023;10(4):678–85.
11. Bonnasse-Gahot L, Berestycki H, Depuiset M-A, Gordon MB, Roché S, Rodriguez N, Nadal J-P. Epidemiological modelling of the 2005 French riots: a spreading wave and the role of contagion. Sci Rep. 2018;8(1):107.
12. Cicchelli V, Galland O, de Maillard J, Misset S. Retour sur les violences urbaines de l'automne 2005: Émeutes et émeutiers à Aulnay-sous-Bois. Horizons stratégiques. 2007;1:98–119.
13. Davies TP, Fry HM, Wilson AG, Bishop SR. A mathematical model of the London riots and their policing. Sci Rep. 2013;3(1):1303.
14. Davila A, Colan J, Hasegawa Y. Comparison of fine-tuning strategies for transfer learning in medical image classification. Image Vis Comput. 2024;146: 105012.
15. Dijkstra L, Poelman H, Veneri P. The EU-OECD definition of a functional urban area (Tech. Rep.). Organization for Economic Development and Cooperation. 2019. Retrieved from https://doi.org/10.1787/d58cb34d-en (Data available at https://www.oecd.org/cfe/regionaldevelopment/functional-urban-areas.htm). Accessed Aug 2024.
16. DiPasquale D, Glaeser EL. The Los Angeles riot and the economics of urban unrest. J Urban Econ. 1998;43(1):52–78.
17. Dosovitskiy A. An image is worth 16x16 words: transformers for image recognition at scale; 2020. arXiv preprint arXiv:2010.11929

18. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second international conference on knowledge discovery and data mining. 1996;96:226–31.
19. Gabriel E. Estimating second-order characteristics of inhomogeneous spatiotemporal point processes: influence of edge correction methods and intensity estimates. Methodol Comput Appl Probab. 2014;16:411–31.
20. George Y, Karunasekera S, Harwood A, Lim KH. Real-time spatiotemporal event detection on geotagged social media. J Big Data. 2021;8(1):91.
21. Gurr TR. Why men rebel. Routledge; 2015. p. 22–58.
22. He Z, Tao L, Xie Z, Xu C. Discovering spatial interaction patterns of near repeat crime by spatial association rules mining. Sci Rep. 2020;10(1):17262.
23. Holdo M, Bengtsson B. Marginalization and riots: a rationalistic explanation of urban unrest. Hous Theory Soc. 2020;37(2):162–79. https://doi.org/10.1080/14036096.2019.1578996.
24. Home Office. An overview of recorded crimes and arrests resulting from disorder events in August 2011. Crown Copyright; 2011. p. 4.
25. Lachman SJ. Psychological perspective for a theory of behavior during riots. Psychol Rep. 1996;79(3):739–44. https://doi.org/10.2466/pr0.1996.79.3.739.
26. Le Parisien (2023, June 30). Mort de Nahel : des scènes de pillages et de saccages en plein Paris. (https://www.leparisien.fr/faits-divers/mort-de-nahel-a-paris-des-scenes-de-pillages-et-de-saccages-30-06-2023-AZNCXGFSUFHQLL2STJBMB3KMEU.php). Accessed Sept 2024.
27. L'obs (2023, June 30). Mort de Nahel à Nanterre : Macron va présider une cellule de crise après une troisième nuit d'émeutes. (https://www.nouvelobs.com/societe/20230630.OBS75142/mort-de-nahel-a-nanterre-macron-va-presider-une-cellule-de-crise-apres-une-troisieme-nuit-d-emeutes.html). Accessed Sept 2024.
28. Moran M, Waddington D. Riots: an international comparison. London: Palgrave Macmillan; 2016. p. 1–14.
29. Newburn T. The causes and consequences of urban riot and unrest. Annu Rev Criminol. 2021;4(1):53–73.
30. Nicoletti L, Verma T, Warnier M, Santi P. Exploring the impact of hashtag activism on political processes using geolocated social media data (Unpublished master's thesis). TU Delft. (2021). (https://repository.tudelft.nl/record/uuid:a4e3c609-c4ea-48cb-b268-d5b77549efe6). Accessed Jan 2025.
31. Olteanu A, Castillo C, Diaz F, Kıcıman E. Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data. 2019;2:13.
32. Owen S. Monitoring social media and protest movements: ensuring political order through surveillance and surveillance discourse. Social Identities. 2017;23(6):688–700.
33. Panagiotopoulos P, Barnett J, Bigdeli AZ, Sams S. Social media in emergency management: Twitter as a tool for communicating risks to the public. Technol Forecast Soc Chang. 2016;111:86–96.
34. Prabhu A, Guhathakurta D, Subramanian M, Reddy M, Sehgal S, Karandikar T, et al. Capitol (pat) riots: a comparative study of Twitter and Parler; 2021. arXiv preprint arXiv:2101.06914
35. Procter R, Vis F, Voss A. Reading the riots on Twitter: methodological innovation for the analysis of big data. Int J Soc Res Methodol. 2013;16(3):197–214.
36. PyTorch Developers (2024). Torchvision: Datasets, transforms, and models specific to computer vision. (https://github.com/pytorch/vision). Accessed Jan 2025.
37. Recht B, Roelofs R, Schmidt L, Shankar V. Do imagenet classifiers generalize to imagenet? In: International conference on machine learning; 2019. p. 5389–400.
38. Redmon J, Farhadi A. Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 7263–71.
39. Ripley BD. The second-order analysis of stationary point processes. J Appl Probab. 1976;13(2):255–66.
40. Smith M, Miller S. The ethical application of biometric facial recognition technology. Ai & Soc. 2022;37(1):167–75.
41. Spierenburg L, Cats O, van Cranenburgh S. Data related to the paper Studying social unrest through the lens of social media. 4TU.ResearchData; 2024. https://doi.org/10.4121/649e8f5d-8e40-4ab7-9d07-b5ef53d810f0.v1)
42. Stodden V. Privacy, big data, and the public good: frameworks for engagement. Cambridge University Press; 2014. p. 112–35.
43. Sénat. Rapport d'information fait au nom de la commission des lois constitutionnelles, de législation, du suffrage universel, du règlement et d'administration générale (1), investie des pouvoirs d'une commission d'enquête, sur les émeutes survenues à compter du 27 juin 2023. Session ordinaire de 2023–2024 (p. 10, 162); 2024. https://www.senat.fr/rap/r23-521/r23-5211.pdf.
44. Tan M. Efficientnet: Rethinking model scaling for convolutional neural networks; 2019. arXiv preprint arXiv:1905.11946
45. Timoneda JC. Where in the world is my tweet: detecting irregular removal patterns on Twitter. PLoS ONE. 2018;13(9):e0203104.
46. Tonkin E, Pfeiffer HD, Tourte G. Twitter, information sharing and the London riots? Bull Am Soc Inf Sci Technol. 2012;38(2):49–57.
47. Ultralytics (2024). Yolo 8: yolov8n-face. https://github.com/ultralytics/ultralytics. Accessed Aug 2024.
48. Viola P, Jones MJ. Robust real-time face detection. Int J Comput Vision. 2004;57:137–54.
49. Vis F. Twitter as a reporting tool for breaking news: journalists tweeting the 2011 UK riots. Digit J. 2013;1(1):27–47.
50. Vosoughi S, Roy D, Aral S. The spread of true and false news online. Science. 2018;359(6380):1146–51. https://doi.org/10.1126/science.aap9559.
51. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data. 2016. https://doi.org/10.1186/s40537-016-0043-6.
52. Wilkinson SI. Riots. Annu Rev Polit Sci. 2009;12:329–43. https://doi.org/10.1146/annurev.polisci.12.041307.075517.
53. Wilson SL. Detecting mass protest through social media. J Soc Media Soc. 2017;6(2):5–25.
54. Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett. 2016;23(10):1499–503.

55. Zimmer F, Scheibe K, Stock M, Stock WG. Fake news in social media: bad algorithms or biased users? J Inf Sci Theory Pract. 2019;7(2):40–53.

Publisher's Note