Civil Unrest on Twitter (CUT):
A Dataset of Tweets to Support Research on Civil Unrest

Justin Sech*, Alexandra DeLucia*, Mark Dredze*, Anna L Buczak†

*Center for Language and Speech Processing, Johns Hopkins University
†Johns Hopkins University Applied Physics Laboratory
{jsech1, aadelucia, mdredze}@jhu.edu, Anna.Buczak@jhuapl.edu

Abstract

We present CUT, a dataset for studying Civil Unrest on Twitter. Our dataset includes 4,381 tweets related to civil unrest, hand-annotated with information related to the study of civil unrest discussion and events. Our dataset is drawn from 42 countries from 2014 to 2019. We present baseline systems trained on this data for the identification of tweets related to civil unrest. We include a discussion of ethical issues related to research on this topic.

1 Introduction

From the tomb-builder strikes in 1159 BCE Egypt[1] to the Black Lives Matter protests in 2020 CE U.S.[2], humanity has utilized protests and other demonstrations to register grievances and affect change in government or society. While some basic elements of protest remain unchanged for thousands of years, the ability to organize and execute civil unrest activities has been transformed by social media.

Twitter has played a central role in several recent civil unrest related activities, with the most noted being the Arab Spring, a series of pro-democratic uprisings, protests, and armed rebellions from 2010 to 2012 in Tunisia, Morocco, Syria, Libya, Egypt and Bahrain, that led to regime changes[3]. Some sociologists believe that the distrust in official country press due to censorship led to civilians turning to each other on social media for independent news (Smidi and Shahin, 2017; Soengas-Pérez, 2013). In addition to spreading news, civilians also use social media to share opinions on new policies or recent events (e.g. political debates), and to share information for upcoming events (e.g. protests). The use of social media to plan protests was a key motivator for the Planned Protest module in the

is event: specific (y)   sentiment: neither
unrest: no   topic: no
timing: past   participation: neither

is event: nonspecific (y)   sentiment: support
unrest: yes   topic: yes
timing: current   participation: neither

Figure 1: Example tweets and their annotations in the Civil Unrest on Twitter (CUT) dataset.

EMBERs civil unrest forecasting system (Muthiah et al., 2015; Ramakrishnan et al., 2014).

We study civil unrest discussions on Twitter for the exact same reason Tunisians turned to Twitter—we believe civilian voices are an important source of information about the state of a country. Ramakrishnan et al. (2014) hinted at this with their paper title, "Beating the News," for what is news other than reports of the people to the people? While news articles can inform about the presence of civil unrest, we (and other researchers who utilize Twitter data) seek to find information about events and opinions from Twitter before official news reports

---

[1] https://www.ancient.eu/article/1089/the-first-labor-strike-in-history/

[2] https://blacklivesmatter.com

[3] https://www.history.com/topics/middle-east/arab-spring

(Osborne and Dredze, 2014).

We discuss ethical concerns of analyzing civil unrest data in §5.

A challenge to studying civil unrest in social media is finding it. Tweets cover a wide range of topics, and identifying those directly relevant to a protest can be challenging. Therefore, to support the study of civil unrest from Twitter, we present the Civil Unrest on Twitter (CUT) dataset: a collection of 4,381 tweets with annotations for a variety of information related to civil unrest. Examples of annotated tweets are shown in the appendix (Figure 1). Tweets are labeled for the following: whether a tweet refers to a protest/strike/riot, if general unrest/dissatisfaction is conveyed, the time of the tweet with respect to the event, the user's stance, whether the event topic is present, if the user intends to participate, and event-specific hashtags. These annotations are useful for a variety of tasks, such as stance detection and event extraction (Mohammad et al., 2016; Zong et al., 2020). As an example use case, we create a model that distinguishes between tweets that discuss civil unrest events from those that do not. This filtration model is a popular processing step in civil unrest detection and forecasting pipelines (Islam et al., 2020; Alsaedi et al., 2017; Edouard, 2018; Korolov et al., 2016; Ranganath et al., 2016).

We make the following contributions:

- CUT: A dataset of 4,381 English Tweets from 42 African, Middle Eastern, and Southeast Asian countries (2014-2019), annotated for a variety of information of interest with respect to civil unrest.

- Baseline classifiers that determine if a tweet is related to a civil unrest event.

Our dataset and code are available at `https://github.com/AADeLucia/JHU-CUT`.

2 Related Work

Several studies have examined specific events on social media. Examples include tracking information related to public health, such as COVID-19 (Zong et al., 2020; Paul and Dredze, 2017), and riots, such as the London Riots (Alsaedi et al., 2017).

Several studies have specifically considered building datasets related to the task of civil unrest detection. Alsaedi et al. (2017) label a sample of 5,000 tweets from the Middle East from October to November 2015, however they were labelled for general events and not only disruptive events (e.g. weather). Islam et al. (2020) use a list of keywords to filter tweets, and then manually label a sample 10,500 tweets from 178 countries from November 26, 2017 to June 25, 2018 to verify their "informative" vs "uninformative" filtration model. De Silva and Riloff (2014) incorporated profile information (i.e. organization or user) to predict protest-related tweets from a collection of 6,000 English and Spanish disease and civil-unrest keyword-filtered tweets. Edouard (2018) created an Event2012/FSD corpus with roughly 155K tweets from late 2011 and late 2012, and labelled the tweets as event/non-event related according to a list of events from Wikipedia Current Event Portal. Korolov et al. (2016) purchased third-party curated tweets related to the 2015 Baltimore protests and manually annotated those to filter related tweets in a larger set from the same time period in large U.S. cities. Ranganath et al. (2016) collected geotagged messages from Nigeria and manually labelled 2,686 tweets for their relation to the 2015 general election and Boko Haram insurgencies.

Other work collects tweets from a known event with location and hashtag filters. Wang et al. (2015) collected 6.5 million tweets from the dates and locations that were affected by Hurricane Sandy (October 22nd, 2012–November 2nd, 2012, northeastern US). Littman (2018) collected 7.6 millions tweets from the 2017 "Unite the Right" protest in Charlottesville, Virginia with event-specific hashtags (e.g. #defendCville, #HeatherHeye).

Other work uses external information from news articles and other sources to categorize groups of tweets, e.g. country and day, but not to identify individual tweets related to an event (Korkmaz et al., 2016; Chen and Neill, 2014).

A drawback of these efforts is that they focus on specific events or locations, rather than producing a more general dataset that can be used to identify civil unrest tweets from new or emerging events. Our goal is to produce a more general dataset from a large number of countries (42) over several years (2014 to 2019).

3 Dataset Creation

We present a dataset to support the study of civil unrest on Twitter. Our data set contains English tweets with annotations related to civil unrest produced through annotations from Amazon Mechani-

cal Turk.

Twitter Data   We selected a sample of tweets from 2014 to 2019 that were collected from the Twitter streaming API based on filters to collect geolocated data. Our geolocation filters included African, Middle Eastern, and Southeast Asian countries (see Table 4 in the appendix.) We filtered the data set to include English only tweets as identified by `langid` (Lui and Baldwin, 2012).[4] We excluded retweets. Geolocation is present in every tweet due to the method of collection.

We select tweets based on their inclusion of an English language keyword related to civil unrest. Using an approach similar to that of Muthiah et al. (2015) and Ramakrishnan et al. (2014), we used a combination of manual and automated methods to create a large set of 709 keywords, which include terms such as "unemployment," "police," and "extremist." The full list of keywords appears with the released dataset.

In total, we include 4,415 tweets in the dataset for annotation.[5] 34 Tweets are removed from the dataset after annotators report them as non-English, resulting in the final dataset of 4,381 Tweets.

Annotations   Our goal was to collect a wide range of annotations that could potentially be helpful in the study of civil unrest on Twitter. Annotators were asked several questions about each tweet:

1. Does this Tweet discuss a protest, march, riot, or strike?

    (a) At the time of this Tweet, is the referenced event currently in progress, in the past, or an upcoming event?
    (b) Does this Tweet support or oppose the event in question?
    (c) Does this Tweet state a specific topic of the event that reflects the intent of the protesters?
    (d) Does this Tweet describe participation/intent to participate in the event?
    (e) If this Tweet contains hashtags specific to the event, list the hashtags.

2. Does this Tweet indicate civil or political unrest, frustration, or dissatisfaction? For example, dissatisfaction with government policy, economic situation, etc.

Questions 1(b)-1(e) were only answered if the answer to (1) was 'specific' or 'nonspecific', and Question 1(a) was only answered if (1) was 'specific.' A screenshot of the survey is in the appendix (Figure 2).

Survey Setup   We obtained annotations using Amazon Mechanical Turk. Our HIT contained 10 tweets, and each HIT was annotated by 3 workers.

To ensure a balanced inclusion of different countries and time periods in the annotated set, we selected tweets uniformly by country and year, i.e. each country had the same probability of having a tweet included in the annotated set.

To ensure annotation quality we release the HITs in batches and perform a quality check by inserting pre-annotated tweets into each HIT. These 100 quality check tweets were manually annotated by two annotators (one an author of this paper). Conflicting annotations were adjudicated by the author annotator. If an annotator incorrectly annotates a quality check tweet, then their work is set aside to be inspected by the author annotator. If their work is considered unsatisfactory (i.e. the author had reason to believe that the annotator was simply clicking through) then their annotations are removed[6]. Workers were paid $0.40/HIT, for 500 HITs with 3 annotators each, for a total cost of $200.

Our first batch of annotations had a very low rate of civil unrest related tweets (7%). This is likely because of the breadth of keywords used to filter tweets and polysemy (e.g. "guns" could refer to weapons or muscles). This issue was also encountered De Silva and Riloff (2014), where 80% of the tweets collected through keyword filtering alone did not discuss events. Therefore, we sought to bias future annotation rounds towards more civil unrest related tweets. We trained a Random Forest classifier on the tweets from the first round of annotations using features of unigram counts.

While the classifier only achieved an F1 of 0.502, we found that the highly scored tweets were much more likely to be about civil unrest.

We then included the keyword feature importance as weights for sampling the next batch of

---

[4]While Twitter includes language identification in tweets, this feature was introduced after the start of our data collection. For consistency, we used `langid` for all tweets.

[5]The odd number is due to the presence of pre-labeled tweets in HITs for quality control, and a handful of tweets that were sampled multiple times.

[6] All ties were broken by author annotations or 2 volunteer annotators.

Rank	Keyword	Importance
1	protest	0.198
2	protests	0.117
3	protesters	0.077
4	strike	0.047
5	violence	0.020
6	fight	0.019
7	campaign	0.015
8	detained	0.014
9	demonstrations	0.014
10	protesting	0.013

Table 1: Top 10 important keywords for sampling tweets obtained from a random forest model.

Question (IAA)	Annotations		
1 (0.430)	yes, specific 539	yes, non-specific 151	no 3,691
1(a) (0.478)	current 381	past 111	future 47
1(b) (0.325)	support 196	oppose 69	neither 425
1(c) (0.312)	yes 322	no 364	unclear 4
1(d) (0.183)	yes 65	no 4	neither 621
2 (0.168)	yes 1,951	no 2,446	

Table 2: Annotation results from the Mechanical Turk survey. 4,415 tweets were labelled by three annotators for a variety of civil unrest-related questions (§3). 34 tweets were labelled as "not English" and removed. The inter-annotator agreement (IAA) was calculated using Fleiss' kappa for 3 annotators, before the "unsatisfactory" workers were removed. Due to the overriding nature of our adjudications, recalculating agreement would be inaccurate.

tweets. If a tweet contained words from the top of the important keyword list, it had a higher chance of being selected. The top keywords are in Table 1. The concept of not treating all keywords as equal was also brought up in Islam et al. (2020), where they categorized their civil unrest "keyword dictionary" into ranked categories based "negative impact of an unrest event on civil life." The third and final batch was sampled the same way as batch two (using the keyword weights).

For each question we selected the majority label from the three annotators. If no majority label existed, then the answer was adjudicated by the authors[7].

Annotated Dataset Table 2 shows statistics for the final dataset. Despite efforts to increase the number of civil unrest-related tweets, only 690 of 4,381 tweets were about events, but 1,951 tweets did contain signs of general unrest. Annotators labeled 34 Tweets from the original 4,415 dataset as "not English," and we removed those from the final dataset.

After manually inspecting some of the provided hashtags, we decided that the hashtag question contained the least quality answers. The answers were mostly blank. We suspect this is due to answering the hashtag free-response question is more time consuming than the other multiple choice questions. For greater coverage, we will include all listed hashtags in the released dataset instead of the hashtags listed by at least 2 annotators.

In terms of annotator agreement, question 2 had the lowest agreement with a Fleiss' kappa of 0.168 and question 1a the highest (0.478) (See Table 2). Lower agreement rates are not uncharacteristic for labeling tweets. One potential difficulty in our setting is that our questions are very specific. While

other datasets ask about "event vs. no event," we ask for more details.

4 Civil Unrest Classification

Using our annotated dataset we created a baseline model for predicting if a tweet was related to civil unrest, i.e. predicting if the label was "yes, a specific event" or "yes, in a non-specific fashion" for question 1 (690 of 4,381 tweets or 16%). The resulting classifier can be used to identify large amounts of data around specific events for further study (Islam et al., 2020; Alsaedi et al., 2017; Edouard, 2018; Korolov et al., 2016; Ranganath et al., 2016).

We considered two logistic regression classifiers. 1) Unigram counts of all tokens in a tweet. 2) Counts of the civil unrest keywords only. Both of these methods used the scikit learn implementation of logistic regression and `CountVectorizer` (Pedregosa et al., 2011). The unigram models were regularized with L2 loss and were evaluated with 5-fold cross validation (same folds across experiments).

All methods preprocessed tweets with the `littlebird` implementation of the BERTweet tokenizer (DeLucia, 2020). This tokenizer was used to allow easy extendability and future comparison with a BERTweet-based model.

Table 3 shows that the keyword based logistic

---
[7]See footnote 6

Features	F1	Precision	Recall
Keywords	0.782	0.894	0.697
Unigrams	0.775	0.892	0.687

Table 3: Logistic regression results on predicting if a tweet is related to civil unrest. Standard deviation across 5-fold cross validation (F1) for keywords is 0.023, unigrams is 0.030. F1 is for the positive class only (i.e. events).

regression model outperformed the unigram model (0.782 F1 vs 0.775 F1). This out-performance is slight, but still worthwhile, especially since the keyword model is smaller (only 700 features instead of 15K) and converges faster (roughly 50 iterations versus 100). Additionally, since the civil unrest keywords were used to sample tweets their inclusion as features adds further information to the keyword-based model.

5 Ethical Considerations

Many studies of event detection on Twitter have explored the implications for public health (e.g. spread of infectious diseases) (Paul and Dredze, 2017) and natural disasters (Wang et al., 2015), which offer benefits in combating harmful events. However, civil unrest presents a more complex cost benefit trade off as it can yield insights into what issues are most important to a population, but can also be used to monitor or track individuals who participate in these events. Deciding what constitutes civil unrest versus unjustified violence requires a value judgement, which could easily degrade into weaponizing against dissenting opinion. Additionally, non-government actors could use predictions of unrest to squash disapproving voices. Moreover, frequently, marginalized voices have found solace and organization using social media (Xiong et al., 2019; Ince et al., 2017) and predicting civil unrest could unintentionally lead to actions such as further policing of overpoliced communities.

With this in mind, we should consider Twitter data not just as text data, but as people. Several proposals for protecting people including avoiding reverse identification (Ayers et al., 2018; Benton et al., 2017)) and data anonymization tools (Nguyen-Son et al., 2012). We believe that the numerous studies of civil unrest that further our understanding of complex societal issues are convincing evidence that there is much to be gained

from developing data resources in support of this topic. At the same time, we must remain vigilant in our evaluation of research efforts to ensure they remain supportive of these goals.

6 Conclusion

We have presented the Civil Unrest on Twitter (CUT) dataset and a baseline classifier trained on the data for identifying tweets related to a civil unrest event. Future work can build on our multi-faceted annotations to expand the study of communities and how they express concern about complex societal issues through civil unrest.

Acknowledgments

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.

John W. Ayers, Theodore L. Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *npj Digital Medicine*, 1(1):1–2. Number: 1 Publisher: Nature Publishing Group.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Feng Chen and Daniel B. Neill. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD*

'14, pages 1166–1175, New York, New York, USA. ACM Press.

Lalindra De Silva and Ellen Riloff. 2014. User Type Classification of Tweets with Implications for Event Recognition. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 98–108, Baltimore, Maryland. Association for Computational Linguistics.

Alexandra DeLucia. 2020. AADeLucia/littlebird. Original-date: 2020-04-24T00:15:07Z.

Amosse Edouard. 2018. *Event detection and analysis on short text messages*. Ph.D. thesis.

Jelani Ince, Fabio Rojas, and Clayton A. Davis. 2017. The social media response to black lives matter: how twitter users interact with black lives matter through hashtag use. *Ethnic and Racial Studies*, 40(11):1814–1830.

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.

Gizem Korkmaz, Jose Cadena, Chris J. Kuhlman, Achla Marathe, Anil Vullikanti, and Naren Ramakrishnan. 2016. Multi-source models for civil unrest forecasting. *Social Network Analysis and Mining*, 6(1):50.

Rostyslav Korolov, Di Lu, Jingjing Wang, Guangyu Zhou, Claire Bonial, Clare Voss, Lance Kaplan, William Wallace, Jiawei Han, and Heng Ji. 2016. On predicting social unrest using social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 89–95, San Francisco, CA, USA. IEEE.

Justin Littman. 2018. Charlottesville Tweet Ids. Publisher: Harvard Dataverse type: dataset.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

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*.

H. Nguyen-Son, Q. Nguyen, M. Tran, D. Nguyen, H. Yoshiura, and I. Echizen. 2012. Automatic anonymization of natural languages texts posted on social networking services and automatic detection of disclosure. In *2012 Seventh International Conference on Availability, Reliability and Security*, pages 358–364.

Miles Osborne and Mark Dredze. 2014. Facebook, twitter and google plus for breaking news: Is there a winner? In *ICWSM*.

Michael J. Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183.

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

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.

Suhas Ranganath, Fred Morstatter, Xia Hu, Jiliang Tang, and Huan Liu. 2016. Predicting Online Protest Participation of Social Media Users. *AAAI Press*, page 7.

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.

Haoyu Wang, E. Hovy, and Mark Dredze. 2015. The hurricane sandy twitter corpus. In *AAAI Workshop: WWW and Public Health Intelligence*.

Ying Xiong, Moonhee Cho, and Brandon Boatwright. 2019. Hashtag activism and message frames among social movement organizations: Semantic network

analysis and thematic analysis of twitter during the #metoo movement. *Public Relations Review*, 45(1):10 – 23.

Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. 2020. Extracting COVID-19 Events from Twitter. *arXiv:2006.02567 [cs]*. ArXiv: 2006.02567.

AO	100	BD	109	BF	100	BI	105	BJ	106	CD	101	CF	91
CI	101	CM	103	DZ	104	EG	104	ET	108	GH	110	GN	95
KE	108	KH	105	LK	102	LR	102	LY	107	MA	105	MG	102
ML	97	MM	105	MW	103	MZ	96	NA	109	NE	100	NG	134
NP	109	PK	123	SD	106	SL	106	SN	105	SO	111	TD	89
TG	98	TH	107	TN	111	TZ	107	UG	100	ZA	129	ZM	94

Table 4: Country breakdown of the 4,415 annotated Tweets. 34 tweets were removed after annotation because they were marked as "not English."

A Dataset Breakdown

This section is a more detailed breakdown of the Twitter data used for annotation in §3. Tweets from 42 countries were annotated. The number of tweets per country is in Table 4.

Figure 2: Questions presented to Amazon Mechanical Turk workers. Grayed-out questions are disabled unless previous questions are given specific answers. The questions are displayed for each Tweet, with 10 Tweets per HIT.

B Annotation Details

A screenshot of the survey presented to annotators is in Figure 2.