Study of Manifestation of Civil Unrest on Twitter

Abhinav Chinta[*] and **Jingyu Zhang**[*] and **Alexandra DeLucia** and **Mark Dredze**
Center for Language and Speech Processing, Johns Hopkins University
{achinta3, jzhan237, aadelucia, mdredze}@jhu.edu

Anna L Buczak
Johns Hopkins University Applied Physics Laboratory
Anna.Buczak@jhuapl.edu

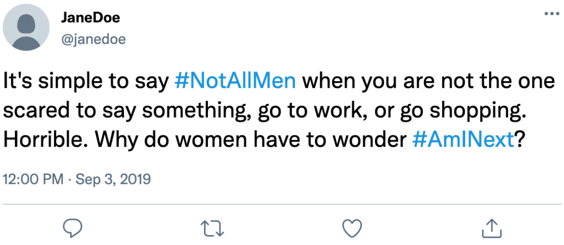
Abstract

Twitter is commonly used for civil unrest detection and forecasting tasks, but there is a lack of work in evaluating *how* civil unrest manifests on Twitter across countries and events. We present two in-depth case studies for two specific large-scale events, one in a country with high (English) Twitter usage (Johannesburg riots in South Africa) and one in a country with low Twitter usage (Burayu massacre protests in Ethiopia). We show that while there is event signal during the events, there is little signal leading up to the events. In addition to the case studies, we train n-gram-based models on a larger set of Twitter civil unrest data across time, events, and countries and use machine learning explainability tools (SHAP) to identify important features. The models were able to find words indicative of civil unrest that generalized across countries. The 42 countries span Africa, Middle East, and Southeast Asia and the events range occur between 2014 and 2019.

1 Introduction

Citizens utilize public demonstrations, protests, and in extreme cases, riots, to express dissatisfaction over the current political or social state in their country. Some of these movements successfully achieve government reforms, many do not. These movements are often driven from the ground up, by grassroots advocacy that snowballs into social and political change. Understanding the factors that drive sociopolitical change can help policy makers and advocates petition for and advance their causes.

Since these causes emerge from the public, studying them requires data on public attitudes, perceptions, and actions around previous movements. In particular, understanding the factors behind civil unrest based on social media posts can reveal important trends. However, while social media data reflects many public attitudes, it's unclear how sites such as Twitter capture activities around civil unrest. Previous work has considered the task of predicting protests or riots using social media, often with the aid of news sources (Alsaedi et al., 2017; Ramakrishnan et al., 2014; Osborne and Dredze, 2014; Islam et al., 2020; Edouard, 2018; Korolov et al., 2016; Ranganath et al., 2016; De Silva and Riloff, 2014; Littman, 2018).



JaneDoe
@janedoe

It's simple to say #NotAllMen when you are not the one scared to say something, go to work, or go shopping. Horrible. Why do women have to wonder #AmINext?

12:00 PM · Sep 3, 2019

Figure 1: Example tweets from Johannesburg riots in South Africa. The tweets were found through their use of event-related hashtags. Paraphrased for privacy.

In this paper we consider how these activities manifest themselves on Twitter. Specifically, we consider events across a range of countries to understand how Twitter is used by different populations. Our analysis has two facets: (1) qualitative case studies focusing on two specific events in countries with different civil unrest rates and Twitter usage patterns and (2) a quantitative analysis to find indicators of civil unrest that generalize across countries and events.

For the case studies, we look at the 2019 Johannesburg riots in South Africa and the 2018 Burayu Massacre and subsequent protests in Ethiopia. We chose these events because they were significant within those countries and received global attention. Their scale should mean they are reflected on social media. Furthermore, these countries have high (South Africa) and low (Ethiopia) Twitter usage, and the majority of their tweets are in English. For each event we ask: (1) is the event discussed on Twitter? (2) is there any noticeable buildup to the

[*] Equal contribution.

event? (3) *who* is talking about the event? (4) do tweets reflect the issues that motivated the events?

We then take a quantitative approach to find generalizable indicators of civil unrest across time, events, and countries. This focus on generalizability is motivated by previous work that detects/forecasts unrest across countries or regions (Alsaedi et al., 2017; Ramakrishnan et al., 2014; Islam et al., 2020; Korolov et al., 2016; Ranganath et al., 2016), and the desire to have a single model perform well across countries. For this analysis we introduce n-gram models trained on a new dataset, Global Civil Unrest on Twitter, or G-CUT, which spans 42 countries from Africa, the Middle East, and Southeast Asia, and riots and protests from 2014 to 2019. We evaluate model generalizability by checking not only the overall performance, but also the per-country performance. Also, we use AI explainability tools (SHAP) to analyze the top indicators and ensure they are not country- or event-specific.

All data (tweet IDs), models, and code will be made publicly available at `https://github.com/AADeLucia/civil-unrest-case-study`.

2 Data

We analyze Twitter since it is a widely used social media platform, with many public posts about major events in real-time. For ground-truth data on civil unrest we use the Armed Conflict Location & Event Data Project (ACLED) (Raleigh et al., 2010), a manually curated database tracking civil unrest events.

2.1 Twitter Collection

We collected geotagged tweets from the Twitter streaming API from 2014 to 2019 (inclusive) for 42 countries from Africa, the Middle East, and Southeast Asia. We selected English tweets using `langid` (Lui and Baldwin, 2012).[1] Appendix Table 4 lists the number of tweets per country.

To focus on relevant tweets we filter our data using the BERTweet civil unrest tweet classifier from Sech et al. (2020).[2] We keep tweets with probabilities above 0.5. Previous work has achieved this filtration using keywords (Muthiah et al., 2015; Ramakrishnan et al., 2014) or focus on post-event data collection (Alsaedi et al., 2017). The number of tweets by country before and after filtering for language and relevance are shown in Figure 2. Only a small number of these countries have a large number of English tweets, and we consider many to be "low data" countries. This low amount of data can be due to low country internet usage in general, low popularity of Twitter in the region, or low prevalence of English. The language breakdown of each country is in Appendix Figure 8.

2.2 Civil Unrest Ground Truth

We use a combination of ACLED and Wikipedia to determine whether a riot or protest occurred. We are focused on events that are reflected in these sources and also appear on Twitter. We select ACLED instead of other civil unrest event databases (e.g., GDELT (Leetaru and Schrodt)) because it is freely available, manually curated by regional experts, and provides simple event categorization.

From ACLED we select events of the types Riots and Protests for our automated analysis. Data was downloaded from 2014-2019 using the data export tool[3]. The events in ACLED range from small- to large-scale, but the scope of the event is not clearly reflected in ACLED. For the case studies we utilize Wikipedia, since it contains information about large-scale events that likely appear on Twitter. The ACLED coverage information for each country is in Figure 2. Most countries have at least 500 events in the 6 year period.

3 Case Studies

We focus on two countries with differing levels of available English tweets and civil unrest levels according to ACLED: South Africa and Ethiopia. South Africa is a high-data country (i.e., has an abundance of English tweets) with high rates of civil unrest (almost 2,000 events in the 6 year period), and Ethiopia is a low data country with relatively lower rates of civil unrest (almost 600 events). These two countries are not representative of all countries, but do provide insight on issues that may arise when using Twitter data from high and low data countries for civil unrest detection and forecasting.

For each country, we select a single event and ask: (1) is the event discussed on Twitter? (2)

[1] Twitter provided language identification was not available for the early years in our dataset.

[2] The model has a test F1-score of 0.81 for identifying civil unrest-related tweets.

[3] `https://acleddata.com/data-export-tool/`

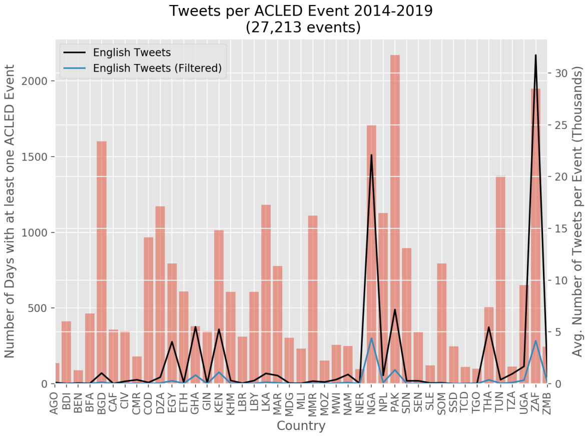


Figure 2: The average number of tweets in the dataset per ACLED event for the selected 42 countries. Counts for English tweets and filtered English tweets are shown. Events are limited to ACLED types Riots and Protests.

is there any noticeable buildup in Twitter activity prior to the event? (3) *who* is talking about the event? We limit our study to events with a clear start and end date (e.g., a specific protest) as opposed to general sociopolitical movements with a buildup over years (e.g., a Black Lives Matter (BLM) protest in the US versus studying the BLM movement). For (1) we look at the ratio of civil unrest-related tweets to all tweets (the relevance classifier discussed in §2) and then proxy the Twitter discussion through popular hashtags. The analysis for (2) is important for systems that aim to forecast civil unrest. To approximate event "buildup" we expand our scope from the day of the event to a few days before and after. This analysis looks at the emergence of event-related hashtags and whether Twitter is used in event buildup or is more reactionary. For the question of *who* is talking (3), we use user verification status and rates of likes and retweets as a heuristic for influential users. Only users active in the one week time frame are included.

After the case studies, we test how well post-event knowledge can help with data filtration. Other work uses location, time, keyword, or hashtag information gathered through post-event analysis (Littman, 2018; Alsaedi et al., 2017). For this experiment, we use the provided ACLED descriptions of the events to query event-related tweets.

3.1 South Africa: Johannesburg Riots

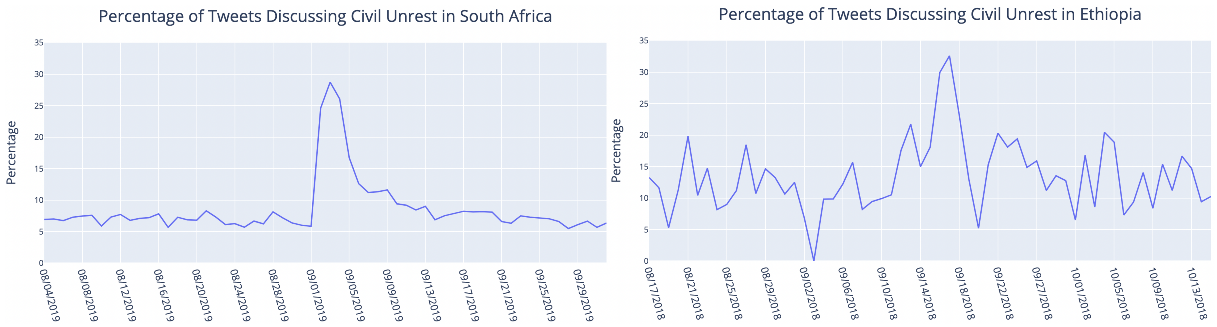
South Africa is prominent in our dataset due to the widespread use of English and the popularity of the platform in the country. We examine the 2019 Johannesburg riots, a large-scale event that took place in Johannesburg, South Africa, during the period September 1-5, 2019. The riots were the culmination of two unrelated incidents: (1) the murder of a taxi driver by an immigrant and (2) the sexual assault and murder of a college student, Uyinene Mrwetyana, on August 24, 2019.[4] The close timing of the incidents caused uproar about xenophobia and gender-based violence in the country.

We first explore if the event is discussed on Twitter. Figure 3a shows the ratio of civil unrest-related tweets (i.e., filtered tweets) during the week of the riots. The ratio is highest during the riots (September 2-5). On September 3 almost 30% of all tweets were civil unrest-related, though the civil unrest score from the filtration model does not determine if the tweets discuss the event in question. We next use hashtags to measure popular topics and find that the top hashtags correspond to the riots (Figure 4a). Hashtags are counted using only the filtered tweets, but the top hashtags from all tweets are in Appendix Figure 9a. All hashtags were lowercased to ensure accurate counts across hashtag variation. Some of the hashtags are popular enough to be in both filtered and unfiltered analysis.

Conversations about both the rise of xenophobia and gender-based violence are visible from the hashtags, some even calling to "shut down" the country in a state of emergency (e.g., `#shutdownsouthafrica`). The gender-based violence discussion centers around the victim's name and other slogans which questions the safety of women in the country (e.g., `#aminext` and `#uyinenemrwetyana`). The xenophobia discussion centers around hashtags containing "xenophobia", such as `#xenophobiansouthafrica`. The locations of large protests are also identified, with `joburgcbd` (Johannesburg central business district) and `#pretoriacbd` (Pretoria central business district). The riots started in Pretoria and then spread to Johannesburg, and this is shown in the hashtag popularity for the week. An example tweet from the event is in Figure 1.

While the rise in civil unrest tweets corresponds to the dates of the protest, we see that there is no gradual rise in the unrest. This may be due to bias in data collection or the time needed to rally around

[4] `https://en.wikipedia.org/wiki/2019_Johannesburg_riots`



(a) South Africa Johannesburg Riots, September 1-5th 2019.

(b) Ethiopia Burayu Massacre, September 14-16 2018.

Figure 3: Ratio of civil unrest-related tweets in South Africa and Ethiopia around the time of large-scale protests. In both countries, the civil unrest-related tweets spike around 30%, but that spike corresponds to thousands of tweets in South Africa and only 150 in Ethiopia.

central hashtags. The hashtags remained popular after the event, indicating people were still discussing the event and/or dealing with the aftermath.

The hashtag analysis lends insight as to what is being discussed, and we are also interested in *who* is leading the conversation. We find these "influential" users based on verification status and the number of likes their tweets received over the one week period. We found that verification status did not necessarily mean they were popular users, hence the inclusion of number of likes to measure engagement. After manual analysis of the screen names and descriptions (not included here for privacy concerns), we found that these users were entertainers, reporters, and activists. The entertainment accounts did post content about the protests, often in support. News organizations were not among the top 10 accounts, which confirms the discussion is led by the people and not organizations. This is important to know for systems that use social media to "beat the news" (Ramakrishnan et al., 2014).

3.2 Ethiopia: Burayu Massacre Protests

Compared to South Africa, Ethiopia has considerably less English Twitter (see Appendix Table 4), despite having almost double the population[5]. Despite the low prevalence of English in the country,[6] it is the language of choice for the majority of tweets (see Appendix Table 8). A large-scale civil unrest event from Ethiopia is the Burayu massacre, which took place on 14-16 September 2018 in Burayu, a town near the capital, Addis Ababa. During the three day period there were clashes between ethnic groups and targeted looting of business. The activities led to 23 deaths and many injured and displaced people.[7] On September 17, thousands of people marched in the capital in protest of the poor government handling of the looting and violence.

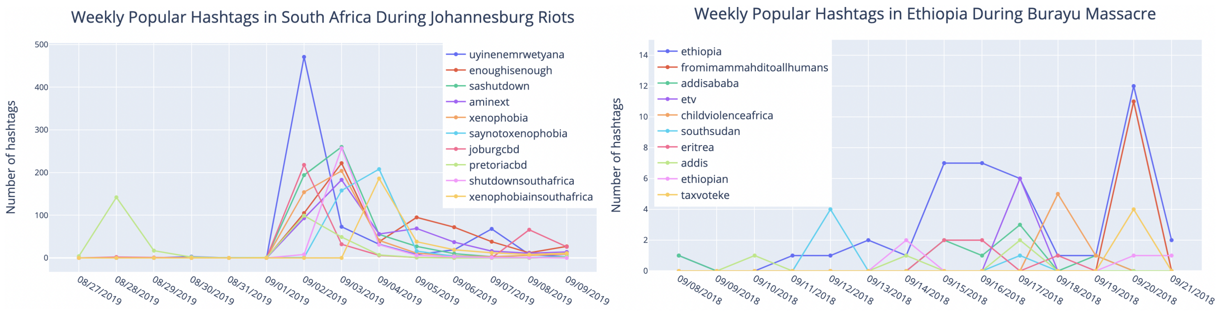
From Figure 3b, we see a clear spike in civil unrest tweets on the last day of the looting and on the day of the protest in the capital. However, the perceived volatility of civil unrest discussion in Ethiopia is due to the low data. Even at the peak where over 30% of tweets are civil unrest-related, that corresponds to only 150 tweets. The low data is also obvious when analyzing hashtags, where even the most popular only appeared 12 times (Figure 4b). Out of the top hashtags, none directly use keywords about the situation (i.e., no "massacre", "death", "violence", or "Burayu"), and event related hashtags were only determined after manual analysis of tweets. The hashtags containing relevant information were `#etv, #addisababa`, `#addis`, which reference the Ethiopian public broadcast station and the capital where protests occurred. The other top hashtags were irrelevant to the event, but still civil unrest-adjacent in nature (e.g., `#southsudan` and `#eritrea` were discussing a peace treaty between South Sudan and rebel groups hosted in Ethiopia).[8] The discussions surrounding `#etv` were interesting since they concerned the coverage of the protesters and treatment from the police. An example tweet found through

[5] https://www.worldometers.info/world-population/population-by-country/

[6] https://www.cia.gov/the-world-factbook/countries/ethiopia/#people-and-society

[7] https://en.wikipedia.org/wiki/Burayu_massacre

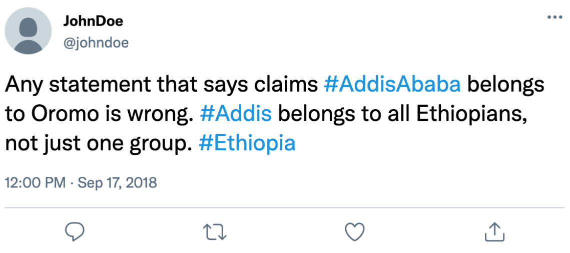
[8] https://www.reuters.com/article/us-southsudan-unrest/south-sudans-president-rebel-leader-sign-peace-deal-idUSKCN1LS2PW



(a) South Africa

(b) Ethiopia

Figure 4: Popular hashtags in South Africa and Ethiopia from the one week period of the Johannesburg riots and Burayu Massacre, respectively. Hashtag frequency is calculated from filtered tweets only.



JohnDoe
@johndoe

Any statement that says claims #AddisAbaba belongs to Oromo is wrong. #Addis belongs to all Ethiopians, not just one group. #Ethiopia

12:00 PM · Sep 17, 2018

Figure 5: Example tweet from the events following the Burayu massacre in Ethiopia. Tweet was queried based on its event-related hashtags. Paraphrased for privacy.

hashtag analysis is in Figure 5.

The influential users around the time of the protest were mostly activists. Only a few directly referenced the massacre, and those tweets mostly condemned the violence, especially against women.

3.3 Filtering Tweets using Event Descriptions

In our case studies we observed an increase in civil unrest tweets during the event time period. However, the civil unrest filtration method was general and could not limit the tweets to those discussing the specific event of interest. Other work uses post-event knowledge such as keywords or hashtags (Littman, 2018; Alsaedi et al., 2017), but we leverage the ACLED event descriptions to query related tweets. We test this method on tweets from the Johannesburg Riots.

For the query process we created embedding representations of each tweet from September 3, 2019 in South Africa (the peak of the protests) and of the aggregated event descriptions from the same day in ACLED, and then used cosine similarity between event and tweet representations to find the relevant tweets. We used the average Twitter GloVe embedding (Pennington et al., 2014) to represent the tweets and the ACLED event. Tweets and event descriptions were pre-processed with the `GloVeTweetTokenizer` from `littlebird` (DeLucia, 2020). Although the embeddings are not meant for the formal writing style in the descriptions, we decided to prioritize the tweet representation.

As seen in Table 1, the tweets with the highest cosine similarity to the event description explicitly discussed the protests, indicating success. This is a useful approach for those who want an automated method to filter a general collection of tweets. As seen from the case studies, even reducing tweets to those that are just civil unrest related was not enough to only have event-specific tweets. And gathering tweets through post-hoc analysis as in (Littman, 2018; Alsaedi et al., 2017) requires extensive knowledge of keywords and hashtags, not just the time and location of the event. This automated method can help discover more search terms and event-specific hashtags. An important note is while the most similar tweets were relevant, they only had a cosine similarity of 0.3 with the event description. An automated approach would have to tune this similarity threshold.

A scaled-up version of this experiment covering the entire month of September is in Appendix C. The matching results were subpar, indicating events could not be accurately paired with their corresponding days. We believe this is due to the sheer amount of data hiding an event signal, since South Africa has many tweets.

4 General Indicators of Civil Unrest

In §3 we determined that for two specific events in different countries (South Africa and Ethiopia), there is signal in the country's Twitter activity that a civil unrest event occurred. Those case studies

ACLED Event Descriptions
Foreign nationals demonstrated and barricaded roads in Rosettenville (City of Johannesburg, Gauteng). Some sources reported demonstrators held weapons, in case they were attacked. The demonstration was in response to nationals attacking foreign nationals.
Students, mainly female, demonstrated at the University of Cape Town, disrupting classes. The demonstration was in response to the rape and murder of student, calling for more safety for students.
Most Similar Tweets
Is the @SAPS investigation of the alleged murder of the Pretoria taxi driver by undocumented immigrants over yet?
Glad something is happening but maybe they should have waited longer for investigations... but where are the police?!

Table 1: Example ACLED event description and corresponding tweets, according to highest cosine similarity for South Africa on September 3, 2019. Only some of the 8 event descriptions for this day are shown. The tweets all have cosine similarity around 0.3 to the event representation. Tweets are paraphrased for privacy.

were very manual, which is not practical for large-scale analyses across events, time, and countries. Also, the identified relevant tweets and hashtags from the two events were very different (e.g., `#etv` vs `#shutdownsouthafrica`). Are there general trends/indicators of civil unrest across many countries and events? We answer this question by training an interpretable[9] n-gram model and analyzing its per-country performance and feature importance with SHAP, a popular explainability tool SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017).

4.1 Experimental Setup

We formulate the civil unrest event detection task as a binary classification problem to predict whether an event occurred in a particular country on a particular day. The ground truth is the ACLED labels from §2, where a day is "positive" if at least one event occurred. For easily understandable features, we use n-gram (token) counts from Scikit-learn (Pedregosa et al., 2011). All tweets are grouped by their country and day of origin and assigned a positive or negative label depending on if that country and day is present in ACLED.[10] Only filtered tweets are included (see §2), and days without corresponding tweets are removed. To make computation tractable, we only consider the top 10,000 tokens from the training set. The tweets were tokenized with `littlebird`. Each day in a country is represented as the raw count of the aggregated tweet tokens. For temporal generalizability, we split the Twitter data by year, where 2014–2016/2017/2018–2019 are the train/validation/test sets.

We use a random forest model because it is both simple enough to be interpretable and powerful enough to capture relevant features from the data. For comparison, we utilize a random baseline that predicts the positive class with the probably of the positive class in the train set. Since 29% of the country/date samples were positive from 2014–2016, the random classifier predicts the positive class 29% of the time. The models were evaluated with precision-recall metrics. To encourage model generalizability across countries we remove location-specific tokens from the data, focusing on country names and locations mentioned in ACLED. Details are in Appendix §D. In addition to combating country bias by removing location words, we also sampled the dataset to ensure equal representation of countries.

The n-gram model achieved an F1 score for the positive class of 0.45, with precision close to 0.60. This score greatly outperforms the random baseline, which only achieves an F1 of 0.29 (see Table 3). The non-debiased n-gram model performed slightly better than the n-gram model (0.50, an improvement of 0.05 F1), however this increase in performance came at the cost of less informative features. We attribute this low F1 score to noisy data and labels. While the tweets are filtered to those civil unrest-related, as discussed in §3, that does not mean the tweets are all related to the ACLED event(s) reported on the day. In addition, while ACLED is a human-curated dataset, the event granularity varies from small-scale events in a single town to large-scale events in populous

[9] "Interpretable" referring to the simplicity of the model and easily understandable features, i.e., token counts

[10] "Country of origin" is present in all tweets since only geotagged tweets are in the dataset.

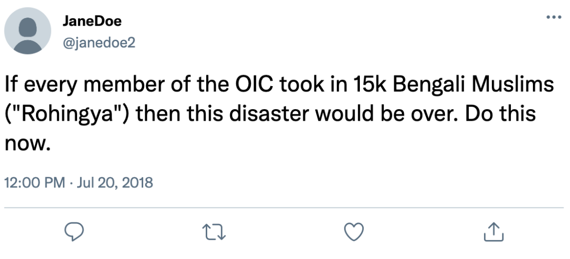
cities. Both cases provide a positive label for the day in a country, despite the probable difference in signal. Once the model was trained, we used SHAP values to find the most important features, or tokens, according to the model.

The SHAP value of a feature estimates the marginal contribution of that feature to the model output across different combinations of other features, essentially a feature "importance" (Lundberg et al., 2020). A negative SHAP value means the feature pushes the prediction towards the negative class, and a positive value pushes it towards the positive class. Although SHAP operates on a single instance, the aggregation of SHAP values can provide insights on the overall impact of a feature. The *magnitude* of a feature's impact is the sum of its SHAP values across many examples, regardless of its positive/negative impact (absolute value). We chose SHAP because it worked better with the sparse count-based features than LIME (Ribeiro et al., 2016).

4.2 Features Indicative of Civil Unrest

The magnitude of a feature's SHAP values across many examples provides insight into features the model deems important to decision making. In this case, a feature is the raw number of times a token appeared in all the tweets for a country for a day. To evaluate country generalizability we check the top features for a specific country, Myanmar, and all 42 countries in the dataset (including Myanmar). The SHAP values are from all Myanmar samples in the test set and 500 samples from all countries. All samples are pulled from the test set, years 2018–2019. The top features are in Table 2.

In general, the majority of the words deemed informative by the model for all 42 countries are related to human rights and civil unrest (e.g., "government", "casualties"). The majority of top features are not country-specific (other than "rohingya"), indicating that the country debiasing was a success. There is high overlap between the top tokens for all countries and those for Myanmar. However, in Myanmar we see features indicative of the Rohingya crisis (i.e., "rohingya", "burma", "terrorist", "military").[11] Unlike the international features, these are very specific and clearly describe the events in the country for 2018-19. This high overlap is interesting considering Myanmar comprised only 2.4% of the test set.



JaneDoe
@janedoe2

If every member of the OIC took in 15k Bengali Muslims ("Rohingya") then this disaster would be over. Do this now.

12:00 PM · Jul 20, 2018

Figure 6: Tweet representative of civil unrest in Myanmar. Queried using aggregate SHAP feature importance as token weight. Paraphrased for privacy.

To find tweets representative of civil unrest in Myanmar, we use the average SHAP feature values as weights for each token and query the tweets with this weight system. The result provided very relevant tweets, as seen in Figure 6.

We also evaluate whether the country "debiasing" improved the generalization of the model. From Table 3 we see the non-debiased model (with location stopwords) outperforms the debiased model across train, validation, and test sets. However, when looking at the distribution of scores for each country, we see the debiasing did improve the model's performance on previously low scoring countries (Appendix Figure 10). Essentially debiasing spreads the model's performance more equally across countries, instead of only performing very well on a minority. In a specific example from Ethiopia, the effects from debiasing are clear (Figure 7). Although we learned from the case study in §3 that Ethiopians rallied around location-specific hashtags for the Burayu massacre, the features learned by the model could not transfer to other countries (i.e., "addis", "ethiopia"). The features after debiasing are much more general, including "media", "inhuman", and "injured".

We aimed to find features that generalize across countries, and from the international analysis on all countries, we did find features beyond the obvious keywords "protest" and "war." It was also promising that the international features overlapped with the important Myanmar features.

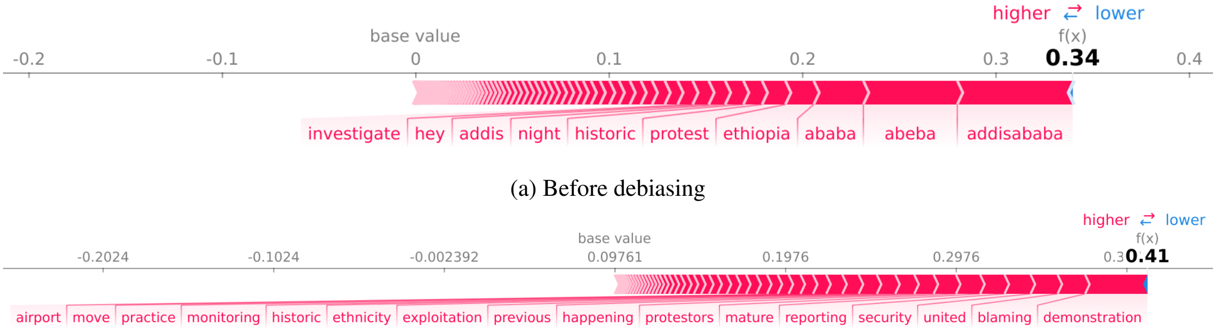
5 Ethical Considerations

Twitter is a commonly used social media source for studying and forecasting civil unrest. The main ethical concerns over this use of Twitter data is user privacy and bias in event information. For user privacy, we paraphrased tweet content to reduce

[11] https://www.unicef.org/emergencies/rohingya-crisis

All Countries	Myanmar
human, **country**, **world**, **years**, **media**, **hope**, situation, **government**, read, **rohingya**, **feel**, **rights**, casualties, true, police, watch, **time**, day, call, national	burma, **rohingya**, **world**, **media**, **country**, **human**, military, **years**, terrorist, **hope**, **feel**, **government**, group, war, day, **time**, today, plz, **rights**, official

Table 2: Top 20 features according to SHAP values from (1) 500 samples from all countries and (2) all Myanmar samples. Samples drawn from the test set (2018-19). The majority of the important tokens directly relate to human crisis and civil unrest, and the top Myanmar features describe the Rohingya crisis. Importance determined by the magnitude of SHAP values. The all countries sample is class-balanced.



(a) Before debiasing

(b) After debiasing

Figure 7: Individual SHAP output on days of the Burayu Massacre protests in Ethiopia before and after country debiasing (i.e., removing location-specific words). While the overall model performance dropped 0.05 F1, the most important features became more generalizable and informative.

	F1	Precision	Recall
N-gram	0.45	0.58	0.37
N-gram (biased)	0.50	0.61	0.42
Random	0.29	0.32	0.27

Table 3: F1 scores for the civil unrest detection task. Ground truth labels are from ACLED Riots and Protests. Only test scores are shown (2018-2019). See Appendix Table 5 for train and validation scores.

reverse identification of users (Ayers et al., 2018).

The bias in event information can stem from the language of study, user misinformation, and outside information from inaccurate geotags (tweet location). Language can proxy status in certain countries, providing different viewpoints depending on the studied language. Also, tweet content is not fact-checked and may contain misleading or incomplete information. This misinformation is not necessarily malicious, but could warp observers' perceptions of events. This is very important to note if Twitter is being used to identify event information and details, as opposed to a reputable news source. Claims made by Twitter users should be attempted to be verified before being treated as fact.

Misleading information can also come from users outside of the country. Inaccurate geotags can come from the use of a VPN, which masks a user's true location. VPNs are commonly used to circumvent government-authorized social media shutdowns, especially in countries present in our analysis.[12] A user could potentially tweet about the situation in their country, but it would be aggregated with the tweets for a different country based on the geotag information.

6 Conclusion

There are ample real-time systems and research using Twitter to forecast civil unrest, but not many studies step back and evaluate the *manifestation* of civil unrest on Twitter, especially across countries. In this work we presented a mix of qualitative in-depth case studies of events in two countries, and a quantitative large-scale automatic detection of civil unrest events across 42 countries. This evaluation of civil unrest on Twitter is not comprehensive, but demonstrates some concerns of using Twitter for civil unrest detection across countries.

From the case studies of the Johannesburg riots

[12] https://www.top10vpn.com/cost-of-internet-shutdowns/

in South Africa and the Burayu massacre protests in Ethiopia, we find there is presence of event-related discussion for both events. In South Africa there was clear development of event-specific hashtags, but the hashtags in Ethiopia were very generic, mostly using location words. Gradual rise in civil unrest-related discussion before the large event was not able to be identified in either country, indicating users see Twitter as a reactionary platform, i.e., share their opinions in real-time or after the event. This is an important finding for work that uses Twitter data to forecast civil unrest events days or weeks in advance.

While the case studies did not unearth immediately generalizable patterns indicative of civil unrest, our n-gram model trained on Twitter data from 42 countries did find generalizable token-based patterns. Despite the low F1 score (0.5 for the highest performing non-country debiased model), the top tokens were related to civil unrest and human rights crises.

For future work we would expand the case studies to other languages, possibly even comparing event manifestation between languages. Also, we would explore explainability techniques for more powerful embedding-based models. We also only focused on large-scale events and did not evaluate the model's ability to detect small-scale events.

Acknowledgments

The authors thank Arya McCarthy, Keziah Naggita, Sougol Gheissi, Rachel Wicks, Elizabeth Salesky, and the anonymous reviewers for their feedback. This work relates to Department of Navy 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.

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.

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.

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.

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.

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

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.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67. Number: 1 Publisher: Nature Publishing Group.

Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30.

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*.

M. Osborne and Mark Dredze. 2014. Facebook, Twitter and Google Plus for Breaking News: Is There a Winner? In *ICWSM*.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, San Francisco California USA. ACM.

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.

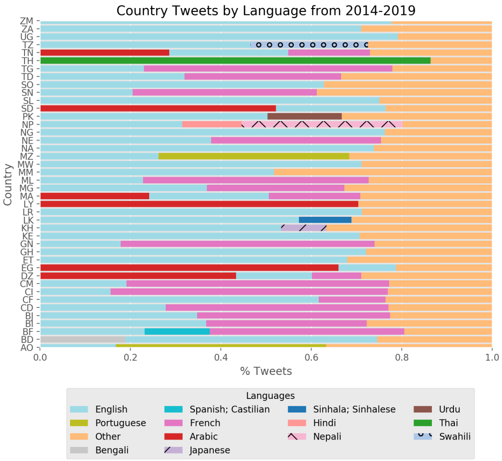


Figure 8: The language breakdown (in percentages) of Twitter data from all 42 countries. All languages that covered less than 10% of the country's tweets grouped as "other". Languages were identified with `langid` (Lui and Baldwin, 2012).

A Twitter Dataset Details

The exact number of tweets per country in the Twitter dataset are in Table 4. The country is identified by its ISO 3166 alpha-3 code. The civil unrest filtration removes many tweets, but ensures the case study qualitative analysis only looks at relevant data. We also include a language breakdown of all tweets from the initial collection in Figure 8. The choice of English covers many countries, but expansion into Arabic and French would be beneficial to future work.

B Case Study Unfiltered Analysis

§3 analyzed the Twitter content around the selected events in South Africa and Ethiopia using the restricted set of filtered tweets. As discussed in §2, the "filtered" tweets are those identified by the (Sech et al., 2020) BERTweet filtration model as being related to civil unrest (i.e., score above 0.5). As seen in Figure 9, the filtration step was necessary to reduce dataset size for the manual qualitative analysis presented in this work.

In South Africa, the hashtags from unfiltered tweets contain more general content such as references to popular TV shows (`#bbnaija`, `#uyajola99`, `#idolsa`) and tourism (`#southafrica`, `#capetown`). However, the protest-centered hashtags were so popular that they remain popular even in the unfiltered list. In Ethiopia, there is overlap between the filtered and unfiltered top hashtags, but the very informative `#etv` is not on the list.

C Comparing Aggregated Twitter Content and Event Descriptions

The tweet representations and the event descriptions were matched based on cosine similarity and the matching was evaluated with mean reciprocal rank (MRR). MRR is a score between $(0, 1]$, where 1 means the correct answer is ranked first (i.e., most similar), but the score approaches 0 the lower the rank of the correct answer.

We tried two different representations: (1) n-gram-based and (2) embedding-based (Twitter GloVe (Pennington et al., 2014)). The filtered tweets over the period of a month in South Africa were aggregated by day (September 2019), represented as a vector of either (1) raw token counts or (2) average GloVe embeddings, and then compared to the ACLED event descriptions from the same time period. Tweets were preprocessed with `littlebird` (DeLucia, 2020). Only days with events were included and if more than one event happened on a single day then the descriptions were aggregated into one. Unfortunately neither representation provided good results, as both had MRR scores around 0.15. The GloVe representation and N-gram representation has scores of 0.14 and 0.15, respectively. This indicates events could not be accurately paired with their corresponding days. We believe this is due to the sheer amount of data being aggregated, since South Africa has many tweets. In the case of the embedding representation, each day is the average tweet representation, which in turn is the average of its token representations.

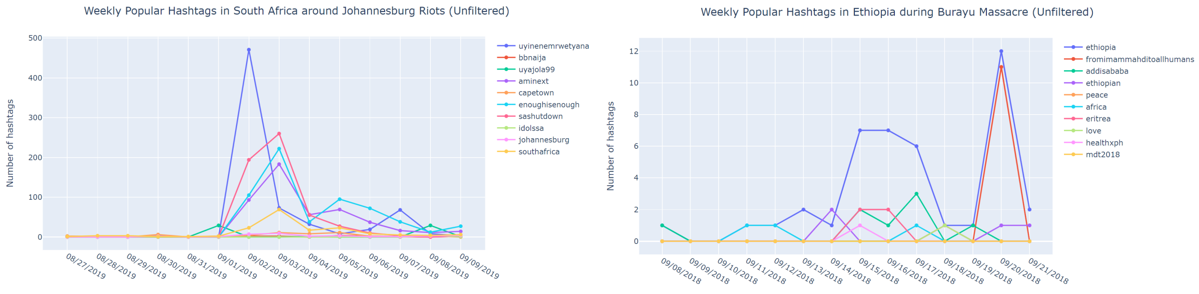
D Country Debiasing

As discovered from the initial civil unrest prediction results, the model learned country and location tokens as proxies for civil unrest. While the model performs reasonably well on the prediction task, the important features indicate it cannot generalize to new data from other countries.

We debias the model by re-training it on a "scrubbed" version of the tweets with removed location words. We create two location stopword lists. One stopword list is manually curated from the country, state/province, and city names in the ACLED data. Following the Ethiopia example from Figure 7, this list includes "ethiopia" and "ad-

AGO	289,373 / 16,639	BDI	25,041 / 6,097	BEN	150,348 / 16,489
BFA	50,197 / 4,327	BGD	2,318,969 / 330,947	CAF	28,723 / 2,860
CIV	564,197 / 34,286	CMR	833,795 / 71,420	COD	251,430 / 27,920
DZA	1,561,362 / 150,084	EGY	8,845,187 / 608,977	ETH	261,202 / 59,111
GHA	11,772,277 / 1,711,557	GIN	68,168 / 4,652	KEN	11,837,021 / 2,451,866
KHM	658,744 / 86,021	LBR	114,329 / 20,267	LBY	674,195 / 55,177
LKA	2,312,676 / 320,593	MAR	2,155,938 / 243,772	MDG	118,673 / 10,622
MLI	77,523 / 6,782	MMR	552,406 / 81,130	MOZ	349,321 / 31,634
MWI	780,767 / 101,891	NAM	1,881,238 / 238,720	NER	58,960 / 5,220
NGA	48,954,857 / 9,660,532	NPL	1,789,592 / 239,200	PAK	15,927,538 / 2,966,772
SDN	743,925 / 55,151	SEN	684,381 / 40,840	SLE	131,235 / 19,460
SOM	215,610 / 60,061	TCD	26,951 / 2,577	TGO	58,253 / 8,254
THA	14,661,060 / 980,846	TUN	944,833 / 87,903	TZA	2,216,871 / 248,856
UGA	3,274,687 / 638,372	ZAF	72,155,722 / 9,323,649	ZMB	1,706,438 / 246,815

Table 4: English tweets per country. The counts are aggregated over the 6 year period 2014-2019. The overall English tweet count and the filtered English tweet counts are shown (separated by “/"). Countries are identified by their ISO 3166 alpha-3 code.



(a) South Africa

(b) Ethiopia

Figure 9: Popular hashtags in South Africa and Ethiopia from the one week period of the Johannesburg riots and Burayu Massacre, respectively. Hashtag frequency is calculated from all English tweets, not only filtered tweets.

dis ababa"[13], but not adjectives such as “ethiopian". The second stopword list is automatically created by using the same civil unrest prediction setup, but we changed the prediction task to tweet country origin. The stopword list is comprised of the features that SHAP identifies as having > 0.005 contribution towards the prediction. Example top features from the most identifiable countries are in Table 7. We compared the two feature sets of 10,000 tokens produced by the two debiasing methods. The union of the two feature sets have a size of 10,630 while the size of intersection is 9,374, thus giving a Jaccard similarity of 0.882.

The country prediction model (same setup as the civil unrest prediction model with n-grams and random forest) achieved an F1, precision, and recall of $0.8, 0.69$ and 0.97, respectively. For comparison, we evaluated the country prediction performance after removing the ACLED location words. That model achieved scores of $0.09/0.13/0.07$ F1/precision/recall. At first glance it appears that only the ACLED location stoplist is sufficient, but the model was able to identify a handful of countries with very high accuracy. The top features for this model are in Table 6. However, due to poorer model performance with the SHAP stoplist, we used the ACLED location stoplist for the prediction experiments in §4. We believe the SHAP stoplist removed too many informative words.

[13] The alternative spacings of the tokens are also included, such as “addisababa".

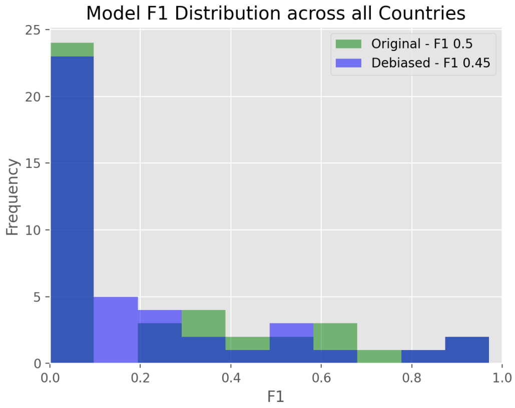


Figure 10: Distribution of F1 scores across 42 countries for the n-gram civil unrest detection task. The original “biased" and “debiased" models are compared.

	F1	Precision	Recall
N-gram	0.73	0.85	0.64
	0.46	0.56	0.39
	0.45	0.58	0.37
N-gram (with location stopwords)	0.84	0.87	0.81
	0.53	0.57	0.51
	0.50	0.61	0.42
Random	0.27	0.27	0.27
	0.27	0.28	0.27
	0.29	0.32	0.27

Table 5: Precision-recall metrics from the n-gram model trained to predict civil unrest in a country for a day. Ground truth are riot and protest labels from ACLED. Scores shown for each model are train, validation, and test, respectively (top to bottom).

Kenya	**Nigeria**	**Pakistan**
kenyan	hit	hit
hit	job	basic
basic	basic	job
job	beaten	dear
force	dear	qatar
dear	qatar	peak
broken	broken	broken
carry	beef	force
ktn	condolence	love
beaten	salaries	blood
South Africa	**Ghana**	**Uganda**
hit	basic	force
basic	force	time
job	hit	dear
branches	broken	hit
kzn	time	broken
beaten	mining	shame
condolence	shame	basic
love	twitter	water
broken	dear	twitter
piss	barking	barking

Table 6: Top 10 features with the largest magnitude of SHAP value for countries that are easy to predict even after removing ACLED location stopwords.

Kenya	**Nigeria**	**Pakistan**
ruto	abeg	pti
news	pls	kpk
reason	stop	pakistanis
hii	guys	army
guys	security	pakistan’s
court	jonathan	asif
security	court	stop
years	years	court
long	army	years
stop	wike	security
South Africa	**Ghana**	**Uganda**
soweto	ghanaians	reason
zuma	citicbs	news
saps	mahama	ugandans
guys	news	croozefmnews
anymore	reason	mps
there’s	smh	court
court	what’s	dead
years	court	life
reason	ndc	govt
hayi	govt	shd

	No debias	**Debiased**
Kenya	0.97/1.00/0.95	0.97/0.99/0.95
Nigeria	0.97/1.00/0.94	0.97/1.00/0.94
Pakistan	0.97/1.00/0.95	0.97/0.99/0.95
South Africa	0.97/1.00/0.93	0.97/1.00/0.93
Ghana	0.99/1.00/0.98	0.96/0.96/0.97
Uganda	0.98/0.99/0.96	0.88/0.84/0.93

Table 7: Top 10 features with the largest magnitude of SHAP value for countries that are easy to predict (top), as well as the country prediction scores before and after debiasing (bottom).