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

Alexandra DeLucia*, **Mark Dredze***, and **Anna L Buczak**§

*Center for Language and Speech Processing, Johns Hopkins University
§Johns Hopkins University Applied Physics Laboratory

CASE @ RANLP 2023

Citizens use public demonstrations, protests, and riots, to **express dissatisfaction** over the current political or social state in their country. Since these causes emerge from the public, **studying them requires data on public attitudes, perceptions, and actions** around previous movements.

Why study civil unrest using Twitter?

Large-scale

Millions of users tweet from all over the world every second

"From the people"

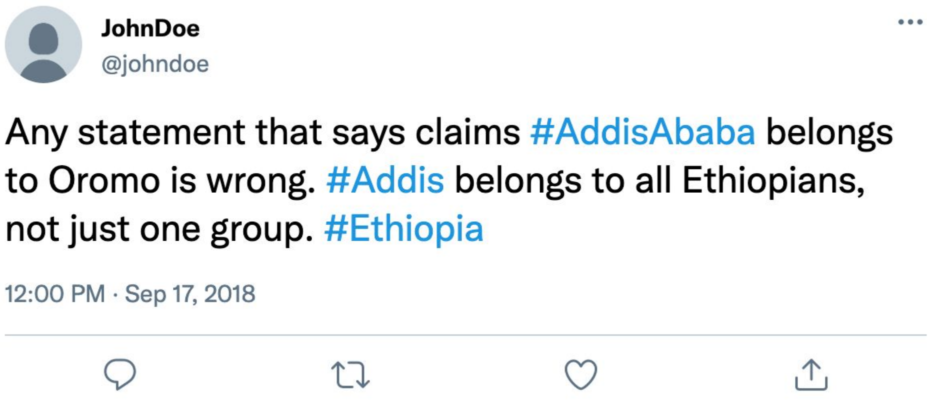
Unlike news articles, tweets are typically from individuals and express their personal beliefs and opinions

Influential in past events

Prior work has shown that citizens used social media in large events such as the Arab Spring and London Riots

Twitter is noisy

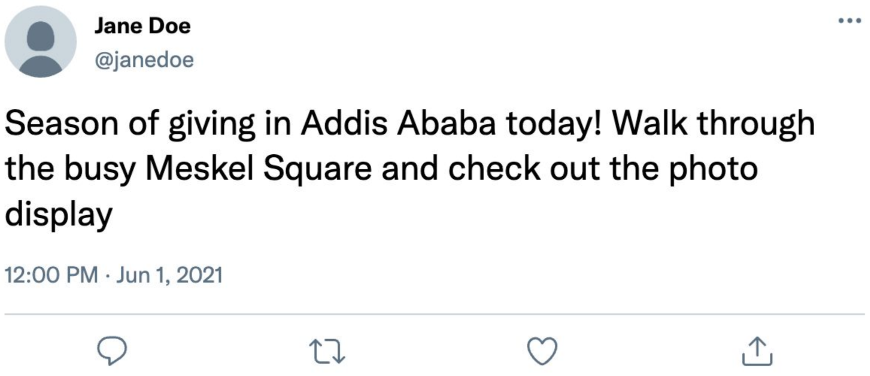
Even during a large-scale protest or riot, there are can be irrelevant tweets and conversations that occur



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



Jane Doe
@janedoe

Season of giving in Addis Ababa today! Walk through the busy Meskel Square and check out the photo display

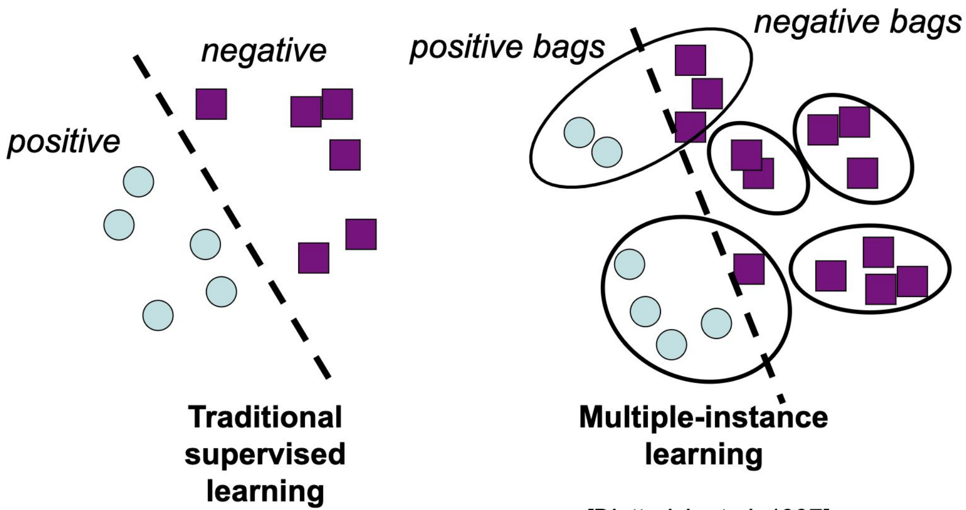
12:00 PM · Jun 1, 2021

- Prior work requires a two-step process: filtration *then* prediction
- Can we avoid the filtration step with a more holistic approach?

(Author, Year)

Solution: a multi-instance approach (MIL)

- MIL is a form of weakly supervised learning
- Training **instances** are grouped in **bags**
- A label is provided at the **bag-level**
- The depicted approach is the **standard assumption**



[Dietterich et al. 1997]

Collective assumption is better for noisy Twitter

Standard Assumption	Collective Assumption
- Instances are independent **All negative bags contain only negative instances** - Positive bags contain at least one positive instance	- More than one instance is needed to identify a positive bag **Negative bags can contain positive and negative instances** - Positive bags are identified by the distribution or **aggregation** of their instances

Research Questions

Problem:

Given tweets from a country on a specific day, can we detect that a civil unrest event occurred?

1. Does an MIL approach outperform a standard machine learning approach?
2. Can we incorporate instance-level knowledge for a better model?
3. How well does the model perform across different countries?
4. Are the most important instances for prediction useful for downstream tasks?

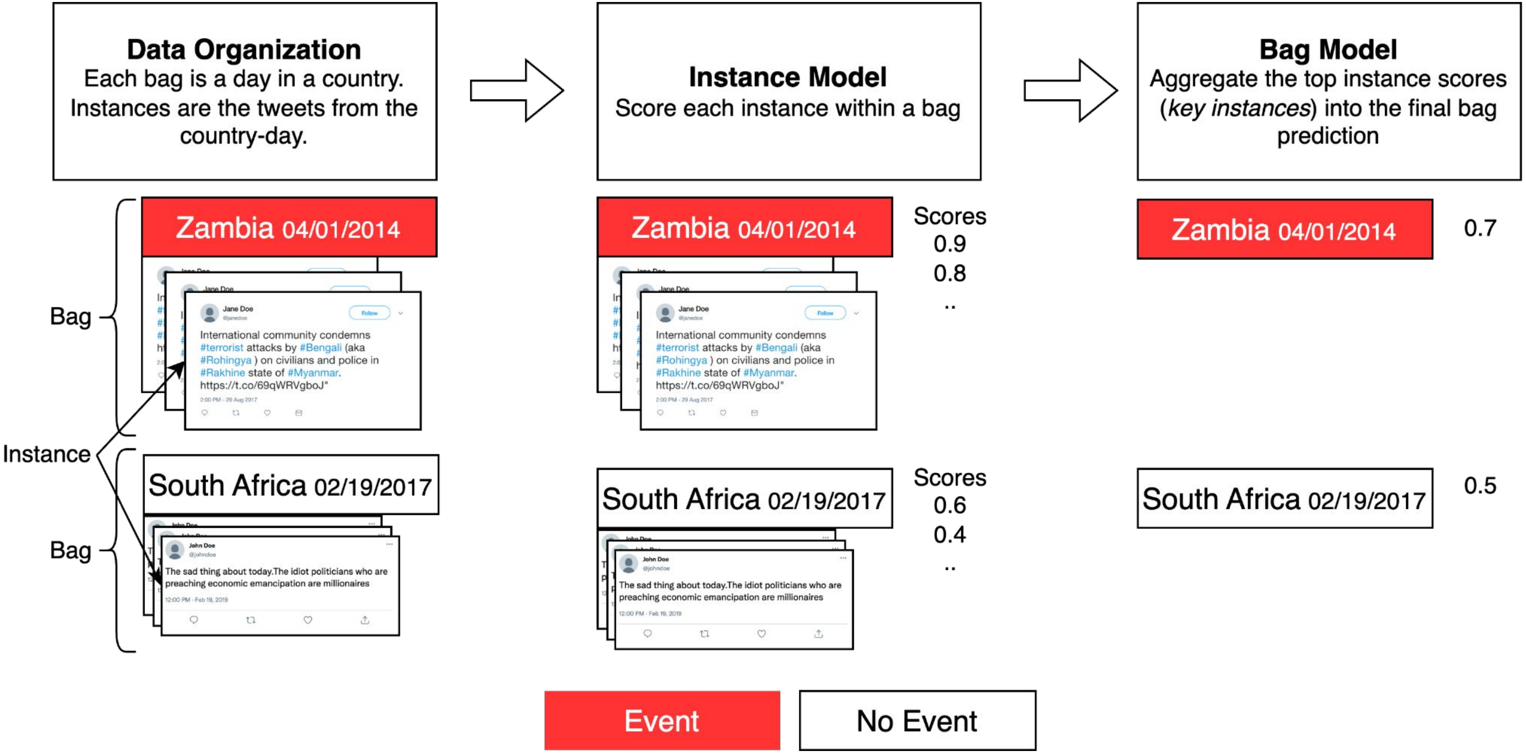
Research Questions

Problem:

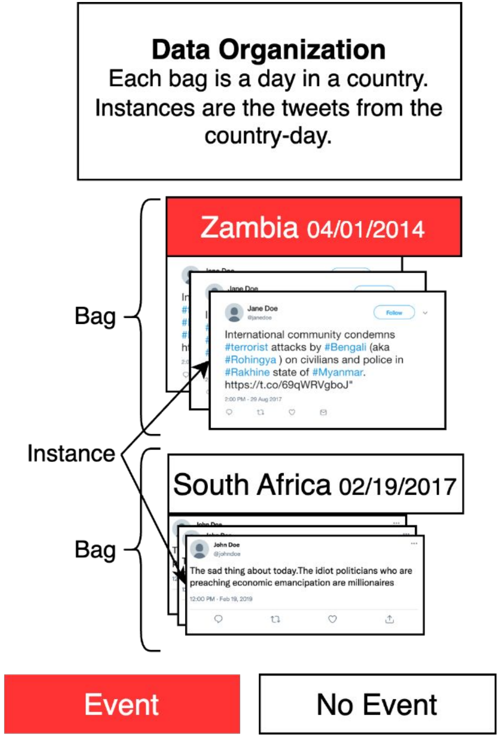
Given tweets from a country on a specific day, can we detect that a civil unrest event occurred?

1. **Does an MIL approach outperform a standard machine learning approach?**
2. Can we incorporate instance-level knowledge for a better model?
3. **How well does the model perform across different countries?**
4. **Are the most important instances for prediction useful for downstream tasks?**

MIL Approach



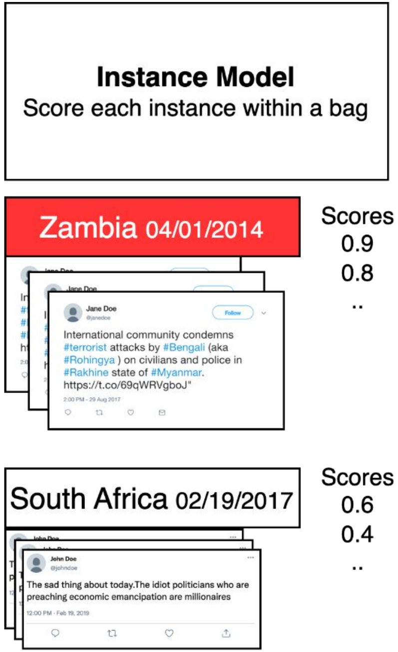
Dataset: Global Civil Unrest on Twitter (G-CUT)[1]



- 200M English tweets from 2014-2019 from Twitter streaming API
- 42 countries in Africa, the Middle East, and Southeast Asia
- Tweets are identified by their country of origin (geotagged) and the date they were created
- Ground-truth labels are from the Armed Conflict Location & Event Data Project (ACLED, "Riots and Protests" label)[2]
- An instance is a single tweet and a bag is a collection of tweets from the same country on a specific day ("country-day")

1. Study of manifestation of civil unrest on Twitter. (Chinta et al., W-NUT 2021)
2. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. (Raleigh et al., Journal of Peace Research. 2010)

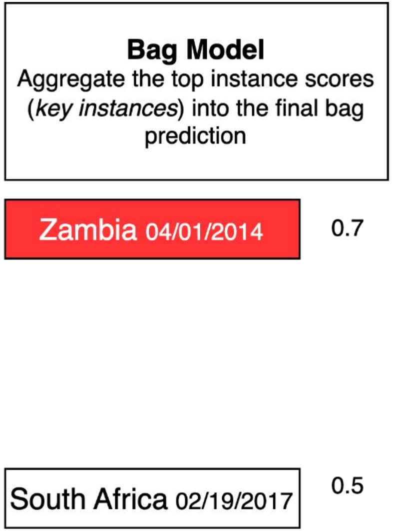
Instance Model: Predict tweet-level scores



- **Represent a single instance (tweet)** and predict a score for the tweet
- Scores are then fed into the bag model
- Fine-tune BERTweet[1] on Civil Unrest on Twitter (CUT) dataset[2]
- **CUT** is a manually annotated dataset of tweets for whether a tweet discusses civil unrest

1. BERTweet: A pre-trained language model for English Tweets (Nguyen et al., EMNLP 2020)
2. Civil Unrest on Twitter (CUT): A Dataset of Tweets to Support Research on Civil Unrest. (Sech et al., W-NUT 2020)

Bag Model: Average key instance scores



- The instance scores are provided to the bag model to calculate the final prediction for the bag
- **Bag prediction is the average of the highest-scoring instances**
- The top instances are referred to as **key instances** because they can be used to **explain the model's prediction**[1]
- Number of key instances is determined by hyperparameter **η** to determined the ratio of instances that can be considered

1. A Multiple Instance Learning Framework for Identifying Key Sentences and Detecting Events. (Wang et al., CIKM 2016)

MIL Training

- Maximum of 100 instances per bag
- Trained for 50 epochs with batch size of 20, AdamW optimizer, and 1e-5 learning rate
- The best key instance ratio (η) was found to be 0.4

Standard ML comparisons

Country-random

Same as Random but positive rate is based on the positive rate of each country. From Chinta et al. 2021.

AVG-Bag

Represent each country-day as the average instance embedding from the instance model Features used with a random forest classifier.

AVG-Bag (BERTweet)

Same as AVG-Bag but represented tweets with BERTweet instead of CUT-finetuned instance model

Ngram

Unigram counts as features with a random forest classifier. From Chinta et al. 2021.

Random

Predict positive class based on positive rate from train set. From Chinta et al. 2021.

Experiment

- All models are trained and tested on the G-CUT dataset
- Evaluated with (weighted) F1, precision, and recall
- A prediction is correct if the model predicts a civil unrest event occurred on a country-day (i.e., bag) that matches the ACLED ground truth

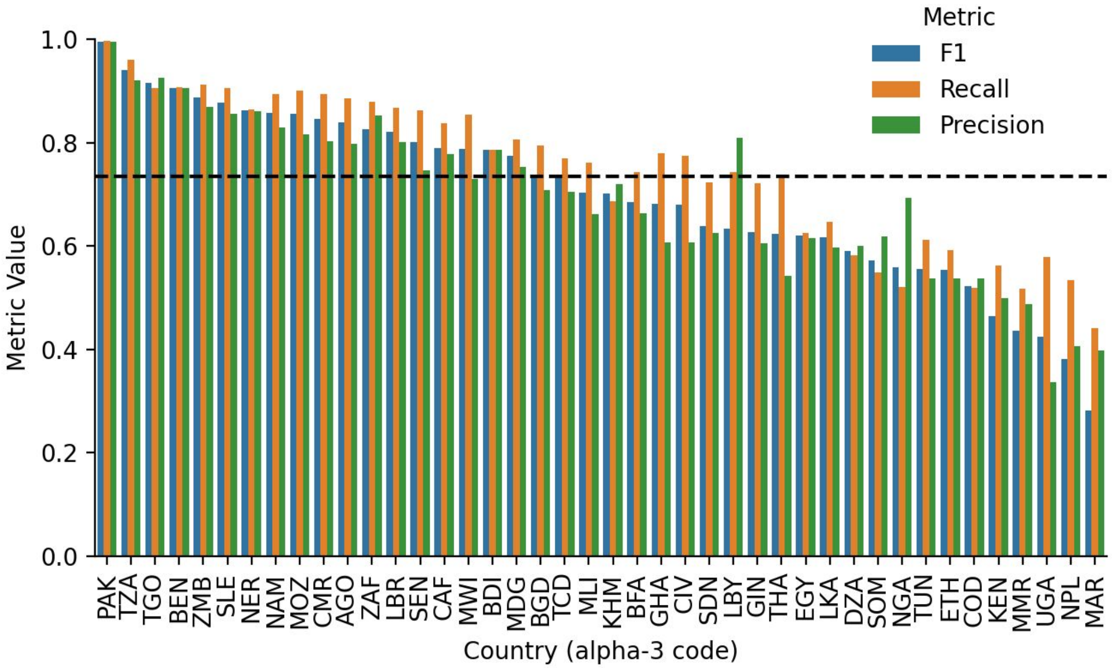
MIL model outperforms standard ML approaches

Model	F1	Precision	Recall
MIL (η=0.4)	**0.73**	**0.73**	**0.74**
AVG-Bag	0.48	0.33	0.88
AVG-Bag BERTweet	0.38	0.58	0.29
Ngram	0.48	0.64	0.38
Random	0.31	0.33	0.28
Country-random	0.50	0.54	0.46

- MIL approach outperformed other aggregation models and baselines
- Country-Random model proved to be a strong baseline
- AVG-Bag BERTweet performed worse than the AVG-Bag model, indicating importance of civil unrest pretraining from instance model

Model performance differs by country

- 50% of countries have an F1 score below the aggregated score
- Clear gap in performance between countries with the highest (Pakistan, 1.0 F1) and lowest (Morocco, 0.28 F1) scores
- Partly explained by country positive rate and presence in train set



Per-country F1 results of top MIL model on the test set

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Downstream tasks with key instances

- Example key instances identified by MIL compared to the ACLED event description
- Key instances can be used for summarization, event extraction, etc. Left for future work.

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

Model	Bag Score	Tweet Score	Tweet
MIL ($\eta = 0.4$)	0.53	0.99	@realDonaldTrump What about Saudi attacks ?
		0.99	The Joint Opposition (JO) is planning to carry out a huge mass protest called "Jana-balaya Kolabata" against the Government targeting Colombo on the 5th September 2018 from 1400 Hrs.
		0.98	Over 5,000 policemen from various units armed with all riot controlling mechanisms will remain standby to face the...

Summary

- Evaluated a multi-instance learning (MIL) approach to civil unrest detection on Twitter and compared to other standard machine learning (ML) methods
- MIL formulation worked well, achieving an F1 score of 0.73 on detecting an event occurred in a country on a specific day (as identified by ACLED/G-CUT)
- There is room for model improvement and key instance analysis in downstream tasks

Thank you for your time

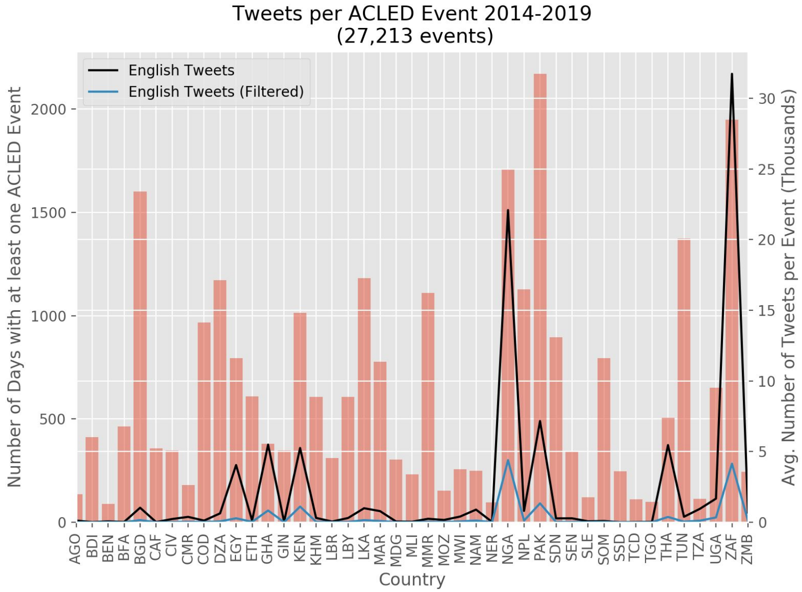
https://www.cs.jhu.edu/~aadelucia/assets/research/mil_twitter_case2023.pdf

aadelucia@jhu.edu

alexir563

Dataset: Global Civil Unrest on Twitter (G-CUT)[1]

- 200M English tweets from 2014-2019 from Twitter streaming API
- 42 countries in Africa, the Middle East, and Southeast Asia
- Tweets are identified by their country of origin (geotagged) and the date they were created
- Ground-truth labels are from the Armed Conflict Location & Event Data Project (ACLED, “Riots and Protests” label)[2]



1. Abhinav Chinta, et al. 2021. Study of manifestation of civil unrest on Twitter. In Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021), pages 396–409, Online. Association for Computational Linguistics.
2. Clionadh Raleigh, et al. 2010. Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature. Journal of Peace search. Publisher: SAGE Publications. Sage UK: London, England

Full loss objective

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

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

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

η	MIL		
	F1	Precision	Recall
0.0	0.71	0.73	0.74
0.1	0.73	0.73	0.74
0.2	0.73	0.73	0.74
0.3	0.72	0.74	0.74
0.4	**0.73**	**0.73**	**0.74**
0.5	0.73	0.73	0.74
0.6	0.72	0.73	0.74
0.7	0.72	0.74	0.74
0.8	0.72	0.73	0.74
0.9	0.72	0.73	0.74
1.0	0.72	0.72	0.73

Number of key instances does not have an effect on MIL performance

- Key instance ratio (η) had little impact on results which might be due to the high variance in the number of tweets per bag
- η=0.4 had the highest performance (**MIL (best)**)
- η>0 outperformed MIL-max (η=0) indicating an advantage in basing predictions on more than a single tweet

Bag information is needed alongside instances for accurate bag prediction

- Evaluate simply averaging the top η key instances **without training embeddings for bag prediction**
- Lack of training with bag labels leads to a performance worse than all other models
- MIL-I (max, η=0) results are skewed by high recall, thus overpredicting positive bags

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

Incorporating instance supervision hurts model performance

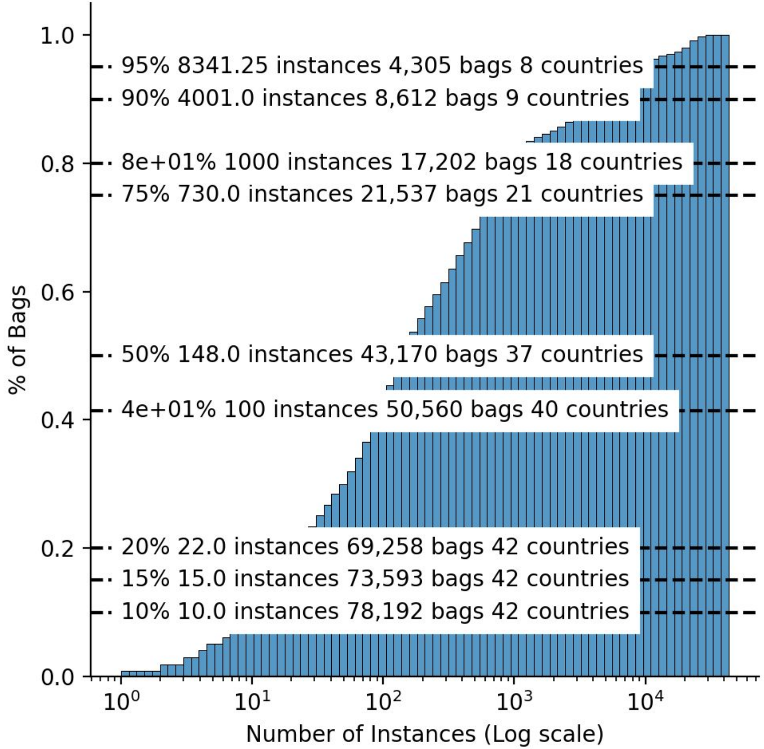
- Instance loss β is more impactful on the model than the number of key instances (η)
- As β increases, performance decreases, confirming the conflict of optimizing for both instance and bag-level classification
- No model with β>0 performs as well as MIL (best)
- MIL-BI model (β=0.25) achieves an F1 of 0.72 on the test set

β	F1	Precision	Recall
0.0	**0.73**	0.73	0.74
0.25	0.72	**0.74**	**0.74**
0.5	0.71	0.73	**0.74**
0.75	0.70	0.73	0.73
1.0	0.67	0.73	0.72

Results on test set with η=0.4. β=0.0 is equivalent to MIL best

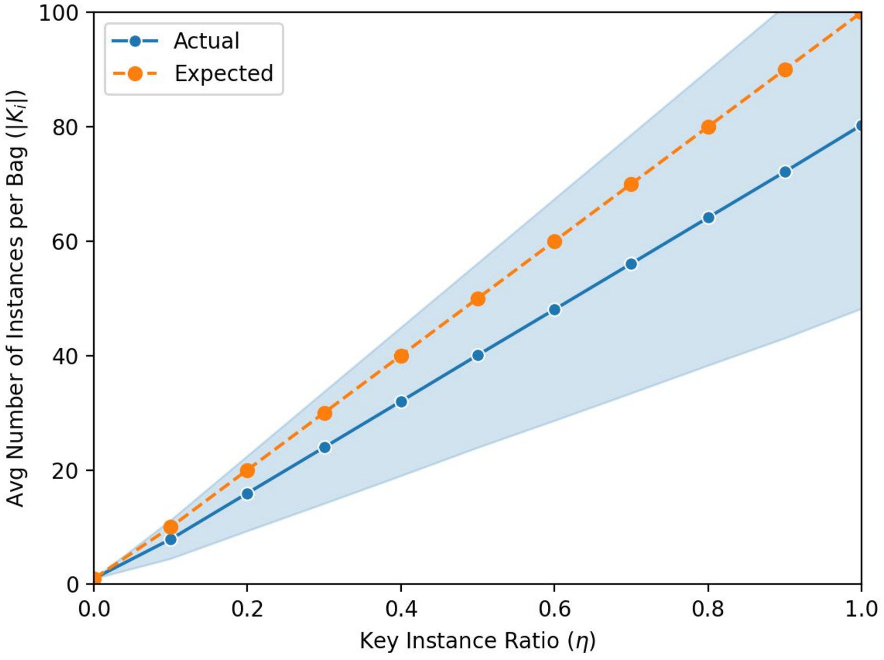
Number of instances per bag

- Decided on minimum 10 instances and dropped bottom 10% of bags
- Retained 78,192 samples (91% of the original dataset) from all 42 countries



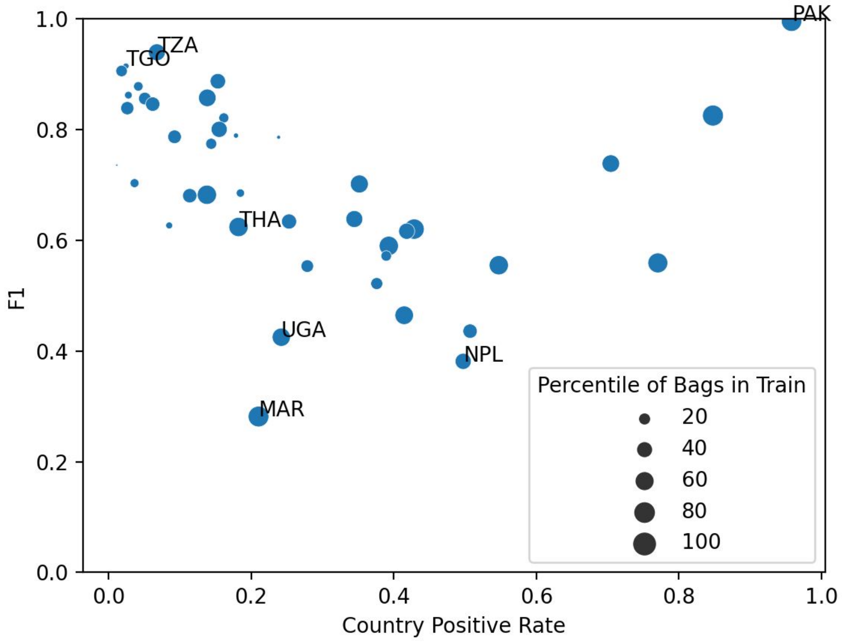
Number of instances per bag

- Bags contain less than the expected number of key instances
- Expected = η * tweets in a bag



Model performance differs by country

- Highest performing country (PAK) is highly prevalent and has high positive rate
- Countries with either very high or very low levels of civil unrest in the train set generally perform better than those in the middle (40-60% positive events)
- Morocco (MAR) is an outlier



Positive tweets in negative bags

Jane Doe
@janedoe

Somalia's militant Islamist group al-Shabab has shot dead two people it accused of being gay.

12:00 PM · Jan 11, 2017

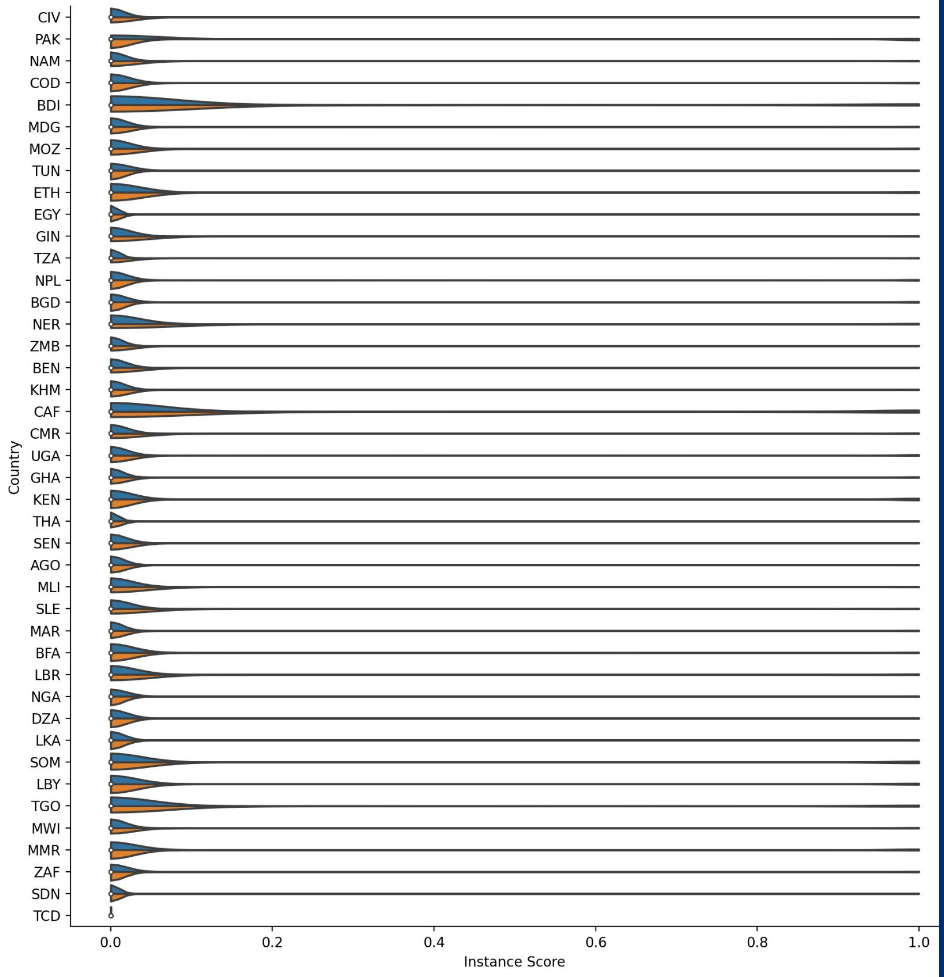
User3
@user3

Some of issues we need Govt to address:non prioritisation of National Health insurance scheme. #Ugbudget17 @USER @HealthVoice_UG

Jane Doe
@janedoe

The sad thing about today.The idiot politicians who are preaching economic emancipation are millionaires

12:00 PM · Feb 19, 2017



Key instance score distribution

- Distribution of instance scores grouped by country across days with (orange) and without an event (blue)
- Majority of tweets are not unrest-related (score<0.5)
- Little visible difference in civil-unrest related tweets on days with and without events.
- Noise is a strong indication of why civil unrest prediction on the country-day level is difficult

JOHNS HOPKINS

WHITING SCHOOL of ENGINEERING