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

Justin Sech
Center for Language and Speech Processing
Johns Hopkins University

Alexandra DeLucia
Center for Language and Speech Processing
Johns Hopkins University

Mark Dredze
Center for Language and Speech Processing
Johns Hopkins University

Anna L Buczak
Applied Physics Laboratory
Johns Hopkins University

Introduction

Motivation

We believe civilian voices are an important source of information about the state of a country. Studying Twitter dialogue helps us find information about events and opinions before official news reports[1].

Research Goals

- Collect a wide range of annotations potentially helpful in the study of civil unrest on Twitter
- Build a baseline classification model using collected annotations

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.

Twitter Data Collection

- Sample of Tweets from 2014-2019 collected from Twitter streaming API
- Filtered by geolocation to include African, Middle Eastern, and Southeast Asian countries
- English Tweets only as identified by langid[2]
- Excluded retweets
- Filtered Tweets with a set of 709 keywords including works like "unemployment", "police", and "extremist"

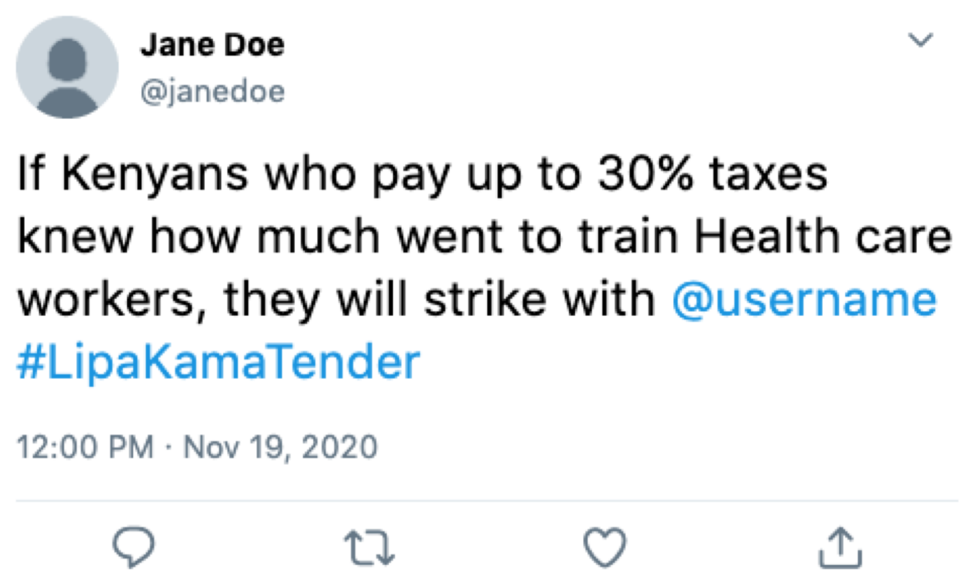
Annotation Questions

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.**

Annotated Dataset

Selected Annotated Tweets

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



`is unrest:` **nonspecific (y)**
`unrest: yes`
`timing:` **current**
`sentiment:` **neither**
`topic:` **yes**
`participation:` **neither**

Annotation Results and Inter-Annotator Agreement

- Selected the majority label for each question from three Amazon Mechanical Turk annotators, all ties were adjudicated by the authors
- Inter-annotator agreement (IAA) calculated using Fleiss' kappa

Question (IAA)	Annotations		
1 (0.430)	**yes, specific 539**	**yes, non-specific 153**	no 3,691
1(a) (0.478)	current 381	past 111	future 47
1(b) (0.325)	support 196	oppose 69	neither 425
1(d) (0.183)	yes 322	no 364	unclear 4
2 (0.168)	**yes 1,951**	no 2,446	

Civil Unrest Classification

Classification Task

- Created baseline classifiers to predict if a Tweet was related to civil unrest (predicting if Question 1 label was "yes, a specific event" or "yes, in a non-specific fashion")
- Positive class was 690 out of 4,381 Tweets (16%)

Model 1: Unigram counts with Logistic Regression

- Features were unigram counts of all tokens in a Tweet collected by CountVectorizer[3]
- Preprocessed Tweets with littlebird[4] implementation of the BERTweet tokenizer
  - Easily extendable for future comparison with a BERTweet-based model
- Regularized with L2 loss and evaluated with 5-fold cross validation

Model 2: Civil Unrest Keywords with Logistic Regression

- Features were counts of the civil unrest keywords only
- Same keywords used for filtering the initial Twitter API stream
- Identical pre-processing and evaluation steps as Model 1

Results

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

The keyword model's slight outperformance is worthwhile considering it is much smaller (700 features versus 15k) and converges faster than the unigram model (roughly 50 iterations versus 100)

Acknowledgements

The authors thank Ben Mitsunaga for testing the Twitter survey, and the Mechanical Turkers who annotated our dataset. We also thank Zach WoodDoughty, Alicia Nobles, and Arya McCarthy for their comments and advice. We thank Arya again for assisting with annotation tie-breaking. 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

1. Miles Osborne and Mark Dredze. 2014. Facebook, twitter and google plus for breaking news: Is there a winner? In ICWSM.
2. 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
3. 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.
4. Alexandra DeLucia. 2020. AADeLucia/littlebird. Original-date: 2020-04-24T00:15:07Z.