Electoral Violence Datasets

Xiao Yang Craig Macdonald Iadh Ounis

1 Introduction

Electoral violence is a common theme in developing countries all around the world where they destabilize basic standards for democratic elections. Violence against candidates, voters, journalists and election officials can reduce voters' choices and suppress the vote. Nowadays, social media platforms such as Twitter are popular as a medium for reporting and discussing current news and events, including political events. In particular, by comparing Twitter and newswire for breaking news, Petrovic et al. [6] found that Twitter leads newswire in reporting political events. Such a conclusion indicates that Twitter is useful for monitoring and studying political events, including elections. Therefore, we collect Twitter posts that are topically related to three selected elections: the 2015 Venezuela parliamentary election, 2016 Philippines general election and 2016 Ghana general election. Using human annotators and trained classifiers, we built two datasets in tweet-level and incident-level. Tweet-level dataset is consist of annotated tweets, however the incident-level dataset contains grouped tweets and the reported incident details by each group of tweets. Our datasets enable further electoral violence studies based on social media data, which can provide valuable insights on explaining and mitigating electoral violence.

2 Method

2.1 Twitter API

We use the publicly accessible Twitter Stream API to collect Twitter posts that are topically related to our target elections. By setting keywords and Twitter accounts, Twitter Stream API collects Twitter posts that contain one of the keywords or posted by one of the Twitter user accounts. Both of the keywords and Twitter user accounts are manually selected by our political science experts through browsing the Twitter posts related to the target election. We collect Twitter posts that published by Twitter users during the period of one month before and after the election dates. The example keywords used for each election are listed in Table 1. A full list of the used keywords and the followed Twitter user accounts is available separately.

Election	Keywords
Venezuela	7D,6DGanaChávez,AbajoALalzquierda,CuentaRegresiva El6DGanaChaveze,ElCambioEstaEnLaEsquina,eleccionesAN guachiman6d,laManitoNiDeVaina,MiQuerenciaEsVenezuela,SOSVzla pasoloquepase,YoDefiendoMiRevolucion,VenezuelaQuiere,victoriaPerfecta
Philippines	PHVote,Halalan2016,PiliPinas2016,VotePH2016,PiliPinasDebates2016 RoxasRobredo,IVotePH,Elections2016,PHVoteDuterte,Eleksyon2016 MIRIAM2016,Binay,Poe,philippineelection2016,Philippines vote Philippines fraud,Philippines kill,Philippines violence
Ghana	national democratic party,national democratic congress,new patriotic party convention people's party,progressive people's party,all people's congress NDP,NDC,NPP,CPP,PPP,APC,EIBElection,PulseElection,GHElection GhanaElection,GhanaDecides,ElectionHQ,Ghana2016

Table 1: Examples of keywords used with the Twitter API to collect tweets related to each election

Election	Query terms
Venezuela	eleccion,violencia,votar,pistola,armas,ametralladora,ataque electora,muerto,miedo,muerte,asesinato,disparar,fraude muere,delincuente,herido,agreden,asesinar,guachiman,protesta
Philippines	violence,attack,dead,fraud,assault,protest,intimidation,unrest gunshot,racial,die,kill,threat,vote buying,murder,corrupt terrorize,ambush,explosion,shoot,fire,harass,injure,burn selling vote,cheating,election
Ghana	poll,vote rigging,politic,fraud,assault,protest,intimidation unrest,corrupt,kill,gunshot,injured,threat,cheat,security rally,death,attack,violence,burn,clash,ballot,campaign

Table 2: Query terms used for sampling tweets from whole collections

2.2 Pooling and Annotation

In order to permit human assessors to identify relevant (election-related) tweets without having to judge millions of tweets, we adopt a TREC-style pooling methodology [7]. In particular, we allow assessors to suggest queries, in response to which an IR system ranks the tweets each day, and k top-ranked tweets are added to the *pool* of tweets to be assessed. When ranking tweets, we use the Terrier IR platform [5] and the DFReeKLIM [4] weighting model that is designed for microblog retrieval. We select only the top $k = 7$ ranked tweets per query term per day, because this gives a tweet collection with reasonable size for our human annotators. The sampled tweets are merged into one pool and judged by 5 experts who label a tweet as: "Election-related" or "Not Election-related". If a tweet is labelled as "Election-related", we then ask our human annotators to further categorise the tweets into three categories: "Electoral violence", "Electoral malpractice" or "None of them". We also provided a "Date" option to allow the annotators to state the date of the "Electoral violence" or "Electoral malpractice" incidents. For tweets without a majority agreement, an additional expert of politics was used to further clarify their categories. Query terms used for each election are listed in Table 2.

2.3 Incident Identification

In order to leverage the entire collection for each election, we train convolutional neural networks (CNN) using the sampled and annotated datasets generated in Section 2.2. The learned CNN model is then applied to the entire collection of each election to automatically detect tweets topically related to electoral violence or malpractice.

Afterwards, tweets classified as electoral violence and malpractice are clustered using K-means. By browsing tweets in each clusters, incident information including:

- incident type
- description
- date
- location
- URL link to the news report
- number of victims
- gender of victims
- number of death

is extracted by our human annotators. Each identified incident is further validated by our annotator through checking newswire reports or the actual Twitter posts reporting the incident.

Venezuela	
Type	Description
Violence	Lilian Tintori was attacked by violent group at the entry to Cojedes.
Violence	Opposition politician Luis Manuel Diaz was kill on stage.
Malpractice	Witness reported electoral irregularities in Valera.
Malpractice	Journalists fired for expressing their support for the candidate of MUD.
Philippines	
Type	Description
Violence	Campaign manager was shot dead in Sto. Tomas, Batangas.
Violence	7 killed, another wounded in the election day Cavite ambush.
Malpractice	Vote buying in Cagayan de Oro reported.
Malpractice	Police arrested 7 people for vote buying in two Bulacan towns.
Ghana	
Type	Description
Violence	NDC supporters attacked journalist S. Dogbe at parliament.
Violence	1 dead, 14 other injured in NDC and NPP clashes in Chereponi.
Malpractice	EC disqualified president aspirants and unfairly introducing new guidelines.
Malpractice	NPP calling for investigation into the illegal printing of ballot papers.

Table 3: Example of identified incidents

Election	Language	Election-related			Non-Election	Total
		Violence	Malpractice	None		
Venezuela	Spanish	294	101	1,879	3,474 (60%)	5,747
Philippines	English	193	152	1,410	2,408(58%)	4,163
Ghana	English	185	71	998	1,999(61%)	3,253

Table 4: Statistics of the tweet-level datasets

3 Dataset

In this section, we present general information and statistics of our datasets. For each election, we have a tweet-level dataset that each tweet is annotated in the way described in Section 2.2. In addition, an incident-level dataset is available for each election that describes electoral violence and malpractice incidents identified from our election collections.

3.1 Venezuela Election

We target the 2015 Venezuela parliamentary election that was held on 6 December 2015 to elect the 164 deputies and three indigenous representatives of the National Assembly.

Tweet-level Dataset The general statistics such as the dominated language and number of tweets in each category are shown in Table 4. Since the official and *de facto* language in Venezuela is Spanish, this Venezuela dataset is dominated by Spanish language. We observe that within the Twitter post annotated as "election-related", there are much more posts reporting "violence" compared to "malpractice". We also show the distribution of the annotated Twitter posts over time in Figure 1, which shows that most tweets were published by users around and ahead of the election day of 6 December. We notice that there are some peaks at particular dates, which indicates some salient events (e.g. killing of opposition politician Luis Diaz on 25 November).

Incident-level Dataset Using the trained classifier, 58,152 out of 6,652,280 Twitter posts from the whole collection are classified as "electoral violence" while 4,933 as "malpractice". Our incident-level dataset covers significant events during the Venezuela election. For example, killing of opposition politician Luis Diaz [1] in the 2015 Venezuela parliamentary election is observed. In total, we have identified 61 incidents including 47 electoral violence incidents and 14 electoral malpractice incidents. more examples of the identified electoral violence and malpractice incidents are listed in Table 3. As shown in Figure 2, most incidents, particularly the violent incidents, occurred on the election day for the Venezuela election.

Figure 1: Number of tweets per day (Venezuela)

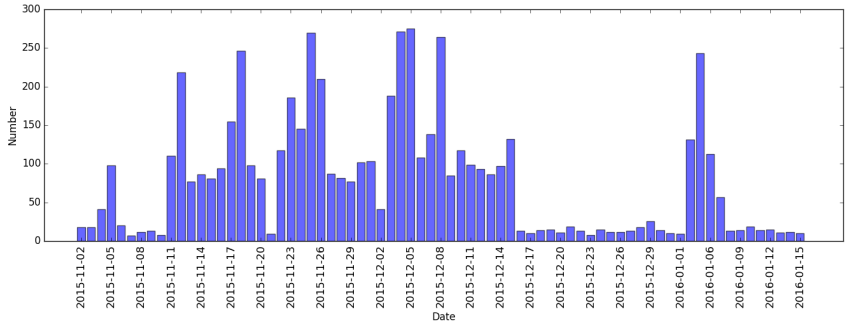
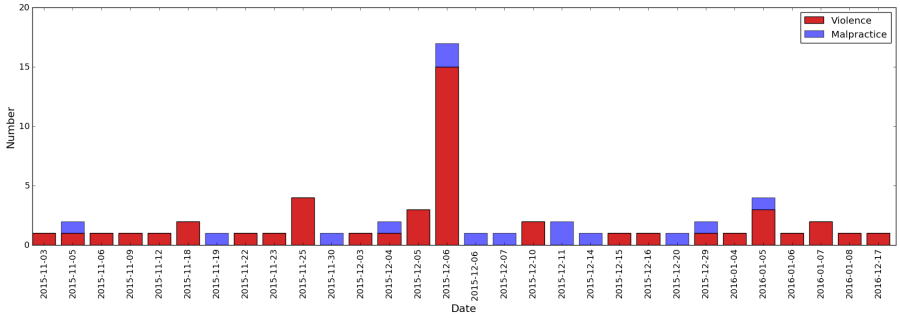


Figure 2: Number of incidents per day (Venezuela)



3.2 Philippines Election

We target the 2016 Philippines general election that was held on 9 May 2016 for executive and legislative branches for all levels of government - national, provincial, and local, except for the barangay officials. The Philippine presidential and vice presidential elections of 2016 was held as part the General election.

Tweet-level Dataset The general statistics such as the dominated language and number of tweets in each category are shown in Table 4. The dominated language is English since one of the official language in Philippines is English and most query terms that used in pooling are English. We also notice that the number of "violence" and "malpractice' tweets are very similar, which shows that electoral malpractice plays an important role in the Philippines election as electoral violence. Figure 3 shows the distribution of the annotated Twitter posts over time. Compared to the distribution of the Venezuela dataset (shown in Figure 1), Twitter posts are not concentrated on particular dates in Philippines dataset.

Incident-level Dataset Using the trained classifier, 8,396 out of 1,880,844 Twitter posts from the whole collection are classified as "electoral violence" while 7,966 as "malpractice". In total, we have identified 77 incidents, including 51 electoral violence incidents and 26 electoral malpractice incidents. Our incident dataset covers salient violent events during the Philippines election, for

Figure 3: Number of tweets per day (Philippines)

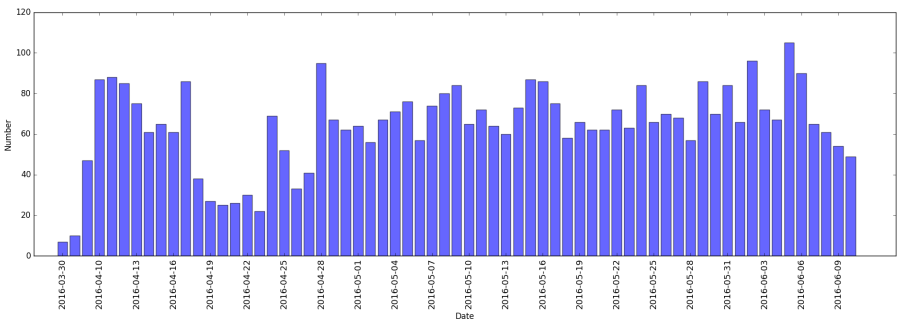


Figure 4: Number of incidents per day (Philippines)

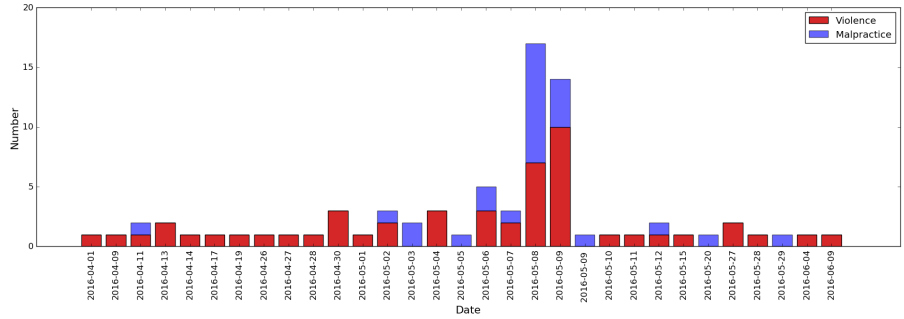
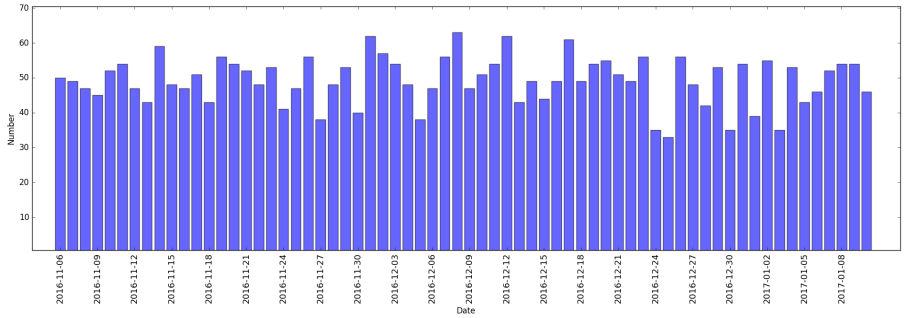


Figure 5: Number of tweets per day (Ghana)



example the Cavite ambush that leads to 7 people shot dead on the election day [2]. More examples of electoral violence and malpractice incidents are listed in Table 3. As shown in Figure 4, most incidents, both the violent and malpractice incidents, occurred on the election day and the day before in the Philippines election. This confirms our observation on the tweet-level dataset that malpractice plays an important role as the electoral violence.

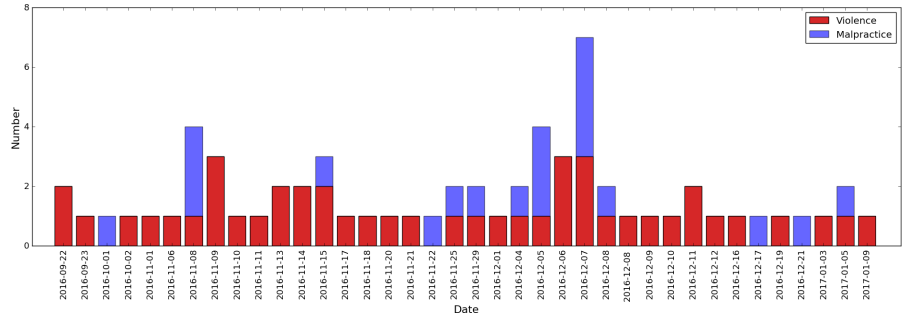
3.3 Ghana Election

We target the 2016 Ghana General election that was held on 7 December 2016 to elect a President and Members of Parliament. The election had originally been scheduled for 7 November 2016, but the date was later rejected by Parliament.

Tweet-level Dataset The general statistics such as the dominated language and number of tweets in each category are shown in Table 4. The dominated language is English since the official language is English and the used query terms in pooling are mostly English. We observe the similar pattern to the Venezuela dataset that there there are more "violence" than "malpractice" in the election. From the distribution of the annotated Twitter post over time (shown in Figure 5), it is clear that there are no particular peaks occurred during the sampled dates. This is similar to the Philippines dataset but different to the Venezuela dataset, which may indicate consistent discussion about election events.

Incident-level Dataset Using the trained classifier, 24,844 out of 3,297,160 Twitter posts from the whole collection are classified as "electoral violence" while 6,917 as "malpractice". In total, we have identified 65 incidents, including 45 electoral violence incidents and 20 electoral malpractice incidents. Salient violent events during the Ghana election are observed in our dataset, for example the clashes between the supporters of two parties NDC and NPP in Chereponi [3]. More examples of electoral violence and malpractice incidents are listed in Table 3. As shown in Figure 6, the violent and malpractice incidents are likely to occurred a few days around the election day in the Ghana election. This observation is similar to the Philippines incident-level dataset that malpractice plays an important role as the electoral violence.

Figure 6: Number of incidents per day (Ghana)



Acknowledgements

We acknowledge the efforts of our colleagues who dedicated time to making reviewing tweets: Sarah Birch, Paul Cockshott, David Muchlinski, Fatma Elsafoury and Inaki Sagarzazu.

References

[1] Venezuela opposition politician luis manuel diaz killed. http://www.bbc.co.uk/news/world-latin-america-34929332, November 2015. [Accessed: 2016-05-15].

[2] 7 killed, another wounded in election day cavite ambush. http://www.rappler.com/nation/politics/elections/2016/132410-cavite-ambush-2016-philippine-elections, May 2016. [Accessed: 2017-04-15].

[3] One killed as ndc, npp supporters clash at chereponi. http://pulse.com.gh/politics/election-2016-one-killed-as-ndc-npp-supporters-clash-at-chereponi-id5858497.html, December 2016. [Accessed: 2017-04-15].

[4] Gianni Amati, Giuseppe Amodeo, Marco Bianchi, Giuseppe Marcone, Fondazione Ugo Bordoni, Carlo Gaibisso, Giorgio Gambosi, Alessandro Celi, Cesidio Di Nicola, and Michele Flammini. FUB, IASI-CNR, UNIVAQ at TREC 2011 microblog track. In *Proc. of TREC*, 2011.

[5] Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. From puppy to maturity: Experiences in developing terrier. In *Proc. of OSIR workshop at SIGIR*, volume 60, 2012.

[6] Sasa Petrovic, Miles Osborne, Richard McCreadie, Craig Macdonald, and Iadh Ounis. Can twitter replace newswire for breaking news? In *Proc. of ICWSM*, 2013.

[7] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.