Political Science Research and Methods (2020), page 1 of 18
doi:10.1017/psrm.2020.32

ORIGINAL ARTICLE

We need to go deeper: measuring electoral violence using convolutional neural networks and social media

David Muchlinski[1]*, Xiao Yang[2], Sarah Birch[3], Craig Macdonald[4] and Iadh Ounis[4]

[1]International Affairs, Georgia Institute of Technology, Atlanta, GA, USA, [2]Literature Service Group, European Bioinformatics Institute, Cambridge, UK, [3]Political Economy, King's College London, London, UK and [4]School of Computing Science, University of Glasgow, Glasgow, UK
*Corresponding author. Email: david.muchlinski@inta.gatech.edu

(Received 14 June 2018; revised 17 July 2019; accepted 30 August 2019)

Abstract
Electoral violence is conceived of as violence that occurs contemporaneously with elections, and as violence that would not have occurred in the absence of an election. While measuring the temporal aspect of this phenomenon is straightforward, measuring whether occurrences of violence are truly related to elections is more difficult. Using machine learning, we measure electoral violence across three elections using disaggregated reporting in social media. We demonstrate that our methodology is more than 30 percent more accurate in measuring electoral violence than previously utilized models. Additionally, we show that our measures of electoral violence conform to theoretical expectations of this conflict more so than those that exist in event datasets commonly utilized to measure electoral violence including ACLED, ICEWS, and SCAD. Finally, we demonstrate the validity of our data by developing a qualitative coding ontology.

Keywords: Text and content analysis

Elections are the most common means by which citizens select and provide legitimacy to their political leaders. Unfortunately, electoral politics has become intertwined with violence across much of the world (Dunning, 2011). Research into the causes of electoral violence has recently become more systematic, examining the conditions under which incumbents are likely to use violence to influence the electoral process (Hafner-Burton *et al.*, 2014), the effects of electoral institutions on electoral violence (Fjelde and Höglund, 2016), and the conditions under which ethnic diversity contributes to such conflict (Butcher and Goldsmith, 2017).

Despite the increased interest in electoral violence, the concept remains theoretically underdeveloped and conceptually vague (Staniland, 2014). Inherent in most definitions of electoral violence is the *temporal* link between violence and elections and the *causal* link between the two. Electoral violence is conventionally understood as violence that takes place contemporaneously with the electoral cycle. The causal link, which is often more implicit, limits electoral violence to that which is in some way connected to the electoral process, as opposed to violence that takes place during the electoral process but has no direct bearing on the election. We follow Birch and Muchlinski (2020, 3) who define electoral violence as, "coercive force, directed toward electoral actors and/or objects, that occurs in the context of electoral competition."

Electoral violence is often conceptualized at quite high levels of aggregation utilizing blunt categories including whether there were post-election protests, whether "civilians were killed in significant numbers," and whether government forces harassed opposition candidates (Hyde and

Marinov, 2012). Most studies of electoral violence tacitly assume that violence which occurs contemporaneously with an election is also related to the election, but there are legitimate reasons to be skeptical. Commonly utilized datasets on this phenomenon are hand coded and conceptual ambiguity can easily creep into published measures as coders impose their own subjective biases into the data-generating process (Brass, 1997). While no recorded measure of electoral violence is free from error, the fact that many measures of this concept rely heavily on the timing of violence to justify its coding leaves substantial uncertainty regarding whether such violence would still have occurred in the absence of any election.

Other studies (Daxecker, 2012, 2014) have used disaggregated event datasets such as ACLED (Raleigh *et al.*, 2010), ICEWS (Boschee *et al.*, 2015), and SCAD (Salehyan *et al.*, 2012) to develop measures of electoral violence. Because they utilize event-based datasets, these studies may be able to more accurately assess the relationship between violence and elections by including only, for instance, violence between opposition and incumbent parties. These datasets, however, which rely on major international news media reports tend to under estimate the true number of violent events, introducing another possible source of bias into measures of electoral violence (Hendrix and Salehyan, 2015; Weidmann, 2015).

We propose utilizing an alternative source of data to develop conceptually clear measurements of electoral violence. Social media platforms such as Twitter catalog reports on political violence, and these data have previously been used to predict political instability (Ramakrishnan *et al.*, 2014). Compared to traditional news reports, Twitter reports on major news events equally well, but contains a longer tail of minor events often not covered by traditional print media sources (Jackoway *et al.*, 2011; Petrovic *et al.*, 2013). This study joins a growing field of research using social media to document conflict dynamics (Doyle *et al.*, 2014; Steinert-Threlkeld *et al.*, 2015). Most research using social media to estimate political violence has focused on large-scale, high-intensity violence like civil unrest and violent protest, but no attempt has been made to estimate occurrences of electoral violence using social media.

Our contributions are twofold. Methodologically, we introduce a way to more accurately estimate the link between elections and violence utilizing a convolutional neural network to estimate a form of political violence directly from unstructured text. While we chose to estimate electoral violence, our method is general and can be applied to estimate any concept. Substantively, we demonstrate that the combination of social media data and our machine learning platform develops more accurate estimates of electoral violence than those that currently exist. This is due to the superior classification ability of our convolutional neural network. Scholars who adopt our methodology to measure electoral violence will thus be able to draw more statistically valid correlations between such violence and variables theorized to bring about its occurrence. This is important for advancing not only scholarly knowledge about this destructive form of conflict, but can also assist policy makers to forecast this violence and develop policies to ameliorate its effects.

We want to make clear at the outset that we are aware of the possible problems inherent in utilizing social media as a source of data to measure instances of possible violence around elections. We are well aware that not all events reported by social media may have actually occurred. Therefore, we corroborate our estimates of electoral violence using local print media sources. We also examine the external and concept validity of our estimates by measuring the temporal trends of violence reported across elections against other established datasets. To be sure our data are truly related to the elections under study, we qualitatively code events discovered by the neural network and existing datasets, and compare these results. Finally, we provide in the supplementary materials evidence documenting each event discovered by our methodology. We are also aware that our data collection and analysis pipeline depends heavily on individual access to the Internet and social media. This access is geographically uneven, and is often subject to government censorship. While the methodology proposed here may not be applicable to all elections, when it can be utilized it is able to estimate electoral violence with a level of detail which is unmatched.

This paper is structured in the following way. The next section argues existing sources of data measuring electoral violence do a good job measuring the temporal link between elections and violence, but the causal link between the two remains ambiguous. We argue that social media offers a useful alternative source of data to establish this relationship. The next section introduces our natural language processing model as well as our convolutional neural network. The results section discusses how our method of detecting events in text enhances estimation of electoral violence compared to other previously utilized text analysis and machine learning methods. We also qualitatively demonstrate that our machine learning pipeline is vastly more accurate than existing datasets in assessing the relationship between electoral violence and the electoral process. We conclude with some remarks about the use of social media to estimate political violence and the use of neural networks for the collection of these data.

1. Estimating electoral violence from social media

Using textual sources of data to develop estimates of political violence using automated methods is not a new endeavor. Datasets like ICEWS are created by fully automated systems built to search for and record specific events in newswire reports (Boschee *et al.*, 2015). Nor is utilizing social media data a foreign concept to scholars of political violence. Zeitzoff (2011) collected social media data from Twitter to analyze temporal violent dynamics between Israel and Hamas during the 2008–2009 Gaza conflict, and Ramakrishnan *et al.* (2014) used social media data to forecast civil unrest across multiple countries.

Thanks to these automated methods and massive sources of textual data, scholars of political violence now have access to massive datasets measuring political cooperation and conflict. It is hard to overstate the impact this new form of data has had on the field. It is the size of these new datasets, with millions of observations coded from international media outlets and spanning decades, that has allowed scholars to understand the minute details of political violence that previous data were unable to distinguish. As the collection and use of this text-as-data has proceeded, however, its limitations have become clearer.

Datasets constructed by automated methods may be systematically under counting the true number of violent events (Weidmann, 2016; Cook *et al.*, 2017). Media organizations cannot be everywhere at once. Much political violence occurs where these organizations lack established bureaus to report these events (Earl *et al.*, 2004). Perpetrators of political violence also go to some lengths to obfuscate their use of violence to make sure they do not leave a record of their activity (Zeitzoff, 2011). This is especially likely to affect estimates of electoral violence as such violence does not often rise to a level which will draw international media attention.

Other datasets utilize reports by international organizations to develop broad measures of electoral violence (Hyde and Marinov, 2012). These datasets have also done much to improve our knowledge, but the broad categories with which they measure electoral violence often obscure the identity of the perpetrators and victims and the tactics employed, misrepresent the nature of the event itself, or otherwise provide measures of this violence at quite high levels of generality and aggregation (Staniland, 2014). It can be difficult to determine whether a violent event was related to an election because reports used to generate these data generally do not report on each violent event that occurred, but rather describe elections as "generally peaceful," or "not peaceful." As a result, most datasets that measure electoral violence, though they posit a causal relationship between violence around the election and the electoral contest itself assume this relationship rather than making it explicit.

This is problematic. While electoral violence is indeed a broad category of violence perpetrated by many different actors with a diversity of motivations (Staniland, 2014), it is unknown to what extent current data on these events are actually electoral in nature. Under reporting of this violence is also another unanswered question. While it is possible to utilize methods to uncover more empirically accurate distributions of political violence from text (Cook *et al.*, 2017), these methods do not provide for us any information about these other events, including whether they were related to the election.

We propose a solution to these problems by utilizing a different source of data entirely: social media. Social media networks facilitate collective action for political activity (Larson *et al.*, 2016). Given the ability of social media to facilitate collective action, the digital footprints left by individuals involved in these activities provide researchers with relevant data that can be used to discover the relationship of an event to the election (Schrodt *et al.*, 2013). Social media can also assist in fleshing out the obscure details of electoral violence where power asymmetries force combatants to utilize nontraditional means of violence which may go unreported by traditional news organizations (Zeitzoff, 2011).

Given the massive amount of content contained in textual data, automated document classification has become a popular method of coding information due to its inherent efficiency and flexibility (Grimmer and Stewart, 2013). Automated methods code massive amounts of information regarding political violence, including outbreaks of civil and international conflict (D'Orazio *et al.*, 2014), and have identified perpetrators of mass atrocities (Bagozzi and Koren, 2017). These algorithms, including neural networks, have achieved accuracy beyond that of previously utilized textual analysis methods, such as parsers (Beieler, 2016; Lin *et al.*, 2016).

We hypothesize that the accurate estimation of electoral violence will be enhanced by utilizing social media and neural networks for two reasons. First, because most event datasets were not created to measure electoral violence, we expect these datasets to under estimate this violence. Second, convolutional neural networks have produced state-of-the-art results in many computational linguistics tasks, out-performing other commonly utilized machine learning methods (Goldberg, 2016). We argue the combination of disaggregated reporting using social media and advances in computational linguistics will allow scholars to more accurately estimate the occurrence of electoral violence. With more accurate discovery of these events, better statistical models can be constructed to inform scholars of the mechanisms underlying such violence and its impacts on society.

2. Data collection and preprocessing

We use the publicly available Twitter Streaming API to collect Twitter posts related to electoral violence. We collected these tweets from a two-month period around elections in three countries: Venezuela in 2015, Ghana in 2016, and the Philippines in 2016. We chose these countries because they have some of the largest levels of social media penetration in their respective regions.[1] While tweets collected from the Philippines and Ghana were almost exclusively written in English, tweets from Venezuela were in Spanish.[2] We chose the two-month window in order to analyze trends in both pre and post-electoral violence, a choice commonly made in the literature (Hyde and Marinov, 2012; Hafner-Burton *et al.*, 2014).

We utilized a keyword search related to the election and electoral violence. A table with the keywords utilized in our search is given in the supplementary materials. Because the size of the tweet-based datasets resulting from this keyword search are very large, we used a computerized platform to select a random sample of tweets from each country and manually code them.

[1]For instance, Venezuela's social media penetration (the percentage of Internet users who use social media) is 68 percent, the Philippine's social media penetration is 37 percent, and Ghana's social media penetration is 40 percent; https://www.statista.com/statistics/754520/venezuela-penetration-social-networks/, https://cliqafrica.com/wp-content/uploads/2017/01/2016-Final-Ghana-Social-Media-Rankings-Report-CliQAfrica-Ltd.pdf, https://www.statista.com/statistics/490378/mobile-messaging-user-reach-philippines/, accessed 14 May 2018).

[2]For the purposes of coding the training data, Spanish tweets were automatically translated into English. Quality of the automatic translations were checked by two Spanish speakers, one author who is fluent in Spanish, and another native speaker.

Each author coded the same random sample of tweets, and a report of inter-coder reliability is provided in the supplementary materials. In total, our Venezuela election training data consist of a random sample of 5747 Spanish tweets. The Philippine training data consisted of a random sample of 4163 English tweets. The training data for the election in Ghana consisted of 3235 English tweets. A table with the statistics of these samples is provided in the supplementary materials. Tweets were hand-coded according to a two-tier classification scheme. First, a tweet was coded as election related or not election related. Then, out of those tweets that were coded as related to the election, a tweet was further coded as referencing violence or not. Thus, all tweets that were coded as violent were coded as violent with respect to the election. This two tired ontology ensured that all tweets labeled as violent referenced electoral violence rather than other forms of violence that were not related to the election. These hand coded data were used to train the convolutional neural network. To collect tweets, we adopt an informational retrieval and pooling methodology (Voorhees and Harman, 2005) as shown in Figure 1.

The keyword search collected a large number of tweets each day. In order to manually code them, we used a search and pooling methodology to identify a reduced set that were mostly likely to be concerned with electoral violence for each day. In particular, we used the Terrier information retrieval platform (Ounis *et al.*, 2006; Macdonald *et al.*, 2012) to rank tweets that well match a set of electoral violence related search terms.[3] In particular, we configured Terrier to rank tweets using the DFReeKLIM weighting model (Amati *et al.*, 2011), which is specifically designed for the analysis of text-sparse Twitter data. In this way we constructed three training datasets collected from each of the three countries. We did this for each election so that there is one training dataset of tweets for the Venezuela election, one for the Ghanaian election, and one for the Philippine election. While our natural language processing model allows for multilingual data sources, we assume that there may be systematic differences in the ways in which people tweeted about elections in each country, therefore the neural network was trained separately for each election.

2.1 Data preprocessing using word embeddings

Once the training datasets were collected, they were preprocessed to remove stopwords and capitalization and stemmed using the English and Spanish Snowball stemmer. Then the hand-labeled tweets used to train the neural network were transformed into real-valued vectors to produce word embeddings (Collobert *et al.*, 2011; Mikolov *et al.*, 2013). The software used to create these embeddings is called *word2vec* and is freely available.[4] Because word embeddings have not widely been utilized as a natural language processing tool in political science, a quick explanation is in order.

The commonly utilized method to transform words into numeric vectors is to assign each word a one-hot vector in $\aleph^{|V|}$, where $|V|$ is the vocabulary size of the text. Repeating this process for all words across n documents results in the creation of a $V \times N$ document-term matrix where words that appear in a given document are given a value of 1, and 0 otherwise. Representing words in this way leads to substantial data sparsity, increasing the data required in order to train statistical models (Mandelbaum and Shalev, 2016). The one-hot encoding of words also discards much linguistic information regarding the surrounding syntactic and semantic context of a given word in a sentence. Methods that can retain this kind of information are able to use this information to increase classification accuracy (Bengio *et al.*, 2003; Collobert *et al.*, 2011).

One such natural language processing method is word embeddings. Word embeddings are a set of language modeling and feature leaning techniques where words and phrases from the vocabulary of the textual data are mapped to vectors of real numbers (Collobert *et al.*, 2011; Mikolov *et al.*, 2013). The basic idea behind word embeddings is to create a more meaningful

[3]Terrier is available from http://terrier.org. We used version 4.1, but any more recent version would also be suitable.
[4]The website hosting this software is https://deeplearning4j.org/.

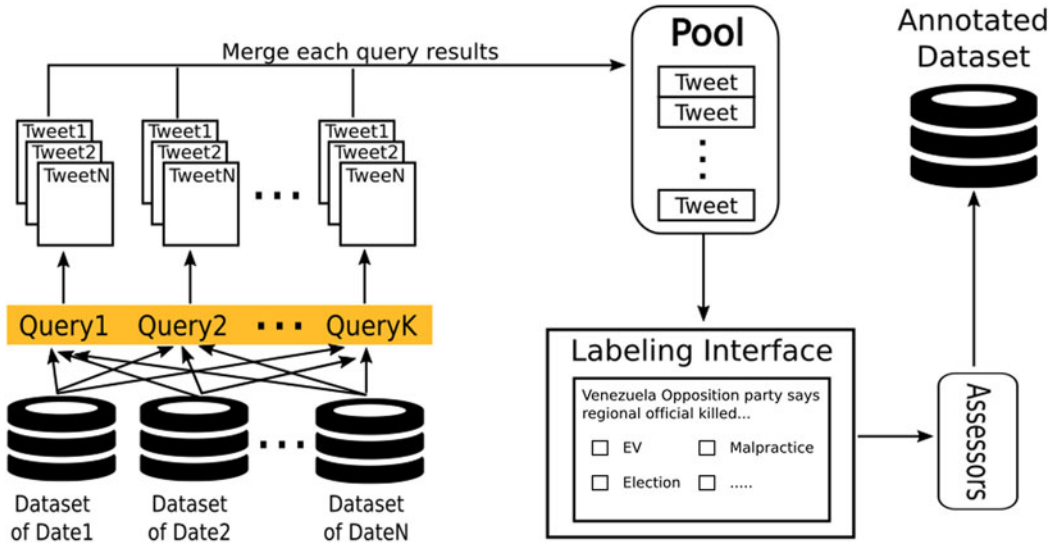

Figure 1. The information retrieval and pooling methodology used to generate tweet-based datasets.

numeric representation of the text that contains both information regarding the word itself as well as information regarding the linguistic *context* of that word.

In *word2vec*, word embeddings are randomly generated by maximizing the average log probability of the linguistic context *c* given word *w* (Goldberg and Levy, 2014). Rather than consider words as atomized features to be represented as a series of ones or zeros, *word2vec* transforms these sparse word representations into dense real-valued vectors.

Conceptually, the contextual meaning of a word is not determined by viewing a word in isolation; one also needs an understanding of the surrounding linguistic context. Word embeddings are assigned a real-valued vector such that words which appear in similar contexts cluster together in the embedding space. The assignment of vector values to words and the dimensions of the embedding space are meaningful only in the context of the embeddings themselves. While each word is given a vector representation, the values of these vectors have no valuation attached to them. Words like violence, death, and assault will cluster together in the embedding space because *word2vec* recognizes that these words co-occur more frequently together than do other words like fraud, cheat, or vote. This allows our neural network to learn not only which words are predictive of electoral violence, but also to learn other words in similar contexts that also report on violent events. Some visualizations of word embedding space are provided in the supplementary materials.

3. Describing the convolutional neural network

Convolutional neural networks are quite complex, and the number of hyperparameters that are used to train the network can represent an extreme case of Gelman and Loken (2013)'s "garden of forking paths." This section introduces the basic components of our neural network and explains their functions. A secondary subsection describes our parameterization of the neural network.[5]

[5]Replication code and data for this research can be found at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/KFVTMI. We wish to note that there is substantial debate surrounding the extent to which complicated methodologies which rely on the setting of multiple hyper-parameters and even starting seed values, like neural networks, replicate exactly the same every time (Ferro *et al.*, 2016; Ferro and Kelly, 2018; Muchlinski *et al.*, 2019). Though perhaps the exact number of events discovered, the number of tweets captured, and metric values may differ during future replications, the neural network will out-perform the support vector machine if our methodology is followed.

Neural networks apply non-linear transformations to the input data, allowing nearly any relationship between the response and predictor variables. This makes neural networks ideal tools for textual analysis tasks as they can learn functional mapping between any word in a vocabulary and the probability that a text references a violent event. Given we do not assume to know the entire vocabulary which is predictive of such events, we allow the neural network to learn these linguistic features directly from the data itself. Here we introduce the basics of neural networks and expand these fundamentals to describe our convolutional neural network in the following subsections.

3.1 The basics of a neural network

A neural network used for the classification of a binary variable is a nonlinear and interactive extension of the familiar logistic regression model (Beck *et al.*, 2000). Logistic regression fits one function to estimate the relationship between a dataset of features **X** and the probability that a tweet references a violent event, call this π_i. A neural network can fit N approximations of this relationship. Statistically, we begin by assuming the data **Y**, which represents observations regarding electoral violence, are defined according to a known statistical distribution.

$$Y_i \sim \text{Bernoulli}$$

The standard logistic regression model expresses the relationship between **X** and π as

$$\pi_i = \text{logit}(X_i\beta) = \frac{1}{1 + \mathrm{e}^{-X_i\beta}},$$

where i denotes the i-th tweet in the dataset. A neural network extends the logistic regression model in the following way:

$$\pi_i = \text{logit}[\gamma_0 + \gamma_1\text{logit}(X_i\beta_1) + \gamma_2\text{logit}(X_i\beta_2) + \cdots + \gamma_N\text{logit}(X_i\beta_N)]. \tag{1}$$

$$\pi_i = \text{logit}[\gamma_0 + \gamma_1\text{logit}(\pi_1) + \gamma_2\text{logit}(\pi_2) + \cdots + \gamma_N\text{logit}(\pi_N)]. \tag{2}$$

The γ terms in these equations are weights representing how much confidence the network attaches to a probability estimate of electoral violence. More generally, we can write the weighted product of $\gamma_n\pi_n$ as a single weight matrix **W**, and replace the logistic functional form with a more general form **x**. Rewriting Equations 1 and 2 with more general notation, we obtain:

$$f(\mathbf{x}) = \pi_i = \mathbf{x_1W_1} + \cdots + \mathbf{x_nW_n}. \tag{3}$$

The functional form **x** is estimated directly from the data by computation units in the network called "neurons." Neurons are mathematical functions that apply nonlinear transformations of the data to various parts of the network. In our network, we use a type of neuron called a Rectified Linear Unit or ReLU, which passes tweets to other layers of the network if and only if the neuron receives sufficient evidence that a given tweet references electoral violence. The network learns which linguistic features of a tweet reference violence through its learning procedure, called backpropagation, which passes errors in classification backward through the network. This backpropagation provides the neurons with information regarding whether a tweet was misclassified or correctly classified. If a tweet was correctly classified, the information passed to the neurons by backpropagation does not substantially alter the weight matrix for a tweet. If, however, a tweet was misclassified, the neural network will update its information regarding which features

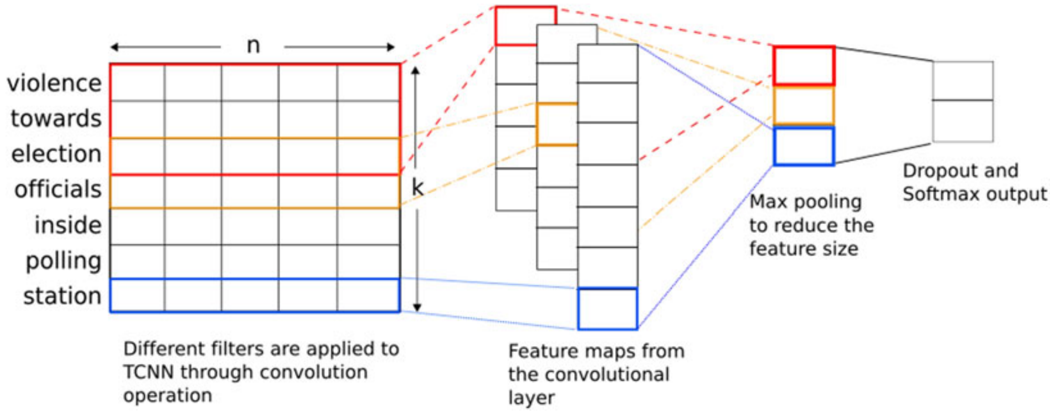

Figure 2. Visual depiction of the convolutional neural network adapted from Kim (2014).

are indicative of electoral violence, attaching different weight to different features. This process continues by gradient descent until a global minimum is reached.

3.2 Explaining the convolutional neural network

A neural network like the one outlined above represents the simplest architecture of a neural network. Convolutional neural networks apply further transformations to the data. These transformations are applied in different *layers* of the network. Our convolutional neural network is based upon the architecture described in Kim (2014) and Severyn and Moschitti (2015), and consists of a convolutional layer, a max pooling layer, a dropout layer, and a softmax layer. Each of these are explained in turn. To facilitate ease of understanding, a diagram of our network is presented in Figure 2. Our neural network was run using a standard Windows desktop computer with an Intel CPU 3.6 GHz i7 processor and 16 GB of RAM. Such a setup is readily available and inexpensive, making this methodology competitive with current event data projects in political science.

3.2.1 The convolutional layer

Consider the tweet "Ten people are dead in election day violence." A vectorized representation of this tweet generated from *word2vec* is input to the first layer of the network, the convolutional layer. This layer passes a series of filters over the tweet. These filters read the tweet, learning which features indicate whether or not a tweet references violence by generating a series of feature maps. These are d-dimensional vectors of tweet features—in essence vectorized words or n-grams—that are learned to be representative of violence. To take the example shown in Figure 2, three feature maps may be generated from this tweet. The first is "ten people dead," the second is "dead in election," and the third is the word "violence." These feature maps can be different lengths and are generated endogenously. They are passed to the next layer of the network, the Max Pooling Layer, which concatenates these features maps together, then reduces them into a sparser, but more easily learned, representation.

3.2.2 The max pooling layer

The feature maps generated by the convolutional layer are diverse. Variation in the length of the feature maps may increase computation time and reduce classification accuracy as the network has a greater number of parameters to learn. The max pooling layer uses dimensionality reduction to concatenate the feature maps into a single vector, then removes all but the most salient features of these maps, speeding up computation time by reducing the number of features the network has to learn.

In our example, the max pooling layer takes the three feature maps generated by the convolutional layer: “ten people dead,” “dead in election,” “violence,” and reduces them into a smaller vector that reads “people dead violence.” Though this reduced tweet is not grammatically correct, it greatly assists the neural network to classify such a tweet as referencing violence. As error rates are passed back through the max pooling layer by backpropagation, fewer weight matrices are updated—because there are fewer words in the tweet—and the neural network will learn this “tweet” references violence more quickly and accurately than if it had to update weights for all features in the longer tweet “ten people dead in election day violence.”

3.2.3 Dropout and softmax layers

The dropout layer acts like an ensemble learner. With a certain probability p it keeps a set of neurons in the network active and switches the others off. We set p equal to 0.5 such that through each iteration of training, the network used a randomly selected half of its computational units to classify tweets. This allows us to avoid overfitting. At the end of the training procedure, the results of training are averaged over all training iterations. Since each training iteration used a random configuration of neurons, our training results represent the weighted average of thousands of network configurations, reducing bias. We further used $L2$ regularization to control overfitting.

The softmax layer is the final output layer. It uses a variation of the logistic function to classify the tweets passed by the max pooling layer into mutually exclusive categories of electoral violence or not electoral violence. Concluding with our running example, the concatenated tweet “people dead violence” is passed from the max pooling to the softmax layer. The softmax layer compares the features in this reduced-form tweet—“people dead violence”—to the hand coded class label and assigns a probability to the category of violence. Once this probability is assigned, backpropagation updates the weight matrices for each individual feature in the reduced tweet to reflect new information that features like “people,” “dead,” and “violence” are predictive of the class of violence.

3.3 The parameterization of our neural network and word embedding model

To train the convolutional neural network and support vector machine, we use five fold cross-validation, such that in each fold, three partitions are used for training, one partition for validation, and one partition for test. Afterward, the overall performance on the test instances is assessed by averaging the scores across all folds. The support vector machine was initialized using the *LinearSVC* model in the Python *Scikit-learn* library (Pedregosa *et al.*, 2011) and the parameter c was tuned using five fold cross-validation. Our neural network was also coded in Python using the *Tensorflow* library (Abadi *et al.*, 2016). For all the experiments conducted with the neural network, we use three filter sizes $m = \{1, 2, 3\}$, stride $s = 1$, window size $W = \{1, 5, 10\}$, and the dimension size of the word embedding model was set to $D = \{200, 500, 800\}$. For each filter size, 200 filters are applied to the convolutional layer, producing 600 feature maps in total. Window size set to 10 and embedding size equal to 800 produced the best results, so we set those values as default parameter values for all elections. For our word embedding model, we set the batch size to 50, minimum word frequency to 5 and iterations to 5. As the distribution of tweet classes were imbalanced, we also set negative sampling to 10 as an additional parameter and conduct experiments by varying negative sampling size $ns = \{2, 10\}$. The class imbalance in the training data is shown in the supplementary materials. We note here that tweets describing electoral violence are between 5 and 6 percent of all tweets across all three elections. To correct for this class imbalance, a weighted cross-entropy loss function was used to give a larger weight to the minority class for the neural network. For the support vector machine, we set the class weights parameter of the model to “balanced” in the Scikit-Learn library.

Robustness tests of our results across various combinations of window size and word embedding dimension sizes, are extensively covered in Tables 3 and 4 of Yang *et al.* (2018) for the Venezuelan election and the Philippine election, respectively. Robustness tests displaying the results of varying the negative sampling size are reported in Table 5 of Yang *et al.* (2018).[6] The neural network consistently outperforms the support vector machine as measured by the *F*-1 score across every window size and every dimension size of the word embedding model for both elections.

4. Results: estimating electoral violence

Here we describe the classification accuracy of our neural network compared to a baseline, derive the total number of violent events for each election, and compare the number of events discovered by our neural network to those reported in other event datasets including ACLED, ICEWS, and SCAD. We chose a support vector machine for the baseline model because this algorithm has been shown to accurately classify textual data referencing various forms of political violence (D'Orazio *et al.*, 2014). We further ensure that the events our neural network has discovered actually occurred using two methods. First, by verifying the veracity of each event using local media sources. Tweets reporting violent events often contain other media, including linked news reports that we can independently verify. The second is to create a qualitative coding ontology which we apply to all data estimated by our neural network as well as all violent data occurring during the two-month electoral period in ACLED, ICEWS, and SCAD. Qualitatively coding these data gives us greater insight into whether a violent event that occurred was causally related to the election. To code this information qualitatively, we rely on linked news stories in our tweets, but because we do not have access to the textual sources underlying the data in the other datasets, we are forced to make judgments about how likely these events were related to the election.[7]

We use several metrics to compare classification accuracy between the neural network and support vector machine including precision, recall, and the *F*-1 score. Precision, is defined as *true positives*/(*true positives* + *false positives*), while recall is given by *true positives*/(*true positives* + *false negatives*). The *F*-1 score is the harmonic mean of precision and recall.

To briefly summarize, we find that our neural network more accurately classifies tweets that report actual electoral violence compared to a support vector machine. The neural network identifies thousands more violent tweets in the data, allowing us to discover many violent events that would have gone undiscovered by utilizing other methods. We further find substantial concept validity of our data by measuring the temporal distribution of electoral violence and by qualitatively coding our observations.

4.1 Comparing classification accuracy

After the neural network and support vector machine were trained, the parameters of each algorithm are saved, and classification accuracy is assessed using a hold-out test dataset of tweets. Because each model was trained separately for each election, test set accuracy was also assessed for each election separately. Table 1 compares the classification accuracy of the neural network compared to the support vector machine. Because it combines information from both precision and recall, we utilize the *F*-1 score as our primary metric of classification accuracy. As is clear

[6]The article containing our robustness checks, though written by some of our coauthors does not have the same focus as this manuscript. Yang *et al.* (2018) examined the ability of a convolutional neural network to accurately classify tweets related to electoral violence and malpractice. This was a purely experimental paper, and the current manuscript has the empirical goal of extending the work of Yang *et al.* (2018) by comparing the classification ability of the neural network and support vector machines to prominent datasets in political science which have been used to study electoral violence.

[7]ACLED and SCAD contain some notes about each event taken from the underlying text, and we use these notes to assist our qualitative coding of that data as well.

Table 1. Classification accuracy for electoral violence tweets

Country	Classifier	Precision	Recall	*F*-1
Venezuela	SVM	71.3	72.0	71.5
	CNN	74.3	75.4	74.6
Philippines	SVM	67.1	76.1	70.9
	CNN	78.7	74.0	75.9
Ghana	SVM	75.1	77.6	76.0
	CNN	82.6	72.9	77.1

from the table, the neural network more accurately discovers electoral violence in social media as shown by the higher *F*-1 scores across all elections. These differences, further, are statistically significant using McNemar's test for the elections in the Philippines ($p = 0.0153$) and Venezuela ($p = 0.0218$), but are not significant for the Ghanaian election ($p = 0.0736$)

Compared to the support vector machine's performance on the entire tweet-level data (training and test sets), the neural network identifies 27,282 (47 percent) additional tweets referencing violence during the Venezuelan election, 1135 (13 percent) additional tweets for the Philippine election, and 18,868 (76 percent) tweets for the Ghanaian election. These are large numbers, and it is likely that among these tens of thousands of additional tweets, there are tweets referencing violent events that the support vector machine has not discovered.

In effect, this replicates under-reporting bias all over again. Given the support vector machine fails to identify over 40,000 tweets that actually report on violent events, we cannot be confident that this method will be able to provide an accurate accounting of the total number of such events for each election. However, it may also be the case that the neural network is simply producing many more false positives. What is needed is a way to determine if the two algorithms detect a different number of violent events, rather than simply detecting a different number of tweets referencing violence. If the tweets classified by the neural network reference a larger number of violent events, we can more accurately determine the level of electoral violence for each election.

4.2 Discovering the number of violent events with clustering

Because multiple tweets may report on the same event, counting each tweet as a single event would provide an inflated estimate of violence across our three elections. To discover how many events actually occurred we utilize K-means clustering which partitions observations into k clusters, where k is chosen by the researcher. For each election, we set $k = 100$, 100 being a large enough number such that all violent events could potentially be observed.[8] With word embeddings, tweets which report on the same event will contain similar linguistic information, and thus have similar numerical values. Tweets with similar values will cluster closely together, while tweets reporting on different events should cluster further away in the data space. Partitioning this space into clusters assists in the discovery of individual violent events.

The results of our clustering analysis are shown in Table 2, which shows the number of events discovered by each algorithm across all elections, as well as the difference in events discovered between the neural network and the support vector machine. The neural network discovers an additional 15 violent events in Venezuela, 11 in the Philippines, and 3 in Ghana compared to the support vector machine.

[8]The exact choice of k in our analysis does not matter as long as it is sufficiently large to capture all relevant events. The idea is to have a reasonable number of clusters for authors to manually validate the events. A small number will lead to clusters with mixed events but a very large number will necessitate that researchers spend more time to check the homogeneity of event clusters. For some experiments demonstrating how the choice of k affects inter- and intra-tweet cluster homogeneity, see the supplementary materials.

Table 2. Number of violent events per election

Country	Classifier	Number of violent events	Difference
Venezuela	SVM	32	
	CNN	47	+ 15
Philippines	SVM	36	
	CNN	47	+ 11
Ghana	SVM	42	
	CNN	45	+ 3

4.3 Electoral violence in social media: measuring concept validity

Here we examine the concept validity of our estimates of electoral violence by examining temporal trends in violence during each election compared to that of violent events recorded in other established event datasets including ACLED, ICEWS, and SCAD. The objective of electoral violence is to influence the electoral process (Höglund, 2009). Because violence can be strategically deployed to affect voting patterns, electoral violence tends to increase in frequency as election-day approaches (Harish and Little, 2017). Therefore, we should expect to discover an increase in violence in the days immediately surrounding each election, as electoral actors strategically deploy violence in order to affect the results of the election according to their particular ends.

Figure 3 shows temporal trends in violence across all three elections. The temporal trends discovered by our neural network are quite different from those recorded elsewhere. This suggests that our method detects a different type of violence. For each election, our neural network detects substantial increases in electoral violence in the days immediately surrounding each election (election day is represented as 0 on the x-axis), a trend no other dataset picks up, except for ACLED in the Ghanaian election. This suggests that our estimates of electoral violence have good concept validity. By measuring violence that peaks on election-day, our algorithm is accurately estimating political violence that is directly related to the electoral process.

4.4 Qualitatively coding electoral violence

To ensure the neural network has discovered violent events that are correlated with the electoral process, we developed a qualitative coding ontology of all events discovered by our neural network as well as all events recorded by ACLED, ICEWS, and SCAD. We separate events into six mutually exclusive categories: strongly related to the election, probably related to the election, probably not related to the election, not related to the election, related to the election but not violent, and the final category being not enough information to code. We develop a qualitative codebook to separate events into these mutually exclusive categories. It is available in the supplementary materials. For data collected by our neural network, we relied on linked news articles to determine the association of each event to each election. When tweets contained no linked news article, we could not determine if there was sufficient information to determine the causal relation of an event to the election.

Events were coded as strongly related to the election if at least one actor had strong connections to the electoral process, such as being a political party or activist, *and* if it could be corroborated through a description of the event that the motive for the violence was related to the election.[9] Events are probably related to the election if at least one actor could be linked to the electoral process, but if the motive for engaging in the violence remained unclear. Events for

[9]ICEWS, which is the only alternative source of event data for two elections, does not contain any additional descriptions of events, like notes, which can be used to get a better understanding of the event. Fortunately, ACLED and SCAD do contain such information, and we use this additional data to assist in our qualitative coding of data from ACLED.

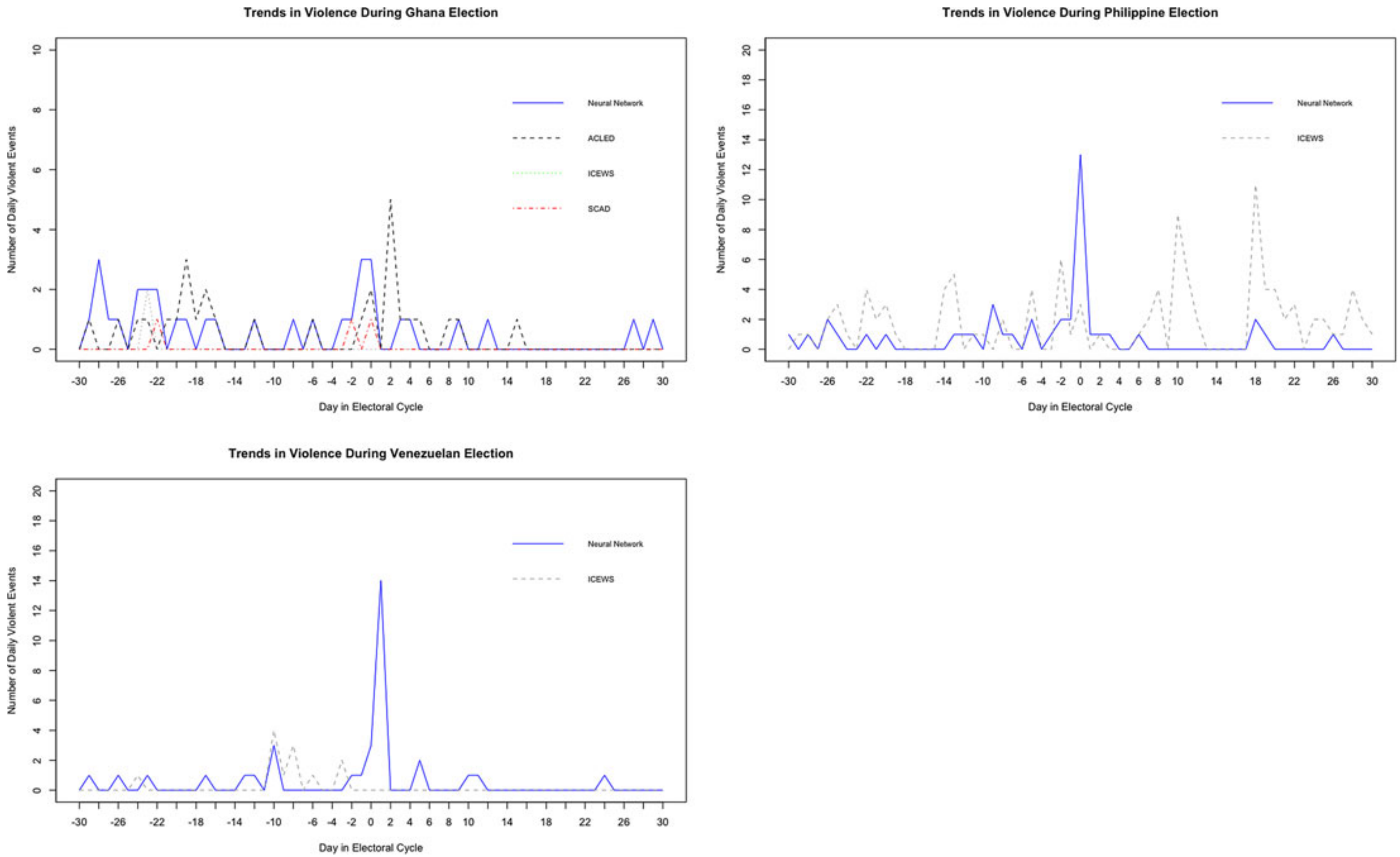

Fig. 3. Temporal trends of electoral violence.

which the identity of either actor is ambiguous (i.e., "vigilante militia," or "civilians") and for which the motive is unclear are coded as probably not related to the election. Events are coded as not related to the election if no actor has an identification that can be clearly traced to the electoral process, and if the motive for the incident is clearly not related to the election. Events could also be coded as related to the election, but the event was not violent in nature. An example would be if members of a political party staged a rally, and no violence broke out. Finally, there were often events which could not be corroborated using alternative sources of data, or the identity of at least one actor was completely unknown (i.e., the identity of the perpetrator or victim in ACLED, ICEWS, or SCAD was left blank). These events were coded as not having enough information to code.

The qualitative coding of the data collected by our neural network confirms our earlier results, with an important caveat. The neural network is able to determine with a high degree of confidence whether events are correlated with the electoral process, but only for English language tweets. For the Venezuela election, there is little difference in the percentage of tweets qualitatively classified as strongly or probably related to the election compared to ICEWS or our neural network. We discovered 47 violent incidents, compared to 16 recorded by ICEWS. The proportions of events that are related to the election among the two datasets, however, are similar. The neural network suggests that 53 percent of events discovered are strongly or probably related to the Venezuelan election. By contrast, 56 percent of events in ICEWS fall into the same two categories. While our neural network detects a greater number of violent events, it does no better at classifying the relationship of those events to the electoral process.

When the neural network is trained on English language tweets, however, it is able to far surpass other datasets in determining which violent events are related to the electoral process. These results are shown in Table 3. For instance, our qualitative coding of the Philippine election demonstrates most recorded violence in ICEWS is not related to the electoral process. ICEWS records 106 violent events, compared to 51 discovered by our neural network. Seventy-eight percent of all events recorded by the neural network during the Philippine election are either strongly or probably related to the election. By contrast, only 7 percent of all observations recorded in ICEWS could be considered to be probably election related, and no event could be considered to be strongly related to the election. The vast majority of violence contained in ICEWS are false positives reporting the killings of drug dealers or users, or military actions against rebel groups like Abu Sayyaf. The first category of events are clearly unrelated to the election. The second could plausibly be related to the electoral process, but the insurgency against such rebels has long predated the 2016 election, so such violence is quite unlikely to have electoral causes.

Similar results hold for the election in Ghana. Our neural network discovered 45 violent events, compared to 29 in ACLED, 2 in ICEWS, and 3 in SCAD.[10] ACLED does a comparatively good job in correctly identifying electoral violence. It suggests 44 percent of the 29 events are strongly or probably related to the election. Our neural network, however, is twice as accurate in correctly identifying true positives. ACLED incorrectly classifies 40 percent of its events as violent when there is considerable evidence in ACLED itself to suggest they were peaceful. Adding the additional 16 percent of events that are not related to the election, ACLED's false positive rate is 56 percent—12 percent higher than its true positive rate.

These results demonstrate that our machine learning platform is vastly more accurate in correctly identifying electoral violence as compared to existing event datasets. This suggests statistical models of electoral violence developed using these event datasets should be interpreted with caution because rates of misclassification on the dependent variable appear to be substantial. Scholars working in this field may wish to utilize alternative sources of information to measure electoral violence. Social media is one useful source of text, but many more may exist. Finally, our

[10]We report only the comparison with ACLED in Table 3 due to these small sample sizes.

Table 3. Rates of qualitative classification for each election according to different datasets

Election	Strongly	Probably	Probably not	Not related	Not violence	No info
Venezuela (NN)	0.33	0.20	0.04	0.09	0.15	0.20
Venezuela (ICEWS)	0.06	0.50	0.19	0.25	0.00	0.00
Philippines (NN)	0.53	0.25	0.02	0.06	0.02	0.10
Philippines (ICEWS)	0.00	0.07	0.08	0.74	0.00	0.10
Ghana (NN)	0.66	0.18	0.00	0.02	0.07	0.07
Ghana (ACLED)	0.40	0.04	0.00	0.16	0.40	0.00

results show that the choice of machine learning algorithm matters for measuring violent events in text. The support vector machine possibly under counted the true rate of violence during these three elections, contributing to another possible source of statistical bias. While we have demonstrated our neural network is able to estimate electoral violence more accurately than existing methods, our machine learning method can be applied to different, and much broader, classes of political phenomena. While convolutional neural networks can be quite complex, they are useful to the broader community of political methodologists or any researcher who simply wishes to measure data developed from unstructured text more accurately.

5. Conclusions

Election-related violence plagues countries around the world. It impedes the peaceful transition of power and can prevent citizens from exercising their constitutionally protected rights to choose their elected leaders. Despite a proliferation of recent research into this phenomenon, the concept of electoral violence still remains ill-defined and most studies assume, rather than validate, that violence occurring during elections actually seeks to affect the electoral process in some way. We have developed a new method to collect, code, and validate data mined from social media to estimate trends in electoral violence during three elections in different countries. We have demonstrated that our machine learning pipeline more accurately measures electoral violence compared to existing datasets and other state of the art machine learning algorithms. We show that the trends in violence uncovered by our neural network peak on or near election day, and we demonstrate through qualitative coding that the data we have collected have a stronger causal connection to the electoral process compared to existing data in ACLED, ICEWS, and SCAD.

Electoral violence can take a variety of forms, is perpetrated by many different actors, and often falls short of erupting into full-fledged civil conflict. Thus, it can be difficult to correlate the presence of any violent event that occurs to the election itself. We have provided scholars with a method of moving past this technical barrier. Because it is a more direct type of reporting, often from observers of the event itself, social media may offer a more straight forward way to discover violent events. We have shown that word embeddings, further, provide machine learning classifiers greater accuracy in identifying instances of violence in text. These tools currently show the most promise in enhancing natural language processing pipelines, like ours, and classifiers trained using such embeddings have proven to be more accurate than commonly utilized tools across the discipline (Beieler, 2016). Our results demonstrate that word embeddings outperform traditional bag-of-words approaches to textual analysis. The ability of word embeddings to encode not just about the word itself, but its linguistic relationship to other parts of the text, enhances classification accuracy, assisting the discovery of violent events. Our neural network classifier, further, has been demonstrated to be a more accurate algorithm for identifying instances of violence in social media text compared to other machine learning algorithms, like support vector machines, that have previously been utilized for similar tasks.

We are aware of the limitations of our methodology. Because we derive our data on electoral violence from social media, Internet and social media access is a prerequisite for measuring electoral violence. Scholars utilizing our methodology will not be able to measure electoral violence in countries where citizen access to the Internet is limited. Scholars must also remain vigilant against the spread of misinformation throughout social media networks, and validate the information they gather against alternative sources. The performance of our methodology may deteriorate somewhat in countries where English is not the primary language used across social media platforms. Further application of this methodology to multilingual datasets of tweets is warranted to resolve this possible limitation.

Our results also demonstrate that the granularity by which media is reported matters. Both ACLED and our neural network utilize national, regional, and local reporting sources. Of the three event datasets, ACLED seems to be more accurate in identifying electoral violence, though it is extremely difficult to determine if this result would hold across additional elections. Because ACLED does not contain data on Venezuela or the Philippines for this time period, a more thorough comparison is not possible within the scope of this project. While ACLED is the most accurate in identifying the nature of electoral violence, it is interesting to note that our neural network discovered 14 additional violent events during the Ghanaian election. A detailed analysis of why our neural network discovered more events, even holding the locality of news reporting more or less constant, is unfortunately outside the scope of this project. Despite our ignorance on this issue, we can heartily advise scholars to, where possible, utilize the most disaggregated source of reporting that is relevant for their research needs.

Perhaps unexpectedly, we have also uncovered a result suggesting that the choice of algorithm used to discover violent events in text matters. Given the inherent costs of failing to accurately diagnose potential conflicts, including electoral violence, we suggest scholars utilize the most accurate methods available to ameliorate any possible source of under-reporting bias. Though neural networks are quite complex, and the process by which they produce their estimates are a subject of much current research, they are worth using for tasks in which the box of causality can remain black. If researchers only wish to recover the most accurate estimates of violence from text, it makes sense to use the most accurate method. Of course a sophisticated machine learning algorithm cannot substitute for the watchful eye of an expert researcher, but it can be a powerful tool in the right hands. Given our success in estimating electoral violence, we invite scholars of political violence more generally to embrace this new technology and take a dive in the deep end.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/psrm.2020.32

Acknowledgements. We are very grateful to Patrick Brandt, Vito D'Orazio, Ben Bagozzi, the other participants at the APSA Event Data Mini-Conference Program 2018, and the PSRM reviewers and editors for very helpful comments on previous versions of this article. The authors acknowledge support from the Economic and Social Research Council, Grant No. ES/L016435/1.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M and Kudlur M (2016) Tensorflow: a system for large-scale machine learning. *OSDI* **16**, 265–283.

Amati G, Amodeo G, Bianchi M, Marcone G, Bordoni FU, Gaibisso C, Gambosi G, Celi A, Di Nicola C and Flammini M (2011) FUB, IASI-CNR, UNIVAQ at TREC 2011 microblog track. In *Proceedings of TREC.*

Bagozzi BE and Koren O (2017) Using machine learning methods to identify atrocity perpetrators. In *2017 IEEE International Conference on Big Data.*

Beck N, King G and Zeng L (2000) Improving quantitative studies of international conflict: a conjecture. *American Political Science Review* **94**, 21–35.

Beieler J (2016) Generating politically-relevant event data. *arXiv preprint arXiv:1609.06239.*

Bengio Y, Ducharme R, Vincent P and Janvin C (2003) A neural probabilistic language model. *Journal of Machine Learning Research* **3**, 1137–1155.

Birch S and Muchlinski D (2020) The dataset of countries at risk of electoral violence. *Terrorism and Political Violence* **32**(2), 217–236.

Boschee E, Lautenschlager J, O'Brien S, Shellman S, Starz J and Ward M (2015) Icews coded event data. *Harvard Dataverse* **12**.

Brass PR (1997) *Theft of an Idol: Text and Context in the Representation of Collective Violence*. Princeton University Press.

Butcher C and Goldsmith BE (2017) Elections, ethnicity, and political instability. *Comparative Political Studies* **50**, 1390–1419.

Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K and Kuksa P (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12**, 2493–2537.

Cook SJ, Blas B, Carroll RJ and Sinha S (2017) Two wrongs make a right: addressing underreporting in binary data from multiple sources. *Political Analysis* **25**, 223–240.

Daxecker UE (2012) The cost of exposing cheating: international election monitoring, fraud, and post-election violence in Africa. *Journal of Peace Research* **49**, 503–516.

Daxecker UE (2014) All quiet on election day? International election observation and incentives for pre-election violence in African elections. *Electoral Studies* **34**, 232–243.

D'Orazio V, Landis ST, Palmer G and Schrodt P (2014) Separating the wheat from the chaff: applications of automated document classification using support vector machines. *Political Analysis* **22**(2), 224–242.

Doyle A, Katz G, Summers K, Ackermann C, Zavorin I, Lim Z, Muthiah S, Butler P, Self N, Zhao L, *et al.* (2014) Forecasting significant societal events using the embers streaming predictive analytics system. *Big Data* **2**, 185–195.

Dunning T (2011) Fighting and voting: violent conflict and electoral politics. *Journal of Conflict Resolution* **55**, 327–339.

Earl J, Martin A, McCarthy JD and Soule SA (2004) The use of newspaper data in the study of collective action. *Annual Review of Sociology* **30**, 65–80.

Ferro N and Kelly D (2018) SIGIR initiative to implement ACM artifact review and badging. In *ACM SIGIR Forum*, vol. 52, ACM, pp. 4–10.

Ferro N, Fuhr N, Järvelin K, Kando N, Lippold M and Zobel J (2016) Increasing reproducibility in IR: findings from the Dagstuhl seminar on "reproducibility of data-oriented experiments in e-science." In *SIGIR Forum*, vol. 50, pp. 68–82.

Fjelde H and Höglund K (2016) Electoral institutions and electoral violence in sub-Saharan Africa. *British Journal of Political Science* **46**, 297–320.

Gelman A and Loken E (2013) The garden of forking paths: why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

Goldberg Y (2016) A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research* **57**, 345–420.

Goldberg Y and Levy O (2014) word2vec explained: deriving Mikolov *et al.'s* negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Grimmer J and Stewart BM (2013) Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* **21**, 267–297.

Hafner-Burton EM, Hyde SD and Jablonski RS (2014) When do governments resort to election violence? *British Journal of Political Science* **44**, 149–179.

Harish S and Little AT (2017) The political violence cycle. *American Political Science Review* **111**, 237–255.

Hendrix CS and Salehyan I (2015) No news is good news: mark and recapture for event data when reporting probabilities are less than one. *International Interactions* **41**, 392–406.

Höglund K (2009) Electoral violence in conflict-ridden societies: concepts, causes, and consequences. *Terrorism and Political Violence* **21**, 412–427.

Hyde SD and Marinov N (2012) Which elections can be lost? *Political Analysis* **20**(2), 191–210.

Jackoway A, Samet H and Sankaranarayanan J (2011) Identification of live news events using Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ACM, pp. 25–32.

Kim Y (2014) Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Larson J, Nagler J, Ronen J and Tucker J (2016) Social networks and protest participation: evidence from 93 million Twitter users.

Lin Y, Shen S, Liu Z, Luan H and Sun M (2016) Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, vol. 1, pp. 2124–2133.

Macdonald C, McCreadie R, Santos R and Ounis I (2012) From puppy to maturity: experiences in developing Terrier. In *Proceedings of OSIR at SIGIR*, pp. 60–63.

Mandelbaum A and Shalev A (2016) Word embeddings and their use in sentence classification tasks. *arXiv preprint arXiv:1610.08229*.

Mikolov T, Sutskever H, Chen K, Corrado G and Dean J (2013) Distributed representations of words and phrases and their compositionality. *arXiv preprint*.

Muchlinski DA, Siroky D, He J and Kocher MA (2019) Seeing the forest through the trees. *Political Analysis* **27**, 111–113.

Ounis I, Amati G, Plachouras V, He B, Macdonald C and Lioma C (2006) Terrier: a high performance and scalable information retrieval platform. In *Proceedings of the OSIR Workshop*, pp. 18–25.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, *et al.* (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830.

Petrovic S, Osborne M, McCreadie R, Macdonald C, Ounis I and Shrimpton L. (2013) Can Twitter replace newswire for breaking news?. *Seventh international AAAI conference on weblogs and social media*.

Raleigh C, Linke A, Hegre H and Karlsen J (2010) Introducing ACLED: an armed conflict location and event dataset: special data feature. *Journal of Peace Research* **47**, 651–660.

Ramakrishnan N, Butler P, Muthiah S, Self N, Khandpur R, Saraf P, Wang W, Cadena J, Vullikanti A, Korkmaz G, *et al.* (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*, ACM, pp. 1799–1808.

Salehyan I, Hendrix CS, Hamner J, Case C, Linebarger C, Stull E and Williams J (2012) Social conflict in Africa: a new database. *International Interactions* **38**, 503–511.

Schrodt PA, Yonamine J and Bagozzi BE. (2013) Data-based computational approaches to forecasting political violence. In Subrahmanian VS (ed). *Handbook of Computational Approaches to Counterterrorism*. New York, NY: Springer Science & Business Media, pp. 129–162.

Severyn A and Moschitti A (2015) Unitn: training deep neural convolutional neural networks for Twitter sentiment classification. In *Proceedings of SemEval*, pp. 464–469.

Staniland P (2014) Violence and democracy. *Comparative Politics* **47**, 99–118.

Steinert-Threlkeld ZC, Mocanu D, Vespignani A and Fowler J (2015) Online social networks and offline protest. *EPJ Data Science* **4**, 19.

Voorhees EM and Harman DK (eds) (2005) *TREC: Experiment and Evaluation in Information Retrieval*, vol. **63**. Cambridge: MIT Press.

Weidmann NB (2015) On the accuracy of media-based conflict event data. *Journal of Conflict Resolution* **59**, 1129–1149.

Weidmann NB (2016) A closer look at reporting bias in conflict event data. *American Journal of Political Science* **60**, 206–218.

Yang X, Macdonald C and Ounis I (2018) Using word embeddings in Twitter election classification. *Information Retrieval Journal* **21**, 183–207.

Zeitzoff T (2011) Using social media to measure conflict dynamics: an application to the 2008–2009 Gaza conflict. *Journal of Conflict Resolution* **55**, 938–969.

Cite this article: Muchlinski D, Yang X, Birch S, Macdonald C, Ounis I (2020). We need to go deeper: measuring electoral violence using convolutional neural networks and social media. *Political Science Research and Methods* 1–18. https://doi.org/10.1017/psrm.2020.32