Using Machine Learning Methods to Identify Atrocity Perpetrators

Benjamin E. Bagozzi
Dept. of Political Science & IR
University of Delaware
Newark, USA
Email: bagozzib@udel.edu

Ore Koren
Dept. of Political Science
University of Minnesota
Minneapolis, USA
Email: koren044@umn.edu

***Abstract*—"Big data" on atrocities events are now widely analyzed in the social sciences. Unfortunately, these data often contain incomplete information on the identities of atrocity perpetrators. This study addresses this deficiency by developing a machine learning approach for the accurate recovery of unknown perpetrator identities within existent atrocities datasets. In doing so, it demonstrates how to transform and standardize a large number of auxiliary variables into text-compatible data. It next shows how to leverage this information to train a series of classifiers on observed atrocities data. After identifying the ideal set of machine learning algorithms and evaluating their performance in this context, this study then uses an ensemble of the best performing algorithms to classify all unknown atrocity perpetrators included within a prominent atrocities dataset, validating the results with external data from the Iraq conflict.**

***Keywords*-atrocities, political violence, machine learning**

I. Introduction

The intentional killing of civilians by armed combatants for political motives—which we term 'atrocities'—represents one of the most pernicious attributes of modern warfare. There is also ample evidence to suggest that this practice remains widespread [1], [2]. In light of these realities, there have been several recent efforts to collect large, fine grained datasets on atrocity events. For instance, the Armed Conflict Location and Event Data (ACLED) project [3] records several forms of 'violence against civilians' within developing countries in Africa and Asia. Likewise, [2] codes and provides annual data on one-sided (government and rebel) violence against civilians during intrastate wars, whereas the geo-referenced event dataset [4] now codes one-sided violence at a near-global scale. Lastly, and perhaps most extensively, the Political Instability Task Force's (PITF's) recently created Worldwide Atrocities Dataset records geo-located atrocity events involving at least five civilian causalities at the daily level for all countries other than the U.S., from 1995-present [5].

These data collection efforts are commendable, and each dataset mentioned above has facilitated numerous theoretical insights into the determinants of atrocities [6], [7], [8]. Yet, while the above data projects have exerted considerable effort in coding the identity of atrocity-event perpetrators, many of their recorded atrocity-events lack information on the identity of atrocity perpetrators. Consider for example, the PITF Worldwide Atrocities Dataset mentioned above. This dataset uses NGO and international newswire reports to human code atrocity perpetrators' identities according to an eight-category perpetrator ontology: state, transnational, non-state (no state sanction), non-state (state sanctioned), multiple (state), multiple (non-state), multiple (state and non-state), and unknown. Of those atrocity events recorded by the PITF for the years 1995-2013, nearly 32% of all incidents were recorded as unknown, making this designation the second most frequent perpetrator classification within the PITF data. ACLED similarly provides detailed perpetrator information within its recorded cases of violence against civilians. Of the 29,000 recorded African cases of violence against civilians during the years 1997-2014, ACLED reports 6,533 (22.5%) as having unknown perpetrators.

Incomplete information on the identities of 'political violence perpetrators' within event datasets is both a well known problem, and an understandable one. Indeed, high levels of missingness in perpetrator identities have now been shown to be a systematic problem for political violence data [9], [10]. In terms of the mechanisms generating this missingness, we note that atrocities, like many forms of contemporary political violence, often occur in rural areas, conflict zones, and/or countries lacking in press freedoms. These factors limit the completeness of political violence reporting, and NGO or news reports thereof [11], [12]—which are the primary sources used to identify and code atrocity events by the aforementioned data projects. Similarly, atrocity perpetrators often have incentives to obscure their identities in order to avoid subsequent prosecution or to instill added fear among civilians, and civilian victims may not be willing to openly speak to the press about perpetrator identities for fear of retribution. This further limits the identification of atrocity perpetrators within media reports of atrocities.

Nevertheless, accurately identifying the perpetrators of atrocities is important for both theory testing and international advocacy. Numerous scholars posit theories of atrocities that hinge on the identity of their perpetrators, and go on to test such theories with data similar to that described above [13], [6], [7], [8]. In such contexts, missing identity-

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

information on atrocity perpetrators requires that these cases be dropped, which, if they are not missing completely at random (MCAR), can bias one's estimates and conclusions.

Similarly, efforts to prosecute atrocities by domestic or international legal bodies rely on investigators' abilities to accurately identify atrocity perpetrators. And naming and shaming strategies by the media and human rights campaigners also entail that when atrocities are committed, the proper perpetrators can be identified, and shamed, so as to compel them to curtail their actions [14], [15]. If media-reported atrocities lack perpetrator information, the effectiveness of naming and shaming tools will be severely undermined. Hence, the development of methods for the recovery of atrocity perpetrators' identities is needed to (i) enable scholars to better leverage the recorded atrocities events in existing datasets for scientific research and (ii) provide more accurate atrocity information to lawmakers, investigators, and advocacy groups interested in preventing atrocities, or punishing their perpetrators.

This study develops an approach for the identification of unknown atrocity perpetrators in such circumstances. It begins with the recognition that, while atrocities datasets often lack information on perpetrator identity, they nevertheless include a wealth of meta-data on the severity, location, timing, victims, and details of any given atrocity event. For example, ACLED includes information on the identity of atrocity-victims, in addition to information on the geographic location and timing of relevant events, and the news report source. The PITF's atrocity data likewise contain information on the number of civilians killed and injured, a summary of the event, the geographic location, and the method(s) of violence used, among other variables. Using the PITF atrocities data, this study therefore develops an approach to integrate all available contextual information information, which is often both text-based and numeric, into a unified input format. These input data are then used as predictors within an ensemble of supervised machine learning (ML) techniques, so as to classify atrocity perpetrators within in-sample and out-of-sample settings.

In applying our proposed ML approach to the PITF data below, we also compare the performance of this approach to a recently proposed *multiple imputation* (MI) strategy [10]. We find that ML offers a number of advantages over MI for the task of accurately recovering perpetrator identities. Following these comparisons, we identify the optimal ML classifiers for the PITF's atrocities cases. We then apply this subset of ML classifiers—along with MI—to the PITF's atrocities cases with unknown perpetrators, so as to provide future researchers with credible perpetrator information on these events. We find in these cases that the vast majority of unknown perpetrator cases are attributable to non-state actors, rather than state-based or state-affiliated actors, which has important implications for both our theoretical understanding of atrocities, and analyses thereof. After discussing these implications, we validate our unknown perpetrator classifications against an external dataset on insurgent violence against civilians in Iraq (2004-2010).

II. Overview of Approach

Atrocities data are typically recorded at the event level, with each row in one's data corresponding to a single event that was identified from a news source or NGO/government report. Additional variables then separately record atrocity-information associated with "who" did "what" to "whom" and "where." Auxiliary information pertaining to a given event or its coding is then included within additional variable fields for the sources coded, synopses of the event itself, and other details such as the tactic(s) used, and numbers of civilians injured or killed. Depending on the data set, coding system, and coding sources, these variables will have varying levels of both detail and missingness. For our purposes, atrocity-perpetrator information corresponds to the "who" aspect mentioned above, and is typically coded as a categorical identifier for whether the perpetrator of a given atrocity was a state, rebel, militia, civilian, or international actor—alongside designations for cases where the perpetrator is "unknown."

We contend that proper recovery of perpetrator identities in cases where these actors were recorded as "unknown" is essential for both atrocities researchers and analysts. Supervised ML methods offer one means of doing so. In essence, our supervised ML approach turns missing perpetrator identities into a classification problem. A variety of different ML algorithms are then trained on the cases that *do not* have missing perpetrator information, so as to calibrate and identify the best classifiers for accurately predicting these categorical classifications "in sample." Under this framework, the large number of contextual variables for each atrocities event, including measures of victim information, geolocation, timing, textual summaries of the events, and sources used for coding, can be used as input data to classify perpetrators' identities. In these instances, one can further divide samples into training and test data for a robust evaluation of how well each classifier performs, while also minimizing threats of overfitting. After identifying the best performing classifiers, an ensemble (agreement) of these classifiers' predictions can then be derived for the missing perpetrator cases in one's data, so as to recover the most likely perpetrators involved in these instances. In this regard, our research compliments a variety of recent ML applications to the study of political violence [16], [17] and to atrocities events coding [18].

One challenge for our proposed approach concerns how best to incorporate the full set of additional contextual variables as "features" within a single ML classification task. This is a challenge because—within both the PITF examined below and the related datasets mentioned above—some contextual atrocities variables are numeric, whereas

others are textual (and are at times multiple sentences in length), and still others are categorical. Moreover, each type of contextual variable can exhibit its own pattern of missingness. Given this potential heterogeneity in variable properties, we propose a data processing step below where one "standardizes" all variables for an ML classification task by converting all variables into a single document term matrix (DTM), including NAs, while appending any numeric variable values to variables' labels in order to ensure that similar numeric entries (across variables) do not get mistaken for equivalent values during the DTM-construction and classification tasks. This strategy ensures that missingness is leveraged as an informative trait rather than listwise deleting that case. This overall approach also helps to ensure that fine grained numeric information, and rich qualitative descriptions, can be simultaneously leveraged under a single ML classification framework.

We discuss our ML approach in detail below. Before doing so, we note that one alternative means of handling unknown perpetrator identities is MI [10]. In most social science applications of MI, missing categorical values are typically imputed with continuous predictions under a multivariate normal model, with these resultant imputations either left "as is" or "discretized" after imputation. More recent MI methods increasingly allow one to impute categorical variables directly using, for instance, multinomial logit [10]. We posit that each of these approaches is likely to underperform in categorical classification relative to supervised ML. Numerous ML algorithms have been explicitly designed to outperform common parametric models in polytomous classification via innovations such as penalized maximum likelihood, classification trees, and neural networks.

As a consequence of these innovations, ML algorithms' superior performance in prediction relative to standard parametric models has been now convincingly argued and demonstrated in conflict research [16], [17]. Moreover, as elaborated upon below, the practice of ensembling several distinct ML algorithms' predictions into a single (more accurate) prediction of each missing case is likely to further improve upon the strengths of ML over MI for the task of recovering categorical perpetrator identities—given ML ensembles' widely noted advantages over individual classifiers in these regards [19], [20], [21]. Hence, while MI may be ideal for handling perpetrator identities in instances where perpetrator variables are included in a subsequent regression analysis,[1] ML is likely preferable in instances where one prioritizes the *accurate* recovery of unknown values.

There are four instances where recovering accurate perpetrator identity values with (ensemble) ML is preferable to handling missing perpetrator identities via MI. First, as noted above, *accurately* identifying human rights violators is critical to successful international naming and shaming efforts. Second, the accurate recovery of perpetrator identity is also often most relevant at the initial event *dataset creation* stage [18]. MI is limited in such settings given that event data generators will not have access to (or be able to anticipate) the additional variables that will ultimately be included within regressions analyzing their atrocities data. As researchers must include all regression variables within MI to avoid bias [22], the use of MI to recover perpetrator identities at the initial data generation stage risks introducing bias, and thus has limited applicability.

Third, researchers often make use of perpetrator identities not as control or explanatory variables, but rather, to initially subset event datasets into smaller, perpetrator-specific samples. For example, [8] develop a theory concerning the effects of droughts on armed *rebel* actors' incentives to use violence against civilians. They test this theory by subsetting the PITF data mentioned above to a sample containing *only* those atrocity-events arising from rebel perpetrators, omitting all unknown perpetrator cases. Likewise, [7] posit that government and rebel actors will each be more likely to target civilians that have ethnic ties to their opponents. To test these expectations, they analyze separate samples (and models) of state and rebel perpetrated violence. Because perpetrator identities are used in the above studies to define samples *prior to analysis*, MI has limited applicability, relative to alternate approaches that prioritize the a priori accurate classification of perpetrator.

Fourth, ML is also more tractable than MI in recovering unknown perpetrator identities when one's predictors include unstructured text. Such text is common in atrocities event datasets, where multiple variable fields often contain summaries of the news stories and sources used for coding each event, or related hard-to-quantify contextual variables (e.g., estimates of the numbers of civilians injured or killed). While MI can handle categorical variables, the number of categorical variables generated by DTMs of text (which generally equal the number of total unique words appearing across all text—often in the thousands) limits the applicability of MI due to computational and runtime concerns. Indeed, as [23] note, common MI methods work well for applications with 30-40 variables but are "especially poorly suited" for datasets with "many more variables" (pg. 562). ML methods do not face these constraints, and are hence more applicable in situations where text *and* numeric data are available as potential predictors of perpetrator identities.

III. Data Description

We apply our ML approach to the PITF Worldwide Atrocities Dataset, which is a recently developed global event dataset. Atrocities are defined by the PITF [5] as "implicitly or explicitly political, direct, and deliberate violent action resulting in the death of noncombatant civilians" (pg. 3). The PITF uses a primary set of seven international news and NGO sources to collect and code a reasonably systematic

[1]Given MI's incorporation of uncertainty at the analysis stage.

sample of atrocities occurring worldwide, beginning in 1995, and then employs human coders to accurately record each atrocity's traits and geo-location. While the PITF data continues to be released in near real time, the present analysis focuses on the 1995-2013 period, which was the most recent data available at the beginning of this study. The PITF dataset records information on both atrocity campaigns, and atrocity incidents where five or more noncombatant deaths occurred.[2] In the interest of comparability across different cases and regions, only incidents are considered below.

Beyond the occurrence of an atrocity event or campaign, the PITF data also record an extensive number of additional variables on each atrocity's characteristics. These additional variables are fully defined in [5] and include information, when available, on the date of an atrocity, its sub-national location, the identities of its victim(s) and perpetrator(s), the modes of violence and tactics used, the number of civilians killed and injured, a summary of the event, and the sources used to code an event, amongst other meta-data

Our sample contains a total of 7,127 atrocity cases. The main quantity of interest for the study at hand is the *identity* of a given atrocity's perpetrator(s). As mentioned earlier, perpetrators within the PITF data are coded within a "perpetrator" variable according to the following eight categories (i) state, (ii) transnational, (iii) non-state (no state sanction), (iv) non-state (state sanctioned), (v) multiple (state), (vi) multiple (non-state), (vii) multiple (state and non-state), and (viii) unknown. Of the atrocity events recorded by the PITF for the years 1995-2013, nearly 32% of all incidents were recorded as unknown, which accordingly makes this designation the *second most frequent* perpetrator classification within the PITF data. We present a frequency histogram of atrocities by perpetrator for our entire sample in Figure 1.

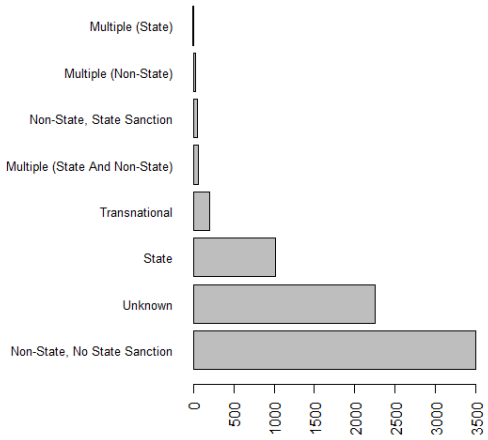

Figure 1. Atrocities by Perpetrator, 1995-2013

A. Preprocessing

Aside from our main perpetrator identity variable, all remaining PITF variables included in our analysis were transformed to the DTM-level for use as features within the classification tasks described below. DTMs are commonly used for text-as-data oriented classification tasks, wherein one seeks to classify a document-level coding based upon the text features of each document [20], [21], [24]. For such DTMs, each row corresponds to a unique document and each column corresponds to each unique word appearing across one's document-corpus. The cell values of the DTM accordingly denote the number of appearances of each word (column) within each document (row). For our application, we do not have a set of exclusively text-based features, but rather a mix of text-based variables[3] and numeric variables that we intend to use for classification. Moreover, because virtually all numeric variables included within the PITF data are categorical,[4] converting each and every variable into "text data," and then restructuring all PITF atrocities variables into a single DTM is not only appropriate, but ideal for our anticipated classification tasks.

In these regards, care must be taken to ensure that numeric or categorical values (e.g., `Scorched.Earth.Tactics= 1`) and/or common terms within one variable field are not interpreted by the ML algorithm as having the same meaning as equivalent values within other variable fields (e.g., `Deaths.Contested= 1`). To address this issue, we preprocessed our data frame prior to converting it into a DTM by appending a variable's label to each numeric (or common term) variable value where appropriate, such that the previous example yields two distinct string-values for the number "1": "Scorched.Earth.Tactics1" and "Deaths.Contested1." This was especially beneficial for ensuring that our DTM's "unique terms" included distinct values for the many binary variables' corresponding binary values in the dataset. Further, where appropriate, "NA" or "missing" records were also treated as unique strings (e.g., "Scorched.Earth.TacticsNA") and included as a feature-value for classification.

We then sought to ensure that common N-gram phrases were not broken up into individual words within our DTM when these phrases corresponded to unique constructs. This

[2]Per [5], incidents are defined as as "perpetrated by members of a single organization or communal group, or by members of multiple organizations or groups reportedly acting in concert, in a single locality within a 24-hour period" (pg. 6). Campaigns are a residual category for atrocities that lack sufficient information for the identification of incidents.

[3]Which, in some cases, correspond to single words, and in other cases, include lengthy passages of text, such as the `Description` field, which contains 2-3 sentence descriptions describing the event in narrative English.

[4]Moreover, even the numeric variables that are not categorical contain categorical values (including NAs) alongside numeric values, rendering them effectively categorical for our tasks.

was achieved by converting all common N-gram phrases[5] to unigrams via dashed-lines. Note however that we did not implement this step for variables containing longer text passages that corresponded to narratives or comments, rather than categorical (N-gram) classes. After converting each variable into appropriate 'text data,' all variables were restructured into a DTM for analysis, where a "document" in this case corresponds to a single atrocity case, and that document's "words" correspond to the combined string-converted values on each variable for that atrocity-case. This is in line with published supervised text-as-data applications [21]. All told, our final DTM had 7,127 "documents" and 2,947 unique terms.

IV. ML and MI Applications

Supervised ML methods have different strengths and weaknesses for any particular application. As such, researchers increasingly recommend ensemble methods, which allow one to leverage multiple ML or predictive algorithms within a single classification or forecasting task, so as to maximize tradeoffs between efficiency and accuracy [20]. We follow this approach here in using [21] to evaluate the following ML algorithms for the classification of our perpetrator identities: LASSO, stabilized linear discriminant analysis (SLDA), support vector machine (SVM), maximum entropy, bagging, boosting, random forests, classification trees, and neural networks. After evaluating the out-of-sample performance of each algorithm, we retain the six best preforming algorithms, and use their ensemble to classify the atrocities cases in our sample that exhibit missing perpetrator identities.

We begin by evaluating the aforementioned classifiers within our subsample of atrocities events with known perpetrators. This subsample corresponds to the 68% of our atrocities sample that recorded a perpetrator as *not* being "unknown." The resultant outcome variable that is used in the first step of our classification tasks thus corresponds to a seven-category polytomous (i.e., categorical) *known* perpetrator variable, and encompasses 4,870 (out of 7,127) total observations. For these 4,870 retained cases, we further divide this sample into 80% training data (N =3,847) and 20% test data (N =1,023).

We compare this ML approach to the MI approach proposed by [10] below. These comparisons are not intended to suggest that [10]'s MI approach is inferior to our proposed ML approach. Rather, our claim is that ML is a preferred method for the specific task of accurately recovering perpetrator identities from unknown-perpetrator cases. If the goal is instead to make inferences within a standard regression model that includes perpetrator identity as a control or explanatory variable, we recommend that researchers follow [10]. For our MI comparisons, we retain a subset of the predictor variables mentioned above, as not all ML predictors are compatible with MI.[6] For the variables retained within our MI approach, we treat missing values as missing within the MI model, rather than as unique categorical character-strings (as we do in ML). Following [10], we use these retained variables to "train" a set of 100 MI models on the known perpetrator data mentioned above, while artificially deleting the identities of our test cases so that our MI method generates predictions for these cases under a similar out-of-sample framework to that of our ML approach. While doing so, we follow [10] to treat perpetrator identity, and all other relevant dichotomous or polytomous variables, as categorical within our MI-routines, and then ensemble the MI predictions for perpetrator identity to identify MI's modal perpetrator identity prediction across all 100 MI datasets for comparison to ML.

A. Known Perpetrator Classification

Before leveraging our ensemble approach, we first classify our known atrocity-perpetrators separately by training each of our nine candidate algorithms on the training sample referenced above. Based on the results from these training exercises, we predict the known perpetrator identity classes in our (held-out) test sample. We then follow the same steps using our aforementioned MI strategy, and compare each approach's out-of-sample perpetrator predictions to the "true" perpetrator identities in this test sample.

To do so, we follow past work in this arena [20], [21] to extract out-of-sample classification measures of (i) precision, (ii) recall, (iii) f-scores, and (iv) the proportion of all cases correctly classified (CCR) for each ML algorithm used, as well as for our MI approach. Precision measures the correct predictions of a given class (e.g., perpetrator identity = "state") as a share of an algorithm's total (correct and incorrect) predictions for that same class. Recall instead quantifies the proportion of a given class' (i.e., perpetrator identity's) cases that were classified as such by an algorithm. F-scores provide a weighted average of precision and recall, with higher values reflecting better overall accuracy for a given algorithm. CCR reports the proportion of all cases—*across all classes*—that were correctly classified by a given algorithm, and thus places relatively more weight on accurate classification of one's more common classes. We report these out-of-sample classification statistics—averaged across all seven perpetrator identity classes for precision, recall, and f-scores—in Table I.

Table I offers several revealing insights. First, in terms of precision, recall, and f-scores, none of our classifiers do a particularly commensurate job in classifying *all* categories

[5]For example, the `Intent` variable coded each atrocity for its intent based on an six category categorical string-indicator. For each categorical string-value, we omitted spaces so that each was considered as a single unigram, such that, "Intent Apparent But Not Stated" was converted to "Intent-Apparent-But-Not-Stated."

[6]Omitting text-based features that exhibit 50+ unique values (factors).

Table I
OUT-OF-SAMPLE ALGORITHM PERFORMANCE

Approach	Precision	Recall	F-score	CCR
LASSO	0.484	0.321	0.346	0.868
SLDA	0.393	0.373	0.380	0.853
SVM	0.444	0.339	0.363	0.877
Maximum Entropy	0.379	0.351	0.360	0.865
Bagging	0.419	0.306	0.331	0.861
Boosting	0.381	0.309	0.327	0.853
Random Forests	0.384	0.289	0.311	0.871
Classification Trees	0.359	0.257	0.280	0.810
Neural Network	0.230	0.246	0.237	0.832
Multiple Imputation	0.377	0.187	0.199	0.729

of the perpetrator variable. The recall values in Table I range from a low of 0.187 (MI) to a high of 0.371 (SLDA), implying that our classifiers correctly predict a perpetrator's identity in 19%-37% of the cases, on average across all seven classes—with MI faring worst in these regards. Similarly, the precision values in Table I imply that our ML algorithms' predictions of a given class (i.e., perpetrator identity) are on average correct in 23%-48% of all instances in which that algorithm predicted an observation as having that class. MI performs better than the two worst performing algorithms in this case, but still only places eighth best overall.

Given the highly unbalanced nature of the seven (non-missing) perpetrator identities in our training (and test) dataset, the modest overall performance of our ML and MI classifiers in these instances is unsurprising. Indeed, four of our seven perpetrator classes together encompass fewer than 3% of all cases within our samples and are thus exceptionally difficult to classify. The difficulty in classifying these perpetrators brings down the average levels of precision, recall, and f-scores for all ML and MI classifiers in Table I. By contrast, CCR—which is less sensitive to algorithm underperformance in rare classes—indicates that our ML algorithms correctly classify 81%-88% of all training cases, whereas MI only correctly classifies 73% of these cases.

We therefore also examine our precision, recall, and f-score metrics separately for our seven perpetrator classes. We find that our disaggregated precision, recall, and f-scores each exhibit superior performance across the most prominent classes, relative to the aggregated metrics presented above. For example, if we examine the performance of our classifiers for the three most abundant perpetrator identities ("State," "Non-State," and "Transnational," representing 97% of all data), the average levels of precision, recall, and f-scores across all nine ML classifiers are 0.768, 0.679, and 0.723; each far higher than the comparable levels obtained when we average classifier performance across all seven perpetrator classes (of 0.386, 0.310, and 0.326). By comparison, the equivalent values for MI remain far lower, at 0.545, 0.403, and 0.403. Below, we additionally evaluate the potential further improvements in accuracy that can be obtained by (i) leveraging ensembles of our aforementioned algorithms and/or (ii) classifying a binary, "state" versus "non-state" perpetrator variable, rather than the seven category perpetrator variable examined here. Before turning to these extensions, we first give more attention to the relative strengths of our classifiers for the primary seven-category perpetrator identity variable.

The f-scores in Table I provide the most direct means of comparing the relative strengths of each algorithm in these regards. Here, we find that six of our ML algorithms—LASSO, SLDA, SVM, maximum entropy, bagging, and boosting—noticeably outperform the remaining three algorithms—random forests, classification trees, and neural networks, as well as MI. This is evidenced by the noticeable drop-off between the "best" six algorithms (with boosting, the lowest performing of these six, reporting an f-score of 0.327 and CCR of 0.853), relative to the f-scores and CCRs obtained from the three aforementioned "worst" performing algorithms (with f-scores ranging from 0.237-0.311, and CCRs ranging from 0.810-0.871) and MI (f-score=0.199; CCR=0.729). Among the best performing algorithms, SLDA, SVM, and maximum entropy are the strongest performing of all algorithms based on f-scores and CCR. Even so, many of our remaining "best" performing algorithms outperform SVM, maximum entropy, and SLDA in *precision*, most notably LASSO and bagging. SLDA performs best on *recall*, though the remaining five "best" performing ML algorithms exhibit fairly comparable recall.

In sum, the ML algorithms discussed above are able to classify our seven-category perpetrator identity variable with a notable degree of accuracy, and with far higher accuracy than MI. Yet, several algorithms are also more accurate than others for our classification tasks, and each algorithm generally performs much better in classifying our most common perpetrator classes. As such, it is also likely that an ensemble of our ML algorithms will offer further improvements in recall and coverage for our perpetrator identities variable. In this context, an 'ensemble' corresponds to a consensus agreement coding of the predicted classifications that are obtained from two or more of our algorithms. The predicted class assigned by an ensemble accordingly represents the most frequent class predicted across our all algorithms evaluated. If all algorithms disagree for a given case, the algorithm with the highest predicted probability—and its associated predicted class—will be assigned as the ensemble prediction for that case. The use of an ensemble in these respects is in line with comparable political science applications of supervised classification [20], [21], which have found that ensembles substantially improve classification accuracy.

We hence examine the added benefits of ensemble agreement for the classifiers used in our application. To do so, we again leverage CCR, as defined above, and also calculate (i) a measure of total coverage, which corresponds to the percent of cases that meet the ensemble threshold divided by the total cases, and (ii) the number of algorithms that

were in consensus agreement for each case. Together, these metrics allow us to incrementally subset our training sample to only include cases of consensus agreement involving at 2-9 algorithms, and to evaluate overall accuracy (via CCR) and consensus agreement at each of these eight thresholds. These quantities are reported in Table II.

Table II
ENSEMBLE AGREEMENT

N-Ensemble	Coverage	CCR
$n \geq 2$	1.00	0.88
$n \geq 3$	1.00	0.88
$n \geq 4$	1.00	0.88
$n \geq 5$	0.99	0.88
$n \geq 6$	0.94	0.91
$n \geq 7$	0.89	0.92
$n \geq 8$	0.83	0.94
$n \geq 9$	0.72	0.96

We find that the use of 2-through-5 classifier ensemble agreement does not yield a noticeable improvement (in CCR) over using SVM alone, which was the best preforming individual classifier (in CCR). That is, when using SVM, we found that this algorithm correctly classified 88% of all cases, and we find that the cases in our sample that saw at least 2-to-5 ensemble agreement similarly classified roughly 88% of all cases correctly. This lack of improvement in CCR is likely a function of the relatively high CCR obtained under SVM alone, which offers little additional room for improvement. However, as we examine the subsets of our cases that saw ensemble agreement involving six or more algorithms, we begin to observe additional improvement in overall accuracy. For example, the CCR when one uses all nine classifiers is exceptional, with 96% of all cases correctly classified, although in this case the proportion of cases that met this ensemble threshold has declined to 72%. To identify an ensemble that balances coverage against the benefits (to accuracy) of adding additional classifiers, we follow the 90% inter-coder reliability standard that is often invoked in the social sciences [21]. This leads us to choose a six algorithm ensemble, as this ensemble agreement level classifies 94% of our perpetrator cases with 91% accuracy.

B. Robustness Tests

To assess the robustness of our findings, we first reevaluate our results when using four-fold cross-validation. In doing so, we find very similar patterns to those discussed above, and also find that our ML algorithms continue to strongly outperform MI in classification. We then examine the potential improvements in accuracy that are obtained from classifying a binary, "state" versus "non-state" perpetrator variable, rather than the seven-category perpetrator variable examined earlier. Results suggest that our nine algorithms, and ensembles, have noticeably higher accuracy in classifying our binary perpetrator identity variable than either (i) polytomous ML (ensemble) classification or (ii) our application of MI to the binary classification case (which performs even worse than in the polytomous case).

Given that our MI comparison models omitted 15 text-based variable-features that contained $\geq$50 unique factors, we also compared our MI approach to a set of smaller ML specifications that only included the predictors used within our MI models. These underspecified ML algorithms were slightly less accurate than those reported in Table I, but were still far more accurate than MI on relevant metrics. This suggests that the differences in accuracy between MI and ML highlighted earlier are not wholly attributable to the withholding of text variables from the MI model.

C. Unknown Perpetrator Classification

Our main findings suggest that an ensemble of the following six algorithms performs best in classifying our known perpetrator cases: LASSO, SLDA, SVN, maximum entropy, bagging, and boosting. The present section accordingly returns to our full dataset, and divides this dataset into two samples: a training sample corresponding to *all* known perpetrator cases (N=4,870) and a second virgin sample corresponding to all unknown perpetrator cases (N=2,258). We then separately deploy (i) an ensemble of LASSO, SLDA, SVN, maximum entropy, bagging, and boosting and (ii) our MI approach for the task of classifying all "unknown" perpetrator cases in our virgin sample. After classifying the unknown perpetrator cases into our seven known perpetrator classes using the ensemble and MI approaches, we examine each approach's frequencies of assigned perpetrator classes within our virgin "unknown" perpetrator sample. We also consider the frequency of ML ensemble agreements for this virgin sample in Figure 2.

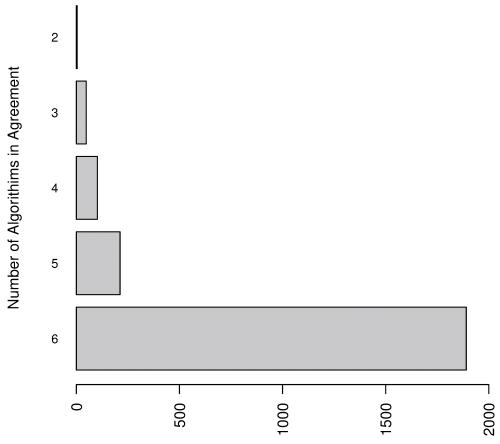

Figure 2. Consensus Agreements (ML Approach)

The virgin classifications are similar in the aggregate

across the MI and ensemble ML approaches, thereby underscoring the utility of *both* approaches in recovering unknown perpetrator identities. For instance, we find that our MI and ML classifications of unknown perpetrators each yield remarkably higher shares of non-state perpetrators than that found in the known perpetrator sample: 72% of all known perpetrator cases in the PITF dataset had the designation of "Non-State, No State Sanction" whereas 93% of all unknown cases were classified as "Non-State, No State Sanction" by our ML approach. In addition, our virgin sample saw only 6% of all cases classified as "State" perpetrators by our ML approach—far lower than the 21% of our known cases recorded as being "State" perpetrators within our known perpetrator sample. These results suggest that the vast majority of unattributed atrocities (1995-2013) were committed by non-state actors. This in turn implies that studies of *non-state* perpetrated atrocities that discard unknown cases may substantially bias their results due to the fact that "unknown" cases are not missing at random. By that same token, studies of *government* perpetrated atrocities that discard unknown atrocities cases may not be substantially biased, given the very small share of unknown cases that correspond to this class of perpetrator.

There are several reasons to view these observed patterns as valid. Based on Figure 2, the vast majority (93%) of virgin cases saw at least five algorithms in consensus agreement, and 84% of all virgin cases saw *all six* algorithms in consensus agreement. When interpreted alongside Table II—where we found that instances of six-algorithm consensus agreement accurately classified 91% of all corresponding cases within our test sample—these levels of agreement suggest that our algorithms are fairly confident, and accurate, in classifying our unknown atrocities cases. Second, we also *do not* find that our out-of-sample classifications of known perpetrator cases yield comparably skewed results towards the "non-state" perpetrator class. Thus, we are unlikely to have classified a high number of unknown perpetrators as "non-state" due to ML and MI severely overpredicting "non-state" cases over "state" cases.

Third, the fact that we find similar patterns (i) across MI and ML and (ii) when applying these same virgin-classification steps to a binary perpetrator variable further underscores the robustness of our overall classification findings. Fourth, extant research suggests that rebel violence against civilians is significantly higher than government violence during our general period of analysis [6], [2]—a pattern that our full PITF dataset now matches much more closely after these virgin classifications are added in. Fifth, previous research indicates that non-state actors are more likely to operate in rural areas, where they disproportionately rely on violence to induce civilian compliance [8]. Given that rural areas are more likely to have higher instances of reporting bias and related information deficiencies [11], [12], it is thus unsurprising that the majority of missing perpetrator cases correspond to rebel perpetrators.

These patterns also sharpen our theoretical understandings of atrocities. Past research suggests that non-state actors frequently resort to civilian killings as a means of offsetting imbalances in power and capacity between themselves and government forces [13], [25], [26]. This implies that the *frequency* of atrocities should be much higher among non-state actors, who are generally weaker than opposing government forces in terms of military mobilization and technological capacity [27]. Our finding that a majority of unknown atrocities are perpetrated by non-state actors lends support to this argument and suggests that this strategy may be more prevalent than previously thought. Likewise, power asymmetries within intrastate conflicts imply that state actors are often likely to control more territory and have better information on enemy collaborators than will non-state groups [7]. In these situations, state forces will often be better able (and hence more likely) to use discriminate violence against specific targets. However, because non-state groups frequently have fewer resources to allocate to collecting information [28], [27] and are likely to operate in territories they do not control [25], they must rely on *indiscriminate* violence to a much greater extent—a pattern confirmed by the disproportionate rates of non-state perpetrated atrocities uncovered by our analysis.

Finally, note that a key justification for the use of violence against civilians by both state and non-state actors is the ability to signal strength and resolve to opponents or other third-party observers [29], [25], [30]. In these situations, perpetrators *want* to have their identity known, and would therefore gladly take responsibility for their actions. Perpetrating atrocities in anonymity, however, does not allow for such credible signals. From this perspective, our finding that the majority of unknown atrocities are actually perpetrated by non-state actors thus lends support to *less* "strategic" perspectives, such as lack of control over one's own troops, or intermittent pressures to secure resources from civilians by violent means [8], [31]. As such, our results suggest that while both state and non-state actors may use atrocities to signal resolve and strength, it is predominately non-state actors that commit "unclaimed" atrocities, possibly for reasons associated with a lack of control or resource (e.g., food) shocks. This, again, is in line with rebels' tendencies towards operating in more rural and/or disputed areas [6] and oft-lower resources and technological capacities relative to government forces [28], [27].

D. External Validation

To verify that the patterns found within our virgin classification exercise—and our ML approach more generally—are accurate, we next compare our ML-adjusted perpetrator estimates to the National Counterterrorism Center's Iraq Georeferenced Worldwide Incident Tracking System (WITS) Data (2004-2010). The WITS dataset is one of the most

comprehensive and detailed publicly available datasets on attacks against civilians perpetrated by non-state groups, and has been used in similar assessments in past research [10]. Further, the WITS data are likely to exhibit higher levels of ground truth than the PITF data given the WITS dataset's coding "from open sources manually using commercial subscription news services, the USG's Open Source Center (OSC), local news websites reported in English, and, as permitted by the linguistic capabilities of the team, local news websites reported in foreign languages" [32]. To guarantee that analysts do not overclassify attacks or code attacks against combatants as attacks against civilians, the WITS uses computer programs to "flag" incidents that might occur due to human error, which allows analysts to update their entries [32].

The Iraq WITS dataset only records attacks perpetrated by non-state actors. Given that our ML algorithms classify the majority of attacks by unknown perpetrators as being carried out by non-state actors, these Iraq WITS data thereby allow us to most appropriately test whether our ML approach over-classifies or correctly-classifies unknown atrocities as non-state atrocities. To ensure that the WITS data match the PITF's atrocities cases as closely as possible, we subset the WITS data to only include incidents where five or more fatalities were reported. We then merge these incidents to those coded by PITF as occurring in Iraq between 2004-2010, which is the temporal range covered by the Iraq WITS dataset. Taken together, these steps allow us to draw comparisons between three different Iraq 'datasets' of non-state perpetrated atrocities during the 2004-2010 period: (1) the Iraq WITS dataset; (2) the PITF dataset with all unknown perpetrator events removed; and (3) a version of the PITF dataset that also adds all additional PITF atrocities cases whose unknown perpetrators were classified as non-state by our ML algorithms to dataset (2).

We then merge datasets (1)-(3) to a monthly-PRIO-GRID [33] for Iraq. This leads our final validation data to contain non-state perpetrated atrocities counts for each 0.5x0.5° grid-cell month in Iraq (2004-2010). Comparing these count data, we find our non-ML corrected PITF events to be correlated with the WITS events at 0.51, whereas our ML-corrected PITF events are instead correlated with the WITS events at 0.91. Next, in dichotomizing each grid-month event count indicator and examining how well each dichotomized PITF measure then classifies our WITS cases, we further find that our ML-adjusted PITF records exhibit superior levels of precision (0.791>0.772), recall (0.461>0.217), f-score (0.582>0.339) and CCR (0.962>0.952) than the unadjusted PITF cases. These comparisons indicate that ML accurately recovers non-state atrocity perpetrators in Iraq.

We then repeat these comparisons using the MI-adjusted PITF data. Within our disaggregated grid-month level framework, the MI-adjusted PITF counts exhibit a comparable WITS-correlation to that of our ML-adjusted counts, and one that is far superior to that of the unadjusted PITF counts. These findings underscore the applicability and appropriateness of MI as a competitive approach to recovering missing perpetrator information in this application, in support of [10]. Nevertheless, we also find that our dichotomized grid-month MI-corrected atrocities cases classify our binary WITS records with moderately lower levels of precision, recall, f-score, and CCR—relative to ML—suggesting that ML remains a preferred method in this case.

V. Conclusion

This paper evaluates whether supervised ML techniques can accurately recover the identities of unknown atrocity perpetrators within datasets of atrocities events. Many theories of atrocities (and tests thereof) hinge on the identity of atrocities-perpetrators [13], [6], [7], [8]. International human rights and legal communities likewise require reliable information on perpetrators for successful prosecution and/or censuring. However, as many as one third of all atrocities incidents in commonly used atrocities datasets lack information on perpetrators' identities. We have shown that ML methods offer one way forward in reliably identifying (unknown) atrocity perpetrators within contemporary atrocities events datasets. Our application to the PITF Worldwide Atrocities Dataset then demonstrates that this proposed strategy can accurately recover the identities of atrocities perpetrators within both out-of-sample and in-sample settings.

Substantively, our findings indicate that a majority of contemporary unclaimed atrocities have been perpetrated by non-state actors. This supports past research [6], [2], while also suggesting that rebel violence against civilians may be even more prevalent than previously thought. Future research should evaluate the causal mechanisms underlying these trends by exploring whether this pattern is attributable to rebel groups' relatively (i) lower levels of discipline and supplies, (ii) stronger incentives to conceal their use of violence against proximate civilians, or (iii) higher tendencies to operate in rural areas prone to media reporting bias. Methodologically, our analysis illuminates an important avenue for future research: the integration of MI and ML. Indeed, fully integrating ML and MI to address missing social science data via approaches such as [34] would likely allow future researchers to better leverage these approaches' relative strengths, while avoiding their respective weaknesses.

Acknowledgment

Bagozzi's contribution is partly based upon work supported by the National Science Foundation under grant no. SBE-SMA-1539302. Koren's contribution was supported by a Jennings Randolph Peace Scholar Award from the USIP.

References

[1] B. Valentino, P. Huth, and D. Balch-Lindsay, "'Draining the sea': mass killing and guerrilla warfare," International Organization, vol. 58, pp. 375-407, 2003.

[2] K. Eck and L. Hultman, "One-sided violence against civilians in war: insights from new fatality data," Journal of Peace Research vol. 44, pp. 233-246, 2007.

[3] C. Raleigh, A. Linke, H. Hegre, and J. Karlsen, "Introducing ACLED: an armed conflict location and event dataset," Journal of Peace Research, vol. 47, pp. 651-660, 2010.

[4] M. C. Croicu and R. Sundberg, "UCDP georeferenced event dataset codebook version 2.0," Department of Peace and Conflict Research, Uppsala University, 2015.

[5] P. A. Schrodt and J. Ulfelder, "Political instability task force worldwide atrocities event data collection codebook version 1.0B2," 2009.

[6] C. Raleigh, "Violence against civilians: a disaggregated analysis," International Interactions, vol. 38, pp. 462-481, 2012.

[7] H. Fjelde and L. Hultman, "Weakening the enemy: a disaggregated study of violence against civilians in africa," Journal of Conflict Resolution, vol. 58, pp. 1230-1257, 2014.

[8] B. E. Bagozzi, O. Koren, and B. Mukherjee, "Droughts, land appropriation, and rebel violence in the developing world," Journal of Politics, vol 79, pp. 1057-1072, 2017.

[9] B. Arva and J. Beieler. 2014. "Dealing with missing data in group-level studies of terrorism." APSA Annual Meeting.

[10] V. Bauer, K. Ruby and R. Pape, "Solving the problem of unattributed political violence," Journal of Conflict Resolution, forthcoming.

[11] C. Davenport and P. Ball, "Views to a kill: exploring the implications of data selection in the case of guatemalan state terror, 1977-1995," Journal of Conflict Resolution, vol. 46, pp. 427-450, 2002.

[12] N. B. Weidmann, "On the accuracy of media-based conflict event data," Journal of Conflict Resolution, vol. 59, pp. 1129-1149, 2015.

[13] R. M. Wood, "From loss to looting? battlefield costs and rebel incentives for violence," International Organization, vol. 68, pp. 979-999, 2014.

[14] M. Krain, "Jaccuse! does naming and shaming perpetrators reduce the severity of genocides or politicides?" International Studies Quarterly, vol 56, pp. 574-589, 2012.

[15] J. H. R. DeMeritt, "International organizations and government killing: does naming and shaming save lives?" International Interactions, vol. 38, pp. 597-621, 2012.

[16] D. Muchlinski, D. Siroky, J. He and M. Kocher, "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data," Political Analysis, vol. 24, pp. 87-103, 2015.

[17] D. W. Hill Jr. and Z. M. Jones, "An empirical evaluation of explanations for state repression," American Political Science Review, vol. 108, pp. 661-687, 2014.

[18] M. Solaimani, S. Salam, A. M. Mustafa, L. Khan, P. T. Brandt, and B. Thuraisingham, "Near real-time atrocity event coding," IEEE Intelligence and Security Informatics (ISI), 2016.

[19] T. G. Dietterich, "Ensemble methods in machine learning," MCS 2000: Multiple Classifier Systems, pp. 1-15, 2000.

[20] L. Collingwood and J. Wilkerson, "Tradeoffs in accuracy and efficiency in supervised learning methods," Journal of Information Technology & Politics, vol. 9, pp. 298-318, 2012.

[21] T. P. Jurka, L. Collingwood, A. E. Boydstun, E. Grossman and W. van Atteveldt, "RTextTools: a supervised learning package for text classification," The R Journal, vol. 5, pp. 6-12, 2013.

[22] G. King and J. Honaker and A. Joseph and K. Scheve, "Analyzing incomplete political science data: an alternative algorithm for multiple imputation," American Political Science Review, vol. 95, pp. 49-69, 2001

[23] J. Honaker and G. King, "What to do about missing values in time-series cross-section data," American Journal of Political Science, vol. 54, pp. 561-581, 2010.

[24] T. Hughes, "Assessing minority party influence on partisan issue sttention in the US House of Representatives, 1989-2012," Party Politics, pp. 1-12, 2016.

[25] L. Hultman, "The power to hurt in civil war: the strategic aim of RENAMO violence," Journal of Southern African Studies, vol. 35, pp. 821-834, 2009.

[26] I. Sánchez-Cuenca, I. and L. De la Calle, "Domestic terrorism: the hidden side of political violence," Annual Review of Political Science, vol. 12, pp. 31-49, 2009.

[27] G. Clayton, "Relative rebel strength and the onset and outcome of civil war mediation," Journal of Peace Research, vol. 50, pp. 609-622, 2013.

[28] R. M. Wood, "Rebel capability and strategic violence against civilians," Journal of Peace Research, vol. 47, pp. 601-614, 2010.

[29] A. H. Kydd and B. F. Walter, "The strategies of terrorism," International Security, vol. 31, pp. 49-80, 2006.

[30] J. R. Hollyer and B. P. Rosendorff, "Why do authoritarian regimes sign the convention against torture? signaling, domestic politics and non-compliance," Quarterly Journal of Political Science, vol. 6, pp. 275-327, 2011.

[31] O. Koren and B. E. Bagozzi, "Living off the land: The connection between cropland, food security, and violence against civilians," Journal of Peace Research., vol 54, pp. 351-364. 2017.

[32] J. Wigle, "Introducing the worldwide incidents tracking system (WITS)," Perspectives on Terrorism, vol. 4, 2010.

[33] A. F. Tollefsen, H. Strand, and H. Buhaug, "PRIO-GRID a unified spatial data structure," Journal of Peace Research, vol. 49, pp. 363-374, 2012.

[34] Y. Deng, C. Chang, M. S. Ido and Q. Long, "Multiple imputation for general missing data patterns in the presence of high-dimensional data" Scientific Reports, vol. 6, pp, 2016.