Forecasting civil wars

Jorge Rodríguez Peña
MSc Data Science, INM430
City, University of London
London, United Kingdom

***Abstract*—Conflict forecasting is an emerging field in conflict science. This work aims to dive into the field by using monthly aggregated event data from the ICEWS dataset. Sectors were classified as Government, Opposition, Insurgents and People. The results showed that escalation and relaxation of the conflict could be determined from the data. The model could explain the interaction between sectors at the beginning of a civil war, with violent interaction between Government, People and Insurgents and Opposition taking a more passive role with verbal confrontation. The model could predict the start of a civil war (AUC 0.7) but is worse at determining the ending (AUC 0.6). Predictions for February 2020 were also estimated and compared with registered critical events.**

***Index Terms*—conflict, civil wars, event data, machine learning, forecasting, ICEWS, PITF**

I. Introduction

History is most of the time conceived as a retrospective field: with the information available today, descriptions of the past can be formed. This is a misconception, since it obviates that experiences can help understand the present and future experiences. That is the case of conflict science. Conflict science studies conflict in all its forms: World Wars, mobs, riots, coups, etc. In the recent years, forecasting has gained importance in conflict science becoming its own field [1].

It is sometimes stated that the theory should be the base in conflict forecasting theory [2]. However, recent studies prove that theoretical models are imprecise in it forecasts [3]. Moreover, the elements studied in theory, like per capita income and natural resources, change slowly to make predictions out of them.

With the arrival of big data, some models have proven to be successful without considering theory and simply focusing on event data, particularly in civil war forecast [4]. Internal conflict can be predicted using internal events of the country and the interactions between its sectors.

Civil wars cost thousands of lives and force people to run away from their home countries. The Chadian Civil War of 2005 to 2010 costed about 7000 lives and 200000 refugees [5]. A powerful civil war forecast model can be useful both at anticipating conflict and to extract theory from the data. Early detection of conflict escalation, combined with intervention and mediation, can help avoid it or end it sooner by detecting key sectors and events.

II. Analytical questions and data

A. Analytical questions

The aim of this work is to generate a civil war forecasting model from event data. The model will be analysed to determine the most relevant predictors and to extract useful information from them.

The model must determine:

- What actors are relevant at the beginning, during and at the end of the conflict.
- What actions taken by the actors are relevant
- How the escalation of the conflict takes place

The model must generate predictions. Two types of predictions will be made:

- To tell if a country will start a civil war in the next month or not.
- To tell if a country currently in a civil war will end its civil war the next month or not.

B. About the data

The Integrated Crisis Early Warning System (ICEWS) is a repository that contains counts of daily events worldwide. The ICEWS dataset records the "who did what to whom and where" for events across the globe. The data is stored in the Harvard Dataverse [6]. Events from 1995 to 2018 will be used.

The ICEWS dataset contains two event coding columns: CAMEO and Intensty. The Conflict and Mediation Event Observations code (CAMEO code) is a coding system for event data. There are 20 different codes for 20 different types of events and each has different levels of sub-coding. The intensity value is obtained from the CAMEO code. It ranges from -10 to 10, with -10 the most hostile events and 10 the most cooperative events.

The State Failure Problem dataset, from the Political Instability Task Force (PITF), will be used to stablish the start and ending of a civil war. It contains civil wars from 1955 to 2018. The PITF collects data of political conflicts and state failure. The dataset can be found in the Center of Systemic Peace website [7].

III. Analysis

A. Data preparation

To obtain the final dataset, the following steps will be performed:

1) Select only internal events (i.e., events with the same Source and Target Country).

2) Map each source and target to four sectors: Government, Opposition, Insurgents and People. The mapping was performed with the
3) help ICEWS Dictionaries [6]. Events not mapped were removed.
4) Generate dummy variables to code source-target interaction, with 1 if the event occurred between the specified source-target pair and 0 otherwise.
5) Add ISO 3166-1 alpha-3 country code to avoid double counting a country (for example, North Korea and Republic of Korea).

This way the answer for "who (source) did what (CAMEO code/Intensity) to whom (target) where (ISO3) and when (Year-Month)" is available in the data.

After cleaning the data, between 75% and 80% of the total events were removed for each year, keeping 3.5 millions of events. Figure 1 shows how numbers of events registered evolve in time. Although some temporal differences are observed, no month is an outlier in the number of events, so all year-month aggregations will be used.

Data from the PITF was simplified to contain only the ISO3 country code and the start and ending date of the conflict.

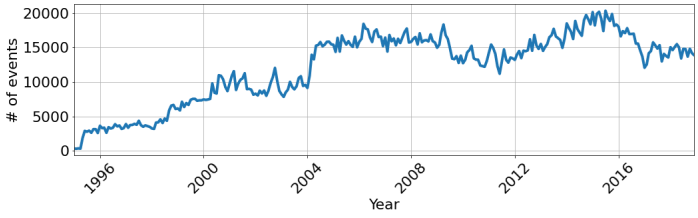

Fig. 1. Evolution of the number of events in the dataset in time.

B. Data derivation

Country-month counts of Source-Target-CAMEO triads will be used for the model. For example

Government-People-01: 15

means that, during that month, the Government made 15 public statements (CAMEO code 01) directed to People.

To obtain the results, the dataset will be aggregated by country and month. Only the first two numbers of the code will be used, so that our model sticks to the basic 20 CAMEO codes. A column with the total intensity was added.

This model tries to explain what CAMEO codes are more relevant to each type of monthly aggregation (Periods of peace, start, ongoingness and ending of a civil war) with the number of events per Source-Target-CAMEO triad. This way, the "who(source) does what(CAMEO code) to whom(target)" is present in the data and theoretical results can be extracted from them.

Some countries did not contain events for certain months. These missing monthly aggregations were considered missing and replaced by 0, considering that no events were registered.

As for outliers, the model should handle noisy data, since not all civil wars are equal. Moreover, removing outliers can remove certain aggregations where a civil war is taking place due to the imbalance of the data.

Civil wars were added merging the ICEWS and the PITF datasets. Values for periods of peace starting, ongoing and ending civil wars were coded into the dataset and shifted up by a month. That way, each monthly aggregation contains the outcome of the following month and thus establishes the predictor.

C. Construction of models

Two models will be trained, each with a different objective.

The first model will aim to predict the start of a conflict in the absence of one. This model can be useful to avoid conflict escalation in countries currently in peace, using its forecast probabilities and extracting key predictors that might help mitigate it.

The second model will aim to predict the end of a conflict when one is ongoing. This model can be useful to mitigate conflict by looking at key predictors, so that key sectors are identified to intervene.

Random forests will be used to make predictions. The choice of random forest relies on in their properties. Firstly, random forests select the predictor with information gain, so variables that do not improve the model will be removed. This will help in case the selected predictors in the previous step are not relevant to the model. Secondly, by increasing the number of trees in the forest it is possible to adequate data imbalance, which is the case for this work. Thirdly, random forests are known to handle noisy data better than other models. For this work, this is also an advantage, considering the huge amount of data the model will be handling.

Both random forests will be tuned with 10.000 trees to reduce the importance of the noise in the model and a maximum depth of 4 to avoid over-fitting. This is done to prevent over-fitting.

D. Validation of results

To identify relevant actors and events that generate a civil war or end it, t-test will be performed to determine if the differences in a predictor for each case are statistically significant. The threshold will be set at 0.05, meaning that any p-value of the t-test above 0.05 will not be considered statistically significant.

The discovery of relevant variances inside predictors will provide answers to the theoretical question of "who does what to whom and when" to start a civil war or to put it to an end. This completely inverts the process: rather than using theory to make predictions, how the predictions are calculated will be used to extract theory. Predictor importance for each model will be used to further deep into the theory.

Validation of the models will be done using the Receiving Operator Characteristic Area Under the Curve (ROC-AUC). The goal is to predict probabilities of a civil war starting or ending. For a biased dataset as this one, the accuracy of the

model can be excessive if it only predicts the majority class. The AUC uses the prediction score instead of the predicted class and therefore is better for the purposes of this work. Each model will be validated using cross-validation, out of bag samples and a test set.

IV. Experimental results

A. Findings and reflections

Escalation of the conflict is observed on the model (see Figure 3), with the distributions of the mean intensity evolving with the conflict. The month prior to the civil war, assaults from people and insurgents (CAMEO code 18) towards government were answered with repression (CAMEO codes 18 and 19). The opposition tends to confront the government with verbal attacks and criticism (CAMEO codes 01 and 11). The Government and Insurgents are also less likely to negotiate (CAMEO codes 02, 03 and 10), suggesting the evolution of verbal conflict to material conflict. This simplified interaction is described in Figure 2.

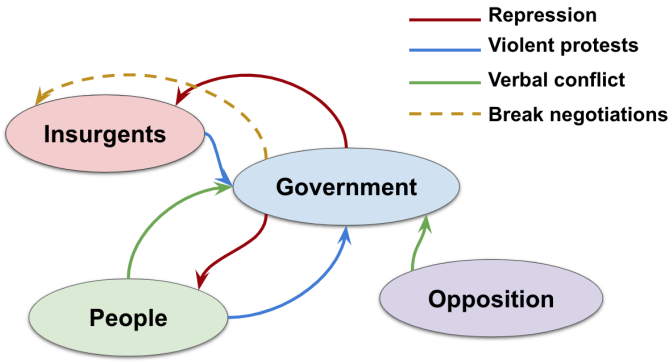

Fig. 2. Simplified interactions between sectors at the beginning of a civil war.

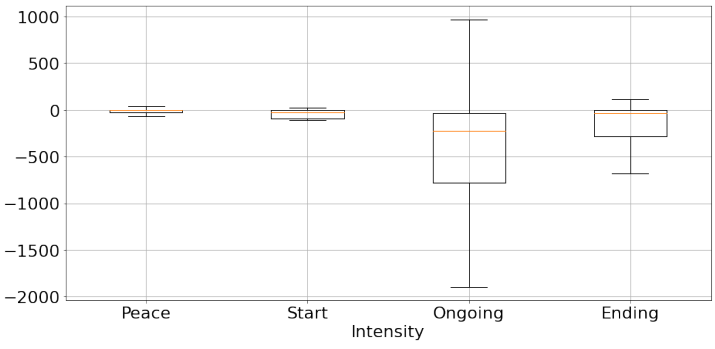

Fig. 3. Distribution of the Intensity for each period of time.

The opposite occurs at the end of a civil war. The total intensity decreases, suggesting the relaxation of the conflict. Criticism and disapproval are less common (CAMEO codes 01 and 11), the assaults and fights decrease (CAMEO codes 18 and 19), suggesting the wear of the conflict, as do the number of demands. The latter can mean that sectors are less likely to demand due to conflict relaxation or that cooperation, instead of demands, is needed to put the conflict to an end.

The AUCs obtained for the models are in Table I. AUCs for predicting the start of a civil war are above 0.5 and over 0.7, showing little variance. This means the model can tell the differences between periods of peace and civil wars and, therefore, can produce predictions. The second model has a worse performance. The AUCs are close to 0.6 and show higher variance. The ROCs shown in Figures 4 and 5 where it is straightforward that the second model is proximate to the random model.

TABLE I
AUCs for each model

Models	Validation methods		
	5-fold CV	*Out of bag samples*	*Test set*
Start forecaster	0.77	0.72	0.72
Start forecaster	0.64	0.58	0.66

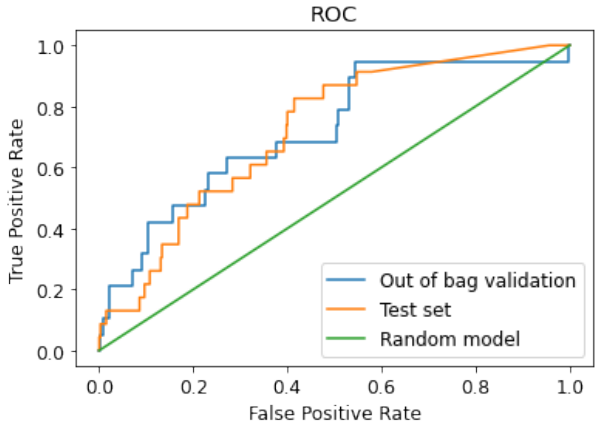

Fig. 4. ROCs for the first model.

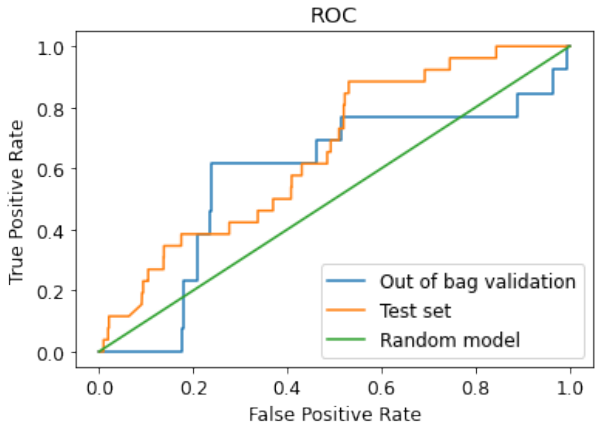

Fig. 5. ROCs for the second model.

The model was used to predict the chances of a civil war starting in Feabruary 2020. The results for the top 5

probabilities are listed below, with critical events that took place in February 2020:

- India (0.15%): Riots in Delhi ended up with 53 deaths [8].
- Philippines (0.019%): Opposition senators were indicted with "conspiracy to commit sedition" [9].
- Iraq (0.015%): An anti-government protest camp was raided and 8 people were shot dead [10].
- Lebanon (0.012%): A new cabinet formed after months without one and with the October Revolution resuming [11].
- Hong Kong (0.004%): The Anti-Extradition Law Amendment Bill Movement are ongoing in January. However it was paused on February due to the pandemic. The model was not trained to detect this [12].

B. Further work

Simplifying the interactions as it was done in Figure 2 can lead to a simpler model with higher performance at predicting the start of a civil war. Moreover, it can be implemented to build a network that explains relations and interactions between sectors. This network can be used to construct theoretical models of intra-state interaction.

Increasing the number of sectors or decreasing them can improve the model or give more relevant information. Military is sometimes involved in civil wars and coups, so maybe considering military and security forces as an extra sector that may or may not depend on the government can be helpful.

However, the current model can determine key interactions that generate conflict, and these can be modelled into theory (as shown in Figure 2). It can also make predictions to determine the start of a civil war with accurate results.

WORD COUNTS

- Abstract: 123 words
- Introduction: 244 words
- Analytical questions and data: 297 words
  - Analytical questions: 119 words
  - About the data: 178 words
- Analysis: 884 words
  - Data preparation: 187 words
  - Data derivation: 246 words
  - Construction of models: 237 words
  - Validation of results: 214 words
- Experimental results: 531 words
  - Findings and reflections: 388 words
  - Further work: 143 words

REFERENCES

[1] Van Holt, Tracy & Johnson, Jeffrey & Moates, Shiloh & Carley, Kathleen. (2016). The Role of Datasets on Scientific Influence within Conflict Research. PloS one. 11. e0154148. 10.1371/journal.pone.0154148.

[2] Halvard Buhaug, Lars-Erik Cederman, Kristian Skrede Gleditsch, Square Pegs in Round Holes: Inequalities, Grievances, and Civil War, International Studies Quarterly, Volume 58, Issue 2, June 2014, Pages 418–431, https://doi.org/10.1111/isqu.12068

[3] Ward, Michael & Greenhill, Brian & Bakke, Kristin. (2010). The perils of policy by p-value: Predicting civil conflicts. Journal of Peace Research. 47. 363-375. 10.1177/0022343309356491.

[4] Blair, Robert & Sambanis, Nicholas. (2020). Forecasting Civil Wars: Theory and Structure in an Age of "Big Data" and Machine Learning. Journal of Conflict Resolution. 64. 002200272091892. 10.1177/0022002720918923.

[5] Vicenç Fisas. Anuario 2009 de procesos de paz Archived 3 March 2016 at the Wayback Machine. Barcelona: Icaria Editorial, pp. 75. ISBN 978-84-9888-076-2

[6] Harvard Dataverse. 2020. ICEWS Coded Event Data. [online] Available at: https://doi.org/10.7910/DVN/28075 [Accessed 19 December 2020]

[7] Systemicpeace.org. 2020. INSCR Data Page. [online] Available at: http://www.systemicpeace.org/inscrdata.html [Accessed 19 December 2020].

[8] en.wikipedia.org. 2020. 2020 Delhi Riots. [online] Available at: https://en.wikipedia.org/wiki/2020_Delhi_riots [Accessed 20 December 2020].

[9] eGMA News Online. 2020. QC Court Issues Arrest Orders Vs. Trillanes, 9 Others. [online] Available at: https://www.gmanetwork.com/news/news/nation/725992/qc-court-issues-arrest-warrant-vs-trillanes-9-others/story/?just_in [Accessed 20 December 2020].

[10] France 24. 2020. Seven Killed As Rival Protesters Clash In Iraq's Najaf. [online] Available at: https://www.france24.com/en/20200205-seven-killed-as-rival-protesters-clash-in-iraq-s-najaf [Accessed 20 December 2020].

[11] en.wikipedia.org. 2020. 2019–20 Lebanese Protests. [online] Available at: https://en.wikipedia.org/wiki/2019%E2%80%9320_Lebanese_protests#Protests_resume [Accessed 20 December 2020].

[12] 3n.wikipedia.org. 2020. 2019–20 Hong Kong Protests. [online] Available at: https://en.wikipedia.org/wiki/2019%E2%80%9320_Hong_Kong_protests#COVID-19_crisis [Accessed 20 December 2020].