Explainable Tracking of Political Violence Using the Tsetlin Machine

Hsu-Chiang Hsu
Dept. of Informatics & Network Systems
University of Pittsburgh
hsh40@pitt.edu

Bree Bang-Jensen
Dept. of Political Science
University of Pittsburgh
bree.bang.jensen@gmail.com

Michael P. Colaresi
Dept. of Political Science
University of Pittsburgh
mcolaresi@pitt.edu

Panos K. Chrysanthis
Dept. of Computer Science
University of Pittsburgh
panos@cs.pitt.edu

Vladimir I. Zadorozhny
Dept. of Informatics & Network Systems
University of Pittsburgh
viz@pitt.edu

2023 International Symposium on the Tsetlin Machine (ISTM) | 979-8-3503-4477-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISTM58889.2023.10454920

***Abstract*—In this paper we introduce a framework that utilizes an architecture based on the Tsetlin Machine to output explainable rules for the prediction of political violence. The framework includes a data processing pipeline, modeling architecture, and visualization tools for early warning about notable events. We conducted an experimental study to explain and predict a one of the most notable events, - a civil war. We observed that the rules that we produced are consistent with theories that emphasize the continuing risks that accumulate from a history of conflict as well as the stickiness of civil war.**

***Index Terms*—tracking political violence, explainable prediction system, data streaming, data warehousing, logical interpretable learning**

I. Introduction

Through a qualitative understanding of conflict dynamics, human rights violations and protections, and ethnic politics and relations, policy analysts and conflict researchers build theoretical and qualitative models of the underlying grievances and alliances that structure war and peace [1]. For example, the violence in Yemen over the last several decades is explained by experts as emerging from the treatment of Zaydis, who are part of the Shia branch of Islam, by the Sunni government (see "Government of Yemen (North Yemen) - Ansarallah" Uppsala Conflict Data Program, 2023, https://ucdp.uu.se/statebased/10855). However, to date, there has been no system that directly allowed policy makers and researchers to both inform their qualitative perspectives of these grievances and policies with quantitative data at a high spatial and conceptual resolution, as well as share the resulting maps of these crucial patterns that guide decision-making, theorizing, and predictions with others.[1]

[1]The closest system is supplied by the Ethnic Power Relations (EPR) https://icr.ethz.ch/data/epr/ project which, when compared to the pipeline described below, does not operate on groups of allegations of human rights violations or reports of protections.

In our prior work, we developed a *Human Rights Violation Exploration, Analytics, and Warning* system (HR-VEAW) [2] that allows users to visualize the rich spatial and conceptual information relevant to making sense of both the escalation of instability, as well as to how negotiators might wind down tensions, and with whom. We process textual data from human rights reports and other social media that communicates both historical and contemporaneous information on who is alleged to have violated or protected a broad array of rights and behaviors for specific groups. This enables us to not only look at changes in patterns of violations and protections in aggregate over time or across both space and time, but fundamentally explore which groups are being targeted or privileged by the government and other actors and on what specific dimensions.

HR-VEAW implements a scalable information processing pipeline combining traditional database technologies with data streaming, NLP and sentiment-aspect representations, and data visualization (Figure 1). Because of the sensitive and important context of the political information flowing through HR-VEAW, it is crucial that policy makers and researchers understand why the system produces predictions that signal warnings (or not). Additionally, it is crucial to maximize both the usefulness of the warning signals as well as the explainability of the system as either a lower performing system or an un-trusted black-box will undercut the potential positive impact of this system.

In this paper we report on extending HR-VEAW with interpretable and explainable ML based on the Tseitlin Machine [3], [4], which helps users to explain social instability and conflicts and guides decision-making, theorizing, and predictions. In general, the literature has some ambiguity in definitions of interpretability and explainability. For example, a model is considered "interpretable" in the sense of we understand why the model fit the parameters it did, or "explainable" in the sense of we understand why the model created the predictions it did. From this point of view, the Tsetlin Machine is both interpretable and explainable (as we further discuss in Section

IV). In this paper we will be using those terms interchangeably.

The paper is organized as follows. After presenting the overall architecture of HR-VEAW system in the next section, we elaborate on the TM logical interpretable learning in Section III. Section V explains the data transformations in the data streaming pipeline to binarize the source data for further data interpretation with Tsetlin Machine. Section VI provide an experimental study and Section VII concludes.

II. System Overview

An overview of the data flow of the HR-VEAW system is illustrated in Figure 1. As it can be seen, there are two phases in the pipeline. The first phase is *Data Acquisition*, in which data from different sources are continuously transformed by the ETL (Extract, Transform, Load) processes and stored into a data warehouse. The second phase is *Data Analysis*, in which data visualization and interpretation are used to explore the data and to discover patterns of social instability for early warnings.

A. Data Acquisition

This phase consists of a variety of ETL processes for different data sources. Human rights reports, news, and tweets, etc. go through the human rights text parser PULSAR [5] so that structured information such as victims and human rights aspects can be extracted. Social and economic statistics data from the ViEWS project [6] is extracted by specialized queries. Geographic information for the regions appearing in the human rights reports are retrieved from the online geographic database OpenStreetMap (OSM) [7]. All extracted and transformed data are then stored in a data warehouse.

The whole data acquisition phase is structured along the producer/consumer paradigm and topic channels, powered by Apache Kafka [8]. There are two major advantages of using Kafka in HR-VEAW. First, Kafka can process high throughput data streams from different sources and can be easily integrated with the ETL processes as an event driven message bus. Furthermore, it is also highly durable. While the data warehouse only stores the structured data after the ETL processes commits it, Kafka can persist data/messages during the whole data acquisition phase, so that messages are never lost and can be retrieved at a later time as needed.

B. Data Analysis

In this phase, different types of data from the data warehouse could be generated at different aggregation levels and forms suitable for the data interpretation and data visualization modules.

The data visualization module is designed to help the data analyst explore and understand the data. It allows the user to visualize human right conditions across different dimensions such as region, time, victim and human rights aspect. The interactive visualization can help the user discover interesting patterns in the data, such as inequalities across dimensions.

The data interpretation module aims to find the logical relationship between different features in the data. The final module of HR-VEAW provides early warning of future conflicts based on the data analysis results from the data interpretation and data visualization modules.

In this paper we focus on the data interpretation module. One way to find the logical relationship within the data is to booleanize some of the aggregated information from the data warehouse to get the corresponding binary indicators. These binary indicators can serve as learning features in machine learning methods, in particular in the Tsetlin Machine (TM) [3], [4]. to conduct logical interpretable learning to discover relationship, pattern, and rules in the data. Next, we elaborate on the process of the logical interpretable learning with TM.

III. Interpretable Learning with the Tsetlin Machine

The TM implements advanced logical interpretable learning that aims at discovering (learning) logical clauses from an arbitrary truth table. We assume that rows of the truth table are produced one by one from an incoming data stream. As a new row arrives, the learner tries to guess the clauses by deciding on whether to include certain literals in a clause. Intuitively, the more incoming rows the learner explores, the more accurate its guesses should be.

The TM decomposes problems into self-contained patterns that are expressed in propositional logic. The patterns are based on disjunctive normal form. Thus, the patterns that the TM builds are interpretable, similar to the branches of a decision tree [9]. In our explorations below, we work with domain experts to probe the usefulness of the discovered patterns.

More formally, a basic TM takes a vector $X = (x_1, \ldots, x_o)$ of binary features as input, to be classified into one of two classes, $y = 0$ or $y = 1$. Together with their negated counterparts, $\bar{x}_k = \neg x_k = 1 - x_k$, the features form a literal set $L = \{x_1, \ldots, x_o, \bar{x}_1, \ldots, \bar{x}_o\}$. A TM pattern is formulated as a conjunctive clause C_j, formed by ANDing a subset $L_j \subseteq L$ of the literal set.

The number of clauses is a parameter. Half of the clauses are assigned positive polarity, half are of negative polarity. During the classification the positive clauses vote for a positive output ($y = 1$) and the negative clauses vote for a negative output ($y = 0$). The classification decision is based on a majority vote. In case of a tie, the decision can be randomized, default, or based on prior knowledge of the contexts.

Each literal in a clause is associated with a Tsetlin Automaton (TA) [10], which decides whether to *exclude* or *include* that literal in the clause. A state number of TA impacts the probability of inclusion or exclusion of the literal in a clause: the higher the state number, the higher the probability of the literal inclusion. The TA performs state transitions under different conditions and a feedback mechanism is designed to decide which literals to include/exclude. Figure 2 shows the feedback matrix, explaining how the states (i.e., inclusion counters) are updated for each literal in a clause of positive and negative polarity. For example, it may suggest including a literal in a positive clause by incrementing its state if the

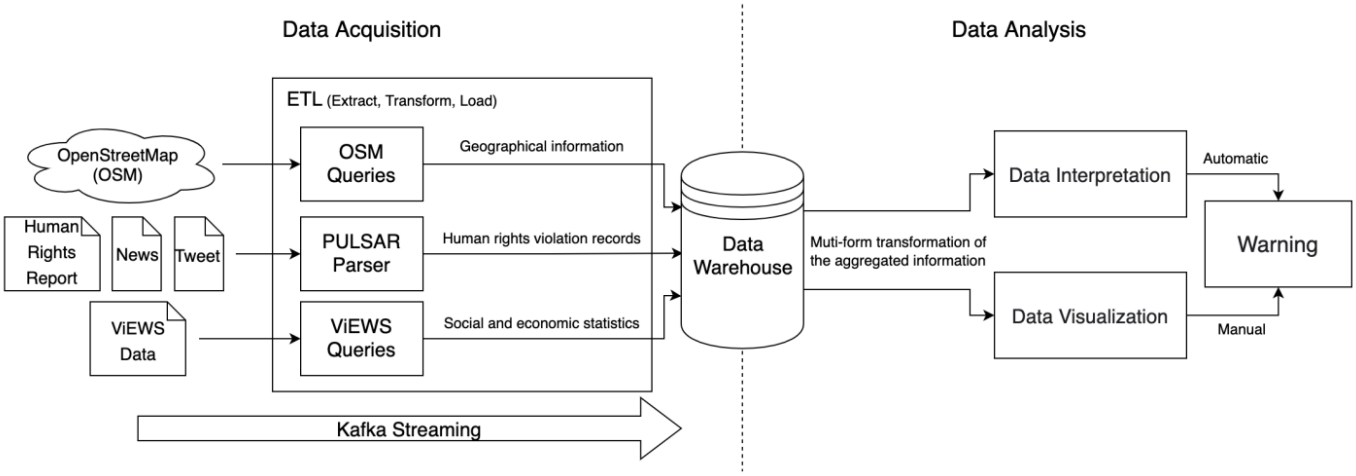

Fig. 1. Overall System Overview

value of the literal is 1 and the output is also 1 (i.e., the literal "agrees" with a positive output). Type I feedback is designed to produce frequent patterns, while Type II feedback increases the discrimination power of the patterns, combating false positives. At the end of the input data stream, the final conjunctive clause will likely to include literals with the largest state numbers of their corresponding TAs.

Thus, the process of TM learning accumulates pieces of evidence from the incoming data stream in order to generate most likely clauses explaining the process behind the data stream. See [3] for further details.

IV. Related Work

Methods of eXplainable AI (XAI) may differ based on methodology and usage [11], [12]. From the methodology point of view, explanation can be either backpropagation-based or perturbation-based [13]. Backpropagation-based methods utilize partial derivatives of the activations (backpropagating gradients) to explain the obtained learning results [14]. Perturbation-based methods explore various combinations (perturbations) of input features, providing explanation in one (forward) pass without the backpropagating gradients [12].

From the usage perspective, an explainable method can be either embedded in the model itself (model-intrinsic explanation) or applied as an external algorithm for explanation (model-agnostic post-hoc explanation), e.g., where the predictions of an already existing well-performing neural network model can be explained by an external algorithm [11].

The TM naturally implements an efficient perturbation-based and model-intrinsic explainable learning. Note, that taking the perspective of the policy-maker, explainability also has several distinct and at time opaque meanings. For example, in the domain of conflict forecasting, foreign affairs officers want to know *why* a specific unit, which could be a country, other administrative region, or a precise "grid-cell" of space defined within a lattice following lines of latitude and longitude, is predicted to be a higher risk than either that same unit in the past, or another unit at a similar point in the future. The answer to this question takes the form of propositions that identify which input features moving in a particular way (eg lower GDP growth, higher protest values, or an election occurring in the presence of previous ethnic grievances), dependent on the computed parameters/weights/rules and their composition, produce that difference. The complexity of these functions, from the point of view of human understanding, can make the accurate identification of these propositions difficult or impossible [15].

Another related but distinct question asked by those making decisions about the distribution of aid or peacekeepers is *what* process led the specific forecasting system or pipeline to make the predictions that were produced. The set of answers to this question do not condition on the computed model representation, but instead ask what led to this specific computed model (eg that had a high weight for lower GDP growth, etc) as compared to a different representation of the problem, in the first place. Here we largely follow [15] in looking at the explainability of predictions for units within given systems. For systems with this level of explainability, we can also compare their inferences across model representations as a second step.[2]

In a useful review, Marcinkevics and Vogt [16] provide several reasons when and why explainable[3] models can be useful, even if they sacrifice some quantitative performance as measured by a specific score or loss. In particular, they cite Doshi-Velez and Kim [17] who argue that interpretable models are the most useful in settings where the true underlying model has not yet been discovered. This is clearly the case in most or all social science problem areas, and particularly in the case of predicting and forecasting political violence.

[2]Another way of putting this is that if a model itself is not explainable in terms of why predictions differ across cases, either real or synthetic, it will be difficult or impossible to explain how it produced those specific predictions versus a different — perhaps even less explainable — system.

[3]The authors discuss both interpretable and explainable models but the discussion below applies across multiple definitions of these concepts.

polarity	*output*	*clause*	*literal*	*FEEDBACK*	
+	0	0	0		
+	0	0	1		
+	0	1	0	+1	Type II
+	0	1	1		
+	1	0	0	-1	Type Ib
+	1	0	1	-1	Type Ib
+	1	1	0	-1	Type Ib
+	1	1	1	+1	Type Ia
-	0	0	0	-1	Type Ib
	0	0	1	-1	Type Ib
-	0	1	0	-1	Type Ib
-	0	1	1	+1	Type Ia
-	1	0	0		
-	1	0	1		
-	1	1	0	+1	Type II
-	1	1	1		

Objectives: Encourage memorizing literal if literal is 1 and clause is 1 and output matches polarity
Discourage memorizing literal if either literal or clause is 0 and output matches polarity
Encourage memorizing literal if literal is 0 and clause is 1 and output does not match polarity

Positive
Type Ia: if output=1 & clause=1 & lit =1, then state increment (+1)
Type Ib: if output=1 & clause=0 or lit =0, then state decrement (-1)
Type II: if output=0 & clause=1 & lit=0, then state increment (+1)

Negative:
Type Ia: if output=0 & clause=1 & lit =1, then state increment (+1)
Type Ib: if output=0 & clause=0 or lit =0, then state decrement (-1)
Type II: if output=1 & clause=1 & lit=0, then state increment (+1)

Fig. 2. Explanation of the TM feedback in the process of logical interpretable learning.

In these ”open-M” settings, interpretability allows others to build on the partial insights of the given model, rather than taking the predictions as is. Similarly, in these ”complicated” settings – where the state of technology does not allow for the optimization of benefits and the minimization of harms to be simultaneously internalized by the model in a valid way – explainability allows qualitative knowledge outside of the model to guide ethical and safe decisions that are partially, but not deterministically based on the model's inferences.

Therefore, despite the definitional imprecision in the literature, the computed output of the Tsetlin Machine can be appreciated as explainable in the most basic sense – whereby the model outputs parameters that are themselves explanations/sets of rules for why units of interest have distinct predictions over time and/or why two cases have distinct predictions in a given period. Further, the Tsetlin Machine outputs rules that describe predictions both for positive and negative classes.

V. Explainable Data Streaming

As we mentioned in Section II the data interpretation module of HR-VEAW explores logical relationship between different features in the data. The data stream is continuous and the process of data interpretation is also continuously providing early warning of future conflicts. Before entering the data interpretation module, the data stream undergoes several notable transformations, starting from a non structured and often textual data representation. Next, HR-VEAW performs application-driven data aggregation followed by various boolinizations of the aggregated information to get important binary indicators. These binary indicators are used as input and output variables in the Tsetlin Machine learning.

More specifically, we base our analysis of multiple features (variables) that we obtain from the HR-VEAW stream (e.g., number of prior civil wars for countries at different years in different regions, duration of those wars, etc.). For each of those variables we compute summary statistics at the level of the cohort (e.g., mean, median, min, max, etc.) The cohort is formed by the historic data that has already been streamed through HR-VEAW. The expectation is that different countries from different regions and at different time periods behave differently with respect to those features.

We are binarizing the data features/variables for each country/year by comparing the statistics for the country/year with

CivilWar	Year	Country	prior_war	duration	region	global_mean	duration_mean	WLastYear	Bmean	Dmean
1	2006	Afghanistan	1	1	Neast	0.651852	4.383459	1	1	0
1	2007	Afghanistan	2	2	Neast	0.742647	4.383459	1	1	0
1	2008	Afghanistan	3	3	Neast	0.838235	4.383459	1	1	0
1	2009	Afghanistan	4	4	Neast	0.933824	4.383459	1	1	0
1	2010	Afghanistan	5	5	Neast	1.022059	4.383459	1	1	1
...	...	...	...	...	...	...	...	...	...	...

Fig. 3. Explaining data preparation

the statistics for the cohort, and use those binarized features to train the Tsetlin Machine. The positive (e.g., in favor of a civil war) and negative (e.g., against civil war) rules produced by the TM are further used for assessing incoming data stream (raising an early warning flag).

Studies of political violence have identified a number of other variables which may influence civil war risk, which we monitor and utilize in a similar way. These include *per capita income, population, percent mountainous terrain, export revenue from oil, new state, whether a state is contiguous, democracy index, ethnic fractionalization, religious fractionalization, number of human rights violations conducted by various perpetrators toward certain ethnic groups, etc*. In the next section we consider an experimental study illustrating this approach.

VI. Experiments

A. Data Sources

The following experiments draw on [18]'s article on civil war prediction, which is of particular interest because it was part of a lively debate in *The Journal of Conflict Resolution* on the role of theory versus big data and machine learning in predicting civil war onset (see also [19], [20]. [20] argue that a model which focuses on procedural factors, such as state response to protest demands, outperforms models involving more predictors, including structural factors such as inequality and resource endowment. Thus, this model and data set offer an ideal benchmark for evaluating alternate modeling approaches.

This data is global and covers the period between January 2001 and December 2015. The output variable, civil war, is drawn from the expanded version of the [21] article on defining and measuring civil war based on eleven criteria. The authors argue that this data offers advantages over currently used alternatives because conflicts are coded as continuous from onset to termination, even if the state is below a certain threshold of battle-related deaths for a given year. For instance, in the UCDP-PRIO dataset, Myanmar is coded as experiencing seven different civil wars between 2000-2015 where in Sambanis (2004), this would be coded as a single conflict. As a result, civil war onset is less common in this dataset than in alternatives, whereas incidence may be higher. The onset year is coded as the year in which the conflict causes at least 500 to 1,000 deaths, unless cumulative deaths over the next three years reach 1,000, in which case the year may also be coded as the onset year.[4]

B. Data Preparation

We base our analysis on the number and duration of civil wars at the country level globally during 2001-2015 We computed the following summary statistics at the level of the cohort: *mean number of civil wars*, and *mean duration of a civil war*. The expectation is that different countries from different regions and at different time periods behave differently with respect to those features. Thus, the input data we use comprise those features (civil war statistics) for each country and year in our data warehouse.

We binarize the input variables for each country/year observation by comparing the mean for the country/year with the mean for the cohort, computed from the training set. Every variable is set to 1 if it is above the respective population mean, and to 0 otherwise. This way, for each country/year pair we obtain two input binary variables (mean number of prior civil wars *Bmean* and mean duration of of those wars *Dmean*. Additionally, we introduced another binary variable for recent history of civil war *WLastYear*, which is set to 1 if there was a war last year in that country, 0 otherwise. Below we will denote negation of those variables with a preceding N character (e.g., *NBmean* for *not Bmean*, *NDmean* for *not Dmean*, *NWLastYear* for *not WLastYear*. The only output variable is *CivilWar*, which is set to 1 if there is a civil war that year, 0 otherwise. We set up the learning for TM to produce for single positive rule (in favor of a civil war). Figure 3 illustrates this process for five observations for Afganistan for years 2006-2010.

[4]We utilize this coding because it was the target of previous research. However, in the future, the forecasting of fatalities in distinct time-steps, instead of the use of retrospective coding that takes 3-years (two years past the focus year's time-stamp) to be reliably recorded, are likely more useful to policy-makers [22].

C. Data Use

In the first experiment, we use the data of all countries for 5 years (2001-2005) for training (2450 observations), and 5 years (2006-2010) for testing (2455 observations). In the second experiment, we use the data of all countries for 10 years (2001-2010) for training (2905 observations), and 5 years (2011-2015) for testing (2460 observations). We conduct those experiments for four different regions: Africa, Asia, the Middle East and Latin America.

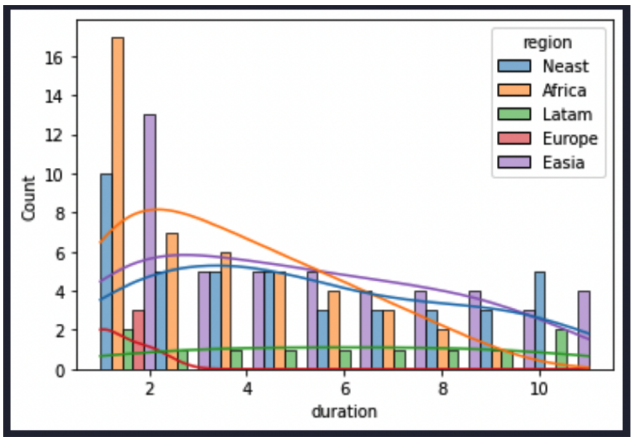

Fig. 4. Distribution of duration of civil wars in years per regions

D. Experimental Results

Figure 4 displays a plot of civil war duration, with the duration of the war (as of the end of the data) on the x axis and the frequency of each duration on the y axis, grouped by region. This describes the spells of sequential 1's (observations of civil war) in our data. The Africa region experiences the greatest frequency of civil wars during the study period, but these civil wars tend to be shorter in duration than in Latin America ("Latam" in the figure), the Middle East ("Neast") or Asia ("Easia"). In contrast, the Latin America region experiences one very long civil war. During this period, the European region experiences the least civil war overall, with two short conflicts.

Next we analyze the output of the interpretable learning system trained on two windows of the available data. We utilize confusion matrices and two types of accuracy, precision and recall measures. The interpretable learning system produces three outputs, predictions of positive (civil war), negative (no civil war) and uncertain conditions. These predictions are arrayed on the rows of the confusion matrix. The actual values are represented in the columns, hence the 3x2 structure. We can calculate accuracy, precision and recall when we ignore the uncertain predictions or include them as predicting civil war given the asymmetric cost of missing war versus peace.

Figure 5 shows the Africa results for 2005-=2010. The interpretable learning rules are $BMean\&WLastYear\&NDMean$, which in words means that the model predicts a civil war when the previous number of civil war observations is larger than the previously observed regional mean and when there was a war in the last year and when the duration of the current spell is less than the mean of previous durations. In addition, for this region, the rules of a negative prediction are $NBMean\&NWLastYear$ – the number of previous civil war observations is less than the mean for the region and the last observations was peace. These rules are consistent with theories that emphasize a history of conflict ($BMean$) as well the stickiness of civil war ($WLastYear$). The addition of $NDMeans$ in Africa suggests that while conflict is likely to continue year to year, this effect is potentially strongest in the earlier years of the conflict, relative to previously observed civil war durations. As we will see below, duration seems to operate distinctly across regions. The safest countries, fitting the negative rule, also conform to qualitative intuition whereby the absence of a history of conflict ($NBMean$) and current conflict ($NWLastYear$) lead to predictions of the absence of civil war.

Notice that a country-year observation could fall outside of both the positive and negative rules, and thus be uncertain. For example, in the Africa rules, an observation with the values $BMean\&NWLastYear$, $BMean\&DMean$, and $NBmean\&WLastYear$ would lead to a value of uncertain for the prediction, because they are cases that do not fit either of these positive or negative rules. The confusion matrix for Africa is illustrative. We see that out of 166 observations where the model made positive or negative prediction 164 of them or .988 were correct. Even when we include the 14 observations with uncertain predictions as positive, 176 out of 180 or .977 observations were correctly classified. The precision for Africa, excluding uncertain predictions is .818 and with uncertain predictions is .76. Recall is 1.0 for this sample. The rules for Africa in the 2011 to 2015 window were the same as the previous window, with precision of .826 when uncertain cases are categorized as positive and a recall of .904. These are impressive results relative to previous findings, considering the simplicity of the rules [20], [22].

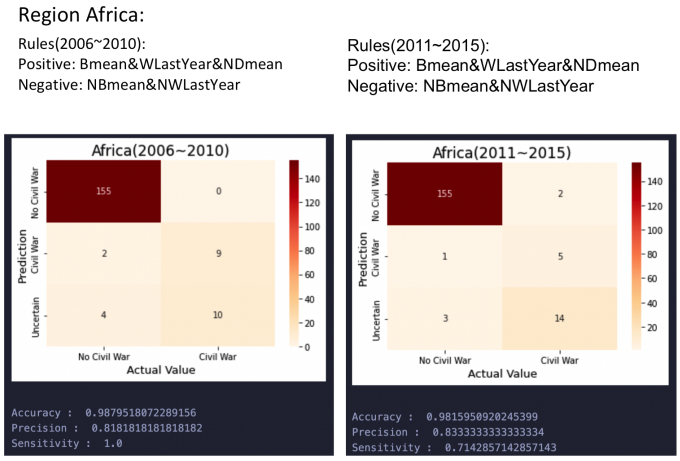

Fig. 5. Rules and prediction results for Africa

Region Easia:

Rules(2006~2010):
Positive: Bmean&WLastYear
Negative: NBmean&NWLastYear

Rules(2011~2015):
Positive: Bmean&WLastYear&NDmean
Negative: NBmean&NWLastYear

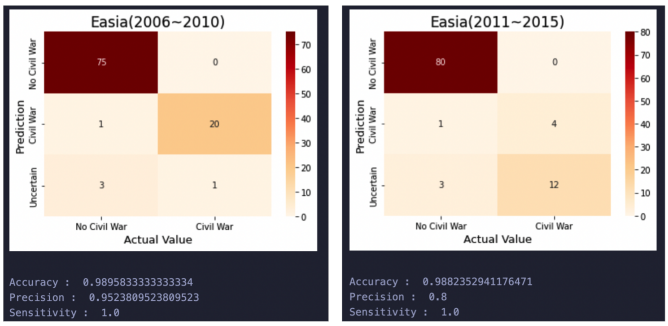

Fig. 6. Rules and prediction results for Eastern Asia

Region Latam:

Rules(2006~2010):
Positive: Bmean&WLastYear
Negative: NBmean&NWLastYear

Rules(2011~2015):
Positive: Bmean&WLastYear
Negative: NBmean&NWLastYear

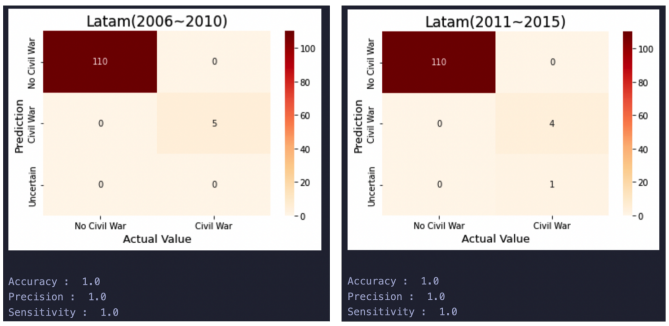

Fig. 8. Rules and prediction results for Latin America

Region Neast:

Rules(2006~2010):
Positive: Bmean&WLastYear
Negative: NBmean&NWLastYear

Rules(2011~2015):
Positive: Bmean&WLastYear&Dmean
Negative: NBmean&NDmean

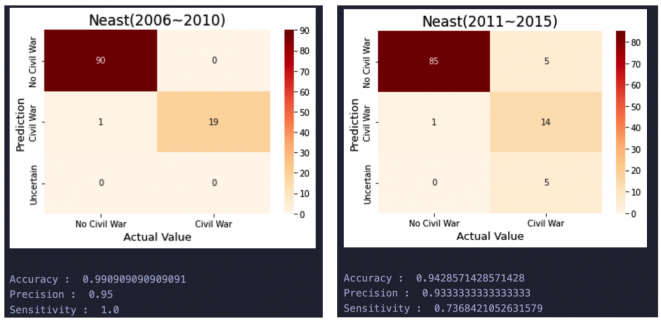

Fig. 7. Rules and prediction results for Middle East

As we look across the other regions (Figures 6,7,8), we see similar performance but with some differences in rules across regions. Specifically, unlike in Africa, the Middle East included longer durations $DMean$ in the positive rules, meaning that when there was a history of conflict ($BMean$) and proximate conflict $WLastYear$, a longer duration of a war spell $DMean$ is part of the positive rule (for 2011-2015). There are many reasons to expect different conflict dynamics across space and time [22] and thus it is not surprising that we have some variation in rules.

VII. Conclusion

We introduced and explored an advanced framework that utilizes the Tsetlin Machine to produce explainable rules for the prediction of civil war. Our framework also helps users to explain social instability and conflicts and to guide decision-making, theorizing, and predictions.

The framework includes a powerful data processing pipeline that produces a continuous data stream and provides continuous data interpretation for early warning of future conflicts. The data stream undergoes several notable transformations, starting from a non structured and often textual data representation followed by application-driven data aggregation. After that we apply various booleanizations of the aggregated information to get important binary indicators, which are used as input and output variables in the Tsetlin Machine learning

We found out that the computed output of the Tsetlin Machine can be appreciated as explainable in the most basic sense – whereby the model outputs parameters that are themselves explanations/sets of rules for why units of interest have distinct predictions over time and/or why two cases have distinct predictions in a given period.

We reported on an experimental study to explain and predict a most notable event, a civil war. We observed that the rules that we produced are consistent with theories that emphasize a history of conflict as well the stickiness of civil war. In future research we plan to explore wider range of input features (such as per capita income, population, percent mountainous terrain, export revenue from oil, etc.), as well as encoding other notable events (output features) that include more specific information, including which groups are the victims and perpetrators of political fatalities. Our goal is to have a flexible system that can produce warnings for both major and minor social unrest, significant violations of human rights, spikes in various kinds of social injustice, and inequality. We will also analyze the benefits of utilizing region as a feature within the interpretable learning framework, increasing the number of clauses, as well as comparing performance on the same tasks to other distinct algorithms.

Acknowledgment

The authors would like to thank Xiaozhong Zhang, Jiawei Xu, and Haocheng Wang for their work on the first version of the HR-VEAW system. This work was partially supported by NSF grant SES-2017614 and reflects only the authors' opinions.

REFERENCES

[1] L.-E. Cederman, N. B. Weidmann, and K. Gleditsch, "Horizontal inequalities and ethnonationalist civil war: A global comparison," *American Political Science Review*, vol. 105, no. 3, p. 478–495, 2011.

[2] X. Zhang, J. Xu, M. Keskin, M. Colaresi, and V. Zadorozhny, "HR-VEAW: A human rights violation exploration analytics, and warning system," in *Proceedings of the Workshops of the EDBT/ICDT 2022 Joint Conference, Edinburgh, UK, March 29, 2022*, ser. CEUR Workshop Proceedings, vol. 3135. CEUR-WS.org, 2022.

[3] O.-C. Granmo, "The Tsetlin Machine - A Game Theoretic Bandit Driven Approach to Optimal Pattern Recognition with Propositional Logic," *arXiv preprint arXiv:1804.01508*, 2018.

[4] R. Saha, O.-C. Granmo, V. Zadorozhny, and M. Goodwin, "A relational tsetlin machine with applications to natural language understanding," *CoRR abs/2102.10952. To appear in Journal of Intelligent Information Systems*, 2021.

[5] B. Park, K. Greene, and M. Colaresi, "Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects," *American Political Science Review*, vol. 114, no. 3, pp. 888–910, 2020.

[6] H. Hegre *et al.*, "Views: a political violence early-warning system," *Journal of peace research*, vol. 56, no. 2, pp. 155–174, 2019.

[7] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[8] J. Kreps *et al.*, "Kafka: A distributed messaging system for log processing," in *Proceedings of the NetDB*, vol. 11, 2011, pp. 1–7.

[9] G. T. Berge, O.-C. Granmo, T. O. Tveit, M. Goodwin, L. Jiao, and B. V. Matheussen, "Using the Tsetlin Machine to Learn Human-Interpretable Rules for High-Accuracy Text Categorization with Medical Applications," *IEEE Access*, vol. 7, pp. 115 134–115 146, 2019.

[10] M. L. Tsetlin, "On behaviour of finite automata in random medium," *Avtomat. i Telemekh*, vol. 22, no. 10, pp. 1345–1354, 1961.

[11] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," *arXiv:2006.11371v2 [cs.CV]*, 2020.

[12] G. Ferrettini, J. Aligon, and C. S. e Dup, "Improving on coalitional prediction explanation," *Proc. of ADBIS*, 2020.

[13] G. R. Lal, "Explainable neural networks: Recent advancements," *https://towardsdatascience.com/recent-advancements-in-explainable-neural-networks-2cd06b5d2016*, 2021.

[14] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv:1704.02685[cs.CV]*, 2019.

[15] E. Baillie, P. D. L. Howe, A. Perfors, T. Miller, Y. Kashima, and A. Beger, "Explainable models for forecasting the emergence of political instability," *PLOS ONE*, vol. 16, no. 7, pp. 1–18, 07 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0254350

[16] R. Marcinkevics and J. E. Vogt, "Interpretability and explainability: A machine learning zoo mini-tour," *arXiv:2012.01805v2[cs.LG]*, 2020.

[17] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2020.

[18] R. A. Blair and N. Sambanis, "Forecasting civil wars: Theory and structure in an age of "big data" and machine learning," *Journal of conflict resolution*, vol. 64, no. 10, pp. 1885–1915, 2020.

[19] A. Beger, R. K. Morgan, and M. D. Ward, "Reassessing the role of theory and machine learning in forecasting civil conflict," *Journal of Conflict Resolution*, vol. 65, no. 7-8, pp. 1405–1426, 2021.

[20] R. A. Blair and N. Sambanis, "Is theory useful for conflict prediction? a response to beger, morgan, and ward," *Journal of Conflict Resolution*, vol. 65, no. 7-8, pp. 1427–1453, 2021.

[21] N. Sambanis, "What is civil war? conceptual and empirical complexities of an operational definition," *Journal of conflict resolution*, vol. 48, no. 6, pp. 814–858, 2004.

[22] H. Hegre, P. Vesco, and M. Colaresi, "Lessons from an escalation prediction competition," *International Interactions*, vol. 48, no. 4, pp. 521–554, 2022.