

Building Concise Logical Patterns by Constraining Tsetlin Machine Clause Size

K. Darshana Abeyrathna^{1,3}, Ahmed A. O. Abouzeid¹, Bimal Bhattacharai¹, Charul Giri¹, Sondre Glimsdal¹, Ole-Christoffer Granmo¹, Lei Jiao¹, Rupsa Saha¹, Jivitesh Sharma¹, Svein A. Tunheim¹ and Xuan Zhang²

¹Centre for Artificial Intelligence Research (CAIR), University of Agder, Grimstad, Norway

²Norwegian Research Centre (NORCE), Grimstad, Norway

³DNV, Oslo, Norway

{ole.granmo; lei.jiao}@uia.no

Abstract

Tsetlin machine (TM) is a logic-based machine learning approach with the crucial advantages of being transparent and hardware-friendly. While TMs match or surpass deep learning accuracy for an increasing number of applications, large clause pools tend to produce clauses with many literals (long clauses). As such, they become less interpretable. Further, longer clauses increase the switching activity of the clause logic in hardware, consuming more power. This paper introduces a novel variant of TM learning – Clause Size Constrained TMs (CSC-TMs) – where one can set a soft constraint on the clause size. As soon as a clause includes more literals than the constraint allows, it starts expelling literals. Accordingly, oversized clauses only appear transiently. To evaluate CSC-TM, we conduct classification, clustering, and regression experiments on tabular data, natural language text, images, and board games. Our results show that CSC-TM maintains accuracy with up to 80 times fewer literals. Indeed, the accuracy increases with shorter clauses for TREC, IMDb, and BBC Sports. After the accuracy peaks, it drops gracefully as the clause size approaches a single literal. We finally analyze CSC-TM power consumption and derive new convergence properties.

1 Introduction

The TM [Granmo, 2018] is a novel approach to machine learning where groups of Tsetlin automata (TAs) [Tsetlin, 1961] produce logical (Boolean) expressions in the form of conjunctive clauses (AND-rules). As opposed to the black-box nature of deep neural networks, TMs are inherently interpretable. Indeed, they produce models based on sparse disjunctive normal form, which is comparatively easy for humans to understand [Valiant, 1984]. Additionally, the logical representations combined with automata-based learning make TMs natively suitable for hardware implementation, yielding low energy footprint [Wheeldon *et al.*, 2020].

TMs now support various architectures, including convolution [Granmo *et al.*, 2019], regression [Abeyrathna

et al., 2020c], deterministic [Abeyrathna *et al.*, 2020b], weighted [Abeyrathna *et al.*, 2020a], autoencoder [Bhattacharai *et al.*, 2023b], contextual bandit [Seraj *et al.*, 2022], relational [Saha *et al.*, 2022], and multiple-input multiple-output [Glimsdal and Granmo, 2021] architectures. The independent nature of clause learning allows efficient GPU-based parallelization, providing almost constant-time scaling with reasonable clause amounts [Abeyrathna *et al.*, 2021]. Several schemes enhance vanilla TM learning and inference, such as drop clause [Sharma *et al.*, 2023] and focused negative sampling [Glimsdal *et al.*, 2022]. These TM advances have enabled many applications: keyword spotting [Lei *et al.*, 2021], aspect-based sentiment analysis [Yadav *et al.*, 2021b], novelty detection [Bhattacharai *et al.*, 2022b; Bhattacharai *et al.*, 2021], semantic relation analysis [Saha *et al.*, 2021], text categorization [Yadav *et al.*, 2021a; Bhattacharai *et al.*, 2022a; Yadav *et al.*, 2022], game playing [Giri *et al.*, 2022], battery-less sensing [Bakar *et al.*, 2022b; Bakar *et al.*, 2022a], recommendation systems [Borgersen *et al.*, 2022], and knowledge representation [Bhattacharai *et al.*, 2023a].

The TM is somewhat related to Logistic Circuits [Liang and Van den Broeck, 2019a], which are Probabilistic Circuits for classification with logical expressions. However, Logistic Circuits use local search to build tractable Bayesian models, learning to classify through stochastic gradient descent. The TM, on the other hand, constructs pure logical AND-rules from online TA-driven learning with global convergence properties [Jiao *et al.*, 2023; Zhang *et al.*, 2022]. Compared to binary and real-valued Logistic Circuits, the TM outperform them accuracy-wise on MNIST and Fashion-MNIST [Sharma *et al.*, 2023]. The pure logical structure of TMs further allows model compression [Maheshwari *et al.*, 2023], yielding a tiny memory footprint.

While TMs match or surpass deep learning accuracy for an increasing number of applications, large clause pools tend to produce longer clauses containing many literals (input features and their negation). As such, they become less interpretable. Further, longer clauses require more memory and increase the switching activity in hardware, consuming more power. There is currently no direct way to control the size of the clauses learned. The challenge lies in coordinating the decentralized TAs, each independently learning whether to include a specific literal per clause. Indeed, since the TAs seek frequent discriminative patterns, the clauses can become

arbitrarily long. In short, current learning schemes seem inefficient when it comes to the size of the clauses they produce.

This paper introduces a novel variant of TM learning – Clause Size Constrained TMs (CSC-TMs) – where one can set a soft constraint on the clause size. CSC-TM revises the TA feedback policy for including literals into the clauses. Specifically, the new policy discourages including additional literals once the length of a clause surpasses a predefined constraint. The TAs instead immediately start expelling literals from the offending clause by reinforcing “exclude” actions. Accordingly, oversized clauses only appear transiently. Otherwise, the TM feedback scheme is left unchanged. The salient property of our approach is that the limited collections of literals ending up in the clauses maintain high discrimination power. Even with significantly constrained clause size, the performance of CSC-TM is not compromised compared with the other TM variants.

Paper Contributions: The contributions of the paper can be summarized as follows:

- We propose CSC-TM to constrain the size of the clauses by introducing a new policy for training TMs.
- We demonstrate that CSC-TM can indeed constrain clause size within an explicit limit, without compromising accuracy in classification, regression, and clustering.
- Using several examples, we show that the shorter clauses become more interpretable.
- We prove analytically that CSC-TM can converge to the intended basic operators when properly configured.
- We describe how constraining the length of the clauses is beneficial for power consumption in embedded hardware solutions due to the reduced switching activity of the clause logic.

2 Training Tsetlin Machines with Constrained Clause Length

In this section, we outline the difference between the vanilla TM¹ and CSC-TM.

A TM processes a vector $\mathbf{X} = [x_1, \dots, x_o]$ of propositional (Boolean) features as input, to be classified into one of two classes, $y = 0$ or $y = 1$. Negating these features produces a set of literals L that consists of the features and their negated counterparts: $L = \{x_1, \dots, x_o, \neg x_1, \dots, \neg x_o\}$.

A TM uses conjunctive clauses to represent sub-patterns. The number of clauses is given by a user set parameter n . For a two-class classifier², half of the clauses gets positive polarity (+). The other half gets negative polarity (−). Each clause $C_j^p, j \in \{1, 2, \dots, n/2\}, p \in \{-, +\}$, then becomes:

$$C_j^p(\mathbf{X}) = \bigwedge_{l_k \in L_j^p} l_k. \quad (1)$$

Here, j is the index of the clause, p its polarity, while L_j^p is a subset of the literals $L, L_j^p \subseteq L$. For example, the clause

¹For those who are not familiar with TM learning, please refer to: https://cair.github.io/ijcai_2023_clause_rationing.html.

²A multi-class classifier gets n clauses per class.

$C_1^+(\mathbf{X}) = \neg x_1 \wedge x_2$ has index 1, polarity +, and consists of the literals $L_1^+ = \{\neg x_1, x_2\}$. Accordingly, the clause outputs 1 if $x_1 = 0$ and $x_2 = 1$, and 0 otherwise.

The clause outputs are combined into a classification decision through summation and thresholding using the unit step function $u(v) = 1$ if $v \geq 0$ else 0:

$$\hat{y} = u\left(\sum_{j=1}^{n/2} C_j^+(\mathbf{X}) - \sum_{j=1}^{n/2} C_j^-(\mathbf{X})\right). \quad (2)$$

Namely, classification is performed based on a majority vote, with the positive clauses voting for $y = 1$ and the negative for $y = 0$. The classifier $\hat{y} = u((x_1 \wedge \neg x_2) + (\neg x_1 \wedge x_2) - (x_1 \wedge x_2) - (\neg x_1 \wedge \neg x_2))$, e.g., captures the XOR-relation.

For training, a dedicated team of TAs composes each clause C_j^p . Each TA of clause C_j^p decides to either *Include* or *Exclude* a specific literal l_k in the clause. A TA makes its decision based on the feedback it receives in the form of Reward, Inaction, and Penalty. There are two types of feedback associated with TM learning: Type I Feedback and Type II Feedback. Type I Feedback stimulates formation of frequent patterns, which suppresses false negative classifications. Type II Feedback, on the other hand, increases the discrimination power of the patterns, counteracting false positive classifications.

The difference between vanilla TM and CSC-TM lies in Type I Feedback. Type II Feedback remains the same for both schemes. Table 1 shows how CSC-TM modifies Type I Feedback to constrain clause size. The modification is marked by the red box. As seen, we introduce an additional condition for triggering the two leftmost feedback columns. These two columns make the clause mimic frequent patterns by reinforcing inclusion of “1”-valued literals with probability $\frac{s-1}{s}$ and by reinforcing exclusion of “0”-valued literals with probability $\frac{1}{s}$. The two rightmost columns, on the other hand, exclusively reinforce exclusion of literals.

CSC-TM requires that the size $\|C_j^p(\mathbf{X})\|$ of the clause is within a constraint \mathbf{b} to give access to the two leftmost columns. Accordingly, as soon as the number of literals in the clause surpasses the constraint \mathbf{b} , only *Exclude* actions are reinforced. The reason is that all the TA feedback then comes from the two rightmost columns, which reward *Exclude* and penalizes *Include* independently with probability $\frac{1}{s}$. As a result, the clause starts expelling literals when oversized, which means that oversized clauses only appear transiently.

In the following sections, we analyze the impact that CSC-TM has on convergence. We further investigate how constraining clause size affects accuracy in classification, regression, and clustering. Finally, we discuss effects on hardware complexity and energy consumption.

3 Convergence Analysis

Here we analyse the convergence property of two basic operators, namely the XOR and the OR operators, using the CSC-TM. Specifically, we show how the literal budget influences their convergences. We select those two operators for analysis mainly due to their simplicity and representativeness.

$C_j^p(\mathbf{X}) \wedge (\ C_j^p(\mathbf{X})\ \leq b)$		1		0	
$x_k/\neg x_k$		1	0	1	0
TA: Include Literal	P (Reward)	$\frac{s-1}{s}$	NA	0	0
	P (Inaction)	$\frac{1}{s}$	NA	$\frac{s-1}{s}$	$\frac{s-1}{s}$
	P (Penalty)	0	NA	$\frac{1}{s}$	$\frac{1}{s}$
TA: Exclude Literal	P (Reward)	0	$\frac{1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$
	P (Inaction)	$\frac{1}{s}$	$\frac{s-1}{s}$	$\frac{s-1}{s}$	$\frac{s-1}{s}$
	P (Penalty)	$\frac{s-1}{s}$	0	0	0

Table 1: Type I Feedback for CSC-TM. The feedback is for a single TA that decides whether to Include or Exclude a given literal $x_k/\neg x_k$ into C_j^p . NA means not applicable. s is a hyper-parameter greater than 1.

x_1	x_2	Output
0	0	0
1	1	0
0	1	1

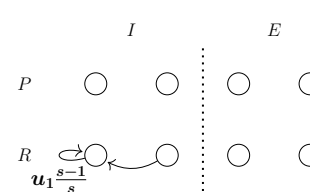
Table 2: A sub-pattern in ‘‘XOR’’ case.

3.1 XOR Operator

Here we study the convergence of the XOR operator when only one literal is given, i.e., $(\|C_j^i(\mathbf{X})\| = 1)^3$. We can then show that if the budget is not sufficient to represent a sub-pattern, the sub-pattern cannot be learnt. For XOR, a sub-pattern means $[x_1 = 1, x_2 = 0]$ or $[x_1 = 0, x_2 = 1]$, i.e., any one input pattern that can trigger $y = 1$. Clearly, the sub-patterns in XOR are mutual exclusive, and one literal cannot capture fully any sub-pattern of XOR. In what follows, we will show how the TM reacts upon training samples of XOR.

As already proven in [Jiao *et al.*, 2023], the vanilla TM can converge almost surely to the intended sub-pattern, i.e., $C_j^i = \neg x_1 \wedge x_2$, when the training samples in Table 2 is given. However, when $(\|C_j^i(\mathbf{X})\| = 1)$ is given in addition, the only absorbing state of the system, i.e., $C_j^i = \neg x_1 \wedge x_2$, disappears, making the system recurrent. More specifically, for vanilla TM, according to [Jiao *et al.*, 2023], when the training samples for $x_1 = 0, x_2 = 1, y = 1$ is given to the system and when $\text{TA}_1^3 = \text{Exclude}$, $\text{TA}_2^3 = \text{Include}$, and $\text{TA}_4^3 = \text{Exclude}$, the following transition⁴ holds for TA_3^3 .

Condition: $x_1 = 0,$
 $x_2 = 1, y = 1,$
 $\text{TA}_4^3 = \text{Exclude}.$
 Therefore, we have
 Type I Feedback
 for literal $x_2 = 1,$
 $C_3 = \neg x_1 \wedge x_2 = 1.$



³Here we follow the index of clause used in [Jiao *et al.*, 2023], where i in C_j^i is the index of class rather than the clause polarity. In [Jiao *et al.*, 2023], we did not specify the polarity of the clause in the proof.

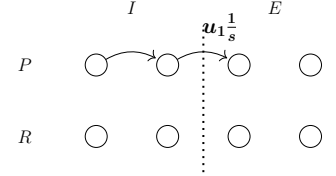
⁴It is the 2nd transition of **Case 1** in Subsection 3.2.1 in [Jiao *et al.*, 2023]. The transition diagram is derived based on the current status of the system and the input training samples. For details please refer to [Jiao *et al.*, 2023].

Here the superscript of TA_j^3 is the clause index and the subscript is the TA index. TA_1^3 has two actions, i.e., *Include* or *Exclude* x_1 . Similarly, TA_2^3 corresponds to *Include* or *Exclude* $\neg x_1$. TA_3^3 and TA_4^3 determine the behavior of the x_2 and $\neg x_2$, respectively. C_3 is the 3rd clause and here the class index i in the clause notation is removed for simplicity. P and R here mean penalty and reward respectively while I and E denote *Include* and *Exclude* respectively. u_1 is a constant in $[0, 1]$.

Clearly, for vanilla TM, the new training sample $x_1 = 0, x_2 = 1, y = 1$ will encourage TA_3^3 to be included, reinforcing C_3 being in the form $\neg x_1 \wedge x_2$. However, when the constraint $(\|C_j^i(\mathbf{X})\| = 1)$ is given in addition, the transition of TA_3^3 changes to:

Condition: $x_1 = 0,$
 $x_2 = 1, y = 1,$
 $\text{TA}_4^3 = \text{Exclude}.$

Therefore, we have
 Type I Feedback for
 literal $x_2 = 1, C_3 =$
 $\neg x_1 \wedge x_2 \wedge 0 = 0.$



The above change will make the only absorbing state, i.e., $C_3 = \neg x_1 \wedge x_2$, disappear. Understandably, given one literal that has already been included, the system will not encourage more literals to be included. Therefore, the TM, given the clause length being 1, cannot almost surely capture any sub-pattern in XOR. The above analysis also confirms that the newly added length constraint operates as expected, i.e., it indeed discourages more literals to be included once the length budget is reached. The complete proof can be found here⁵.

3.2 OR Operator

Now we study the OR operator, aiming at showing the fact that when the budget of the literals in a clause is sufficient to represent a sub-pattern (or a group of sub-patterns), the TM can learn the intended sub-pattern (or the intended group of sub-patterns). Before we study the convergence for the OR operator, let us revisit its nature. There are three sub-patterns that can trigger $y = 1$, i.e., $[x_1 = 1, x_2 = 1]$, $[x_1 = 0, x_2 = 1]$, and $[x_1 = 1, x_2 = 0]$. To represent each of the sub-pattern explicitly (or individually), we need two literals. However, two sub-clauses can also be represented jointly by one literal. Clearly, $C = x_1$ can cover both $[x_1 = 1, x_2 = 1]$ and $[x_1 = 1, x_2 = 0]$ while $C = x_2$ can cover both $[x_1 = 1, x_2 = 1]$ and $[x_1 = 0, x_2 = 1]$. This gives the TM possibility to learn the intended OR operator with clauses that has one literal, in collaboration with the hyper-parameter⁶ T .

Based on the analysis in [Jiao *et al.*, 2021], we understand that if $T = \lfloor \frac{m}{2} \rfloor$, and when T clauses learn x_1 and the other T clauses learn x_2 , the system is absorbed. This absorbing state learns the intended OR operator and also coincides with the requirement for the CSC-TM, which indicate that the OR

⁵https://cair.github.io/ijcai_2023_clause_rationing.html.

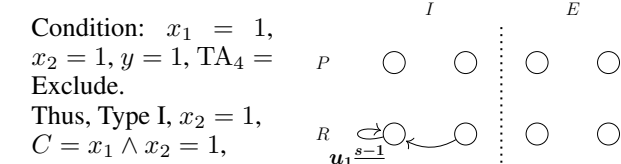
⁶The hyper-parameter T is utilized to guide different clauses to learn distinct sub-patterns. The details can be found in [Jiao *et al.*, 2023].

operator can possibly be learnt by the CSC-TM given literal length budget 1. In what follows, we show that the other absorbing states for the OR operator, i.e., with clauses that require more than one literal, will not be encouraged due to the newly added length constraint.

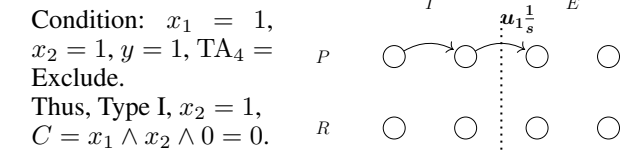
Similar to the analysis for the XOR case, once we revisit the transitions for the absorbing states with 2 literals, we realized that the absorbing states are not absorbing any more. More specifically, the Type I Feedback will encourage the included literal to be excluded. For example, in vanilla TM, for the sub-pattern below:

$$\begin{aligned} P(y = 1 | x_1 = 1, x_2 = 1) &= 1, \\ P(y = 0 | x_1 = 0, x_2 = 0) &= 1, \end{aligned} \quad (3)$$

the transition of TA_3 when its current action is *Include* and $TA_1 = \text{Include}$ and $TA_2 = \text{Exclude}$, namely



is replaced in CSC-TM by



Clearly, the state, i.e., $x_1 \wedge x_2$ is not absorbing any more, and the new constraint encourages the TA with included literal, in this case TA_3 , to move towards “Exclude”. Similar cases apply also to $[x_1 = 1, x_2 = 0]$ and $[x_1 = 0, x_2 = 1]$. Based on the above analysis, we can conclude that the CSC-TM can still learn OR operator but only with clauses that include 1 literal.

Note that although the length of the clauses are constrained, it may still happen that the length of a clause is over the budget. First, the length of a clause may be over the budget when the system update is blocked. Consider an extreme case for the OR operator when T clauses have x_1 and $T - 1$ clauses follow x_2 . In this situation, due to the randomness, a clause may become $\neg x_1 \wedge x_2$ based on a single training sample. In this situation, any future system update is blocked by T and the system is absorbed. Nevertheless, although this event may happen, the probability is very low. The reason is that it requires that both TAs happen to be in the boundary states at the *Exclude* side. This must happen at the same time as a training sample triggers the transitions in both TAs toward the *Include* side.

Another reason of being over the budget is that the Type II Feedback is not constrained and can produce clauses with more literals than the budget. In this case, the TAs of the included literals will all be in the boundary states, and the literals will be quickly swapping between being included or excluded in the clause, until the training stops. Type II Feedback ensures that all the literals are explored. During literal exploration, although the length of the clause can be large, the accuracy stays low until an accurate literal configuration

is found. Accordingly, the included literals at boundary due to Type II Feedback do not necessarily contribute positively to the classification. This can be observed from the numerical results for image processing when the literal budget is low (demonstrated for CIFAR-2 and MNIST with convolution in Table 3). Specifically, the convolution needs more literals so that the clause position also can be stored. With too few literals, the clauses precision suffers, triggering a large degree of Type II Feedback. The clauses will then be unable to settle because of the tight literal budget. As a result, the TAs of the included literals will not be able to progress deeply into the *Include* side of the their state space. Consequently, Type II Feedback will persistently continue to experiment with new candidate literals, without finding a sufficiently sparse high-accuracy configuration.

4 Empirical Results

In this section, we investigate the performance of CSC-TM, focusing on accuracy and interpretability. To this end, we evaluate classification, regression, and clustering performance on various datasets spanning natural language, images, board games, and tabular data. The experiments use a CUDA implementation of CSC-TM and runs on Intel Xeon Platinum 8168 CPU at 2.70 GHz and a Nvidia DGX-2 with Tesla V100 GPU. We describe the details of each task in respective subsections and summarize the findings in Tables 3, 4, and 5.

4.1 Natural Language Processing

We first evaluate CSC-TM on five NLP datasets: BBC sports [Greene and Cunningham, 2006], R8 [Debole and Sebastiani, 2005], TREC-6 [Chang *et al.*, 2002], SemEval 2010 Semantic Relations [Hendrickx *et al.*, 2009], and ACL Internet Movie Database (IMDb) [Maas *et al.*, 2011].⁷ Starting from a maximal constraint of 1, we progressively increase the literal constraint to 64, recording the resulting accuracy and the average number of literals included per clause. For BBC Sports, we notice that a literal constraint of 8 yields the maximum accuracy of 99.1%. Similarly, the maximum accuracy for TREC and R8 are achieved with literal counts 32 and *all*, respectively. We further observe that incorporating all the literals reduces the accuracy in BBC and TREC. For all the datasets, the average literal count drops significantly with literal budgeting. For instance, the literal constraint of 32 for BBC Sports gives 2.64 literals per clause on average, whereas the average is 44.14 without constraints. Accordingly, the clause length is considerable shortened, and the clauses can be quickly glanced by humans for interpretation. Similar trends can also be seen for the other data sets.

Consider as an example the results from R8 in Table 3. Notice how a constraint as strict as 4 still almost maximizes accuracy. Indeed, the NLP results overall show that CSC-TM allows us build concise and accurate propositional rules for better understandability. The improved interpretability is showcased in Figure 1, where we retrieve a few typical clauses from the “Football” class of BBC Sports. The literal-constrained clauses (in red) only contain 2-3 literals, and they

⁷For TM hyperparameters in BBC Sports, TREC, and R8, we use 8000 clauses, a voting margin T of 100, and specificity s of 10.0.

Budget →	Accuracy (Literals per Clause)							
	≤1	≤2	≤4	≤8	≤16	≤32	≤64	All
BBC Sports	98.65 (1.59)	98.65 (1.65)	98.65 (1.71)	99.1 (1.93)	98.2 (2.15)	98.2 (2.64)	98.2 (3.35)	94.14 (44.14)
TREC	91.8 (1.05)	91.6 (1.07)	92.4 (1.07)	90.2 (1.08)	90.4 (1.1)	93.2 (1.12)	90.6 (1.16)	85.8 (96.24)
R8	95.08 (1.09)	94.54 (1.13)	95.08 (1.14)	94.9 (1.19)	95.08 (1.26)	95.26 (1.41)	95.08 (1.7)	95.8 (21.84)
California Housing (5-bins)	59.25 (1.08)	62.28 (1.46)	64.2 (3.09)	65.02 (6.66)	65.24 (12.92)	65.26 (17.15)	65.29 (18.48)	65.79 (20.22)
SEMEVAL	93.65 (1.23)	93.00 (1.26)	92.25 (1.32)	93.2 (1.41)	92.95 (1.62)	93.00 (2.88)	93.1 (5.67)	93.63 (142.97)
IMDb	81.58 (1.08)	81.51 (1.22)	81.28 (1.28)	82.01 (1.42)	83.44 (1.75)	85.67 (2.76)	87.73 (4.05)	84.23 (27.02)
CIFAR-2	69.82 (30.8)	79.65 (30.1)	87.01 (13.5)	91.21 (6.8)	93.23 (10.8)	93.99 (20.1)	94.24 (34.1)	94.18 (60.4)
MNIST	92.09 (1.0)	97.42 (1.4)	98.34 (3.0)	98.40 (5.8)	98.42 (10.8)	98.38 (19.9)	98.33 (33.0)	98.32 (47.7)
MNIST w/conv.	40.94 (18.6)	60.55 (15.3)	95.93 (5.8)	99.22 (7.1)	99.33 (13.4)	99.29 (23.5)	99.30 (34.2)	99.28 (40.3)
Energy Performance (Regression)	5.65 (1.0)	2.44 (1.9)	1.05 (3.9)	0.86 (6.2)	0.78 (9.3)	0.66 (11.2)	0.63 (11.3)	0.59 (11.5)
Hex (Reinforcement Learning)	67.59 (2.7)	74.69 (2.4)	77.79 (3.0)	79.49 (4.2)	81.23 (6.1)	81.60 (9.3)	82.17 (10.2)	81.43 (13.8)

Table 3: Performance on multiple data sets for a literal budgets of 1, 2, 4, 8, 16, 32, 64, and all (no constraint) literals. Other than for Regression analysis, each field reports the worst maximum accuracy across 5 independent runs, followed by, in brackets, the average number of literals used by the TM. For Energy Performance, we report MAE instead of accuracy.

Clauses	Literal Count
$C_1 = [club \wedge football \wedge \neg six]$	3
$C_2 = [olympics \wedge title]$	2
⋮	⋮
$C_m = [goal \wedge winger]$	2
$C_1 = [manchester \wedge \neg australia \wedge \dots \wedge \neg century]$	329
$C_2 = [\neg rugby \wedge \neg ship \wedge \dots \wedge arsenal]$	331
⋮	⋮
$C_m = [barcelona \wedge matches \wedge \dots \wedge injury]$	5

Figure 1: Interpretability of clauses with constrained clause size for class “football” from BBC Sports. The rows highlighted in red are clauses from CLC-TM, while the yellow rows are from vanilla TM.

clearly relate to the “Football” class. The vanilla clauses (in yellow), however, contain a much larger number of literals and rely extensively on feature negation.

4.2 Image Processing

We evaluate our approach on two image datasets: MNIST and CIFAR-2 (a two-class variant of CIFAR-10 that groups vehicle images and animal images into two separate classes)⁸. For MNIST, we perform experiments utilizing both vanilla

⁸For hyperparameters, we adopt 8000 clauses per class, a voting margin T as 10000, and specificity s as 5.0 in the MNIST experiments. For CIFAR-2, the number of clauses is 8000, T is 6000, and s is 10.0.

and convolutional TM with constrained clause length. Observe from Table 3 how a constraint as small as 8 still yields competitive accuracy. Specifically, the maximum accuracy is obtained in CIFAR-2, MNIST, and MNIST w/conv. with literal constrains 64, 16, and 16, respectively. Also notice that CSC-TM on average keeps the number of literals per clause well below the set constraints. For example, for the latter constraints, the corresponding average number of literals per clause are respectively 34.1, 10.8, and 13.4. Without clause length constraints, however, the corresponding average number of included literals are 60.4, 47.7, and 40.3. Finally, notice how setting a too tight literal budget (below 4) for MNIST w/conv. and CIFAR-2 increases the average number of literals used. This can be explained by CSC-TM not finding sufficiently accurate patterns. As a result, it stays in literal exploration mode throughout the epochs. In conclusion, we observe that the maximum accuracy can be achieved using significantly fewer literals per clauses using CLC-TM. This allows us to significantly reduce computational complexity and increase the readability of the clauses.

To investigate how the number of clauses interact with the literal constraint, we now measure the effect of jointly varying the number of clauses and literal budget. From Table 4, we observe a graceful degradation of accuracy as the number of clauses drops to 250 and the literal budget falls to 1. Also notice that fewer clauses produces fewer literals per clause on average. We believe this is the case because when fewer clauses are available, they must become less specialized to solve the task. However, when looking at attaining maximum accuracy, we observe that 1000 clauses require more literals than 4000 (32 vs 16 literals on average per clause). The

reason may be that fewer clauses need to be more specific to compensate and maintain accuracy. As we increase the number of clauses, each clause includes fewer literals, solving the task as an ensemble. In conclusion, CSC-TM allows a fine-grained trade-off between the length of clauses and the number of clauses.

It is interesting to compare the CSC-TM performance against logistic circuits [Liang and Van den Broeck, 2019b]. The binary version of the logistic circuits classifier achieves 97.8% accuracy on MNIST, while the real-valued version obtains 99.4%. The CSC-TM achieves 98.42% test accuracy for MNIST for configurations of 8000 clauses and a literal budget of 16. With convolution the CSC-TM obtains 99.33% test accuracy, also with a literal budget of 16.

4.3 Self-Supervised Learning

For the self-supervised learning task, we evaluate how the clause literal budget influences both training time and interpretability. Here, we evaluate the previously proposed Label-Critic TM (LCTM) [Abouzeid *et al.*, 2022], which is a novel architecture to self-learn data samples’ labels without knowing the ground truths. The LCTM architecture runs on top of the standard CUDA TM implementation and starts by randomly initializing the data labels. Thereafter, it performs hierarchically clustering while learning the sub-patterns and their associated labels. Eventually, the learned sub-patterns represent interpretable clusters, each associated with a single supporting and a single discriminating clause. As a result, the method is interpretable, however, can still benefit from smaller clauses.

Table 5 shows the empirical results from different clause literal budgets on a subset of the MNIST dataset. Here, LCTM is to learn the labels and sub-patterns of the MNIST samples, associated with the digits “One” and “Zero”. The results capture how the literal constraint influences both the training time and the interpretability. The interpretability metrics are as follows. *Supporting interpretability* is the percentage of the positive polarity clauses that a human verifies as recognizable. See Figure 2 for examples of clauses that are deemed interpretable and not interpretable. Similarly, *discriminating interpretability* is the percentage of negative polarity clauses that are recognized by humans.

As shown in Table 5, when the budget was reduced to 1,200, the LCTM was both faster and more interpretable. Clearly, constraining the size of clauses yielded both increased interpretability and learning speed in our evaluations.

4.4 Regression

We use the Energy Performance dataset to evaluate regression performance based on [Abeyrathna *et al.*, 2020c]. The results are reported in terms of Mean Average Error (MAE). In brief, Table 3 shows that the MAE decreases, i.e., the performance is better, as the literal budget is increased from 1 to 64. Again, the degradation of performance is graceful, and one can trade off clause size against MAE.

4.5 Board Game Winner Prediction

We here use the Hex game as an example of reinforcement learning with CSC-TM, where the task is to predict the win-



Figure 2: Example of four clusters deemed human-interpretable (Good) and four clusters not being interpretable (Bad).

ner (value) of any given board configuration. To investigate how the prediction accuracy varies for CSC-TM, we compare the vanilla TM [Giri *et al.*, 2022] with the CSC-TM for distinct literal budgets. The details of the experiment setup can be found in [Giri *et al.*, 2022]. Bottom row of Table 3 summarises how the accuracy varies with the literal budget. We observe that a literal budget of 64 reaches the maximum accuracy of 82.17% against an accuracy of 81.43% for all the literals. However, it is to be noted that the smaller literal budgets provide relatively poor accuracy. We believe this is the case because describing a Hex board configuration accurately generally requires information on a sufficient number of piece positions due to the nature of the game.

5 Hardware Complexity and Energy Consumption Analysis

TM hardware accelerators will typically be implemented by either Field Programmable Gate Arrays (FPGAs) or Application Specific Integrated Circuits (ASICs). In most cases the dominating part of the energy consumption is related to the switching of digital circuits. It should be noted, however, that the static power consumption due to transistor leakage current for high performance processes can reach up to 30% of total power [Dally *et al.*, 2016]. With implementations in low power processes the static power consumption will be less.

The dynamic power, P , consumed by a digital circuit with load capacitance C , operating frequency f , supply voltage V_S , and an activity factor α (transitions per clock cycle) is given by Eq. (4) [Dally *et al.*, 2016],

$$P = 0.5 \times C \times V_S^2 \times f \times \alpha. \quad (4)$$

Limiting the number of literals will reduce the α value in several gates that implement the clause logic. Only those gates

Budget→	Accuracy (Literals per Clause)							
	≤1	≤2	≤4	≤8	≤16	≤32	≤64	All
MNIST w/250 clauses	88.20 (1.1)	93.02 (1.4)	95.96 (2.1)	96.65 (2.8)	96.92 (3.2)	97.03 (3.6)	97.06 (3.8)	97.04 (4.0)
MNIST w/500 clauses	89.05 (1.0)	94.28 (1.3)	96.82 (2.1)	97.50 (3.2)	97.67 (4.5)	97.74 (6.2)	97.78 (7.8)	97.77 (9.0)
MNIST w/1000 clauses	90.22 (1.0)	95.26 (1.2)	97.67 (2.6)	98.03 (4.7)	98.10 (8.4)	98.14 (14.6)	98.09 (22.8)	98.05 (31.2)
MNIST w/2000 clauses	90.88 (1.0)	96.23 (1.2)	98.01 (2.7)	98.26 (5.3)	98.26 (9.7)	98.33 (17.5)	98.22 (28.5)	98.2 (40.1)
MNIST w/4000 clauses	91.68 (1.0)	97.08 (1.3)	98.19 (2.9)	98.37 (5.6)	98.4 (10.5)	98.35 (19.1)	98.32 (31.4)	98.29 (45.2)
MNIST w/8000 clauses	92.09 (1.0)	97.42 (1.4)	98.34 (3.0)	98.40 (5.8)	98.42 (10.8)	98.38 (19.9)	98.33 (33.0)	98.32 (47.7)

Table 4: Performance on MNIST with different numbers of total clauses, where each clause has a literal budget of 1, 2, 4, 8, 16, 32, 64, and all (no constraint) literals. Each field reports worst maximum accuracy across 5 independent runs, followed by, in brackets, the average number of literals used by the TM.

Literals #	Supporting Interpretability (%)	Discriminating Interpretability (%)	Clusters #	Speed up
400	84.92 ± 5.21	32.17 ± 6.34	36 ± 7.04	1.1 ×
800	85.48 ± 7.80	33.02 ± 7.21	35.8 ± 6.76	1.3 ×
1, 200	88.84 ± 2.71	39.05 ± 8.03	34.6 ± 9.54	1.3 ×
1, 568 (all)	86.51 ± 1.33	33.63 ± 7.59	33.8 ± 7.34	1 ×

Table 5: LCTM performance on MNIST with different literal budgets. Mean and standard deviation are calculated over 5 independent runs.

that process the included literals will switch and consume energy. The reduction in energy consumption of the clause logic can therefore roughly be estimated to $\frac{b}{l_{ave}}$, where b is the clause size constraint and l_{ave} is the average number of literals in the model without the length constraint. The exact savings will depend on the dataset.

The classical approach [Wheeldon *et al.*, 2020] for implementing clause logic is shown in Figure 3. Here each literal l_1, \dots, l_{2o} is either included or excluded by the associated include signal i_1, \dots, i_{2o} (active low) by using OR-gates. The include signals will typically all be simultaneously available from a register. The outputs from the OR-gates are then fed to a wide-input AND-gate, which will normally be implemented by several smaller AND-gates connected in a tree-structure to reduce path delay. Clearly, for a certain fixed application, given a smaller number of included literals, we can reduce the number of OR-gates to b , and use an AND gate with only b inputs. In this way, the hardware complexity and power consumption can be reduced. For a general case, where the TM needs to be programmable in distinct applications, we need to have sufficient amount of available literals. Nevertheless, CSC-TM still has benefits in reduced switching activity, thus saving power.

It should be noted that it is only the energy consumption related to the clause logic that is affected by constraining clause size. The ensuing hardware processing, e.g., with clause weighting and summation is not affected. However, for systems with a huge number of clauses, the clause logic will occupy a significant part of the digital circuitry, and reducing its switching activity can enable significant energy savings.

An important system level benefit of literal budgeting is the time required for a model to be loaded from external or on-

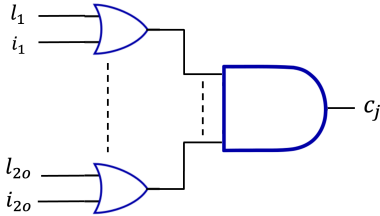


Figure 3: Hardware implementation of clause logic.

chip memory into registers in the ASIC or FPGA. During this time the system’s processor and the data transfer itself will consume energy. With less included literals, the model size and thus the load time can be reduced with a suitable encoding scheme, such as Run Length Encoding (RLE) [Bakar *et al.*, 2022a]. Reduction of the model size is also beneficial for other embedded TM solutions based on low-power microcontrollers.

6 Conclusions and Future Work

In this paper, we proposed CSC-TM — a novel TM mechanism that constrains the size of clauses. We argued how limiting the number of literals significantly reduce switching activity in hardware, and thereby power consumption. We further analyzed and confirmed the convergence of CSC-TM. Experimental results showed that CSC-TM can achieve the same or even better accuracy with shorter clauses, providing better interpretability. Future work includes introducing other kinds of constraints, with the intent of supporting constrained machine learning in general.

Acknowledgments

This work is supported in part by the project “Logic-based Artificial Intelligence Everywhere: Tsetlin Machines in Hardware” and funded under the grant number 312434 of the Research Council of Norway. This work is also supported in part by the project “Unleashing the Sustainable Value Creation Potential of Offshore Ocean” and funded under the grant number 328724 of the Research Council of Norway.

Contribution Statement

All authors have contributed equally in terms of conceptualization, experimentation and production of this paper.

References

- [Abeyrathna *et al.*, 2020a] K. Darshana Abeyrathna, Ole-Christoffer Granmo, and Morten Goodwin. Extending the Tsetlin Machine With Integer-Weighted Clauses for Increased Interpretability. *arXiv preprint arXiv:2005.05131*, 2020.
- [Abeyrathna *et al.*, 2020b] K. Darshana Abeyrathna, Ole-Christoffer Granmo, Rishad Shafik, Alex Yakovlev, Adrian Wheeldon, Jie Lei, and Morten Goodwin. A Novel Multi-Step Finite-State Automaton for Arbitrarily Deterministic Tsetlin Machine Learning. In *the 40th International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI-2020)*. Springer International Publishing, 2020.
- [Abeyrathna *et al.*, 2020c] K. Darshana Abeyrathna, Ole-Christoffer Granmo, Xuan Zhang, Lei Jiao, and Morten Goodwin. The Regression Tsetlin Machine - A Novel Approach to Interpretable Non-Linear Regression. *Philosophical Transactions of the Royal Society A*, 378, 2020.
- [Abeyrathna *et al.*, 2021] K. Darshana Abeyrathna, Bimal Bhattarai, Morten Goodwin, Saeed Gorji, Ole-Christoffer Granmo, Lei Jiao, Rupsa Saha, and Rohan K. Yadav. Massively Parallel and Asynchronous Tsetlin Machine Architecture Supporting Almost Constant-Time Scaling. In *International Conference on Machine Learning (ICML)*, 2021.
- [Abouzeid *et al.*, 2022] Ahmed Abouzeid, Ole-Christoffer Granmo, Morten Goodwin, and Christian Webersik. Label-Critic Tsetlin Machine: A Novel Self-supervised Learning Scheme for Interpretable Clustering. In *International Symposium on the Tsetlin Machine (ISTM)*, pages 41–48. IEEE, 2022.
- [Bakar *et al.*, 2022a] Abu Bakar, Tousif Rahman, Alessandro Montanari, Jie Lei, Rishad Shafik, and Fahim Kawsar. Logic-based Intelligence for Batteryless Sensors. In *the Annual International Workshop on Mobile Computing Systems and Applications (HotMobile)*, pages 22–28. Association for Computing Machinery, 2022.
- [Bakar *et al.*, 2022b] Abu Bakar, Tousif Rahman, Rishad Shafik, Fahim Kawsar, and Alessandro Montanari. Adaptive Intelligence for Batteryless Sensors Using Software-Accelerated Tsetlin Machines. In *the 20th Conference on Embedded Networked Sensor Systems*. ACM, 2022.
- [Bhattarai *et al.*, 2021] Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Measuring the novelty of natural language text using the conjunctive clauses of a Tsetlin machine text classifier. In *International Conference on Agents and Artificial Intelligence*, 2021.
- [Bhattarai *et al.*, 2022a] Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment. In *the 13th Conference on Language Resources and Evaluation*, pages 4894–4903, 2022.
- [Bhattarai *et al.*, 2022b] Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Word-level Human Interpretable Scoring Mechanism for Novel Text Detection Using Tsetlin Machines. *Applied Intelligence*, 52:17465–17489, 2022.
- [Bhattarai *et al.*, 2023a] Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. An interpretable knowledge representation framework for natural language processing with cross-domain application. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, pages 167–181, 2023.
- [Bhattarai *et al.*, 2023b] Bimal Bhattarai, Ole-Christoffer Granmo, Lei Jiao, Rohan Yadav, and Jivitesh Sharma. Tsetlin Machine Embedding: Representing Words Using Logical Expressions. *arXiv preprint arXiv:2301.00709*, 2023.
- [Borgersen *et al.*, 2022] Karl Audun Kagnes Borgersen, Morten Goodwin, and Jivitesh Sharma. A Comparison Between Tsetlin Machines and Deep Neural Networks in the Context of Recommendation Systems. *arXiv preprint arXiv:2212.10136*, 2022.
- [Chang *et al.*, 2002] Eric Chang, Frank Seide, Helen M Meng, Zhuoran Chen, Yu Shi, and Yuk-Chi Li. A System for Spoken Query Information Retrieval on Mobile Devices. *IEEE Transactions on Speech and Audio processing*, 10(8):531–541, 2002.
- [Dally *et al.*, 2016] William J. Dally, Harting R. Curtis, and Tor M. Aamodt. *Digital Design Using VHDL: a Systems Approach*. Cambridge University Press, 2016.
- [Debole and Sebastiani, 2005] Franca Debole and Fabrizio Sebastiani. An analysis of the relative hardness of Reuters-21578 subsets. *Journal of the American Society for Information Science and technology*, 56(6):584–596, 2005.
- [Giri *et al.*, 2022] Charul Giri, Ole-Christoffer Granmo, Herke Van Hoof, and Christian D. Blakely. Logic-based AI for Interpretable Board Game Winner Prediction with Tsetlin Machine. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2022.
- [Glimsdal and Granmo, 2021] Sondre Glimsdal and Ole-Christoffer Granmo. Coalesced Multi-Output Tsetlin Machines with Clause Sharing. *arXiv preprint, arXiv:2108.07594*, 2021.
- [Glimsdal *et al.*, 2022] Sondre Glimsdal, Rupsa Saha, Bimal Bhattarai, Charul Giri, Jivitesh Sharma, Svein Anders

- Tunheim, and Rohan Kumar Yadav. Focused Negative Sampling for Increased Discriminative Power in Tsetlin Machines. In *2022 International Symposium on the Tsetlin Machine (ISTM)*, pages 73–80, 2022.
- [Granmo *et al.*, 2019] Ole-Christoffer Granmo, Sondre Glimsdal, Lei Jiao, Morten Goodwin, Christian W. Omlin, and Geir Thore Berge. The Convolutional Tsetlin Machine. *arXiv preprint arXiv:1905.09688*, 2019.
- [Granmo, 2018] Ole-Christoffer Granmo. The Tsetlin Machine - A Game Theoretic Bandit Driven Approach to Optimal Pattern Recognition with Propositional Logic. *arXiv preprint arXiv:1804.01508*, 2018.
- [Greene and Cunningham, 2006] Derek Greene and Pádraig Cunningham. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering. In *International Conference on Machine Learning (ICML)*, pages 377–384. ACM Press, 2006.
- [Hendrickx *et al.*, 2009] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics, 2009.
- [Jiao *et al.*, 2021] Lei Jiao, Xuan Zhang, and Ole-Christoffer Granmo. On the Convergence of Tsetlin Machines for the AND and the OR Operators. *arXiv preprint <https://arxiv.org/abs/2109.09488>*, 2021.
- [Jiao *et al.*, 2023] Lei Jiao, Xuan Zhang, Ole-Christoffer Granmo, and K. Darshana Abeyrathna. On the Convergence of Tsetlin Machines for the XOR operator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6072–6085, 2023.
- [Lei *et al.*, 2021] Jie Lei, Tousif Rahman, Rishad Shafik, Adrian Wheeldon, Alex Yakovlev, Ole-Christoffer Granmo, Fahim Kawsar, and Akhil Mathur. Low-Power Audio Keyword Spotting Using Tsetlin Machines. *Journal of Low Power Electronics and Applications*, 11, 2021.
- [Liang and Van den Broeck, 2019a] Yitao Liang and Guy Van den Broeck. Learning logistic circuits. In *Proceedings of the 33rd Conference on Artificial Intelligence (AAAI)*, jan 2019.
- [Liang and Van den Broeck, 2019b] Yitao Liang and Guy Van den Broeck. Learning logistic circuits. In *the AAAI Conference on Artificial Intelligence*, volume 33, pages 4277–4286, 2019.
- [Maas *et al.*, 2011] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. In *the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.
- [Maheshwari *et al.*, 2023] Sidharth Maheshwari, Tousif Rahman, Rishad Shafik Senior member, Alex Yakovlev, Ashur Rafiev, Lei Jiao, and Ole-Christoffer Granmo. Re-dress: Generating compressed models for edge inference using tsetlin machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2023.
- [Saha *et al.*, 2021] Rupsa Saha, Ole-Christoffer Granmo, and Morten Goodwin. Using Tsetlin Machine to Discover Interpretable Rules in Natural Language Processing Applications. *Expert Systems*, page e12873, 2021.
- [Saha *et al.*, 2022] Rupsa Saha, Ole-Christoffer Granmo, Vladimir Zadorozhny, and Morten Goodwin. A Relational Tsetlin Machine with Applications to Natural Language Understanding. *Journal of Intelligent Information Systems*, 2022.
- [Seraj *et al.*, 2022] Raihan Seraj, Jivitesh Sharma, and Ole Christoffer Granmo. Tsetlin Machine for Solving Contextual Bandit Problems. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [Sharma *et al.*, 2023] Jivitesh Sharma, Rohan Kumar Yadav, Ole-Christoffer Granmo, and Lei Jiao. Drop Clause: Enhancing Performance, Robustness and Pattern Recognition Capabilities of the Tsetlin Machine. In *the AAAI Conference on Artificial Intelligence (AAAI)*, 2023.
- [Tsetlin, 1961] Michael Lvovitch Tsetlin. On Behaviour of Finite Automata in Random Medium. *Avtomat. i Telemekh.*, 22(10):1345–1354, 1961.
- [Valiant, 1984] Leslie G Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
- [Wheeldon *et al.*, 2020] Adrian Wheeldon, Rishad Shafik, Tousif Rahman, Jie Lei, Alex Yakovlev, and Ole-Christoffer Granmo. Learning Automata based Energy-efficient AI Hardware Design for IoT. *Philosophical Transactions of the Royal Society A*, 2020.
- [Yadav *et al.*, 2021a] Rohan Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. Enhancing interpretable clauses semantically using pretrained word representation. In *the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021.
- [Yadav *et al.*, 2021b] Rohan Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. Human-Level Interpretable Learning for Aspect-Based Sentiment Analysis. In *the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.
- [Yadav *et al.*, 2022] Rohan Kumar Yadav, Lei Jiao, Ole Christoffer Granmo, and Morten Goodwin. Robust Interpretable Text Classification against Spurious Correlations Using AND-rules with Negation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.
- [Zhang *et al.*, 2022] Xuan Zhang, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. On the Convergence of Tsetlin Machines for the IDENTITY- and NOT Operators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6345–6359, 2022.