CAIR CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH

Explainable Tsetlin Machine Framework for Fake News Detection with Credibility Score Assessment

BIMAL BHATTARAI

University of Agder

UiA University of Agder Norway

Introduction

- Fake news are the news that intentionally spread misinformation and deceives people for political and financial gain.
- Most people depends on online platform for news than newspapers.
- Encourages biases, false hopes, increase distrust and can trigger violence.
- Language models such as GPT-3 enable the automatic generation of realistic-looking fake news accelerating growth.
- The explainability of decision can help to overcome many challenges and performance improvements.

Background

- Many works on detecting false online advertising, fake consumer reviews, and spam emails.
- Typical detection technique use text-based linguistic or visual features.
- Recent deep-learning methods uses features based on social platform-specific features such as tweets, retweets etc.
- Existing deep learning methods lacks transparency and interpretability.

Problem Statement

- Fake news can be generated using advance models/algorithms.
- Intentionally created to deceive a user.
- Narrate real events with fake claims.
- Handcrafted and data specific features fails.

labeled data

$$X = (x_i, y_i),\ x_i \in \mathcal{R}^s$$

classifier function

$$\mathcal{F} : X \rightarrow y \in \{0, 1\}$$

Credibility

$$\mathcal{F} : X \rightarrow (y, \mathcal{Q})$$

Tsetlin Machine

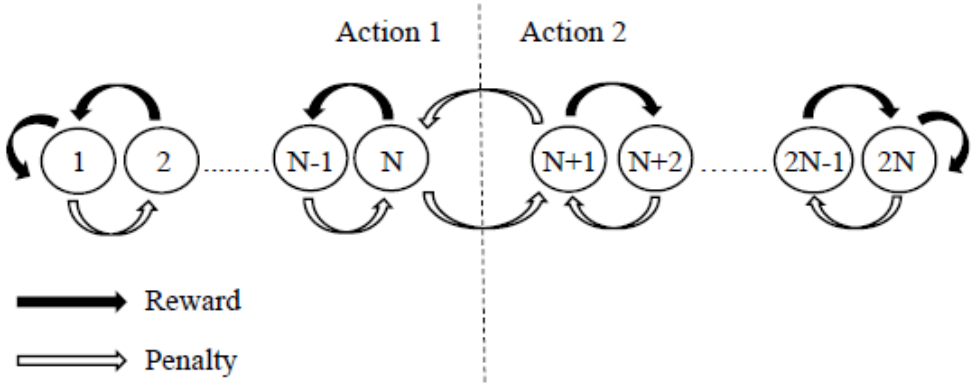
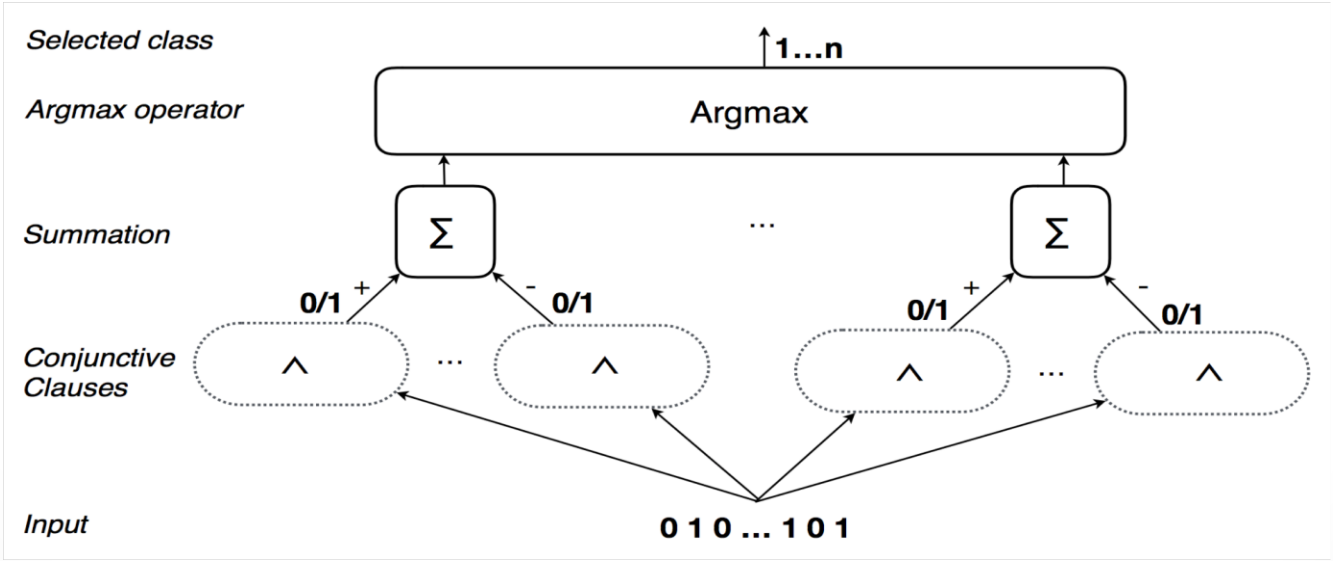


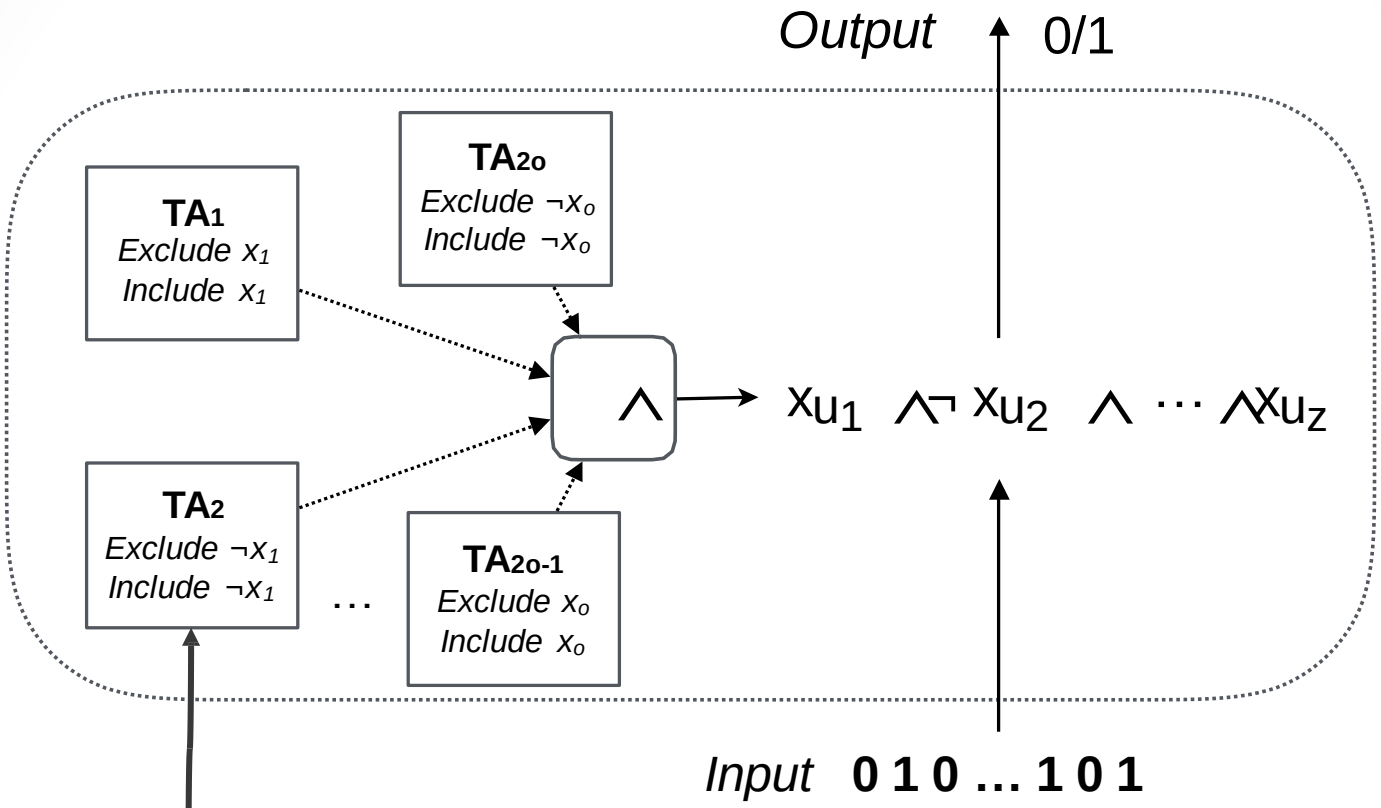
Fig. 1: Transition graph of a two-action Tsetlin Automaton.

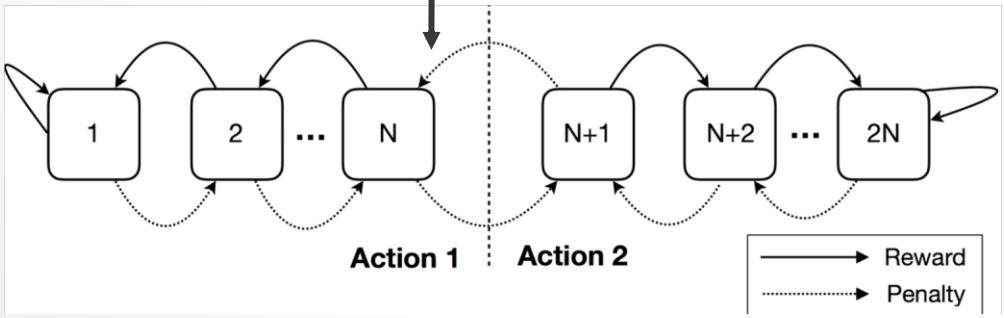


Why Tsetlin Machine?

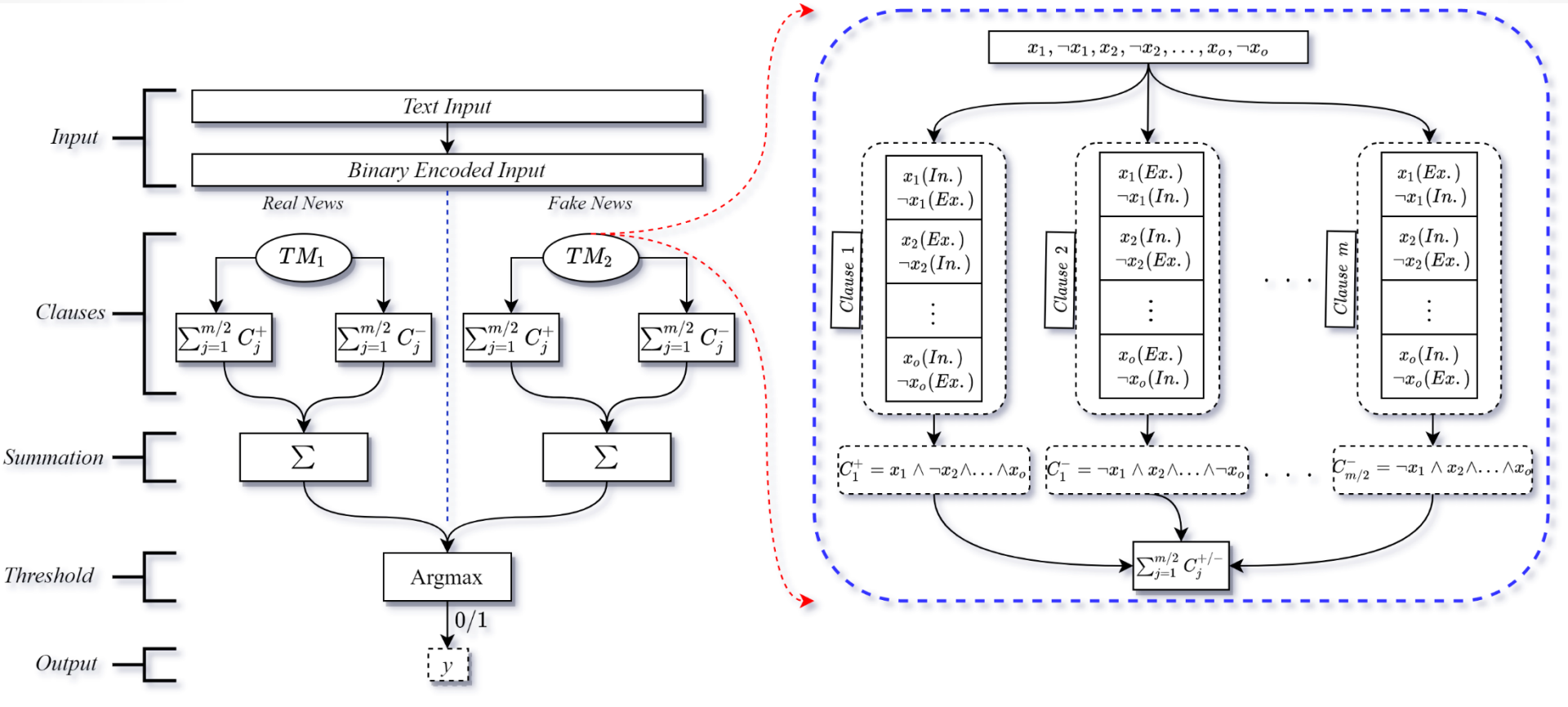
- Tsetlin Machine (TM) is a recent rule-based approach to solve tasks like pattern recognition and data regression.
- TM has promising properties regarding computational simplicity, transparency and interpretability, when compared to deep learning.
- TM has previously performed well in some natural language processing (NLP) applications.

TM clause formation



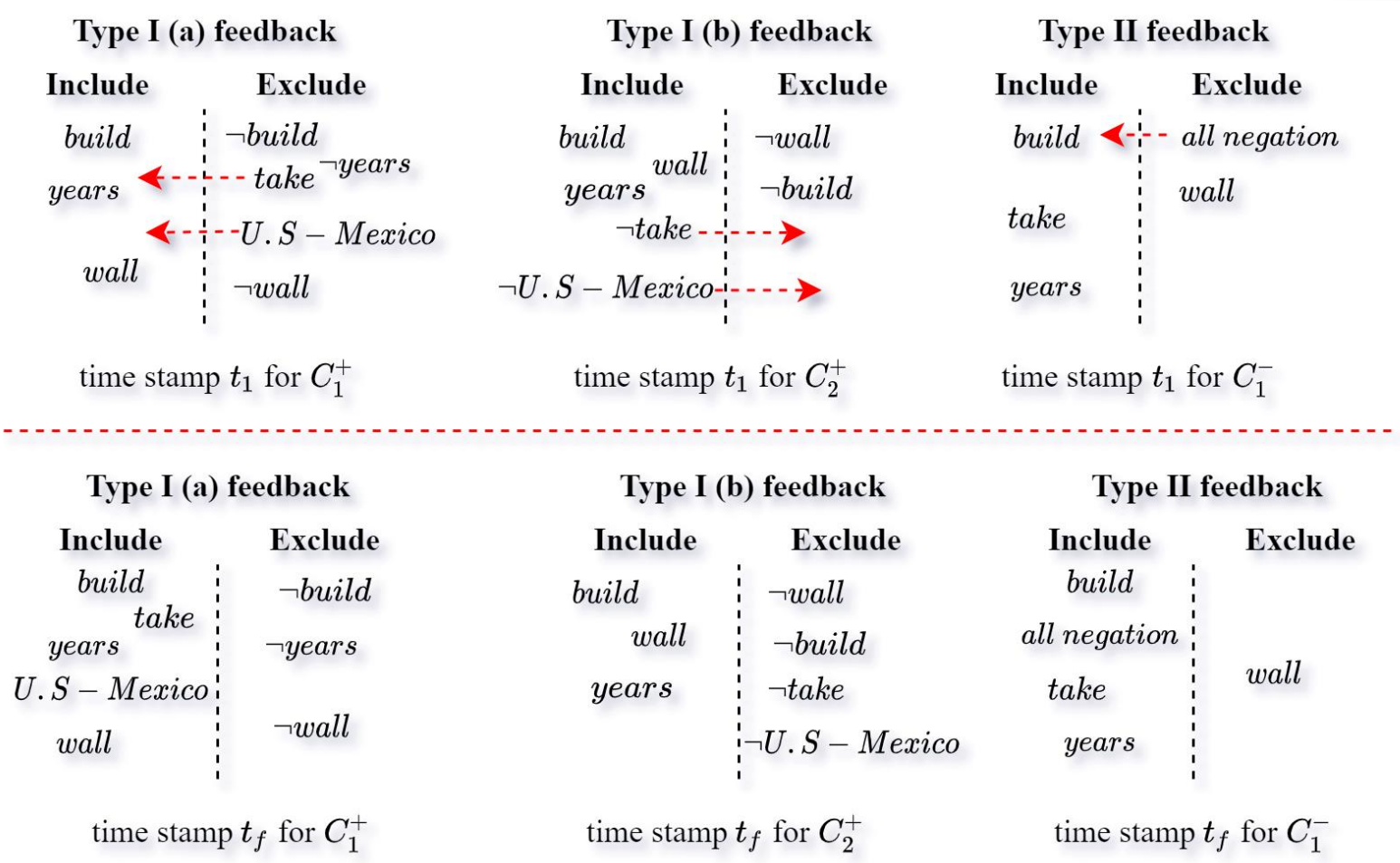


Explainable TM Framework



TM Learning

X= [“Building a wall on the U.S-Mexico border will take literally years.”], with output target “true news” i.e., y = 1.]



TM Formulation

Input $$X = (x_1, \ldots, x_o)$$

Literal set $$L = \{x_1, \ldots, x_o, \bar{x}_1, \ldots, \bar{x}_o\}$$

Clause formation $$C_j^+(X) = \bigwedge_{l_k \in L_j^+} l_k = \prod_{l_k \in L_j^+} l_k.$$

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

XOR case $$\hat{y} = u\left(x_1\bar{x}_2 + \bar{x}_1 x_2 - x_1 x_2 - \bar{x}_1\bar{x}_2\right)$$

Experiments

Dataset	#Real	#Fake	#Total
PolitiFact	563	391	954
GossipCop	15,338	4,895	20,233

Table 1: Dataset statistics.

Preprocessing:

- Includes tokenization, lemmatization and feature selection.
- Bag of Words approach.
- Chi-square test statistics as a feature selection technique.

Results

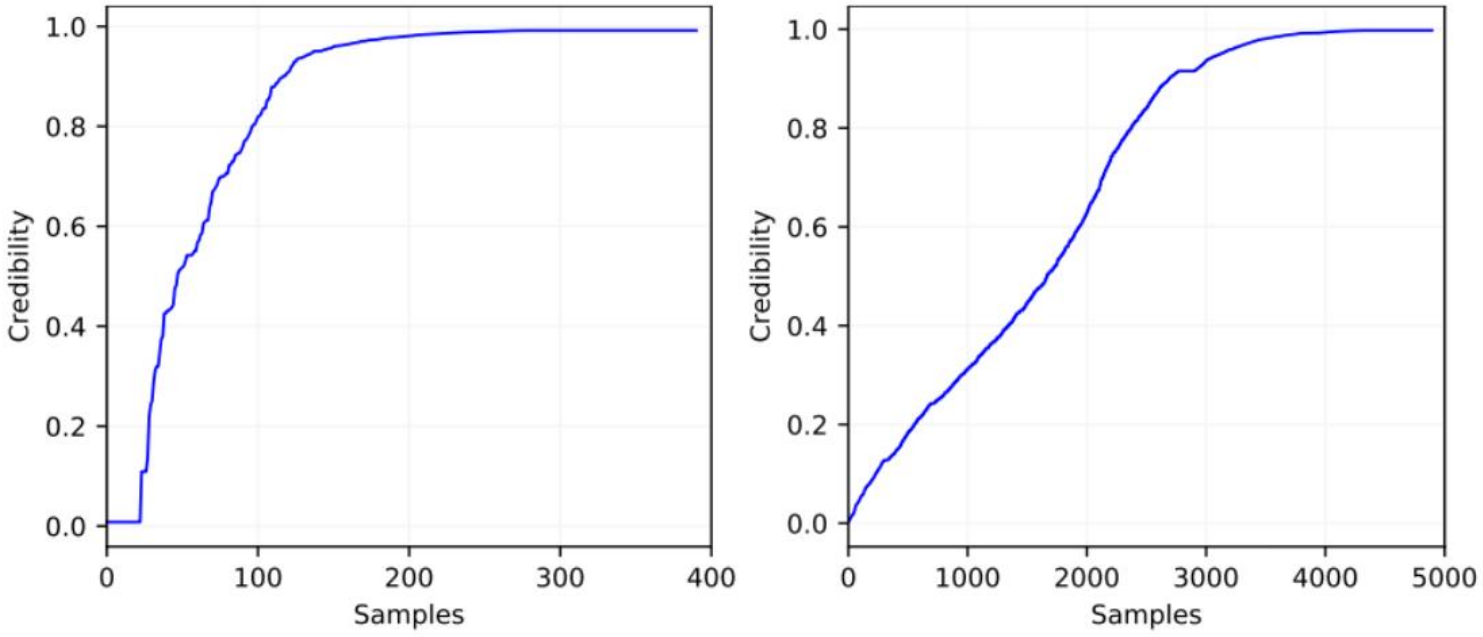
Table 2: Performance comparison of our model with 8 baseline models.

Models	PolitiFact		GossipCop	
	Acc.	F1	Acc.	F1
RST	0.607	0.569	0.531	0.512
LIWC	0.769	0.818	0.736	0.572
HAN	0.837	0.860	0.742	0.672
CNN-text	0.653	0.760	0.739	0.569
LSTM-ATT	0.833	0.836	0.793	0.798
LR	0.642	0.633	0.648	0.646
SVM	0.580	0.659	0.497	0.595
Naïve Bayes	0.617	0.651	0.624	0.649
RoBERTa-MWSS	0.825	0.805	0.803	0.807
BERT	0.88	0.87	0.85	0.79
XLNet	**0.895**	0.90	**0.855**	0.78
TM	0.871±0.24	**0.901 ± 0.001**	0.842 ±0.03	**0.896± 0.004**

Credibility Assessment

The credibility is defined as "the classification confidence of the TM"

$$\mathcal{Q}_i = \frac{1}{1 + \exp^{-k\left(v_i^F - v_i^T\right)}}.$$



(a) PolitiFact ($k = 0.012$). (b) GossipCop ($k = 0.02$).

Figure 3: Credibility assessment for fake news

Explainability

PolitiFact							
True				Fake			
Plain	times	Negated	times	Plain	times	Negated	times
trump	297	candidate	529	congress	136	trump	1252
said	290	debate	413	tax	104	profession	1226
comment	112	civil	410	support	70	navigate	1223
donald	110	reform	369	senate	64	hackings	1218
story	78	congress	365	president	60	reported	1216
medium	63	iraq	361	economic	57	arrest	1222
president	48	lawsuit	351	americans	49	camps	1206
reported	45	secretary	348	candidate	48	investigation	1159
investigation	38	tax	332	debate	44	medium	1152
domain	34	economy	321	federal	41	domain	1153

Table 3: Top ten Literals captured by clauses of TM for PolitiFact.

GossipCop							
True				Fake			
Plain	times	Negated	times	Plain	times	Negated	times
source	357	stream	794	season	150	insider	918
insider	152	aggregate	767	show	103	source	802
rumors	86	bold	723	series	79	hollywood	802
hollywood	80	refreshing	722	like	78	radar	646
gossip	49	castmates	721	feature	70	cop	588
relationship	37	judgment	720	video	44	publication	579
claim	33	prank	719	said	33	exclusively	551
split	32	poised	718	sexual	32	rumor	537
radar	32	resilient	714	notification	25	recalls	535
magazine	30	predicted	714	character	25	kardashian	525

Table 4: Top ten Literals captured by clauses of TM for GossipCop.

Explainability Comparison

Table 1: Top ten features captured by ELI5.

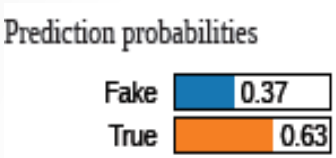
PolitiFact				GossipCop			
Fake		True		Fake		True	
Features	Weights	Features	Weights	Features	Weights	Features	Weights
trump	1.779	tax	0.757	source	4.343	season	2.758
president	0.685	health	0.606	insider	3.769	episode	1.744
domain	0.670	congress	0.560	hollywood	2.114	series	1.273
donald	0.534	senate	0.508	rumors	1.917	video	1.216
email	0.480	hotline	0.484	report	1.866	shared	1.105
meme	0.363	economy	0.421	radar	1.713	related	1.074
reported	0.369	americans	0.395	magazine	1.625	watch	1.016
story	0.365	energy	0.388	gossip	1.544	netflix	1.003
fake	0.347	reform	0.340	romance	1.506	like	0.958
investigation	0.345	iraq	0.269	claims	1.393	dress	0.958

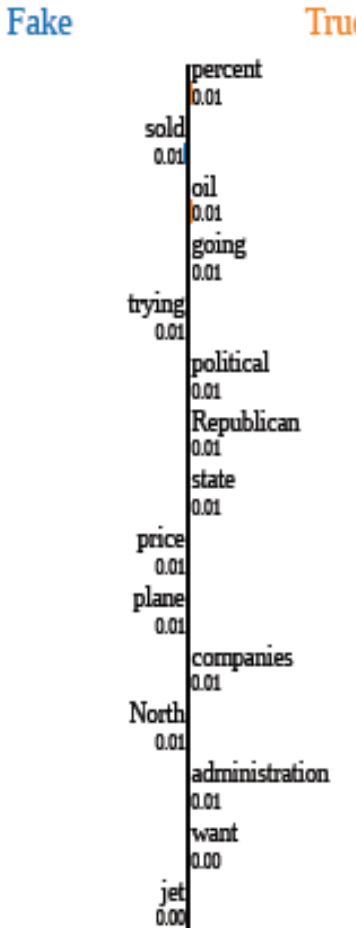
Table 2: Feature weights captured by TM from a single test instance.

PolitiFact				GossipCop			
Fake		True		Fake		True	
Features	Weights	Features	Weights	Features	Weights	Features	Weights
sold	1145	percent	366	fake	606	like	686
trying	1010	oil	259	president	602	video	767
price	1017	going	288	celebrities	653	images	935
North	1000	political	284	pictures	644	network	931
jet	1142	republican	299	worried	600	check	887
plane	1105	state	174	start	239	look	758
-	-	companies	257	Beyoncé	620	USA	1020
-	-	administration	235	networks	672	created	907
-	-	want	148	-	-	able	818

Explainability

PolitiFact - LIME





Text with highlighted words

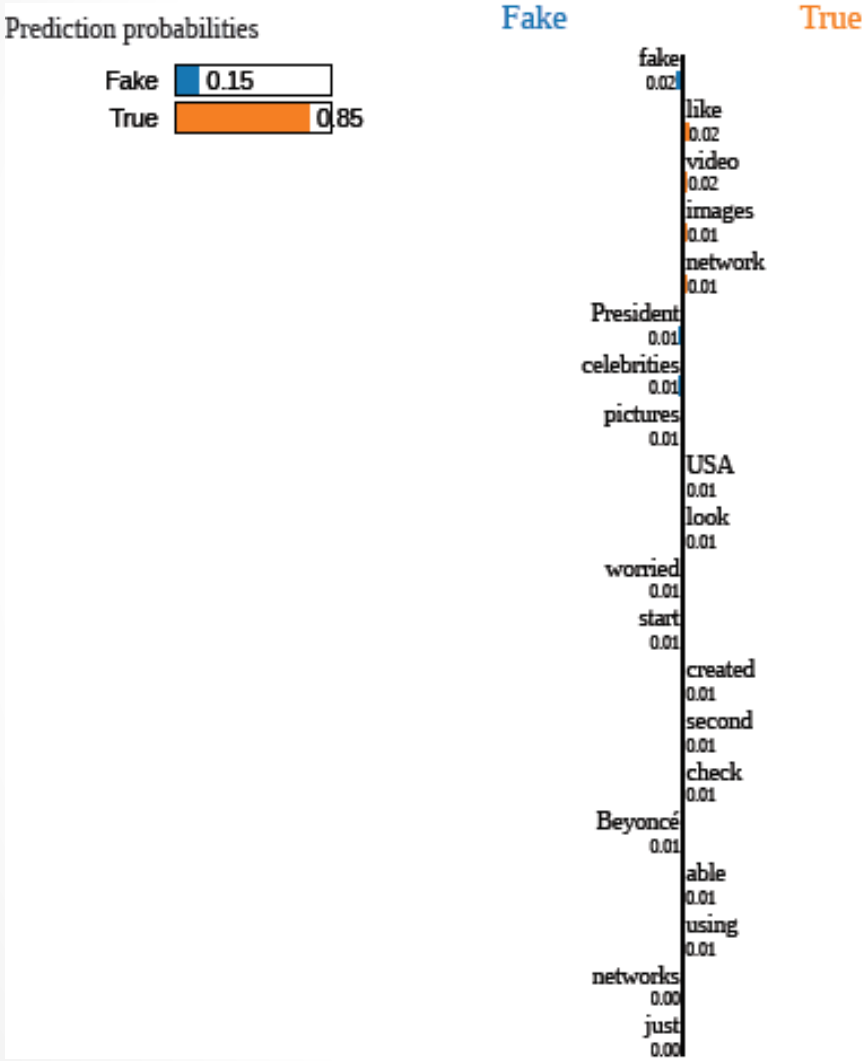
The state has tried selling its unwanted jet online four times and failed. So last week, the Palin administration signed a contract with an Anchorage aircraft broker who thinks he can succeed where eBay couldn't The eBay thing didn't work out very well, said Dan Spencer, director of administrative services for the Department of Public Safety. He's the person charged with trying to get rid of the infamous Westwind II.

The administration made a deal last week with Turbo North Aviation, promising the broker a 1.49 percent cut of the selling price.

Former Gov. Frank Murkowski bought the jet, which cost the state

Explainability

GossipCop - LIME



Conclusion

- Tsetlin Machine employs clauses to capture the lexical and semantic features based on word patterns in a document.
- The credibility assessment is performed for ranking fake news based on classification confidence.
- The explainability of model is highlighted using a case study.

CAIR CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH

Thank you!