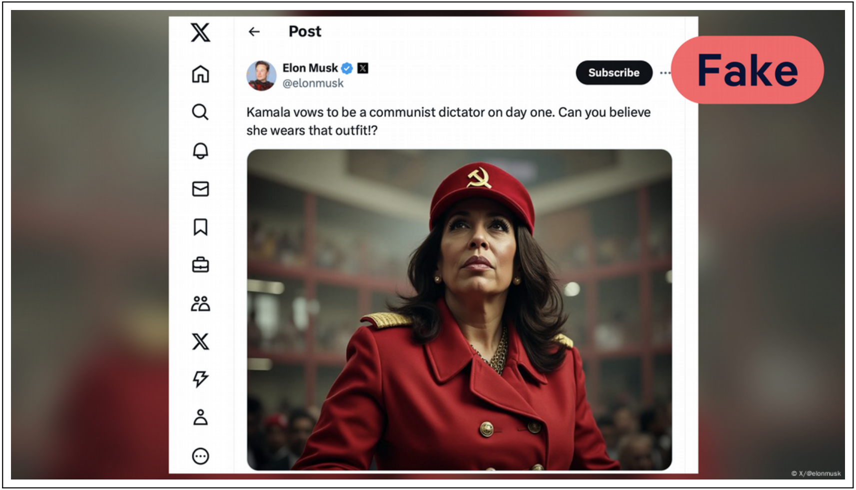
Adaptive Attacks on Misinformation Detection Using Reinforcement Learning

Piotr Przybyła
piotr.przybyla@upf.edu

Universitat Pompeu Fabra
Instytut Podstaw Informatyki PAN

December 19th 2024



Source: `https://www.dw.com/en/fact-check-what-role-did-disinformation-play-in-the-us-election/`

Credibility assessment as text classifiaction

How to build automatic content filtering?

- gather examples of both classes from the Internet,
- experts can provide credibility labels,
- use well-trodden framework of binary classification,
- deploy!

Applicable to credibility, but also inflammatory, violent, illegal content.

Examples:

- fake news [Przybyła, 2020],
- hoaxes [Kumar et al., 2016],
- bot-generated content [Rangel and Rosso, 2019],
- rumours [Han et al., 2019],
- false claims [Graves, 2018],
- hyperpartisan or biased reporting [Kiesel et al., 2019],
- propaganda techniques [da San Martino et al., 2020].

Universitat Pompeu Fabra Barcelona

Reuters World Business Markets Sustainability More

Technology

Exclusive: Twitter leans on automation to moderate content as harmful speech surges

By **Katie Paul** and **Sheila Dang**

December 6, 2022 4:41 AM GMT+7 · Updated 2 years ago

Aa

Source: `https://www.reuters.com/technology/twitter-exec-says-moving-fast-moderation-harmful-content-surges-2022-12-03/`

Adversarial scenario

Consider the following scenario:

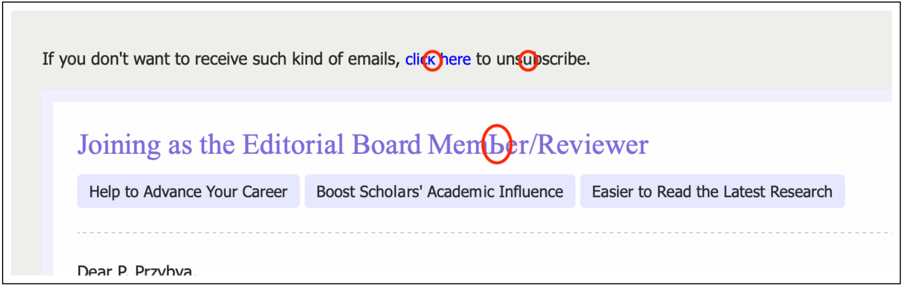
1. Social network $\mathbb{Y}$ uses content filtering predominantly based on ML,
2. Foreign state disseminates messages:
   Radioactive dust approaching after fire in a Ukrainian power plant! [Mierzyńska, 2020]
3. The message gets recognised as misleading and blocked.

What will the author do?

1. ~~Give up.~~
2. Try out different rephrasings until they found a variant thet gets through, e.g.:
   Radioactive dust *coming* *after fire in a Ukrainian power plant!*

→ **adversarial example**

Example


If you don't want to receive such kind of emails, click here to unsubscribe.

Joining as the Editorial Board Member/Reviewer

Help to Advance Your Career | Boost Scholars' Academic Influence | Easier to Read the Latest Research

Dear P. Przybyła,

Adversarial examples

Universitat Pompeu Fabra Barcelona

Let us define:

- Training set X_{train} and attack set X_{attack}, consisting of examples (x_i, y_i): features x_i and label y_i,
- Victim model f, predicting label $\hat{y}_i$ based on the example features: $\hat{y}_i = f(x_i)$,
- Modification function m, transforming x_i into adversarial example $x_i^* = m(x_i)$, guaranteeing:
  - change in victim's decision: $f(m(x_i)) \neq f(x_i)$,
  - preserving similarity to the original example: $m(x_i) \approx x_i$

Note: $y_i = 1$ for non-credible information and 0 for credible.

Research motivation

Universitat Pompeu Fabra Barcelona

Why do we want to look for adversarial examples?

- to assess the **robustness** of classifiers before their implementation in sensitive use-cases,
- to train more robust classifiers (**adversarial training**),
- for better understanding of the principles of the popular architectures.

$\rightarrow$ Find the vulnerabilities of the system **before** the malicious actors do!

Universitat Pompeu Fabra Barcelona

Why Adapt?

Attackers

Universitat Pompeu Fabra Barcelona

Most attackers, e.g. BERT-ATTACK [Li et al., 2020], work on the same principle:

1. Observe an input text x, e.g.
   $x =$ *Water causes death! 100%! Stop drinking now!*
   and the classifier response, e.g. $f(x) = 1$ // misinformation
2. Heuristically choose one token, e.g. *causes*
3. Make modification m by replacing it with a similar token according to a dictionary or language model or visual similarity, e.g. *provokes, inflicts, caυses, cause*
4. If $f(m(x)) \neq f(x)$: we have a success $\rightarrow$ proceed to next example
5. Otherwise, go back to step 2., unless all moves have been tried.

Examples

Universitat Pompeu Fabra Barcelona

Id., task, type	Original example	Adversarial example
EX1 PR Synonymous	Puerto Rico's housing secretary, Fernando Gil, says the number of **homes** destroyed by the hurricane totals about 70,000 so far, and homes with major damage have amounted to 250,000 across the island.	Puerto Rico's housing secretary, Fernando Gil, says the number of **houses** destroyed by the hurricane totals about 70,000 so far, and homes with major damage have amounted to 250,000 across the island.
EX2 FC Typographic	Sabbir Khan. Sabbir's second movie, Heropanti starring Tiger Shroff & Kriti Sanon**,** released on 23 may 2014. → Sabbir Khan directed a movie.	Sabbir Khan. Sabbir's second movie, Heropanti starring Tiger Shroff & Kriti Sanon**?** released on 23 may 2014. → Sabbir Khan directed a movie.
EX3 PR Grammatical	Fastiggi and Goldstein have managed to make the problem even worse in their attempt to **explain** it away.	Fastiggi and Goldstein have managed to make the problem even worse in their attempt to **explained** it away.

[Przybyła et al., 2023]

Adaptive attacks

upf. Universitat Pompeu Fabra Barcelona

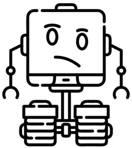
- Instead of forgetting successes and failures between examples, let's learn from them.
- This will allow us to have better AEs later.
- Reflects long-term nature of misinformation spreaders, i.e. Russia's *Internet Research Agency* (see photo).

Universitat Pompeu Fabra Barcelona

How to Adapt?

Reinforcement Learning

→ RL [Sutton and Barto, 2018] is a process, in which a model (*agent*) learns an optimal behaviour (*policy*) in an *environment* by performing *actions* and observing results (*rewards*).
→ The goal is to find a strategy that maximises the received profits (minimise losses).



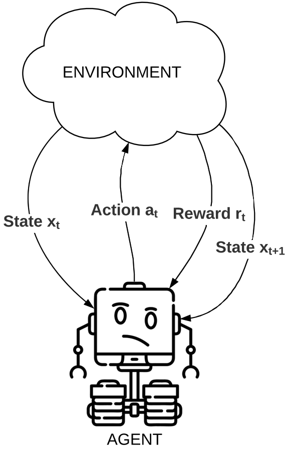
Universitat Pompeu Fabra Barcelona

RL procedure

- for all time steps $t = 0 \ldots \infty$:
  - observe current state $x_t \in \mathcal{X}$,
  - perform action $a_t \in \mathcal{A}$,
  - observe reward $r_t \in \mathbb{R}$ and next state x_{t+1},
  - learn from experience $\langle x_t, a_t, r_t, x_{t+1} \rangle$

How to define policy π that dictates action $a_t = \pi(x_t)$?

How to learn from experience?



Q-learning

Universitat Pompeu Fabra Barcelona

In Q-learning [Watkins, 1989], the Q function expresses the discounted (with factor γ) reward from taking action a in state x and then following policy π:

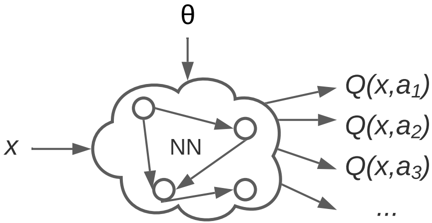
$$Q^{\pi}(x, a) = \mathbb{E}_{\pi}\left[\rho(x, a) + \sum_{t=1}^{\infty} \gamma^{t} r_{t} | x_0 = x, a_0 = a\right]$$

Knowing Q, we can perform the greedy policy with respect to it:

$$\forall_x \pi(x) = \arg\max_a Q(x, a)$$

Fitted Q-learning

How can we know Q? We can approximate it with a neural network [François-Lavet et al., 2018]:



Update rule:

- $Q'(x_t, a_t) = r_t + \gamma \max_a Q(x_{t+1}, a)$
- $L_Q = [Q'(x_t, a_t) - Q_t(x_t, a_t)]^2$
- use the loss L to update the weights θ

Universitat Pompeu Fabra Barcelona

What is XARELLO?

XARELLO

XARELLO (eXploring Adversarial examples using REinforcement Learning Optimisation) consists of:

- **Environment**, mapping the AE search into RL steps,
- **Optimiser**, a neural network for estimating $Q(s, a)$,
- **Attacker**, using Q values to choose a sequence of steps making up adversarial examples.

Unrelatedly, **Xarel·lo** is also a grape variety used in great Catalan wines.

Image source: https://www.cava.wine/en/origin-cava/authorised-grape-varieties/

Environment

Universitat Pompeu Fabra Barcelona

- environment state s includes:
  - $x_{i,j}^{(t)}$ – the current form (in step t) of the i-th target text,
  - $f(x_i)$ – victim's decision for the original text.
- action $a = (j, k)$ with the positions of the changed token j and the replacement candidate z_k from a pre-computed list $z_1, z_2, \ldots z_K$.
- reward r:
  - 1, if victim changed its decision,
  - -1, if attempting to modify a non-word token,
  - otherwise, $[f_p(x_i^{(t)}) - f_p(x_i^{(t-1)})] \times [1 - 2 \times f(x_i)]$.

Universitat Pompeu Fabra Barcelona

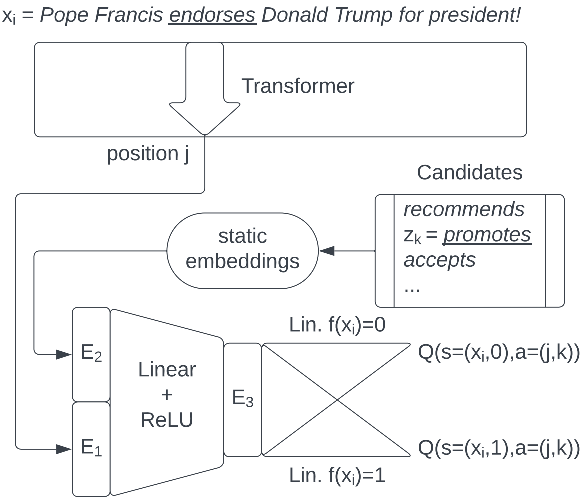
Example episode

1. state $s = (x^{(0)}_{15,j} = \textit{Drinking orange juice causes DEATH!}, f(x_{15}) = 1)$
2. action $a_0 = (j \sim \textit{causes}, k \sim \textit{provokes})$
3. state $s = (x^{(1)}_{15,j} = \textit{Drinking orange juice provokes DEATH!}, f(x_{15}) = 1)$
4. reward $r_0 = 0.15$ (P(MISINFO): 75%$\rightarrow$ 60%)
5. action $a_1 = (j \sim \textit{Drinking}, k \sim \textit{Consuming})$
6. state $s = (x^{(2)}_{15,j} = \textit{Consuming orange juice provokes DEATH!}, f(x_{15}) = 1)$
7. reward $r_1 = -0.08$ (P(MISINFO): 60%$\rightarrow$ 68%)
8. action $a_2 = (j \sim \textit{provokes}, k \sim \textit{brings})$
9. state $s = (x^{(3)}_{15,j} = \textit{Consuming orange juice brings DEATH!}, f(x_{15}) = 1)$
10. reward $r_2 = 1$ (P(MISINFO): 68%$\rightarrow$ 47%)

Adversarial example was found!

Optimiser

Universitat Pompeu Fabra Barcelona



Attacker

In **adaptation** phase:

- in each *episode*, the attacker can make max 5 *steps* (changes),
- for each text example, 5 episodes are performed,
- all text examples are processed for 20 *epochs*,
- to encourage exploration:
  - with probability ϵ, a random action is chosen,
  - with probability $1 - \epsilon$, a greedy action is chosen,
  - *epsilon* falls from 100% to 10% during the warmup (30% of epochs)

In **attack** phase:

- always the greedy action is chosen,
- optimiser is frozen,
- for each text, episodes of increasing lengths are performed (10 e. of 5 s, 5 e. of 10 s, 2 e. of 25 s, 1 e. of 50 s)

In training, the memory of previous experiences [Mnih et al., 2015] is used.

Universitat Pompeu Fabra Barcelona

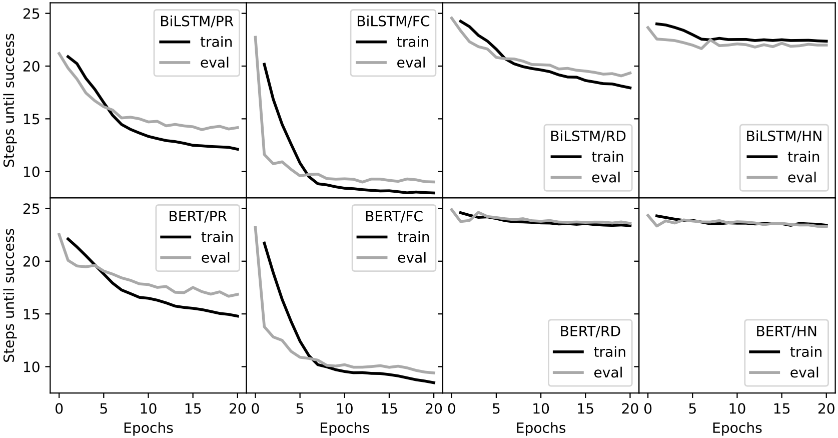
Does It Work?

Evaluation schema

- Using BODEGA framework, four misinformation detection tasks: → News bias assessment (HN), Propaganda detection (PR), Fact checking (FC), Rumour detection (RD),
- Three victim classifiers: BiLSTM, BERT and GEMMA-2B,
- Measures of success:
  - Confusion score (1 if AE found, 0 otherwise),
  - Semantic similarity score (0-1) using BLEURT [Sellam et al., 2020],
  - Character similarity score (0-1) using edit distance [Levenshtein, 1966],
  - BODEGA score, product of the above,
  - Number of queries to victim classifier
- Baselines: *DeepWordBug* [Gao et al., 2018], *BERT-ATTACK* [Li et al., 2020] and XARELLO raw (without adaptation).

See `https://github.com/piotrmp/BODEGA` and [Przybyła et al., 2023].

Adaptation process



Results: propaganda

Measure	Victim: BiLSTM DWB	B-A	XARELLO raw	XARELLO full	Victim: BERT DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.292	0.527	0.466	**0.632**	0.278	0.429	0.360	**0.512**
conf.	0.382	0.800	0.928	0.990	0.363	0.697	0.769	0.962
sem.	0.795	0.716	0.595	0.698	0.794	0.678	0.562	0.606
char.	0.960	0.914	0.791	0.884	0.962	0.902	0.772	0.834
queries	27.4	61.4	61.4	**15.0**	**27.4**	80.2	89.8	30.2

Measure	Victim: GEMMA DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.143	0.460	0.474	**0.697**
conf.	0.190	0.724	0.899	0.986
sem.	0.786	0.695	0.605	0.748
char.	0.958	0.906	0.813	0.920
queries	27.3	77.5	59.5	**14.9**

Note: ordering.

Results: fact-checking

Measure	Victim: BiLSTM DWB	B-A	XARELLO raw	XARELLO full	Victim: BERT DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.484	0.598	0.640	**0.817**	0.440	0.535	0.559	**0.773**
conf.	0.575	0.857	0.938	1.000	0.531	0.770	0.862	0.995
sem.	0.855	0.728	0.733	0.837	0.843	0.726	0.708	0.800
char.	0.984	0.954	0.917	0.975	0.982	0.953	0.902	0.970
queries	54.4	132.8	56.0	**5.0**	54.3	146.7	74.1	**7.4**

Measure	Victim: GEMMA DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.074	0.566	0.577	**0.775**
conf.	0.091	0.832	0.904	0.995
sem.	0.829	0.718	0.698	0.802
char.	0.983	0.939	0.902	0.969
queries	53.9	192.2	66.3	**7.3**

Note: query numbers.

Results: rumour detection

	Victim: BiLSTM				Victim: BERT			
			XARELLO				XARELLO	
Measure	**DWB**	**B-A**	**raw**	**full**	**DWB**	**B-A**	**raw**	**full**
BODEGA	0.164	0.292	0.244	**0.650**	0.159	0.181	0.145	**0.227**
conf.	0.243	0.790	0.537	0.973	0.229	0.439	0.333	0.436
sem.	0.682	0.409	0.514	0.694	0.701	0.429	0.500	0.580
char.	0.991	0.890	0.842	0.957	0.991	0.961	0.830	0.870
queries	232.8	985.5	617.8	**84.0**	**232.7**	774.3	763.5	631.7

	Victim: GEMMA			
			XARELLO	
Measure	**DWB**	**B-A**	**raw**	**full**
BODEGA	0.104	0.300	0.228	**0.314**
conf.	0.152	0.725	0.434	0.492
sem.	0.694	0.433	0.590	0.678
char.	0.991	0.951	0.865	0.934
queries	**239.0**	703.1	665.7	538.9

Note: model ordering.

Results: hyperpartisan news

Measure	Victim: BiLSTM DWB	B-A	XARELLO raw	XARELLO full	Victim: BERT DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.406	**0.636**	0.496	0.612	0.223	**0.601**	0.340	0.341
conf.	0.527	0.980	0.760	0.848	0.287	0.965	0.560	0.583
sem.	0.771	0.656	0.689	0.737	0.777	0.638	0.644	0.607
char.	0.998	0.988	0.933	0.975	0.998	0.972	0.918	0.937
queries	396.2	487.9	445.7	**256.1**	**395.9**	648.4	599.8	564.4
Avg: BODEGA	0.337	0.513	0.461	**0.678**	0.275	0.436	0.351	**0.463**
queries	177.7	416.9	295.2	**90.0**	**177.6**	412.4	381.8	308.4

Measure	Victim: GEMMA DWB	B-A	XARELLO raw	XARELLO full
BODEGA	0.240	**0.546**	0.485	0.528
conf.	0.307	0.905	0.752	0.757
sem.	0.783	0.622	0.676	0.715
char.	0.998	0.965	0.930	0.963
queries	385.9	943.0	427.7	**373.6**
Avg: BODEGA	0.141	0.468	0.441	**0.578**
queries	**176.5**	478.9	304.8	233.7

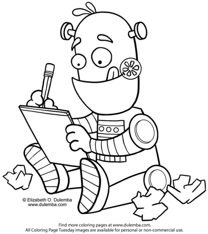
Note: large search space task

Universitat Pompeu Fabra Barcelona

Qualitative analysis

Manual analysis of changes that XARELLO has learnt to make:

- Changing sub-word tokens, resulting in non-words with graphical similarity to originals
  vocations → *vassations*,
  hypocritically → *hypoclipically*
- Replacing emotionally charged fragments with more general words
  his aggressive behaviour → *his own behaviour*,
  type of injustice → *type of work*
- Often failing to preserve grammatical structure
  spread and worsen → *slow and badn*,
  reported on a gaping hole in → *reported on a in*



Limitations

- Attacker failing to learn enough in modifying news articles,
  → large search space requires a different design of adaptation phase,
- Simple *action* model: only single word-replacements considered to reduce search space,
  → more complex operations can be included, such as deletions [Garg and Ramakrishnan, 2020] or multi-word replacements [Przybyła et al., 2025]
- RL process requires many parameters, only some were tested in evaluation,
  → easy to expand, but the computational cost will be substantial,
- Only specific tasks were tested, but the setup is applicable to any text classification,
  → other misinformation-detection tasks, content filtering (hate speech), other languages etc.

Thank you!

The work was possible thanks to co-authors:
→ Euan McGill and Horacio Saggion from UPF TALN,

and funders:
→ Marie Skłodowska-Curie Postdoctoral Fellowship programme,
→ Polish high-performance computing infrastructure PLGrid (ACK Cyfronet AGH)

This work is part of the ERINIA project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101060930. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

More about ERINIA at `https://www.upf.edu/web/erinia` and XARELLO at `https://doi.org/10.18653/v1/2024.wassa-1.11`


Thank you!

References (1)

Universitat Pompeu Fabra Barcelona

Piotr Przybyła. Capturing the Style of Fake News. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, volume 34, pages 490–497, New York, USA, 2020. AAAI Press. doi: 10.1609/aaai.v34i01.5386. URL https://aaai.org/ojs/index.php/AAAI/article/view/5386.

Srijan Kumar, Robert West, and Jure Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *25th International World Wide Web Conference, WWW 2016*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016. ISBN 9781450341431. doi: 10.1145/2872427.2883085. URL https://dl.acm.org/doi/10.1145/2872427.2883085.

Francisco Rangel and Paolo Rosso. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

Sooji Han, Jie Gao, and Fabio Ciravegna. Neural language model based training data augmentation for weakly supervised early rumor detection. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, pages 105–112. Association for Computing Machinery, Inc, 2019. ISBN 9781450368681. doi: 10.1145/3341161.3342892. URL https://dl.acm.org/doi/10.1145/3341161.3342892.

Lucas Graves. Understanding the Promise and Limits of Automated Fact-Checking. Technical report, Reuters Institute, University of Oxford, 2018. URL https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves{_}factsheet{_}180226FINAL.pdf.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 Task 4: Hyperpartisan News Detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL https://aclanthology.org/S19-2145.

References (2)

Universitat Pompeu Fabra Barcelona

Giovanni da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-2020)*, pages 1377–1414, 2020. URL http://propaganda.qcri.org/annotations/definitions.htmlhttp://arxiv.org/abs/2009.02696.

Anna Mierzyńska. Chmura znad Czarnobyla - kolejna dezinformacja, którą straszono Polaków. Wiemy, skąd się wzięła, 2020. URL https://oko.press/radioaktywna-chmura-znad-czarnobyla-kolejna-dezinformacja-ktora-straszono-polakow.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500.

Piotr Przybyła, Alexander Shvets, and Horacio Saggion. Verifying the Robustness of Automatic Credibility Assessment. *arXiv preprint arXiv:2303.08032*, mar 2023. URL https://arxiv.org/abs/2303.08032v1.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

C.J.C.H. Watkins. *Learning from Delayed Rewards*. PhD thesis, University of Cambridge, 1989. URL https://www.cs.rhul.ac.uk/{~}chrisw/new{_}thesis.pdf.

Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, and Joelle Pineau. An Introduction to Deep Reinforcement Learning. *Foundations and Trends in Machine Learning*, 11(3-4): 219–354, 2018. ISSN 1935-8237. doi: 10.1561/2200000071. URL http://dx.doi.org/10.1561/2200000071.

References (3)

Universitat Pompeu Fabra Barcelona

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature 2015 518:7540*, 518(7540): 529–533, feb 2015. ISSN 1476-4687. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *Proceedings - 2018 IEEE Symposium on Security and Privacy Workshops, SPW 2018*, pages 50–56. IEEE, 2018. ISBN 9780769563497. doi: 10.1109/SPW.2018.00016.

Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.498. URL https://aclanthology.org/2020.emnlp-main.498.