A Multi-Criteria Complexity Evaluation of Cyberattack Datasets in Industrial Control Systems

Saddam Hussain, Ali Tufail, Abul Ghani Haji Naim
School of Digital Science, Universiti Brunei Darussalam
21h8564@ubd.edu.bn, ali.tufail@ubd.edu.bn, ghani.naim@ubd.edu.bn

Abstract

The complexity of Industrial Control Systems (ICS) datasets plays a crucial role in determining effective detection strategies. The paper introduces a multi-dimensional and formal technique of measuring the complexity of a dataset, depending on twelve complexity measures in the dimensions of features-based, neighborhood-based, linearity-based, and topological measures. The measures enable class separability, local ambiguity, and complexity of decision boundaries to be evaluated in a classifier-independent manner, offering valuable insight into the structural and statistical challenges of ICS data. Additionally, we further employ Evaluation based on Distance from the Average Solution (EDAS), a Multi-criteria Decision-Making (MCDM) technique, that allows positive as well as negative deviations in the average performance for the ranking and comparison of datasets in terms of their intrinsic complexity. Findings show a lot of differences in complexity between the datasets benchmark and CISS2019.A1, where the most separable datasets are Dataset 8.3 and Dataset 7.3, and the most difficult ones to classify are Dataset 2.1 and CISS2019.A1(4).

1 Introduction

The severity of ICS security concerns has grown in terms of sophistication and frequency of cyber attacks and potential of cyber intrusion causing digital harm, and the most potent harm caused by Colonial Pipeline ransomware (1). Modern ICS systems are no longer stand-alone systems because they have been connected to external networks, corporate IT systems, and mobile interfaces, which makes them very vulnerable to attack. The synergy between operational technology and information technology overhead has greatly compromised the boundaries of traditional security. The development of automation, remote control and artificial intelligence gives organizations substantial strategic benefits; however, these innovations also threaten security in a very serious manner (2).

Machine learning is becoming one of the key tools against such threats, leaving it with the potential to detect cyberattacks in the ICS environment. The Secure Water Treatment (SWaT) system (3) and the Critical Infrastructure Security Showdown (CISS) (4) testbed generate realistic, labeled datasets for training and evaluating ML systems. However, when studying the field of cyberattack detection on the SWaT system, one of the key issues that occurs repeatedly is that a significant number of ML classifiers face a ceiling in performance where it is hard to improve them, despite the adoption of more advanced methods. This limitation is partly attributed to the intrinsic complexity of the data itself. The major contributors are class imbalances, overlapping features, non-linearity, and local distributional shifts (5). These factors, which are common in ICS logs and sensor input, can also restrict the modern classifiers from delivering better performance. The major contributions of this work are listed below:

- The fundamental contribution of this work is to compute the complexity of ICS datasets, SWaT and CISS.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- This study employed four measurements for complexity analysis, such as Features, Neighbourhood, Topological, and Linearity-based.
- An advanced ranking-based MCDM technique used to rank the datasets based on complexity.
- A total of 38 datasets are used and compared their complexities progressively.

2 Methodology

In this section, we present the methodology used to compute the complexity of a dataset using a wide range of measures of data separability and structure. The datasets are followed by the computation of twelve complexity measures categorized into feature-based $(D1 - D4)$, neighbourhood-based $(M1 - M4)$, linearity-based $(P1 - P3)$, and topological $(Q1)$ metrics (6; 7). The steps of each of the measures of complexity are explained below in detail and with mathematical formulations. For more clarification, the proposed methodology is graphically reported in Figure 1.

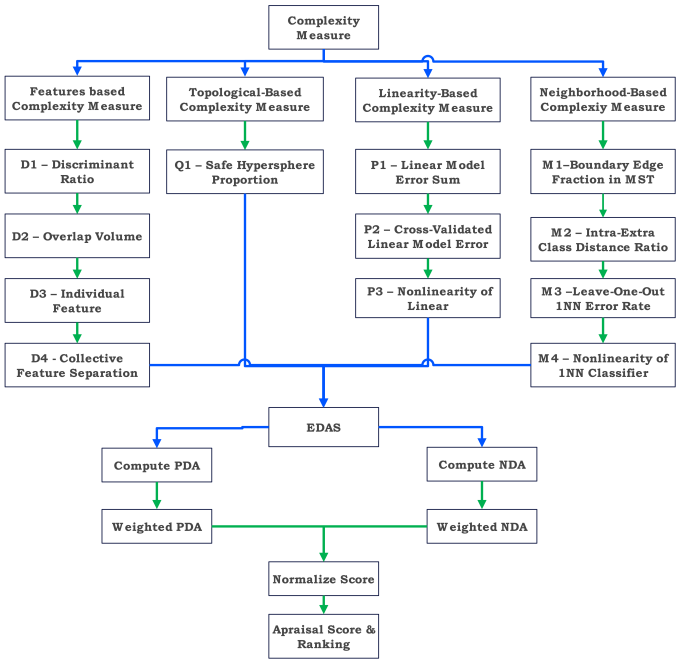

Figure 1: Proposed Methodology

2.1 Feature-Based Complexity Metrics:

These matrices assess the discriminative ability of particular features or combinations to separate classes. Here we have listed all the features-based on measurements. ***D1*** *– Discriminant Ratio,* for feature j is:

$$D1_j = \frac{\sum_k s_k (m_{j,k} - m_j)^2}{\sum_k \sum_{z_{kj} \in k} (z_{kj} - m_{j,k})^2}, \tag{1}$$

D2 *– Overlap Volume:* Let $[u_1, v_1]$ and $[u_2, v_2]$ represent the feature ranges for two distinct classes along a given dimension. The range of overlap between these two ranges is defined to be:

$$D2_j = \frac{\max(0, \min(v_1, v_2) - \max(u_1, u_2))}{\max(v_1, v_2) - \min(u_1, u_2)}, \quad \text{final metric aggregates} \quad D2 = \prod_{j=1}^{m} D2_j \tag{2}$$

D3 *– Individual Feature Efficiency:* Let $S_j = [\min(k_1), \max(k_1)] \cap [\min(k_2), \max(k_2)]$, where S_j denotes the interval in which the two classes k_1 and k_2 overlap for feature j.

$$D3_j = \frac{\text{count}(z_j < \min(S_j) \vee z_j > \max(S_j))}{T}, \quad \text{Final score selector} \quad D3 = \max_j(D3_j) \tag{3}$$

D4 *- Collective Feature Separation:* The joint discriminative power of several features is estimated via (D4) which repeatedly selects the features with the best separation of the instances into non-overlapping regions.

$$D4 = \frac{\text{count(Separated Points)}}{T} \tag{4}$$

2.2 Neighbourhood-Based Complexity Metrics:

These measures serve to explore local relations of instances to their most similar neighbours that can be indicative of boundary ambiguity and decision complexity.

***M1**–Boundary Edge Fraction in MST:*

Let B_e be the number of such edges, while M2 and M3 represent the intra/extra classes distance ratio (Intra-Extra Class Distance Ratio) and the Leave-One-Out 1NN Error rate, respectively. Each of them is formalized as follows:

$$M1 = \frac{B_e}{T-1}, \quad M2 = \frac{\sum_p g_{\text{between}}(z_p)}{\sum_p g_{\text{within}}(z_p)}, \quad M3 = \frac{1}{T}\sum_{p=1}^{T} \mathbb{I}(w_p \neq \arg\min_{q \neq p} \|z_p - z_q\|). \tag{5}$$

***M4** – Nonlinearity of 1NN Classifier:* For the purpose of measuring the complexity of the decision boundary learned by the 1NN classifier, we sample interpolated points between pairs of same class which are randomly sampled.

$$M4 = \frac{\text{count}(\hat{w}_{\text{interp}} \neq w_{\text{interp}})}{U} \tag{6}$$

where U is the number of generated interpolations. High values indicate nonlinear or fragmented decision boundaries, reflecting increased structural complexity in the data.

2.3 Linearity-Based Complexity Metrics:

In this section, we see how the linear models capture the underlying data patterns. ***P1** – Linear Model Error Sum*, ***P2** – Cross-Validated Linear Model Error*, ***P3** – Nonlinearity of Linear*. A linear model is trained on the entire dataset, with its predictions $\hat{w}_p$ compared against the true labels w_p. This gives us a total error as:

$$P1 = \sum_{p=1}^{T} |w_p - \hat{w}_p|, \quad P2 = \frac{1}{K}\sum_{k=1}^{K} \frac{\text{count}(\hat{w}_k \neq w_k)}{T_k}, \quad P3 = \frac{\text{count}(\hat{w}_{\text{interp}} \neq w_{\text{interp}})}{U} \tag{7}$$

2.4 Topological-Based Complexity Metric:

The metric uses the local structure of the data by examining the hyperspheres around each point and determines the relative safety of these regions when it comes to the consistency of the classes.
***Q1** – Safe Hypersphere Proportion:*

$$\forall q : \|z_q - z_p\| < t_p \Rightarrow w_q = w_p, \text{ Final metric } Q1 = \frac{1}{T}\sum_{p=1}^{T} \mathbb{I}(\text{hypersphere around } z_p \text{ is safe}) \tag{8}$$

2.5 Evaluation Based on Distance from Average Solution (EDAS)

In order to rank the datasets based on the measures of complexity, the EDAS technique is used (6; 8), where the datasets are ranked according to their distance to the average solution across all measures.

EDAS is an MCDM technique that ranks alternatives by taking their distance from an average solution (9). To compute average, PDA and NDA, the following mathematical equation can be used:

$$AV_j = \frac{1}{k}\sum_{i=1}^{k} m_j(D_i), \; PDA_{ij} = \frac{|m_j(D_i) - AV_j|}{AV_j}, \; NDA_{ij} \qquad = \frac{|m_j(D_i) - AV_j|}{AV_j} \tag{9}$$

While weighted PDA, NDA and their normalized form can be computed as:

$$SP_i = \sum_{j=1}^{12} w_j \cdot PDA_{ij}, \; SN_i = \sum_{j=1}^{12} w_j \cdot NDA_{ij}, \; NSP_i = \frac{SP_i}{\max(SP)}, \; NSN_i = 1 - \frac{SN_i}{\max(SN)} \tag{10}$$

Calculate the Final EDAS Score: Combine the normalized scores to get one EDAS score:

$$EDAS_i = \frac{NSP_i + NSN_i}{2} \tag{11}$$

This final score reflects a balanced evaluation of how far and in what direction each dataset deviates from the average across all complexity measures.

Rank Datasets: Sort the datasets in descending order of their EDAS scores:

$$\text{Ranking} = \text{argsort}(-EDAS_i) \tag{12}$$

By using EDAS, we obtain a more robust and nuanced ranking of datasets based on their complexity profiles, as it accounts for both how far and in what direction each dataset deviates from the average across all complexity measures.

3 Results and Discussion

Before the final results given in Table 1, we computed the complexity of all datasets using a comprehensive set of twelve data complexity measures, categorized into feature-based (D1–D4), neighbourhood-based (M1–M4), linearity-based (P1–P3), and topological (Q1) metrics. Feature-based measures assess discriminative power through the discriminant ratio (D1), overlap volume (D2), individual feature efficiency (D3), and collective feature separation (D4). Neighbourhood-based metrics capture local structure via boundary edge fraction in MST (M1), intra-extra class distance ratio (M2), 1NN leave-one-out error (M3), and 1NN nonlinearity (M4). Linearity-based measures evaluate the suitability of linear models through linear model error sum (P1), cross-validated linear error (P2), and nonlinearity of linear classifier (P3). The topological measure (Q1) evaluate local homogeneity via the safe hypersphere proportion.

Table 1 presents datasets ranked using PDA and weighted PDA (WPDA), where a lower numerical rank reflects stronger overall performance, i.e., lower complexity. At the top, Dataset 8.3 and CISS2019.A1 (2) both achieve Rank 1, driven by consistently high scores in key metrics such as D1, M3, and P1. These datasets show low values in complexity indicators, which indicates that there is high class separation, low 1NN error, and linear models are well-fitted.

Close behind are datasets such as 7.3, 10.2 and 8.2, which also have good performance in several measures related to M- and P- but are slightly lower in PDA values than the leaders.

On the other hand, datasets with the lower ranks, especially from Rank 20 to Rank 35, make little contribution on most of the metrics, implying weaker performance on PDA and weighted PDA evaluations. These datasets display high values in D1, M1, M2, M3, and P2 exposing poor discriminability of the features, dense decision boundaries, local heterogeneity, and poor generalization of linear models. WPDA essentially combines these multidimensional insights, and provides an understandable score for comparison between datasets.

4 Conclusion

This study offers a multidimensional framework for evaluating the intrinsic complexity of ICS cyberattack detection data sets. It combines twelve proven complexity measures of discriminability of features, neighbourhood structure, linearity, and topology with EDAS based aggregation with Weighted Positive and Negative Distance from Average (WPDA and WNDA) to allow an interpretable evaluation of the suitability of a dataset. Results show that Dataset 8.3 and CISS2019.A1 (2) achieves Rank 1, which shows good class separability, low 1NN error and high linear fidelity, which are good for detection modeling. On the other hand, Dataset 2.1 and CISS2019.A1 (4) are the most difficult, because of the poor feature discriminability, dense decision boundaries, neighbourhood ambiguity and weak linear generalization. These insights are useful to select datasets and address the need for specific preprocessing or advanced modeling in industrial cybersecurity.

Acknowledgment

The authors thank "iTrust, Centre for Research in Cyber Security, Singapore University of Technology and Design for providing the CISS2019 and SWaT2021 datasets".

Table 1: PDA and Weighted PDA Calculations Sorted by Final Rank

Datasets	D1	D3	D4	Q1	D2	M1	M2	M3	M4	P1	P2	P3	Rank
Dataset 8.3	15.5017	0.0189	0.0002	0	0	0.7651	0	1	1	1	0.2833	1	1
Dataset 7.3	17.019	0.0189	0.0002	0	0	0.7651	0	1	0.8943	1	0	1	2
Dataset 10.2	0	0.0073	0.0002	0	0	0.1387	0.4206	0.4142	0	0	0.4563	0	3
Dataset 8.2	0	0.0153	0.0002	0	0	0.1387	0.3153	0.1946	0	0	0.6965	0	4
Dataset 7.4	0	0.0062	0.0002	0	0	0.1387	0.3416	0.4142	0	0	0.2709	0	5
Dataset 3.4	0	0.0153	0	0	0	0.1387	0.7893	0	0	0	0.248	0	6
Dataset 2.5	0	0	0.0002	0	0	0.1387	0.1309	0.4142	0	0	0.5946	0	7
Dataset 6.2	0	0.0031	0.0002	0	0	0.1387	0.0783	0.4142	0	0	0.3617	0	8
Dataset 4.5	0	0	0.0002	0	0	0.3736	0.0519	0.5607	0	0	0	0	9
Dataset 2.2	0	0.0124	0.0002	0	0	0.1387	0.4206	0	0	0	0.3555	0	10
Dataset 5.3	0	0.0057	0.0002	0	0	0.1387	0.1046	0	0	0	0.4848	0	11
Dataset 2.3	0	0.0094	0	0	0	0.1387	0.1573	0.4142	0	0	0	0	12
Dataset 10.3	0	0.015	0.0002	0	0	0.1387	0.1309	0.1946	0	0	0.0688	0	13
Dataset 4.2	0	0.0124	0.0002	0	0	0.1387	0.2363	0	0	0	0.7676	0	14
Dataset 6.3	0	0.0076	0	0	0	0.1387	0	0.4142	0	0	0.1293	0	15
Dataset 6.4	0	0.0153	0	0	0	0.1387	0.1573	0.1946	0	0	0	0	16
Dataset 9.2	0	0.0062	0.0002	0	0	0	0.3679	0.4142	0	0	0	0	17
Dataset 10.4	0	0.0048	0.0002	0	0	0.1387	0.4733	0.1946	0	0	0	0	18
Dataset 3.1	0	0	0.0002	0	0	0.0604	0.1046	0.1213	0	0	0	0	19
Dataset 4.4	0	0.0025	0.0002	0	0	0.1387	0	0	0	0	0.1371	0	20
Dataset 6.1	0	0	0.0002	0	0	0	0.0256	0.2678	0	0	0	0	21
Dataset 3.2	0	0.0127	0.0002	0	0	0.1387	0.3679	0.1946	0	0.4615	0	0	22
Dataset 4.3	0	0.002	0.0002	0	0	0	0.0783	0	0	0	0.2581	0	23
Dataset 9.1	0	0.0028	0.0002	0	0	0	0.3679	0	0	0	0.0234	0	24
Dataset 5.2	0	0.015	0.0002	0	0	0.1387	0	0.4142	0	0	0	0	25
Dataset 5.1	0	0.0109	0.0002	0	0	0.1387	0.2099	0.4142	0	0	0	0	26
Dataset 7.2	0	0.013	0	0	0	0.1387	0.3153	0	0	0	0	0	27
Dataset 1	0	0.0106	0	0	0	0.2953	0.4996	0	0	0	0	0	28
Dataset 2.4	0	0.0086	0.0002	0	0	0.1387	0	0	0	0	0	0	29
Dataset 4.1	0	0	0.0002	0	0	0	0	0	0	0	0.4899	0	30
Dataset 8.1	0	0.0162	0.0002	0	0	0	0	0	0	0	0.1007	0	31
Dataset 10.1	0	0.0106	0.0002	0	0	0	0	0	0	0	0	0	32
Dataset 7.1	0	0	0	0	0	0	0	0	0	0	0	0	33
Dataset 3.3	0	0	0	0	0	0.1387	0	0	0	0	0.6775	0	34
Dataset 2.1	0	0.0122	0.0002	0	0	0	0	0	0	0	0.0985	0	35
CISS2019.A1 (2)	0.1111	0.0001	0.0001	0	0	0.3333	0	0.3333	0	0	0.84	0	1
CISS2019.A1 (3)	0	0	0.0001	0	0	0.3333	0.3636	0.3333	0	0	0.424	0	2
CISS2019.A1 (1)	0	0.0007	0.0001	0	0	0.3333	0	0.3333	0	0	0.1298	0	3
CISS2019.A1 (4)	1	0	0	0	0	0	0	0	0	0	0	0	4

References

[1] R. Grubbs, J. Stoddard, S. Freeman, and R. Fisher, "Evolution and trends of industrial control system cyber incidents since 2017," *Journal of Critical Infrastructure Policy*, vol. 2, no. 2, pp. 45–79, 2021.

[2] A. Kok, A. Martinetti, and J. Braaksma, "The impact of integrating information technology with operational technology in physical assets: a literature review," *IEEE Access*, 2024.

[3] iTrust, Centre for Research in Cyber Security, "SWaT.A8 June 2021 Dataset." `https://itrust.sutd.edu.sg/`, 2023. Singapore University of Technology and Design.

[4] iTrust, Centre for Research in Cyber Security, "CISS 2019 Dataset." `https://itrust.sutd.edu.sg/`, 2019. Critical Infrastructure Security Showdown (CISS), Singapore University of Technology and Design.

[5] J. Eberlein, D. Rodriguez, and R. Harrison, "The effect of data complexity on classifier performance," *Empirical Software Engineering*, vol. 30, no. 1, p. 16, 2025.

[6] M. Keshavarz Ghorabaee, E. K. Zavadskas, L. Olfat, and Z. Turskis, "Multi-criteria inventory classification using a new method of evaluation based on distance from average solution (edas)," *Informatica*, vol. 26, no. 3, pp. 435–451, 2015.

[7] K. Erwin and A. Engelbrecht, "Feature-based complexity measure for multinomial classification datasets," *Entropy*, vol. 25, no. 7, p. 1000, 2023.

[8] S. Chakraborty, P. Chatterjee, and P. P. Das, "Evaluation based on distance from average solution (edas) method," in *Multi-Criteria Decision-Making Methods in Manufacturing Environments*, pp. 183–189, Apple Academic Press, 2023.

[9] S. Hussain, A. Tufail, H. A. A. G. Naim, M. A. Khan, and G. Barb, "Evaluation of computationally efficient identity-based proxy signatures," *IEEE Open Journal of the Computer Society*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: In our approach, we build upon existing work by utilizing the BERT model, incorporating the contextual information of both the head and tail entities. This allows us to leverage richer contextual cues from both ends of the relationship, enhancing the model's ability to understand and capture the nuances in the data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.