NEURAL INFORMATION PROCESSING SYSTEMS

OSTAR: Optimized Statistical Text-classifier with Adversarial Resistance

Yuhan Yao[1,2], Feifei Kou[1,2*], Lei Shi[3], Xiao Yang[1], Zhongbao Zhang[1], Suguo Zhu[4], Jiwei Zhang[1], Lirong Qiu[1,2], Haisheng Li[5]

[1] School of Computer Science (National Pilot School of Software Engineering), BUPT

[2]Key Laboratory of Trustworthy Distributed Computing and Service, BUPT, Ministry of Education

[3]State Key Laboratory of Media Convergence and Communication, CUC

[4]College of Computer Science and Technology, HDU

[5]School of Computer and Artificial Intelligence, BTBU

1 Motivation

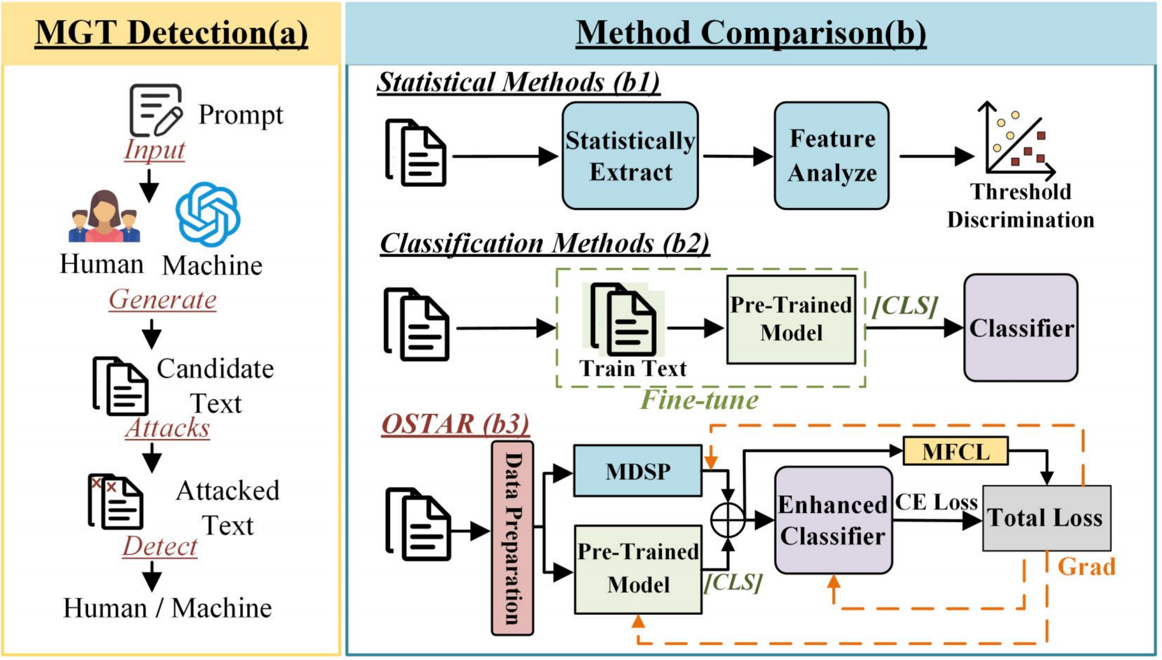
2 Approach

3 Results

4 Conclusion

1 Motivation

MGT Detection Task



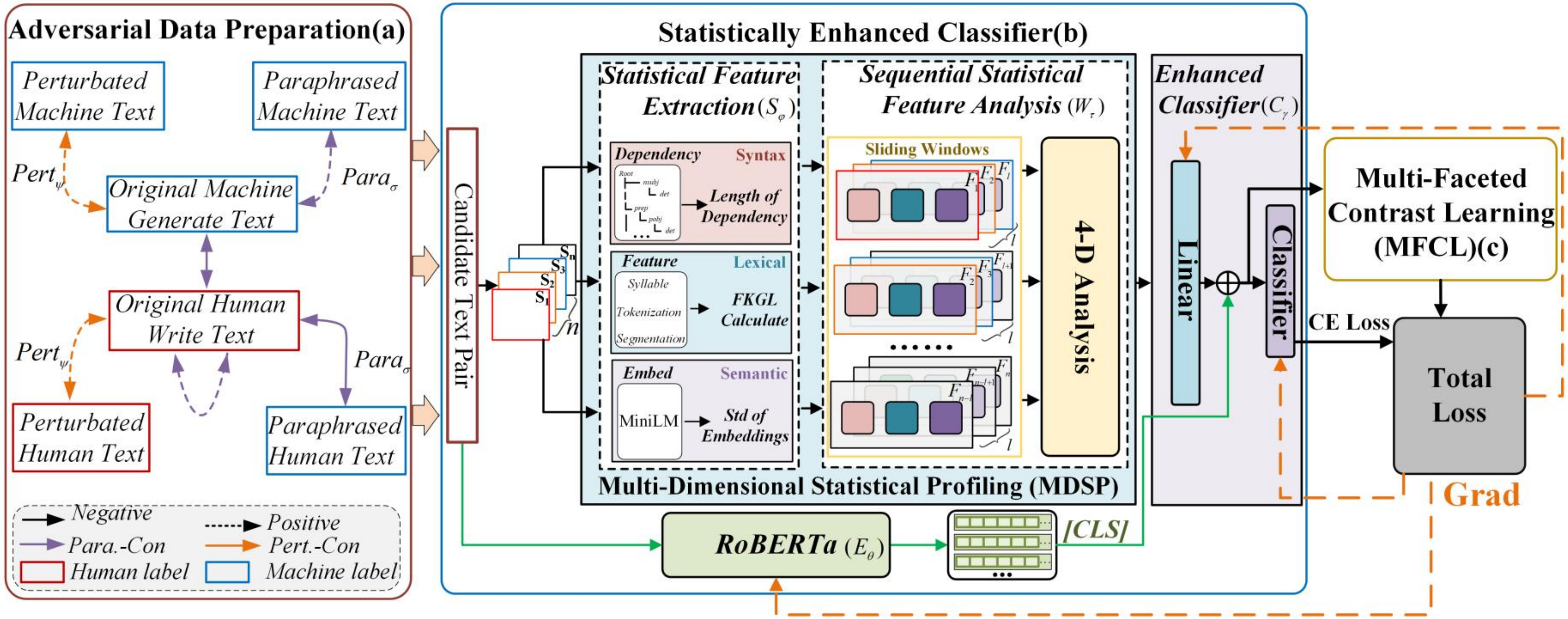
Motivation: Large language models (LLMs) have elevated machine-generated text (MGT) to near-human quality, yet its proliferation risks misinformation and undermines creativity. Real-world adversarial attacks evade detection, **demanding robust methods that adapt dynamically.**

Limitations: **Statistical methods** use intrinsic features but lack adaptability due to rigid thresholds, while **classifier-based methods** overfit superficial patterns and fail under distribution shifts.

Our proposed method **OSTAR** combines the advantages of statistical methods and classifier-based methods, **categorizes real-world adversarial attacks into perturbations and paraphrases**, and employs multi-faceted contrastive learning to achieve joint optimization for adversarial environments.

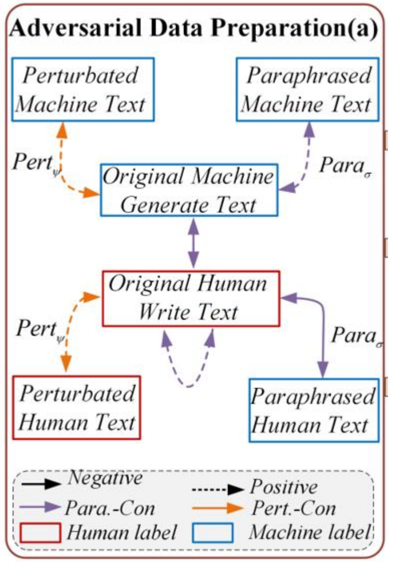
2 Approach

Overview of Our Method

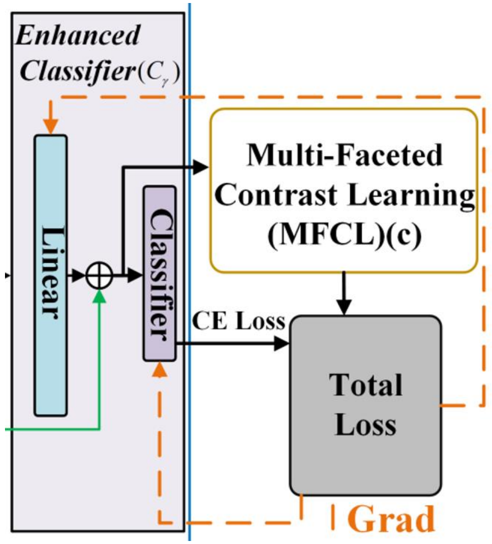


- The OSTAR framework is designed for robust machine-generated text detection, featuring three core components: Data Preparation, Statistically Enhanced Classifier, and Multi-Faceted Contrast Learning.

NEURAL INFORMATION PROCESSING SYSTEMS



Data-Preparation

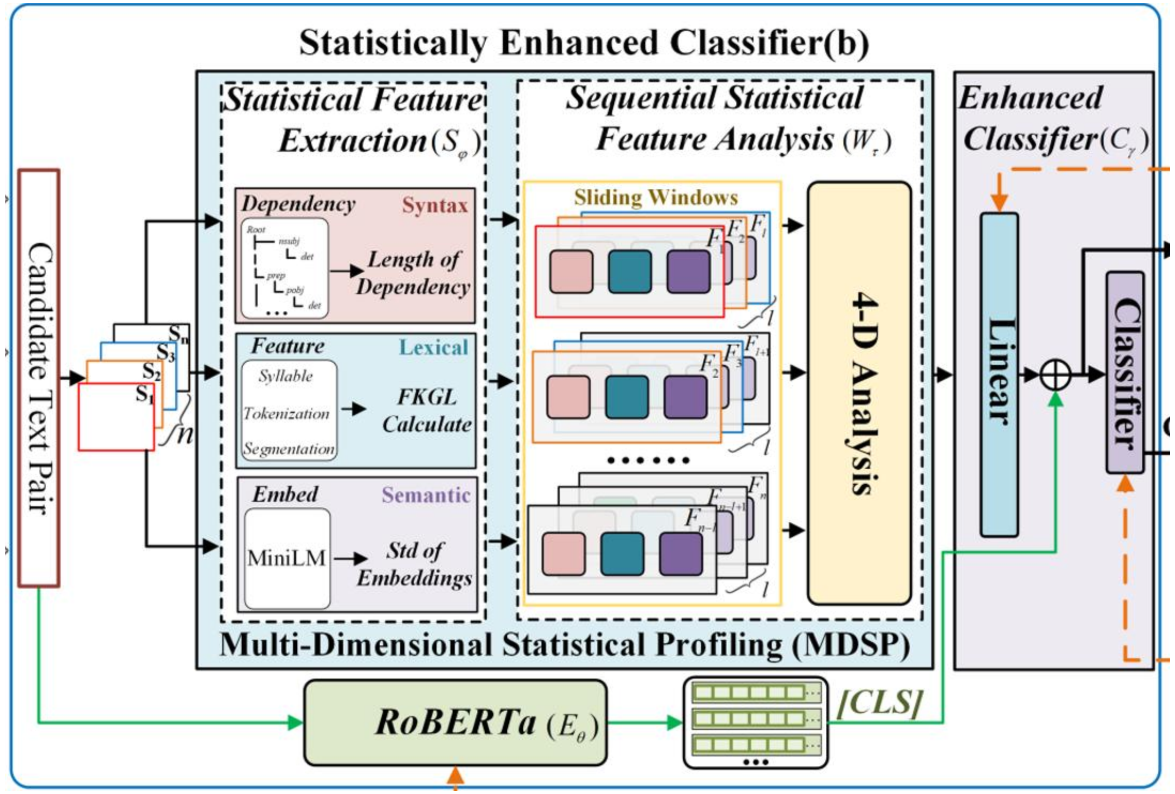


Multi-Faceted Contrast Learning

$$\mathcal{L}_{\text{MFCL}} = \lambda_1 \cdot \underbrace{\sum_{i=1}^{M} \sum_{p \in \mathcal{P}(i)} \log \frac{e^{S_{ip}/\tau}}{\sum_{p' \in \mathcal{P}(i)} e^{S_{ip'}/\tau} + \sum_{n \in \mathcal{N}(i)} e^{S_{in}/\tau}}}_{\mathcal{L}_{\text{Para}}} + \lambda_2 \cdot \underbrace{\sum_{a \in \mathcal{A}(i)} \beta_{ia} r S_{ia}}_{\mathcal{L}_{\text{Pert}}}$$

- **Data-Preparation:** We categorizes texts into original human-authored, machine-generated, and attacked variants (perturbations and paraphrases). Using perturbation sources (e.g., character-level changes) and paraphrasing tools, it dynamically constructs contrastive learning pairs during training epochs. This process ensures the model encounters diverse attack scenarios, enhancing its adaptability to distribution shifts.
- MFCL divides into paraphrase contrast (Para), which aligns samples based on text ownership, and perturbation contrast (Pert), which emphasizes similarity with adversarially modified texts.

NEURAL INFORMATION PROCESSING SYSTEMS



The classifier is augmented with **Multi-Dimensional Statistical Profiling (MDSP)** to capture intrinsic text features. MDSP extracts syntactic, lexical, and semantic statistics and analyzes them via sliding windows with 4-D metrics. The projected statistical features are then concatenated with CLS embeddings from pre-trained models like RoBERTa, **forming an enhanced classifier that leverages stable, invariant patterns for improved detection accuracy.**

3

Results

Robustness Comparison

Methods	DeepFake			CheckGPT			HC3		
	ACC	Recall	F1	ACC	Recall	F1	ACC	Recall	F1
GPT-2	87.29	90.58	88.04	81.92	83.01	80.74	90.86	90.75	89.41
RoBERTa	91.68	91.57	91.66	88.77	87.82	88.78	94.32	94.31	94.32
CoCo	88.03	89.59	87.58	84.55	84.90	85.97	98.42	99.31	98.50
RADAR	55.49	55.49	58.05	63.04	63.26	63.01	89.57	89.57	90.39
Watermark	86.21	90.45	88.91	75.69	**97.06**	72.26	94.88	94.75	95.13
Binoculars	78.22	82.41	76.39	86.90	89.74	87.12	92.44	95.13	91.95
PECOLA	86.29	86.19	86.29	84.58	84.96	84.51	99.23	99.25	99.24
OSTAR	**91.94**	**92.38**	**92.36**	**90.37**	90.12	**90.23**	**99.55**	**99.78**	**99.55**

Table1. Performance on original dataset

Methods	DeepFake			CheckGPT			HC3		
	Ori.	Pert.	Para.	Ori.	Pert.	Para.	Ori.	Attack	Para.
GPT-2	88.04	74.23	73.41	80.74	70.58	72.56	89.41	82.72	81.63
RoBERTa	90.12	77.10	79.00	88.78	80.62	81.59	94.32	90.27	90.95
CoCo	87.58	69.54	76.95	85.97	70.38	74.58	98.50	90.09	90.98
RADAR	58.05	48.54	47.11	63.01	60.21	67.42	49.78	47.52	58.47
Watermark	88.91	66.35	47.01	72.26	55.16	50.07	95.13	69.05	68.16
Binculars	76.39	45.42	51.23	87.12	52.32	54.54	91.95	72.34	78.68
PECOLA	86.29	78.13	60.08	84.51	62.64	60.71	98.35	65.09	68.82
OSTAR	**92.36**	**81.27**	**81.46**	**90.23**	**84.48**	**86.04**	**99.55**	**95.72**	**97.52**

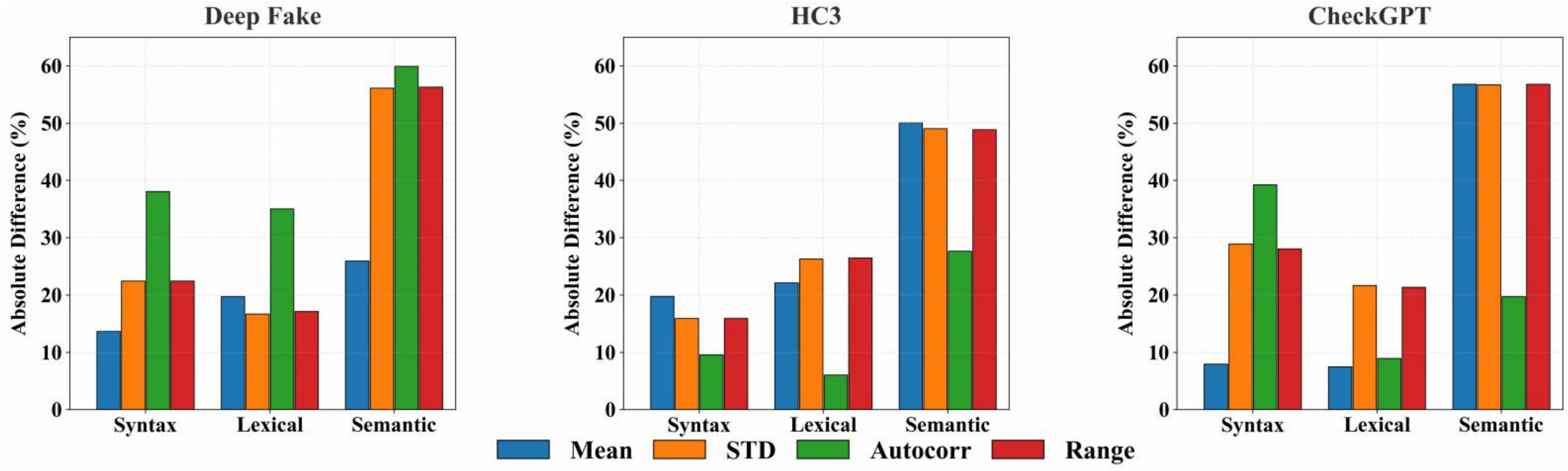
Table2. Performance on adversarial dataset

As shown in the table, OSTAR demonstrates exceptional performance in machine-generated text detection, achieving the highest accuracy, recall, and F1-score on original datasets (Table 1), such as 99.55% F1 on HC3, while under adversarial attacks (Table 2), it maintains robust results with minimal F1 degradation—only 11.09% on DeepFake—outperforming all baseline methods in both scenarios.

Model	Orginal		Pert.		Para.	
	ACC	F1	ACC	F1	ACC	F1
OSTAR (Plain)	90.34	90.12	81.10	77.10	82.61	79.00
OSTAR (Feature Extract)	90.72	90.58	81.67	77.08	82.97	79.27
OSTAR (Feature Extract+Analysis)	**92.17**	**92.87**	82.51	80.25	83.88	80.17
OSTAR	91.94	92.36	**84.34**	**81.27**	**84.75**	**81.46**

Table3. Ablation Study

The ablation study (Table 3) confirms both of OSTAR's components are crucial: the MDSP module boosts performance on original data, while the MFCL ensures robustness under attacks. The full framework shows minimal performance drop against attacks, proving its effectiveness in adversarial environments.

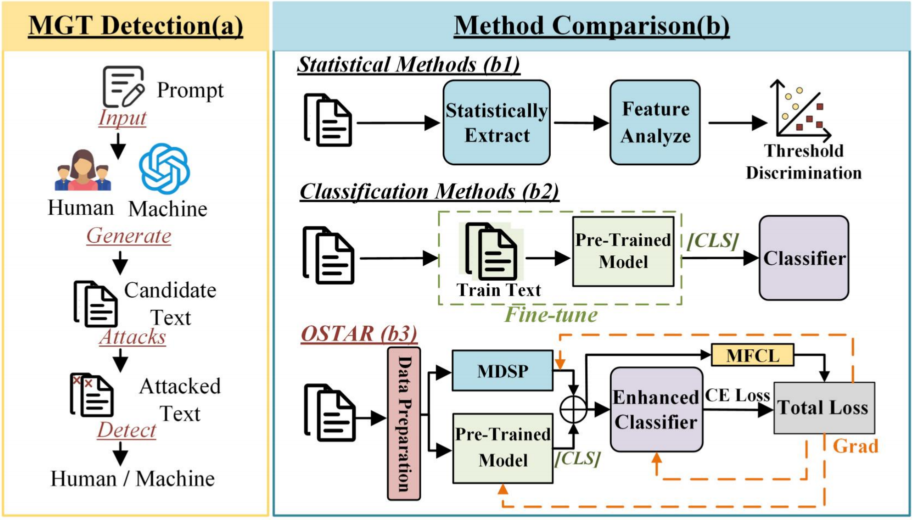


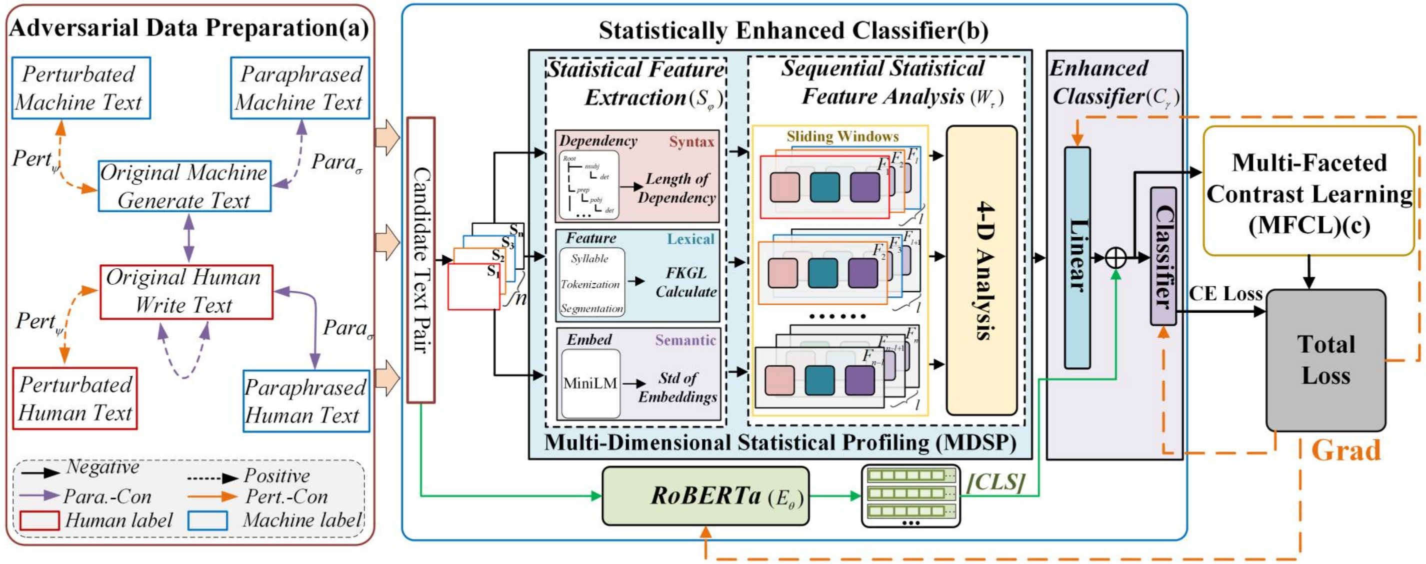
MDSP performance on three datasets

This figure illustrates MDSP's discriminative capability across three datasets, with a 30.95% average feature discrepancy between human and machine texts, highlighting MDSP's robust discriminative power by capturing intrinsic linguistic patterns, which remain stable under adversarial conditions and significantly enhance the OSTAR framework's detection accuracy by providing invariant features that complement neural embeddings.

4 Conclusion

NEURAL INFORMATION PROCESSING SYSTEMS





In this paper, We propose OSTAR, a robust machine-generated text detection framework that synergizes the intrinsic invariant feature extraction of statistics-based methods with the dynamic adaptability of classifier-based approaches. Specifically, the MDSP module manually extracts and analyzes multi-dimensional statistical features for enhanced classification, while MFCL addresses adversarial environments by categorizing attacks into Perturbation and Paraphrase and improving robustness through multi-perspective feature alignment. Extensive experiments on three datasets with various attacks validate OSTAR's effectiveness and robustness.

Thank you for your listening!

Email Address: koufeifei000@bupt.edu.cn
Wechat Address:

Feifei Kou :

寇菲菲
北京 海淀

扫一扫上面的二维码图案，加我为朋友。

Yuhan Yao :

yyh
北京 昌平

扫一扫上面的二维码图案，加我为朋友。