Investigating Universal Adversarial Attacks Against Transformers-based Automatic Essay Scoring Systems

Igor C. Silveira, André Barbosa, Daniel S. C. Lopes, **Denis D. Mauá**

BRACIS 2024
Nov 18, 2024

Outline

Automatic Essay Scoring

- First (rule-based) approaches date back to 60's, with some level of success
- Ambitious goal: annotate essay with useful feedback (useful means improving learning goals, assessing student performance, etc)
- Less ambitious goal: annotate essays with scores
- Huge potential to scale up personalized education, removing bias
- As any other ML system: sensitive to spurious correlations and prone to malicious usage (especially when stakes are high, such as in automatic grading).

Adversarial attacks

- Standard attack: change in input in an (human) imperceptible way that drives output as desired (e.g., replacing words with synonyms so as to increases the score).
- Universal adversarial attack: input-unrelated rules that drive output as desired (e.g., increasing text length with non-filler content).

This work: Investigate whether Transformers-based AES systems are susceptible to universal adversarial attacks that might occur in a classroom setting/educational enviroment?

Classroom setting: User model

- Students have little or no knowledge about the predictive model (so cannot tweak input as in standard attacks);
- Students submit am essay and receive immediate feedback in form of a score;
- Students interact multiple times (allows experimentation to create a model of the system behavior/explore vulnerabilities);
- Students might exchange information about vulnerabilities.

Dataset and models

- Transformer-based predictive models are trained on the AES-ENEM dataset [Silveira et al. 2024]
  - student essays scraped from publicly available mock ENEM exams
  - Each essay annotated by professional graders on with scores Per-competence scores in 0–200 w.r.t. 5 different competences
- Evaluated 3 different architectures:
  - BERT (from the AES-ENEM paper)
  - Phi-3 (Decoder model – new)
  - Google's Gemini (LLM Agent – new)

Model Performances

Model	Size	C1	C2	C3	C4	C5
LR	72	0.23	**0.40**	0.47	0.34	0.22
Phi-3	14B/892M train.	**0.46**	0.35	**0.52**	0.29	**0.61**
Gemini	≥70B?	0.41	**0.40**	0.40	0.36	0.35
BERTs	110M–330M	0.29–0.37	0.23–0.37	0.42–0.50	0.28–**0.42**	0.26–0.53

Table: QWK performance for different competencies.

Note: LR is competitive for C2–C4, no model clearly outperforms others for all competencies; Phi-3 performs poorly on C4

Deriving realistic universal attacks – Methodology

- Simulates malicious student learning of system's vulnerabilities
- NILC-Metrix generates a large number of interpretable textual features (coherence, fluency, cohesion, complexity, etc)
- Linear Regressor trained on features
- Most relevant features used to derive suitable universal attacks

Deriving realistic universal attacks – Feature analysis

Most relevant features:

Competence 1	Competence 2	Competence 3	Competence 4	Competence 5
adverbs	adverbs	adverbs	content words	adjective ratio
adjective ratio	adjective ratio	adjective ratio	function words	verbs
noun ratio	noun ratio	noun ratio	cau neg conn ratio	adverbs
verbs	verbs	verbs	content density	noun ratio

Student Model Hypothesis: Use of adverbs and adjectives increases score

Deriving realistic universal attacks

To increase the number/ratio of words of certain part-of-speech class we consider:

(a) listing words of that class (irrespective of cohesion or coherence)

(b) replicating the previous list to produce paragraphs

(c) pre-crafting a more natural sentence that employs words listed (but that is not relevant or coherent with text)

Deriving realistic universal attacks – Example

feature	(a) listing	(b) paragraph	(c) pre-crafted sentence
adverbs (1)	list of adverbs (1a)	4x the previous (2a)	many copies of sentence overusing adverbs (3a)
adjectives (2)	list of adjectives (1b)	4x the previous (2b)	many copies of sentence overusing adjectives (3b)
both (3)	list of both (1c)	4x the previous (2c)	many copies of sentence overusing both (3c)

Attack 1a: "Well, badly, enormously, certainly, wrongly, rapidly, slowly, fairly, unfairly".

Attack 1b: Repeat list in four different paragraphs.

Attack 1c: "Undeniably, progressing slowly, leisurely, carefully, silently while deeply breathing and thinking intensively about the given problem", inserted 10 times in 4 different paragraphs.

Results

Att.	Model	C1	C2	C3	C4	C5	Total	Att.	C1	C2	C3	C4	C5	Total	Att.	C1	C2	C3	C4	C5	Total
1a	LR	200	200	200	200	**0**	800	2a	200	200	200	200	**0**	800	3a	200	200	200	200	**0**	800
	BERT	120	80	40	120	**0**	360		120	80	40	120	**0**	360		120	120	40	120	**0**	400
	Gemini	120	120	80	120	120	560		120	120	120	120	160	640		120	80	80	120	120	520
	Phi-3	**0**	40	40	**0**	**0**	80		80	120	40	**0**	**0**	240		80	40	40	**0**	**0**	160
1b	LR	200	200	200	200	**0**	800	2b	200	200	200	200	**0**	800	3b	200	200	200	200	**0**	800
	BERT	120	120	80	120	**0**	440		120	120	80	120	**0**	440		120	120	120	120	**0**	480
	Gemini	80	80	80	80	80	400		80	80	80	80	80	400		80	40	40	40	80	280
	Phi-3	80	120	80	120	**0**	400		80	120	120	120	**0**	440		80	120	80	120	**0**	400
1c	LR	200	200	200	200	200	1000	2c	120	200	200	200	200	920	3c	200	200	200	200	200	1000
	BERT	160	160	160	160	**0**	640		160	160	160	160	0	640		160	160	160	160	40	680
	Gemini	40	40	0	40	40	160		**0**	**0**	**0**	**0**	**0**	**0**		40	**0**	**0**	**0**	**0**	40
	Phi-3	80	120	40	80	**0**	320		80	40	40	80	**0**	240		80	40	40	80	**0**	240

Table: Per-competence scores in $\{0,40,80,120,160,200\}$

- Expectedly, Logistic Regression is easily fooled by attacks
- Phi-3 is less sensitive but still often assigns above average score
- BERT and Gemini often assign high grades for certain attacks
- Attacks for Competence 5 are seldom successful (even for LR)

Deriving realistic universal attacks – Competence 5

- Competence 5 evaluates reflection/conclusion
- Given this, hypothesized semantic-related Attack 4: a sentence that resembles a conclusion overusing adjectives and adverbs appended to 7 paragraphs.

"Consequently, it is up to the fair and democratic Federal Government to rapidly approve laws that rapidly reduce the occurrence of these horrendous problems. Following, the dear Brazilian population must abide by the undeniable laws, and the fast police must arrest those that committed any inhuman crime."

Results – Attack 4

Model	C1	C2	C3	C4	C5	Total
LR	200	200	200	200	**0**	800
BERT	160	160	160	160	160	800
Gemini	40	40	**0**	40	40	160
Phi-3	160	120	120	120	40	560

- LR least sensitive to attack for C5 (but maxing for all others!)
- BERT assigned close to maximum for all competences
- Phi-3 assigned above average for all competences but(!) C5
- Gemini was least sensitive overall

Conclusion

- Universal attacks can deem an AES system useless
- They are easily conceived by non-expert users with repeated use
- BERT and SoTA Phi3 models are very susceptible to such attacks
- Gemini is robust to repetitions but not to small sentences
- Fine-tuning leads to more vulnerable systems (noted in literature)
- Warning: **Cautious deployment of AES models in the wild.**

- Phi-3 model available at HuggingFace, code to generate attacks available at Github.