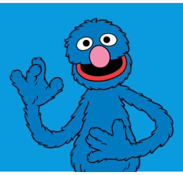


Robustness Analysis of Grover for Machine-generated News Detection

[1]School of Computing Technologies, RMIT University, Australia
[2]Defence Science and Technology Group, Australia

Rinaldo Gagiano[1] | Maria Myung-Hee Kim[2] | Xiuzhen Zhang[1] | Jennifer Biggs[2]

BACKGROUND & RESEARCH QUESTIONS

- Current language models can **produce neural fake news at scale**
- **Grover** is a model for both generation and detection of neural fake news
- Detecting the difference between machine and human-produced articles can **reduce the risk** of neural fake news spreading online
- Grover, serving as a defence mechanism against neural fake news, would **need to be robust against adversarial efforts**

RQ1 ~ Can adversarial attacks with minimal alterations on input articles, deteriorate the performance of Grover's discriminator?

RQ2 ~ What components of Grover's discriminator are affected by adversarial attacks?

RQ3 ~ How do adversarial attacks affect the classification score produced by Grover's discriminator?

ADVERSARIAL ASSESSMENT

Experiment Dataset:
100 *Machine-generated* articles

Adversarial Attacks:
1) Upper/Lower Flip
2) Homoglyph
3) Whitespace
4) Misspelling

Attack Parameters:
- Allow one alteration per iteration
- Iterate through the entire article

How many articles out of the 100 target articles had at least one alteration that resulted in a misclassification

Attack	Alterations	Misclassifications (Proportion)	Affected Articles
U/L Flip	212,224	4,295 (2.02%)	96%
Homoglyph	157,532	6,914 (4.39%)	97%
Whitespace	46,036	1,447 (3.14%)	85%
Misspelling	43,789	4,281 (9.78%)	94%

Grover is highly susceptible to adversarial efforts

ERROR ANALYSIS

Ten most affected words from all false negative cases (changed the classification from 'Machine' to 'Human')

Most affected words are all 'Stop-Words'

Affected Word	Frequency	Proportion	POS
that	1639	8.92%	IN
the	1533	8.34%	DT
to	516	2.81%	TO
and	334	1.82%	CC
with	321	1.75%	IN
in	298	1.62%	IN
of	279	1.52%	IN
for	257	1.40%	IN
from	236	1.28%	IN
The	202	1.10%	DT

IN ~ Preposition, DT ~ Determiner, TO ~ To, CC ~ Coordinating Conjunction

INPUT ENCODING

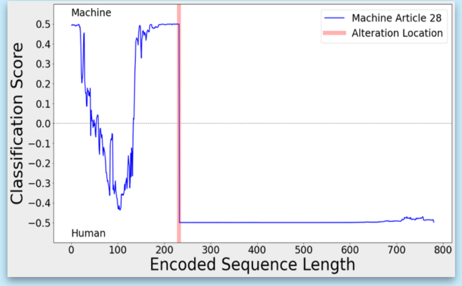
Grover uses a byte-pair encoder splitting input into subword units and assigns a pairing ID

Original	Vector IDs		Altered
A	33		A
Romanian	34345		Romanian
		10497	hosp
hospital	4437	1027	It
		283	al
will	482		will
face	1987		face
a	258		a
fine	3735		fine
for	330		for

Uppercasing of letter 'i' in 'hospital' changes subword unit allocation as 'hospItal' is broken into 'hosp', 'It', 'al'

CUMULATIVE CLASSIFICATION SCORE

Recording each classification score as word vectors are fed to Grover allows a cumulative classification score to be recorded



- Cumulative classification score of a misclassified Machine article
- 'that' altered into 'thaT' by U/L Flip attack
- Classification score dropped a total of 0.98 at attack location

- Average score variations of a subset of True Positive and False Negative cases
- FN cases had a much higher average variation in classification score

Average Score Variation		
Attack	TP Subset	FN Subset
U/L Flip	0.12	0.76
Homoglyph	0.17	0.81
Whitespace	0.04	0.70
Misspelling	0.21	0.69
Average	**0.14**	**0.74**

CONCLUSION

- **Singular character changes** could cause Grover to fail
- Adversarial attacks affected up to **97% of target articles**
- Identified **vulnerable words** to focus attack alterations
- **Grover's encoder is highly sensitive** to particular perturbations causing downstream effects in classification assignment
- Developed a **novel visualisation method** to interpret adversarial attacks affects and identified large variations in classification scores
- False negative cases had large score variations ultimately **affecting the final prediction produced**

R. Gagiano is supported by Defence Science and Technology Group Graduate Industry Placement. X. Zhang is supported by the ARC Discovery Project DP200101441.