

How do humans perceive adversarial text?

A reality check on the validity and naturalness of word-based adversarial attacks

Salijona Dyrmishi, Salah Ghamizi, Maxime Cordy

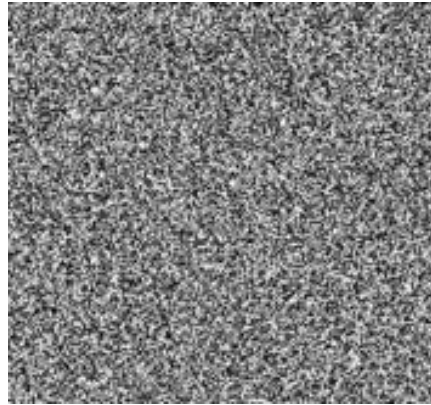
University of Luxembourg



Adversarial attacks against Machine Learning (ML)



Cute Dog
97%

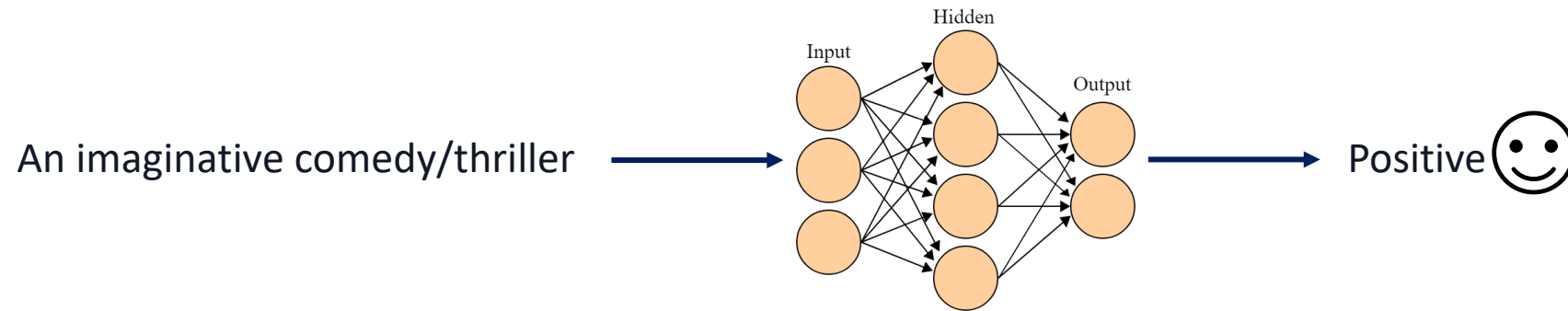


Carefully crafted
adversarial noise

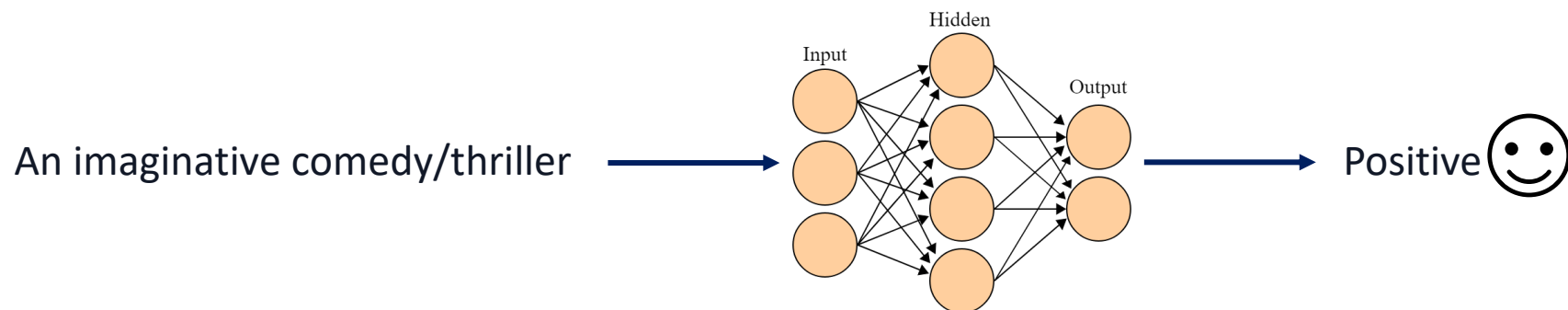


Angry Cat
82%

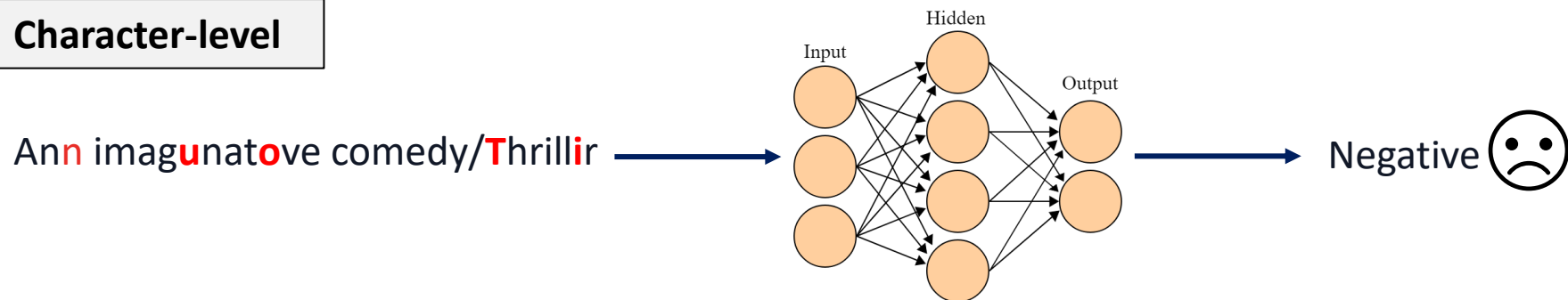
Adversarial attacks against text-based models



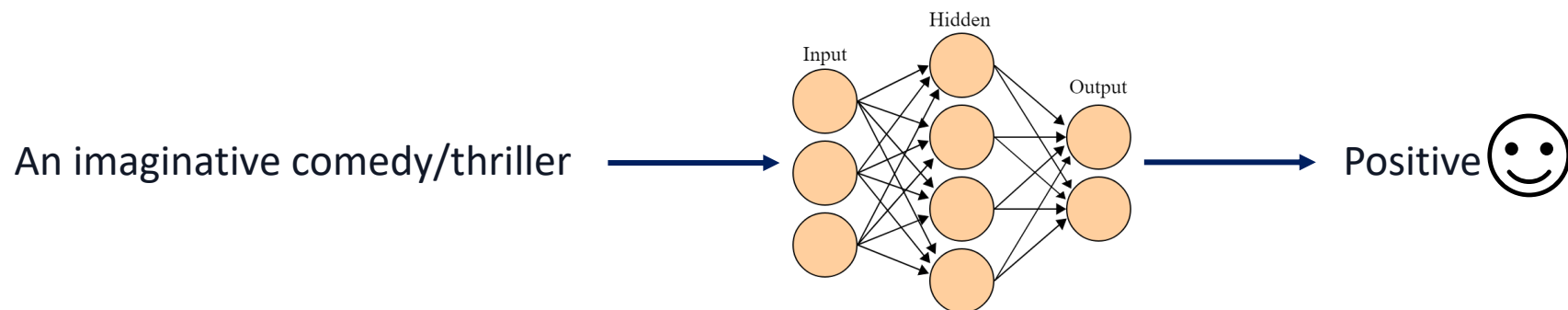
Adversarial attacks against text-based models



Character-level



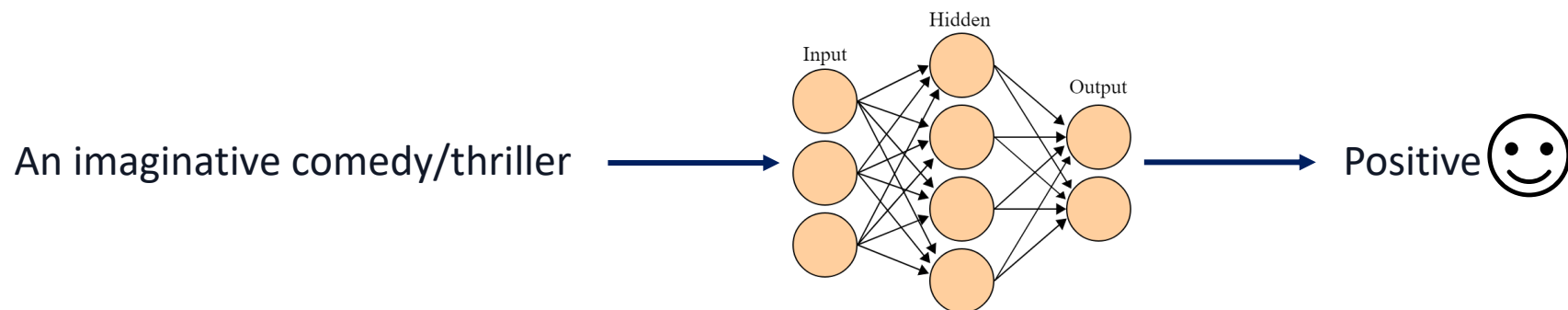
Adversarial attacks against text-based models



Word-level

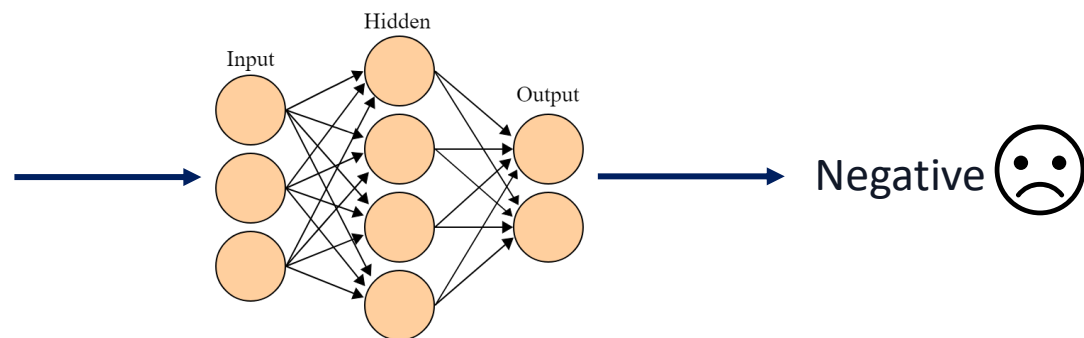
A creative comedy/thriller

Adversarial attacks against text-based models



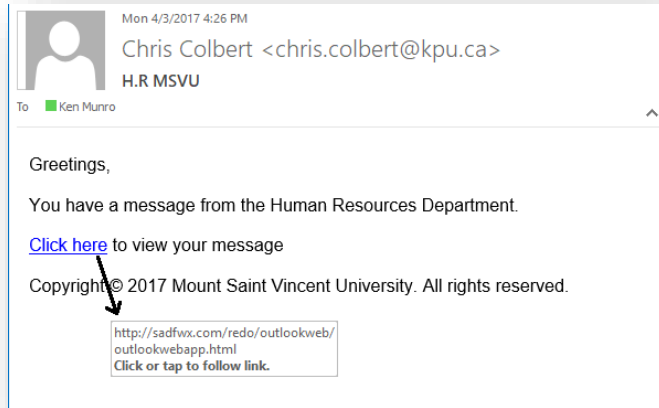
Paraphrase

**A brilliantly crafted and
captivating imaginative
comedy/thriller experience.**



Human in the loop

Phishing



Fake news



Offensive language

Desired properties of adversarial text

Valid & Natural

I **love** this movie  I **like** this movie

Invalid

I **love** this movie  I **hate** this movie

Unnatural

I love this **movie**  i LoVe tHiss cinEmatiC

Attack name/paper	Type	Evaluation				Participants	Attacks studied	
		Validity	Naturalness					
			S.	D.	G.	M.		
Hotflip (Ebrahimi et al., 2017)	Word based	✓	X	X	X	X	3	1
Alzantot(Alzantot et al., 2018)		✓	X	X	X	X	20	1
Input-reduction(Feng et al., 2018)		✓	X	X	X	X	N/A	1
Kuleshov(Kuleshov et al., 2018)		✓	X	X	X	X	5	1
Bae(Garg and Ramakrishnan, 2020)		✓	✓	X	✓	X	3	2
Pwvs(Ren et al., 2019)		✓	✓	X	X	X	6	1
Textfooler (Jin et al., 2019)		✓	X	X	✓	✓	2	1
Bert-attack(Li et al., 2020b)		✓	X	X	✓	X	3	1
Clare (Li et al., 2020a)		✓	X	X	X	X	5	2
PSO (Zang et al., 2019)		✓	✓	X	X	X	3	1
Fast-alzantot (Jia et al., 2019)		X	X	X	X	X	0	0
IGA (Wang et al., 2019)		X	X	X	X	X	0	0
Textbugger (Li et al., 2018)	Character based	✓	X	✓	X	X	297	1
Pruthi (Pruthi et al., 2019)		✓	X	X	X	X	N/A	1
DeepWordBug (Gao et al., 2018)		X	X	X	X	X	0	0
Morris et al. (2020a)	Independent	X	✓	X	✓	✓	10	2

Table 1: Human evaluation performed on quality of adversarial examples by existing literature. The terms abbreviated are Suspiciousness(S.), Detectability(D.), Grammaticality(G.), Meaning(M.). N/A indicates information is not available.

Attack name/paper	Type	Evaluation				Participants	Attacks studied	
		Validity	Naturalness					
			S.	D.	G.	M.		
Hotflip (Ebrahimi et al., 2017)	Word based	✓	X	X	X	X	3	1
Alzantot(Alzantot et al., 2018)		✓	X	X	X	X	20	1
Input-reduction(Peng et al., 2018)		✓	X	X	X	X	N/A	1
Kuleshov(Kuleshov et al., 2018)		✓	X	X	X	X	5	1
Bae(Garg and Ravi, 2017)		✓	X	X	X	X	3	2
Pwvs(Ren et al., 2019)		✓	✓	X	X	X	6	1
Textfooler (Jin et al., 2020)		✓	X	X	✓	✓	2	1
Bert-attack(Li et al., 2020b)		✓	X	X	✓	X	3	1
Clare (Li et al., 2020a)		✓	X	X	X	X	5	2
PSO (Zang et al., 2019)		✓	X	X	X	X	5	1
Fast-alzantot (Jia et al., 2019)		X	X	X	X	X	0	0
IGA (Wang et al., 2019)		X	X	X	X	X	0	0
Textbugger (Li et al., 2018)	Character based	✓	X	✓	X	X	297	1
Pruthi (Pruthi et al., 2019)		✓	X	X	X	X	N/A	1
DeepWordBug (Gao et al., 2018)		X	X	X	X	X	0	0
Morris et al. (2020a)	Independent	X	✓	X	✓	✓	10	2

3 studies do not involve humans in their evaluation

Naturalness evaluated only through few criteria or not at all

Less than 10 participants

Effect of perturbation size and language proficiency not considered

Table 1: Human evaluation performed on quality of adversarial examples by existing literature. The terms abbreviated are Suspiciousness(S.), Detectability(D.), Grammaticality(G.), Meaning(M.). N/A indicates information is not available.

Attack name/paper	Type	Evaluation					Participants	Attacks studied
		Validity	Naturalness					
			S.	D.	G.	M.		
Hotflip (Ebrahimi et al., 2017)	Word based	✓	X	X	X	X	3	1
Alzantot(Alzantot et al., 2018)		✓	X	X	X	X	20	1
Input-reduction(Peng et al., 2018)		✓	X	X	X	X	N/A	1
Kuleshov(Kuleshov et al., 2018)		✓	X	X	X	X	5	1
Bae(Garg and Rastogi, 2019)		✓	X	X	X	X	3	2
Pwvs(Ren et al., 2019)		✓	✓	X	X	X	6	1
Textfooler (Jin et al., 2020)		✓	X	X	✓	✓	2	1
Bert-attack(Li et al., 2020b)		✓	X	X	✓	X	3	1
Clare (Li et al., 2020a)		✓	X	X	X	X	5	2
PSO (Zang et al., 2019)		✓	X	X	X	X	5	1
Fast-alzantot (Jia et al., 2019)		X	X	X	X	X	0	0
IGA (Wang et al., 2019)		X	X	X	X	X	0	0
Textbugger (Li et al., 2018)	Character based	✓	X	✓	X	X	297	1
Pruthi (Pruthi et al., 2019)		✓	X	X	X	X	N/A	1
DeepWordBug (Gao et al., 2018)		X	X	X	X	X	0	0
Morris et al. (2020a)	Independent	X	✓	X	✓	✓	10	2

3 studies do not involve humans in their evaluation


Naturalness evaluated only through few criteria or not at all

Less than 10 participants

Effect of perturbation size and language proficiency not considered

Table 1: Human evaluation performed on quality of adversarial examples by existing literature. The terms abbreviated are Suspiciousness(S.), Detectability(D.), Grammaticality(G.), Meaning(M.). N/A indicates information is not available.

An extensive study on human perception of adversarial texts

		Validity	Naturalness					Studies
			S.	D.	G.	M.		
Hotflip (Ebrahimi et al., 2017)	 378 participants 9 word-level attacks	✓	X	X	X	X	3	1
Alzantot(Alzantot et al., 2018)		✓	X	X	X	X	20	1
Input-reduction(Feng et al., 2018)		✓	X	X	X	X	N/A	1
Kuleshov(Kuleshov et al., 2018)		✓	X	X	X	X	5	1
Bae(Garg and Ramakrishnan, 2020)		✓	✓	X	✓	X	2	2
Wang et al., 2019)		✓	✓	X	X	X	1	1
Wang et al., 2019)		✓	X	X	✓	✓	1	1
Bert-attack(Li et al., 2020b)		✓	X	X	✓	X	3	1
Chen et al., 2020a)		✓	X	X	X	X	2	2
PSO (Zang et al., 2019)	Character based	✓	✓	X	X	X	297	1
Fast-alzantot (Jia et al., 2019)		X	X	X	X	X	0	0
IGA (Wang et al., 2019)		X	X	X	X	X	0	0
Textbugger (Li et al., 2018)	Independent	✓	X	✓	X	X	297	1
Pruthi (Pruthi et al., 2019)		✓	X	X	X	X	N/A	1
DeepWordBug (Gao et al., 2018)		X	X	X	X	X	0	0
Morris et al. (2020a)	Independent	X	✓	X	✓	✓	10	2
Our study		✓	✓	✓	✓	✓	378	9

3000 texts
(original and adversarial)

Table 1: Human evaluation performed on quality of adversarial examples by existing literature. The terms abbreviated are Suspiciousness(S.), Detectability(D.), Grammaticality(G.), Meaning(M.). N/A indicates information is not available.

Evaluated aspects

Validity



Naturalness



Suspiciousness



Detectability



Grammaticality



Meaningfulness

Results: Validity

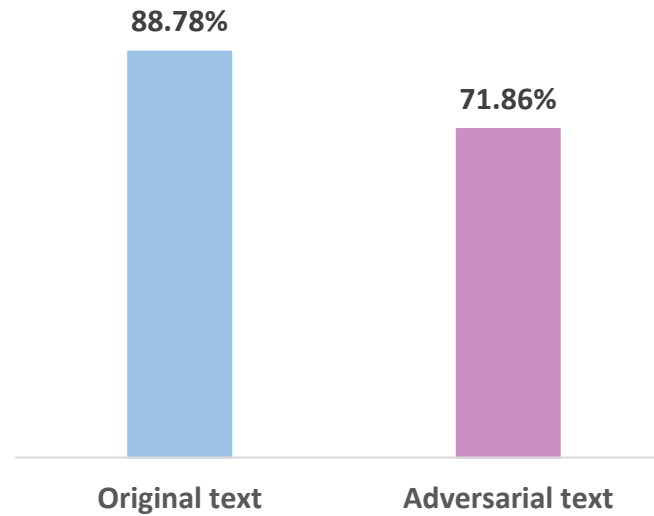


Fig 1. Percentage of correctly labelled texts according to their ground truth

Naturalness: Suspicion

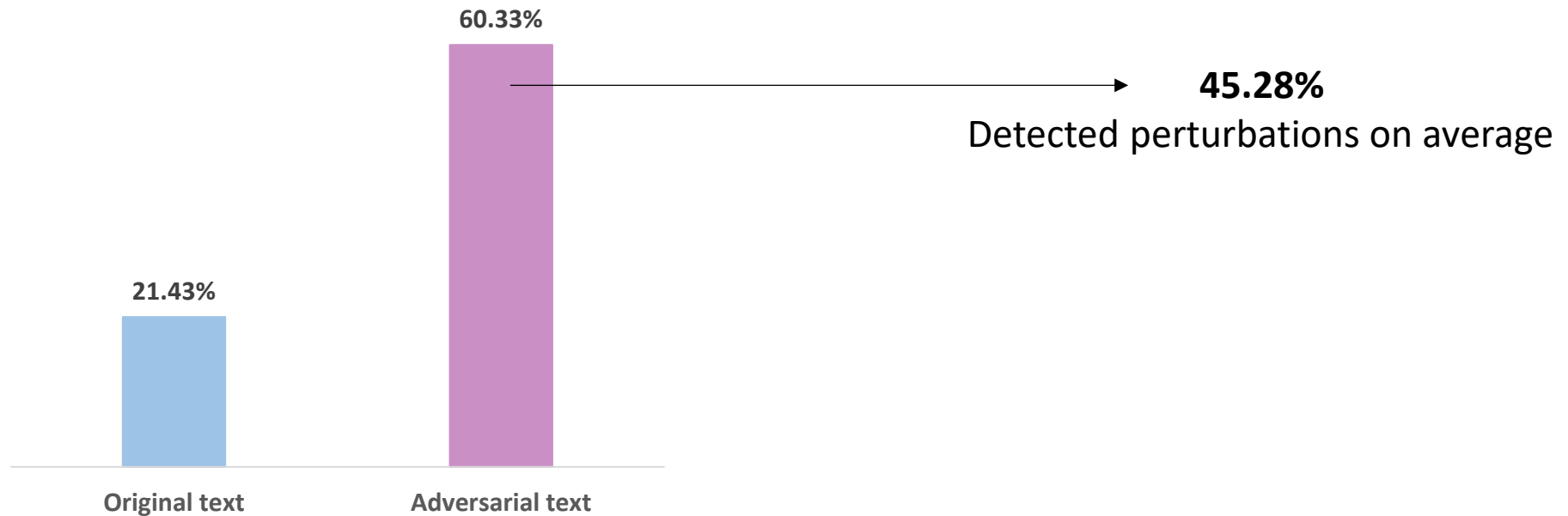


Fig 2. Percentage of texts that were suspected to be computer altered

Naturalness: Grammaticality

45.28%

Adversarial texts contain errors not present in their original counterpart

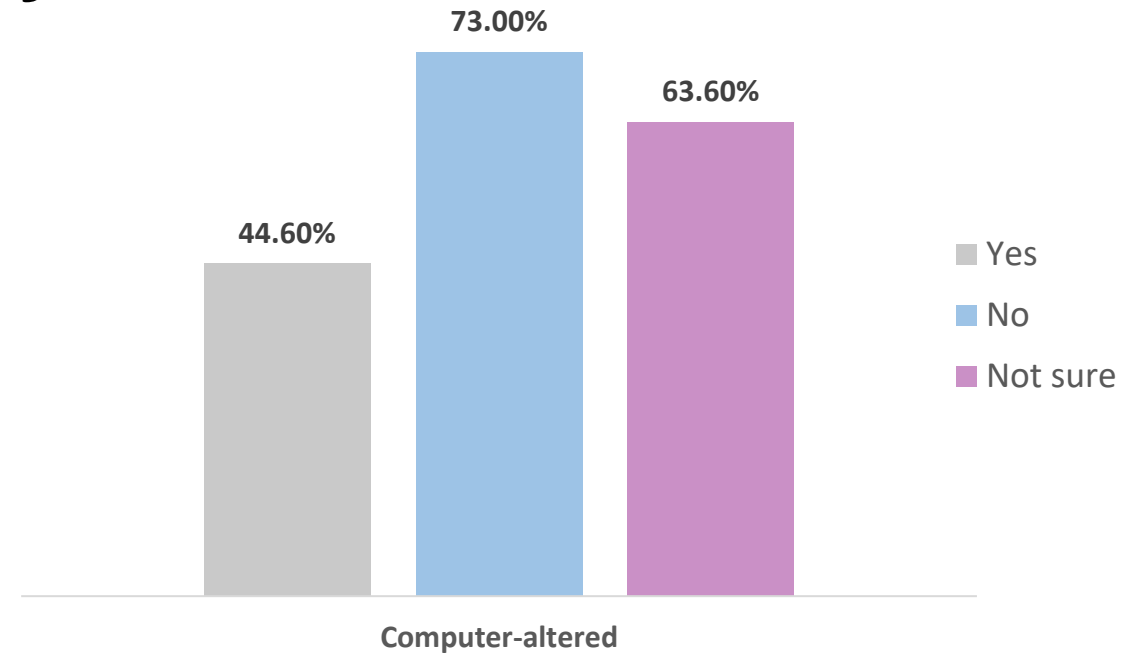


Fig 3. Percentage of adversarial texts labelled as computer-altered according to grammar errors.

Naturalness: Meaning

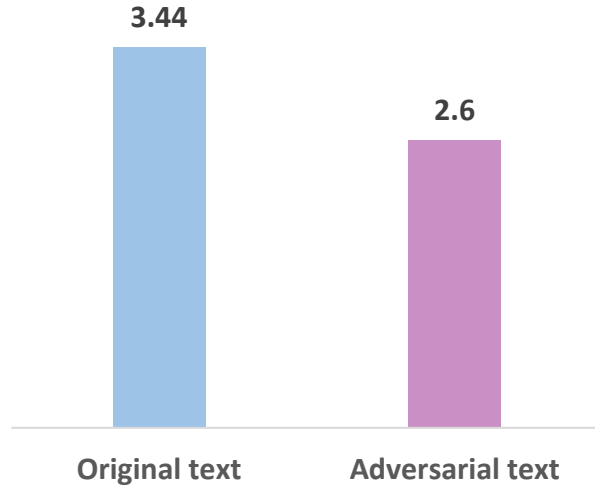


Fig 4. Meaning clarity rating on a 1-4 Likert scale

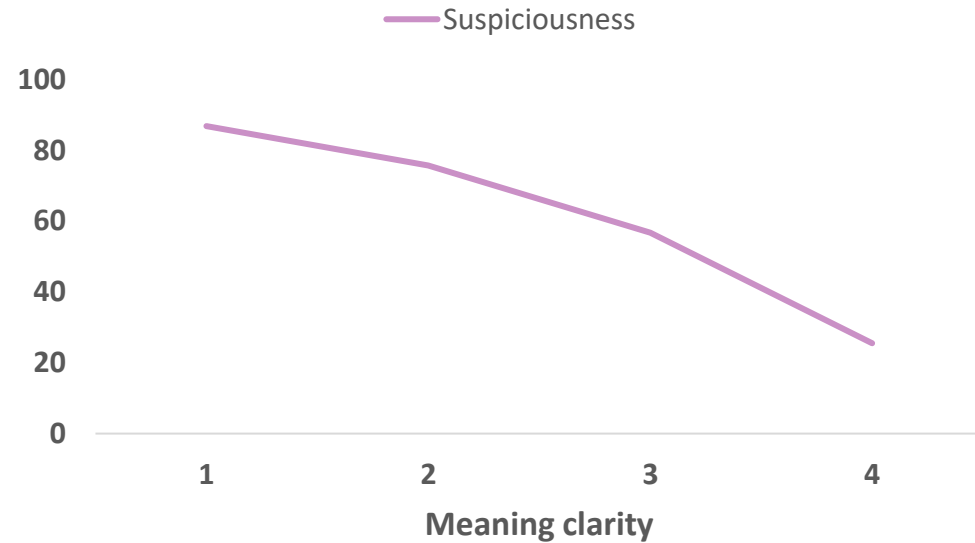


Fig 5. Percentage of adversarial text suspected to be computer altered according to meaning clarity.

Extra investigation

- Individual attacks
- Language proficiency effect
- Perturbation size effect



Evaluating the human perception of adversarial text requires extra attention in NLP systems where a human is involved in the loop.

ACL 2023, 9-14 July

