

INTRODUCTION

Natural language processing models are known to be vulnerable to adversarial examples and synonym substitution based attacks are widely adopted for generating textual adversarial examples. Various defense methods have been proposed to mitigate the threat of textual adversarial examples, *e.g.* adversarial training, input transformations, detection, *etc*.

In this work, we propose a simple yet effective detection method called **RS**&V to resist adversarial attacks.

MOTIVATION

- **Replacement sequence.** We regard the optimization process of synonym-based attacks as searching a specific sequence for word replacement, in which the words mutually influence each other and contribute together to mislead the target classifier.
- Hypothesis. We can eliminate the perturbation if we break the mutual interaction of the words in the replacement sequence.
- **Observation.** Randomly substituting words with its synonyms could consistently and significantly **improve the robust accuracy** against adversarial examples while maintaining the high clean **accuracy** under various substitution rates.
- **RS&V.** We detect adversarial examples to **vote** the prediction label by accumulating the logits of k samples generated by randomly substituting the words in the input text with synonyms.

THE FRAMEWORK OF RS&V



Detecting Textual Adversarial Examples through Randomized Substitution and Vote

Xiaosen Wang, Yifeng Xiong, Kun He School of Computer Science and Technology, Huazhong University of Science and Technology

ALGORITHM

Algorithm 1 The RS&V Algorithm

Input: Input text $x = \{w_1, w_2, ..., w_n\}$, target classifier f, substitution rate *p*, number of votes *k*, stopword selection portion *s*.

- **Output:** Detection result and restored label 1: Calculate the stopword set \mathcal{W} containing the top s high fre-
- quency words in the training set
- 2: Initialize converted text set $\mathcal{X} = \emptyset$
- 3: for $i = 1 \rightarrow k$ do
- Initialize a new text $x_i = x$
- Randomly sample $n \cdot p$ words for \mathcal{P} from x_i/\mathcal{W} 5:
- for each word $w_t \in \mathcal{P}$ do 6:
- Randomly select a synonym $\hat{w}_t^{j} \in \mathcal{S}(w_t)$ 7:
- Substitute $w_t \in x_i$ with \hat{w}_t^j 8:
- end for 9:
- $\mathcal{X} = \mathcal{X} \cup x_i$ 10:
- 11: **end for**
- 12: Calculate the prediction label for input text x: $\bar{y} = \arg \max f(x)$ ▷ Vote & Detection
- 13: Calculate the voted label: $\bar{y}_v = \arg \max \sum_{i=1}^k f(x_i)$
- 14: if $\bar{y} = \bar{y}_v$ then
- **return** False, \bar{y} 15:
- 16: **end if**
- 17: return True, \bar{y}_v

Randomized Substitution

Benign sample

> Adversarial example

EXPERIMENTAL RESULTS

Table 1: The classification accuracy (%) and F1 score (%) of various detection methods for Word-CNN and BERT on three datasets. N/A denotes the normally trained model without the detection module.

$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	Dataset	Model	Method	Clean	GA		PWWS		PSO		Textfooler		HLA		Average	
$ \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$					Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	AG's News	CNN	N/A	92.1	43.6	_	37.1	_	36.4	_	24.8	_	41.9	_	36.8	_
AG's News FGWS FGWS 91.3 91.3 76.2 84.1 80.3 90.2 75.8 85.1 83.4 76.2 92.2 86.0 86.0 77.5 86.0 88.3 86.0 88.1 94.8 94.8 86.0 88.1 94.8 94.8 86.0 94.8 86.1 94.8 86.0 AG's News BERT M/A 94.9 68.5 - 74.9 - 59.2 - 61.3 - 62.0 - 65.2 BERT DISP 94.5 85.3 77.8 85.2 77.7 88.0 93.0 81.1 85.6 84.2 84.6 RoBERTa DISP 94.6 87.2 82.9 88.8 91.7 94.2 92.5 96.2 92.3 90.8 92.8 N/A 93.5 71.9 - 78.7 - 66.6 8.5 85.2 88.3 85.3 85.9 84.9 82.6 77.3 86.8 85.9 87.4 RoBERTa DISP 93.4 86.6 88.9 90.3 87.0 </td <td>DISP</td> <td>91.6</td> <td>77.3</td> <td>83.7</td> <td>76.5</td> <td>85.0</td> <td>78.0</td> <td>85.8</td> <td>69.0</td> <td>80.8</td> <td>79.2</td> <td>85.9</td> <td>76.0</td> <td>84.2</td>			DISP	91.6	77.3	83.7	76.5	85.0	78.0	85.8	69.0	80.8	79.2	85.9	76.0	84.2
$ \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$			FGWS	91.3	76.2	80.3	75.8	83.4	76.2	84.0	77.5	88.3	80.0	85.8	77.1	84.4
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$			RS&V	91.3	84.1	90.2	85.1	92.2	86.8	94.0	86.0	94.8	88.1	94.8	86.0	93.2
News BERT DISP 94.5 85.3 77.8 85.2 70.3 83.8 80.9 83.0 81.1 85.6 84.2 84.6 RGWS 94.6 87.2 82.9 88.8 82.5 87.7 88.0 89.2 91.5 89.1 90.5 88.4 RoBERTa N/A 93.5 71.9 - 78.7 - 66.8 - 74.6 - 68.8 - 72.2 RoBERTa DISP 93.4 84.8 75.8 84.8 65.3 85.9 84.9 82.8 67.3 86.8 85.3 85.0 ROBERTa DISP 93.4 84.6 75.8 84.8 65.3 85.9 84.9 82.8 67.3 86.8 85.3 85.0 82.8 67.3 86.6 88.7 88.9 87.4 88.9 87.4 89.3 88.9 87.4 88.9 89.3 61.2 62.0 77.4 49.3 87.4 68.9		BERT	N/A	94.9	68.5		74.9	—	59.2	_	61.3	—	62.0	—	65.2	_
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$			DISP	94.5	85.3	77.8	85.2	70.3	83.8	80.9	83.0	81.1	85.6	84.2	84.6	78.9
$ \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$			FGWS	94.6	87.2	82.9	88.8	82.5	87.7	88.0	89.2	91.5	89.1	90.5	88.4	87.1
$ \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$			RS&V	94.6	90.5	90.4	91.3	88.8	91.7	94.2	92.5	96.2	92.3	90.8	92.8	92.1
RoBERTa DISP FGWS 93.4 93.1 84.8 75.8 84.8 65.3 85.9 84.9 82.8 67.3 86.8 85.3 85.0 RS&V 93.1 89.6 88.9 90.3 87.0 91.5 94.7 91.2 92.6 91.5 94.7 90.8 RS&V 93.4 89.6 88.9 90.3 87.0 91.5 94.7 91.2 92.6 91.5 94.7 90.8 CNN N/A 87.2 6.2 - 1.5 - 2.7 - 0.6 - 17.4 - 5.7 PGWS 86.5 64.7 82.8 64.7 84.0 69.7 87.5 72.6 90.0 72.8 87.4 68.9 RS&V 86.3 79.6 94.0 80.2 94.8 80.9 95.0 79.2 94.1 81.7 94.5 80.3 IMDB BERT N/A 91.9 15.4 - 26.7 7.5		RoBERTa	N/A	93.5	71.9		78.7	—	66.8	_	74.6	—	68.8	—	72.2	_
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$			DISP	93.4	84.8	75.8	84.8	65.3	85.9	84.9	82.8	67.3	86.8	85.3	85.0	75.7
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$			FGWS	93.1	86.2	80.0	87.2	74.6	87.0	86.6	88.5	85.2	88.3	88.9	87.4	83.1
$ \begin{tabular}{ c c c c c c c c c c c c c c c c c c c$			RS&V	93.4	89.6	88.9	90.3	87.0	91.5	94.7	91.2	92.6	91.5	94.7	90.8	91.6
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$	IMDB	CNN	N/A	87.2	6.2	_	1.5	_	2.7	_	0.6	_	17.4	—	5.7	_
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$			DISP	87.2	48.8	68.3	43.1	64.8	53.3	74.4	39.3	61.2	62.0	77.4	49.3	69.2
IMDB RS&V 86.3 79.6 94.0 80.2 94.8 80.9 95.0 79.2 94.1 81.7 94.5 80.3 IMDB BERT DISP 91.8 64.3 77.5 63.7 72.2 68.7 84.1 62.0 77.6 74.5 86.7 66.6 FGWS 92.5 80.6 90.9 79.5 88.2 82.0 92.9 83.0 93.2 84.8 94.0 82.0 82.0 RS&V 92.1 87.8 96.0 88.2 95.4 88.5 96.9 89.1 97.2 89.9 97.2 88.7 RoBERTa N/A 94.2 18.3 - 29.9 - 7.0 - 34.3 - 21.8 - 22.3 RoBERTa DISP 93.9 66.3 77.4 64.1 70.0 67.4 81.7 68.7 73.3 76.7 86.1 68.6 RS&V 94.4 81.0 90.1 82.0 89.1 83.2 92.9 86.6 92.7 85.7 93.4			FGWS	86.5	64.7	82.8	64.7	84.0	69.7	87.5	72.6	90.0	72.8	87.4	68.9	86.3
IMDB BERT N/A 91.9 15.4 - 26.7 - 5.6 - 9.5 - 15.7 - 14.6 DISP 91.8 64.3 77.5 63.7 72.2 68.7 84.1 62.0 77.6 74.5 86.7 66.6 FGWS 92.5 80.6 90.9 79.5 88.2 82.0 92.9 83.0 93.2 84.8 94.0 82.0 82.0 RoBERTa N/A 94.2 18.3 - 29.9 - 7.0 - 34.3 - 21.8 - 22.3 RoBERTa N/A 94.2 18.3 - 29.9 - 7.0 - 34.3 - 21.8 - 22.3 RoBERTa DISP 93.9 66.3 77.4 64.1 70.0 67.4 81.7 68.7 73.3 76.7 86.1 68.6 FGWS 94.4 81.0 90.1 82.0 89.1 83.2 92.9 86.6 92.7 85.7 93.4 83.7			RS&V	86.3	79.6	94.0	80.2	94.8	80.9	95.0	79.2	94.1	81.7	94.5	80.3	94.5
$ \begin{array}{ c c c c c c c c c c c c c c c c c c c$		BERT	N/A	91.9	15.4	_	26.7	_	5.6	_	9.5	_	15.7	_	14.6	_
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$			DISP	91.8	64.3	77.5	63.7	72.2	68.7	84.1	62.0	77.6	74.5	86.7	66.6	79.6
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$			FGWS	92.5	80.6	90.9	79.5	88.2	82.0	92.9	83.0	93.2	84.8	94.0	82.0	91.8
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$			RS&V	92.1	87.8	96.0	88.2	95.4	88.5	96.9	89.1	97.2	89.9	97.2	88.7	96.5
RoBERTaDISP FGWS93.9 94.466.3 81.077.4 90.164.1 82.070.0 89.167.4 83.281.7 92.968.7 86.673.3 92.776.7 85.786.1 93.468.6 83.7CNNN/A DISP FGWS69.8 69.837.4 37.467.1 67.135.6 35.6- 63.82.6 39.3- 70.635.9 35.966.5 85.745.0 90.876.3 96.990.3CNNN/A FGWS FGWS69.0 68.049.7 49.782.6 82.648.2 48.280.9 80.949.4 49.482.2 82.240.6 40.672.1 39.939.9 75.075.0 45.6Yahoo!N/A N/A 76.776.7 13.8 76.713.8 76.7- 25.6 76.48.9 76.7- 76.717.9 76.7- 76.711.5 76.7- 		RoBERTa	N/A	94.2	18.3	—	29.9	_	7.0	-	34.3	—	21.8	—	22.3	_
$\begin{array}{c ccccccccccccccccccccccccccccccccccc$			DISP	93.9	66.3	77.4	64.1	70.0	67.4	81.7	68.7	73.3	76.7	86.1	68.6	77.7
RS&V94.689.495.988.894.791.097.491.496.590.896.990.3 N/A 69.54.7-5.6-2.6-3.9-4.3-4.2DISP69.837.467.135.663.839.370.635.966.545.076.338.6FGWS68.049.782.648.280.949.482.240.672.139.975.045.6RS&V69.3 63.092.662.192.863.293.361.693.262.691.962.5 Yahoo!N/A76.713.8-25.6-8.9-17.9-11.5-15.5			FGWS	94.4	81.0	90.1	82.0	89.1	83.2	92.9	86.6	92.7	85.7	93.4	83.7	91.6
$CNN = \begin{bmatrix} N/A & 69.5 & 4.7 & - & 5.6 & - & 2.6 & - & 3.9 & - & 4.3 & - & 4.2 \\ DISP & 69.8 & 37.4 & 67.1 & 35.6 & 63.8 & 39.3 & 70.6 & 35.9 & 66.5 & 45.0 & 76.3 & 38.6 \\ FGWS & 68.0 & 49.7 & 82.6 & 48.2 & 80.9 & 49.4 & 82.2 & 40.6 & 72.1 & 39.9 & 75.0 & 45.6 \\ RS\&V & 69.3 & 63.0 & 92.6 & 62.1 & 92.8 & 63.2 & 93.3 & 61.6 & 93.2 & 62.6 & 91.9 & 62.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ DISP & DISP & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & 76.7 & 13.8 & - & 25.6 & - & 8.9 & - & 17.9 & - & 11.5 & - & 15.5 \\ \hline N/A & $			RS&V	94.6	89.4	95.9	88.8	94.7	91.0	97.4	91.4	96.5	90.8	96.9	90.3	96.3
CNNDISP FGWS RS&V69.8 68.0 37.4 		CNN	N/A	69.5	4.7	_	5.6	_	2.6	_	3.9	_	4.3	_	4.2	-
FGWS 68.0 49.7 82.6 48.2 80.9 49.4 82.2 40.6 72.1 39.9 75.0 45.6 RS&V 69.3 63.0 92.6 62.1 92.8 63.2 93.3 61.6 93.2 62.6 91.9 62.5 Yahoo! N/A 76.7 13.8 - 25.6 - 8.9 - 17.9 - 11.5 - 15.5			DISP	69.8	37.4	67.1	35.6	63.8	39.3	70.6	35.9	66.5	45.0	76.3	38.6	68.9
RS&V 69.3 63.0 92.6 62.1 92.8 63.2 93.3 61.6 93.2 62.6 91.9 62.5 N/A 76.7 13.8 $-$ 25.6 $-$ 8.9 $-$ 17.9 $-$ 11.5 $-$ 15.5 Yahoo! DED $ -$ <			FGWS	68.0	49.7	82.6	48.2	80.9	49.4	82.2	40.6	72.1	39.9	75.0	45.6	78.6
N/A 76.7 13.8 - 25.6 - 8.9 - 17.9 - 11.5 - 15.5 Yahoo! DIOD T	Yahoo! Answers		RS&V	69.3	63.0	92.6	62.1	92.8	63.2	93.3	61.6	93.2	62.6	91.9	62.5	92.8
		BERT	N/A	76.7	13.8		25.6	_	8.9	_	17.9		11.5	_	15.5	_
Answers BERT DISP 76.7 50.0 74.5 50.5 68.6 53.8 80.4 51.7 74.8 56.0 81.9 52.4			DISP	76.7	50.0	74.5	50.5	68.6	53.8	80.4	51.7	74.8	56.0	81.9	52.4	76
FGWS 75.7 62.2 88.0 62.7 85.9 62.4 88.7 66.0 90.4 65.5 91.1 63.8			FGWS	75.7	62.2	88.0	62.7	85.9	62.4	88.7	66.0	90.4	65.5	91.1	63.8	88.8
RS&V 75.8 67.4 92.3 68.7 91.0 69.7 93.7 71.6 94.1 70.0 93.9 69.5			RS&V	75.8	67.4	92.3	68.7	91.0	69.7	93.7	71.6	94.1	70.0	93.9	69.5	93.0
N/A 74.7 19.8 – 33.7 – 15.2 – 41.7 – 19.6 – 26.0		RoBERTa	N/A	74.7	19.8	_	33.7	_	15.2	_	41.7	—	19.6	_	26.0	-
RoBERTa DISP 74.7 48.0 68.7 50.4 61.3 50.9 75.1 53.7 57.6 55.2 78.6 51.7			DISP	74.7	48.0	68.7	50.4	61.3	50.9	75.1	53.7	57.6	55.2	78.6	51.7	68.3
FGWS 74.8 62.3 87.5 64.9 85.9 63.3 88.5 67.2 86.3 65.7 90.1 64.7			FGWS	74.8	62.3	87.5	64.9	85.9	63.3	88.5	67.2	86.3	65.7	90.1	64.7	87.7
RS&V 76.0 66.4 90.3 66.8 86.7 68.1 92.2 68.3 86.8 68.7 92.8 67.7			RS&V	76.0	66.4	90.3	66.8	86.7	68.1	92.2	68.3	86.8	68.7	92.8	67.7	89.8

CONCLUSION

We propose a novel detection method RS&V against synonym substitution based adversarial attacks for text classification.

- the model architectures.
- Impact. RS&V identifies the fragility of textual adtext without degrading clean accuracy.



• Novelty. We identify that randomized synonym substitution could destroy the mutual interaction among words in the replacement sequence for adversarial attacks. Based on this observation, we propose RS&V to effectively detect adversarial examples.

• Effectiveness. Empirical evaluations demonstrate that RS&V could achieve better detection performance than existing baselines while maintaining a high performance on benign samples.

• **Generality.** RS&V is generally applicable to all existing deep neural networks without any additional training or modification on

versarial examples, which might inspire more defense and detection methods by pre-processing the

