Appendix

A RETVec Model Details

Table 7 details the hyperparameter settings for the RETVec model architecture, as described in Section 3.

Hyperparameter	RETVec
Max word length	16
Per-character encoding dim	24
Activation	GeLU
# of projection layers	1
Projection layer dim	32
# of fully-connected layers	2
Fully-connected layer dim	256
Spatial dropout rate	0.0625
Dropout rate	0
Embedding activation	Tanh
Embedding dim	256
Similarity dim	256

Table 7: RETVec model hyperparameter details.

B Benchmarking Models

In this section, we provide more detailed model hyperparameters for the evaluation models used in Section 5 and Section 8.

RNN We used a Stacked-LSTM architecture, with the following hyperparameters:

- Dim: 256
- Layers: 4
- Dropout rate: 0.1

DPCNN We use the architecture described in [14], with the following hyperparameters:

- Filters: 256
- Layers: 6
- Kernel size: 3
- Final dropout: 0.5
- Activation: ReLU

BERT-Mini We use the architecture as described in [6] for BERT-Mini, with the following hyperparameters:

- Layers: 4
- Hidden dim: 256
- Intermediate dim: 1024
- Self-attention heads: 4
- Dropout rate: 0.1
- Activation: GeLU

BERT-Base We use the architecture described in [6] for BERT-Base, with the following hyperparameters:

- Layers: 12
- Hidden dim: 768
- Intermediate dim: 3072
- Self-attention heads: 12
- Dropout rate: 0.1
- Activation: GeLU

C RETVec Ablation Studies

In this section, we present our ablation study results for the RETVec model design and hyperparameter selection. Results are reported on the Multilingual Amazon Reviews dataset following the methodology described in Section 10.

C.1 Embedding Dimension

Detailed results on how RETVec's embedding layer dimension affects classification performance are reported in Table 8.

Embedding Dim	Pre-training Loss	Test Accuracy			
		RNN	CNN	BERT	AVG
64	0.0284	92.8%	91.8%	92.5%	92.4%
100	0.0267	93.2%	92.1%	92.4%	92.6%
128	0.0260	93.3%	92.2%	92.5%	92.7%
200	0.0253	93.4%	92.3%	92.6%	92.7%
256	0.0248	93.6%	92.3%	92.8%	92.9%
300	0.0245	93.6%	92.3%	92.6%	92.8%
384	0.0237	93.6%	92.2%	92.6%	92.8%
512	0.0225	93.7%	92.3%	92.8%	92.9%

Table 8: Ablation study results on the effect of the RETVec pre-trained model's embedding dimension on classification performance. **Bold** denotes the hyperparameter selected for the final RETVec model.

C.2 Model Architecture

Detailed results on the effect of RETVec architecture type on classification performance are reported in Table 9.

Architecture	Pre-training Loss	Test Accuracy			
		RNN	CNN	BERT	AVG
MLP	0.0248	93.6%	92.3%	92.8%	92.9%
MLP + BERT	0.0129	93.2%	92.0%	92.4%	92.5%
MLP + T5	0.0120	93.5%	92.2%	92.5%	92.7%
MLP + GAU	0.0133	93.5%	92.2%	92.8%	92.8%
MLP + LSTM	0.0179	93.4%	91.9%	92.5%	92.6%
MLP + CNN	0.0214	93.5%	92.1%	92.6%	92.7%

Table 9: Ablation study results on RETVec model architecture type on classification performance. **Bold** denotes the hyperparameter selected for the final RETVec model.

C.3 Maximum Word length

Detailed results on how RETVec's maximum input word length affects classification performance are reported in Table 10.

Word Len	Pre-training Loss	Test Accuracy			
		RNN	CNN	BERT	AVG
8	0.0428	93.6%	92.3%	92.7%	92.8%
10	0.0326	93.5%	92.4%	92.6%	92.8%
12	0.0267	93.5%	92.2%	92.7%	92.8%
14	0.0259	93.5%	92.2%	92.6%	92.8%
16	0.0248	93.6%	92.3%	92.8%	92.9%
20	0.0242	93.5%	92.2%	92.6%	92.7%
24	0.0234	93.6%	92.3%	92.7%	92.9%
28	0.0246	93.5%	92.2%	92.7%	92.8%
32	0.0235	93.6%	92.2%	92.6%	92.8%

Table 10: Ablation study results on the effect of maximum word length on RETVec classification performance. **Bold** denotes the hyperparameter selected for the final RETVec model.

C.4 Pre-training Loss Hyperparameters

Detailed results on the effect of Multi-Similarity loss hyperparameters on RETVec classification performance are reported in Table 11. We also experimented with Circle Loss [26] and report the results in Table 12.

Hyperparameter				Test Accuracy			
α	β	λ	Pre-training Loss	RNN	CNN	BERT	AVG
2	20	0.5	0.0432	93.5%	92.1%	92.4%	92.7%
2	20	1.0	0.2643	92.9%	91.2%	92.6%	92.2%
2	40	0.5	0.0455	93.6%	92.1%	92.6%	92.7%
2	40	1.0	0.3610	92.9%	91.2%	92.4%	92.2%
2	80	0.5	0.0464	93.5%	92.3%	92.6%	92.8%
2	80	1.0	0.3583	92.6%	91.1%	92.4%	92.0%
4	20	0.5	0.0270	93.4%	92.0%	92.6%	92.7%
4	20	1.0	0.1537	93.0%	91.3%	92.4%	92.2%
4	40	0.5	0.0242	93.5%	92.4%	92.6%	92.8%
4	40	1.0	0.1919	92.8%	91.3%	92.4%	92.2%
4	**80**	**0.5**	0.0248	93.6%	92.3%	92.8%	92.9%
4	80	1.0	0.1851	92.8%	91.2%	92.4%	92.1%

Table 11: Ablation study results on the effect of Multi-Similarity loss hyperparameters on RETVec classification performance. $\epsilon = 0.1$ is fixed for all experiments. **Bold** denote the hyperparameters selected for the final RETVec model.

Circle-Loss Hyperparameter			Test Accuracy			
Scale Factor γ	Relaxation Factor m	Pre-training Loss	RNN	CNN	BERT	AVG
64	0.3	7.55	93.3%	91.9%	92.6%	92.6%
64	0.4	2.77	93.5%	92.0%	92.4%	92.7%
64	0.5	0.63	93.6%	92.4%	92.6%	92.9%
128	0.3	12.85	93.3%	91.9%	92.6%	92.6%
128	0.4	5.06	93.6%	92.1%	92.6%	92.8%
128	0.5	1.18	93.6%	92.3%	92.6%	92.9%
256	0.3	24.97	93.3%	91.9%	92.5%	92.5%
256	0.4	9.49	93.5%	92.0%	92.4%	92.7%
256	0.5	2.59	93.7%	92.4%	92.5%	92.9%
512	0.3	49.01	93.2%	92.0%	92.6%	92.6%
512	0.4	19.88	93.3%	92.2%	92.7%	92.7%
512	0.5	5.26	93.6%	92.4%	92.7%	92.9%

Table 12: Ablation study results for pre-training RETVec with various Circle Loss [26] hyperparameter settings.

C.5 Model Capacity

Detailed results on how the number and dimension of fully-connected dense layers in the RETVec model affects classification performance are presented in Table 13.

# Layers	Dim	Pre-training Loss	Test Accuracy			
			RNN	CNN	BERT	AVG
0	-	0.0445	92.8%	91.6%	92.3%	92.2%
1	128	0.0383	93.5%	91.8%	92.4%	92.6%
1	256	0.0312	93.5%	91.8%	92.4%	92.6%
1	384	0.0284	93.6%	92.0%	92.6%	92.7%
1	512	0.0258	93.6%	92.2%	92.8%	92.8%
2	128	0.0334	93.4%	91.9%	92.6%	92.6%
2	**256**	0.0248	93.6%	92.3%	92.8%	92.9%
2	384	0.0201	93.5%	92.3%	92.6%	92.8%
2	512	0.0177	93.7%	92.3%	92.6%	92.9%
3	128	0.0314	93.4%	91.9%	92.5%	92.6%
3	256	0.0213	93.6%	92.4%	92.5%	92.8%
3	384	0.0175	93.6%	92.2%	92.7%	92.8%
3	512	0.0149	93.5%	92.4%	92.6%	92.8%

Table 13: Ablation study results on the effect of RETVec model capacity (number and dimension of the fully-connected layers) on classification performance. **Bold** denote the hyperparameters selected for the final RETVec model.

C.6 Spatial Dropout Rate

Detailed ablation study results on the amount of spatial dropout in the RETVec model and its effect on classification performance are presented in Table 14. Increments of $1/16$ were used because it corresponds to dropping out one character of the input on average, since RETVec's model accepts an input of up to 16 characters per word.

Spatial Dropout	Pre-training Loss	Test Accuracy			
		RNN	CNN	BERT	AVG
0.00%	0.0122	93.4%	91.4%	92.5%	92.4%
6.25%	0.0248	93.6%	92.3%	92.8%	92.9%
12.50%	0.0465	93.4%	92.0%	92.3%	92.6%
18.75%	0.0722	92.7%	91.5%	91.8%	92.0%
25.00%	0.0967	92.7%	91.4%	91.2%	91.8%

Table 14: Ablation study results on the effect of spatial dropout rate on the RETVec input character encoding. **Bold** denotes the value selected for the final RETVec model.

C.7 Pre-Training Objectives

We evaluated combining RETVec's pre-training objective (Multi-Similarity loss) with other objective functions and pre-training tasks as well. Specifically, we experimented with the following objectives: 1) augmentation position prediction, 2) augmentation position and type prediction, 3) decoding (predicting the character encoding of the input token), and 4) denoising (predicting the character encoding of the original, non-augmented token). Table 15 reports the results of our experiments on different pre-training objectives.

Objectives	Similarity Loss	Total Loss	RNN	CNN	BERT	AVG
Similarity, Augmentation Position Detection	0.0261	0.2228	93.4%	92.0%	92.8%	92.7%
Similarity, Augmentation Position and Type Prediction	0.0238	0.0970	93.2%	92.2%	92.6%	92.7%
Similarity, Decoding	0.0278	0.0384	93.5%	92.1%	92.6%	92.8%
Similarity, Denoising	0.0241	0.1088	93.3%	91.8%	92.7%	92.6%
Similarity	0.0248	0.0248	93.6%	92.3%	92.8%	92.9%

Table 15: Ablation study results on combining different pre-training objectives with similarity loss for RETVec pre-training.

D RETVec Pre-training Dataset Augmentations

Below, we provide the full list of character-level augmentations (broken down into four categories) used to generate typo-augmented words for the RETVec pre-training dataset, as described in Section 3.3.

- Deletion
- Insertion
  - Repeated character insertion
  - n-grams based prefix insertion for $n = 3, 4, 5$
  - n-grams based suffix insertion for $n = 3, 4, 5$
  - Random ASCII character insertion
  - Language alphabet-based random character insertion
  - Random punctuation insertion
  - Random punctuation prefix
  - Random punctuation suffix
  - Random BMP Unicode insertion
  - Random emoji prefix
  - Random emoji suffix
- Substitution
  - Case substitution
  - n-grams based substitution for $n = 3, 4, 5$
  - QWERTY keyboard typo substitution
  - Homoglyphs substition
  - Random ASCII character substitution
  - Language alphabet-based random character subsitution
  - Random puctuation substitution
  - Random BMP Unicode substitution
- Transposition
  - Neighboring character swap
  - 3-character block random shuffle

E RETVec Pre-training Hyperparameters

We train RETVec using Multi-Similarity loss with hyperparameters $\alpha = 4$, $\beta = 40$, $\epsilon = 0.1$ and $\lambda = 0.5$. Detailed pre-training hyperparameters are reported in Table 16.

Hyperparameter	**Pre-training**
Training steps	500k
Batch size	1024
Adam ϵ	1.00e-7
Adam β_1	0.9
Adam β_2	0.999
Weight decay	0
Peak learning rate	0.001
End learning rate	0.0001
Warmup steps	10000
Decay function	Cosine

Table 16: RETVec pre-training optimizer hyperparameters.

Dataset	Vectorizer	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
AG News	Whitespace	92.0%	91.6%	91.0%	90.2%	88.6%	86.8%	83.8%	79.3%	72.3%	63.5%	49.1%
	SentencePiece	91.8%	91.0%	89.9%	88.5%	86.2%	82.8%	78.7%	74.0%	67.9%	60.9%	51.5%
	BPE	91.6%	90.5%	89.1%	87.6%	84.8%	80.6%	76.1%	70.2%	63.8%	56.2%	46.3%
	fastText	92.8%	92.1%	91.6%	90.7%	89.1%	87.8%	85.1%	82.5%	77.2%	70.5%	59.7%
	RETVec-raw	91.0%	90.3%	89.6%	88.3%	87.3%	85.5%	83.7%	81.8%	78.2%	74.7%	68.5%
	RETVec	92.6%	92.1%	91.6%	91.1%	90.5%	89.9%	88.7%	87.3%	86.0%	84.2%	81.2%
Yelp P.	Whitespace	93.2%	92.5%	91.6%	90.5%	88.6%	86.7%	84.1%	80.4%	75.8%	69.7%	60.9%
	SentencePiece	93.1%	91.5%	89.7%	87.7%	85.1%	81.9%	78.7%	75.1%	71.2%	67.4%	62.6%
	BPE	93.0%	91.7%	90.2%	88.5%	86.2%	83.5%	80.5%	77.2%	73.1%	69.0%	64.2%
	fastText	94.1%	93.5%	92.8%	92.0%	90.7%	89.3%	87.9%	85.8%	83.1%	80.2%	75.6%
	RETVec-raw	92.4%	91.7%	90.9%	89.8%	88.6%	87.2%	85.7%	83.7%	81.2%	78.6%	74.5%
	RETVec	92.7%	92.2%	91.6%	90.9%	90.0%	89.2%	88.0%	86.8%	85.1%	83.5%	80.9%
Multilingual Amazon P.	Whitespace	92.7%	92.1%	91.6%	91.1%	90.2%	89.1%	87.9%	86.2%	83.7%	80.9%	75.0%
	SentencePiece	89.6%	88.5%	87.5%	86.2%	84.5%	82.6%	80.7%	78.7%	76.1%	73.9%	70.7%
	BPE	88.9%	87.7%	86.6%	85.4%	83.6%	81.8%	80.1%	77.9%	75.3%	73.4%	70.3%
	fastText	86.2%	85.5%	84.8%	84.0%	82.9%	81.0%	80.1%	77.9%	75.9%	73.9%	70.2%
	RETVec-raw	92.3%	91.6%	90.8%	90.0%	89.0%	87.7%	86.3%	84.8%	82.7%	80.6%	77.0%
	RETVec	92.9%	92.5%	92.0%	91.7%	91.3%	90.6%	90.2%	89.5%	88.6%	87.7%	86.1%
MASSIVE	Whitespace	70.0%	58.9%	58.8%	58.4%	55.8%	51.7%	49.2%	43.9%	37.1%	34.8%	17.3%
	SentencePiece	69.0%	58.5%	58.4%	57.8%	55.8%	52.7%	50.7%	46.7%	41.8%	40.9%	29.6%
	BPE	65.5%	54.6%	54.4%	54.1%	52.0%	48.6%	46.9%	42.7%	38.4%	37.0%	26.4%
	fastText	16.7%	15.5%	16.1%	15.2%	14.8%	14.3%	13.9%	13.7%	13.1%	13.0%	12.5%
	RETVec-raw	69.6%	59.7%	59.7%	59.2%	57.4%	54.4%	52.7%	49.0%	44.8%	43.6%	32.5%
	RETVec	73.2%	65.7%	65.7%	65.5%	63.9%	61.8%	60.4%	57.5%	54.1%	52.9%	43.5%

Table 17: Random mixed typo resilience results (0% to 100% word typo rate) for each classification dataset and vectorizer. Following the methodology described in Section 6, test accuracy on each dataset is reported and results are averaged across the three model architectures we benchmarked in 5.

F Typo Resilience Evaluation

Detailed results for random mixed typo resilience across every dataset and vectorizer can be found in Table 17.

G Adversarial Resilience Evaluation

We report adversarial attack resilience results for all vectorizers, classification models, and adversarial attack algorithms we benchmarked in Table 18. The TextAttack [20] framework was used to conduct all three types of adversarial attacks.

Model	Vectorizer	TextBugger			Pruthi			DeepWordBug		
		Original Acc	Acc under Atk	Atk Success %	Original Acc	Acc under Atk	Atk Success %	Original Acc	Acc under Atk	Atk Success %
LSTM	Whitespace	90.6%	9.9%	89.1%	90.6%	84.2%	7.1%	90.6%	9.9%	89.1%
	SentencePiece	90.3%	0.8%	99.1%	90.3%	68.1%	24.6%	90.3%	0.8%	99.1%
	BPE	88.1%	3.3%	96.3%	88.1%	73.3%	16.8%	88.1%	3.3%	96.3%
	fastText	92.7%	14.4%	84.5%	92.7%	83.3%	10.1%	92.7%	14.4%	84.5%
	RETVec-raw	90.8%	22.3%	75.4%	90.8%	74.8%	17.6%	90.8%	22.3%	75.4%
	RETVec	91.8%	23.7%	74.2%	91.8%	80.9%	11.9%	91.8%	23.7%	74.2%
CNN	Whitespace	90.9%	17.6%	80.6%	90.9%	83.8%	7.8%	90.9%	9.0%	90.1%
	SentencePiece	90.3%	2.9%	96.8%	90.3%	72.2%	20.0%	90.3%	3.5%	96.1%
	BPE	89.3%	31.8%	64.4%	89.3%	54.1%	39.4%	89.3%	43.5%	51.3%
	fastText	91.9%	17.3%	81.2%	91.9%	74.9%	18.5%	91.9%	15.4%	83.2%
	RETVec-raw	86.9%	30.1%	65.4%	86.9%	59.6%	31.4%	86.9%	37.7%	56.6%
	RETVec	91.4%	34.3%	62.5%	91.4%	77.4%	15.3%	91.4%	45.0%	50.8%
BERT	Whitespace	89.4%	9.8%	89.0%	89.4%	83.6%	6.5%	89.4%	3.3%	96.3%
	SentencePiece	89.8%	2.8%	96.9%	89.8%	70.8%	21.2%	89.8%	3.7%	95.9%
	BPE	90.8%	8.2%	91.0%	90.8%	78.2%	13.9%	90.8%	1.2%	98.7%
	fastText	92.6%	22.9%	75.3%	92.6%	80.5%	13.1%	92.6%	18.1%	80.5%
	RETVec-raw	93.0%	30.8%	66.9%	93.0%	82.2%	11.6%	93.0%	38.9%	58.2%
	RETVec	93.7%	30.1%	67.9%	93.7%	84.6%	9.7%	93.7%	40.5%	56.8%

Table 18: Detailed adversarial resilience results on AG News. Results are reported on the same randomly selected 1000 examples from the AG News test split, following the methodology described in Section 7.

H Pre-training and Fine-tuning BERT

Table 19 details the hyperparameter settings used for pre-training and fine-tuning BERT-Base models. Table 20 shows detailed results on the GLUE benchmark, including the models' average performance and standard deviation for each GLUE task.

Hyperparameter	Pre-training	Fine-tuning
Training steps	100k steps	20 epochs
Batch size	64	32
Sequence length	512	512
Adam ϵ	1e-8	1e-8
Adam β_1	0.9	0.9
Adam β_2	0.999	0.999
Weight decay	0.01	0.01
Max learning rate	5e-5	2e-5
End learning rate	0	0
Warmup steps	10000	First 5% of steps
Decay function	Linear	None

Table 19: Pre-training and fine-tuning hyperparameters for BERT-Base models described in Section 8.

Vectorizer	MNLI	QNLI	QQP	RTE	SST-2	MRPC	CoLA	STS-B	GLUE Avg
SentencePiece	80.6 (0.1)	88.4 (0.3)	90.1 (0.0)	**66.1 (1.0)**	90.8 (0.3)	85.4 (0.15)	**50.3 (0.4)**	**82.0 (0.5)**	**79.2 (0.3)**
RETVec-raw	**82.0 (0.5)**	**89.5 (0.1)**	**90.4 (0.1)**	64.5 (0.9)	**91.5 (0.1)**	86.3 (0.9)	47.9 (1.1)	79.1 (0.2)	78.9 (0.4)
RETVec	80.9 (0.4)	88.9 (0.3)	90.4 (0.1)	65.0 (0.7)	90.7 (0.2)	**86.9 (0.4)**	47.2 (0.7)	79.6 (0.2)	78.7 (0.3)

Table 20: Detailed results on GLUE Benchmark for pre-trained BERT-Base models using RETVec compared to SentencePiece. Each model is trained three times with different seeds, and the average and standard deviation is reported here. **Bold** indicates best results, underline indicates second best.

I fastText Word Dataset

Table 21 contains statistics on word length computed on the fastText word dataset using words from all 157 available languages.

Avg	Median	Std	p90	p95	p99	p99.9
8.4	7.9	4.6	13.0	**15.0**	20.8	**36.1**

Table 21: Word length statistics computed on all fastText words from 157 languages. p90 denotes the 90th percentile, p95 denotes the 95th percentile, and so on.