Supplementary Material for:
A Survey on Data Augmentation for Text Classification

MARKUS BAYER, MARC-ANDRÉ KAUFHOLD, and CHRISTIAN REUTER, PEASEC,
Technical University of Darmstadt

3 TEXTUAL DATA AUGMENTATION METHODS

3.1 Data Space

3.1.2 Word Level.

3.1.2.2 Synonym Replacement.

Table 2. Overview of Different Approaches of the Synonym Replacement Method

	Synonym Database	**Replacement Method**	**Synonym Selection**	**Model Base**	**Dataset**	**Improvements**
[57]	WordNet	Headword replacement	Not stated	Logistic Regression	TempEval Reuters (12) Wikipedia (1)	−1 (F1) −0.6 −0.1
[61], [63]	mytheas (LibreOffice) WordNet-based	Randomly chosen number of words based on geometric distribution.	Randomly based on geometric distribution.	Character CNN	 AG News DBP. Yelp P. Yelp F. Yahoo A. Amazon F. Amazon P.	[61]/[63] (Acc.) −0.38 /−0.57 +0.05 /+0.13 −0.03/ +0.36 +0.22/0.65 +0.1/0.1 −0.17/−0.17
[39]	WordNet	Substitutable words are nouns, verbs, adjectives, or adverbs that are not part of a named entity. Each word is replaced with a certain probability.	The remaining probability of substitution is shared among the synonyms based on a language model score.	CNN	MR CR Subj SST MR/CR CR/MR	+0.8 (Acc.) +1.2 +0.5 +0.1 0.9 0.3
[9]	WordNet	Only adverbs and adjectives, sometimes nouns, more rarely verbs.	Most similar companion information of the synonym with the context of the chosen word.	XGBoost MLP (2 hidden layer)	IMDB	+0.5 (Acc.) +4.92
[59]	WordNet	No pronouns, conjunctions, prepositions, and articles for replacement. Choosing uniform randomly.	Uniform random	CNN with word embedding	Toxic Comment Classification	−0.09/−0.21 (AUC)
[60]	HIT IR-Lab Tongyici Cilin (Extended) (Chinese)	No time words, prepositions, and mimetic words. Chi-square statistics method.	Chi-square statistics method	Character CNN-SVM	Hotel R. Laptop R. Book R.	∼+1 (Acc.) ∼+1 ∼+0.25

(Continued)

0360-0300/2022/12-ART146 $15.00
https://doi.org/10.1145/3544558

Table 2. Continued

	Synonym Database	Replacement Method	Synonym Selection	Model Base	Dataset	Improvements
[62]	WordNet	Verbs, nouns, and their combination. Geometric distribution.	Geometric distribution	DNN	AG News Sentiment Hate Speech	∼+0.4 (Acc.) ∼+0 ∼−0.8
[64]	WordNet & Thesaurus. com	For Minibatch: Augmentation with probability. Replacement of those words that belong to certain POS-tags. One replacement of a word per sentence that maximizes loss.	Synonym that maximizes the loss.	Kim CNN	TREC	+1.2 (Acc.)
[2]	WordNet	No stop words. Choosing n random words to be replaced (SR) or from which the synonyms are inserted at a random position (RI)	Uniform random	CNN	Classification tasks (500) (2000) (5000) (full)	SR / RI (Acc.) ∼+1.9/∼+2.0 ∼+1.2/∼+0.9 ∼+0.7/∼+0.6 ∼+1.0/∼+0.9
[1]	WordNet	Replacement of a word based on a certain probability.	Temperature hyperparameter learned while training.	CNN	SST-5 SST-2 Subj MPQA RT TREC	−0.6 (Acc.) +0.5 +0 +0.2 +0.1 −0.4
[40]	WordNet	Replacement of a word based on a certain probability.	Temperature hyperparameter learned while training.	TextRCNN	ICS NEWS	−0.26 (Macro F1) +1.63
[49]	Not stated	Filtering words according to their POS-tag. Selecting a fixed or variable number of words.	Specific or variable number of synonyms.	LSTM-CNN	Tan NLPCC	Results only in combination with other methods
[65]	WordNet	Not stated	Not stated	BERT	SST-5 (40) IMDB (40) TREC (40)	−0.87 (Acc.) −0.87 +0.01
[66]	WordNet	No stop words. 10% of documents randomly selected.	Not stated	M-BERT	CodiEsp-D CodiEsp-P	+0.6 (F1) −0.7 (F1)
[37]	WordNet	Keywords replaced are ordered by their RAKE score (e.g., the probability of being a keyword).	Randomly selected. Replacement only with same POS-tag.	No model (intrinsic evaluation with different metrics)	Yelp-LR (small subset of Yelp Reviews)	+0.015 (SBLEU) −0.018 (UTR) −0.02 (TTR) −0.016 (RWords) 0 (SLOR) −0.007 (BPRO) +0.001 (SStd) 0 (SDiff)
[46]	WordNet	No stop words. Uniform random replacement until 20% of the words in a sentence are changed.	Uniform random	CNN	Yelp P.	Only against other data augmentation methods

3.1.2.3 Embedding Replacement.

Table 3. Overview of Different Approaches of the Embedding Replacement Method

	Replacement Selection	Embedding Selection	Model Base	Dataset	Embedding Model	Improvements
[70]	Not stated	K-nearest-neighbor and cosine similarity	Logistic regression	Petpeeve dataset	UrbanDictionary W2V Twitter W2V GoogleNews W2V	+0.3 (F1) +1.7 +2.4
[62]	Random	Random with probability proportional to cosine similarity.	DNN	AG News Sentiment Hate Speech	Wikipedia W2V Wikipedia W2V GloVe Twitter	~0 (Acc.) ~+0.5 ~−0.3
[48]	Every word	Cosine similarity threshold + POS-tag matching	CNN+LSTM/GRU	HON RSN-1 RSN-2	Word2Vec Hate Speech FastText Wikipedia GoogleNews W2V GloVe Common Crawl GloVe Common Crawl GloVe Common Crawl	−22.7 (Macro F1) +1.0 −3.3 +0.3 −0.2 0
[49]	1. Method: Filtering words according to their POS-tag. Selecting a fixed or variable number of words. 2. Method: Replacing adverbial phrases (Chinese related).	Own similarity measure and specific or variable number of replacements	LSTM-CNN	Tan	W2V self-pretrained	Results only in combination with other word level augmentation methods
[39]	Substitutable words are nouns, verbs, adjectives, or adverbs that are not part of a named entity. Each word is replaced with a certain probability.	Embeddings are found with the counter-fitting method. Each candidate is replaced with a probability. The remaining probability of substitution is shared among the embeddings based on a language model score.	CNN	MR CR Subj SST MR/CR CR/MR	GoogleNews W2V	−0.6/−4.2 (Acc) +0.1/−3.7 +0.2/−1.4 −0.4/−4.2 +1.9/+0.4 +0.1/−3.0
[41]	Not stated	Cosine similarity	Random Forest, Naïve Bayes, SVM	Vietnamese comments	W2V Vietnamese	Results only in combination
[20]	Random sampling with probabilities proportional to the neighbors each word has within the counter-fitted embedding space + exclusion of common articles and prepositions.	1. K-nearest-neighbors with Euclidean distance + counter-fitting method. 2. Google LM to filter out words. 3. Selection of the word that will maximize the target label prediction probability.	LSTM	IMDB	GloVe	Adversarial training: No improvements but safer model
[71]	Only for multi-piece words. Random probability for replacement.	Random embedding of the k-nearest-neighbor	Small transformer model	Various GLUE tasks	GloVe	No augmentation baseline comparisons
[72]	No stop words or symbolic and numerical data	Cosine similarity threshold of 0.97	Manhattan LSTM model	Thai text similarity task	Thai2fit (Thai language)	+1.71

3.1.2.4 *Replacement by Language Models.*

Table 4. Evaluation Results of the State of the Art Language Substitution Method c-BERT

Publication	Method	Dataset	Improvements (Accuracy)
[74]	c-BERT	SST-5	+0.8 (CNN)/+1.3 (RNN)
		SST-2	+0.2 (CNN)/+0.5 (RNN)
		Subj	+0.5 (CNN)/+0.4 (RNN)
		MPQA	+0.5 (CNN)/+0.7 (RNN)
		RT	+0.8 (CNN)/+0.6 (RNN)
		TREC	+0.8 (CNN)/+0.2 (RNN)
[76]	c-BERT with consistency training	MLNI-m	+0.4 (RoBERTa-Base)
[42]	c-BERT	ATIS	−1.9 (BERT)/−0.8 (SVM)/−5.8 (LSTM)
		TREC	+1.1 (BERT)/+1.1 (SVM)/+6.5 (LSTM)
		WVA	+0.2 (BERT)/0.5 (SVM)/+2.4 (LSTM)
[65]	c-BERT integrated in reinforcement learning scheme	SST-5 (42)	+1.17 (BERT)/+2.19 (normal c-BERT)
		IMDB (45)	+1.97 (BERT)/+1.97 (normal c-BERT)
		TREC (45)	+0.73 (BERT)/+0.87 (normal c-BERT)
[71]	c-BERT and embedding substitution for multiple-pieces words	MNLI-m	+2.3 (TinyBERT)
		MNLI-mm	+1.9 (TinyBERT)
		MRPC	+3.4 (TinyBERT)
		CoLA	+21.0 (TinyBERT)

3.1.4 Document Level.

3.1.4.1 Round-trip Translation.

Table 5. Overview of the Round-trip Translation Approaches

	Translation Model	Languages	Filtering	Model	Dataset	Improvements
[92]	Google's NMT [96]	en → fr → en	No filtering	Convolution and self-attention model	SQuAD	+1.5 (EM)/+1.1 (F1)
[9]	Google Translate API	Not stated	Excluding identical instances. Similarity threshold based on lengths.	XGBoost MLP 2 hidden layer	IMDB	+0 (Acc.) +5.8
[69]	Google Translate API	en → fr, es, de, hi → en	No filtering	NBSVM CNN LSTM BiLSTM CNN-LSTM LSTM-CNN CNN-BiLSTM BiLSTM-CNN	Aggression Detection	+0.19 (Macro F1) +5.31 +7.39 +5.6 +5.94 +19.45 +14.33 +6.87
[94]	Google Translate	Randomly selected	No filtering	Fusion CNN	TREC Incident Streams track	~−1.2 (F1)
[62]	Google Translate API & Amazon translate	en → fr, de → en	"We ensured that the [...] texts carry the same meaning as the source text"	DNN	AG News Hate Speech	~+0.33 (Acc.) ~−2.3
[50]	WMT'14 English-French translation model	en → fr → en	No filtering	Randomly initialized transformer	Yelp-5	+1.65 (Acc.)
[76]	WMT19 and released in FairSeq	en → de → en	No filtering	RoBERTa	MLNI-m	+0.9 (Acc.)
[97]**	Translation models from Britz et al. [98]	en → de, zh → en	No filtering	BERT	MNLI QNLI QQP RTE SST-2 MRPC CoLA STS-B	+0 (Acc.) +0.2 (Acc.) +0.4 (Acc.) +3.6 (Acc.) +0.7 (Acc.) +0 (F1) +2.3 (Mcc) +0.6 (Corr.)
[99]*	Not stated	Not stated	No filtering	Transformer base with consistency training	MNLI QNLI QQP RTE SST-2 MRPC CoLA STS-B	+0.9 (Acc.) +0.6 (Acc.) −0.2 (Acc.) +5.1 (Acc.) +0.7 (Acc.) +2.6 (F1) +1.4 (Mcc) +0.4 (Corr.)
[100]	MarianMT	en → fr, de, es → en Chained: en → es → fr → en	Word sense disambiguation: Retaining of those in which the target word occurs exactly once (in both original and augmented instance).	MT-DNN	SemEval-2013 + SemEval-2015 + Senseval-2 + Senseval-3	No baseline comparisons

*Trained with consistency training.
**Trained with contrastive learning.

3.1.4.2 *Generative Methods.*

Table 6. Overview of Text Generation Methods

Publication	Method	Model	Dataset	Improvements
[40]	VAE CVAE + prior sampling CVAE + posterior sampling	Ensemble of BiLSTM, TextCNN, TextRCNN, and FastText with XGBoost as top-level classifier	ICS (Zh) News Category Dataset (EN) ICS (Zh) News Category Dataset (EN) ICS (Zh) News Category Dataset (EN)	+0.04 (F1) +2.02 −0.13 +1.55 −0.06 +1.88
[101]	VAE CVAE + prior sampling CVAE + posterior sampling	BiLSTM	Movie Movie + Live Entertainment Movie Movie + Live Entertainment Movie Movie + Live Entertainment	+4.0 (Macro F1) −0.5 +5.9 +1.7 +5.6 +0.6
[117]	CVAE	BERT	SNIPS (few shot) SNIPS FBDialog (few shot) FBDialog	+8.00 +0.06 (Acc.) +7.42 +0.0
[103]	Transformer-based sentence editor	CNN CNN CNN CNN LSTM LSTM	Subj (20%) Subj (100%) SST-2 (20%) SST-2 (100%) Amazon Reviews (1%) Amazon Reviews (4%)	+1.71 (Acc.) +1.62 +0.87 −0.84 +1.12 +0.41
[48]	RNN LM with random start word priming	CNN+LSTM + GloVe++	HON RSN-1 RSN-2	−1.8 (Micro-F1) +8.2 −7.4
[66]	CNN-LSTM LM with 30% of a given sentence for priming	CNN-LSTM	CodiEsp-P	+3.1 (F1)
[49]	seqGAN	LSTM + pretrained embeddings CNN + pretrained embeddings LSCNN + pretrained embeddings	Tan's task	+1.06 (F1) +0.9 +0.8
[105]	CS-GAN (GAN, RNN and reinforcement learning)	CNN	Amazon-5000 Amazon-30000 Emotion-15000 NEWS-15000	+1.6 (Acc.) −0.21 +0.77 +2.25
[106]	GPT-2 for rarer instances without filtering	Logistic regression/biLSTM/Bi-attentive classification + ELMo + GloVe	Alerting Information Feed Prioritization	No comparative results
[42]	CVAE	BERT	ATIS (5) TREC (5) WVA (5)	+7.3 (Acc.) +0.8 −1.8
[42]	LAMBADA – GPT-2 generation and classifier filtering	BERT	ATIS (5) ATIS (20) ATIS (50) ATIS (100) TREC (5) WVA (5)	+22.4 (Acc.) ~0 ~+2.0 ~+0.5 +4.0 +1.4
[107]	PREDATOR – DistilGPT2 generation and classifier filtering	BERT CNN BERT	AG-NEWS CyberTrolls SST-2	+0.61 (Acc.) +0.45 +1.63

(Continued)

Table 6. Continued

Publication	Method	Model	Dataset	Improvements
[44]	GPT-2 with conditional fine-tuning, special prompting, and document embedding filtering	ULMFit	SST-2 (100) SST-2 (700) Layoff Management Change Mergers & Acquisitions Flood Wildfire Boston Bombings Bohol Earthquake West Texas Explosions Dublin New York	+15.53 (Acc.) −0.19 (Acc.) +4.84 (F1) +3.42 (F1) +1.42 (F1) +0.25 (F1) +0.44 (F1) +2.44 (F1) +2.05 (F1) +3.81 (F1) −2.54 (F1) +0.44 (F1)
[108]	GPT-2 with conditional fine-tuning, special prompting, and classifier filtering	RoBERTa FlauBERT	MediaEval CLS-FR	+0.55 (micro-F1) +0.57
[109]	GPT-2 with a reinforcement learning component for class conditional generation.	XLNet	Offense Detection (20%) Offense Detection (40%) Sentiment Analysis (20%) Sentiment Analysis (40%) Irony Classification (20%) Irony Classification (40%)	+1.3 (F1) +4.3 +1.2 +1.4 +1.0 +2.3
[112]	GPT-3 with prompt-based generation and pseudo-labeling	BERT (base) BERT (large)	COLA (0.1%, 0.3%, 1.0%) TREC6 (0.1%, 0.3%, 1.0%) CR (0.1%, 0.3%, 1.0%) SUBJ (0.1%, 0.3%, 1.0%) MPQA (0.1%, 0.3%, 1.0%) RT20 (0.1%, 0.3%, 1.0%) SST-2 (0.1%, 0.3%, 1%, full) SST-2 (0.1%, 0.3%, 1.0%)	+7.9, 3.2, −2.4 +15.6, 17.1, −6.5 +11.0, 17.3, 8.9 +1.3, −1.8, −1.2 +12.9, 13.4, 3.8 +6.2, 13.6, 17.5 +20.9, 19.3, 5.7, 2.9 +23.7, 14.6, 3.0

3.2 Feature Space

3.2.2 *Interpolation Methods.*

3.2.2.2 *Mixup Interpolation.*

Table 8. Overview of Different Approaches of the Replacement Method "Mixup Interpolation"

Method	Technique for textual application	Model	Datasets	Improvements
mixup by Marivate and Sefara [62]	Not stated	DNN	AG News Sentiment 140 Hate Speech	+0.2 (Acc.) +0.4 +0
[117]	Interpolation of the BERT CLS output	BERT-base-english-uncased	SNIPS (few shot) SNIPS FBDialog (few shot) FBDialog	+8.36 (Acc.) +0.0 +7.92 +0.08
[76]	Interpolation of the embedding matrices	RoBERTa-base	MNLI-m	+0.6 (Acc.)
wordMixup by Guo et al. [137]	Interpolation of zero-padded word embeddings	CNN	Trec SST-1 SST-2 Subj MR	+1.6 (Acc.) +1.9 +0.2 +0.3 +1.5
senMixup by Guo et al. [137]	Interpolation on the final hidden layer	CNN	Trec SST-1 SST-2 Subj MR	+1.2 (Acc.) +2.3 +0.3 +0.5 +0.8
Nonlinear Mixup by Guo [138]	Nonlinear interpolation of padded word embeddings	CNN	Trec SST-1 SST-2 Subj MR	+2.6 (Acc.) +3.0 +2.3 −0.5 +3.6
Mixup-Transformer by Sun et al. [136]	Interpolation after last layer of the transformer	BERT-large	CoLA SST-2 MRPC STS-B QQP MNLI-mm QNLI RTE	+2.68 (Corr.) +0.81 (Acc.) +1.72 (Acc.) +0.89 (Corr.) +0.42 (Acc.) −0.01 (Acc.) +0.13 (Acc.) +0.37(Acc.)
TMix by Chen et al. [95]	Interpolation of the m-th BERT layer (7, 9, and 12 randomly chosen per batch)	BERT-base-uncased + average pooling + two-layer MLP	AG News (10) AG News (2500) DBPedia (10) DBPedia (2500) Yahoo! (10) Yahoo! (2500) IMDB (10) IMDB (2500)	+4.6 (Acc.) +0.2 +1.6 +0.0 +2.4 +0.3 +1.8 +0.5
TMix evaluated by [112]	Interpolation of the m-th BERT layer (7, 9, and 12 randomly chosen per batch)	BERT-base	SST-2 (0.1, 0.3, 1.0%) COLA (0.1, 0.3, 1.0%) TREC6 (0.1, 0.3, 1.0%) CR (0.1, 0.3, 1.0%) SUBJ (0.1, 0.3, 1.0%) MPQA (0.1, 0.3, 1.0%) RT20 (0.1, 0.3, 1.0%)	−0.2, −1.5, −2.1 +0.8, 2.4, −0.7 −0.2, −1.4, +2.4 −0.1, −0.5, −3.3 −0.5, +0.4, −0.1 +0.2, 2.9, 0.0 +2.3, 0.6, −1.9

(Continued)

Table 8. Continued

Method	Technique for textual application	Model	Datasets	Improvements
Intra-LADA [140]	Interpolation of an instance with a randomly reordered version of itself	BERT-base-multilingual-cased + linear layer	CoNLL (5%) CoNLL (100%) GermEval (5%) GermEval (100%)	+0.24 (F1) +0.03 (*) +0.29 +0.04 (*)
Inter-LADA [140]	Interpolation of the nearest neighbors and sometimes randomly selected instances	BERT-base-multilingual-cased + linear layer	CoNLL (5%) CoNLL (100%) GermEval (5%) GermEval (100%)	+1.32 (F1) +0.64 +0.49 +0.33
Intra-Inter-LADA [140]	Combination of Intra- and Inter-LADA	BERT-base-multilingual-cased + linear layer	CoNLL (5%) CoNLL (30%) GermEval (5%) GermEval (30%)	+1.57 (F1) +0.59 +0.53 +0.78

*Included in the pretraining.

3.3 Training Strategies

Table 9. Overview of the Contrastive Learning Works Using Data Augmentation

Publication	Method	Model	Dataset	Improvement
CERT [97]	Continued contrastive self-supervised learning (MLM) with round-trip translation.	Continued BERT training compared to BERT	MNLI QNLI QQP RTE SST-2 MRPC CoLA STS-B	+0 (Acc.) +0.2 (Acc.) +0.4 (Acc.) +3.6 (Acc.) +0.7 (Acc.) +0 (F1) +2.3 (Mcc) +0.6 (Corr.)
CODA [76]	Supervised contrastive learning with adversarial training and round-trip translation.	RoBERTa-base finetuning compared to RoBERTa-base	MNLI-m QNLI SST-2 RTE MRPC	+0.5 (Acc.) +0.8 +0.5 +3.3 +1.5
CLEAR [47]	Contrastive self-supervised learning (MLM) from scratch with word deletion, span deletion, random reordering, synonym substitution (also in combination). Span deletion + random reordering is shown here.	Transformer architecture by Vaswani et al. [160] trained from scratch compared to RoBERTa-base	MNLI QNLI QQP RTE SST-2 MRPC CoLA STS-B	−0.5 (Acc.) −0.3 (Acc.) +1.8 (Acc) +6.5 (Acc) +0.1 (Acc) +2.8 (F1) +8.2 (Mcc) +0.4 (Corr)
ConSERT [144]	Continued contrastive self-supervised learning (MLM) with adversarial training, token shuffling, cutoff, dropout.	Continued BERT-large training compared to SBERT-large-NLI	STS12 STS13 STS14 STS15 STS16 STSb SICK-R	+0.99 (S. Corr.) +3.9 +2.83 +2.85 +2.5 +2.31 +4.89
C²L [51]	Supervised contrastive learning with counterfactual augmentation.	BERT-base finetuning compared to BERT-base	CF-IMDb CF-IMDb Revised CF-NLI CF-NLI RP CF-NLI RH CF-NLI RP & RH	+0.7 (Acc.) +3.3 −1.2 +2.2 +1.3 +1.8

REFERENCES

[1] S. Kobayashi. 2018. Contextual augmentation: Data augmentation bywords with paradigmatic relations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2. 452–457. DOI: 10.18653/v1/n18-2072

[2] J. Wei and K. Zou. 2020. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*. 6382–6388. DOI: 10.18653/v1/d19-1670

[3] K. Kafle, M. Yousefhussien, and C. Kanan. 2017. Data augmentation for visual question answering. In *INLG'17*. 198–202. DOI: 10.18653/v1/w17-3529

[4] S. Longpre, Y. Wang, and C. DuBois. 2020. How Effective is task-agnostic data augmentation for pretrained transformers? In *EMNLP'20*. 4401–4411. DOI: 10.18653/v1/2020.findings-emnlp.394

[5] C. Shorten and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data* 6, 1, (2019). DOI: 10.1186/s40537-019-0197-0

[6] Q. Wen et al. 2020. Time series data augmentation for deep learning: A survey. 2020. arXiv:2002.12478. Retrieved from http://arxiv.org/abs/2002.12478.

[7] L. Taylor and G. Nitschke. 2019. Improving deep learning with generic data augmentation. In *SSCI'18*. 1542–1547. DOI: 10.1109/SSCI.2018.8628742

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2323. DOI: 10.1109/5.726791

[9] C. Coulombe. 2018. Text data augmentation made simple by leveraging NLP cloud APIs. http://arxiv.org/abs/1812.04718.

[10] R. Müller, S. Kornblith, and G. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems* 32.

[11] M. Banko and E. Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *ACL'01*. 26–33. DOI: 10.3115/1073012.1073017

[12] L. Palen and K. M. Anderson. 2016. Crisis informatics-new data for extraordinary times: Focus on behaviors, not on fetishizing social media tools. *Science* (80-.) 353, 6296 (2016), 224–225. DOI: 10.1126/science.aag2579

[13] M. Bayer, M.-A. Kaufhold, and C. Reuter. 2021. Information overload in crisis management: Bilingual evaluation of embedding models for clustering social media posts in emergencies. *ECIS'21*.

[14] M. A. Kaufhold, M. Bayer, and C. Reuter. 2020. Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Inf. Process. Manage.* 57, 1 (2020). DOI: 10.1016/j.ipm.2019.102132

[15] M.-A. Kaufhold, A. S. Basyurt, K. Eyilmez, M. Stöttinger, and C. Reuter. 2022. Cyber threat observatory: Design and evaluation of an interactive dashboard for computer emergency response teams. In *ECIS'22*. 6–18.

[16] M.-A. Kaufhold et al. 2021. CYWARN: Strategy and technology development for cross-platform cyber situational awareness and actor-specific cyber threat communication. In *Workshop of Mensch und Computer '21*. 1–9.

[17] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. 2021. Deep learning–based text classification. *ACM Comput. Surv.* 54, 3 (April 2021). DOI: 10.1145/3439726

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15 (2014), 1929–1958.

[19] S. Geman, E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Comput.* 4, 1 (1992), 1–58. DOI: 10.1162/neco.1992.4.1.1

[20] A. Hernández-García and P. König. 2018. Data augmentation instead of explicit regularization. arXiv:1806.03852. Retrieved from http://arxiv.org/abs/1806.03852.

[21] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *ACL'18* 2 (2018), 31–36. DOI: 10.18653/v1/p18-2006

[22] M. Alzantot, Y. Sharma, A. Elgohary, B. J. Ho, M. B. Srivastava, and K. W. Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. 2890–2896. DOI: 10.18653/v1/d18-1316

[23] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. 2016. Understanding data augmentation for classification: When to warp? In *DICTA'16*. DOI: 10.1109/DICTA.2016.7797091

[24] P. Y. Simard, D. Steinkraus, and J. C. Platt. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *ICDAR'03*, 958–963. DOI: 10.1109/ICDAR.2003.1227801

[25] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE IROS'17*. 23–30. DOI: 10.1109/IROS.2017.8202133

[26] H. Inoue. 2018. Data augmentation by pairing samples for images classification. arXiv:1801.02323. Retrieved from http://arxiv.org/abs/1801.02929.

[27] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. 2020. Random erasing data augmentation. *AAAI'20*. 13001–13008. DOI: 10.1609/aaai.v34i07.7000
[28] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. In *Proceedings of the International Symposium on Biomedical Imaging*. 289–293.
[29] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. 2018. Emotion classification with data augmentation using generative adversarial networks. In *Lecture Notes in Computer Science*. 349–360. DOI: 10.1007/978-3-319-93040-4_28
[30] N. Carlini et al. 2020. Extracting training data from large language models. arXiv:2012.07805. Retrieved from http://arxiv.org/abs/2012.07805.
[31] X. Cui, V. Goel, and B. Kingsbury. 2014. Data augmentation for deep neural network acoustic modeling. In *ICASSP'14*. 5582–5586. DOI: 10.1109/ICASSP.2014.6854671
[32] N. Jaitly and G. E. Hinton. 2013. Vocal tract length perturbation (VTLP) improves speech recognition. In *Proceedings of the 30th International Conference on Machine Learning* 90, 42–51.
[33] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales. 2014. Data augmentation for low resource languages. In *INTERSPEECH'14*. 810–814.
[34] A. Hannun et al. 2014. Deep speech: Scaling up end-to-end speech recognition. arXiv:1412.5567. Retrieved from http://arxiv.org/abs/1412.5567.
[35] Z. Xie et al. 2017. Data noising as smoothing in neural network language models. In *ICLR'17*.
[36] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. 2015. Audio augmentation for speech recognition. In *INTERSPEECH'15*. 3586–3589.
[37] S. Feng et al. 2021. A survey of data augmentation approaches for NLP. In *ACL'21*. 968–988. DOI: 10.18653/v1/2021.findings-acl.84
[38] Y. Belinkov and Y. Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *ICLR'18*.
[39] S. Y. Feng, V. Gangal, D. Kang, T. Mitamura, and E. Hovy. 2020. GenAug: data augmentation for finetuning text generators. In *Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 29–42. DOI: 10.18653/v1/2020.deelio-1.4
[40] T. Miyato, A. M. Dai, and I. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *ICLR'17*.
[41] Y. Li, T. Cohn, and T. Baldwin. 2017. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2. 21–27. DOI: 10.18653/v1/e17-2004
[42] S. Qiu et al. 2020. EasyAug: An automatic textual data augmentation platform for classification tasks. In *Companion of the World Wide Web Conference*. 249–252. DOI: 10.1145/3366424.3383552
[43] T. H. Huong and V. T. Hoang. 2020. A data augmentation technique based on text for Vietnamese sentiment analysis. In *ACM International Conference Proceeding Series*. 1–5. DOI: 10.1145/3406601.3406618
[44] A. A. Tavor et al. 2020. Do not have enough data? Deep learning to the rescue!. In *Proceedings of the 34th AAAI Conf. Artif. Intell.* 7383–7390. DOI: 10.1609/aaai.v34i05.6233
[45] V. Kumar, A. Choudhary, and E. Cho. 2020. Data augmentation using Pre-trained transformer models. arXiv:2003.02245. Retrieved from http://arxiv.org/abs/2003.02245.
[46] M. Bayer, M.-A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter. 2022. Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *Int. J. Mach. Learn. Cybernet.* (2022). DOI: 10.1007/s13042-022-01553-3
[47] S. T. Luu, K. Van Nguyen, and N. L.-T. Nguyen. 2020. Empirical study of text augmentation on social media text in vietnamese. arXiv:2009.12319. Retrieved from http://arxiv.org/abs/2009.12319.
[48] O. Kashefi and R. Hwa. 2020. Quantifying the evaluation of heuristic methods for textual data augmentation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT'20)*. 200–208. DOI: 10.18653/v1/2020.wnut-1.26
[49] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun, and H. Ma. 2020. CLEAR: Contrastive learning for sentence representation. http://arxiv.org/abs/2012.15466.
[50] G. Rizos, K. Hemker, and B. Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the International Conference on Information and Knowledge Management*. 991–1000. DOI: 10.1145/3357384.3358040
[51] X. Sun and J. He. 2020. A novel approach to generate a large scale of supervised data for short text sentiment analysis. *Multimed. Tools Appl.* 79, 9–10 (2020), 5439–5459. DOI: 10.1007/s11042-018-5748-4
[52] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong, and Q. V. Le. 2019. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1904.12848.
[53] S. Choi, M. Jeong, H. Han, and S.-W. Hwang. 2022. C 2 L: Causally contrastive learning for robust text classification. www.aaai.org.

[54] Y. Cheng, L. Jiang, and W. Macherey. 2020. Robust neural machine translation with doubly adversarial inputs. In *ACL'19*. 4324–4333. DOI: 10.18653/v1/p19-1425

[55] X. Wang, H. Pham, Z. Dai, and G. Neubig. 2020. Switchout: An efficient data augmentation algorithm for neural machine translation. In *EMNLP'18*. 856–861. DOI: 10.18653/v1/d18-1100

[56] J. Andreas. 2020. Good-Enough compositional data augmentation. In *ACL'20*. 7556–7566. DOI: 10.18653/v1/2020.acl-main.676

[57] D. Guo, Y. Kim, and A. Rush. 2020. Sequence-Level mixed sample data augmentation. In *EMNLP'20*. 5547–5552. DOI: 10.18653/v1/2020.emnlp-main.447

[58] G. Kurata, B. Xiang, and B. Zhou. 2016. Labeled data generation with encoder-decoder LSTM for semantic slot filling. In *INTERSPEECH*. 725–729, 2016. DOI: 10.21437/Interspeech.2016-727

[59] O. Kolomiyets, S. Bethard, and M. F. Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 2. 271–276.

[60] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *Int. J. Lexicogr.* 3, 4 (1990), 235–244. DOI: 10.1093/ijl/3.4.235

[61] A. V. Mosolova, V. V. Fomin, and I. Y. Bondarenko. 2018. Text augmentation for neural networks. In *CEUR Workshop Proceedings*, 104–109.

[62] X. Wang, Y. Sheng, H. Deng, and Z. Zhao. 2019. Charcnn-svm for chinese text datasets sentiment classification with data augmentation. *Int. J. Innov. Comput. Inf. Control* 15, 1 (2019), 227–246. DOI: 10.24507/ijicic.15.01.227

[63] X. Zhang, J. Zhao, and Y. Lecun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, vol. 2015-Janua. 649–657.

[64] V. Marivate and T. Sefara. 2020. Improving short text classification through global augmentation methods. In *Lecture Notes in Computer Science* 12279, 385–399. DOI: 10.1007/978-3-030-57321-8_21

[65] X. Zhang and Y. LeCun. 2015. Text understanding from scratch. http://arxiv.org/abs/1502.01710.

[66] M. Jungiewicz and A. Smywiński-Pohl. 2019. Towards textual data augmentation for neural networks: Synonyms and maximum loss. *Comput. Sci.* 20, 1 (2019), 57–84. DOI: 10.7494/csci.2019.20.1.3023

[67] Z. Hu, B. Tan, R. Salakhutdinov, T. Mitchell, and E. P. Xing. 2019. Learning data manipulation for augmentation and weighting. In *Advances in Neural Information Processing Systems*, Vol. 32. 15764–15775.

[68] A. Ollagnier and H. Williams. 2020. Text augmentation techniques for clinical case classification. In *Workshop Proceedings (CLEF'20)*, vol. 2996. 22–25. https://temu.bsc.es/codiesp/index.php/2019/09/19/resources/.

[69] Z. S. Harris. 1954. Distributional structure. *WORD* 10, 2–3 (1954), 146–162. DOI: 10.1080/00437956.1954.11659520

[70] J. Firth. 1962. A synopsis of linguistic theory. In *Studies in Linguistic Analysis*. 1–32.

[71] S. T. Aroyehun and A. Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Trolling, Aggress. Cyberbullying (TRAC'18)*. 90–97.

[72] W. Y. Wang and D. Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Conference on Empirical Methods in Natural Language Processing*. 2557–2563. DOI: 10.18653/v1/d15-1306

[73] T. Phreeraphattanakarn and B. Kijsirikul. 2021. Text data-augmentation using text similarity with manhattan siamese long short-term memory for thai language. *J. Phys. Conf. Ser.* 1780, 1 (2021), 12018. DOI: 10.1088/1742-6596/1780/1/012018

[74] N. Mrkšić et al. 2016. Counter-fitting word vectors to linguistic constraints. In *NAACL HLT'16*. 142–148. DOI: 10.18653/v1/n16-1018

[75] X. Wu, S. Lv, L. Zang, J. Han, and S. Hu. 2019. Conditional BERT contextual augmentation. In *Lecture Notes in Computer Science* 11539, 84–95. DOI: 10.1007/978-3-030-22747-0_7

[76] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT'19*. 4171–4186.

[77] Y. Qu, D. Shen, Y. Shen, S. Sajeev, J. Han, and W. Chen. 2020. CoDA: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. arXiv 5 (2020), 1–14. http://arxiv.org/abs/2010.08670.

[78] X. Jiao et al. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics Findings of ACL (EMNLP'20)*. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372

[79] J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP'14*. 1532–1543. DOI: 10.3115/v1/d14-1162

[80] F. Gao et al. 2020. Soft contextual data augmentation for neural machine translation. In *ACL'20*. 5539–5544. DOI: 10.18653/v1/p19-1555

[81] A. J. Ratner, H. R. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré. 2017. Learning to compose domain-specific transformations for data augmentation. In *NeurIPS'17*. 3237–3247.

[82] M. Fadaee, A. Bisazza, and C. Monz. 2017. Data augmentation for low-Resource neural machine translation. In *ACL'17*. 2 (2017), 567–573. DOI: 10.18653/v1/P17-2090
[83] G. G. Şahin and M. Steedman. 2018. Data augmentation via dependency tree morphing for low-resource languages. In *EMNLP'18*. 5004–5009. DOI: 10.18653/v1/d18-1545
[84] C. Vania, Y. Kementchedjhieva, A. Søgaard, and A. Lopez. 2020. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *EMNLP-IJCNLP'19*. 1105–1116. DOI: 10.18653/v1/d19-1102
[85] S. Y. Feng, A. W. Li, and J. Hoey. 2020. Keep calm and switch on! Preserving sentiment and fluency in semantic text exchange. In *EMNLP-IJCNLP'19*. 2701–2711. DOI: 10.18653/v1/d19-1272
[86] J. Min, R. T. McCoy, D. Das, E. Pitler, and T. Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. In *ACL'20*. 2339–2352. DOI: 10.18653/v1/2020.acl-main.212
[87] R. Thomas McCoy, E. Pavlick, and T. Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *ACL'19*. 3428–3448. DOI: 10.18653/v1/p19-1334
[88] Y. Goldberg. 2019. Assessing BERT's syntactic abilities. arXiv:1901.05287. Retrieved from http://arxiv.org/abs/1901.05287.
[89] I. Tenney et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *ICLR'19*.
[90] J. F. Steffensen, Interpolation. Courier Corporation, 2006.
[91] H. Shi, K. Livescu, and K. Gimpel. 2021. Substructure Substitution: Structured data augmentation for NLP. arXiv:2101.00411. Retrieved from http://arxiv.org/abs/2101.00411.
[92] H. Kim, D. Woo, S. J. Oh, J.-W. Cha, and Y.-S. Han. 2021. ALP: Data augmentation using lexicalized PCFGs for few-shot text classification. arXiv:2112.11916. Retrieved from http://arxiv.org/abs/2112.11916.
[93] M. Aiken and M. Park. 2010. The efficacy of round-trip translation for MT evaluation. *Transl. J.* 14, 1 (2010).
[94] A. W. Yu et al. 2018. QaNet: Combining local convolution with global self-attention for reading comprehension. In *Proceeding of the 6th International Conference on Learning Representations*.
[95] E. Rabinovich, S. Mirkin, R. N. Patel, L. Specia, and S. Wintner. 2017. Personalized machine translation: Preserving original author traits. In *EACL'17*, 1, 1074–1084. DOI: 10.18653/v1/e17-1101
[96] A. Kruspe, J. Kersten, M. Wiegmann, B. Stein, and F. Klan. 2018. Classification of Incident-related Tweets: Tackling imbalanced training data using hybrid CNNs and translation-based data augmentation. In *TREC'18*.
[97] J. Chen, Z. Yang, and D. Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.* 2147–2157. DOI: 10.18653/v1/2020.acl-main.194
[98] Y. Wu et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv (2016), 1–23. http://arxiv.org/abs/1609.08144.
[99] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie. 2020. CERT: Contrastive self-supervised learning for language understanding. Retrieved from 2022 https://github.com/UCSD-AI4H/.
[100] D. Britz, A. Goldie, M. T. Luong, and Q. V Le. 2017. Massive exploration of neural machine translation architectures. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1442–1451. DOI: 10.18653/v1/d17-1151
[101] D. Shen, M. Zheng, Y. Shen, Y. Qu, and W. Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. http://arxiv.org/abs/2009.13818.
[102] N. Malandrakis, M. Shen, A. Goyal, S. Gao, A. Sethi, and A. Metallinou. 2019. Controlled text generation for data augmentation in intelligent artificial agents. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. 90–98. DOI: 10.18653/v1/d19-5609
[103] K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang. 2018. Generating sentences by editing prototypes. In *ACL'18* 6, 437–450. DOI: 10.1162/tacl_a_00030
[104] G. Raille, S. Djambazovska, and C. Musat. 2020. Fast cross-domain data augmentation through neural sentence editing. http://arxiv.org/abs/2003.10254.
[105] L. Yu, W. Zhang, J. Wang, and Y. Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In *AAAI'17*. 2852–2858.
[106] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria. 2018. A generative model for category text generation. *Inf. Sci. (N. Y.)*. 450, C (2018), 301–315. DOI: 10.1016/j.ins.2018.03.050
[107] C. Wang and D. Lillis. 2020. Classification for crisis-related tweets leveraging word embeddings and data augmentation. In *TREC'19*, 8. https://trec.nist.gov/.
[108] H. Queiroz Abonizio and S. Barbon Junior. 2020. Pre-trained data augmentation for text classification. In *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12319, 551–565. DOI: 10.1007/978-3-030-61377-8_38

[109] V. Claveau, A. Chaffin, and E. Kijak. 2021. La génération de textes artificiels en substitution ou en complément de données d'apprentissage (Generating artificial texts as substitution or complement of training data). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles*. 37–49. Available: https://arxiv.org/abs/2110.13016v1.

[110] R. Liu, G. Xu, C. Jia, W. Ma, L. Wang, and S. Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP'20)*. 9031–9041. DOI: 10.18653/v1/2020.emnlp-main.726

[111] I. S. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, and Dario Amodei. 2020. Language models are unsupervised multitask learners. *OpenAI Blog* 1, (May 2020), 1–7. Available: https://github.com/codelucas/newspaper.

[112] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv:1910.01108. Retrieved from http://arxiv.org/abs/1910.01108.

[113] K. M. Yoo, D. Park, J. Kang, S. W. Lee, and W. Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Proceedings of the Findings of the Association for Computational Linguistics, Findings of ACL (EMNLP'21)*. 2225–2239. DOI: 10.18653/v1/2021.findings-emnlp.192

[114] T. B. Brown et al. 2020. Language models are few-shot learners. arXiv:2005.14165. Retrieved from http://arxiv.org/abs/2005.14165.

[115] K. Lee, K. Guu, L. He, T. Dozat, and H. W. Chung. 2021. Neural data augmentation via example extrapolation. arXiv:2102.01335. Retrieved from http://arxiv.org/abs/2102.01335.

[116] B. Ding et al. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *EMNLP'20*. 6045–6057. DOI: 10.18653/v1/2020.emnlp-main.488

[117] E. Chang, X. Shen, D. Zhu, V. Demberg, and H. Su. 2021. Neural data-to-text generation with LM-based text augmentation. In *EACL'21*. 758–768.

[118] V. Kumar, H. Glaude, C. de Lichy, and W. Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing*. 1–10. DOI: 10.18653/v1/d19-6101

[119] E. Schwartz et al. 2018. Delta-encoder: An effective sample synthesis method for few-shot object recognition. arXiv:1806.04734. Retrieved from http://arxiv.org/abs/1806.04734.

[120] C. Zhu, Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu. 2019. FreeLB: Enhanced adversarial training for natural language understanding. arXiv:1909.11764. Retrieved from http://arxiv.org/abs/1909.11764.

[121] P. L. Combettes and J. C. Pesquet. 2011. Proximal splitting methods in signal processing. In *Springer Optimization and Its Applications*. 49, 185–212. DOI: 10.1007/978-1-4419-9569-8_10

[122] T. Goldstein, C. Studer, and R. Baraniuk. 2014. A Field guide to forward-backward splitting with a FASTA implementation. arXiv:1411.3406. Retrieved from http://arxiv.org/abs/1411.3406.

[123] A. Shafahi et al. 2019. Adversarial training for free!. In *NeurIPS'19*, 32.

[124] D. Zhang, T. Zhang, Y. Lu, Z. Zhu, and B. Dong. 2019. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS'19*, 32.

[125] T. Miyato, S. I. Maeda, M. Koyama, K. Nakae, and S. Ishii. 2016. Distributional smoothing with virtual adversarial training. In *ICLR'16*.

[126] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *ACL'20*. 2177–2190. DOI: 10.18653/v1/2020.acl-main.197

[127] D. Wang, C. Gong, and Q. Liu. 2019. Improving neural language modeling via adversarial training. In *ICML'19*. 11387–11397.

[128] X. Liu et al. 2020. Adversarial training for large neural language models. arXiv:2004.08994. Retrieved from http://arxiv.org/abs/2004.08994.

[129] D. Liu et al. 2020. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space. In *EMNLP'20*. 5798–5810. DOI: 10.18653/v1/2020.emnlp-main.467

[130] Z. Wan, X. Wan, and W. Wang. 2021. Improving grammatical error correction with data augmentation by editing latent representation. In *COLING'20*. 2202–2212. DOI: 10.18653/v1/2020.coling-main.200

[131] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. 2016. Generating sentences from a continuous space. In *CoNLL'16*. 10–21. DOI: 10.18653/v1/k16-1002

[132] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, (2002), 321–357. DOI: 10.1613/jair.953

[133] V. Verma et al. 2019. Manifold mixup: Better representations by interpolating hidden states. In *ICML'19*. 11196–11205.

[134] N. Tishby and N. Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *ITW'15*. DOI: 10.1109/ITW.2015.7133169

[135] R. Shwartz-Ziv and N. Tishby. 2017. Opening the black box of deep neural networks via information. arXiv:1703.00810. Retrieved from http://arxiv.org/abs/1703.00810.
[136] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. 2018. MixUp: Beyond empirical risk minimization. In *ICLR'18*.
[137] L. Sun, C. Xia, W. Yin, T. Liang, P. Yu, and L. He. 2021. Mixup-Transformer: Dynamic data augmentation for NLP tasks. In *COLING*. 3436–3440. DOI: 10.18653/v1/2020.coling-main.305
[138] H. Guo, Y. Mao, and R. Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. http://arxiv.org/abs/1905.08941.
[139] H. Guo. 2020. Nonlinear mixup: Out-of-manifold data augmentation for text classification. *AAAI Conf. Artif. Intell.* 34, 04 (2020), 4044–4051. DOI: 10.1609/aaai.v34i04.5822
[140] G. Jawahar, B. Sagot, and D. Seddah. 2020. What does BERT learn about the structure of language? In *ACL'19*. 3651–3657. DOI: 10.18653/v1/p19-1356
[141] J. Chen, Z. Wang, R. Tian, Z. Yang, and D. Yang. 2020. Local additivity based data augmentation for semi-supervised NER. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1241–1251. DOI: 10.18653/v1/2020.emnlp-main.95
[142] N. Reimers and I. Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP'19*. 3982–3992. DOI: 10.18653/v1/d19-1410
[143] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan. 2019. AugMix: A simple data processing method to improve robustness and uncertainty. arXiv:1912.02781. Retrieved from http://arxiv.org/abs/1912.02781.
[144] S. Bonthu, A. Dayal, M. S. Lakshmi, and S. Rama Sree. 2022. *Effective Text Augmentation Strategy for NLP Models*. Springer, Singapore, 521–531.
[145] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 5065–5075. DOI: 10.18653/v1/2021.acl-long.393
[146] Y. Yang et al. 2020. Generative data augmentation for commonsense reasoning. In *EMNLP'20*. 1008–1025. DOI: 10.18653/v1/2020.findings-emnlp.90
[147] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit. 2016. A decomposable attention model for natural language inference. In *EMNLP'16*. 2249–2255. DOI: 10.18653/v1/d16-1244
[148] A. Robinson, R. D. Cook, and S. Weisberg. 1984. Residuals and influence in regression. *J. R. Stat. Soc. Ser. A* 147, 1 (1984), 108. DOI: 10.2307/2981746
[149] P. W. Koh and P. Liang. 2017. Understanding black-box predictions via influence functions. In *ICML'17*, 4, 2976–2987.
[150] I. Chaturvedi, Y.-S. Ong, I. W. Tsang, R. E. Welsch, and E. Cambria. 2016. Learning word dependencies in text by means of a deep recurrent belief network. *Knowl.-Based Syst.* 108 (2016), 144–154. DOI: 10.1016/j.knosys.2016.07.019
[151] W. Zhao, H. Peng, S. Eger, E. Cambria, and M. Yang. 2020. Towards scalable and reliable capsule networks for challenging NLP applications. In *ACL'19*. 1549–1559. DOI: 10.18653/v1/p19-1150
[152] Y. Ma, H. Peng, T. Khan, E. Cambria, and A. Hussain. 2018. Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis. *Cognit. Comput.* 10, 4 (2018). DOI: 10.1007/s12559-018-9549-x
[153] P. K. Jain, W. Quamer, V. Saravanan, and R. Pamula. 2022. Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis. *J. Amb. Intell. Humaniz. Comput.* 1 (2022), 3. DOI: 10.1007/s12652-022-03698-z
[154] M.-A. Kaufhold. 2021. *Information Refinement Technologies for Crisis Informatics: User Expectations and Design Principles for Social Media and Mobile Apps*. Springer Fachmedien Wiesbaden.
[155] J. Bragg, A. Cohan, K. Lo, and I. Beltagy. 2021. FLEX: Unifying evaluation for few-shot NLP. arXiv:2107.07170. Retrieved from http://arxiv.org/abs/2107.07170.
[156] S. Gehrmann et al. 2021. The GEM Benchmark: Natural language generation, its evaluation and metrics. In *GEM'21*. 96–120. DOI: 10.18653/v1/2021.gem-1.10
[157] K. D. Dhole et al. 2021. NL-Augmenter: A framework for task-sensitive natural language augmentation. arXiv:2112.0272v1. Retrieved from https://arxiv.org/abs/2112.02721v1.
[158] Z. Papakipos and J. Bitton. 2022. AugLy: Data augmentations for robustness. arXiv:2201.06494. Retrieved from http://arxiv.org/abs/2201.06494.
[159] I. Solaiman et al. 2019. Release strategies and the social impacts of language models. arXiv:1908.09203. Retrieved from http://arxiv.org/abs/1908.09203.
[160] A. Vaswani et al. 2017. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017-Decem (2017), 5999–6009.