***DAT* #**	***DAT* name**	**Action**	**Data Source Details**	**Data Source Availability Link**	**Augmentation Approach (*Aug*)**
1	Swap Word	swap	-	-	Aug_1. RandomWordAug
2	Delete Word	delete	-	-	
3	Spelling Augmenter	substitute	Pre-defined spelling mistake dictionary	https://github.com/makcedward/nlpaug/blob/master/nlpaug/res/word/spelling/spelling_en.txt	Aug_2. SpellingAug
4	Split Words	split	-	-	Aug_3. SplitAug
5	Synonym Wordnet Subs	substitute	WordNet: Large English lexical database where nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets).	https://wordnet.princeton.edu/	Aug_4. SynonymAug
6	Synonym PPDB Subs		PPDB: Database containing millions paraphrases to improve language processing by making systems more robust to language variability and unseen words. Six sizes, from S up to XXXL are available, we used XXXL, which is the recommended one.	http://paraphrase.org/#/download	
7	Tf_Idf Ins	insert	Trained TF-IDF model on security tool API corpus to augment data using TF-IDF statistics.	https://github.com/makcedward/nlpaug/blob/master/example/tfidf-train_model.ipynb	Aug_5. TfIdfAug
8	Tf_Idf Subs	substitute			
9	Word2vec Googlenews Ins	insert	Google News: Pre-trained vectors trained on part of Google News dataset (about 100 billion words). The model contains 300-dimensional vectors for 3 million words and phrases.	https://code.google.com/archive/p/word2vec/	Aug_6. WordEmbsAug
10	Word2vec Googlenews Subs	substitute			
11	Fasttext Wikinews Ins	insert	Wikinews: 1 million word vectors trained on Wikipedia 2017, UMBC webbase corpus and statmt.orgnews dataset (16B tokens).	https://fasttext.cc/docs/en/english-vectors.html	
12	Fasttext Wikinews Subs	substitute			
13	Fasttext Wiki-News SWord Ins	insert	Wikinews subword: 1 million word vectors trained with subword infomation on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset (16B tokens).	https://dl.fbaipublicfiles.com/fasttext/vectors-english/wiki-news-300d-1M-subword.vec.zip	
14	Fasttext Wiki-News SWord Subs	substitute			

DAT #	*DAT* name	Action	Data Source Details	Data Source Availability Link	Augmentation Approach (*Aug*)
15	Fasttext C. Crawl Ins	insert	Common Crawl: 2 million word vectors trained on Common Crawl (600B tokens).	https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M.vec.zip	Aug_6. WordEmbs Aug
16	Fasttext C. Crawl Subs	substitute			
17	Fasttext C. Crawl SWord Ins	insert	Common Crawl subword: 2 million word vectors trained with subword information on Common Crawl (600B tokens).	https://dl.fbaipublicfiles.com/fasttext/vectors-english/crawl-300d-2M-subword.zip	
18	Fasttext C. Crawl SWord Subs	substitute			
19	Fasttext C. Crawl & Wiki Ins	insert	Common Crawl and Wikipedia: Trained on Common Crawl and Wikipedia using fastText. These models were trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives.	https://fasttext.cc/docs/en/crawl-vectors.html	
20	Fasttext C. Crawl & Wiki Subs	substitute			
21	Glove Wiki+Gword 50d Ins	insert	Wikipedia2014+Gigaword: Trained on Wikipedia 2014 (http://dumps.wikimedia.org/enwiki/20140102/) + Gigaword 5 (https://catalog.ldc.upenn.edu/LDC2011T07) 6B tokens, 400K vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB. This paper considered 50d and 300d, the minimum and maximum dimension size to check the variation.	https://nlp.stanford.edu/projects/glove/ http://nlp.stanford.edu/data/glove.6B.zip	
22	Glove Wiki+Gword 50d Subs	substitute			
23	Glove Wiki+Gword 300d Ins	insert			
24	Glove Wiki+Gword 300d Subs	substitute			
25	Glove C. Crawl 300d Ins	insert	Common Crawl: (42B tokens, 1.9 M vocab, uncased, 300d vectors, 1.75 GB)	http://nlp.stanford.edu/data/glove.42B.300d.zip	
26	Glove C. Crawl 300d Subs	substitute			
27	Glove Twitter 200d Ins	insert	Twitter (2B tweets, 27B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB) This paper considered 200d as it contains maximum dataset size. Downloaded data 2.06 GB	http://nlp.stanford.edu/data/glove.twitter.27B.zip	
28	Glove Twitter 200d Subs	substitute			
29	Bert Base Ins	insert	Bert: 12-layer, 768-hidden, 12-heads, 110M parameters. Trained on lower-cased English text.	https://huggingface.co/transformers/pretrained_models.	Aug_7. Contextual WordEmbs Aug
30	Bert Base Subs	substitute			
31	Distilbert Ins	insert	Distilbert: 6-layer, 768-hidden, 12-heads, 66M parameters. The DistilBERT model distilled from the BERT model bert-base-uncased	https://github.com/huggingface/transformers/tree/master/examples/distillation	
32	Distilbert Subs	substitute			

DAT #	*DAT* name	Action	Data Source Details	Data Source Availability Link	Augmentation Approach (*Aug*)
33	Roberta Base Ins	insert	RoBERTa: 12-layer, 768-hidden, 12-heads, 125M parameters. RoBERTa using the BERT-base architecture.	https://github.com/pytorch/fairseq/tree/master/examples/roberta	Aug$_{7}$. Contextual WordEmbs Aug
34	Roberta Base Subs	substitute			
35	Distilroberta-Base Ins	insert	DistilRoBERTa: 6-layer, 768-hidden, 12-heads, 82M parameters. The DistilRoBERTa model distilled from the RoBERTa model roberta-basecheckpoint.	https://github.com/huggingface/transformers/tree/master/examples/distillation	
36	Distilroberta-Base Subs	substitute			