

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

**Jason Wei, Kai Zou.
EMNLP 2019**

Reviewed by Susang Kim

Contents

1.Introduction

2.Motivation

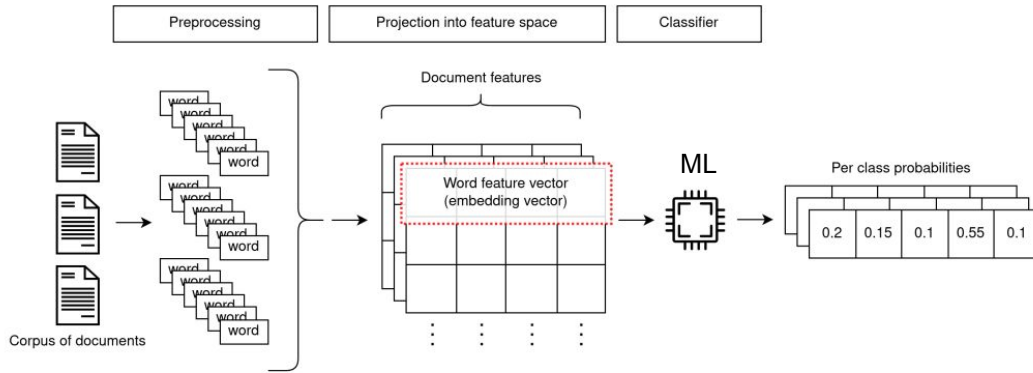
3.Methods

4.Experiments

5.Conclusion & Discussion

✘ AEDA (EMNLP 2021)

1.Introduction (Text Classification)



Text classification is a ML that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize text.

Sentiment analysis (SA) : categorised emotions

Topic labelling (TL) : the task of recognising one or more themes for a piece of text

News classification (NC) : the task of assigning categories to news pieces

Question answering (QA) : the task of selecting an answer to a question.

Named entity recognition (NER) : the task of locating named entities within unstructured text.

(PoS) tagging, speech dependencies and semantic role labelling.

Figure 1. An overview of the two-step procedure adopted by shallow learning methods.

tokenization, cleaning, normalization, stemming, One-Hot Encoding

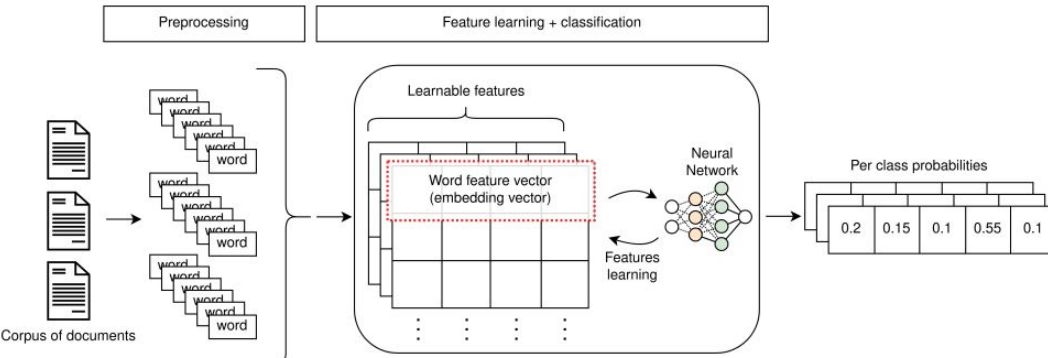


Figure 2. Overview of the training procedure used with deep learning methods.

1.Introduction (Data Augmentation)

Data Augmentation is a techniques that **generate new data points from the data that already exists**. This practice includes making small changes to the data, generating diverse instances, to have improved the performance and outcome of model.

1) Image



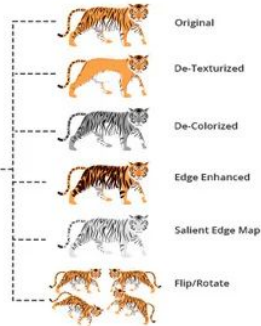
Geometric

Kernel filters

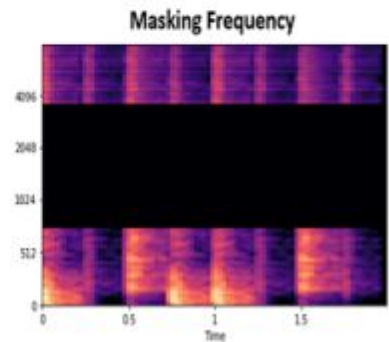
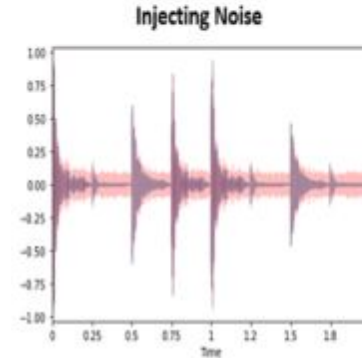
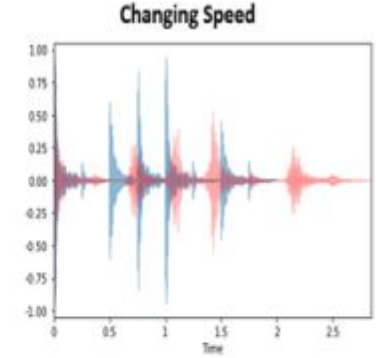
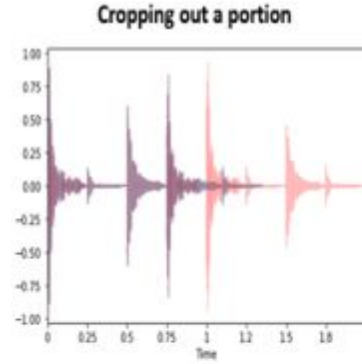


Data Augmentation

Color space



2) Audio



1.Introduction (Data Augmentation - NLP)

1) Easy Data Augmentation (EDA), some easy text transformations are applied. For example, a word is randomly replaced with a synonym. Two or more words are swapped in the sentence.

1. Replacing words with synonyms
2. Words or sentence shuffling
3. Text substitution
4. Random insertion, deletion, or swapping of words

3) Generative Adversarial Networks (GAN)

GAN is an unsupervised learning network that involves automatically discovering and learning the regularities or patterns in input data. The model thus generates new examples that could have been apparently drawn from the original data set.

| | |
|--------------------|--|
| | The pandas have different eyes than bears |
| | The pandas can swim and climb |
| The pandas have... | |
| | The pandas spend a lot of their day eating |

4) Foundation Model(LLM)

2) Back Translation

Back translation, also known as reverse translation is the process of re-translating content from a target language back to its source language. This leads to variants of a sentence that help the model to learn better.

| | | |
|----------------------------------|---------------------------------------|-------------------------------------|
| English: How are you? | Arabic: kayf halukum | English: How are you all |
| English: This is awesome! | Italian: Questo e spettacolare | English: This is spectacular |

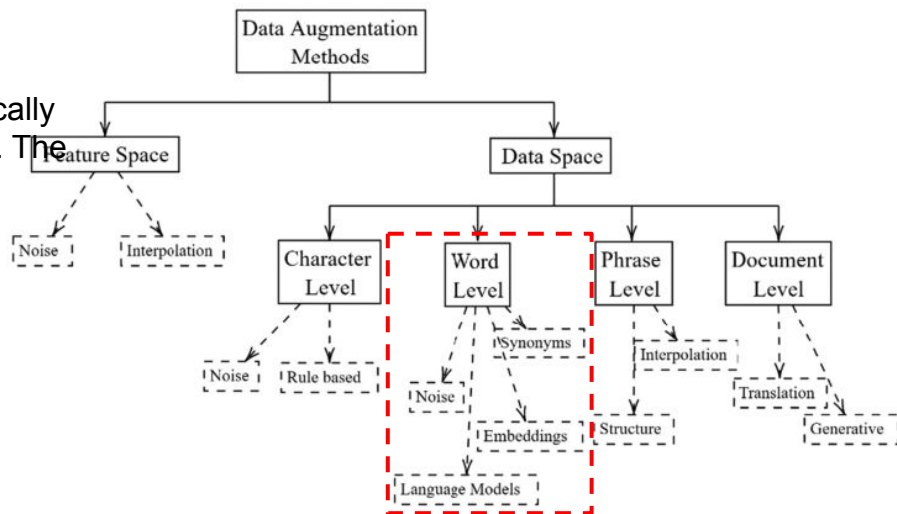
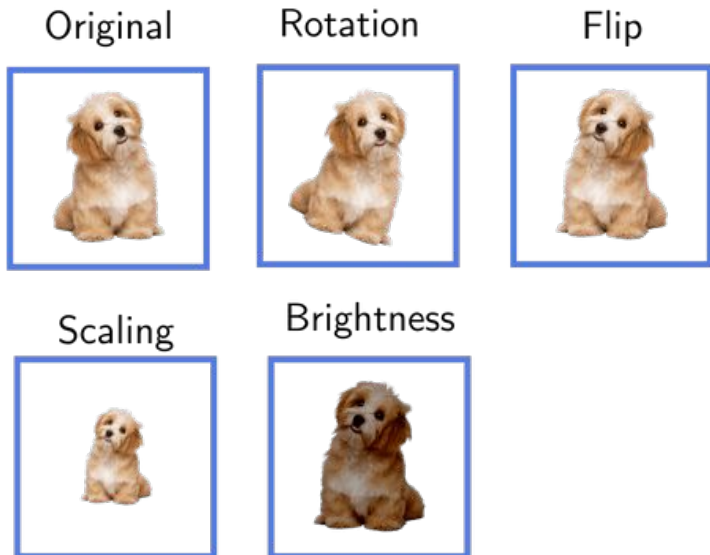


Fig. 1. Taxonomy and grouping for different data augmentation methods.

2.Motivation - Data Augmentation (CV vs NLP)



For vision task, Image Augmentation (Flip, Scaling, Crop, Translation, Rotation, Gaussian Noise) can improve performance.

“눈을 보다” => Look at the eyes, Look at the snow.
 “카메라를 찍다” => Take a picture, Hit the camera.

Are You Happy <-> You Are Happy
 아버지가 방에 들어가신다 <-> 아버지 가방에 들어가신다
 나는 너를 좋아해 <-> 너는 나를 좋아해

For NLP task, different languages have different characteristics, just change the position of a word can completely change its meaning. Therefore, augmentations are quite difficult to apply in NLP.

We present EDA: easy data augmentation techniques for boosting performance on text classification tasks.
 (Not Exploratory Data Analysis)

3.Methods - EDA(Easy Data Augmentation)

| Operation | Sentence |
|-----------|---|
| None | A sad, superior human comedy played out on the back roads of life. |
| SR | A <i>lamentable</i> , superior human comedy played out on the <i>backward</i> road of life. |
| RI | A sad, superior human comedy played out on <i>funniness</i> the back roads of life. |
| RS | A sad, superior human comedy played out on <i>roads</i> back <i>the</i> of life. |
| RD | A sad, superior human out on the roads of life. |

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

In text classification task, randomly select one of the four and apply it.

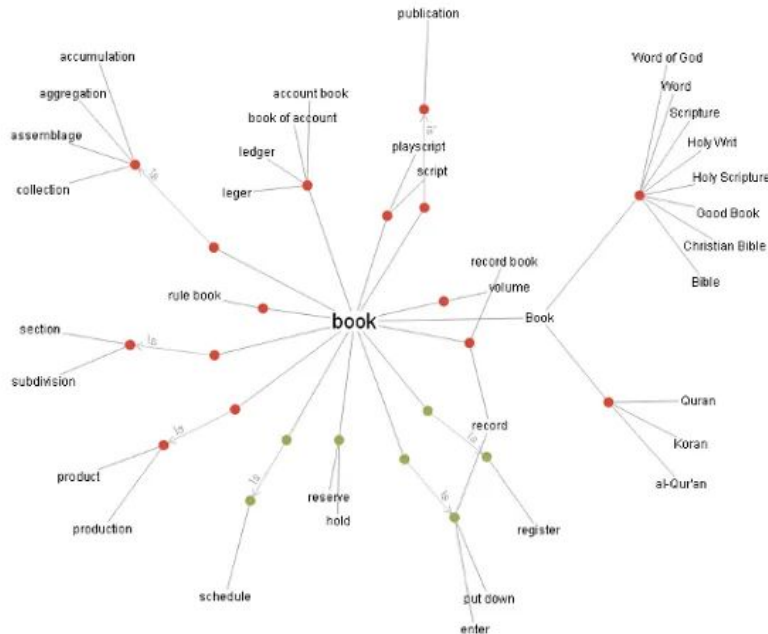
To compensate, we vary the number of words changed, n , for SR, RI, and RS based on the sentence length l with the formula = αl , where α is a parameter that indicates the percent of the words in a sentence are changed (we use $p=\alpha$ for RD).

1. **Synonym Replacement (SR):** Randomly choose n words from the sentence that are not stop words. Replace each of these words with one of its synonyms chosen at random. [WordNet](#)
2. **Random Insertion (RI):** Find a random synonym of a random word in the sentence that is not a stop word. Insert that synonym into a random position in the sentence. Do this n times.
3. **Random Swap (RS):** Randomly choose two words in the sentence and swap their positions. Do this n times.
4. **Random Deletion (RD):** Randomly remove each word in the sentence with probability p .

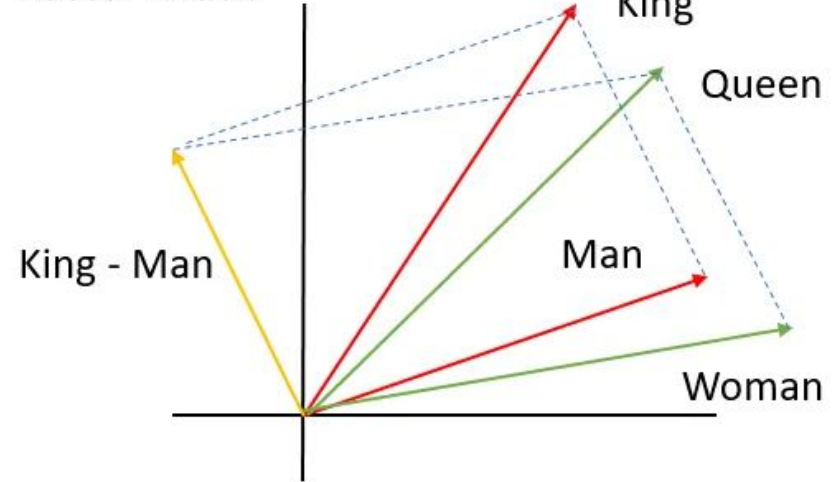
4.Experiments - Word Embedding

Synonym thesaurus: synonym replacements and random insertions were generated using **WordNet**.

Word embeddings: We use 300 dimensional word embeddings trained using **GloVe**



Vector Math



<https://towardsdatascience.com/%EF%B8%8Fwordnet-a-lexical-taxonomy-of-english-words-4373b541cfff>

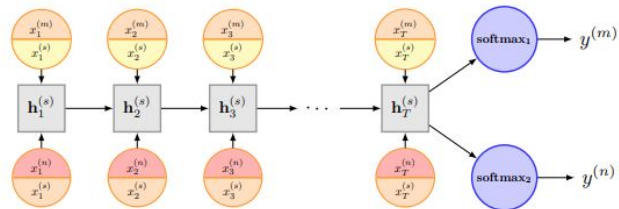
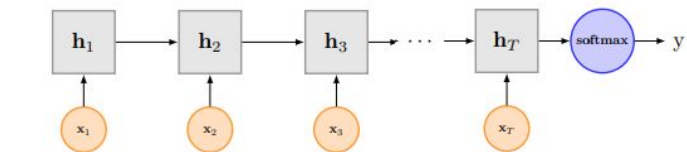
<https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73>

4.Experiments - Text Classification Models(RNN)

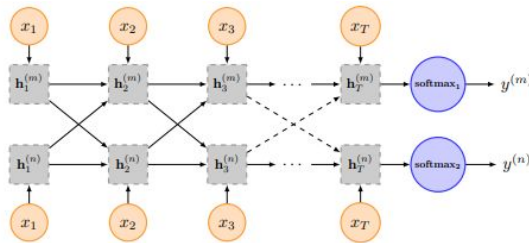
1) LSTM-RNN

$$\mathbf{h}_t = \begin{cases} 0 & t = 0 \\ f(\mathbf{h}_{t-1}, \mathbf{x}_t) & \text{otherwise} \end{cases}$$

$$L(\hat{\mathbf{y}}, \mathbf{y}) = - \sum_{i=1}^N \sum_{j=1}^C \mathbf{y}_i^j \log(\hat{\mathbf{y}}_i^j),$$

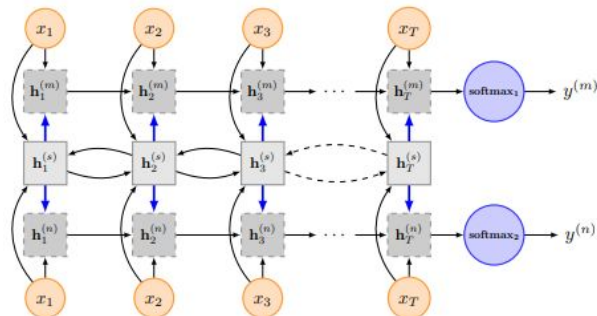


(a) Model-I: Uniform-Layer Architecture



(b) Model-II: Coupled-Layer Architecture

We use the multitask learning framework to jointly learn across multiple related tasks.



(c) Model-III: Shared-Layer Architecture

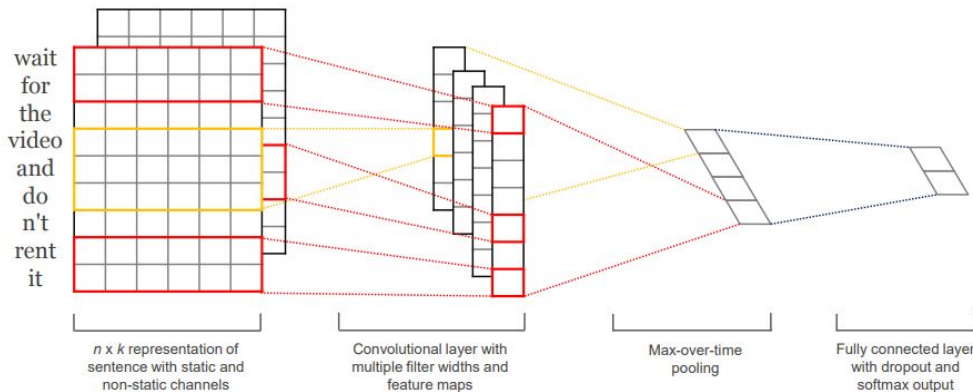
$$\hat{\mathbf{x}}_t^{(m)} = \mathbf{x}_t^{(m)} \oplus \mathbf{x}_t^{(s)}, \quad (9)$$

where $\mathbf{x}_t^{(m)}$, $\mathbf{x}_t^{(s)}$ denote the task-specific and shared word embeddings respectively, \oplus denotes the concatenation operation.

| Model | SST-1 | SST-2 | SUBJ | IMDB | Avg Δ |
|-------------|-------------|-------------|-------------|-------------|--------------|
| Single Task | 45.9 | 85.8 | 91.6 | 88.5 | - |
| SST1-SST2 | 48.9 | 87.4 | - | - | +2.3 |
| SST1-SUBJ | 46.3 | - | 92.2 | - | +0.5 |
| SST1-IMDB | 46.9 | - | - | 89.5 | +1.0 |
| SST2-SUBJ | - | 86.5 | 92.5 | - | +0.8 |
| SST2-IMDB | - | 86.8 | - | 89.8 | +1.2 |
| SUBJ-IMDB | - | - | 92.7 | 89.3 | +0.9 |

4.Experiments - Text Classification Models(CNN)

2) CNNs : We use the publicly available word2vec vectors that were trained on 100 billion words from Google News.



1D convolutional layer of 128 filters of size 5, global 1D max pool layer, dense layer of 20 hidden units with ReLU activation function, softmax output layer.

| Model | MR | SST-1 | SST-2 | Subj | TREC | CR | MPQA |
|------------------|-------------|-------|-------------|------|------|-------------|-------------|
| CNN-rand | 76.1 | 45.0 | 82.7 | 89.6 | 91.2 | 79.8 | 83.4 |
| CNN-static | 81.0 | 45.5 | 86.8 | 93.0 | 92.8 | 84.7 | 89.6 |
| CNN-non-static | 81.5 | 48.0 | 87.2 | 93.4 | 93.6 | 84.3 | 89.5 |
| CNN-multichannel | 81.1 | 47.4 | 88.1 | 93.2 | 92.2 | 85.0 | 89.4 |

CNN-rand: all words are randomly initialized and modified during training.

CNN-static: A model with pre-trained vectors from word2vec. All words including the unknown ones that are randomly initialized are kept static and only the other parameters of the model are learned.

CNN-non-static: Same as above but the pretrained vectors are fine-tuned for each task.

CNN-multichannel: A model with two sets of word vectors. Each set of vectors is treated as a 'channel' and each filter is applied to both channels, but gradients are backpropagated only through one of the channels. Hence the model is able to fine-tune one set of vectors while keeping the other static. Both channels are initialized with word2vec.

4.Experiments - Benchmark Datasets

9.2 Benchmark Datasets

Summary statistics for the five datasets used are shown in Table 5.

| Dataset | c | l | N_{train} | N_{test} | $ V $ |
|---------|-----|-----|-------------|------------|--------|
| SST-2 | 2 | 17 | 7,447 | 1,752 | 15,708 |
| CR | 2 | 18 | 4,082 | 452 | 6,386 |
| SUBJ | 2 | 21 | 9,000 | 1,000 | 22,329 |
| TREC | 6 | 9 | 5,452 | 500 | 8,263 |
| PC | 2 | 7 | 39,418 | 4,508 | 11,518 |

Table 5: Summary statistics for five text classification datasets. c : number of classes. l : average sentence length (number of words). N_{train} : number of training samples. N_{test} : number of testing samples. $|V|$: size of vocabulary.

Random subset of the full training set with
 $N_{train} = \{500, 2,000, 5,000, \text{all available data}\}$.

Five benchmark classification task

(1) **SST-2**: Stanford Sentiment Treebank—an extension of MR but with train/dev/test splits provided and fine-grained labels (very positive, positive, neutral, negative, very negative), re-labeled by Socher et al. (2013).

(2) **CR**: Customer reviews of various products (cameras, MP3s etc.). Task is to predict positive/negative reviews (Hu and Liu, 2004).

(3) **SUBJ**: Subjectivity dataset where the task is to classify a sentence as being subjective or objective (Pang and Lee, 2004)

(4) **TREC**: question dataset—task involves classifying a question into 6 question types (whether the question is about person, location, numeric information, etc.)
question type dataset (Li and Roth, 2002)

(5) **PC**: Pro-Con dataset (Ganapathibhotla and Liu, 2008)

My SLR is on the shelf

by [shortstop24](#), Aug 09 '03

Pros: Great photos, easy to use, good manual, many options, takes videos

Cons: Battery usage; included software could be improved; included 16MB is stingy.

I had never used a digital camera prior to purchasing the Canon A70. I have always used a SLR (Minol ...

[Read the full review](#)

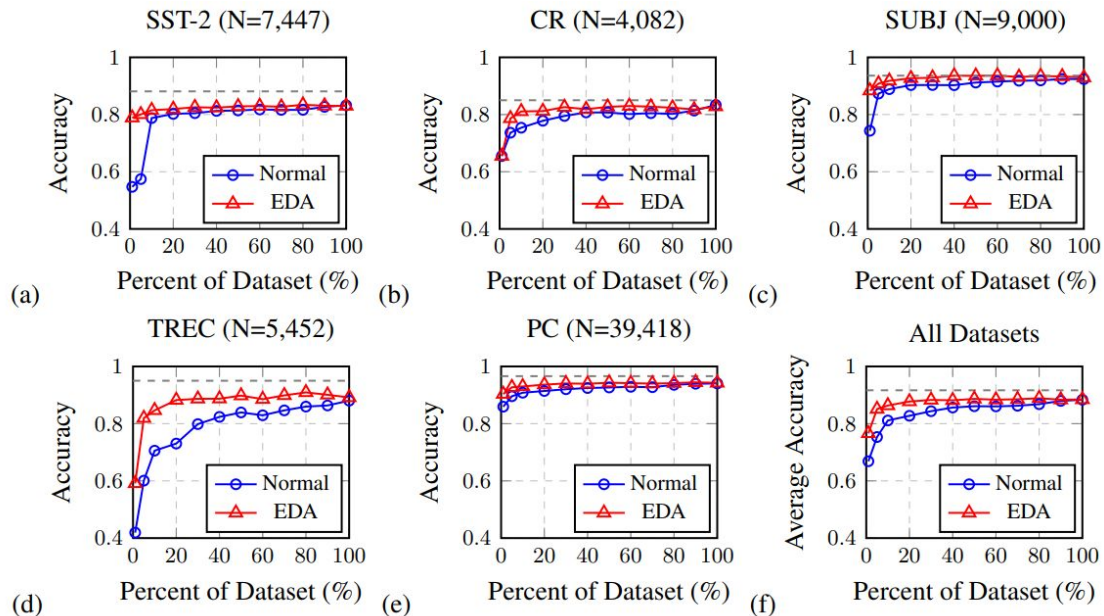
Figure 1: An example review

4.Experiments - EDA Make Gains

| Model | Training Set Size | | | |
|---------|-------------------|-------|-------|-------------|
| | 500 | 2,000 | 5,000 | full set |
| RNN | 75.3 | 83.7 | 86.1 | 87.4 |
| +EDA | 79.1 | 84.4 | 87.3 | 88.3 |
| CNN | 78.6 | 85.6 | 87.7 | 88.3 |
| +EDA | 80.7 | 86.4 | 88.3 | 88.8 |
| Average | 76.9 | 84.6 | 86.9 | 87.8 |
| +EDA | 79.9 | 85.4 | 87.8 | 88.6 |

Table 2: Average performances (%) across five text classification tasks for models with and without EDA on different training set sizes.

Average improvement was **0.8% for full datasets** and **3.0% for Train Set Size=500**.



Both **normal training and EDA training** for the following training set fractions (%): {1, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}.

The best average accuracy without augmentation, **88.3%(100% using training data)**

Models trained using EDA surpassed this number by achieving an average accuracy of 88.6% while only using 50% of the available training data.

4.Experiments - Does EDA conserve true labels?

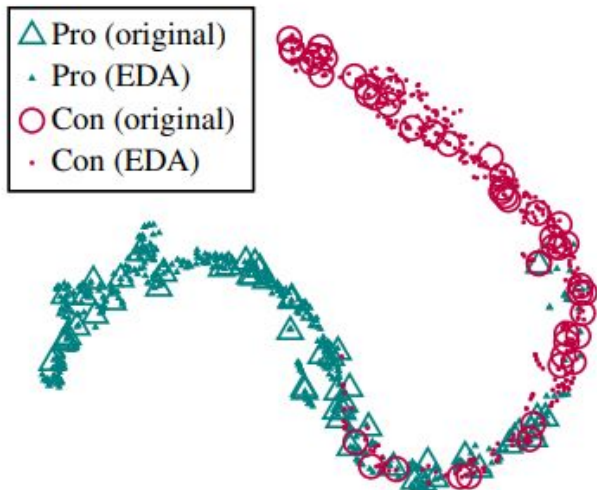


Figure 2: Latent space visualization of original and augmented sentences in the Pro-Con dataset. Augmented sentences (small triangles and circles) closely surround original sentences (big triangles and circles) of the same color, suggesting that augmented sentences maintained their true class labels.

First, we train an RNN on the pro-con classification task (PC) without augmentation. Then, **we apply EDA to the test set** by generating nine augmented sentences per original sentence.

These are fed into the RNN along with the original sentences, and we extract the outputs from the last dense layer. We apply t-SNE (Van Der Maaten, 2014) to these vectors and plot their 2-D representations (Figure 2).

We found that the resulting latent space representations **for augmented sentences closely surrounded those of the original sentences**, which suggests that for the most part, sentences augmented with EDA conserved the labels of their original sentences

4. Experiments - EDA Decomposed

Explore the effects of each operation in EDA.

α is a parameter that indicates the percent of the words in a sentence are changed (varying the parameter $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$)

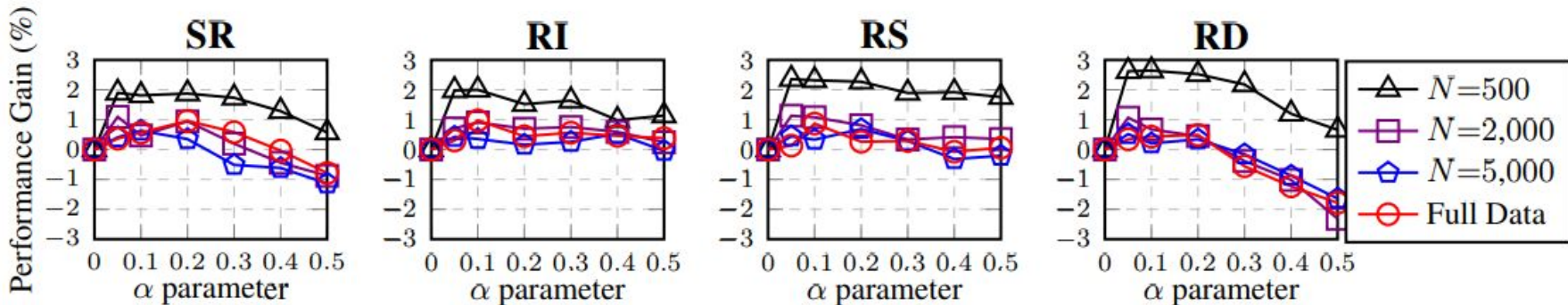


Figure 3: Average performance gain of EDA operations over five text classification tasks for different training set sizes. The α parameter roughly means “percent of words in sentence changed by each augmentation.” SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

For SR, improvement was good for small α , but high α hurt performance, likely because replacing too many words in a sentence changed the identity of the sentence.

4.Experiments - How much augmentation?

How the number of generated augmented sentences per original sentence, n_{aug} , affects performance. we show average performances over all datasets for $n_{aug}=\{1, 2, 4, 8, 16, 32\}$.

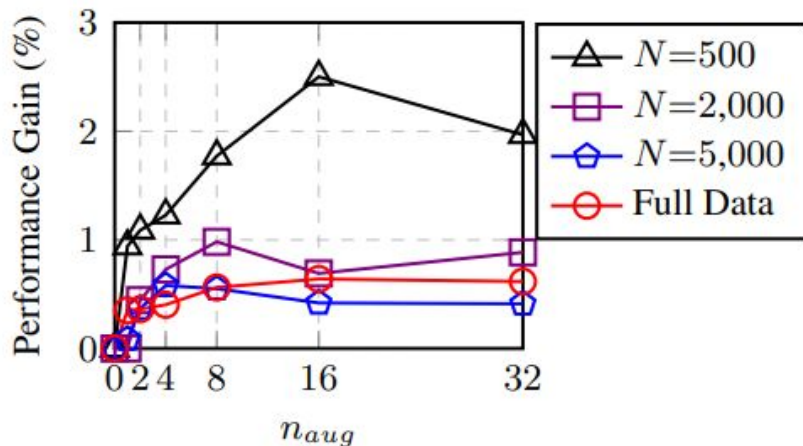


Figure 4: Average performance gain of EDA across five text classification tasks for various training set sizes. n_{aug} is the number of generated augmented sentences per original sentence.

| N_{train} | α | n_{aug} |
|-------------|----------|-----------|
| 500 | 0.05 | 16 |
| 2,000 | 0.05 | 8 |
| 5,000 | 0.1 | 4 |
| More | 0.1 | 4 |

Table 3: Recommended usage parameters.

4.Experiments - Comparison with Related Work

EDA does use a synonym dictionary, WordNet, but the cost of downloading it is far less than training a model on an external dataset, so we don't count it as an "external dataset."

| Technique (#datasets) | LM | Ex Dat |
|-----------------------------------|-----------|-----------|
| Trans. data aug. ¹ (1) | yes | yes |
| Back-translation ² (1) | yes | yes |
| VAE + discrim. ³ (2) | yes | yes |
| Noising ⁴ (1) | yes | no |
| Back-translation ⁵ (2) | yes | no |
| LM + SR ⁶ (2) | yes | no |
| Contextual aug. ⁷ (5) | yes | no |
| SR - kNN ⁸ (1) | no | no |
| EDA (5) | no | no |

Table 4: Related work in data augmentation. #datasets: number of datasets used for evaluation. Gain: reported performance gain on all evaluation datasets. LM: requires training a language model or deep learning. Ex Dat: requires an external dataset.⁹

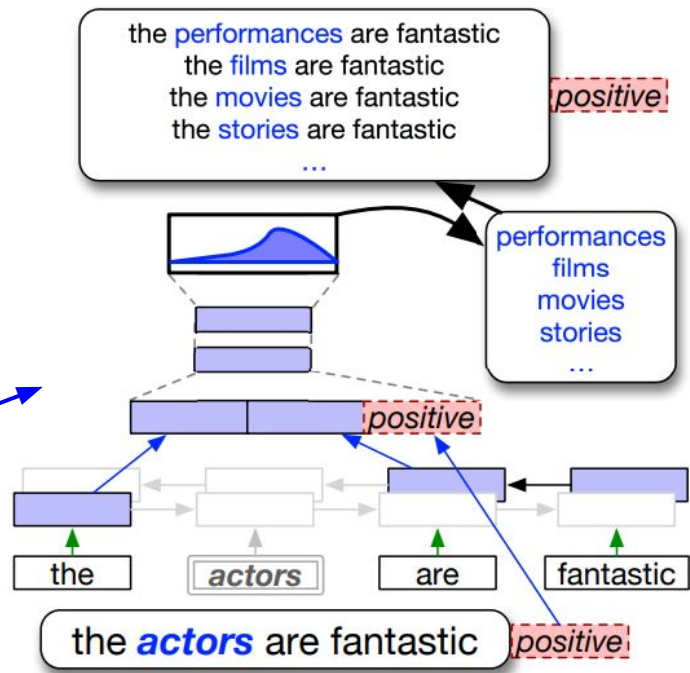


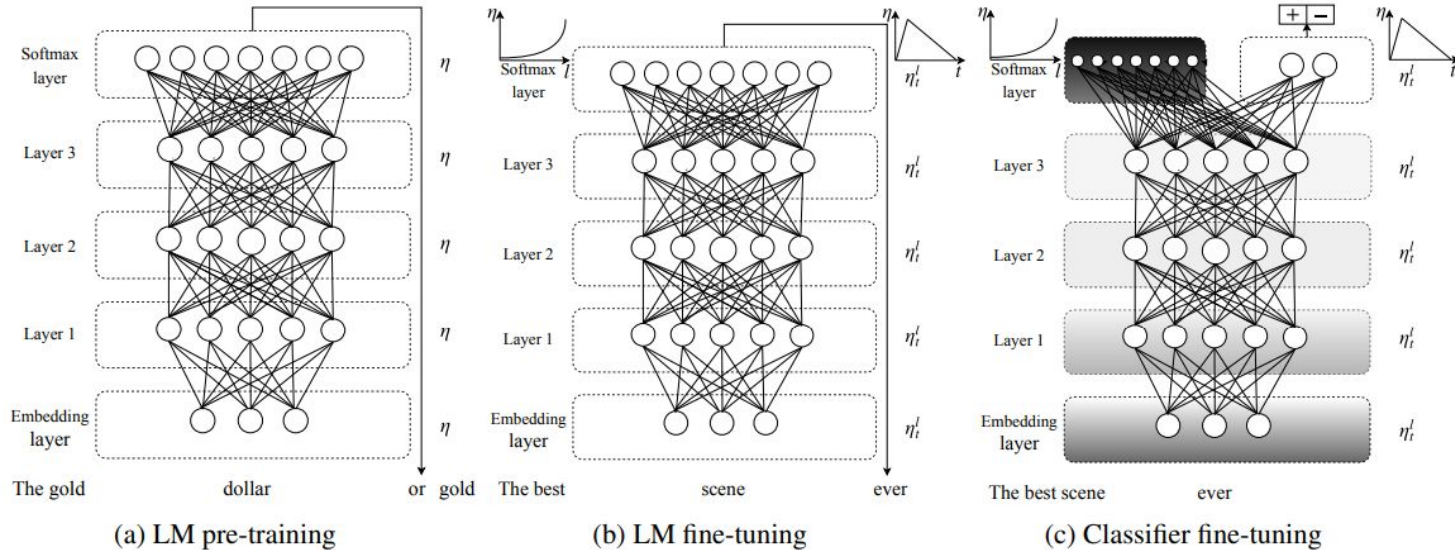
Figure 1: Contextual augmentation with a bi-directional RNN language model, when a sentence "the actors are fantastic" is augmented by replacing only *actors* with words predicted based on the context.

4. Experiments - Discussion and Limitations

EDA's limitations) **Performance gain can be marginal when data is sufficient** the average performance gain for was less than 1% when training with full datasets. And while performance **gains seem clear for small datasets**, EDA might not yield substantial improvements when using pre-trained models.

EDA's **improvement was negligible** when using **ULMFit (Shleifer, 2019)**, and we expect similar results for **ELMo (Peters et al., 2018)** and **BERT (Devlin et al., 2018)**.

Finally, although we evaluate on five benchmark datasets, **other studies on data augmentation in NLP use different models and datasets, and so fair comparison with related work is highly non-trivial.**



5. Conclusions & Discussion

[Conclusions]

Simple data augmentation operations can boost performance on text classification tasks.

EDA substantially boosts performance and reduces overfitting when training on smaller datasets.

[Discussion]

Chinese or other languages? Not yet, but the implementation is **simple use your own**.

Should I use EDA if I'm using a pre-trained model such as **BERT or ELMo**? Models that have been pre-trained on massive datasets probably **don't need EDA**.

Why should I use EDA instead of other techniques such as **contextual augmentation, noising, GAN, or back-translation**? because these require the use of a deep learning model, there is often a **high cost of implementing these techniques** relative to the expected performance gain. we aim to provide a set of simple techniques.

For random insertions, why do you only insert words that are **synonyms, as opposed to inserting any random words**? Data augmentation operations should not change the true label of a sentence, as that would introduce unnecessary noise into the data. Inserting a synonym of a word in a sentence, opposed to a random word, is more likely to be **relevant to the context and retain the original label of the sentence**

AEDA: An Easier Data Augmentation Technique for Text Classification (EMNLP 2021)

AEDA includes only random **insertion of punctuation** marks into the original text.

| | |
|-----------------|--|
| Original | a sad , superior human comedy played out on the back roads of life . |
| Aug 1 | a sad , superior human comedy played out on the back roads ; of life ; . |
| Aug 2 | a , sad . , superior human ; comedy . played . out on the back roads of life . |
| Aug 3 | : a sad ; , superior ! human : comedy , played out ? on the back roads of life . |

Examples of augmented data using AEDA technique. a punctuation mark is picked randomly {".", ":", "?", ":", "!", " ", " "}

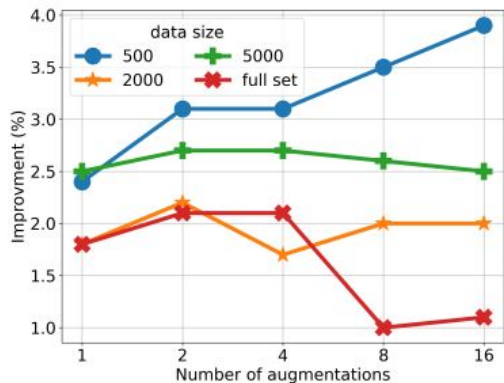
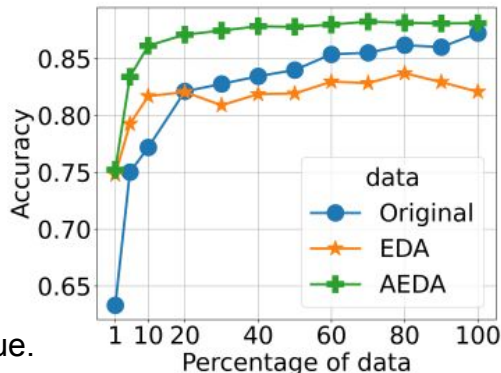


Figure 3: Impact of number of augmentations on the performance of the RNN model trained on various training sizes. Scores are the average of 5 runs over the five datasets. The y axis shows the percentage of improvement.




| Model | Training set size | | | |
|---------|-------------------|-------------|-------------|-------------|
| | 500 | 2,000 | 5,000 | full set |
| RNN | 73.5 | 82.6 | 85.9 | 87.9 |
| +EDA | 76.1 | 81.3 | 85.2 | 86.5 |
| +AEDA | 77.8 | 83.9 | 87.2 | 88.6 |
| CNN | 76.5 | 83.8 | 87.0 | 87.9 |
| +EDA | 77.5 | 82.2 | 84.5 | 86.1 |
| +AEDA | 78.5 | 84.4 | 86.5 | 88.1 |
| Average | 75.0 | 83.2 | 86.5 | 87.9 |
| +EDA | 76.8 | 81.8 | 84.9 | 86.3 |
| +AEDA | 78.2 | 84.2 | 86.9 | 88.4 |

Table 1: Comparing average performance of EDA and AEDA across all datasets on different training set sizes. For each training sample, 16 augmented sentences were added. Scores are the average of 5 runs.

| Model | SST2 | TREC |
|-------|--------------|--------------|
| BERT | 91.10 | 97.00 |
| +EDA | 90.99 | 96.00 |
| +AEDA | 91.76 | 97.20 |

Table 2: Comparing the impact of EDA and AEDA on the BERT model. The model was trained on the combination of the original data and one augmentation for each training sample.

Application (2017 Chatbot)



ChatbotTest

- Intents
- Entities
- Training ^[beta]
- Integrations
- Analytics ^[new]
- Fulfillment
- Prebuilt Agents
- Small Talk
- Docs
- Forum
- Support
- Account
- Logout

피자주문

SAVE

Events

Action

주문

| REQUIRED | PARAMETER NAME | ENTITY | VALUE | IS LIST | PROMPTS |
|-------------------------------------|----------------|------------------|-------------|--------------------------|--------------------|
| <input checked="" type="checkbox"/> | menu | @sys.email | \$cheese | <input type="checkbox"/> | Define prompt S... |
| <input checked="" type="checkbox"/> | size | @sys.date-period | big | <input type="checkbox"/> | Define prompt S... |
| <input type="checkbox"/> | Enter name | Enter entity | Enter value | <input type="checkbox"/> | - |

+ New parameter

Response

DEFAULT

Text response

- 메뉴 골라주세요
- Enter a text response variant

ADD MESSAGE CONTENT

Action

Enter action name...

| REQUIRED | PARAMETER NAME | ENTITY | VALUE | IS LIST | PROMPTS |
|-------------------------------------|----------------|------------------|----------------|--------------------------|-------------------|
| <input checked="" type="checkbox"/> | date | @sys.date | \$date | <input type="checkbox"/> | Define prompts... |
| <input type="checkbox"/> | time-period | @sys.time-period | \$time-period | <input type="checkbox"/> | - |
| <input checked="" type="checkbox"/> | roomnum | @roomnum | \$roomnum | <input type="checkbox"/> | Define prompts... |
| <input type="checkbox"/> | date-period | @sys.date-period | \$date-period | <input type="checkbox"/> | - |
| <input type="checkbox"/> | Enter name... | Enter entity... | Enter value... | <input type="checkbox"/> | - |
| <input type="checkbox"/> | Enter name... | Enter entity... | Enter value... | <input type="checkbox"/> | - |

+ New parameter

User says

Add user expression

오늘 3회의실 3시부터 4시까지 예약해

오늘 3회의실 3시부터 4시까지 예약해줘

내일 회의실 예약해

오늘 8시부터 3회의실 예약해줘

오늘 8시에 2회의실 예약해줘

오늘 8시에 2회의실 예약해

7일에 3시부터 4시까지 2회의실 예약해

3회의실 3시부터 4시까지 예약해줘

5시부터 8시까지 2회의실 예약해

내일 3회의실 예약해

Thanks

Any Questions?

You can send mail to
Susang Kim(healess1@gmail.com)