DeepBT and NLP Data Augmentation Techniques: A New Proposal and a Comprehensive Study

Taynan Maier Ferreira[1,2(✉)] and Anna Helena Reali Costa[1]

[1] Escola Politécnica, Universidade de São Paulo, São Paulo, Brazil
{taynan.ferreira,anna.reali}@usp.br
[2] Data Science Team, Itaú-Unibanco, São Paulo, Brazil

Abstract. Data Augmentation methods – a family of techniques designed for synthetic generation of training data – have shown remarkable results in various Deep Learning and Machine Learning tasks. Despite its widespread and successful adoption within the computer vision community, data augmentation techniques designed for natural language processing (NLP) tasks have exhibited much slower advances and limited success in achieving performance gains. As a consequence, with the exception of applications of back-translation to machine translation tasks, these techniques have not been as thoroughly explored by the wider NLP community. Recent research on the subject also still lacks a proper practical understanding of the relationship between data augmentation and several important aspects of model design, such as hyperparameters and regularization parameters. In this paper, we perform a comprehensive study of NLP data augmentation techniques, comparing their relative performance under different settings. We also propose Deep Back-Translation, a novel NLP data augmentation technique and apply it to benchmark datasets. We analyze the quality of the synthetic data generated, evaluate its performance gains and compare all of these aspects to previous existing data augmentation procedures.

Keywords: Data Augmentation · Natural Language Processing · Back-Translation · Machine learning

1 Introduction

Data Augmentation can be defined as any process of artificially creating new training data by applying class-preserving transformations to the original input data [5].

Partially supported by Itaú-Unibanco, CNPq (grants 25860/2016-7 and 530307027/2017-1), and CAPES (Finance Code 001). Any opinions, findings, and conclusions expressed in this manuscript are those of the authors and do not necessarily reflect the views, official policy or position of Itaú-Unibanco.

R. Cerri and R. C. Prati (Eds.): BRACIS 2020, LNAI 12319, pp. 435–449, 2020.
https://doi.org/10.1007/978-3-030-61377-8_30

In harmony with Statistical Learning Theory, which states that discrepancy between training and generalization error shrinks as the number of training examples increases [9], Data Augmentation has successfully been used in the Machine Learning and Deep Learning communities to artificially inflate data for training and, as a consequence, obtain models with greater generalization power.

Its application by researchers range from Image Processing [18,25], Sound and Speech Recognition [1,21], and Time Series [27] to Natural Language Processing [15,24].

Specifically within the Video Processing community, Data Augmentation has been used successfully for several years now, being part of the training process of models related to some of the greatest achievements in Image Classification tasks, such as the AlexNet [17], All-CNN [23], and ResNet [11] models.

These remarkable accomplishments have led researchers to investigate the underlying theoretical principles governing Data Augmentation, trying to shed some light into its relationship to aspects such as model learning process, decision surface, etc. These researches show that Data Augmentation improve generalization by both increasing invariance and penalizing model complexity [5].

Data Augmentation also can be considered a form of implicit regularization, closely related to explicit regularization techniques such as Weight Decay and Dropout. In fact, the works of [30] and [16] indicate that, under certain circumstances, Data Augmentation and Dropout can be considered equivalent methods. Other studies, on the other hand, state that Data Augmentation exhibit superior performance in comparison to explicit regularization methods [12].

Despite unquestionable success in computer vision tasks, NLP research has not yet benefited as largely from data augmentation systems. General NLP tasks and challenges are often characterized by the low – or often unsuccessful – usage of data augmentation techniques. When analyzing the solutions proposed for some of the SemEval Tasks over the period of 2017–2019, e.g., we observe the following.

1. SemEval-2017 Task 5: there is no mention to the use of Data Augmentation methods by any of the participants [4];
2. SemEval-2018 Task 1: among 75 teams, only 2 teams acknowledge the use of some kind of Data Augmentation procedure [19];
3. SemEval-2019 Task 5: within 74 participants, only one of them indicates using some kind of Data Augmentation in her or his system [2].

We hypothesize that the low adoption of data augmentation techniques within the NLP community is a consequence of its primitive state, still in its infancy when compared to the advanced methods used in image processing tasks.

To address this research gap and provide practitioners and researchers general guidelines on its use, we conduct an in-depth investigation of NLP data augmentation techniques. We compare their output and relative performance under various settings and study their sensibility to different parameters. We also investigate the relationship between data augmentation, which is an implicit regularization technique, with an explicit regularization technique, namely the dropout procedure.

To further advance the state-of-the-art knowledge on the subject, we also present Deep Back-Translation, an unprecedented data augmentation technique for NLP tasks which stacks more intermediate layers of translation between the original and synthesized sentences. We apply Deep Back-Translation to benchmark datasets and compare its outcomes to results generated by previous existing methods. To the best of our knowledge, we are the first to propose and study such a technique for data augmentation.

The remainder of this paper is organized as follows. In Sect. 2 we describe the main and most recent proposals in the field of Data Augmentation for Natural Language Processing. Section 3 presents the new proposed method and our main objectives in this paper, followed by the Experimental Setup at Sect. 4. After presenting the main Results in Sect. 5, we summarize our main contributions and conclude the paper in Sect. 6.

2 Related Work

In this section we highlight several recent researches in the realm of NLP-specific Data Augmentation techniques that relate to the present work.

Easy Data Augmentation (EDA) is a technique first proposed by [26]. This method consists of applying a set of simple operations to the original text in order to generate new synthetic texts. The operations, all randomly applied according to the parameter α, which controls the percentage of words changed in any given sentence, are Synonym Replacement, Random Insertion, Random Swap and Random Deletion. Though maybe original in its proposal as a pure data augmentation technique, the use of this set of operations closely resembles the noise injection procedure proposed by [7] as an auxiliary task to improve neural machine translation models.

Considered crucial to neural machine translation tasks nowadays [10], Back-Translation (BT) is another method for generating auxiliary synthetic data. First introduced by [22], the term Back-Translation was initially conceived specifically within the context of machine translation, whereby monolingual data was leveraged by translating *Target Language* $\rightarrow$ *Source Language* (hence the term *Back*) in order to obtain additional training data for the *Source Language* $\rightarrow$ *Target Language* final translation task. The first implementation of Back-Translation as a data augmentation method for down-stream tasks seems to be the work of [28], which used Back-Translation to rephrase original sentences (i.e. generating paraphrases), producing extra data and obtaining state-of-the-art results on question answering tasks.

Supported by its remarkable success in neural machine translation tasks, there has been an emergence of numerous variations to the traditional Back-Translation method.

Iterative Back-Translation (IterativeBT), proposed in [13], is a process where models are successively trained using data Back-Translated by the previous model. This cyclical training yields generation of models which are able to improve at each iteration.

Noised Back-Translation (NoisedBT)[1], presented by [7] is a variation where noise is injected to the Back-Translated text. In the seminal paper, three types of noise are used: random deletion of words, random replacement of words and random swapping of words. Despite the fact that final noised sentences are not realistic, the authors argue that the superior performance obtained by noise injection could be attributed to the model becoming robust to reordering and substitutions occurring naturally on texts.

Tagged Back-Translation (TaggedBT) [3], heavily influenced by the works of [7] and [14], proposes another hypothesis for the superiority of noise injection in Back-Translation postulated in [7]: instead of increased text diversity, noise injection would instead benefit the final model by signaling which data is synthetic and which is original data.

Despite all of the proposed variations of Back-Translation, few researches have investigated how different choices and parameters of Back-Translation can affect its performance in down-stream tasks. Questions such as the impact of language used for translation (pivotal language), or even the effectiveness of EDA compared to BT, remain still open. The ablation studies in data augmentation performed by [28] seem to be the closest to partially address some of these open issues, though leaving the majority of the questions here proposed still unanswered.

Also, none of the Back-Translation variations take advantage of language translation at a greater level for general NLP Tasks. While IterativeBT is only applicable for Machine Translation tasks, TaggedBT and NoisedBT just add additional handcrafted information to the data. To the best of our knowledge, there has not been proposed any method that leveraged translations one step further when compared to traditional Back-Translation.

3 Deep Back-Translation

In this paper we contribute a new method for data augmentation, named Deep Back-Translation (DeepBT), which adds more layers of intermediate translations between the original text and the final paraphrase. Hence, using capital letters to represent original and final languages (which are always the same) and arrows to represent translations to languages L, while in the original Back-Translation we always have

$$A \rightarrow L \rightarrow A,$$

in DeepBT we could have n intermediate layers of translations:

$$A \rightarrow L_1 \rightarrow L_2 \rightarrow ... \rightarrow L_n \rightarrow A.$$

Figure 1 illustrates the difference with a 2-layer Deep Back-Translation.

For any given Back-Translation procedure (be it the Deep version or any other) we can define the concept of Multiplication Factor, which informs us by

[1] The term Noised Back-Translation was not used in [7], but coined by [3].

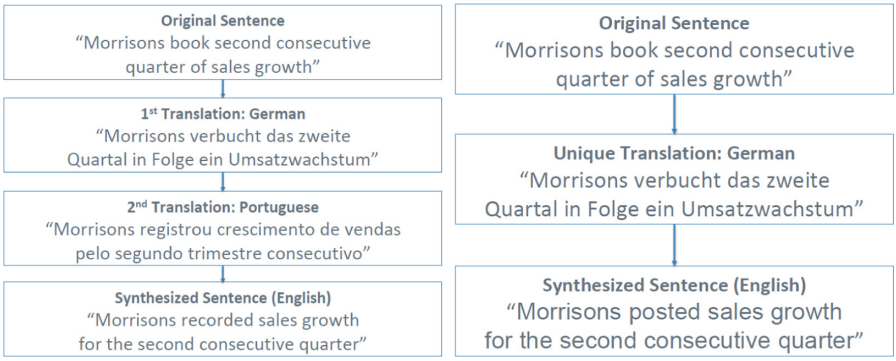


Fig. 1. Graphic representation of Deep- and traditional Back-Translation procedures. In DeepBT (a) we stack several intermediate layers (red) of translation. In traditional Back-Translation (b) there is always only one translation between original and synthesized text.

how much the available data has been multiplied by the data augmentation process.

Following the rationale of the Back-Translation method, which generates paraphrases with the same meaning and label as the original one, the Deep Back-Translation technique is created under the hypothesis that using several intermediate languages between the original and destination one could increase the difference between phrase and paraphrases, while still maintaining original meaning and label. With greater variability in training data we assume that we could reduce overfitting and achieve greater performance.

Therefore, we have three main objectives to be investigated in this paper, all addressing research gaps found in the NLP data augmentation literature.

First, we propose Deep Back-Translation, a new data augmentation technique. The development and evaluation of new NLP-specific data augmentation techniques is a relevant purpose since, as discussed before, it addresses the lack of new methods in this research area, in contrast to the continuous progress in data augmentation techniques in the computer vision community.

Second, and once again drawing inspiration from advances in the latest image processing research, we perform a systematic and comparative study of the main NLP data augmentation techniques. In particular, we aim at analyzing the relationship between data augmentation (an implicit regularization technique) to dropout (an explicit regularization procedure). This type of study has not yet been carried out within the NLP community, not even by the seminal papers that introduced the chosen methods.

Finally, we also want to address a research question related to the impact of linguistic styles in data augmentation techniques. Since NLP data augmentation methods rely, among others, on translation and the use of dictionaries, we want

to investigate whether there are any performances variations when these methods are applied to formal or informal texts.

Hence, with this investigation we hope to be able to advance the state-of-the-art on the subject of data augmentation in NLP and to deepen the understanding of techniques that overcome the bottleneck of limited labeled data.

4 Experimental Setup

We present below the main aspects related to the experimental setup, introducing benchmark datasets and machine learning techniques used. We also summarize steps and procedures involved in obtaining results presented thereafter.

To accomplish the objectives mentioned in Sect. 3, we selected two methods with which DeepBT will be compared: traditional Back-Translation and EDA. These three data augmentation techniques were applied to the same benchmark datasets and using the same set of hyperparameters. Our baseline for comparing the outcomes will be the output of training without any data augmentation method and Mean Squared Error (MSE) was used as the performance metric in all experiments.

To address the first objective, we compare DeepBT to the other data augmentation procedures. To tackle the second objective we analyzed how each of the data augmentation techniques responded to varying dropout values. Finally, the third objective is reached by inspecting response of aforementioned data augmentation methods to each of the benchmark datasets.

Following [26] and others, data augmentation procedures were applied under varying dataset percentage usages (every decile from 10% to 100%), so as to assess the impact of data availability on results and conclusions.

4.1 Benchmark Datasets

We chose two datasets provided by the SemEval 2017 Task 5 challenge [4], in Tracks 1 and 2. Both of the datasets are used in NLP regression tasks within the domain of sentiment analysis.

The dataset of Track 1 (Microblog Messages) consists of messages collected in two different microblog platforms (StockTwits and Twitter) related to stock market events, discussion and assessments. The second dataset, associated to Track 2 (News Statements & Headlines), contains financial news headlines and texts crawled from sources on the Internet.

In both datasets, the label is a continuous sentiment score ranging from -1 (very negative) to $+1$ (very positive), with 0 being neutral. This sentiment score was labeled by domain experts to reflect the point of view of investors regarding negative, neutral or positive prospective trends for companies or stocks.

4.2 Data Augmentation and Preparation Procedure

Here we describe details about the investigated data augmentation methods and how the data was prepared.

Back-Translation and Deep Back-Translation. Since the focus of this paper is not related to the translation model itself and considering the high effort and computational resources required for training a translation language model, we chose to use a widely accepted translation API, namely the Google Cloud Translation API[2]. This allows for rapid and high quality translation for several different languages with minimal associated costs.

So as to use languages from different families, the languages chosen for both Deep- and traditional Back-Translation procedures were German, Russian and Portuguese. Thus, in choosing West Germanic (German), East Slavic (Russian) and Western Romance (Portuguese) languages, we hypothesize that this diversity could bring greater heterogeneity to the paraphrases generated by Back-Translation. For assessing the DeepBT method the experiments were carried out with 2-layer translation settings (e.g. $English \rightarrow Russian \rightarrow German \rightarrow English$).

Easy Data Augmentation (EDA). The EDA method accepts two parameters to control its data augmentation process, namely α and n_{aug}. While the first controls the percentage of words in a sentence that are changed, the last is responsible for indicating the number of augmented sentences in the output. In the seminal paper [26], the authors propose general guidelines regarding optimal α and n_{aug} parameters to be used, depending on the size of the training dataset. For our experiments, to allow for best performance, we follow these guidelines, using $\alpha = 0.05$ and $n_{aug} = 8$ for both datasets.

The EDA procedure was applied using the original code made available by the authors [26].

Data Preparation. Well established NLP data preparation steps were equally employed on both datasets: stop words and punctuation removal, tokenization and lowercasing. In each sentence, the company or cashtag referred to were replaced with a generic token. All implemented models shared the same data preparation process.

The fact that we used original and synthetic data for training and validation in the cross-validation procedure required special attention as to how original and artificial data were distributed in the training phase. The presence, e.g., of original sentence in the training set and of the corresponding synthesized sentence in the validation set would entail the occurrence of data leakage. As a result, cross-validation was carefully designed so as to have original and synthesized sentences always together at the training or at the validations sets.

4.3 Machine Learning Model

We based our experiments in a Convolutional Neural Network (CNN). Choices regarding architecture, hyperparameters and feature representation were done

[2] https://cloud.google.com/translate/docs.

inspired by state-of-the-art models in sentiment analysis tasks [29] and the winning architectures of the SemEval 2017 Task 5 challenge [4,6].

The CNN architecture was composed of 2 convolutional layers followed by a single dense layer. Mean Squared Error (MSE) was picked as the loss function and *Adam* as the optimizer. The activation function for hidden and output layers where the *ReLU* activation function and the *tanh* function respectively. The "Wikipedia 2014 + Gigaword 5" pre-trained *GloVe* was used as our input word embedding [20]. To avoid drawing conclusions that are specific to some arbitrary chosen hyperparameter values [8] and seed, models were trained and results were averaged along the following values of hyperparameters: number of neurons in dense layer ({100, 150}), size of filters ({2, 3}) and dropout value ({0.0, 0.1, 0.2}).

5 Results

We begin showing the result of each method in the generation of artificial text. Next we present results related to each of our three main objectives. We end this section discussing the results and drawing conclusions from them.

5.1 Synthesized Data

Before diving into model results obtained by using each of the techniques, it is interesting to analyze synthesized data outputs of each of the methods so as to gain better insights into final model outcomes. Some examples are shown in Tables 1 and 2 for the Microblog Messages and for the News Statements & Headlines respectively.

Table 1. Synthesized Data generated by different data augmentation techniques in the Microblog Messages Dataset.

Original Sentence	EDA	Back-Translation	Deep Back-Translation
"watching for bounce tomorrow"	"watching for resile tomorrow"	"Watch out for jumping power tomorrow"	"I look forward to jumping tomorrow"
"Bad governance. not confident in core biz"	"in governance not confident bad core biz"	"Bad governance. not confident in core business"	"Bad governance. not confident in core business"
"#OwnItDon'tTradeIt"	"ownitdonttradeit"	"# OwnIt-Don'tTradeIt"	"# OwnItDon'tTradeIt"

We observe that, due to the random noise injected by EDA (random swap, random deletion, etc.) sentences generated by this method normally suffer from lack of correct grammatical or syntactical structure. Both Back-Translation methods, on the other hand, generally yield sentences with correct grammatical structure.

Table 2. Synthesized Data generated by different Data Augmentation techniques in the News Statements & Headlines respectively Dataset.

Original Sentence	EDA	Back-Translation	Deep Back-Translation
"Morrisons book second consecutive quarter of sales growth"	"Morrisons of second consecutive quarter book sales growth"	"Morrisons posted sales growth for the second consecutive quarter"	"Morrisons recorded sales growth for the second consecutive quarter"
"Britain's FTSE lifted by solid Kingfisher"	"britains ftse lifted united kingdom of great britain and northern ireland by solid kingfisher"	"Britain's FTSE lifted by solid kingfishers"	"British FTSE filmed by solid kingfishers"
"Brazil Vale says will appeal ruling to block assets for dam burst"	"says vale brazil will appeal ruling to block assets for dam burst"	"Brazil Vale will appeal to block assets for the dam breach"	"Brazil Vale Appeals Asset Lockout Dam Dam"

When comparing Back-Translation and DeepBT, we see that often the first is characterized by better capture of the true meaning of the sentence.

DeepBT may more easily loose the original ideas expressed by the original sentence, such as in *"I look forward to jumping tomorrow"* or in *"British FTSE filmed by solid kingfishers"*.

Interesting insights arise when analyzing the output of each method regarding formal (News & Headlines) and informal (Microblog Messages) texts. In the second example of the Microblog Messages Dataset we can see that, while EDA maintained the original term *biz*, both of the Back-Translation methods were able to identify this expression and output it as *business*. In contrast, the third example of this same dataset shows that none of the methods were able to capture the meaning of *#OwnItDon'tTradeIt* to generate paraphrases, probably due to the absence of spaces and the use of octothorpe sign (*hashtag*).

5.2 Performance

We now present the results for DeepBT for a variety of factors – such as percent of dataset used, multiplication factor and language of translation – and compare them to the results obtained by other data augmentation techniques.

We start comparing relative performance gain of DeepBT, BT and EDA against the baseline trained on 100% of the Microblog Messages dataset in Fig. 2a. Results show DeepBT achieving greater performance when compared to the traditional Back-Translation, though still inferior to EDA when smaller percentages of the dataset are made available. When the entire dataset is used in training, the three data augmentation techniques converge to similar results.

Figures 2b to 2d exhibit how each data augmentation technique's performance is affected by the Multiplication Factor parameter. We notice that BT and DeepBT yield superior results with greater Multiplication Factor, in contrast to what is observed in EDA. It is also interesting to notice how both translation-based methods start with much worse performance than the baseline with low percentages of dataset use, but rapidly respond to growing data availability.

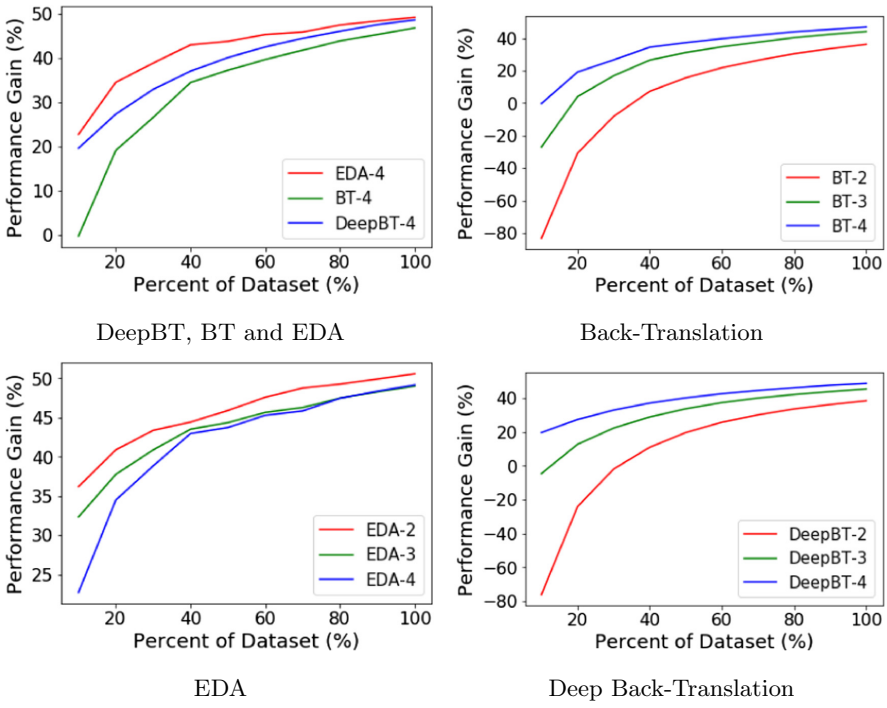
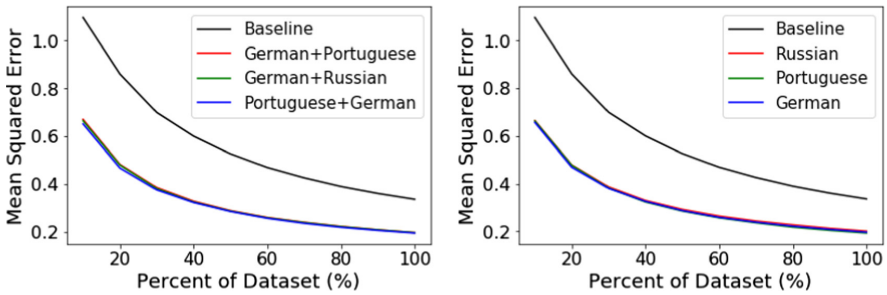


Fig. 2. Comparison of data augmentation techniques and response of each technique to varying Multiplication Factor in the Microblog Messages dataset. Performance gain is measured against the Baseline trained on 100% of the dataset. Labels are in the format "Method-Multiplication Factor".

Figure 3 presents models performance response for each of the chosen languages used for translation in DeepBT and traditional Back-Translation. We observe no performance distinction between any of the chosen languages used for translation purposes.

Table 3 presents how model performance is affected by the dropout hyper-parameter in both datasets. In contrast to the results obtained by [12] in image processing tasks, we obtained stable performance under varying dropout parameter.

An intriguing outcome of our experiments is that EDA had an average performance gain above 20% in all analyzed settings, far superior than the average



Deep Back-Translation Back-Translation

Fig. 3. Comparison of models' performance for different languages choices for Back-Translation and Deep Back-Translation on the News Statements & Headlines dataset. Since we used a 2-layer DeepBT, in 3a each line represents a combination of two languages used in sequenced translation.

gains between 1% and 3% achieved by the original work proposing EDA [26]. We hypothesize that this could be attributable to difference in the characteristics of the dataset used in our experiments compared to the ones used in [26] (like linguistic style of the texts) and compare how each method responds to each dataset. The results are shown in Fig. 4.

While translation-based methods yielded similar results, regardless of the dataset, this was not the case for the EDA method. In the latter, we observed far better performance gains in the News Statements & Headlines dataset (formal language), 15%-20% higher than the outcomes obtained in the Microblog Messages dataset (informal language). Considering that the same methodology was applied to both datasets, which also have similar size, we hypothesize that this contrasting behavior can be attributable to greater stability of translation-based methods in comparison to EDA when facing different linguistic styles. Further experiments should be put forward to confirm this conjecture.

Table 3. Average MSE obtained by data augmentation techniques under different dropout values in the Benchmark Datasets.

Method	Microblog Messages			News Statements & Headlines		
	Dropout			Dropout		
	0.0	0.1	0.2	0.0	0.1	0.2
DeepBT	0.13 ± 0.02	0.12 ± 0.02	0.12 ± 0.02	0.21 ± 0.07	0.19 ± 0.06	0.19 ± 0.06
BT	0.15 ± 0.04	0.13 ± 0.03	0.13 ± 0.02	0.21 ± 0.07	0.18 ± 0.05	0.19 ± 0.06
EDA	0.12 ± 0.02	0.12 ± 0.02	0.12 ± 0.02	0.13 ± 0.02	0.13 ± 0.03	0.13 ± 0.02

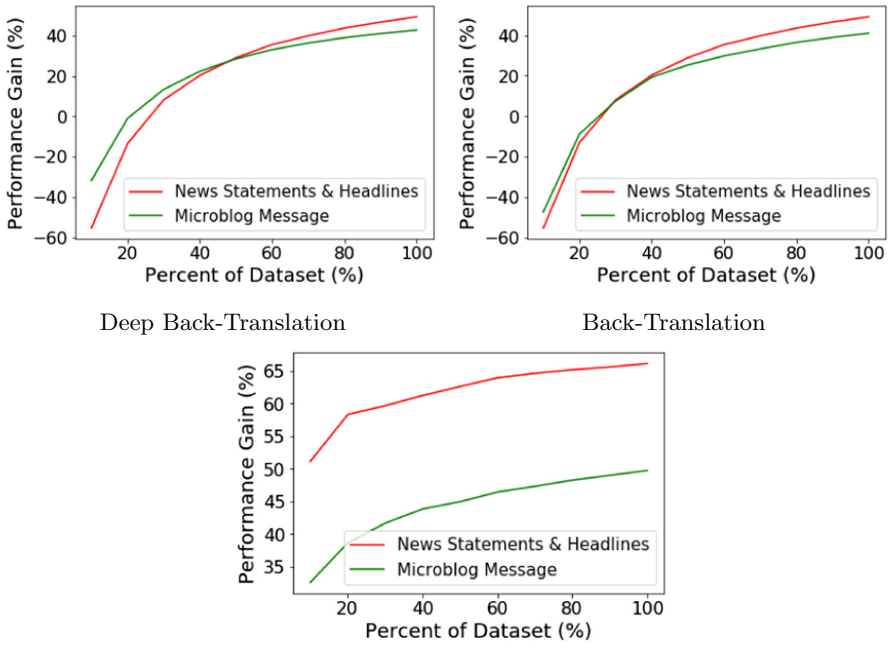


Fig. 4. Comparison of performance gain brought by data augmentation techniques at each dataset. Translation-based methods showed grater homogeneity in outcomes when compared to EDA.

6 Conclusion

In this paper, we performed an in-depth study of data augmentation techniques and presented a new method, performing a systematic evaluation of it. We summarize below our main findings and contributions:

1. **Proposal of new Data Augmentation technique**: with Deep Back-Translation, we add to the state-of-the-art on NLP data augmentation techniques an unprecedented method. Despite its results being somewhat similar to traditional Back-Translation in most experimental setups, this new method yielded superior results in some scenarios, such as in low data availability settings. Furthermore, DeepBT showed greater stability among different datasets when compared to EDA.
2. **Assessing the impact of Dropout in various data augmentation techniques**: to the best of our knowledge, we are the first authors to perform an evaluation of the relationship between these explicit and implicit regularization techniques in a NLP task. This is extremely relevant since the combined use of implicit and explicit regularization techniques is commonplace among

practitioners, despite evidences from the computer vision community indicating that data augmentation could yield better results when used alone [12].

3. **Comparison of the effect of Data Augmentation techniques in Formal and Informal texts**: since NLP data augmentation techniques rely on auxiliary language processing tasks such as translation and synonym replacement, it is interesting to compare their relative behavior in texts with different levels of formality. The observed difference in performance underlines the importance of developing data augmentation techniques that are robust to linguistic preferences prevalent in informal texts, such as contractions, abbreviations, colloquialism, slang, among others.

With this paper, we hope to encourage further use of Back-Translation (and its variations) as auxiliary method for data augmentation in NLP tasks outside of the realm of neural machine translation, where it has already been heavily used.

As future work, we would like to assess Deep Back-Translation at a wider variety of settings, including classification datasets and tasks outside of the sentiment analysis domain. Also, we would like to broaden the scope of comparison of different NLP data augmentation techniques, including e.g. those involving the use of Generative Adversarial Networks. Finally, we would also like to explore data augmentation techniques that are not based on heuristic and handcrafted procedures, but rather learned in a machine learning framework so as to optimize the final model output.

References

1. Bao, F., Neumann, M., Vu, T.: CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition. In: Proceedings of the Interspeech 2019, pp. 2828–2832, September 2019. https://doi.org/10.21437/Interspeech.2019-2293
2. Basile, V., et al.: SemEval-2019 task 5: multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63. Association for Computational Linguistics, Minneapolis, June 2019. https://doi.org/10.18653/v1/S19-2007
3. Caswell, I., Chelba, C., Grangier, D.: Tagged back-translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pp. 53–63. Association for Computational Linguistics, Florence, August 2019. https://doi.org/10.18653/v1/W19-5206
4. Cortis, K., et al.: SemEval-2017 task 5: fine-grained sentiment analysis on financial microblogs and news. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 519–535. Association for Computational Linguistics, Stroudsburg (2017). https://doi.org/10.18653/v1/S17-2089
5. Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., Re, C.: A kernel theory of modern data augmentation. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 1528–1537. PMLR, Long Beach, 09–15 June 2019. http://proceedings.mlr.press/v97/dao19b.html

6. Davis, B., Cortis, K., Vasiliu, L., Koumpis, A., Mcdermott, R., Handschuh, S.: Social sentiment indices powered by X-scores. In: ALLDATA 2016, The Second International Conference on Big Data, Small Data, Linked Data and Open Data, Lisbon, Portugal (2016)
7. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 489–500. Association for Computational Linguistics, Brussels, October–November 2018. https://doi.org/10.18653/v1/D18-1045
8. Ferreira, T., Paiva, F., Silva, R., Paula, A., Costa, A., Cugnasca, C.: Assessing regression-based sentiment analysis techniques in financial texts. In: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional, pp. 729–740. SBC, Porto Alegre (2019). https://doi.org/10.5753/eniac.2019.9329, https://sol.sbc.org.br/index.php/eniac/article/view/9329
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press, Cambridge (2016)
10. Graça, M., Kim, Y., Schamper, J., Khadivi, S., Ney, H.: Generalizing back-translation in neural machine translation. In: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pp. 45–52. Association for Computational Linguistics, Florence, August 2019. https://doi.org/10.18653/v1/W19-5205
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90
12. Hernández-García, A., König, P.: Further advantages of data augmentation on convolutional neural networks. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11139, pp. 95–103. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01418-6_10
13. Hoang, V.C.D., Koehn, P., Haffari, G., Cohn, T.: Iterative back-translation for neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 18–24. Association for Computational Linguistics, Melbourne, July 2018. https://doi.org/10.18653/v1/W18-2703
14. Imamura, K., Fujita, A., Sumita, E.: Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In: Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 55–63. Association for Computational Linguistics, Melbourne, July 2018. https://doi.org/10.18653/v1/W18-2707
15. Kobayashi, S.: Contextual augmentation: data augmentation by words with paradigmatic relations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 452–457. Association for Computational Linguistics, New Orleans, June 2018. https://doi.org/10.18653/v1/N18-2072
16. Konda, K.R., Bouthillier, X., Memisevic, R., Vincent, P.: Dropout as data augmentation. arXiv abs/1506.08700 (2015)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems. NIPS 2012, vol. 1, pp. 1097–1105. Curran Associates Inc., Red Hook (2012)
18. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW), pp. 117–122 (2018)

19. Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: SemEval-2018 task 1: affect in tweets. In: Proceedings of The 12th International Workshop on Semantic Evaluation, pp. 1–17. Association for Computational Linguistics, New Orleans, June 2018. https://doi.org/10.18653/v1/S18-1001
20. Pennington, J., Socher, R., Manning, C.D.: GloVe : global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
21. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017)
22. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 86–96. Association for Computational Linguistics, Berlin, August 2016. https://doi.org/10.18653/v1/P16-1009
23. Springenberg, J., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015). http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a
24. Sugiyama, A., Yoshinaga, N.: Data augmentation using back-translation for context-aware neural machine translation. In: Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pp. 35–44. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/D19-6504
25. Taylor, L., Nitschke, G.: Improving deep learning with generic data augmentation. In: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1542–1547 (2018)
26. Wei, J., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 6382–6388. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/D19-1670
27. Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., Xu, H.: Time series data augmentation for deep learning: a survey (2020)
28. Yu, A.W., et al.: QANET: combining local convolution with global self-attention for reading comprehension. CoRR abs/1804.09541 (2018). https://arxiv.org/pdf/1804.09541
29. Zhang, Y., Wallace, B.C.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In: Proceedings of the 8th International Joint Conference on Natural Language Processing, pp. 253–263 (2017)
30. Zhao, D., Yu, G., Xu, P., Luo, M.: Equivalence between dropout and data augmentation: a mathematical check. Neural Netw. Off. J. Int. Neural Netw. Soc. **115**, 82–89 (2019)