TAYNAN MAIER FERREIRA

DATA AUGMENTATION METHODS IN NATURAL LANGUAGE PROCESSING

São Paulo
2021

TAYNAN MAIER FERREIRA

DATA AUGMENTATION METHODS IN NATURAL LANGUAGE PROCESSING

Dissertation submitted to the Escola Politécnica of the University of São Paulo to obtain the degree of Master of Science.

São Paulo
2021

TAYNAN MAIER FERREIRA

DATA AUGMENTATION METHODS IN NATURAL LANGUAGE PROCESSING

Versão Corrigida

Dissertation submitted to the Escola Politécnica of the University of São Paulo to obtain the degree of Master of Science.

Concentration area:
Computer Engineering

Advisor:
Prof. Dra. Anna Helena Reali Costa

São Paulo
2021

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, ______ de ____________________ de __________

Assinatura do autor: ________________________

Assinatura do orientador: ________________________

Catalogação-na-publicação

Ferreira, Taynan
DATA AUGMENTATION METHODS IN NATURAL LANGUAGE PROCESSING / T. Ferreira -- versão corr. -- São Paulo, 2021.
p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Aprendizado de Máquina 2.Processamento de Linguagem Natural 3.Aumento de Dados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

Prof. Dra. Anna Helena Reali Costa

Prof. Dr. Thiago Alexandre Salgueiro Pardo

Prof. Dra. Aline Marins Paes Carvalho

ACKNOWLEDGMENTS

I would first like to thank my thesis supervisor, Professor Anna Helena, whose guidance and patient support were fundamental since the very beginning of this journey.

My special thanks to my colleagues from the C^2D (Centro de Ciência de Dados) and of Itaú-Unibanco for their collaboration and for all our stimulating discussions.

I deeply appreciate the encouragement and support of my parents Antonio and Martina, and of my brother Renan.

I am very grateful for the friendship and encouragement of my close friends: Felipe Contatto, Lucas Simone, Ignacio Pozo, Josué Bratfich, Renan Cunha and Cristiano Lopes.

My most profound and sincere gratitude to my loving wife Talita, without whose care, love and support I would not be able to accomplish anything – and for the opportunity to share with her the most profound and exciting moments of my life.

Finally and foremost, I would like to praise and thank God, for His graces and sustenance.

RESUMO

Métodos de aumento de dados (AD) – uma família de técnicas desenhada para a geração de dados de treino sintéticos – têm demonstrado resultados notáveis em diversas tarefas de Aprendizado Profundo e Aprendizado de Máquina. Apesar de sua adoção ampla e bem-sucedida dentro da comunidade de visão computacional, técnicas de AD desenhados para tarefas de Processamento de Linguagem Natural (PLN) têm demonstrado avanço muito mais lento e limitado sucesso em ganho de desempenho. Como consequência, com a exceção da adoção de Back-Translation em tarefas de tradução, essas técnicas não tem sido exploradas tão profundamente e de forma ampla pela comunidade de PLN. Não há uma visão unificada ou análise comparativa entre os vários métodos de AD disponíveis. Além disso, ainda não se tem um entendimento prático adequado sobre o relacionamento entre AD e diversos outros aspectos importantes do desenho de um modelo, como dados de treino e parâmetros de regularização. Nesse trabalho, realizamos um profundo estudo de técnicas de AD em PLN, comparando seus desempenhos relativos sob diferentes cenários em tarefas de Análise de Sentimentos. Também propomos Deep Back-Translation, uma nova técnica de AD para PLN. Nós realizamos uma análise qualitativa e quantitativa do dado sintético, avaliamos seu ganho de desempenho e comparamos todos esses aspectos com procedimentos prévios de AD.

Palavras-Chave – Aumento de Dados, Processamento de Linguagem Natural, Back-Translation, Aprendizado de Máquina.

ABSTRACT

Data Augmentation (DA) 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, DA 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. There is no unified view or comparative analysis between the various DA methods available. Furthermore, there still lacks a proper practical understanding of the relationship between DA and several important aspects of model design, such as training data and regularization parameters. In this work, we perform a comprehensive study of NLP DA techniques, comparing their relative performance under different settings in Sentiment Analysis tasks. We also propose Deep Back-Translation, a novel NLP DA technique. We perform qualitative and quantitative analysis of generated synthetic data, evaluate its performance gains and compare all of these aspects to previous existing DA procedures.

Keywords – Data Augmentation, Natural Language Processing, Back-Translation, Machine Learning.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

BC - Between-Class

BigGAN - Big GAN

BT - Back-Translation

CGAN - Conditional GAN

CNN - Convolutional Neural Network

DA - Data Augmentation

DCGAN - Deep Convolutional GAN

DeepBT - Deep Back-Translation

DL - Deep Learning

EDA - Easy Data Augmentation

ERM - Empirical Risk Minimization

GAN - Generative Adversarial Net

GECA - Good-Enough Compositional Data Augmentation

MCMC - Markov Chain Monte Carlo

MHA - Metropolis-Hastings attack

ML - Machine Learning

NLP - Natural Language Processing

NMT - Neural Machine Translation

MSDA - Mixed Sample Data Augmentation

NN - Neural Network

PBA - Population Based Augmentation

RD - Random Deletion

RI - Random Insertion

RS - Random Swap

SA - Sentiment Analysis

SR - Synonym Replacement

t-SNE - t-Distributed Stochastic Neighbor Embedding

WGAN - Wasserstein GAN

CONTENTS

1 INTRODUCTION

When building Machine Learning (ML) models to address supervised learning tasks, one has the objective to be able to predict unseen inputs based on previously seen inputs – i.e., the goal is to minimize the so called generalization error (GOODFELLOW; BENGIO; COURVILLE, 2016), or in other words, reduce overfitting. Many strategies have been developed in order to increase the generalization power of ML models (SHORTEN; KHOSHGOFTAAR, 2019): dropout, batch normalization, transfer learning, pretraining, One-shot and Zero-shot learning are some of them. Data Augmentation (DA), the focus of this work, is yet another strategy to reduce overfitting.

DA can be defined as any process of artificially creating new training data by applying class-preserving transformations to the original input data (DAO et al., 2019). In consonance with one of the most important results from Statistical Learning Theory, which states that discrepancy between training and generalization error diminishes with increasing training examples (GOODFELLOW; BENGIO; COURVILLE, 2016), DA has successfully been used in the ML and Deep Learning (DL) communities to synthetically inflate data for training and, as a result, obtain models with greater generalization power.

Since DA tackles the issue of overfitting from the training dataset itself, it is a general and task-agnostic approach, whose application varies from Image Processing (TAYLOR; NITSCHKE, 2018; MIKOłAJCZYK; GROCHOWSKI, 2018) to Sound and Speech Recognition (BAO; NEUMANN; VU, 2019; SALAMON; BELLO, 2017), from Time Series (WEN et al., 2020) to Natural Language Processing (NLP) (KOBAYASHI, 2018; SUGIYAMA; YOSHINAGA, 2019). Within the Computer Vision community, DA has been successfully used for several years now, being part of the training process of models responsible for some of the greatest achievements in Image Classification tasks, such as the AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), All-CNN (SPRINGENBERG et al., 2015), and ResNet (HE et al., 2016) models.

These remarkable accomplishments have led researchers to investigate the underlying theoretical principles governing DA, trying to shed some light into its relationship to aspects such as model learning process, decision surface, among others. These researches show that DA improves generalization by both increasing invariance and penalizing model complexity (DAO et al., 2019). DA 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 (ZHAO et al., 2019) and (KONDA et al., 2015) indicate that, under certain circumstances, DA and Dropout can be considered equivalent methods. Other studies, on the other hand, state that DA exhibit superior performance in comparison to explicit regularization methods (HERNÁNDEZ-GARCÍA; KÖNIG, 2018).

More recently, state-of-the-art DA techniques in Image Processing have shifted from prior-knowledge-oriented handcrafted transformations to techniques that learn the augmentation transformations themselves. That is, instead of leveraging domain-specific knowledge to define the set of rules for generating artificial data, new methods *learn* the needed transformations. The development of an end-to-end approach, where DA NLP techniques are optimized to output artificial text that is best suited for the learning process of the final model seems a natural and promising research path (EDUNOV et al., 2018). In summary, DA applications in image processing tasks have been an active area of research with new methods being proposed every year, some of them with highly promising results. AutoAugment (CUBUK et al., 2019) and Population Based Augmentation (PBA) (HO et al., 2019) are some of the most recent ones, just to name a few.

Despite unquestionable success in computer vision tasks, NLP research has not yet benefited as largely from DA systems. General NLP tasks and challenges are often characterized by the low – or often unsuccessful – usage of DA techniques. When analyzing the solutions proposed for some of the SemEval Tasks[1] 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 DA methods by any of the participants (CORTIS et al., 2017);

2. SemEval-2018 Task 1: among 75 teams, only 2 teams acknowledge the use of some kind of DA procedure (MOHAMMAD et al., 2018);

3. SemEval-2019 Task 5: within 74 participants, only one of them indicates using some kind of DA (BASILE et al., 2019).

[1] SemEval is a series of international NLP research workshops whose mission is to advance the current state of the art in semantic analysis and to help create high-quality annotated datasets in a range of increasingly challenging problems in natural language semantics. https://semeval.github.io/

We hypothesize that the low adoption of DA techniques in NLP tasks and challenges results from the fact that NLP-specific DA methods have not been so successful as the ones used in the computer vision community. The reason for that may be twofold: on the one hand, DA research has been still primarily focused on computer vision tasks. On the other hand, NLP-specific DA strategies have shown themselves difficult to develop.

To address this research gap and provide practitioners and researchers general guidelines on its use, we conduct an in-depth investigation of some of the most important NLP DA techniques. We compare their output and relative performance under various settings and study their sensitivity to different parameters. We investigate the relationship between DA, which is an implicit regularization technique, with an explicit regularization technique, namely the dropout procedure. We also further discuss and evaluate Deep Back-Translation (DeepBT), the DA technique for NLP tasks we proposed in (FERREIRA; COSTA, 2020). We apply DeepBT to benchmark datasets and compare its outputs to results generated by previous existing methods.

To the best of our knowledge, this is the first work to perform a broad and in-depth investigation of NLP DA methods by reviewing relevant literature and tackling some of the most significant research gaps. We summarize in the next section the main objectives of this work.

1.1 Objectives

We have four main objectives to be accomplished in this work, all addressing research gaps found in the NLP DA literature.

NLP DA literature review: first, we provide readers with an overview of the literature related to DA techniques in NLP tasks. This review starts with a brief introduction to the theoretical framework that supports DA methods, followed by an overview of the main approaches that have been explored as NLP DA techniques.

Research Gaps distillation: second, and based on the aforementioned review, we summarize the main research gaps found on NLP DA literature. We highlight research gaps found not only in terms of techniques, but also in terms of problems tackled and the lack of comparative studies.

DeepBT: we also propose DeepBT, a new DA technique. This new method, alongside traditional Back-Translation (BT) method, is thoroughly studied, exploring how different parameters and design choices affect model performance. The development and evaluation of new NLP-specific DA 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 DA techniques in the computer vision community.

Systematic comparative study: finally, and once again drawing inspiration from the aforementioned research gaps, we perform a systematic and comparative study of the main NLP DA techniques. We do not limit our analysis to model performance, but also perform qualitative and quantitative evaluation of DA methods' outputs. This type of study has not yet been carried out within the NLP tasks.

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

1.2 Organization of the manuscript

The remainder of this manuscript is organized as follows. We present the Theoretical Framework supporting DA in Chapter 2. We follow in Chapter 3 with an overview of the DA research landscape, consolidating and discussing some of the main and most recent proposals in the field of DA, specially those more closely related to NLP. Chapter 4 presents DeepBT, the new NLP DA technique developed in this work. We present the methodology and implementation details of the comparative analysis of NLP DA techniques in Chapter 5, followed by the respective results in Chapter 6. Finally, we close with relevant conclusions and contributions, as well as interesting future work and research paths in Chapter 7.

2 THEORETICAL FRAMEWORK

In this chapter we examine the theoretical background that supports DA methods. We start with a brief overview of Machine Learning and Statistical Learning in Section 2.1, followed by theoretical considerations related to DA in Section 2.2.

2.1 Machine Learning and Statistical Learning

Learning algorithms are procedures that are able to learn from data. More precisely, one can consider a procedure as a learning algorithm whenever it is able to satisfy the following condition: to "learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P improves with experience E" (MITCHELL, 1997). ML is known as the study of learning algorithms.

This above definition is able to embrace the wide variety of applications and problems solved with ML. Task T includes common problems, such as Classification, Regression, Machine Translation, among various others. The performance measure P is responsible for quantifying the algorithms' success in its learning task. Depending on the task T being tackled, among common performance measures we could cite error rate, accuracy, quadratic error and cosine similarity. The experience E determines the source from which the process will learn. In supervised learning, the algorithms' experience E is a dataset containing features associated with a label or target (GOODFELLOW; BENGIO; COURVILLE, 2016).

When building a supervised ML model, the final purpose is to learn from previously seen inputs to be able to best predict unseen inputs. To this end, the ML model is trained on a training set and one tries to minimize the **training error** – the error calculated on observed data in the training phase. The real goal, however, is to minimize the error the model will perform on unobserved data – the so called **generalization error**. The generalization error can be defined as the expected value of the error on inputs drawn from a distribution expected to be found on the prediction phase (GOODFELLOW; BENGIO;

COURVILLE, 2016).

Many practical problems – such as classification, pattern recognition, regression and density estimation, among others – are particular cases of a more general problem related to the process of function estimation from a given collection of data (VAPNIK, 1999). This type of problem can be analyzed under the general statistical framework of minimizing expected loss using observed data, which we develop below.

In supervised learning we aim at finding a function $f \in \mathcal{F}$ from a set of functions that best describes the relationship between random feature vector X and target random vector Y, which follows a fixed, but unknown, joint distribution $P(X, Y)$ (VAPNIK, 1999; ZHANG et al., 2018). To this end, we can define a loss function l that penalizes the difference between the predicted outcome $f(x)$, $x \in X$, and the actual outcome y, with $y \in Y$. Therefore, our goal is to minimize the average of the loss function l over the distribution P, a measure known as *risk functional* or *expected risk*:

$$R(f) = \int l(f(x), y))dP(x, y). \tag{2.1}$$

Since the distribution $P(X, Y)$ is unknown, in order to minimize the risk functional in Eq. 2.1 we can make use of an induction principle called Empirical Risk Minimization (ERM), which approximates the risk functional in Eq. 2.1 by the *empirical risk* functional defined as

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i). \tag{2.2}$$

Hence, for finding the desired function f, we rely on a given set of training data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ formed by a set of independent identically distributed (i.i.d) observations drawn from the joint distribution $P(X, Y)$.

Therefore, solving the learning problem using ERM consists of estimating the function $f \in \mathcal{F}$ that minimizes the *empirical risk*:

$$\arg\min_{f \in \mathcal{F}} R_{emp}(f). \tag{2.3}$$

ERM is equivalent to minimizing the expectation of the loss function with respect to an empirical distribution $P_{emp}(x, y)$ formed by assembling δ functions located on each example (CHAPELLE et al., 2001):

$$dP_{emp}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)\delta_{y_i}(y). \tag{2.4}$$

A natural extension to the ERM framework is to replace the delta functions $\delta_{x_i}(x)$ by some estimate of the density in the vicinity of the point x_i, $P_{x_i}(x)$:

$$dP_{est}(x, y) = \frac{1}{n}\sum_{i=1}^{n} dP_{x_i}(x)\delta_{y_i}(y). \tag{2.5}$$

This more general approach leads us to the definition of *vicinal risk* function, defined as

$$R_{vic}(f) = \int l(f(x), y)dP_{est}(x, y) = \frac{1}{n}\sum_{i=1}^{n} \int l(f(x), y_i)dP_{x_i}(x). \tag{2.6}$$

Consequently, similar to the reasoning involving ERM, solving the learning problem with *Vicinal Risk Minimization* consists of estimating (CHAPELLE et al., 2001):

$$\arg\min_{f\in\mathcal{F}} R_{vic}(f). \tag{2.7}$$

Minimizing empirical risk can lead to overfitting. To avoid this, one can add a complexity penalty $C(f)$ term to the risk function we want to minimize. This process is known as regularization and the modified risk minimization process is the Regularized Risk Minimization principle (MURPHY, 2013):

$$R_{reg}(f) = R_{emp} + \lambda C(f). \tag{2.8}$$

where λ controls the strength of the complexity penalty.

One of the most important conclusions drawn from Statistical Learning Theory is synthesized by the following Theorem (MURPHY, 2013), valid for finite hypothesis spaces $\mathcal{H}$:

Theorem 2.1.1. *For any distribution P, and any dataset $\mathcal{D}$ of size N, drawn from P, the probability p that our estimate of the error rate will be more than ϵ wrong, in the worst case, is upper bounded as follows:*

$$p(\max_{h\in\mathcal{H}} \mid R_{emp}(\mathcal{D}, h) - R(P, h) \mid > \epsilon) \leq 2\ dim(\mathcal{H})\ e^{-2N\epsilon^2}, \tag{2.9}$$

where $dim(\mathcal{H})$ is the dimension of the hypothesis space $\mathcal{H}$.

In other words, the gap between training and generalization error is bounded from above by a quantity that grows with model capacity and decreases as the number of training examples increases (GOODFELLOW; BENGIO; COURVILLE, 2016). That is, the more observations are available at training time, the most the gap between training and generalization error is closed. This theorem leads naturally to the following question: is there any way of obtaining more observations without incurring into the high costs or limitations associated with collecting additional real world data? This is precisely the purpose of DA methods, whose theoretical background we will address in the following section.

2.2 Data Augmentation

In various ML applications it is known that model output should not change when some set of transformations are applied to the input variables (BISHOP, 2006). This property, called **invariance**, is generally task-specific. As an example, handwritten digit classifiers predictions should be invariant to small rotations of the inputs but not to reflection.

When presented with enough data – including numerous inputs-output pairs subjected to these invariant transformations – it is possible for the model itself to learn these invariances. However, considering that the number of combinations grows exponentially with the number of transformations, this is rarely the case due to limited data availability (BISHOP, 2006). One can therefore induce the model to exhibit the necessary invariances without the need of collecting additional training data. There are four main approaches to this end:

1. Adding to the error function a regularization term in order to penalize changes in model output when the input is modified;
2. Building and using features that do not change when the required transformations are applied to the input;
3. Training with models in which invariance properties are naturally present due to its architecture (e.g. Convolutional Neural Networks (CNN), with its mechanisms like weight sharing, is able to incorporate some types of invariance);
4. Finally, the training set can be augmented using artificially created inputs transformed according to the desired invariances.

The last option is exactly the definition of a DA process. Despite being a subject whose theoretical framework has not been yet fully developed or unified, there has been growing interest in understanding the theoretical foundations for DA based on the principles of Statistical Learning. We will briefly comment some of the formal mathematical perspectives used so far, concentrating in the results obtained in (DAO et al., 2019) and (CHEN; DOBRIBAN; LEE, 2019).

In (DAO et al., 2019) the DA process is first modeled as a Markov Chain process. Under this approach, the authors show that Kernel classifiers appear naturally, even when dealing with other types of classifiers. Therefore, the authors investigate the effect of DA in Kernel classifiers using the ERM principle. Under the hypothesis that the applied augmentations are local – and thus not significantly modifying the feature map – and using first- and second-order Taylor approximation to the ERM principle, they draw two main conclusions: the first-order Taylor approximation of the ERM is equivalent to training the model on the average feature of all the transformed versions of the inputs; the second-order approximation is equivalent to applying a data-dependent regularization term to the objective. In conclusion, the authors show that DA methods have two effects: (a) increasing invariance by averaging the features of augmented inputs, and (b) penalizing model complexity via a regularization term based on data variance (DAO et al., 2019).

The DA theoretical framework developed by Chen, Dobriban and Lee (2019) arrives in similar conclusions and goes even further. Studying DA in a group-theoretical formulation, the authors show how DA leads to sample efficient learning and explain connections to other concepts of ML and Statistics such as sufficiency, equivariance, regularization, among others.

Despite not being an entirely closed subject, with still several open questions, the theoretical foundations of DA have been an active area of research with important contributions in recent years. However, the fact that not every aspect of DA has been explained by statistical learning has not hindered the development of multiple techniques and applications of DA. In the next chapter we provide an overview of the NLP DA research landscape, presenting the main methods and applications and closing with a summary of the major research gaps.

3 DATA AUGMENTATION RESEARCH LANDSCAPE

In this section we present a literature review of DA methods, highlighting several recent researches that relate to the present work. We start exploring a variety of existing DA methods, taking special care to always consider and examine works that applied these methods to NLP problems. We close in Section 3.7 summarizing the main research gaps found in the literature.

The following sections are influenced by Shorten and Khoshgoftaar (2019), from whom we adapt their image DA taxonomy to explore augmentation methods in general. Since research is often developed by borrowing concepts and techniques from different areas and applying them to the research topic of interest, we will not restrain our DA methods summary to the ones applied strictly to NLP tasks. Rather, we will examine the subject from a more general perspective, trying to relate research on various topics (image processing, NLP, sound processing, etc) within an unique general taxonomy.

3.1 Basic Manipulations

When applied to image processing tasks, **basic manipulations methods for DA** refer to subjecting original images to label-preserving geometric transformations and other image manipulations. This class of techniques include manipulations such as flipping, color space modifications, cropping, rotation, translation, noise injection, among others.

In the realm of NLP problems, basic manipulations usually involve simple treatment of original text such as deletion or swapping of words or characters. Word replacements could take form based on a thesaurus, which could guide word replacements based on semantic closeness to frequently seen meaning. In (ZHANG; ZHAO; LECUN, 2015), e.g., a thesaurus obtained from the WordNet (FELLBAUM, 2005) is used to replace words according to a probability distribution, augmenting datasets used for text classification tasks. The work of Wang and Yang (2015), on the other hand, replaces words by their k-nearest-neighbors embeddings (such as the popular word2vec Word Embedding (MIKOLOV et al., 2013)), achieving with this strategy statistically significant superior performance on categorization models on social media text.

Easy Data Augmentation (EDA) is yet another example of simple manipulation techniques, which combines several manipulations in a single unified method (WEI; ZOU, 2019). This method consists of applying a set of simple operations to the original text in order to generate new synthetic texts. For any given sentence in the training set, one of the following operations is randomly chosen and performed:

1. **Synonym Replacement (SR):** randomly chosen words are replaced by one of its synonyms (also chosen at random) from the WordNet dictionary. Stop words are not considered in this operation.

2. **Random Insertion (RI):** a random word from the given sentence is chosen (stop words are not considered). A random synonym of this word is then inserted in a randomly defined position in the sentence.

3. **Random Swap (RS):** randomly chosen words are defined and their respective positions are swaped.

4. **Random Deletion (RD):** randomly remove words from the sentence, following some probability parameter.

The operations are all randomly applied according to the parameter α, which controls the percentage of words changed in any given sentence. Though maybe original in its proposal as a pure DA technique, the use of this set of operations closely resembles the noise injection procedure proposed by (EDUNOV et al., 2018) as an auxiliary task to improve Neural Machine Translation (NMT) models.

More closely inspired by two common DA manipulations used in image processing, the work of Şahin and Steedman (2018) proposes NLP-equivalents of cropping (focusing on a particular item) and rotating. Using dependency trees, the authors crop – forming smaller sentences from the original ones – and rotate/swap words around some defined root to inflate data for training part-of-speech tagger models. Despite directly inspired by successful techniques in image processing, these NLP-equivalent manipulations did not perform so well in text data, bringing negligible gains in most of the tested scenarios.

Basic manipulations DA methods are generally characterized by easy of implementation, since there is almost no change at all to the way the problem is postulated and solved besides the simple modifications performed in training data.

3.2 Adversarial Training

Since the seminal paper of Szegedy et al. (2014) showing the counter-intuitive vulnerability of state-of-the-art Neural Networks (NN) to small perturbations in its inputs, the field of *adversarial attack* has been an active area of research, especially within the Deep Learning community. Adversarial examples, synthesized data generated by adding imperceptibly small perturbations in original examples for the sole purpose of misleading trained models, raised the question of the generalization ability of NNs. Following works, such as (GOODFELLOW; SHLENS; SZEGEDY, 2015), helped deepen our understanding on how adversarial examples affect both linear models and NNs. More than that, the authors leveraged the knowledge of the adversarial flaw as an opportunity to fix it, giving rise to *adversarial training*.

Adversarial training consists in the process of augmenting original data with adversarial examples and training the model on this augmented dataset, increasing model's robustness. Adversarial training differs from traditional DA techniques because, whereas traditional methods synthesize data with transformations that are expected to occur in the test set, adversarial training uses inputs unlikely to occur (GOODFELLOW; SHLENS; SZEGEDY, 2015). Instead, this technique uses adversarial examples to expose and correct the flaws in model decision functions and provide additional regularization benefits beyond what is achieved by using dropout exclusively.

Originally created and developed within the computer vision community, adversarial attacks and adversarial training methods elaborated for image processing tasks can not be directly applied to NLP problems. This is due to three main differences between these areas (ZHANG et al., 2020). First, the discrete nature of text, in contrast to the continuous nature of images, poses additional challenges to adversarial training, such as defining appropriate distance measurements and perturbations to text data. Second, small changes applied to images (i.e., perturbations applied to pixels) usually can not be easily identified by humans, despite being able to fool NN models. The same is not true for text data, where small modifications to the original text (such as character or word change) are easily identifiable (i.e., may generate unfluent text). This makes it more difficult to find unperceivable textual adversaries, which hinders adversarial training since fluent adversarial examples are more effective than unfluent ones (ZHANG et al., 2019). Finally, different to the case of text data, small perturbations to images usually do not change the semantics of the original data. On the other hand, in NLP even small changes to the original data can affect the semantics of it, which is against the goal of adversarial attack, that has the objective to fool the NN without altering the ground truth of the input.

Tackling these additional challenges, several works have been developed to apply adversarial training to NLP models. The work by Zhang et al. (2019) proposes MHA, Metropolis-Hastings attack, an adversarial example generator based on Markov Chain Monte Carlo (MCMC) sampling. Experimental results obtained by the authors in two benchmark datasets show that adversarial examples generated by MHA are more fluent and more effective in improving adversarial robustness and classification accuracy after adversarial training. As another example, Morris et al. (2020) reviews several works on the topic to propose a unified framework of adversarial attacks on NLP divided into four main categories: Search Method, Transformation, Goal Function and Constraints. After categorizing relevant research within these categories, the authors then implement a Python software library to implement several NLP adversarial attacks from the literature. Performing the first review of its kind, in (ZHANG et al., 2020) the authors conduct a systematic survey of adversarial attacks research applied to NLP problems, proposing a classification scheme to organize the reviewed literature. By covering all relevant literature since 2017, when the first work applying adversarial training to NLP was published, the authors are able not only to organize state-of-the-art knowledge on the subject, but also to discuss open issues and promising research paths on the field.

3.3 GAN-based Data Augmentation

Generative Adversarial Nets (GANs) consists of a framework of coupled NNs for estimating generative models first proposed by Goodfellow et al. (2014). Since its inception, GAN-related research has been a rapidly evolving investigation topic, with impressive progress being made in training and applications. Several architecture variations have been developed upon the original, such as Conditional GANs (CGAN), Deep Convolutional GANs (DCGAN), Wasserstein GANs (WGAN) and Big GANs (BigGAN), just to name a few (GONOG; ZHOU, 2019; PAN et al., 2019).

Figure 1 shows the basic functioning of any GAN architecture. Inspired by game theory, GANs are composed of a generator network G and a discriminator network D that compete which each other. The goal of the generator is to generate artificial data as close as possible to the distribution of real data. The discriminator networks' goal, on the other hand is, when presented with real and fake data (generated by G), to be able to distinguish one from another.

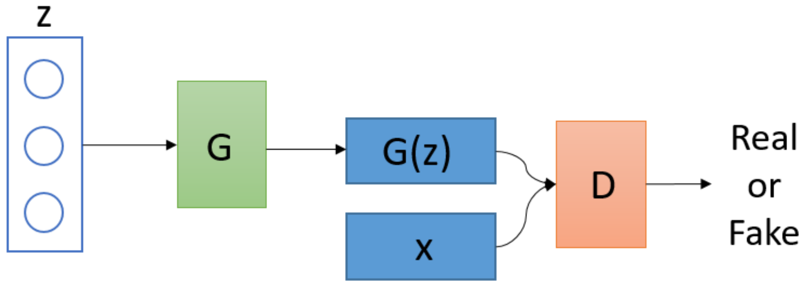


Figure 1: General architecture of a GAN: the NN generator G receives as input a random noise vector z (usually a uniform or normal distribution) and outputs artificial data G(z). The NN generator G is trained so as to output data as close as possible to the original data distribution X, while the NN discriminator D has the task to identify whether G(z) is real or fake data. Adapted from (PAN et al., 2019).

Due to its capability of generating highly realistic synthetic data, GANs have been applied as yet another method of DA in numerous and diverse research topics. In Image processing tasks, augmentation-oriented GANs have been employed, for example, in medical image analysis to generate synthetic skin lesion or liver cancer images (SHORTEN; KHOSHGOFTAAR, 2019). The applications, however, are not constraint to image processing tasks. In (GUAN et al., 2018) the authors develop and propose Medical Text Generative Adversarial Netowrk, a GAN-based framework for creating artificial electronic medical records. After testing this method with a Chinese electronic medical record dataset, the authors show that the synthetic dataset generated presents similar properties to real electronic medical record texts and could therefore be used within a DA approach.

Switching from medical applications to cyber security, in (MIMURA, 2020) authors use GANs to generate fake malware samples for use in models of malware detection. As the experiments show, this approach was able to improve detection accuracy by oversampling the minority class. Finally, citing another application, more related to the applications of this work, in (GUPTA, 2019) the authors used a GAN architecture for solving a Sentiment Analysis (SA) task and synthesizing data on a low resource scenario. Using this approach, the authors observe that training with the GAN-augmented data improves the SA classification model. Furthermore, projecting the GAN-generated data in a lower 2-dimensional space using t-distributed stochastic neighbor embedding (t-SNE) showed that, despite overlapping with original data, artificially created data did not occupy the entire region of the feature space of original data.

3.4 Meta Learning

A great number of DA strategies rely on manually specifying the necessary transformations that generate artificial label-preserving data. These type of procedures have several drawbacks. First, one often has to have significant domain knowledge of the dataset to be able to choose the appropriate transformations, relying purely on experience and intuition. Second, the trial and error process of exploring different augmentation techniques frequently result in prohibitive computing resources or timing, specially when using Deep Learning models. Another unwanted characteristic of handcrafted DA transformations is not being universal. As a consequence, successful data modifications can not be easily applied to other datasets with similar performance improvements. Meta Learning DA techniques consist of a set of various DA methods that try to learn from training data the most appropriate transformations so as to achieve better generalization power.

The work of Hauberg et al. (2016) aligns $\langle source - destination \rangle$ pairs of images of the same class to obtain the transformation that maps one to the other. This transformation is mathematically expressed as a diffeomorphism and is later applied to other images to generate synthetic data. As a consequence, the necessary transformations are obtained from the training data itself.

Smart Augmentation (LEMLEY; BAZRAFKAN; CORCORAN, 2017) is another form of Meta Learning DA. In this technique, a NN is trained to generate augmented data that yields the best performance on the final target NN by minimizing its loss. The augmentation is achieved by blending pairs of images in ways that improve regularization.

In (CUBUK et al., 2019) the authors propose AutoAugment, another strategy to automate the process of generating synthetic data. In this work policies express augmentation operations and Reinforcement Learning is used to find the transformations that yield the best validation accuracy on a target NN. The work of Niu and Bansal (2019) adapts AutoAugment – whose basic architecture is depicted in Fig. 2 – to a dialogue generation setting, and choose operations such as Stopword Dropout and Grammar Errors as the policy search space.

Finally, in (HO et al., 2019) the authors propose Population Based Augmentation, whose goal is to learn a schedule of augmentation policies, instead of a fixed policy as in AutoAugment. This choice is accountable to the far superior computational efficiency of this method when compared to AutoAugment.

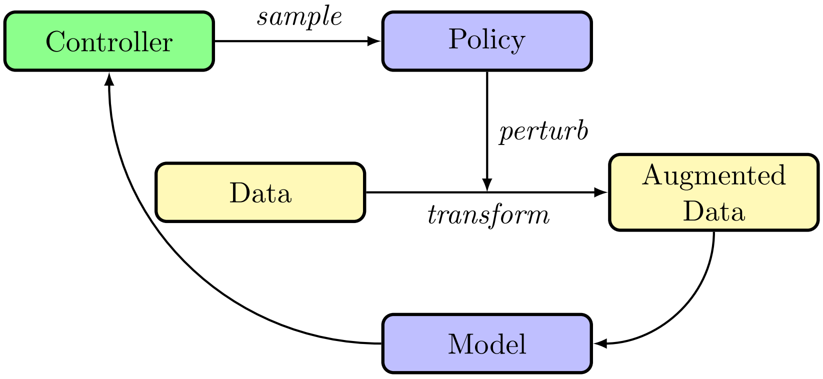


Figure 2: System design of AutoAugment, a technique for generating data-augmented examples using Reinforcement Learning. A Controller samples a Policy to perturb original Data, generating Augmented Data. The artificially generated data is used for training the model, whose performance is fed back to the Controller. Adapted from (NIU; BANSAL, 2019).

3.5 Mixed Sample Data Augmentation (MSDA)

A very recent and promising class of DA methods are mixed sample – also known as mixed-example – techniques. These constitute a class of methods that artificially generate data by combining pairs of inputs drawn from the original set. They can be considered a more general approach to DA, since they are not necessarily label-preserving processes. The first general techniques for augmenting through linear combination of examples are *mixup* (ZHANG et al., 2018) and Between-Class (BC) learning (TOKOZUME; USHIKU; HARADA, 2018b) – which are equivalent and were developed in parallel – along with the improved "BC+" learning method (TOKOZUME; USHIKU; HARADA, 2018a).

In essence, linearity-based methods such as *mixup* are based on Vicinity Risk Minimization and consist of using a specific type of vicinal distribution. Sampling from this distribution produces feature-target vectors

$$\begin{cases} \tilde{x} = \lambda x_i + (1-\lambda)x_j & \text{where } x_i,\ x_j \text{ are raw input vectors} \\ \tilde{y} = \lambda y_i + (1-\lambda)y_j & \text{where } y_i,\ y_j \text{ are one-hot label encodings,} \end{cases} \tag{3.1}$$

where (x_i, y_i) and (x_j, y_j) are feature-target vectors randomly drawn from the original dataset and $\lambda \in [0, 1]$. As shown in Fig. 3, this technique leads to decision boundaries with linear transitioning from class to class. When the *mixup*-hyperparameter α, which controls the strength of interpolation, goes to zero, traditional ERM (Equation 2.3) is recovered.

FMix is another mixed-example technique which uses binary masks obtained by ap-

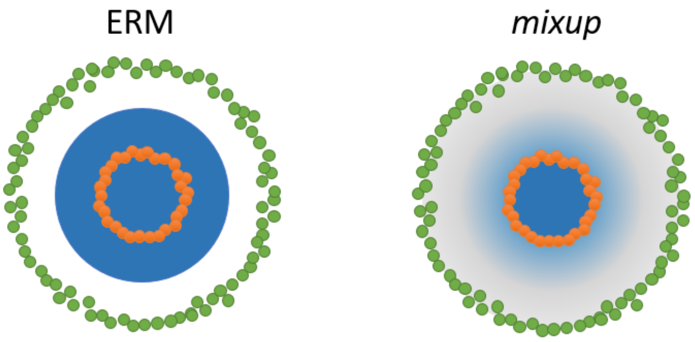


Figure 3: Comparison of using ERM and *mixup* on a toy problem. Green and orange represent classes 0 and 1, respectively. Blue shading indicates the result of $p(y = 1|x)$ using ERM and *mixup*. *Mixup* encourages the model to behave linearly in-between training examples. In this example, *mixup* used $\alpha = 1$. Adapted from (ZHANG et al., 2018).

plying a threshold to low frequency images sampled from Fourier space. First proposed in (HARRIS et al., 2020), the authors test its efficacy in image classification, sound classification and Sentiment Analysis (SA) tasks. The results obtained in several SA tasks show *FMix* and *mixup* obtaining gains compared to the baseline, but there is no clear winner between these DA techniques: their results are in most cases very similar.

More recently, MSDA strategies that combine inputs and targets in non-linear ways were developed and applied to image processing tasks (SUMMERS; DINNEEN, 2019).

Despite proving very effective in improving generalization error, even after the use of other DA techniques, the exact reason why mixed-example methods are so successful remains as an open research question.

The applicability of MSDA techniques to different types of data – image, sound, text, etc. – varies from method to method. BC learning (TOKOZUME; USHIKU; HARADA, 2018a) and *mixup* (ZHANG et al., 2018), e.g., are broader approaches, employable in various types of data. More general and non-linear mixing methods, such as those proposed by Summers and Dinneen (2019), on the other hand, operate only on images. In spite of the existence of NLP-applicable mixed-example DA methods, this field remains almost untouched. To the best of our knowledge, the works that have been developed exploring this family of DA techniques on text are very few and recent – most of them have been published or are in review process in 2020.

The work of (GUO; MAO; ZHANG, 2019) proposes two variations to *mixup* – wordMixup and senMixup. While in wordMixup word embeddings are interpolated, in senMixup sentences embeddings are interpolated to generate new synthesized data. These methods are applied to five different text datasets and perform better than their own no-DA implementation – although not always better than other authors' no-DA implementations.

Inspired by linguistic principles related to compositionality, Andreas (2020) propose "Good-Enough Compositional Data Augmentation" (or GECA). This method, which can be considered a form of mixed-example DA, consists of generating new training examples by replacing fragments with other that appear in similar environment. The proposed protocol has the advantages of being model-agnostic and useful in a variety of tasks.

The work of (GUO; KIM; RUSH, 2020) proposes SeqMix, a sequence-level variation of *mixup* which performs soft combination of input/output sentences from the training set. The random combination of sentences prevents the model from memorizing long segments and encourage them to rely on compositions of subparts to predict the output. Another relevant contribution of this paper is to provide a theoretical framework that unifies several other DA strategies for compositionality (WordDrop, SwitchOut and GECA).

3.6 Back-Translation-based Methods

Considered crucial to NMT tasks nowadays (GRAÇA et al., 2019), Back-Translation (BT) is another method for generating auxiliary synthetic data. First introduced by Sennrich, Haddow and Birch (2016), the term BT 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 DA method for down-stream tasks seems to be the work of Yu et al. (2018), which used BT 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 NMT tasks, there has been an emergence of numerous variations to the traditional BT method.

Iterative BT (IterativeBT), proposed in (HOANG et al., 2018), 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, although the quality of the BT system in use being crucial to the success of the proposed approach. The authors report improvements in NMT systems that apply IterativeBT both in high-resource as in low-resource scenarios.

Noised BT (NoisedBT)[1], presented by Edunov et al. (2018) 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. The authors also report an intriguing finding when comparing model improvements obtained by adding artificially generated text data with those obtained by adding real data: in cases where the domains match, synthetic data can be nearly as effective as real human translated data.

Tagged BT (TaggedBT) (CASWELL; CHELBA; GRANGIER, 2019), heavily influenced by the works of Edunov et al. (2018) and Imamura, Fujita and Sumita (2018), proposes another hypothesis for the superiority of noise injection in BT postulated in (EDUNOV et al., 2018): instead of increased text diversity, noise injection would instead benefit the final model by signaling which data is synthetic and which is original data. To assess the validity of this assumption, the authors trained a translation model where artificially generated source data had no noise injection and was rather tagged with a reserved token. The fact that this method leaded to similar or even slightly higher performance supported the hypothesis of the authors.

3.7 Research Gaps

Given the above literature review, here we want to highlight some of the still open research questions related to NLP DA methods. We summarize below the most patent research gaps found in the NLP DA literature:

1. **Lack of comparative studies between different classes of NLP DA methods**: when proposing new methods, authors restrain their comparative assessments to the same class of DA methods they are contributing to, failing to contrast their findings to other existing categories of DA techniques. Works proposing variations to the traditional BT method generally compare their results only to other BT-based techniques. Despite BT-based methods being commonplace at the time, the work introducing EDA (WEI; ZOU, 2019) did not contrast EDA results against any BT-based method, comparing it only against a no-DA setting. The same is

[1] The term Noised BT was not used in (EDUNOV et al., 2018), but coined in (CASWELL; CHELBA; GRANGIER, 2019)

true for MSDA methods. In (HARRIS et al., 2020) the authors compare FMix, the proposed technique, to *mixup* and CutMix – EDA and BT-based methods were left without mention. As a consequence, there is no unified overview of how all these NLP DA methods compare to each other in any aspect (quality of synthesized data, performance gain, etc.).

2. **Superficial understanding of BT methods**: despite all of the proposed variations of BT, few researches have investigated how different choices and parameters of BT 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 BT when using dropout remain still open. The studies on DA performed by Yu et al. (2018) seem to be the closest to partially address some of these open issues, though leaving the majority of the questions here proposed still unanswered.

3. **Limitations in BT-based methods**: almost none of the BT variations take advantage of language translation at a greater level for general NLP Tasks. While IterativeBT (HOANG et al., 2018) is only applicable for Machine Translation tasks, TaggedBT (CASWELL; CHELBA; GRANGIER, 2019) and NoisedBT (EDUNOV et al., 2018) just add additional handcrafted information to the data. To the best of our knowledge, DeepBT (FERREIRA; COSTA, 2020) is the only BT-based method that leverages translations one step further when compared to traditional BT.

4. **Qualitative and Quantitative assessment of synthesized data:** contrary to what is usual in image processing, NLP DA works rarely address any topic beyond model performance or computational overhead. Qualitative or quantitative evaluations of the artificially generated data – are they comprehensible to a human, maintaining their original meaning? Are they by any metric similar to the original data? – receive little to no attention.

5. **Insufficient diversity in types of problems tackled with DA methods:** research has been extremely narrowly focused, exploring the same types of problems, namely classification tasks, normally image classification problems. Little is known about the effectiveness of DA methods (specially NLP) within other domains. In the words of the authors who proposed *mixup* (ZHANG et al., 2018), the research question whether it is "*possible to make similar ideas work on other types of supervised learning problems, such as regression and structured prediction*" is rarely addressed, remaining as an open research subject.

6. **Lack of NLP-specific knowledge of DA:** as a consequence of the scarcity of NLP DA studies, we still do not have a clear comprehension of any specific considerations or caveats of DA methods when applied to text data. As an example, we do not have an answer of what is the impact, if any, of different linguistic styles on DA methods' performance.

In summary, to the best of our knowledge, there has not been any research covering the entire NLP DA landscape, comparing different methods and consolidating this knowledge. We start to tackle these research opportunities and yet unanswered questions in the following chapters, proposing DeepBT, a new DA method, followed by a comparative analysis of NLP DA methods in SA tasks. Considering the low adoption of DA techniques in the SemEval series of challenges – as discussed in Chapter 1 –, we decided for performing this comparative analysis of NLP DA methods on SemEval SA tasks, which will be presented in Chapter 5.

4 DEEP BACK-TRANSLATION

DA methods can be evaluated by different perspectives, two of the most important being the validity and the diversity of the data they generate when augmenting the original training set (XIE et al., 2020). Robust DA techniques should generate a diverse set of examples to prevent overfitting. On the other hand, this diversity should not hinder synthetic data validity. As a consequence, DA methods should strive for adequate balance within the diversity-validity trade-off in order to be successful (DAI; ADEL, 2020).

Exploring the diversity-validity trade-off, we propose in (FERREIRA; COSTA, 2020) a new method for DA, named DeepBT, which adds more layers of intermediate translations between the original text and the final paraphrase. 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 BT we always have

$$A \rightarrow L \rightarrow A, \tag{4.1}$$

in DeepBT we could have n intermediate layers of translations:

$$A \rightarrow L_1 \rightarrow L_2 \rightarrow ... \rightarrow L_n \rightarrow A. \tag{4.2}$$

Figure 4 illustrates the difference between traditional BT and a 2-layer DeepBT.

Following the rationale of the BT method, which generates paraphrases with the same meaning and label as the original one, the DeepBT technique is created under the hypothesis that using several intermediate languages between the original and destination one could increase diversity on generated paraphrases without hurting validity. With greater diversity in training data we assume that we could reduce overfitting and achieve greater performance.

We implemented our proposal, DeepBT, alongside EDA and traditional BT. The following chapter contains the details about the experimental setup.

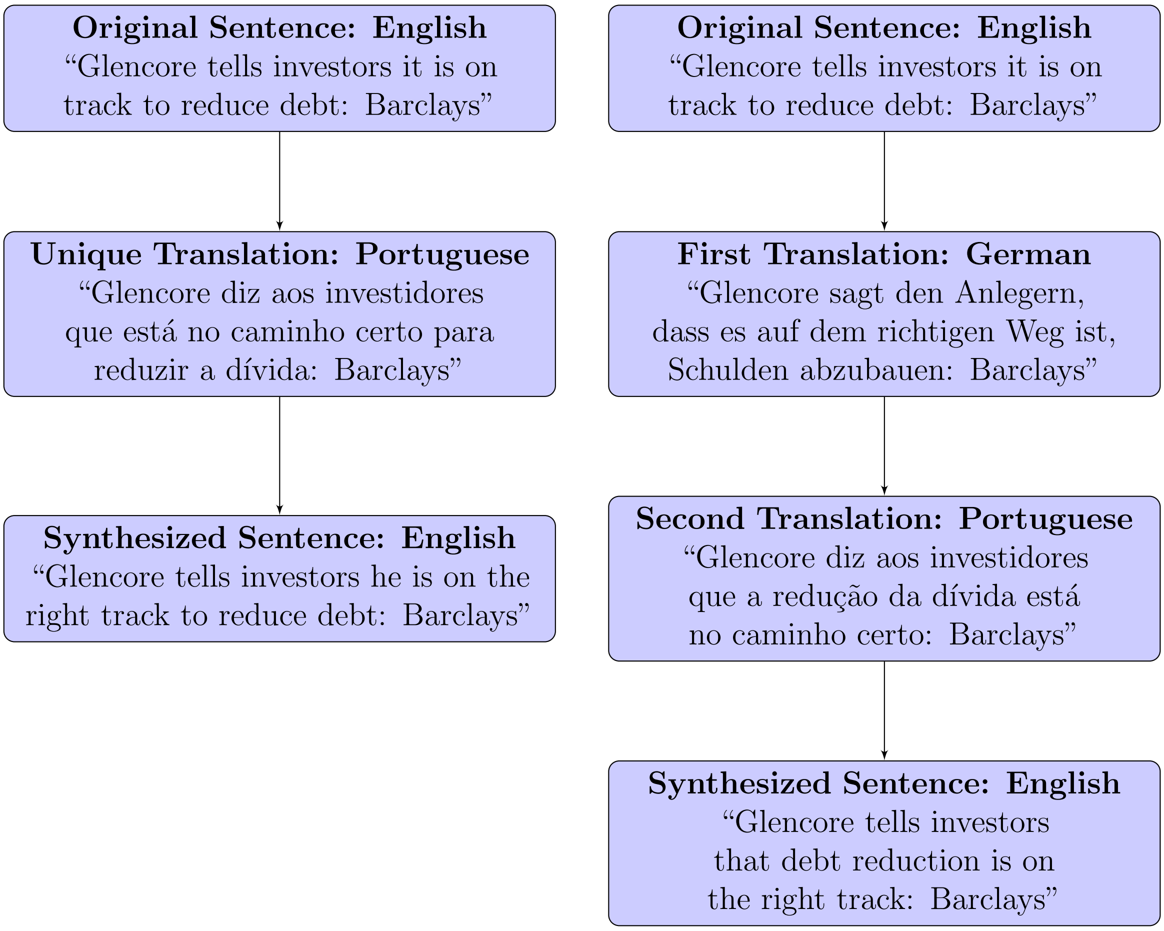

Original Sentence: English
"Glencore tells investors it is on track to reduce debt: Barclays"
Unique Translation: Portuguese
"Glencore diz aos investidores que está no caminho certo para reduzir a dívida: Barclays"
Synthesized Sentence: English
"Glencore tells investors he is on the right track to reduce debt: Barclays"
Original Sentence: English
"Glencore tells investors it is on track to reduce debt: Barclays"
First Translation: German
"Glencore sagt den Anlegern, dass es auf dem richtigen Weg ist, Schulden abzubauen: Barclays"
Second Translation: Portuguese
"Glencore diz aos investidores que a redução da dívida está no caminho certo: Barclays"
Synthesized Sentence: English
"Glencore tells investors that debt reduction is on the right track: Barclays"

Figure 4: Graphic representation of Deep- and traditional BT procedures. In DeepBT (right flowchart) we stack several intermediate layers of translation, whereas in traditional BT (left flowchart) there is always only one translation between original and synthesized text. As in this example, the additional translation layer present in DeepBT is able to generate synthesized text different from the text generated by traditional BT.

5 COMPARATIVE ANALYSIS OF NLP DATA AUGMENTATION TECHNIQUES

We detail below the comparative analysis performed between NLP DA methods, introducing benchmark datasets used, DA and data preparation procedures applied, while also presenting ML architecture characteristics. We also summarize steps and methodology involved in obtaining the results presented thereafter.

In order to evaluate how different DA methods compare to each other, we selected three techniques, representing two of the most prevalent NLP DA family of methods: traditional BT and DeepBT (representing BT-based methods) and EDA (representing simple manipulations methods). These three DA techniques were applied to the same benchmark datasets, using the same data preparation procedure and matching set of hyperparameters. By choosing these methods, we cover the families of NLP DA techniques that were more largely and successfully employed, but that have never been compared against each other. Furthermore, another compelling reason for choosing these methods relies in their broad accessibility: the code for EDA was made publicly available by the authors themselves and the use of Google Translation API can be easily implemented by any researcher – and is free, up to a certain amount of calls. Hence, by choosing these DA techniques, we assure best comparability by exploring methods whose algorithms or implementation are easily available. Finally, the chosen DA methods comprise a set of relatively simple techniques, with very few parameters to choose from, enhancing once again reproducibility.

5.1 Application Domain and Benchmark Datasets

As the application domain for this comparative analysis, we chose one of the most prolific areas of study in NLP: Sentiment Analysis. SA can be defined as any effort in detecting, analyzing and evaluating humans' opinions or emotions towards events, issues or any other topics (YADOLLAHI; SHAHRAKI; ZAÏANE, 2017). This domain is not only compelling because of the growing research interest, but also because of the wide range of applications: they span from consumer industry to public opinion assessment (Liu et al., 2019), from financial market prediction to politics and disaster management (YADAV; VISHWAKARMA, 2020).

In order to define the most appropriate tasks and datasets within the SA domain for this work, we prioritized addressing the lack of diversity in the types of problems tackled with DA methods (as explained in subsection 3.7). As a result, we avoided traditional classification tasks, but decided to tackle a regression task instead.

When selecting the specific dataset to use, we opted for looking for tasks in a long-established and accepted series of computational linguistics evaluations and challenges, namely the SemEval (International Workshop on Semantic Evaluation) (SemEval. . . , 2021). The 2017 edition of SemEval provided the exact tasks and annotated datasets for the objectives we were aiming at.

The becnhmark datasets used for experiments are the ones provided by the SemEval 2017 Task 5 challenge (CORTIS et al., 2017), in Tracks 1 and 2. Both of the datasets consist of NLP regression tasks within the domain of SA. The dataset of Track 1 (Microblog Messages) contains messages related to stock market events collected in two different microblog platforms (StockTwits and Twitter). The dataset of 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.

5.2 Data Augmentation and Preparation Procedure

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

Considering the high effort and computational resources required for training a translation language model – and since the focus of this work is not related to the translation model itself – we opted for using the Google Cloud Translation API[1] for the translations required by the BT-based methods. This allows for rapid and high quality translation for several different languages with minimal associated costs.

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

We wanted to test the hypothesis that the diversity of language families used as pivotal-language in BT-based methods could bring greater diversity to paraphrases generated by these techniques. In order to test this conjecture, we chose as pivotal languages from different families, namely German (West Germanic), Russian (East Slavic) and Portuguese (Western Romance), for both Deep- and traditional BT procedures. For assessing the DeepBT method the experiments were carried out with 2-layer translation settings (e.g. $English \rightarrow Russian \rightarrow German \rightarrow English$).

Recalling the explanation of the EDA method made in Section 3, this technique randomly chooses and performs four different operations on any given sentence of the training set: Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS) and Random Deletion (RD). This method accepts two parameters to control its DA process, namely α and n_{aug}. While the first controls the percentage of words in a sentence that are changed, the latter is responsible for indicating the number of augmented sentences in the output. In the seminal paper of Wei and Zou (2019), 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 (WEI; ZOU, 2019).

All implemented models shared the same data preparation process, which consisted of standard NLP procedures: stop words removal, tokenization and lowercasing. Furthermore, in each sentence, the company or cashtag referred to were replaced with a generic token. Tables 1 and 2 show some examples of original sentences and final sentences after these standard NLP preprocessing procedures were implemented.

Table 1: Comparison between original sentences and prepared sentences after preprocessing steps in Subtask 1 dataset.

Cashtag	**Original Sentence**	**Prepared Sentence**
$FB	“watching for bounce tomorrow”	“watching bounce tomorrow”
$AAPL	“now seems like its helping the downtrend”	“seems like helping downtrend”
$SPY	“trade continuing very nicely from yesterday. looking very strong here	“trade continuing nicely yesterday . looking strong”
$CHK	“reserves are in decline”	“reserves decline”

Datasets of Subtask 1 and Subtask 2 had originally 1.678 and 1.143 records, respectively. Training and test sets were divided in a 80% / 20% split. The fact that we used original and synthetic data for training and testing required special attention as to how

Table 2: Comparison between original sentences and prepared sentences after preprocessing steps in Subtask 2 dataset.

Company	**Original Sentence**	**Prepared Sentence**
Morrisons	“Morrisons book second consecutive quarter of sales growth”	“company book second consecutive quarter sales growth”
Barclays	“Barclays share price subdued as bank faces fresh forex probe”	“company share price subdued bank faces fresh forex probe”
BP	“UPDATE 1-BP reports worst annual loss in at least 20 years, cuts more jobs”	“update 1-company reports worst annual loss least 20 years , cuts jobs”
Shire	“Drugmaker Shire to buy Baxalta for $32 billion after 6-month pursuit”	“drugmaker company buy baxalta $ 32 billion 6-month pursuit”

Table 3: Final number of records used for training and testing in each dataset controlling by original dataset percentage use and multiplication factor.

Percentage	**Subtask 1**			**Subtask 2**		
	Multipl. Factor			**Multipl. Factor**		
	2	**3**	**4**	**2**	**3**	**4**
10%	336	504	672	228	342	456
20%	672	1.008	1.344	458	687	916
30%	1.006	1.509	2.012	686	1.029	1.372
40%	1.342	2.013	2.684	914	1.371	1.828
50%	1.678	2.517	3.356	1.144	1.716	2.288
60%	2.014	3.021	4.028	1.372	2.058	2.744
70%	2.350	3.525	4.700	1.600	2.400	3.200
80%	2.684	4.026	5.368	1.828	2.742	3.656
90%	3.020	4.530	6.040	2.058	3.087	4.116
100%	3.356	5.034	6.712	2.286	3.429	4.572

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 test set would entail the occurrence of data leakage. As a result, splitting data procedure was carefully designed so as to have original and synthesized sentences always together at the training or at the test sets.

5.3 Model Architecture

In order to choose the more appropriate architecture, we analyzed the choices made by the participants of the SemEval 2017 Task 5 challenge. In Subtask 1 (Microblog Messages), 3 of the 6 best scoring teams used CNNs as part of their architecture. In Subtask 2 (News Statements & Headlines), 2 of the 3 best scoring papers used CNNs either as their sole technique or as part of an Ensemble of Methods. Thus, following some of the best performing teams in SemEval 2017 Task 5 challenge (GHOSAL et al., 2017;

KAR; MAHARJAN; SOLORIO, 2017; MANSAR et al., 2017; CORTIS et al., 2017), we based our experiments in a CNN.

Choices regarding architecture, hyperparameters and feature representation were done inspired by state-of-the-art models in SA tasks (ZHANG; WALLACE, 2017) and the winning architectures of the SemEval 2017 Task 5 challenge (CORTIS et al., 2017). 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 (PENNINGTON; SOCHER; MANNING, 2014). To avoid drawing conclusions that are specific to some arbitrary chosen hyperparameter values and seed (FERREIRA et al., 2019), 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).

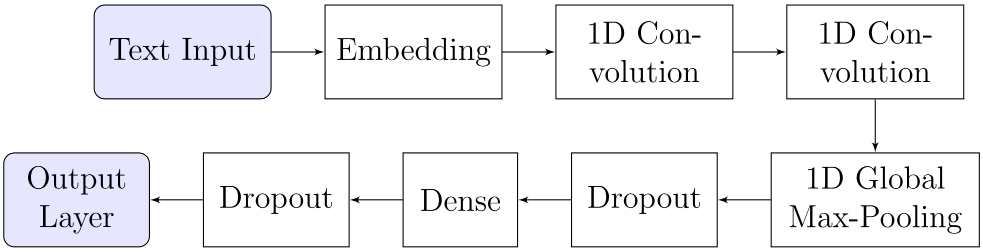


Figure 5: CNN architecture used in our experiments: 2 convolutional layers followed by single dense layer.

6 RESULTS

We begin showing the result of each method in the generation of artificial text. Next we follow presenting and discussing the results related to some of the research gaps exhibited in Section 3.7.

6.1 Synthesized Data

Before diving into model results obtained by using each of the DA techniques, it is interesting to analyze synthesized data outputs of each of them so as to gain better insights into final model outcomes. Some examples are shown in Tables 4 and 5 for the Microblog Messages and for the News Statements & Headlines respectively.

Table 4: Synthesized Data generated by different DA techniques in the Microblog Messages Dataset.

Original Sentence	**EDA**	**BT**	**DeepBT**
"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"	"# OwnItDon'tTradeIt"	"# OwnItDon'tTradeIt"
"absolute garbage still up stores TOTALLY EMPTY stock mispriced"	"stock garbage still up stores totally empty absolute mispriced"	"absolute garbage still. Stores TOTALLY EMPTY. stocks priced incorrectly"	"absolute trash is still on. Stores TOTALLY EMPTY. Stock is underestimated"
"possible double bottom set up"	"possible double set up"	"possible raised floor construction"	"possible double bottom"

We observe that, due to the random noise injected by EDA (random swap, random deletion, i.e., noise that generally do not follow linguistic rules), sentences generated by this method normally suffer from lack of correct grammatical or syntactical structure. Both BT methods, on the other hand, generally yield sentences with correct grammatical structure – though not necessarily preserving original semantics.

When analyzing BT and DeepBT, we observe that the original meaning of the sentence is captured in varying degrees. In *"I look forward to jumping tomorrow"* (Table 4), as well

Table 5: Synthesized Data generated by different DA techniques in the News Statements & Headlines Dataset.

Original Sentence	**EDA**	**BT**	**DeepBT**
"Tesco says recovery on track, asks investors to be patient"	"tesco says recovery on track be investors to asks patient"	"Tesco says recovery on track and asks investors to be patient"	"Tesco says recovery is on track and urges investors to be patient"
"Glencore tells investors it is on track to reduce debt: Barclays"	"glencore tells to it is on track investors reduce debt barclays"	"Glencore tells investors it's on track to cut debt - Barclays"	"Glencore Tells Investors Debt Reduction Is On The Right Path: Barclays"
"Horizonte acquires neighbouring Glencore nickel property in Brazil"	"acquires neighbouring glencore nickel property in brazil"	"Horizons acquires the neighboring Glencore nickel property in Brazil"	"Horizonte acquires Glencore's nickel properties in Brazil"
"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"

as in *"British FTSE filmed by solid kingfishers"* (Table 5), the connotation of the original sentence was clearly lost. We noticed however, that this lost of meaning is often associated to the pivotal language used or to the translation to this specific language – not necessarily to the DA method itself. Let us analyze the sentences generated from *"possible double bottom set up"* in Table 4. In BT, when using German as the pivotal language, we arrived at *"possible raised floor construction"* as the final sentence, also clearly loosing original meaning. When using Russian or Portuguese as pivotal languages however, we arrived at *"double bottom possible"* and *"possible double bottom configuration"* respectively, much closer to original text. Similar result is observed when applying DeepBT to this same sentence. When using Russian and Portuguese as the 2-layer pivotal languages we arrive at *"possible double bottom"*, whereas using German and Russian synthesized *"raised floor construction possible"*, once again far from original connotation. That is, for this specific example, BT and DeepBT both generated reasonable outputs when using only Portuguese or Russian (or both) as pivotal languages, but in both methods the synthesized sentence when using German (alone or in conjunction with other languages) was very different from the original text – indicating possible quality issues with the translations from or to German.

Other interesting insights arise when analyzing the output of each method regarding informal texts (Microblog Messages). In the second example of the Microblog Messages Dataset we can see that, while EDA maintained the original term *biz*, both of the BT methods were able to identify this expression and output it as *business*. That is, the translation API was able to identify the true meaning of this slang and perform correct translations to pivotal language and back to the original language. 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*). In general, when analysing several synthesized texts in both datasets, we observe that – due to occasional existence of expressions not recognizable by a dictionary or translation API – the quality of texts generated by DA methods in the Microblog Messages dataset varied greatly depending on the input.

Additionally to a completely qualitative analysis, it is also valuable to evaluate synthesized data from a quantitative perspective. This inspection could prove insightful for better understanding performance output of each method or configuration.

In order to assess how original and synthesized sentences are similar to each other, we quantified the similarity between those sentences using *cosine similarity*, a standard metric that has been widely applied in web search engines and general information retrieval tasks to compare words or documents against each other (MANNING, 2008). Given two vector representations $\overrightarrow{V}(d_1)$ and $\overrightarrow{V}(d_2)$ of documents d_1 and d_2, respectively, the cosine similarity $sim(d_1, d_2)$ is given by

$$sim(d_1, d_2) = \frac{\overrightarrow{V}(d_1) \cdot \overrightarrow{V}(d_2)}{|\overrightarrow{V}(d_1)||\overrightarrow{V}(d_2)|}. \tag{6.1}$$

In order to represent each record of the dataset (usually one or two sentences) as a single vector, we used another technique that has long been used: the record is represented as a vector obtained by element-wise averaging each of the components of the Word Embeddings of each word. That is, if the record is composed of a sentence with 5 words, this record will be represented as a single 300-dimensional vector obtained by element-wise averaging each of the 5 300-dimensional GloVe Word Embeddings that represent each word.

Table 6 displays average cosine similarity, along with standard deviation, obtained when comparing original and synthesized data in each of the datasets and with each of the DA techniques. For a complementary analysis, Figure 6 shows the Boxplot of cosine similarity when applying each of the DA methods to both datasets.

We discuss and highlight some of the most interesting results. First, we observe relevant differences between the distribution of the similarity between original and synthesized data obtained in the Microblog Messages dataset in comparison with the News Statements & Headlines dataset. Besides presenting slightly inferior average similarity,

Table 6: Average cosine similarity and standard deviation between original and synthesized text for each DA method and dataset.

Dataset	**BT**	**DeepBT**	**EDA**
Microblog Messages	0.88 ± 0.17	0.85 ± 0.18	0.86 ± 0.17
News Statements & Headlines	0.93 ± 0.07	0.90 ± 0.08	0.88 ± 0.09

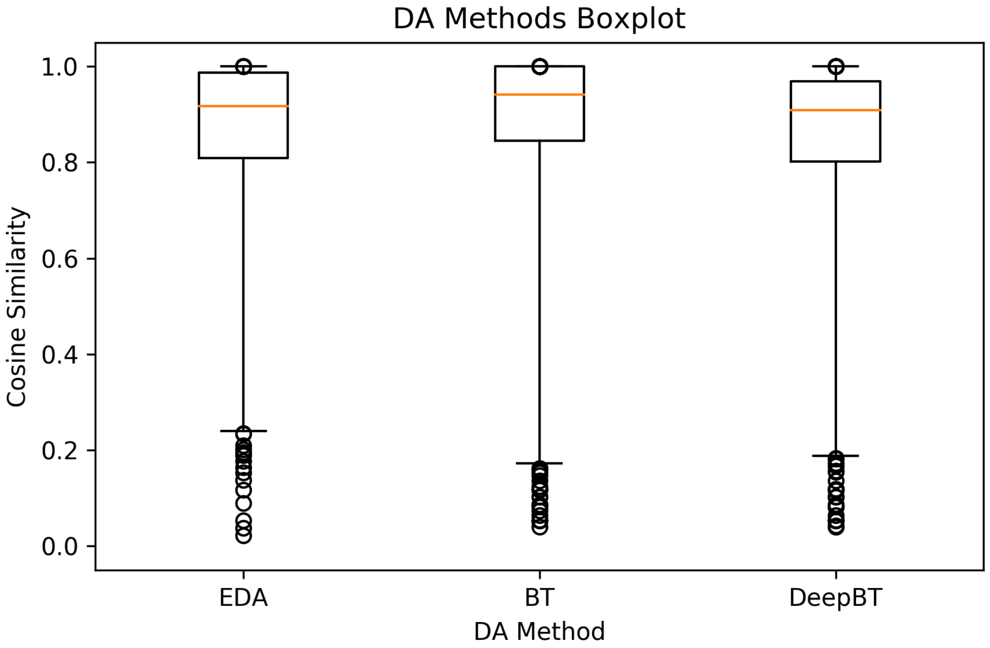
Table 7: Average Cosine Similarity and Standard Deviation between original and synthesized text when using Back-Translation as DA method for different Pivot Languages.

Dataset	**PT**	**DE**	**RU**
Microblog Messages	0.88 ± 0.17	0.90 ± 0.15	0.87 ± 0.17
News Statements & Headlines	0.93 ± 0.06	0.93 ± 0.06	0.91 ± 0.07

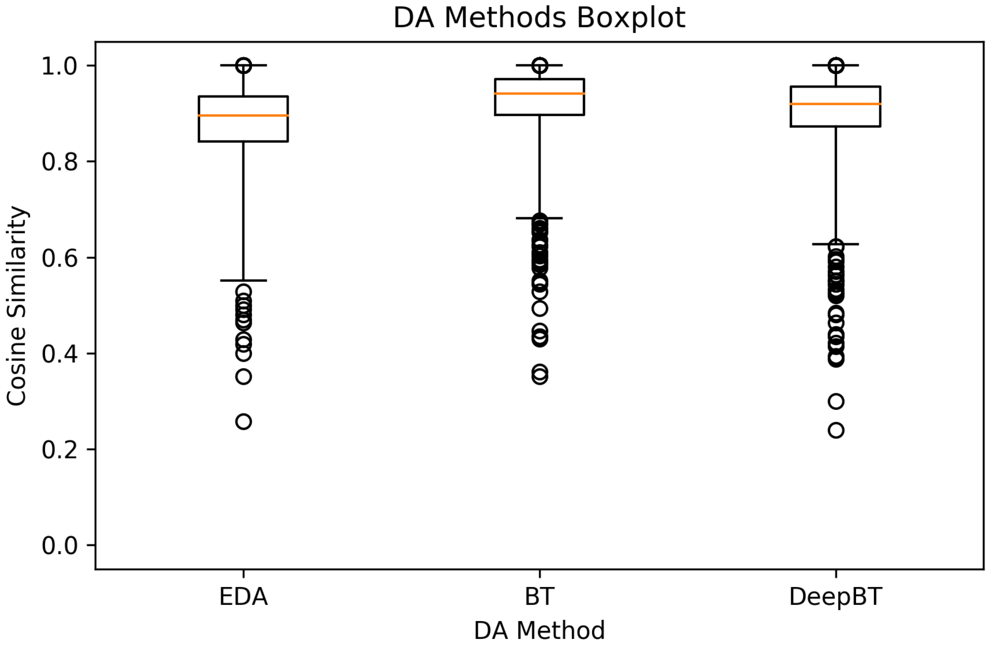
the first dataset presented significant higher standard deviation – and, thus, variability – when compared to the News Statements & Headlines dataset. The boxplots depicted in Figure 6 confirm this difference. Since the DA methods used in both of the datasets were exactly the same, we can infer that this divergence is most probably due to the different linguistic styles of the datasets. This conclusion is in accordance with the previous qualitative observation that DA methods tended to achieve higher stability and quality of synthesized text on the News Statements & Headlines dataset, most probably due to its formal vocabulary and grammar. We also observe that in both datasets DeepBT obtained slightly inferior average similarity when compared to BT-generated data. This result indicate that, as expected, DeepBT synthesized data with a different balance point in the aforementioned diversity-validity trade-off, showing slightly greater variability when compared to traditional BT.

Next, we performed the same analysis, but controlling for the pivotal language in use in order to assess whether choosing different languages could yield significant changes in synthesized data output in translation-based methods. Table 7 and 8 show the result from this investigation for BT and DeepBT, respectively.

From Table 8 onwards, whenever appropriate we will be using PT, DE and RU to represent the pivotal languages Portuguese, German and Russian, respectively. For a more detailed analysis, Figure 7 shows the Boxplot of cosine similarity when applying each of the translation-based DA methods and pivotal languages to both datasets. We observe in both translation-based methods that the pivotal language in use did not present significant impact in the similarity of artificially generated text: cosine similarity presented small variations around similar numbers, always inside the range of one standard deviation.

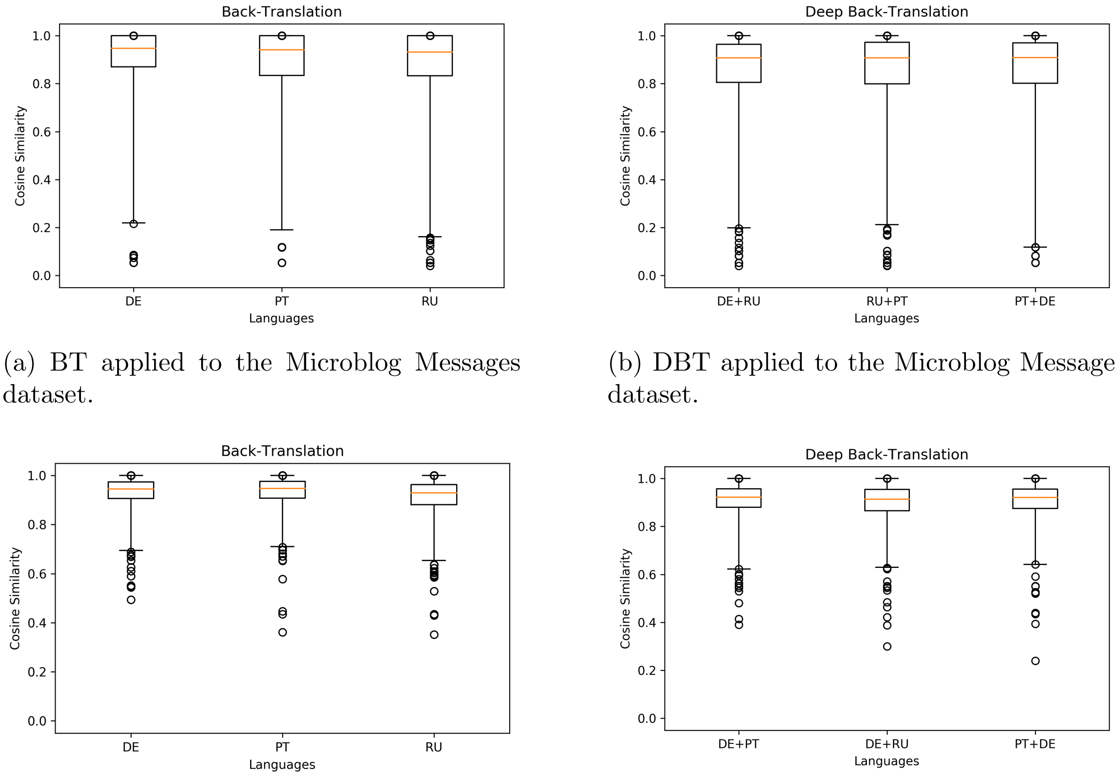


(a) Microblog Message dataset cosine similarity with different DA methods.



(b) News Statements & Headlines Message dataset cosine similarity with different DA methods.

Figure 6: Boxplots show the distribution of cosine similarity obtained in each dataset when using different DA methods. Cosine similarity at the Microblog Messages dataset show greater dispersion when compared to the News Statements & Headlines dataset.



(a) BT applied to the Microblog Messages dataset.

(b) DBT applied to the Microblog Messages dataset.

(c) BT applied to the News Statements & Headlines dataset.

(d) DBT applied to the News Statements & Headlines dataset.

Figure 7: Boxplots show the distribution of cosine similarity obtained in each dataset when using each of the languages as pivot language in translation-based DA methods. Cosine similarity at the Microblog Messages dataset show greater dispersion when compared to the News Statements & Headlines dataset. There is no significant changes in similarity between original and synthesized text when changing pivotal language used.

Table 8: Average Cosine Similarity and Standard Deviation between original and synthesized text when using Deep Back-Translation as DA method for different Pivot Languages.

Dataset	**PT+DE**	**DE+RU**	**RU+PT**
Microblog Messages	0.85 ± 0.18	0.85 ± 0.17	0.85 ± 0.18
News Statements & Headlines	0.90 ± 0.08	0.90 ± 0.08	0.90 ± 0.08

This is an interesting observation since, as the discussion around the last example of Table 4 (*"possible double bottom set up"*), in some examples we observed a clear relation between the quality of synthesized data and the pivotal language used. Since this same behaviour could not be observed in this more quantitative and general analysis, we hypothesize that the degraded quality of synthesized text was either restricted to a relatively small amount of cases or that its occurrence was similarly distributed among the different pivotal languages tested.

In summary, when analyzing the quality of synthesized sentences of each DA method from a qualitative and quantitative standpoint – still not taking in account its ability to assist models' performance – we observe that artificially generated data seems highly dependent on:

- **Quality of translations:** despite not giving rise to significant differences in average similarity of synthesized data, in translation-based methods we observed cases where the same method obtained very different outputs depending on the chosen pivotal language, likely indicating the impact of translation quality into synthesized data.
- **Linguistic style:** final DA outputs seemed sensitive to text formality – in all scenarios investigated their output presented lower average similarity with original data when dealing with informal texts. Hence, DA outputs seemed to be sensitive to proper recognition of words and expressions outside of formal language, like contractions, abbreviations, colloquialism, slang, among others.

Further research could be done in order to validate this observations and deepen our understanding about the quality of synthesized texts within the context of NLP DA methods.

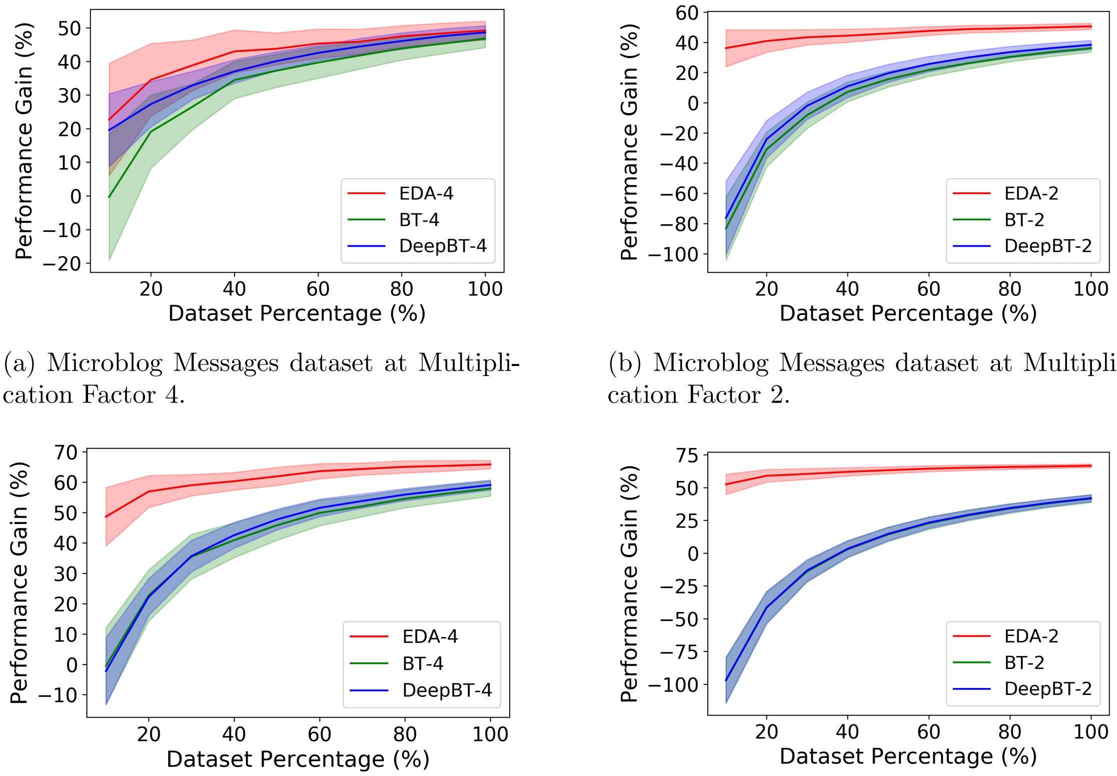
6.2 Performance

Before diving into the analysis of the performance results, it is instructive to first elucidate the details regarding how each outcome was generated, evaluated and compared. The baseline for comparing the outcomes of each of the methods are the output of training without any DA method. In all experiments we used Mean Squared Error (MSE) as the performance metric, matching the loss function used by the NN.

In order to assess the impact of data availability on results and conclusions – and following (WEI; ZOU, 2019) and others – DA procedures were applied having as input varying dataset percentages (every decile from 10% to 100%). Hence, when evaluating performance at a dataset percentage of 20%, 80% of the available dataset was discarded and only the remaining 20% were used as input for the DA technique. For every method we analyzed the impact on results of different **multiplication factors** – a quantity that informs us by how much has the available training data been multiplied by the DA process. Therefore, using a multiplication factor of 3 means that the DA procedure doubled the amount of data, which summed to the original data, add up to 3 times the initial number of records. Consequently, if the original dataset had 1000 sentences, e.g., a dataset percentage of 20% combined with a multiplication factor of 3 would arrive at a final (post-DA) training dataset with 600 records.

We now present and compare the results each of the DA techniques controlling for a variety of factors – such as percentage of dataset used, multiplication factor and language of translation.

We start comparing relative performance gain of DeepBT, BT and EDA against the baseline (trained on 100%) in Fig. 8. In 8a, where we compare each techniques' performance against the Microblog Message dataset at a Multiplication Factor of 4, we observe two very distinct behaviours depending on data availability. When more input data is made available (dataset percentage greater than 50%), all DA methods perform similarly and results variability – represented through standard deviation's shaded area – are low. On the other side of the spectrum (dataset percentage less than 50%), results for each DA techniques show greater variability. Despite some overlap between outcomes, EDA and traditional BT appear respectively as the best- and worse performing methods, DeepBT achieving intermediate results. On the remaining settings (8b to 8d), on the other hand, we observe a different pattern. In these scenarios, BT and DeepBT achieve almost identical results, while EDA performed much better, especially on the low end of data availability.



(a) Microblog Messages dataset at Multiplication Factor 4.

(b) Microblog Messages dataset at Multiplication Factor 2.

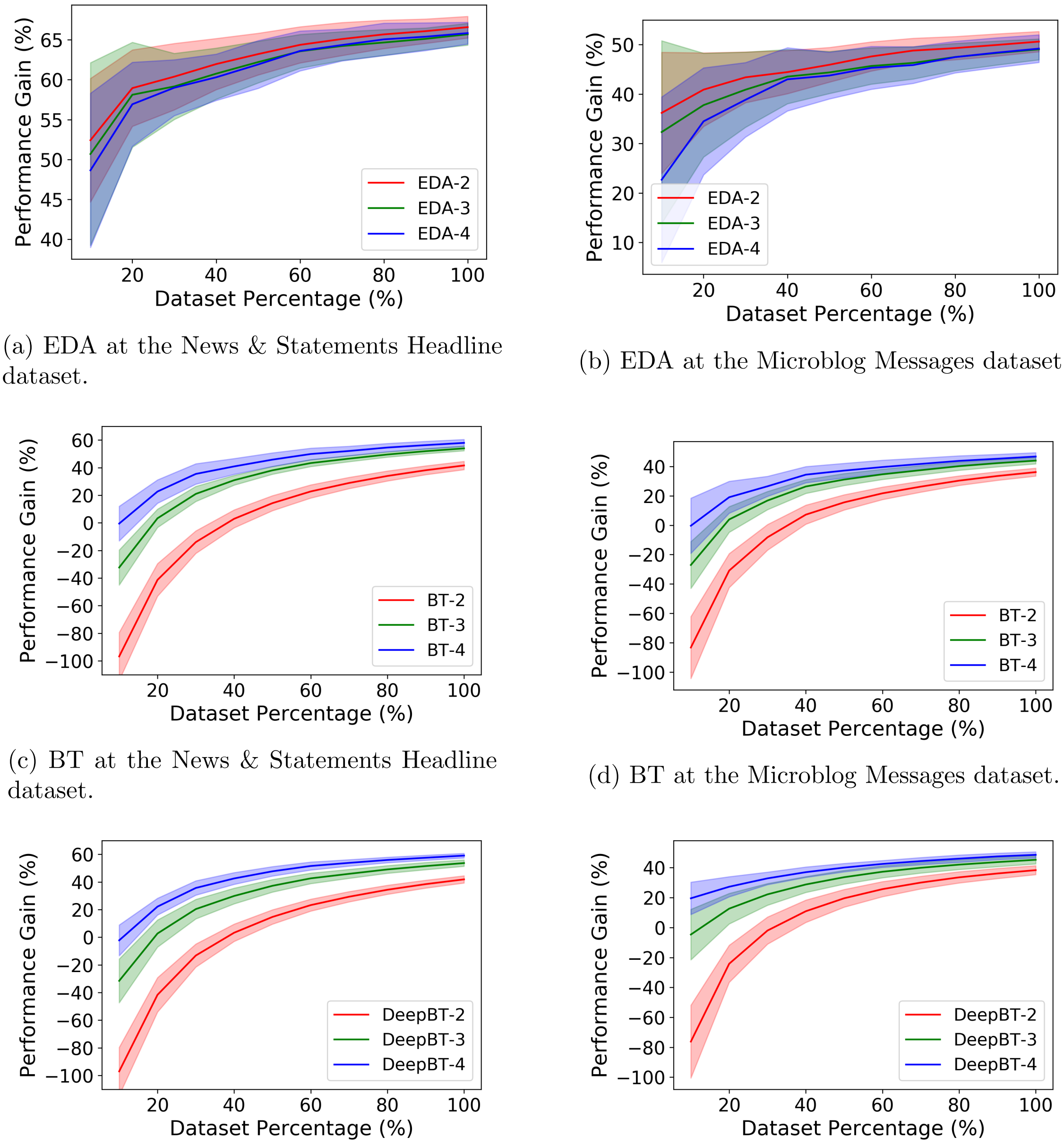
(c) News Statements & Headlines dataset at Multiplication Factor 4.

(d) News Statements & Headlines dataset at Multiplication Factor 2.

Figure 8: Comparison of DA techniques in both datasets. Performance gain is measured against the baseline trained on 100% of the dataset without any DA. Solid lines represent mean output of multiple executions, whereas shaded areas delimit one standard deviation around the mean. Labels are in the format "DA Method – Multiplication Factor". The reader should note that the y-axis are in different scales along the graphs, since using the same scale would hinder proper visualization.

Figures 9d to 9f exhibit how each DA technique's performance is affected by the Multiplication Factor parameter. Despite some overlap between results when the dataset size is doubled or tripled (BT-[2/3] and DeepBT-[2/3]), we notice that both BT and DeepBT achieve greater performance as the Multiplication Factor is increased. When using EDA, on the other hand, we observe great variance in results, the method performing very similarly regardless of the Multiplication Factor in use. 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.

Continuing our analysis about the impact of pivotal languages on translation-based techniques, now from a quantitative perspective, Fig. 10 presents models' performance response for each of the chosen languages used for translation in DeepBT and traditional BT. We do not observe performance divergence between any of the chosen languages used for translation purposes, regardless of the dataset. This result indicates that, despite



(a) EDA at the News & Statements Headline dataset.

(b) EDA at the Microblog Messages dataset.

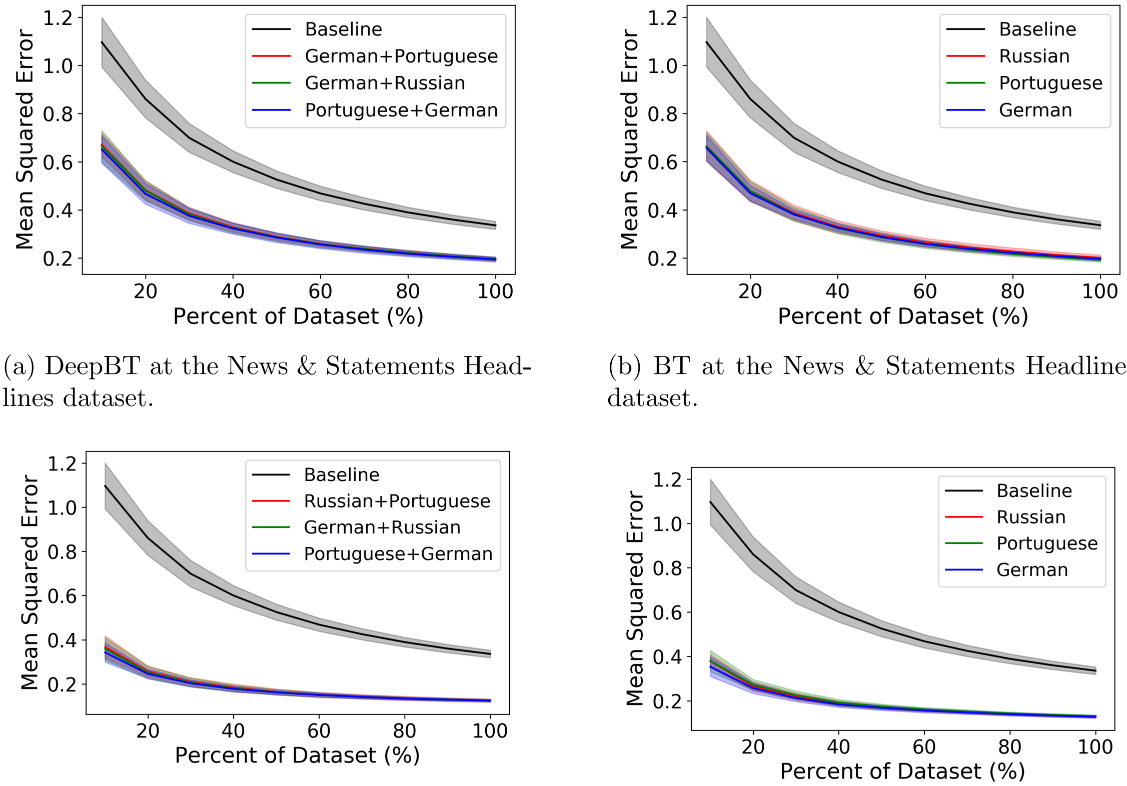
(c) BT at the News & Statements Headline dataset.

(d) BT at the Microblog Messages dataset.

(e) DeepBT at the News & Statements Headline dataset.

(f) DeepBT at the Microblog Messages dataset.

Figure 9: Comparison of DA techniques and response of each technique to varying Multiplication Factor in both datasets. Performance gain is measured against the Baseline trained on 100% of the dataset without any DA. Solid lines represent mean output of multiple executions, whereas shaded areas delimit one standard deviation around the mean. Labels are in the format "Method-Multiplication Factor". The reader should note that the y-axis are in different scales along the graphs, since using the same scale would hinder proper visualization.



(a) DeepBT at the News & Statements Headlines dataset.

(b) BT at the News & Statements Headlines dataset.

(c) DeepBT at the Microblog Messages dataset.

(d) BT at the Microblog Messages dataset.

Figure 10: Comparison of models' performance for different languages choices for BT and DeepBT on the News Statements & Headlines dataset (a and b) and on the Microblog Messages dataset (c and d). Here we compare the Mean Squared Error output by each translation-based DA to the baseline trained on 100% of the dataset without any DA. Since we used a 2-layer DeepBT, in 10a and 10c each line represents a combination of two languages used in sequenced translation.

the qualitative differences observed on some of the synthesized sentences when changing pivotal language (as discussed in 6.1), these variations did not seem to be significant enough to negatively impact performance.

Table 9 presents how model performance is affected by the dropout hyperparameter in both datasets. In contrast to the results obtained by (HERNÁNDEZ-GARCÍA; KÖNIG, 2018) in image processing tasks, we obtained stable performance under varying dropout parameter. This result hold true in both datasets and regardless of the DA method in use, indicating that the combined use of these regularization methods do not seem to bring any significant benefit or drawback.

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 gains between 1% and 3% achieved by the original work proposing EDA (WEI; ZOU, 2019). We come up

Table 9: Average MSE obtained by DA techniques under different dropout values in the Benchmark Datasets.

DA Method	**Microblog Messages**		
	Dropout factor		
	0.0	**0.1**	**0.2**
DeepBT	0.13 ± 0.02	0.12 ± 0.02	0.12 ± 0.02
BT	0.15 ± 0.04	0.13 ± 0.03	0.13 ± 0.02
EDA	0.12 ± 0.02	0.12 ± 0.02	0.12 ± 0.02

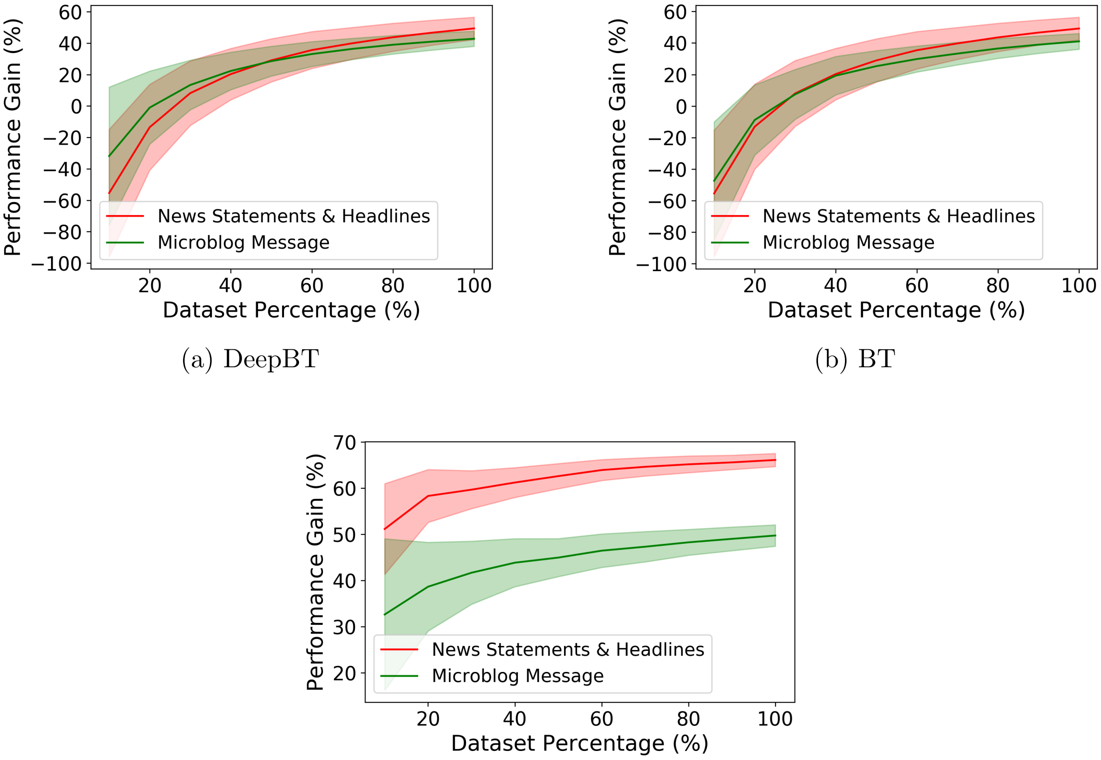
DA Method	**News Statements & Headlines**		
	Dropout factor		
	0.0	**0.1**	**0.2**
DeepBT	0.21 ± 0.07	0.19 ± 0.06	0.19 ± 0.06
BT	0.21 ± 0.07	0.18 ± 0.05	0.19 ± 0.06
EDA	0.13 ± 0.02	0.13 ± 0.03	0.13 ± 0.02

with two hypothesis for this outcome. On one hand, EDA seems to naturally output results with great variance – as shown in Fig. 9b. On the other hand, these results could be attributable to differences in the characteristics of the dataset used in our experiments compared to the ones used by Wei and Zou (2019), like the linguistic nature of the texts. To further explore this hypothesis, we compare how each DA method responds to each of the datasets – which are marked by very distinct linguistic styles. The results are shown in Fig. 11.

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 poorer stability of EDA when facing different linguistic styles when compared to translation-based methods. Further experiments should be put forward to confirm this conjecture.

It is also interesting to compare previous results to the ones obtained by the competing teams at SemEval 2017 Task 5. To make this a valid comparison, we compared our best result achieved when augmenting the whole dataset with the challenge's best result, assuring this way a fair comparison. Table 10 shows how they compare to each other.

In Subtask 1, our best result was achieved when augmenting the original dataset with DeepBT, with a Multiplication Factor of 3 and using German+Russian and Russian+Portuguese as pivotal languages. In Subtask 2, our best result was achieved when



(a) DeepBT

(b) BT

(c) EDA

Figure 11: Comparison of performance gain brought by different DA techniques at each dataset – Microblog Message dataset (green) and News Statements & Headlines (red). Translation-based methods (DeepBT and BT) showed greater homogeneity in outcomes between datasets when compared to EDA, which performed better in the dataset with formal language. The reader should note that the y-axis are in different scales along the graphs, since using the same scale would hinder proper visualization.

Table 10: Comparison between SemEval 2017 Task 5 challenger's best results and our baseline and best results. Result reported in cosine similarity.

Dataset	SemEval 2017 Best Result	Our Baseline	Our DA Best Result	Our DA Performance Lift
Subtask 1	0.778	0.586	0.679 (DeepBT)	15,9%
Subtask 2	0.745	0.489	0.597 (EDA)	22,1%

using EDA and a multiplication factor of 2.

A number of relevant observations are to be taken into account when analyzing this comparison:

1. While the focus of the teams participating in the SemEval 2017 Task 5 challenge was to achieve highest possible test score results, ours was directed to the study of the impact of DA methods on final results. As a result, our efforts were directed to analyzing **relative performance** - with versus without DA, contrasting different DA methods, impact of pivotal language, etc - rather than **absolute performance**.

2. As a consequence of the above, all relevant methodological choices – data preparation steps, model technique and architecture, hyperparameters – were done so as to represent a typical NLP scenario, not so as to necessarily maximize this specific task output.

3. Furthermore, the tackled research questions imposed an inevitable trade-off between optimal performance and adequate comparability. Choices that probably would maximize model output – fine-tuning data preparation steps, model technique, architecture and hyperparameters for each of the subtask datasets and DA techniques, using specific sentiment dictionaries for each of the datasets, etc – would naturally jeopardize comparability, making it difficult to differentiate the influence of each of these factors to results and conclusions.

4. Finally, the SemEval 2017 Task 5 datasets available nowadays and used for this work are not exactly the same that were originally used in the challenge. In e-mails exchanged with the organizers of the challenge, we were informed of the possibility of small discrepancies to the original gold standards due to terms and conditions related to public data disclosure.

In summary, results reported in Table 10 can be considered only an approximation to the real result achieved in case the gold standards were exactly the same as the original ones. Furthermore, it is reasonable to assume that, in a scenario where all the efforts and relevant choices were exclusively directed to absolute performance on this tasks, considerable better results would be achieved.

7 CONCLUSIONS AND FUTURE WORK

In this work, we performed an in-depth study of the most usual DA techniques in NLP, performing a systematic and comparative evaluation of them. We summarize below our main findings and contributions:

1. **NLP DA literature review**: to the best of our knowledge, we are the first to conduct such a comprehensive literature review on this subject. More important, rather than just performing a general overview of several works on the topic, we condensed and highlighted the main research gaps found on literature.

2. **Systematic assessment of NLP DA techniques in SA**: by carrying out an in depth study of various NLP DA methods within SA tasks, we add to the state-of-the-art on NLP DA techniques an unprecedented systematic comparative study, addressing research gaps 1 and 2 discussed in Section 3.7. By bringing additional experimental evidence regarding DeepBT and by deepening the discussion initiated in (FERREIRA; COSTA, 2020), we tackle research gap 3.

3. **Qualitative and quantitative assessment of DA methods' output in Regression tasks**: this work did not restricted itself to pure model performance considerations, but also analyzed synthesized data from a qualitative and quantitative standpoint, helping to enlighten research gap 4. Since experiments were carried out in regression problems, we also addressed research gap 5.

4. **Assessing the impact of Dropout in various DA techniques**: we performed an evaluation of the relationship between these explicit (Dropout) and implicit (DA) regularization techniques in a NLP task. This is a relevant contribution since the combined use of implicit and explicit regularization techniques is commonplace among practitioners – despite the lack of evidences of the benefits of this strategy (HERNÁNDEZ-GARCÍA; KÖNIG, 2018).

5. **Comparison of the effect of DA techniques in Formal and Informal texts**: to assess NLP DA techniques' dependence on linguistic style, we compared their

relative behavior in texts with different levels of formality. The observed difference in performance underlines the importance of developing DA techniques that are robust to linguistic preferences prevalent in informal texts, such as contractions, abbreviations, colloquialism, slang, among others. Once again recalling the list of research gaps shown in section 3.7, with this contribution we helped address gap 6.

In summary, in this work we addressed several of the most important research gaps found in NLP DA literature.

As a result of the investigations conducted during the execution of this master's work, the following articles were published:

1. 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. Porto Alegre, RS, Brasil: SBC, 2019. – **Received "Outstanding Paper Award" prize**

2. FERREIRA, T. M.; COSTA, A. H. R. **DeepBT and NLP Data Augmentation Techniques: A new proposal and a comprehensive study**. In: CERRI, R.; PRATI, R. C. (Ed.). Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I. Springer, 2020. (Lecture Notes in Computer Science, v.12319), p. 435–449. – **Invited by the editors of BRACIS 2020 to submit an extended version of this paper to the Applied Soft Computing (ASOC) Journal**

3. FERREIRA, T. M.; COSTA, A. H. R. **Data Augmentation Techniques in Natural Language Processing**. Applied Soft Computing. Elsevier, 2021 (Submission currently under revision)

With these findings and articles, we hope to encourage further use of DA as an auxiliary method in NLP tasks and to raise NLP DA techniques to greater maturity. Also, we expect that this work could serve as inspiration and starting point to future studies on the subject, so as to tackle some of the remaining open questions on the topic, which we summarize below and that could be explored in future work.

1. **Scarcity of mixed-example research within the NLP domain:** despite unquestionable success in image processing tasks (ZHANG et al., 2018; HENDRYCKS

et al., 2020; HARRIS et al., 2020; SUMMERS; DINNEEN, 2019), to the best of our knowledge there are only four works applying this class of method to NLP (HARRIS et al., 2020; ANDREAS, 2020; GUO; KIM; RUSH, 2020; GUO; MAO; ZHANG, 2019), some of which are still under revision at the time of this writing. Therefore, little is known regarding the efficacy of applying this family of DA techniques to NLP problems.

2. **Efficacy of DA on models using large-scale pretraining**: most studies assess the effectiveness of DA using as baseline models whose parameters are learned from scratch (DAI; ADEL, 2020). Considering the fact that large-scale pretrained models – such as BERT (DEVLIN et al., 2019), XLNet (YANG et al., 2019) and its variations – are used to achieve some of the state-of-the-art results on various NLP tasks, it would be beneficial to better investigate how DA techniques aid the performance of such lage-scale pretrained models. Literature has so far reported mixed results on this topic: while (WEI; ZOU, 2019) and (LONGPRE et al., 2019) reported negligible or no gains when using DA with such type of models, the work of (DAI; ADEL, 2020) obtained relevant gains.

3. **Influence of dataset size on DA effectiveness**: there is amounting evidence that relative gains brought by DA methods decrease rapidly with increasing dataset size: that is, gains brought by DA methods tend to be relevant in smaller datasets (or in fractions of the original full dataset) and marginal on greater datasets (or when using the original full dataset) (DAI; ADEL, 2020; WEI; ZOU, 2019; FERREIRA; COSTA, 2020; ŞAHIN; STEEDMAN, 2018). Further research should deepen our understanding about this behaviour and propose methods where this trend, if present, is less pronounced.

4. **Effectiveness of combined use of different DA techniques**: little work has been done to understand to what extent – if any – the combined use of distinct DA methods could benefit model performance. Two of the few works to perform such an investigation are the works of (WANG et al., 2018) – indicating gains when combining Back-Translation and Switch-Out – and of (DAI; ADEL, 2020)– showing the benefits of the compounded use of DA techniques for Named Entity Recognition tasks. Clearly, if beneficial, the combined use of different DA techniques could bring faster advances to the subject then developing new DA methods from scratch, which constitutes a reasonable motivation for further research.

5. **Theoretical framework and systematic studies for NLP DA techniques**:

some – if not most – of the proposed NLP DA techniques are derived from heuristics, domain knowledge and common sense, instead of formally derived. Research would greatly profit from the development of a formal theoretical framework to support and explain DA methods in NLP. This could allow for better insights into such methods and a proper understanding of the similarities and differences between each technique. One of the few works to derive its proposal – a technique called SwitchOut – from a formal probabilistic framework is the work of (WANG et al., 2018), proving alongside that some of the previously proposed DA NLP methods were special cases of their more general proposition. In regard to systematic studies, (JHA; LOVERING; PAVLICK, 2020) constitutes an interesting work that tries to investigate the issue of generalization in NLP from a very structured and methodical approach. In this work, the authors design a series of toy learning problems to inspect the effectiveness of DA in NLP when training on counterexamples. We strongly believe that approaching DA NLP research using more formal theoretical tools (like (WANG et al., 2018)) or more systematic procedures (like (JHA; LOVERING; PAVLICK, 2020)) constitute promising research paths.

6. **Novel applications for DA**: in almost every work found in the literature, the authors use DA methods with the only purpose of improving model performance. Recent research showed, however, other interesting and valuable applications to DA methods. The works of (LU et al., 2019) and (ZMIGROD et al., 2019), e.g., propose the use of DA for mitigating gender bias in NLP tasks. After observing and quantifying gender bias in language modeling and word embedding, e.g., (LU et al., 2019) proposes Counterfactual Data Augmentation (CDA), a method by which new instances, with interventions on its targeted words, are generated with the purpose of mitigating bias in neural NLP tasks. (ZMIGROD et al., 2019) extends this technique for languages with rich morphology. Both authors achieve gender bias reduction in NLP tasks without sacrificing accuracy or grammaticality. As the aforementioned works show, further research could greatly benefit from expanding the range of applications of DA methods to other aspects other then purely performance issues.

REFERENCES

ANDREAS, J. Good-enough compositional data augmentation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020. p. 7556–7566. Disponível em: ⟨https://www.aclweb.org/anthology/2020.acl-main.676⟩.

BAO, F.; NEUMANN, M.; VU, T. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In: *Proc. Interspeech 2019*. [S.l.: s.n.], 2019. p. 2828–2832.

BASILE, V.; BOSCO, C.; FERSINI, E.; NOZZA, D.; PATTI, V.; PARDO, F. M. R.; ROSSO, P.; SANGUINETTI, M. 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*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 54–63.

BISHOP, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

CASWELL, I.; CHELBA, C.; GRANGIER, D. Tagged back-translation. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*. Florence, Italy: Association for Computational Linguistics, 2019. p. 53–63.

CHAPELLE, O.; WESTON, J.; BOTTOU, L.; VAPNIK, V. Vicinal risk minimization. In: LEEN, T. K.; DIETTERICH, T. G.; TRESP, V. (Ed.). *Advances in Neural Information Processing Systems 13*. MIT Press, 2001. p. 416–422. Disponível em: ⟨http://papers.nips.cc/paper/1876-vicinal-risk-minimization.pdf⟩.

CHEN, S.; DOBRIBAN, E.; LEE, J. H. *A Group-Theoretic Framework for Data Augmentation*. 2019.

CORTIS, K.; FREITAS, A.; DAUDERT, T.; HUERLIMANN, M.; ZARROUK, M.; HANDSCHUH, S.; DAVIS, B. 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)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. p. 519–535.

CUBUK, E. D.; ZOPH, B.; MANE, D.; VASUDEVAN, V.; LE, Q. V. Autoaugment: Learning augmentation policies from data. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 113–123.

DAI, X.; ADEL, H. An analysis of simple data augmentation for named entity recognition. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 3861–3867. Disponível em: ⟨https://www.aclweb.org/anthology/2020.coling-main.343⟩.

DAO, T.; GU, A.; RATNER, A.; SMITH, V.; SA, C. D.; RE, C. A kernel theory of modern data augmentation. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, USA: PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 1528–1537. Disponível em: ⟨http://proceedings.mlr.press/v97/dao19b.html⟩.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Disponível em: ⟨https://www.aclweb.org/anthology/N19-1423⟩.

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*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 489–500.

FELLBAUM, C. Wordnet and wordnets. In: BROWN, K. (Ed.). *Encyclopedia of Language and Linguistics*. Oxford: Elsevier, 2005. p. 665–670. Disponível em: ⟨http://wordnet.princeton.edu/⟩.

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*. Porto Alegre, RS, Brasil: SBC, 2019. p. 729–740. ISSN 0000-0000. Disponível em: ⟨https://sol.sbc.org.br/index.php/eniac/article/view/9329⟩.

FERREIRA, T. M.; COSTA, A. H. R. Deepbt and NLP data augmentation techniques: A new proposal and a comprehensive study. In: CERRI, R.; PRATI, R. C. (Ed.). *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20-23, 2020, Proceedings, Part I*. Springer, 2020. (Lecture Notes in Computer Science, v. 12319), p. 435–449. Disponível em: ⟨https://doi.org/10.1007/978-3-030-61377-8\\_30⟩.

GHOSAL, D.; BHATNAGAR, S.; AKHTAR, M. S.; EKBAL, A.; BHATTACHARYYA, P. IITP at SemEval-2017 task 5: An ensemble of deep learning and feature based models for financial sentiment analysis. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 899–903. Disponível em: ⟨https://www.aclweb.org/anthology/S17-2154⟩.

GONOG, L.; ZHOU, Y. A review: Generative adversarial networks. In: *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. [S.l.: s.n.], 2019. p. 505–510.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: The MIT Press, 2016. ISBN 0262035618.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial nets. In: GHAHRAMANI, Z.; WELLING, M.; CORTES, C.; LAWRENCE,

N. D.; WEINBERGER, K. Q. (Ed.). *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 2014. p. 2672–2680. Disponível em: ⟨http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf⟩.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. [s.n.], 2015. Disponível em: ⟨http://arxiv.org/abs/1412.6572⟩.

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)*. Florence, Italy: Association for Computational Linguistics, 2019. p. 45–52.

GUAN, J.; LI, R.; YU, S.; ZHANG, X. Generation of synthetic electronic medical record text. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. [S.l.: s.n.], 2018. p. 374–380.

GUO, D.; KIM, Y.; RUSH, A. Sequence-level mixed sample data augmentation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020. p. 5547–5552. Disponível em: ⟨https://www.aclweb.org/anthology/2020.emnlp-main.447⟩.

GUO, H.; MAO, Y.; ZHANG, R. Augmenting data with mixup for sentence classification: An empirical study. *CoRR*, abs/1905.08941, 2019. Disponível em: ⟨http://arxiv.org/abs/1905.08941⟩.

GUPTA, R. Data augmentation for low resource sentiment analysis using generative adversarial networks. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2019. p. 7380–7384.

HARRIS, E.; MARCU, A.; PAINTER, M.; NIRANJAN, M.; PRüGEL-BENNETT, A.; HARE, J. *FMix: Enhancing Mixed Sample Data Augmentation*. 2020.

HAUBERG, S.; FREIFELD, O.; LARSEN, A. B. L.; III, J. W. F.; HANSEN, L. K. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In: GRETTON, A.; ROBERT, C. C. (Ed.). *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*. JMLR.org, 2016. (JMLR Workshop and Conference Proceedings, v. 51), p. 342–350. Disponível em: ⟨http://proceedings.mlr.press/v51/hauberg16.html⟩.

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)*. [S.l.: s.n.], 2016. p. 770–778.

HENDRYCKS, D.; MU, N.; CUBUK, E. D.; ZOPH, B.; GILMER, J.; LAKSHMINARAYANAN, B. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

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. (Ed.). *Artificial Neural Networks and Machine Learning – ICANN 2018*. Cham: Springer International Publishing, 2018. p. 95–103. ISBN 978-3-030-01418-6.

HO, D.; LIANG, E.; CHEN, X.; STOICA, I.; ABBEEL, P. Population based augmentation: Efficient learning of augmentation policy schedules. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, USA: PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 2731–2741. Disponível em: ⟨http://proceedings.mlr.press/v97/ho19b.html⟩.

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*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 18–24.

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*. Melbourne, Australia: Association for Computational Linguistics, 2018. p. 55–63.

JHA, R.; LOVERING, C.; PAVLICK, E. *Does Data Augmentation Improve Generalization in NLP?* 2020.

KAR, S.; MAHARJAN, S.; SOLORIO, T. RiTUAL-UH at SemEval-2017 task 5: Sentiment analysis on financial data using neural networks. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 877–882. Disponível em: ⟨https://www.aclweb.org/anthology/S17-2150⟩.

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)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 452–457.

KONDA, K. R.; BOUTHILLIER, X.; MEMISEVIC, R.; VINCENT, P. Dropout as data augmentation. *ArXiv*, abs/1506.08700, 2015.

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 - Volume 1*. Red Hook, NY, USA: Curran Associates Inc., 2012. (NIPS'12), p. 1097–1105.

LEMLEY, J.; BAZRAFKAN, S.; CORCORAN, P. Smart augmentation learning an optimal data augmentation strategy. *IEEE Access*, v. 5, p. 5858–5869, 2017.

Liu, R.; Shi, Y.; Ji, C.; Jia, M. A survey of sentiment analysis based on transfer learning. *IEEE Access*, v. 7, p. 85401–85412, 2019.

LONGPRE, S.; LU, Y.; TU, Z.; DUBOIS, C. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 220–227. Disponível em: ⟨https://www.aclweb.org/anthology/D19-5829⟩.

LU, K.; MARDZIEL, P.; WU, F.; AMANCHARLA, P.; DATTA, A. *Gender Bias in Neural Natural Language Processing*. 2019.

MANNING, P. R. S. C. D. *Introduction to Information Retrieval*. [S.l.: s.n.], 2008.

MANSAR, Y.; GATTI, L.; FERRADANS, S.; GUERINI, M.; STAIANO, J. Fortia-FBK at SemEval-2017 task 5: Bullish or bearish? inferring sentiment towards brands from financial news headlines. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 817–822. Disponível em: ⟨https://www.aclweb.org/anthology/S17-2138⟩.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. In: BENGIO, Y.; LECUN, Y. (Ed.). *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. [s.n.], 2013. Disponível em: ⟨http://arxiv.org/abs/1301.3781⟩.

MIKOłAJCZYK, A.; GROCHOWSKI, M. Data augmentation for improving deep learning in image classification problem. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*. [S.l.: s.n.], 2018. p. 117–122.

MIMURA, M. Using fake text vectors to improve the sensitivity of minority class for macro malware detection. *J. Inf. Secur. Appl.*, v. 54, p. 102600, 2020. Disponível em: ⟨https://doi.org/10.1016/j.jisa.2020.102600⟩.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. ISBN 978-0-07-042807-2.

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*. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 1–17.

MORRIS, J. X.; LIFLAND, E.; YOO, J. Y.; QI, Y. *TextAttack: A Framework for Adversarial Attacks in Natural Language Processing*. 2020.

MURPHY, K. P. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN 9780262018029 0262018020.

NIU, T.; BANSAL, M. Automatically learning data augmentation policies for dialogue 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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 1317–1323. Disponível em: ⟨https://www.aclweb.org/anthology/D19-1132⟩.

PAN, Z.; YU, W.; YI, X.; KHAN, A.; YUAN, F.; ZHENG, Y. Recent progress on generative adversarial networks (gans): A survey. *IEEE Access*, v. 7, p. 36322–36333, 2019.

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)*. [S.l.: s.n.], 2014. p. 1532–1543.

ŞAHIN, G. G.; STEEDMAN, M. Data augmentation via dependency tree morphing for low-resource languages. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 5004–5009. Disponível em: ⟨https://www.aclweb.org/anthology/D18-1545⟩.

SALAMON, J.; BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, v. 24, n. 3, p. 279–283, 2017.

SemEval — International Workshop on Semantic Evaluation. 2021. Accessed: 2021-02-15. Disponível em: ⟨https://semeval.github.io/⟩.

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)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 86–96.

SHORTEN, C.; KHOSHGOFTAAR, T. A survey on image data augmentation for deep learning. *Journal of Big Data*, v. 6, 12 2019.

SPRINGENBERG, J. T.; DOSOVITSKIY, A.; BROX, T.; RIEDMILLER, M. A. Striving for simplicity: The all convolutional net. In: BENGIO, Y.; LECUN, Y. (Ed.). *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*. [s.n.], 2015. Disponível em: ⟨http://arxiv.org/abs/1412.6806⟩.

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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 35–44.

SUMMERS, C.; DINNEEN, M. J. Improved mixed-example data augmentation. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. [S.l.: s.n.], 2019. p. 1262–1270.

SZEGEDY, C.; ZAREMBA, W.; SUTSKEVER, I.; BRUNA, J.; ERHAN, D.; GOODFELLOW, I. J.; FERGUS, R. Intriguing properties of neural networks. In: BENGIO, Y.; LECUN, Y. (Ed.). *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. [s.n.], 2014. Disponível em: ⟨http://arxiv.org/abs/1312.6199⟩.

TAYLOR, L.; NITSCHKE, G. Improving deep learning with generic data augmentation. In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. [S.l.: s.n.], 2018. p. 1542–1547.

TOKOZUME, Y.; USHIKU, Y.; HARADA, T. Between-class learning for image classification. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018. p. 5486–5494. Disponível em: ⟨http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Tokozume\\_Between-Class\\_Learning\\_for\\_CVPR\\_2018\\_paper.html⟩.

TOKOZUME, Y.; USHIKU, Y.; HARADA, T. Learning from between-class examples for deep sound recognition. In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. Disponível em: ⟨https://openreview.net/forum?id=B1Gi6LeRZ⟩.

VAPNIK, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, v. 10, n. 5, p. 988–999, 1999.

WANG, W. Y.; YANG, D. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015. p. 2557–2563. Disponível em: ⟨https://www.aclweb.org/anthology/D15-1306⟩.

WANG, X.; PHAM, H.; DAI, Z.; NEUBIG, G. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 856–861. Disponível em: ⟨https://www.aclweb.org/anthology/D18-1100⟩.

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)*. Hong Kong, China: Association for Computational Linguistics, 2019. p. 6382–6388.

WEN, Q.; SUN, L.; SONG, X.; GAO, J.; WANG, X.; XU, H. *Time Series Data Augmentation for Deep Learning: A Survey*. 2020.

XIE, Q.; DAI, Z.; HOVY, E. H.; LUONG, T.; LE, Q. Unsupervised data augmentation for consistency training. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. [s.n.], 2020. Disponível em: ⟨https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html⟩.

YADAV, A.; VISHWAKARMA, D. K. Sentiment analysis using deep learning architectures: a review. *Artif. Intell. Rev.*, v. 53, n. 6, p. 4335–4385, 2020. Disponível em: ⟨https://doi.org/10.1007/s10462-019-09794-5⟩.

YADOLLAHI, A.; SHAHRAKI, A. G.; ZAÏANE, O. R. Current state of text sentiment analysis from opinion to emotion mining. *ACM Comput. Surv.*, v. 50, n. 2, p. 25:1–25:33, 2017. Disponível em: ⟨https://doi.org/10.1145/3057270⟩.

YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R. R.; LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. In: WALLACH, H.; LAROCHELLE, H.; BEYGELZIMER, A.; ALCHé-BUC, F. d'; FOX, E.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. v. 32, p. 5753–5763. Disponível em: ⟨https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf⟩.

YU, A. W.; DOHAN, D.; LUONG, M.-T.; ZHAO, R.; CHEN, K.; NOROUZI, M.; LE, Q. V. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, abs/1804.09541, 2018. Disponível em: ⟨https://arxiv.org/pdf/1804.09541⟩.

ZHANG, H.; CISSE, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. Disponível em: ⟨https://openreview.net/forum?id=r1Ddp1-Rb⟩.

ZHANG, H.; ZHOU, H.; MIAO, N.; LI, L. Generating fluent adversarial examples for natural languages. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 5564–5569. Disponível em: ⟨https://www.aclweb.org/anthology/P19-1559⟩.

ZHANG, W. E.; SHENG, Q. Z.; ALHAZMI, A.; LI, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.*, Association for Computing Machinery, New York, NY, USA, v. 11, n. 3, abr. 2020. ISSN 2157-6904. Disponível em: ⟨https://doi.org/10.1145/3374217⟩.

ZHANG, X.; ZHAO, J.; LECUN, Y. Character-level convolutional networks for text classification. In: CORTES, C.; LAWRENCE, N.; LEE, D.; SUGIYAMA, M.; GARNETT, R. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. v. 28, p. 649–657. Disponível em: ⟨https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf⟩.

ZHANG, Y.; WALLACE, B. C. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In: *Proceedings of the The 8th International Joint Conference on Natural Language Processing*. [S.l.: s.n.], 2017. p. 253–263.

ZHAO, D.; YU, G.; XU, P.; LUO, M. Equivalence between dropout and data augmentation: A mathematical check. *Neural networks : the official journal of the International Neural Network Society*, v. 115, p. 82–89, 2019.

ZMIGROD, R.; MIELKE, S. J.; WALLACH, H.; COTTERELL, R. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 1651–1661. Disponível em: ⟨https://www.aclweb.org/anthology/P19-1161⟩.