Divide to Better Classify

Yves Mercadier[1(✉)], Jérôme Azé[1], and Sandra Bringay[1,2]

[1] LIRMM UMR 5506, Université de Montpellier, CNRS, Montpellier, France
{yves.mercadier,jerome.aze,sandra.bringay}@lirmm.fr,
http://www.lirmm.fr/
[2] Université Paul-Valéry Montpellier 3, Montpellier, France

Abstract. Medical information is present in various text-based resources such as electronic medical records, biomedical literature, social media, etc. Using all these sources to extract useful information is a real challenge. In this context, the single-label classification of texts is an important task. Recently, in-depth classifiers have shown their ability to achieve very good results. However, their results generally depend on the amount of data used during the training phase. In this article, we propose a new approach to increase text data. We have compared this approach for 5 real data sets with the main approaches in the literature and our proposal outperforms in all configurations.

Keywords: Natural language processing · Document classification · Textual data augmentation

1 Introduction

Medical information is present in various text-based resources such as electronic medical records, biomedical literature, social media, etc. Using all these sources to extract useful information is a real challenge. In this context, the single-label classification of texts is an important task. Recently, in-depth classifiers have shown their ability to achieve very good results. However, their results generally depend on the amount of data used during the training phase.

In this article, we focus on data augmentation methods that can be effective on small data sets. Data augmentation uses limited amounts of data and transforms existing samples to create new ones. Then, an important challenge is to generate new data that retain the same label. More precisely, it is a matter of injecting knowledge by taking into account the invariant properties of the data after particular transformations. The augmented data can thus cover unexplored input space and improve the generalization of the model.

This technique has proven to be very effective for image classification tasks, especially when the training database is limited. For example, for image recognition, it is well known that minor changes due to scaling, cropping, distortion, rotation, etc. do not change the data labels because these changes can occur in real-world observations. However, transformations that preserve text data labels are not as obvious and intuitive.

M. Michalowski and R. Moskovitch (Eds.): AIME 2020, LNAI 12299, pp. 89–99, 2020.
https://doi.org/10.1007/978-3-030-59137-3_9

In this article, we present a new technique for augmenting textual data, called DAIA (**D**ata **A**ugmentation and **I**nference **A**ugmentation). This method is simple to implement because it does not require any semantic resources or a long training phase. We will evaluate DAIA for various medical data sets of different nature and show an improvement over the state of the art, especially on small data sets. To realize our proposition we use the python module Mantéïa. The code is public and accessible at[1].

2 State of the Art

Data augmentation has been used successfully in the field of image analysis [15]. For example, Perez et al. [11] compared several simple techniques such as cropping, rotating and flipping images. They also used more advanced techniques such as GAN (Generative Adversarial Network) to generate images of different styles. The neural network learns which type of augmentation improves the classifier the most. While many solutions exist for image analysis, augmentation methods have been much less studied in the field of text analysis. There are four main approaches that we will describe below:

Approaches using semantic resources were first proposed. For example, Zhang et al. [19] used a thesaurus to replace words with their synonyms in order to create an augmented dataset used for text classification. This increase proved to be inefficient and in some cases even reduced performance. The authors explain that when large amounts of real data are available, models are easily generalized and are not improved by increasing data.

Approaches inspired by the distortions that can be added to images have also been applied to texts. For the classification of texts, [16], in the EDA (Easy Data Augmentation) method, the number of samples is increased by deleting, swapping a word or replacing it with a synonym. Some approaches focused on the choice of words to be changed. [7] analysed the context of the word to find the one to be swapped. The context is defined by the training of an LSTM-type neural network. This approach improved accuracy by 0.5% over five datasets. In the UDA (Unsupervised Data Augmentation) method, [17] replaced words with low information content, identified with a low TF-IDF, with their synonyms while retaining those with high TF-IDF values representing keywords. This heuristic has been tested on six datasets, and the authors have shown that it is possible to reduce the classification error.

Generative approaches have also been explored. [6] has formed GAN models on small datasets and used them to augment the data to improve the generalization of a sentiment classifier. The results improve the accuracy by 1% on two of the datasets.

A final approach is based on increasing the textual data using back-translation [4]. This involves translating an example into a language and then translating the resulting translation into the original language. Shleifer et al [14]

[1] https://github.com/ym001/Manteia/blob/master/notebook/notebook_Manteia_classification_augmentation_run_in_colab.ipynb.

has shown that the back-translation technique has hardly improved with modern classifiers such as UMLfit.

In this article, we will focus on a new approach, simple to implement, which does not require resources such as semantic approaches, or large amounts of computation such as generative approaches, or even access to external resources such as reverse translation approaches. A limitation of distortion approaches found in the literature is that they do not preserve the order of words in sentences. The examples generated are different from those we might find in the real world and do not allow for effective embedding. In the proposed DAIA approach, we will describe three approaches to divide sentences during the learning phase into several sequences, which will be used to augment the textual data. The same approach will be used during the testing phase on the examples to be classified by applying a soft voting technique.

We will show in the experiments that DAIA approach improves the results of text classification over deep classifiers [3], RoBERTa [9], Albert [8], DistilBert [13] or ScienceBert [1]. We will also compare the DAIA proposal to the UDA [17] and EDA [16] distortion approaches, the TextGen generative approach [6] and the back translation [14].

3 DAIA

The DAIA method is structured in two parts: the first is related to training (DA: Data Augmentation) and the second is related to the test phase (IA: Inference Augmentation).

Data Augmentation: In the training phase, we increase the amount of data by dividing the initial text of each sample. We seek to produce new samples without modifying the order between words to the initial sequence. The objective is to not decrease the learning quality of the description of word embeddings. We have conducted preliminary experiments based on different types of divisions, which for lack of space, are not described in this article. We present below only the three methods that we have combined to form the pyramidal division. These three approaches split the initial sentence into n sequences of words which will be associated with the same label as the initial sentence.

- Symmetrical Division 1: We divide the text symmetrically by removing $x\%$ from the text at both ends to generate a new text sequence. For each initial text, we thus obtain an additional text.
- Division 2 by sliding window: We cut the sentence by applying a sliding window of size l which moves m words to the end of the initial text. The number of generated sequences depends on the length of the initial sentence.
- Division 3 into equal parts: The division is done in i equal parts. The results is i new documents plus the original document.

After data augmentation, from each text in the initial training data set, we generated a set of new texts, associated with the same label as the initial

text, and are used as input into the learning model. We tested the advantages and disadvantages of these different types of division and finally proposed a pyramidal division that combines divisions 1 and 3 and is described in Fig. 1. This new division is based on n levels. The increase for level 1 is done by the symmetrical division 1. The increase for level 2 is done by dividing the text into two equal parts and adding the increase obtained in level 1. The increase for level i is done by dividing the text into i equal parts and adding the increase at level $i-1$. For each text, this results in $\frac{n\times(n+1)}{2}$ new labeled segments.

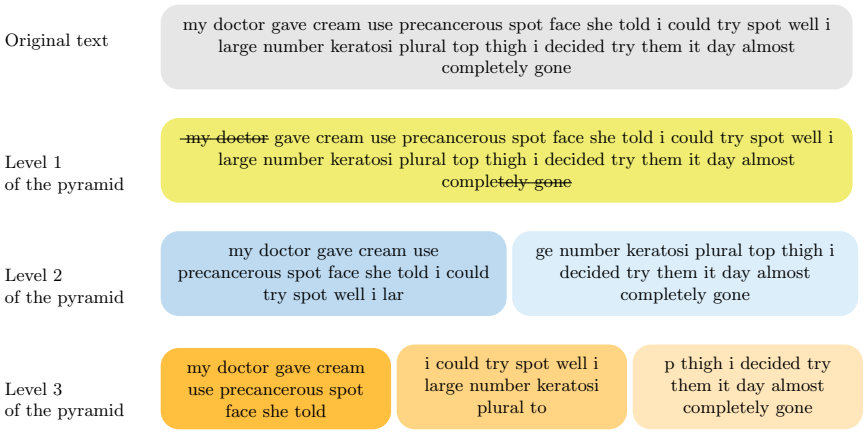

Fig. 1. Description of the pyramidal division. The figure shows the division of the original text into 6 new documents according to the tree levels.

Inference Augmentation: The test phase consists in predicting the labels for new texts from the model learned during the learning phase. We divide the text to be classified according to the same protocol as described for the training phase. Then we give the classifier all the generated sequences. For each sequence, the classifier returns a prediction which is aggregated for the initial text by soft-voting (sum of the values for each predicted class of each element of the set). Thus, for each initial text, one prediction per class is obtained.

4 Presentation of the Experiments

4.1 Data Sets

In order to show the generalisation of our approach, we selected five data sets in the medical field, described in Table 1. The first two data sets correspond to medical publications. The other three data sets correspond to texts written by patients. It is important to note that these data sets are unbalanced.

Table 1. Description of the textual data sets according to the number of classes and documents, the length of the documents in term of words and classification tasks.

Data sets	Classes	Documents	Length	Task
PubMed 200k RCT	5	2 211 861	26.22	Text analysis
WHO COVID-19	4	26 909	166	Provenance analysis
Drugs.com	10	53 766	85.58	Sentiment analysis
eR anorexie	2	84 834	38.24	Sentiment analysis
eR depression	2	531 394	36.76	Sentiment analysis

PubMed 200k RCT[2] Dernoncourt et al. [2] is a database containing more than 200,000 abstracts of articles dealing with randomized controlled trials, with more than 2 million sentences. Each sentence is labeled according to its meaning in the abstract (context, objective, method, result and conclusion).

WHO COVID-19[3], was proposed by the Allen for AI Institute. This data set is composed of more than 29,000 articles about COVID-19 and the corona virus family. In this study we only used articles with an abstract. Each text is labeled according to its source: CZI, PMC, biorxiv, medrxiv.

Drugs.com[4] Gräßer et al. [5] corresponds to patients opinions about drugs. Data were obtained by analyzing online pharmaceutical sites. Each text is labeled with a score from 1 to 10 corresponding to patients' satisfaction.

The eR Depression and eR Anorexia data sets were produced for the CLEF eRisk 2018 challenge[5]. The texts correspond to messages from users in the social network Reddit[6]. [12]. Each text is labeled according to the depression/non-depression and anorexia/non-anorexia classes.

4.2 Data Pre-processing

For each dataset, we applied the following pre-processing: removal of punctuation, special characters, stop words, shift from upper to lower case and lemmatization. For lemmatization, we use the NLTK python module associated with the Wordnet dictionary. Each text can be associated with zero, one or more classes depending on the dataset, used as prediction output.

4.3 Comparison to Other State of the Art Approaches

We compare our proposal to three state-of-the-art methods, which are described in detail below.

[2] https://github.com/Franck-Dernoncourt/pubmed-rct.
[3] https://pages.semanticscholar.org/coronavirus-research.
[4] https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29.
[5] https://early.irlab.org/2018/index.html.
[6] https://www.reddit.com/.

For semantic word distortion approaches, we considered the EDA and UDA approaches. For EDA, we implement the four simple data augmentation techniques described by Wei et al. [16]: replacement by synonyms, random insertion, random exchange, random deletion. For this, we used Wordnet thesaurus of NLTK[7]. For UDA, in order to identify the words with the most informational content, we proceed as Xie et al. [17] who identify these words as being those negatively correlated with their TF-IDF score. For this, they define the probability $min(p(C - TFIDF(xi))/Z, 1)$, where p is a hyper-parameter controlling the variation of augmentation, C is the maximum TF-IDF score for words x_i of a text x and $Z = \sum_i (C - TFIDF(xi))/|x|$. For our experiments, we have chosen $p = 0.9$. Xie et al. [18] do not change the keywords, identified by their frequency. Spotted words which are not key words are replaced by one of the non essential words of the corpus.

As the state-of-the-art text generators used for data augmentation have been exceeded in semantic quality by the generator GPT2 [6], we used the latter according to the following protocol. For each text, an augmented text is constructed by concatenating two parts. An initial part, the seed, corresponding to half of the original text and a second part obtained using the GPT2 generator having taken the seed as input.

Data augmentation by back translation [14] consists of producing paraphrases which globally preserve the semantics of the initial sentence. We use the translation web service Yandex[8] in order to produce these paraphrases. We first translate the texts into Japanese and then translate them back into English and then label them with the original label of the text.

4.4 Assessment Protocol

We proceed as follows for each data set described in Sect. 4.1. We extract 5,000 texts while respecting the weighting of the classes. We then separate the data into two sets: a first set of 1,250 texts (i.e. 25%) is used for the test phase and a second set of 3,750 texts (i.e. 75%) is used for the learning phase while respecting the class stratification. To estimate the quality of learning, we use the accuracy metric, which is calculated as the ratio of the number of labels correctly assigned by the classifier to the total number of labels. All the values produced in this study were calculated on the average of a four-fold cross-validation.

4.5 Hyper-parameters and Training

All hyper-parameters in the networks are set for all data sets. The learning rate is set to 0.00001. All our neural networks are trained by a back-propagation process using a cross-entropy error (*loss*) function. The gradient calculation optimizers are of the *Adam weight* type. In addition, we use a linear learning rate update with warm-up. The experiments are performed on two GeForce RTX-type GPUs.

[7] https://www.nltk.org/howto/wordnet.html.
[8] https://yandex.com/.

5 Results of the Experiments

We have made available some preliminary experiments at this URL[9] in order to compare the tree divisions that allow us to define the pyramidal division that is evaluated in the rest of this section.

5.1 Impact of the Classifier and Comparison to the State of the Art

We're working on the drugs.com dataset. The baseline corresponds to the use of a single classifier without increasing the data. The classifiers compared are the most efficient in the literature: Bert [3], RoBERTa [9], Albert [8], DistilBert [13] or ScienceBert [1]. The DAIA method is applied here for data augmentation during the learning phase and the test phase as detailed in Sect. 3.

The results are presented in the Table 2. The pyramidal division during the learning phase outperforms the other divisions, regardless of the classifier. It is also superior to other approaches in the literature such as EDA and UDA or even reverse translation, for Roberta and Xlnet. Combined with increasing inference during the test phase, DAIA outperforms for all classifiers. The Roberta network over-performs in terms of accuracy.

Table 2. DAIA impact on the accuracy for 6 classifiers and comparison with other data augmentation approaches.

Classifier	Roberta	Bert	Xlnet	Albert	Distilbert	Scibert
Baseline without DA	0.3661	0.3637	0.3760	0.3200	0.3601	0.3392
Distorsion EDA	0.38	0.3669	0.3712	0.3226	0.3689	0.3510
Distorsion UDA	0.3811	0.3632	0.3525	0.3084	0.3660	0.3327
Text generation GPT2	0.3384	0.3252	0.3425	0.3122	0.3204	0.3078
Back translation	0.3798	0.3462	0.3728	0.3426	0.3584	0.3361
DA - symmetrical division	0.3769	0.3491	0.3568	0.3154	0.3359	0.3498
DA - sliding division	0.3494	0.3399	0.3657	0.3266	0.3527	0.3156
DA - pyramid division	0.3877	0.3660	0.3762	0.334	0.3688	0.3590
DAIA	**0.3931**	**0.3755**	**0.3786**	**0.3444**	**0.3693**	**0.3625**

5.2 Impact of Number of Classes

In Table 3, we study the performance of the DAIA according to the number of categories for the five corpora in the study. We selected Roberta as a classifier with a sample of 500 texts. We notice a greater improvement in DAIA performance for datasets with a high number of classes such as Drugs.com (10).

[9] https://github.com/ym001/DAIA/blob/master/Preliminary%20experiment.pdf.

Table 3. Impact of the number of classes on the accuracy for text classification. We used the best classifier Roberta alone and combined with DAIA.

Classifier	Drugs.com	PubMed 200k RCT	WHO COVID-19	eR Depression	eR anoxeria
Roberta	0.2846	0.6413	0.9319	0.8926	0.9079
DAIA	**0.316**	**0.6513**	**0.938**	**0.9066**	**0.9153**

5.3 Impact of the Amount of Data on the Learning Phase

In this experiment, we used the Roberta classifier as a baseline without increasing the data. We show in Fig. 2 that DAIA provides an improvement for all sizes of the eR anorexia (left) and Drugs.com (right) datasets. In particular, from the 500 text datasets, we observe an improvement of 2.7%. The impact is reduced when the dataset reaches the size of 5,000 texts.

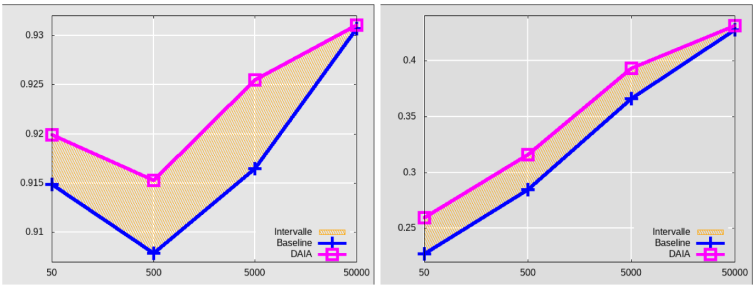

Fig. 2. Impact of DAIA on the accuracy of the text classification according to the size of the corpus on eR Anorexia (left) and Drugs.com (right)

5.4 Impact of the Selected Sequence Size

In Fig. 3, we study the variation in DAIA according to the size of the divisions obtained with approach 3 and the length of the texts in input for the data set drugs.com. Too small a division doesn't make a difference. We got our best result for a level of three combined with a sequence length of 128 words. The size of the input text has little impact in the range studied with a very slight maximum for the 128 word value. We will therefore advise those who wish to test the implementation of the DAIA a level three of the pyramid division with a maximum sequence size of 128. Similar results have been obtained with other data sets.

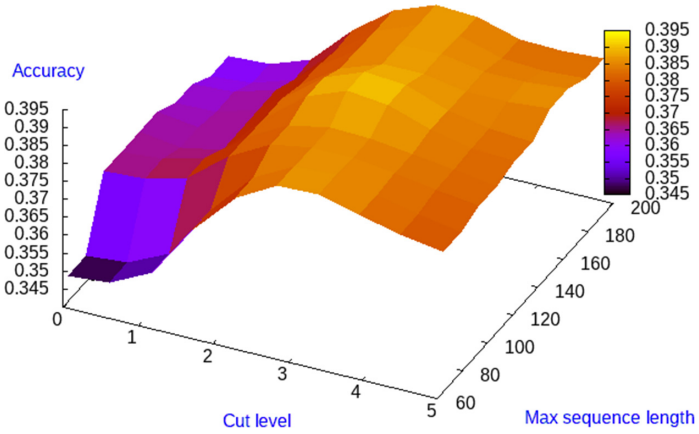

Fig. 3. Accuracy according to the level of pyramidal division and the maximum length of the sequences used as the input of the network.

6 Conclusion

In this study, we proposed a new method to increase textual data, which consists of dividing texts into several segments to increase the variety of training examples while preserving the quality of the learning words embedding. This method, called Data Augmentation and Inference Augmentation, is a distortion approach that does not require any semantic resources, nor any very important training phase, and no external resources. Our DAIA approach has proven to be effective on 5 medical datasets, for 5 classifiers and has been successfully compared to the main approaches in the literature. Comparison with approaches not based on Bert must be conducted to confirm other results. As differences in performance are small, statistical significance tests must also be performed. The impact of the choice of language on the retro-translation must be assessed. Furthermore, it is possible to keep the meaning without keeping the order of words. We could then imagine new experiments consisting in preserving certain levels of syntactic articulation rather than the order of words in order to generate more variability while preserving the semantics of the sentence. In the future, we plan to use multimodal data sets composed of text, sound and images for classification purposes and we will also focus on more complex tasks such as multi-label classification. Finally, DAIA will also be applied to deep active learning heuristics in the medical field [10].

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) EMNLP-IJCNLP 2019, Hong Kong, China, 2019. pp. 3613–3618. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1371
2. Dernoncourt, F., Lee, J.Y.: Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In: Kondrak, G., Watanabe, T. (eds.) IJCNLP 2017. Volume 2: Short Papers, Taipei, Taiwan, 2017, pp. 308–313. Asian Federation of Natural Language Processing (2017)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) NAACL-HLT 2019. Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2019, pp. 4171–4186, Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
4. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. CoRR abs/1808.09381 (2018)
5. Gräßer, F., Kallumadi, S., Malberg, H., Zaunseder, S.: Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In: Kostkova, P., Grasso, F., Castillo, C., Mejova, Y., Bosman, A., Edelstein, M. (eds.) Digital Health, DH 2018, pp. 121–125. ACM, Lyon (2018). https://doi.org/10.1145/3194658.3194677
6. Gupta, R.: Data augmentation for low resource sentiment analysis using generative adversarial networks. CoRR abs/1902.06818 (2019)
7. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations. In: Walker, M.A., Ji, H., Stent, A. (eds.) NAACL-HLT, Volume 2 (Short Papers), New Orleans, Louisiana, USA, 1–6 June 2018, pp. 452–457. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-2072
8. Lan, Z.Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite BERT for self-supervised learning of language representations. http://arxiv.org/abs/1909.11942 (2019)
9. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
10. Maldonado, R., Harabagiu, S.M.: Active deep learning for the identification of concepts and relations in electroencephalography reports. J. Biomed. Inform. **98**, 103265 (2019). https://doi.org/10.1016/j.jbi.2019.103265
11. Perez, L., Wang, J.: The effectiveness of data augmentation in image classification using deep learning. CoRR abs/1712.04621 (2017)
12. Ragheb, W., Moulahi, B., Azé, J., Bringay, S., Servajean, M.: Temporal mood variation: at the CLEF eRisk-2018 tasks for early risk detection on the Internet. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018, Avignon, France, 10–14 September 2018, vol. 2125. CEUR-WS.org (2018)
13. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter https://arxiv.org/abs/1910.01108 (2019)
14. Shleifer, S.: Low resource text classification with ULMFit and backtranslation. CoRR abs/1903.09244 (2019)
15. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019). https://doi.org/10.1186/s40537-019-0197-0

16. Wei, J.W., Zou, K.: EDA: easy data augmentation techniques for boosting performance on text classification tasks. CoRR abs/1901.11196 (2019)
17. Xie, Q., Dai, Z., Hovy, E.H., Luong, M., Le, Q.V.: Unsupervised data augmentation. CoRR abs/1904.12848 (2019)
18. Xie, Z., et al.: Data noising as smoothing in neural network language models. In: ICLR Proceedings of the Conference on Track 2017, OpenReview.net, Toulon (2017)
19. Zhang, X., LeCun, Y.: Text understanding from scratch. CoRR abs/1502.01710 (2015)