Data Augmentation through Back-Translation for Stereotypes and Irony Detection

Tom Bourgeade[1,*], Silvia Casola[2], Adel Mahmoud Wizani[3] and Cristina Bosco[3]

[1]LORIA, University of Lorraine, Nancy, France
[2]MaiNLP & MCML, LMU Munich, Germany
[3]Dipartimento di Informatica, Università di Torino, Turin, Italy

Abstract

Complex linguistic phenomena such as stereotypes or irony are still challenging to detect, particularly due to the lower availability of annotated data. In this paper, we explore Back-Translation (BT) as a data augmentation method to enhance such datasets by artificially introducing semantics-preserving variations. We investigate French and Italian as source languages on two multilingual datasets annotated for the presence of stereotypes or irony and evaluate French/Italian, English, and Arabic as pivot languages for the BT process. We also investigate cross-translation, i.e., augmenting one language subset of a multilingual dataset with translated instances from the other languages. We conduct an intrinsic evaluation of the quality of back-translated instances, identifying linguistic or translation model-specific errors that may occur with BT. We also perform an extrinsic evaluation of different data augmentation configurations to train a multilingual Transformer-based classifier for stereotype or irony detection on mono-lingual data.

Warning: *This paper may contain potentially offensive example messages.*

Keywords

Data Augmentation, Back Translation, Irony Detection, Stereotypes Detection, Low-Resource NLP

1. Introduction

Equipping systems with linguistics-grounded capabilities can be complex. Despite the advancements by Large Language Models (LLMs), the availability of annotated corpora remains crucial. State-of-the-art systems still exhibit shortcomings, for example, when access to context or pragmatics for giving a true comprehension of the features of the involved phenomena is required [1].

Unfortunately, the development of large datasets annotated for specifically complex phenomena can be very time-consuming. When only small corpora are available, data augmentation techniques can be applied [2, 3]. Given a small set of original sample data, data augmentation artificially generates new instances that are similar and comparable to the existing data and can, therefore, be used to train and test systems with an extended dataset.

In this paper, we present experiments for augmenting two small datasets annotated for two diverse, challenging phenomena, namely stereotypes and irony detection. In several works exploring data augmentation, Back-Translation (BT) [4] was shown to be a strong and relatively easy-to-implement baseline [5, 6]. A BT process generally consists of two steps: given one or multiple translation systems, a text in a source language is first translated into a chosen *pivot language*, and the resulting text is then translated back into the source language. The expected output of the BT process is a text that is similar but not the same as the original input, accounting for the linguistic differences intrinsic to the language pair, but also the idiosyncrasies of the chosen translation model(s). This relies on the fact that translation is only partially deterministic: the expected output should have the same meaning as the input, outputs that morphologically or syntactically differ may be considered as correct translations of the input. In BT, the application of (at least) two translations improves the variability between the input and the output text.

The usefulness of a dataset augmented by applying BT depends on the quality of the translated outputs. Outputs too similar to the inputs can cause overfitting when used for training, while with too different outputs, there is a risk of a shift in distribution that is too large, which may negatively impact performance, at least in intra-dataset evaluations. A compromise between these two alternatives must be found. Therefore, an evaluation of the quality of translations and back-translations is important to assess the benefits.

In this paper, we want to investigate the viability of BT as a data augmentation technique for low-resource tasks in various configurations. We use French and Italian as source languages — leveraging two multilingual datasets

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

The work of T. Bourgeade and S. Casola was performed while at Dipartimento di Informatica, Università di Torino, Turin, Italy.

tom.bourgeade@loria.fr (T. Bourgeade); s.casola@lmu.de (S. Casola); adel.mahmoudwizani@edu.unito.it (A. M. Wizani); cristina.bosco@unito.it (C. Bosco)
0000-0002-0247-3130 (T. Bourgeade); 0000-0002-0017-2975 (S. Casola); 0000-0002-8857-4484 (C. Bosco)

with subsets for these languages — and various languages as pivots for the BT process (French/Italian, English, and Arabic). We compare BT with an alternative process for data augmentation, specific for multilingual datasets, which we refer to as "cross-translation", where the data from one language subset is translated and then used as a data augmentation source for another language subset.

Our contributions are (1) an intrinsic qualitative human evaluation of translations and back-translations for stereotypes detection and irony detection datasets in various combinations of source and pivot languages, followed by (2) an extrinsic evaluation of machine learning model performance on these datasets, using these various data augmentation sources.

2. Related Work

BT as a data augmentation method was originally proposed by Sennrich et al. [4], in the context of Neural Machine Translation (NMT), to allow using monolingual data to improve translation quality, particularly when parallel (source and target) training data is scarce.

Since then, several works have explored BT, either as a baseline to evaluate other data augmentation methods against or as the primary augmentation method for low-resource tasks. For example, Kumar et al. [5] evaluated pre-trained conditional generative Transformer models as data augmentation sources and used BT as a baseline. They found that BT achieves relatively high extrinsic performance against simpler approaches such as Easy Data Augmentation (EDA) [7] but also against some Transformer models; it also obtains most of the best scores for semantic fidelity and data diversity.

Xie et al. [6] make use of BT as an augmentation strategy in their semi-supervised Consistency Training approach, in which a model is trained with a loss function combining traditional supervised learning on a limited amount of labeled data, with an unsupervised consistency loss. The latter consists of minimizing a divergence metric between the output distributions for an unlabeled input and a noised version of it, the noise function being the chosen data augmentation method, i.e., for text, BT.

As far as the challenges related to the application of translation to texts with irony or sarcasm, a few papers discussing this task were recently published, among which we can cite [8] and [9].

3. Datasets

We focus on the tasks of stereotypes and irony detection with relevant multilingual datasets. Table 1 summarizes the characteristics of their French and Italian splits, the chosen languages for this study:

Dataset	Lang.	Size (train; test; val)	Positive Class
StereoHoax	Italian	3123 (1841; 1185; 97)	15.11%
	French	9342 (6981; 1993; 368)	12.07%
MultiPICo	Italian	967 (619; 193; 155)	25.34%
	French	1724 (1104; 345; 275)	25.17%

Table 1
Statistics for the datasets used in this work.

• StereoHoax [10] is a contextualized multilingual dataset of tweets annotated primarily for the presence of anti-migrant stereotypes. It consists of replies to tweets containing racial hoaxes (RH), with each message having a "conversation head" (the message containing the source RH) and a direct parent message (if applicable).
• MultiPICo [11] is a disaggregated multilingual dataset of short social media conversations annotated for irony detection through crowdsourcing. Each instance is a *(post, reply)* pair, where the post is a starting message in a thread, and the reply is either a direct reply or a second-level reply.

4. Translation Model

To use BT as a data augmentation method, one crucial decision to make is that of the translation system(s) . Machine Translation (MT) models are in fact not explicitly designed to inject relevant noise into texts to increase the variety of data available. Therefore, a significant part of this beneficial noise will be linked to the idiosyncrasies of the chosen model(s).

In this work, due to the number of different configurations, and thus source-target language pairs we wished to investigate, we decided to limit our selection to intrinsically multilingual models.In a preliminary phase, we thus experimented with the locally runnable Transformer-based multimodal Neural MT model `SeamlessM4T v2` [12] proposed by Meta AI. However, after early evaluations of obtained translations and back-translations, we observed too many issues and violations of important criteria (see section 5). As such, we eventually selected the Google Translate API for our evaluation and experiments, as it seemed to offer the best tradeoffs between translation and back-translation quality, as well as ease of access to the languages chosen for this work (French, Italian, English, and Arabic). It is important to note, however, that the models used by Google Translate themselves make use of BT as a data augmentation technique, as well as M4 Modelling[1]: in practice, this may cause some issues for use in BT, as undesirable artifacts of BT and

[1]https://research.google/blog/recent-advances-in-google-translate/

Massively Multilingual Massive NMT — possibly caused by parameters bottlenecks or languages interferences [13] — may have detrimental effects on the quality of the augmented data.

5. Intrinsic Evaluation

To judge the viability of BT for these two datasets and languages, we perform a human qualitative evaluation of produced back-translations using the following protocol. First, we collect a set of data for both datasets and languages randomly sample 50 instances each for the French and Italian subsets, 25 from the positive class, and 25 from the negative class, for a total of 200 instances. For all the cases examined, we consider the text of the messages and the associated conversational context, which can consist of one or two other messages (an optional direct parent, and the conversation head/original post).

In addition to French and Italian as source and pivot languages, American English and Modern Standard Arabic were also selected on account of the linguistic expertise of the authors. Thus, for the 100 instances in Italian, we apply the following BT settings (<*source*> - <*pivot*> - <*target=source*>): Italian - English - Italian; Italian - French - Italian; Italian - Arabic - Italian. Similarly, for the 100 French instances, we apply the following BT settings: French - English - French; French - Italian - French; French - Arabic - French. We use the Google Translate API due to its ease of use and availability of the chosen source and target languages.

A manual qualitative approach is used for the evaluation of the BT results: 4 language experts (co-authors of this paper) evaluate the quality of the produced back-translations (and intermediate translations, though in a less quantitative capacity). All evaluators are native speakers of one of the source languages (French and Italian), as well as sufficiently proficient (or a native speaker) in the pivot languages (French, Italian, English, and Arabic). They are tasked with comparing the original and back-translated instances, also considering the pivot translation to help understand potential artifacts or errors introduced in the process. Evaluators could assign one label to problematic instances containing a violation of the following associated quality criteria:

- **faithfulness**: a faithful translation accurately conveys the meaning of the original text without introducing errors, omissions, or distortions. Since we focus on texts featuring expressions of stereotype or irony, faithful instances must also preserve these phenomena;
- **preservation of non-translatables**: this criterion is referred to in the translation of numbers, units, measurements, and, in general, non-translatable terms such as proper nouns, brands, trademarks, hashtags, user mentions, emojis, acronyms, and specific cultural references for maintaining clarity, consistency, and legal compliance. This category also includes idiomatic expressions which are especially difficult to translate;
- **fluency**: a text is fluent when it is perceived by a native speaker as reading "natural", in the way they would be expected to have structured it;
- **other**: this last criterion is used to report less frequent violations that cannot be encoded by the other criteria, including incomplete translations, word tokenization, or sentence segmentation.

5.1. Back-Translation Examples

To illustrate violations of these criteria, this section presents example parts of instances in their original (**Og**), translated (**Tr**), and back-translated (**BT**) forms, underlining the relevant spans, when applicable.

In the following example from the Italian subset of MultiPICo, the *fluency* criterion is violated because of the inadequate and unnatural back-translation of the plural expression *"per i primi tempi"* ("for the initial period"), into the singular *"per la prima volta"* ("for the first time"):

Og: "Se rimanere impiegato a 1400 euro è il tuo obiettivo ok, altrimenti è solo per i primi tempi"
Tr: "If staying employed at 1400 euros is your goal, ok, otherwise it's only for the first time"
BT: "Se restare impiegato a 1400 euro è il tuo obiettivo, ok, altrimenti è solo per la prima volta"

This example from French StereoHoax illustrates breaking the *faithfulness* criterion, with Arabic as the pivot language. In this message, the informal vulgar expression *"n'avoir rien à foutre"* (vulgar. "to have nothing to do"), which conveys an implied judgment of laziness towards the described target, cannot be properly translated into Arabic, like most vulgar expressions (a common issue with this pivot language), and loses its proper meaning in the back-translation, *"n'avoir rien à se soucier"* meaning "to have nothing to worry/care about":

Og: "Elle n'a rien à foutre"
Tr: " ليس لديها ما تهتم به "
BT: "Elle n'a rien à se soucier"

In this example from Italian MultiPICo, the violation concerns a *non-translatable*, in the form of the colloquial expression *"<X> della Madonna"*, intended as an idiomatic intensifier (similar to "A hell of a <X>" in American English). In the pivot translation, the idiom fails to be transposed, and "Madonna" is interpreted as part of the proper noun of a non-existent virus ("Madonna virus") and transposed into the back-translation:

Og: "... Gli asiatici stanno tramando qualcosa di losco.... prima gli spaghetti al microonde con ketchup e adesso un virus della madonna ?"

Tr: "... The Asians are up to something shady... first microwaved spaghetti with ketchup and now a Madonna virus?"

BT: "... Gli asiatici stanno tramando qualcosa di losco... prima spaghetti al microonde con ketchup e ora un virus Madonna?"

Another example of a *non-translatable* failing to be preserved is the following, taken from the French subset of StereoHoax. Here, the idiomatic expression *"se tuer/mourrir à la tâche"* (lit. "to kill oneself/die doing a task"), used in its informal variant with *"[se] crever"* (lit. "to burst", informal. "kill [oneself]/die") was translated incorrectly, changing the meaning of the message:

Og: "Oui mais est ce que c'est normal ? Quand yen a un qui a rien foutu et que l'autre s'est crever à la tache ? Non la logique c'est qu'il peuvent cumuler pour arriver à une retraite vivable et qui dépasse le seuil de pauvreté !"

Tr: "Yes but is this normal? When one has done nothing and the other has died? No, the logic is that they can accumulate to achieve a livable retirement that exceeds the poverty line!"

BT: "Oui mais est-ce normal ? Quand l'un n'a rien fait et que l'autre est mort ? Non, la logique est qu'ils peuvent accumuler pour obtenir une retraite viable qui dépasse le seuil de pauvreté !"

5.2. Samples Evaluation

Table 2 presents the quantitative results of this quality evaluation on 200 instances (see section 5). Cases that fall outside the selected criteria (classified under "other") include erroneous translations of grammatical gender, especially when using English as a pivot language, which has been extensively discussed in the literature [14]. Other errors refer to segmentation or punctuation. The preservation of proper punctuation and distinction between different sentences, text chunks, and segments ensures clarity and readability and can impact the quality of translation when using Machine Translation models. Unfortunately, due to the nature of the texts in question, i.e., social media messages, proper content segmentation is difficult to achieve due to the overall poor structure and formatting of the content in question (among many other forms of typographical artifacts and errors).

Regardless of the pivot language, some instances seem to be systematic sources of errors which can be explained by the particularities of the MT used model. For example, in MultiPICo Italian, one instance is *"Non la chiudono tranquillo"*, which should be interpreted as "They won't close it, don't worry" (speaking of the Italian Stock Exchange); however, for all pivot languages, and possibly due to the absence of a comma separating *"tranquillo"*, it is misinterpreted as an adverb and thus incorrectly back-translated to *"silenziosamente"* ("quietly"). Similarly, in MultiPICo French, a message discussing the increasing use of the idiomatic discourse marker/connector *"du coup"* (equivalent to connector "so" in English), has this quoted expression consistently mis-backtranslated to *"tout d'un coup"* ("all of a sudden/suddenly"), despite it not making sense in the context of the message. The use of the expression in quotation marks in this case may have confused the MT model, which otherwise does not struggle with this expression when manually tested.

Overall, English appears to perform best across all the pivot languages in all settings. This is not surprising considering that, for most MT models, English is the most represented language in the training data (both in the source and target language), as well as the language typically used as a pivot to generate augmented instances for lower-resource languages. When using Arabic as a pivot language in our evaluations, we observed some unnatural expressions and constructs that appear "borrowed" from English: for example, in a MultiPICo Italian instance, the word *"gratis"* ("free [of charge/cost]") is mistranslated to حرية ("freedom/liberty"); we thus hypothesize that the MT model used English as a pivot language for the Italian-Arabic language pair, as both terms would indeed likely be mapped to the polysemic and thus ambiguous term "free" in English.

6. Extrinsic Evaluation

To evaluate the effectiveness of BT as a data augmentation method for stereotypes or irony detection, we performed some preliminary experiments with varying configurations. For these experiments, we used the XLM-RoBERTa [15] multilingual Transformer classifier: while for smaller models, monolingual Transformers are generally preferable to multilingual ones, we preferred to use a single model in all configurations. For similar reasons, and due to time and resource constraints, for all experiments, we only automatically fine-tuned the hyperparameters of the models once for each dataset and source language combination (with a total of 4 starting configurations), on the *baseline* training set, that is, without any data augmentation. For more technical details, see Appendix A.

As the positive class (stereotype or irony present) is often the minority class for these and related tasks (see Table 1), we evaluate "balanced" data augmentation con-

BT-setting	faith	n-trs	fluency	other
Ita-Eng-Ita	16%	8%	4%	2%
Ita-Fra-Ita	26%	6%	4%	4%
Ita-Arb-Ita	36%	8%	4%	2%
mean	27%	7%	4%	3%
Fra-Eng-Fra	18%	14%	0%	0%
Fra-Ita-Fra	28%	14%	0%	0%
Fra-Arb-Fra	36%	12%	0%	2%
mean	27%	13%	0%	1%

(a) MultiPICo Back-Translation errors

BT-setting	faith	n-trs	fluency	other
Ita-Eng-Ita	22%	4%	8%	0%
Ita-Fra-Ita	24%	4%	12%	2%
Ita-Arb-Ita	44%	12%	8%	0%
mean	30%	7%	9%	1%
Fra-Eng-Fra	18%	6%	4%	0%
Fra-Ita-Fra	36%	4%	6%	0%
Fra-Arb-Fra	18%	20%	10%	0%
mean	24%	10%	7%	0%

(b) StereoHoax Back-Translation errors

Table 2
Distribution of translation-related errors (**faith**: faithfulness, **n-trs**: non-translatable; see section 5) in 50 sample instances (25 of each class) of each dataset, for all combinations of source and pivot languages (**BT-setting**).

Dataset	Source	*baseline*	OV	BT[Eng]	BT[Fra]	BT[Arb]	XT	BT[Eng]|OV	XT|OV
StereoHoax	Ita	75.44	74.98	74.29	74.34	75.96	46.55	74.58	**76.18**
	Fra	**68.05**	67.36	55.73	64.12	60.8	64.43	65.68	65.85
MultiPICo	Ita	68.21	65.23	65.71	63.56	**68.49**	65.79	61.86	63.48
	Fra	59.73	64.7	64.01	61.24	63.28	64.91	64.09	**65.17**

(a) Results in terms of Macro F1-score.

Dataset	Source	*baseline*	OV	BT[Eng]	BT[Fra]	BT[Arb]	XT	BT[Eng]|OV	XT|OV
StereoHoax	Ita	56.13	56.06	54.55	54.48	**57.55**	0.00	55.36	57.14
	Fra	**43.48**	42.89	34.43	39.75	36.09	39.74	39.84	42.63
MultiPICo	Ita	54.55	46.67	**55.22**	47.71	53.47	48.42	44.86	42.86
	Fra	37.09	45.57	**49.51**	47.53	48.80	48.94	49.00	48.62

(b) Results in terms of Positive class F1-score.

Table 3
Results of our experiments for various data augmentation configurations (see section 6). The best scores for each configuration are highlighted in **bold**.

figurations, in which augmented samples are added to the positive class until it is the same size as the negative class. We evaluated the following configurations:

- *baseline*: the model is trained on the original, unmodified training set (with no balancing of the classes).
- *oversampling* (OV): Oversampling was shown to be a strong baseline in various previous works [16, 17], and we thus evaluate it as an alternative or complement to BT.
- *back-translation from <language>* (BT[<language>]): augmented instances are sampled from back-translations of the original data using <language> as a pivot.
- *cross-translation* (XT): as the datasets used are multilingual and contain subsets in both French and Italian, one language's subset can be translated and used as augmented data for the other.
- *mixed back/cross-translation with oversampling* (BT[<language>]/XT|OV): as the positive classes are, for both phenomena and all languages, less than half the size of the negative class, balancing the two requires sampling more instances from the data augmentation source than there are original positive instances, which could result in injecting translation related biases into the training set. To attempt to mitigate this, we also evaluate sampling 50% from back or cross-translation strategies, with 50% from oversampling the positive class. Note that, given the number of potential configurations, we only evaluate BT[Eng]|OV and XT|OV due to time and resource constraints.

Table 3 displays the results of our experiments in terms of macro F1-scores, as well as positive class F1-scores. Except for StereoHoax French, at least one of the data augmentation configurations outperforms the baseline, though not necessarily BT. Indeed, for both StereoHoax Italian and MultiPICo French, the mixed cross-translation with oversampling (XT|OV) configuration achieves the highest Macro F1-score, though not the best positive class score. This seems to indicate that the variety of data

intrinsic to using a separate language subset of a multilingual dataset can be beneficial, when possible, over that artificially created by a data augmentation technique like BT. Additionally, we only experimented with cross-translation within one linguistic typology (Romance languages). As such, future investigations on whether this extends to cross-typologies XT would be worth pursuing.

Interestingly, we find that the mixture of oversampling and back/cross-translation outperforms the equivalent non-mixed configuration for all datasets and languages except MultiPICo Italian. However, due to its small size (see Table 1), the results on this particular subset may be less significant, given the overall protocol for these experiments, and a protocol that can inject greater amounts of augmented data might be preferable. During initial experiments, however, we found that injecting larger quantities of augmented data (preserving or not the initial label distributions) seemed to consistently negatively impact test-set performance, most likely due to overfitting but also possibly due to the models fitting on the translation model detrimental idiosyncrasies, instead of the characteristics of the phenomena to detect.

Moreover, the performance on the positive class (Table 3b) is not necessarily improved correspondingly with the overall macro F1-score (Table 3a), even when the augmentation is applied solely to this class. In other works on similar phenomena, it is shown that data augmentation and related methods can boost the Out-of-Domain performance of such detection models [17]. The addition of variety in the occurrences of the phenomenon to detect would indeed help in generalizing its detection to other sources of data. Though, as the example of Stereo-Hoax Italian in the cross-translation (XT) configuration shows, care should be taken not to overly shift the data distribution; otherwise, models may fail to learn the particular dataset's positive class entirely. The mixed data augmentation with oversampling configurations seems, however, successful in addressing this potential issue, though more variations in the proportions should be experimented with.

7. Conclusions

In this work, we have investigated using Back-Translation as a data augmentation technique for challenging low-resources tasks like stereotypes and irony detection, in a multilingual context.

Through an intrinsic evaluation of the quality of the augmented instances, we identified modes of failure of Machine Translation, which could negatively impact the data augmentation process. These errors stem from the intrinsic differences between typologies and specific languages or translation model idiosyncrasies themselves potentially learned from methods like BT. Through a preliminary extrinsic evaluation of two multilingual datasets, we found that cross-translation can outperform Back Translation, allowing us to augment one language subset by leveraging the variety of inputs present in the others.

In future work, we aim to expand this study to more numerous and varied source and pivot languages, and different data augmentation configurations, namely, different proportions and selections of injected augmented data. We may also compare Back and Cross-Translation against or alongside other related techniques, such as multitasking learning or Active Learning. We also expect that some improvements can be obtained by mitigating translation failures; this can be done, for example, by leveraging an external LLM to check each step and remove or correct the errors from the final augmented dataset. Finally, it could be also interesting to perform tests with different model types on top of RoBERTa.

Acknowledgment

The work of T. Bourgeade was funded by the project StereotypHate, funded by the Compagnia di San Paolo for the call 'Progetti di Ateneo - Compagnia di San Paolo 2019/2021 - Mission 1.1 - Finanziamento ex-post'.
The work of C. Bosco was partially funded by this same project.

References

[1] S. Menini, A. P. Aprosio, S. Tonelli, Abuse is Contextual, What about NLP? The Role of Context in Abusive Language Annotation and Detection, 2021. URL: http://arxiv.org/abs/2103.14916. doi:10.48550/arXiv.2103.14916. arXiv:2103.14916.

[2] M. Bayer, M.-A. Kaufhold, C. Reuter, A Survey on Data Augmentation for Text Classification, ACM Computing Surveys 55 (2022) 146:1–146:39. URL: https://dl.acm.org/doi/10.1145/3544558. doi:10.1145/3544558.

[3] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, E. Hovy, A Survey of Data Augmentation Approaches for NLP, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 968–988. URL: https://aclanthology.org/2021.findings-acl.84. doi:10.18653/v1/2021.findings-acl.84.

[4] R. Sennrich, B. Haddow, A. Birch, Improving Neural Machine Translation Models with Monolingual Data, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1:

Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 86–96. URL: https://aclanthology.org/P16-1009. doi:10.18653/v1/P16-1009.

[5] V. Kumar, A. Choudhary, E. Cho, Data Augmentation using Pre-trained Transformer Models, in: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, Association for Computational Linguistics, Suzhou, China, 2020, pp. 18–26. URL: https://aclanthology.org/2020.lifelongnlp-1.3.

[6] Q. Xie, Z. Dai, E. Hovy, T. Luong, Q. Le, Unsupervised Data Augmentation for Consistency Training, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 6256–6268. URL: https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

[7] J. Wei, K. Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 6382–6388. URL: https://aclanthology.org/D19-1670. doi:10.18653/v1/D19-1670.

[8] H. Ardi, M. Al Hafizh, I. Rezqy, R. Tuzzikriah, Can machine translations translate humorous texts?, Humanus 21 (2022) 99–112.

[9] Initial exploration into sarcasm and irony through machine translation, Natural Language Processing Journal 9 (2024) 100106.

[10] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A Multilingual Dataset of Racial Stereotypes in Social Media Conversational Threads, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 686–696. URL: https://aclanthology.org/2023.findings-eacl.51.

[11] S. Casola, S. Frenda, S. M. Lo, E. Sezerer, A. Uva, V. Basile, C. Bosco, A. Pedrani, C. Rubagotti, V. Patti, D. Bernardi, MultiPICo: Multilingual Perspectivist Irony Corpus, in: Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2024.

[12] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim, J. Hoffman, M.-J. Hwang, H. Inaguma, C. Klaiber, I. Kulikov, P. Li, D. Licht, J. Maillard, R. Mavlyutov, A. Rakotoarison, K. R. Sadagopan, A. Ramakrishnan, T. Tran, G. Wenzek, Y. Yang, E. Ye, I. Evtimov, P. Fernandez, C. Gao, P. Hansanti, E. Kalbassi, A. Kallet, A. Kozhevnikov, G. M. Gonzalez, R. S. Roman, C. Touret, C. Wong, C. Wood, B. Yu, P. Andrews, C. Balioglu, P.-J. Chen, M. R. Costa-jussà, M. Elbayad, H. Gong, F. Guzmán, K. Heffernan, S. Jain, J. Kao, A. Lee, X. Ma, A. Mourachko, B. Peloquin, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, A. Sun, P. Tomasello, C. Wang, J. Wang, S. Wang, M. Williamson, Seamless: Multilingual Expressive and Streaming Speech Translation, 2023. URL: http://arxiv.org/abs/2312.05187. doi:10.48550/arXiv.2312.05187. arXiv:2312.05187.

[13] A. Mueller, G. Nicolai, A. D. McCarthy, D. Lewis, W. Wu, D. Yarowsky, An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3710–3718. URL: https://aclanthology.org/2020.lrec-1.458.

[14] E. Rabinovich, S. Mirkin, R. Patel, L. Specia, S. Winther, Personalized machine translation: Preserving original author traits, in: Proceedings of the EACL 2017 vol. 1 long papers, 2017.

[15] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[16] M. Juuti, T. Gröndahl, A. Flanagan, N. Asokan, A little goes a long way: Improving toxic language classification despite data scarcity, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2991–3009. URL: https://aclanthology.org/2020.findings-emnlp.269. doi:10.18653/v1/2020.findings-emnlp.269.

[17] C. Casula, S. Tonelli, Generation-Based Data Augmentation for Offensive Language Detection: Is It Worth It?, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3359–3377. URL: https://aclanthology.org/2023.eacl-main.244. doi:10.18653/v1/2023.

eacl-main.244.

[18] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[19] L. Biewald, Experiment tracking with weights and biases, 2020. URL: https://www.wandb.com/, software available from wandb.com.

[20] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, A. Talwalkar, Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization, Journal of Machine Learning Research 18 (2018) 1–52. URL: http://jmlr.org/papers/v18/16-558.html.

A. Technical Details

For all experiments, we used the XLM-RoBERTa-base as provided by the the HuggingFace `transformers` [18] ecosystem (including the `datasets` library for data processing).

Automatic hyperparameters fine-tuning was accomplished using the Weights & Biases [19] AI platform's Bayesian hyperparameters optimization system, with the Hyperband early-stopping algorithm [20]. As mentioned in section 6, only 4 such optimizations were executed, one for each language subset of each dataset, in the *baseline* configuration (no data augmentation).

The learning rate (lr), the hardware training batch size (bs), and the number of gradient accumulation steps (ga), were automatically fine-tuned, and their final values are listed in Table A1. These models were trained for a maximum of 10 epochs, with the best performing epoch checkpoint kept at the end (measured by macro F1-score), with a warm-up ratio of 0.2 (linear warm-up from 0 to the initial learning rate over 20% of the training set), both determined during initial experiments.

Automatic fine-tuning and training of the models was performed on the Google Colab platform, using high-RAM T4 GPU instances, for an approximate total of 50 GPU-hours.

Dataset	Lang.	*lr*	*bs*	*ga*
StereoHoax	French	2.963E-05	16	4
	Italian	1.000E-06	16	1
MultiPICo	French	2.963E-05	16	4
	Italian	2.920E-05	8	1

Table A1
Automatically fine-tuned hyperparameters (***lr***: learning rate; ***bs***: batch size; ***ga***: gradient accumulation steps)