Received: 4 July 2023 | Revised: 29 January 2024 | Accepted: 1 February 2024

DOI: 10.1111/exsy.13568

ORIGINAL ARTICLE

Expert Systems WILEY

Backtranslate what you are saying and I will tell who you are

Marco Siino[1] | Francesco Lomonaco[2] | Paolo Rosso[3,4]

[1]DEI Department, University of Bologna, Bologna, Italy

[2]Department of Psychology, University of Milan Bicocca, Milano, Italy

[3]PRHLT Research Center, Universitat Politècnica de València, Valencia, Spain

[4]ValgrAI–Valencian Graduate School and Research Network of Artificial Intelligence, Valencia, Spain

Correspondence
Marco Siino, Department of Engineering, University of Palermo, Palermo, Italy.
Email: marco.siino@unipa.it

Funding information
European Union NextGenerationEU/PRTR, Grant/Award Number: PCI2022-135008-2; ERDF, Grant/Award Number: PID2021-124361OB-C31

Abstract

With this work, we hypothesize that semantically enriching a user's text corpus using backtranslation and expansion modules can improve performance for author profiling tasks. To perform this textual enrichment, we translate an author's representative text. Translations are made from one language—the source language—into another—the target language—and then back to the original one. Finally, we expand an author's text by integrating the original version with the back-translated one. Our framework includes these backtranslation and expansion modules followed by a SOTA classifier successfully employed for text classification. The framework is tested on three author profiling datasets from the last three years' shared tasks on fake news, hate speech, irony and stereotypes detection hosted at the CLEF conference for the PAN Lab. This work is an extension of our previous one where we just presented our main idea. Here we improve our framework, and we also investigate more languages and more datasets. Finally, a qualitative analysis is provided. The results confirm that the back-translation and expansion add-on modules improve model performance on all three datasets evaluated.

KEYWORDS

author profiling, convolutional neural network, data augmentation, fake news, hate speech, irony, stereotypes, Twitter

1 | INTRODUCTION

Social media's ascent, which nowadays dominates the information and entertainment arena all around the world, has revolutionized the way people communicate online (Joo & Teng, 2017; Subramanian, 2017). However, it is possible that the latent information in this form of communication is not always explicit in the text, which could hinder the performance of NLP classification models. Data Augmentation (DA) is a technique that can generate an alternative representation of the input and eventually improve model performance. Therefore, uncovering this latent information could lead to better results in author profiling tasks (Mangione et al., 2022). In this paper we integrate and explore the concept of backtranslation (Brislin & Freimanis, 1995) and we propose a novel module, to highlight and uncover latent information available in an author's text to improve text classification performance.

Studies presented in Ozolins et al. (2020), Shleifer (2019), Lee et al. (2021) have shown that backtranslation can be employed as a powerful tool for expanding samples in NLP-related tasks. *Round-trip* (or *Back-and-forth* or simply *back*) translation entails converting spoken or written samples from one language into another and then back again. Moreover, to increase the size of a dataset for machine learning and NLP tasks, DA

Abbreviations: AP, author profiling; DA, data augmentation; FNS, fake news spreader; HSS, hate speech spreader; ISS, irony and stereotype spreader.

is a widely utilized approach (Hayashi et al., 2018). This method has been shown to be particularly effective in leveraging the semantic differences between languages and improving the representation of the input (Beddiar et al., 2021; Body et al., 2021). In this study, we specifically focus on a novel strategy of DA. In fact, thanks to the backtranslation module in our framework, we are able to augment each sample while maintaining the same number of dataset samples.

In our setting, each sample is a user's corpus of texts (a Twitter feed), and we hypothesize that semantically enriching the user's text corpus using our proposed modules can improve classification performance. By augmenting each sample with one or multiple translations, we aim to increase the diversity and informativeness of the data to improve the representation of the input, ultimately leading to better classification performance of different NLP models. In this paper, focusing on Author Profiling (AP) tasks, we investigate the effectiveness of backtranslation for expanding samples using English as the original source language and Italian, German, Japanese, and Turkish as the target languages. In a previous work, the Italian was investigated as a target language and only using a single dataset. The domain was related to irony and stereotype detection, and the authors highlighted promising performance compared to the not augmenting version of the framework (Mangione et al., 2022). German was adopted as one of the backtranslation languages by the winner of the "Toxic Comment Classification Challenge"[1] while Japanese and Turkish were chosen for their belonging to different linguistic families. The proposed framework is evaluated through a three-stage empirical experiment. First, a baseline of AP models is established using datasets without our augmentation modules. The second stage involves generating augmented data using backtranslation from English to target languages with one or multiple augmentations and then back to English. The backtranslated sample is then concatenated to the original one. In the final stage, a machine learning model is trained using the enriched data, and its performance is compared with and without the backtranslation module. We evaluate our framework on three different AP datasets (regarding, namely, fake news, hate speech, and irony and stereotypes spreaders). Our results outperform the not augmented baseline, showing that the expansion of samples with multiple languages using backtranslation leads to improved performances in AP tasks. Thanks to the backtranslation module, our framework is able to outperform the results obtained without expanding the samples. All the code used in this article is available on GitHub.[2]

The remainder of the article is organized as follows. Section 2 investigates the relevant literature about backtranslation and AP. In Section 3 we present and discusses the architecture comprehensive of the augmentation module, the expansion module and the classifier. Section 4 presents the experimental evaluation. Here, datasets and classifiers used for comparison are described. Finally, the experimental setup is documented. In Section 5 the results of the experiments are presented, and their implications are discussed. Finally, Section 6 presents our conclusions and discusses some future works.

2 | RELATED WORK

We examine prior studies related to DA techniques (including backtranslation), and AP in this section. We highlight the contributions and limitations of each approach.

2.1 | Data augmentation and backtranslation

This work uses the idea of DA as a general concept that refers to any technique that enriches training and test data (Bayer et al., 2023). A former issue related to DA is *label preservation*, which refers to sample data modifications that maintain class label. To enhance the size of a training dataset for machine learning and NLP problems, DA strategies are frequently utilized. DA is also used to improve the representation of every single sample (Xie et al., 2020). To the best of our knowledge it has not been applied in the three tasks we addressed: profiling fake news, hate speech, irony and stereotypes spreaders. By generating additional samples, particularly when there is a dearth of data, DA can reduce over-fitting and increase the robustness of NLP models. According to Banko and Brill (2001), the choice of classifier does not significantly alter the solution quality in the confusion set disambiguation problem; only the integration of new data can achieve this. The present work falls under the category of backtranslation which is an approach to automatically and randomly performing several operations on text (i.e., random deletion, random swap, random insertion, and synonym replacement) with the help of translation models (Bayer et al., 2023). This choice is motivated by the fact that text translations highlight relevant concepts, not so exposed because of the intricacy of the actual language. The work of Beddiar et al. (2021) finds that DA using backtranslation is an effective methodology. They use DA to increase the size of the dataset, using also a paraphrasing module. In Body et al. (2021) the error rate is reduced by up to 3.4% with statistical significance using binary sentiment classification models. They concatenate a backtranslation to the original samples, and the resulting text is given as input to a recurrent model. Chen et al. (2020) proposed the *MixText (TMix) system* as an expansion strategy. The intuition behind their work is that a new sentence that combines the meaning of the original and enhanced sentences is produced by interpolating the hidden states of the original two sentences. The best performance of an LSTM classifier is the result of random synonym insertion, random deletion, random swap and round-trip translation that are individually integrated and tested in the work of Bonthu et al. (2022).

2.2 | Author profiling

Author Profiling (AP) is a challenging task of growing interest. AP can be applied to infer the age, gender, or other user-specific features based on a corpus of texts, also the personality traits that were related to linguistic features can be detected (Pennebaker et al., 2003). Sharmila Devi and Kannimuthu (2023) proposes a dataset composed of WhatsApp messages code-mixed in the Tamil language. From the standpoint of forensic linguistics, it could be very useful for assessing suspects to identify the linguistic profile of the author of a questionable text purely by text analysis. Furthermore, the ability to detect users prone to spread hate speech messages on social media could prevent the dissemination of such content in advance. Similarly, to prevent the spreading of fake news, a classifier could identify authors who write fake content and rumors online.

To evaluate our proposed framework in real-world contexts, we focus on three different AP tasks of growing interest: fake news, hate speech, and irony and stereotypes spreaders. These three AP tasks, along with the related works, are presented and discussed in the rest of this subsection.

2.2.1 | Fake news

Fake news spreading has been recognized as one of the key issues regarding social media (Rangel et al., 2020). User features can be used to profile authors that disseminate false information (Unsvåg & Gambäck, 2018). The automatic detection of different author profiles from a corpus of text is a task that gained momentum, especially after the COVID-19 pandemic. Leonardi et al. (2021) solve this task related to fake news spreaders by expanding the CoAID dataset presented in Cui and Lee (2020). The scholars offer a stacked Transformer-based neural model that combines a deep learning model with the ability of the Transformer to compute language embeddings. To shift the problem from single fake news to an AP task, fake news spreaders are individuals who spread false or misleading information with the intention of deceiving others. AP can be used to identify these individuals and understand the characteristics of those messages that are considered fake news. By understanding the motivations and characteristics of fake news spreaders (Buda & Bolonyai, 2020), scholars can develop strategies to combat the spread of misinformation and promote more informed online conversations. In Giachanou et al. (2022), the authors profile fake news spreaders using psycholinguistic and linguistic characteristics as input to a CNN. The outcomes of their experiments demonstrate that their suggested model correctly categorizes users as fake news spreaders. Rangel et al. (2020) discusses an earlier version of the AP task where using writers' most recent tweets, the aim is to spot those who are likely to propagate false information. The work of Siino, Di Nuovo, et al. (2022) suggests that when tackling a text classification challenge like the identification of fake news spreaders, pre-trained deep models like Transformers are not necessarily the best option. In their work, the authors propose a CNN able of outperforming pre-trained and traditional machine learning models on the evaluated PAN test set (Rangel et al., 2021).

2.2.2 | Hate speech

Several types of harmful messages are shared every day on social media (Lomonaco et al., 2022). The profiling Hate Speech Spreaders (HSSs) on Twitter task's purpose, hosted at PAN 2021, was to find out whether the person who wrote a certain Twitter thread was likely to propagate hate speech (Rangel et al., 2021). We opted to take the work done in Zhang and Wallace (2017) as a starting point for developing our model. In the work, a CNN was initially used for a text categorization problem. There are hybrid techniques in the literature as well, such as using an SVM for classification and prediction and a CNN to extract textual features (Wang & Qu, 2017). Obviously, hate speech is not always conveyed in an aggressive and explicit way, so also other aspects of discourse must be investigated such as stereotypes (Sanguinetti et al., 2020). Recent work by Kumar et al. (2023) proposed an Autoencoder as a feature extractor that feeds the representation to an SVM to classify each sample in the HSS dataset. In the English HSS dataset, they reach an accuracy of 95$\%$.

2.2.3 | Irony and stereotypes

While several techniques to address irony detection are discussed in Sánchez-Junquera, Rosso, et al. (2021), Sulis et al. (2016), in Croce et al. (2022), Sánchez-Junquera, Chulvi, et al. (2021) every day, new methods for the detection of stereotypes are proposed. However, in order to develop our model, we looked at the top-performing models available together with the ones that were taking part in the PAN-shared tasks (Croce et al., 2022). We also leverage the previous work of Siino, Tinnirello & La Cascia (2022), where an ensemble (named T100), composed of a CNN, a SVM, a Decision Tree DT and a Naive Bayes NB classifier provide the input for a final layer that uses an SVM, to determine the author as ISS or not (nISS).

3 | THE PROPOSED FRAMEWORK

The main components of our proposed framework are three, and they are shown in Figure 1 and discussed in this section. It is worth mentioning that the original input sample passes through the same framework during both the training and test phases. While each component is further discussed in the following subsection, here we introduce all the steps performed as shown in the Figure 1. The input sample is provided as input to the backtranslation module. The backtranslation can be performed using one or more target languages. Then the backtranslated sample is merged with the original one using the expansion module. Finally, the newly expanded sample is provided to the classifier, which provides the final prediction. As already stated, each input sample passes through the pipeline of our framework for both the training and the inference stages.

3.1 | Backtranslation module

The proposed augmentation module is a tool that has been designed to enhance and eventually highlight content relevant to the classification task. Text data is translated into a different language, and then it is translated back into the original language as part of the backtranslation augmentation. Instead of retaining the original context and meaning, this technique creates new textual data with distinct phrases from the original text. To perform the backtranslation in this study, we used the Google Translate API.[3]

The augmentation module is also composed of several subcomponents to pre-process each sample (i.e., all the authors-tag and open-close document tags have been removed). This pre-processing ensures that any irrelevant or noisy text is removed from the sample (Siino et al., 2024). Next, the sample is translated using the translator, which converts the sample into a different language. The backtranslation process is then carried out, which involves translating the text back to its original language (in our case English), with the aim of enriching the semantic content of the text. It is worth noting that the backtranslation could be performed in more than one language. As shown in Figure 1, a sample can be eventually backtranslated using different target languages in parallel. In this case, all the backtranslated versions of the sample are provided to the following expansion module.

3.2 | Expansion module

Inside the expansion module, the backtranslated samples are concatenated with the original one, and the augmented sample is generated. Again, it is worth repeating that also this process applies in both the training and test phase. While in previous works as Beddiar et al. (2021) only a single backtranslation (i.e., with just a single target language) is used, in our framework we allow several parallel backtranslation layers to perform the translation toward one or more target languages. In this case, the expansion module merges the text from the original sample with all the

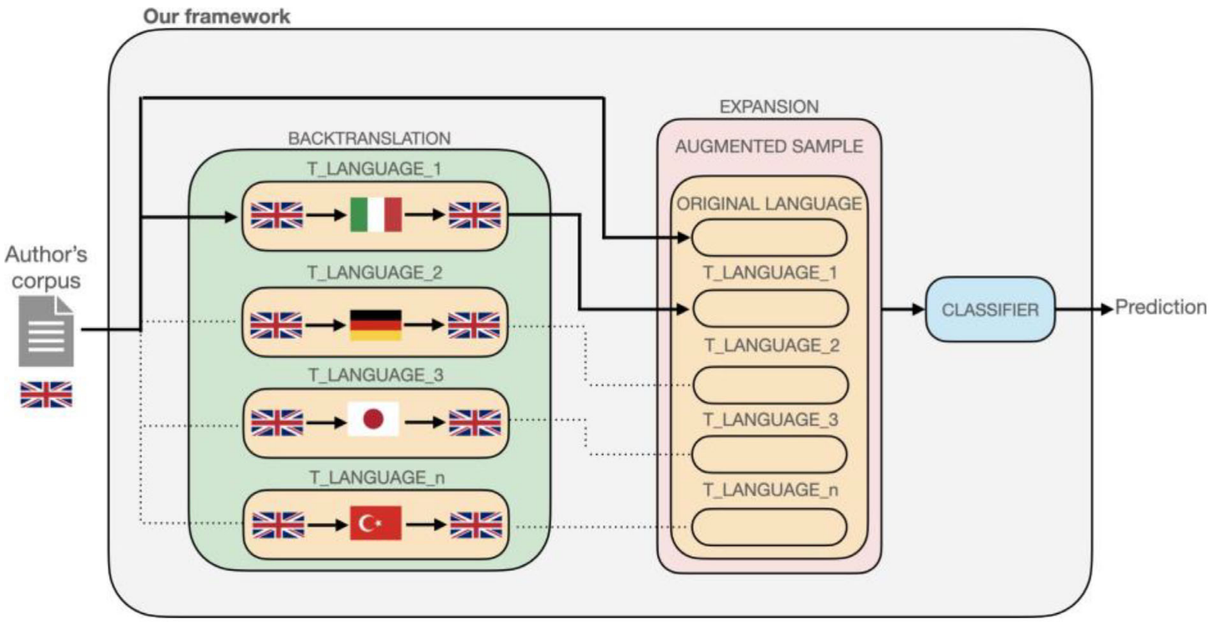


FIGURE 1 Our proposed framework. In the figure, we basically show a backtranslation using the Italian language. It is optional to include other target languages to further improve the quality and the quantity of information available per each sample in the dataset. As shown in the figure, the languages that we used for our experiments were: Italian, German, Japanese, and Turkish.

backtranslated versions. Considering the case where four target languages are used after the expansion module, the length of the expanded sample is around five times more than the original one.

3.3 | Classifier

After the expansion module, the augmented sample is used to train a classifier and also to test its performance. Several state-of-the-art classifiers can be employed in our framework. To evaluate and assess the performance of the two previous modules, we employed four different classifiers.

The classifiers were **RoBERTa**, **GPT-2**, SVM and a CNN. The results were reported in Section 5, and a comparison was made between training on the original and on augmented datasets in each of the four selected languages, and using all of them. As the datasets are balanced between classes' sample sizes, accuracy was chosen as the evaluation metric. As also discussed in Siino, Di Nuovo, et al. (2022), Siino et al. (2021), Yu et al. (2022), Siino, La Cascia, and Tinnirello (2022), CNN-based architectures are able to reach SOTA performance and, also in this study, the CNN model has proved to be the top-performing model over the three different datasets.

4 | EXPERIMENTAL EVALUATION

4.1 | Datasets

In this subsection, we discuss the three datasets used to perform our experiments. The three datasets have been introduced for the three AP tasks held for PAN@CLEF[4] in 2020, 2021 and 2022. They are namely the *Fake News Spreaders (FNS)*, *Hate Speech Spreaders (HSS) and Irony and Stereotype Spreaders (ISS)* datasets. Statistics on the three datasets are shown in Table 1.

4.1.1 | FNS

The FNS dataset was proposed in Rangel et al. (2020). The dataset consists of 300 Spanish (ES) and English (EN) Twitter user accounts, each. Only tweets in English were taken into account for this investigation. Fact-checking was done on the manually gathered tweets. The balanced nature of the dataset is due to the same number of class instances for both labels in the dataset. Authors of half the documents in each language folder have been identified as spreading false information. The remaining 50% are texts from accounts that have not twitted fake news in the past.

4.1.2 | HSS

The multilingual dataset, which includes English and Spanish datasets provided by the task's organizers (Rangel et al., 2021), is made up of 120,000 worth of tweets, with 200 worth of tweets per author, 200 worth of authors for each language training set, and 100 worth of authors for each language test set. Only tweets in English are taken into account in this paper.

4.1.3 | ISS

The PAN organizers' dataset, described in Ortega-Bueno et al. (2022), includes a group of 600 Twitter authors. A set of 200 tweets is given to each author. An author's feed of 200 tweets is contained in a single XML file. 420 worth of authors are included in the organizers' labeled training set. The rest of the 180 ones make up the test set. The training set labels authors as “I” (ISS) or “NI” (nISS), depending on whether they use irony.

TABLE 1 Characterization and statistics of the datasets.

Dataset	Topic	#Total documents	#Documents per sample	#Train samples	#Test samples
FNS	Fake news	50,000	100	300	200
HSS	Hate speech	60,000	200	200	100
ISS	Irony and sarcasm	120,000	200	420	180

4.2 | Backtranslation languages

The proposed framework's performance is assessed considering different languages chosen accordingly to previous works available in the literature.

This work is an extension of previous works (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2023; Siino & Tinnirello, 2023). In our first study (Mangione et al., 2022), we used only Italian as target language for backtranslation and we did not consider other languages. Here we evaluate also other languages and more datasets to further investigate the performance of the framework and to conduct a qualitative analysis. Here we also use Italian as the target language, but we also use it in parallel with other languages. German is the second language we use for our study. This is due because German has been employed as a target language for the backtranslation in several other studies (Beddiar et al., 2021; Behr, 2017; Edunov et al., 2018; Hoang et al., 2018). Furthermore, we want to investigate the performance using two additional languages with *subject-object-verb* words order and very distinctive characteristics in contrast with the structure of Italian and German. They are, namely, the Turkish and the Japanese. Turkish language characteristics include vowel harmony and significant agglutination. Turkish's usual word order is subject-object-verb. Noun classes or grammatical gender do not exist in Turkish. The usage of honorifics in the language makes a clear contrast between levels of courtesy, social distance, age, and familiarity with the addressee. The second-person pronoun and verb forms that relate to one individual are plural. Japanese is a mora-timed an agglutinative language with pure vowels, a phonemic vowel and consonant system, and a pitch-accent that has lexical significance. Normal sentence structure is topic-comment and subject-object-verb, with particles designating the grammatical function of words is the typical word order. The use of sentence-final particles can create inquiries or add emotive or dramatic emphasis. There is no article, no grammatical gender, and no number for nouns. Verbs are conjugated, but not for person, but rather for tense and voice. Adjectives in Japanese can be conjugated as well. Japanese has a sophisticated honorific system that uses verb forms and vocabulary to denote the speaker, listener, and other people's relative position.

4.3 | Classifiers

To test the effectiveness of backtranslation, together with the top-performing CNN, we tested other models, belonging to the set of deterministic and pre-trained models respectively. A brief discussion of all the models is provided in the rest of this section.

1. **CNN** classifies data using a single convolutional layer, a max-pooling layer, and a linear layer in our implementation. The architecture is the same as discussed in Siino et al. (2021). A single 1D-convolution layer is used in this model. There are sixty-four filters of size thirty-six in this layer. The layer then applies convolution to 36-ngram windows with a stride value of 1, which shifts the convolutional filter by one word embedding tensor for each convolution. The activation function used on the output is ReLu and no padding is used. The final output of the linear layer that follows the global average pooling is a single float value.
   A zero-threshold value determines the samples' label to compute the accuracy of the model.
2. **Robustly optimized BERT approach (RoBERTa)** uses a bidirectional encoder to produce contextualized word embeddings and belongs to the class of pre-trained Transformer model. The fine-tuning of the model is performed for the task, which in this case is AP. For this model, presented in Liu et al. (2019), authors conducted a replication study on BERT pre-training and achieved better performance by making modifications to the pre-train phase of a BERT model. These modifications included training the model longer with larger batches, removing the next sentence prediction objective, training on longer sequences, and dynamically changing the masking pattern applied to the training data.
3. **Generative Pre-trained Transformer 2 (GPT-2)**. OpenAI developed GPT-2 (Radford et al., 2019), an open-source large language model. GPT-2 can provide replies to inquiries, translating text, summarizing sections, and producing text. Since it is a general-purpose learner, it was not specifically taught any of these tasks, and its ability to complete them is an extension of its general ability to accurately synthesize the next item in any given sequence.
4. **SVM** is a linear classifier that aims to find the best hyperplane to separate different classes in a high-dimensional space. Based on Chang and Lin (2011), we tested the sklearn SVC implementation.[5] We used a linear kernel type with a value of 0.5 as a regularization parameter

Using different models, to test our augmentation modules, also allowed us to assess the usability of the framework. Different models have different strengths and weaknesses, and they may perform differently on different datasets or tasks. By testing the augmentation module on multiple models, we can better understand its effectiveness and limitations in different scenarios. Moreover, the usage of multiple runs for each model can help to reduce the impact of random initialization and provide more robust evaluation results.

4.4 | Experimental setup

On a Tesla T4 from Google Cloud and an NVIDIA GeForce RTX 2080 GPU on our local system, we ran our experiments using TensorFlow. We used the *Simple Transformers*[6] library to evaluate the large language models. The batch size for all models was equal to 1. Each Transformer used comes from the library of Transformers provided in Wolf et al. (2020). We used early stopping to fine-tune RoBERTa and GPT-2 for 10 epochs in accordance with the test set accuracy. According to the reference study, the best accuracy was typically attained prior to the tenth epoch of fine-tuning. For 20 epochs, the CNN was trained from scratch. Again, there were no advantages at training for more than 20 epochs in this study; on the test set, the top accuracies were consistently obtained before epoch 20. To assess the effectiveness of each model, we adhered to the protocol outlined in Liu et al. (2019). We therefore carried out five random weight initializations. There is no need for multiple runs of the SVM because it uses the implementation covered in the previous section, and it is deterministic. References to the initial implementations of each model and the experimental setups for each architecture are already supplied. On request, all the datasets we used can be obtained.

5 | RESULTS AND DISCUSSION

We tested the performance of our augmentation module on three datasets: FNS, HSS and ISS. For each augmentation combination, we trained all four models, evaluating their performance on the test set. The results for each combination of augmentation and model are presented in Tables 2 and 3.

Looking at the results, the augmentation module ensures that there is at least an augmentation strategy with a performance that is as good as the one of the original (so not augmented) across **all datasets and models**. It is worth noting that *augmented-it* stands for augmentation using Italian, and *augmented-mix* stands for augmentation using all four languages. Overall, **HSS** and **FNS** are the datasets where most of the combinations perform better than without our augmentation modules.

Roberta, CNN, and GPT-2 present significant p-values for the Italian and German augmentation when trained with the FNS dataset. We compute this by performing an unpaired one-tailed t-test, and in the tables the average of the runs is reported too. With RoBERTa, the augmentations with German, Turkish, and all languages mix, always perform better than without our augmentation modules. CNN architecture performs significantly better on the HSS dataset while on the ISS one, the performance of the training with original data can be equalled but not surpassed. Surprisingly, CNN with all languages as expansion cannot outperform other augmentation strategies.

TABLE 2 RoBERTa, CNN, and GPT-2 accuracy for each dataset and augmentation.

	RoBERTa			CNN			GPT-2		
FNS	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**
not-augmented	0,7100	0,6890	0,00	0,7300	0,7200	0,00	**0,6300**	0,6300	0,0000
augmented-it	0,7150	0,7080	0,0296	0,7200	0,7140	0,1140	0,6200	0,6120	0,0004
augmented-de	**0,7200**	0,6930	0,3549	0,7250	0,7160	0,1914	0,6050	0,6050	0,0000
augmented-ja	0,7100	0,6980	0,1548	0,7200	0,7170	0,2297	0,6050	0,6050	0,0000
augmented-tr	0,7100	0,6940	0,2906	**0,7350**	0,7190	0,4420	**0,6300**	0,6140	0,0081
augmented-mix	**0,7200**	0,6970	0,2207	0,7200	0,7140	0,1366	**0,6300**	0,6140	0,0081
HSS	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**
not-augmented	0,5900	0,5660	0,00	0,6500	0,6280	0,00	0,6500	0,6500	0,0000
augmented-it	**0,6400**	0,5700	0,4295	0,6800	0,6440	0,1312	**0,6600**	0,6480	0,3520
augmented-de	0,5900	0,5700	0,3912	0,6500	0,6400	0,1134	0,6500	0,6500	0,0889
augmented-ja	0,6200	**0,5760**	0,3386	**0,7200**	**0,7000**	0,0000	0,6400	0,6400	0,0000
augmented-tr	0,6200	0,5660	0,5000	0,6700	0,6300	0,4383	0,6000	0,5960	0,0001
augmented-mix	0,6100	0,5680	0,4668	0,6700	0,6560	0,0071	0,6100	0,6080	0,0000
ISS	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**	**Best run**	**Average**	***p*-value**
not-augmented	0,8222	0,7967	0,00	**0,9611**	0,9611	0,00	**0,9400**	0,9120	0,0000
augmented-it	0,8222	0,7900	0,3586	**0,9611**	0,9578	0,0352	0,7660	0,7660	0,0000
augmented-de	**0,8333**	0,7944	0,4412	**0,9611**	0,9578	0,1043	0,7660	0,7660	0,0000
augmented-ja	**0,8333**	0,8000	0,4179	0,9556	0,9534	0,0024	0,8700	0,8300	0,0001
augmented-tr	0,8111	0,8011	0,3515	0,9611	0,9545	0,0895	0,7660	0,7660	0,0000
augmented-mix	**0,8333**	0,8022	0,3470	0,9500	0,9500	0,0022	0,9222	0,9222	0,1094

Note: In the first column, the best results are reported while the second one reports the average of the 5 runs. The *p*-value column reports the output of the one-tailed *t*-test to check if there is a statistically significant difference between the not augmented accuracy and the alternatives' accuracies over the 5 runs. Bold values represent the best value of accuracy in each dataset.

TABLE 3 The table reports the values of maximum accuracy reached by the SVM for each dataset and augmentation.

	SVM accuracy		
	FNS	HSS	ISS
not-augmented	0.6300	0.5900	0.9278
augmented-it	0.6450	0.6100	0.9222
augmented-de	**0.6600**	0.6000	0.9167
augmented-ja	0.6300	0.6400	0.9278
augmented-tr	0.6200	0.6200	**0.9333**
augmented-mix	0.6350	**0.6700**	0.9167

Note: Bold means maximum value by columns.

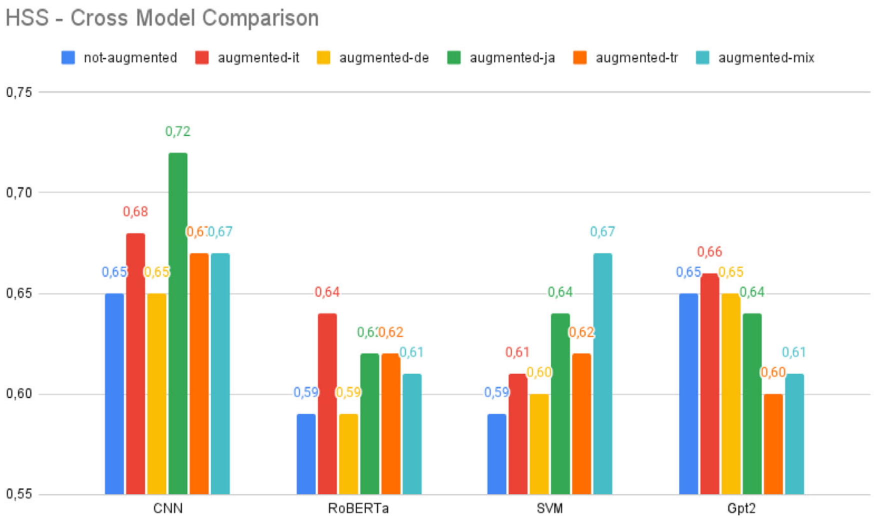
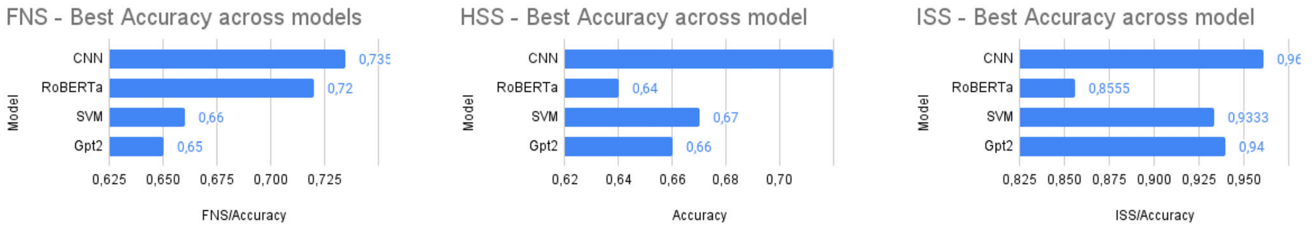


FIGURE 2 Cross-model and cross-augmentation results for the HSS dataset.



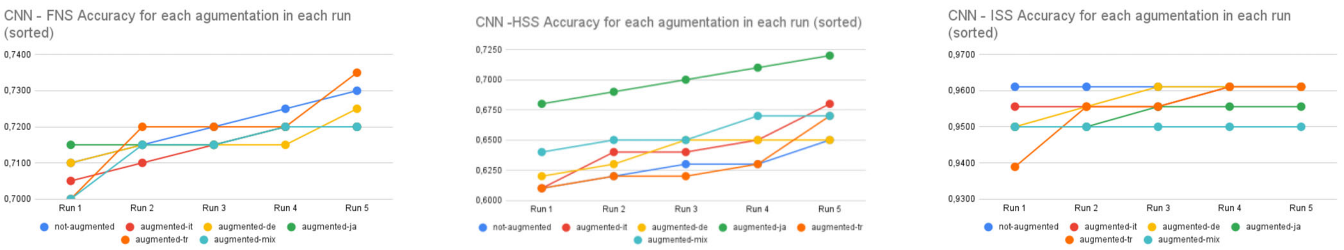
(a) FNS (b) HSS (c) ISS

FIGURE 3 Best accuracies for each model across datasets.

The SVM model seeks to maximize the distance between the decision boundary (hyperplane) and the closest data points from each class. In our experiments for HSS, all the augmentation performs better than the original one; the all language (*mix*) augmentation is better both for HSS and FNS.

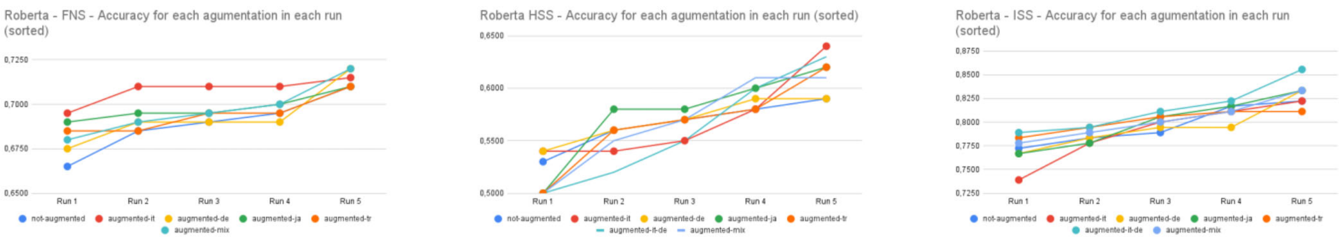
The CNN is the model that overall reaches the most accurate results, especially on the HSS dataset, and this can be seen in Figure 2. Figure 3 confirms that CNN is the best-performing model. CNN outperforms RoBERTa and SVM, which is the second-best on 2 out of 3 datasets. The official results for English,[7] indicate that for HSS the best accuracy is equal to 0.7300, 0.005 less than our best run (Siino et al., 2021), the winner of FNS challenge reaches an accuracy of 0.7500, and finally the best accuracy for ISS is equal to 0.9944.

Figures 4–6 report the (sorted) accuracy (on the validation data, $y-axis$) for each of the 5 runs ($x-axis$) of each model trained with each dataset.



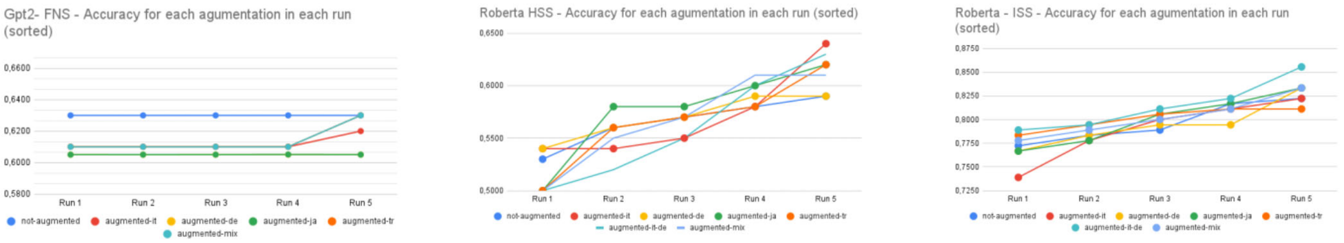
(a) CNN-FNS (b) CNN-HSS (c) CNN-ISS

FIGURE 4 CNN accuracies across different datasets and augmentations.



(a) RoBERTa - FNS (b) RoBERTa - HSS (c) RoBERTa-ISS

FIGURE 5 RoBERTa accuracies across different datasets and augmentations.



(a) GPT-2 - FNS (b) GPT-2 - HSS (c) GPT-2-ISS

FIGURE 6 GPT-2 accuracies across different datasets and augmentations.

The Japanese augmentation in the HSS dataset always outperforms the languages or combinations used for augmentation. On the other hand, in the ISS dataset, CNN reaches a higher level of accuracy (around +10%) w.r.t. RoBERTa. It is worth noting that GPT-2 is not able to perform as well as the other tested models. In this context we use it as a binary classifier on the three datasets and, eventually, this could motivate the lack of performance. It is interesting that GPT-2 on FNS and on ISS is able to reach the highest accuracy without our proposed framework. However, the results are always lower, for all the three datasets, when compared to a CNN. The main reason could be due to the specificity of the task. When performing AP it is not a single and small piece of text to be classified but a feed of texts from the same author. This observation is consistent with the main findings reported in Siino, Di Nuovo, et al. (2022).

5.1 | Qualitative analysis

In this section, we perform some qualitative analyses on the samples augmented after the backtranslation by comparing them with the non-augmented versions. First, we compare augmented versions with significantly improved performance, and then the augmented ones that have not. Specifically, with respect to the HSS dataset, we qualitatively analyze the augmented version in Japanese which allowed a significant increment in performance using a CNN. Then the qualitative analysis will be conducted on ISS where the augmentation has not produced significant increases with any language. Later on, we compare some augmented samples in the case of German with the original versions.

5.1.1 | Japanese on HSS

In Table 4 are shown some samples from the HSS dataset. The samples are backtranslated using Japanese and the changes are highlighted. In case 1*b*, only the word *Queen* is replaced with the word *Mistress*. This is a case of word substitution, where the semantics of the word *Mistress* is more specific and contextualized than the word *Queen*. In fact, the word *Queen* represents a case of polysemy in which the word can refer to both a queen of a kingdom, the popular rock band, a chess piece and, by extension, the concept of Mistress. Thus, a classifier previously trained with other meanings of the word *Queen* may not fully understand the actual meaning used. In contrast, the word *Mistress* has a specific meaning about a woman in a position of authority or control, often in sexual contexts. Also in the cases 2 and 3 some words are replaced (*confident* with *believe* and *man-meat* with *human flesh*). But in case 3, the referent of the discourse is also changed. In case 3*b torture it* becomes *torture you*. Also in case 4) a substitution of words could have made the latent semantics clearer for the classifier. In fact, words *Sophia is on her usual* are replaced with *Sofia shows her usual*. A single word (i.e., shows) replaces "is on her" and this, in the case of a CNN with single-word embedding, allows the expressed concept to be enclosed in a single term. Also in case 5), an interesting substitution (i.e., *to please* in place of *for the amusement*) makes explicit and shortens a concept on a single verb. Furthermore, the plural *races* are replaced by the singular race. It is interesting to note that 7 words present in 5*b*) were not present in 5*a*). In case 6*b*), *hitting* replaces the word *beatings*. And also in this case the two concepts are similar but not the same. In case 7*a*, the plural is replaced with the singular. Therefore, the author's comment loses the generic reference to a set of people and is addressed exclusively to a single subject. Finally, in case 8, the translator corrects a typing error and, therefore, the word in the augmented sample can be traced back to an already learned embedding space.

5.1.2 | German on ISS

With regard to the ISS dataset, as shown by the results, the German-augmented and non-augmented performances with CNN are equivalent. As the examples in Table 5 show, the translation is almost identical. This produces essentially similar classification performances.

TABLE 4 Examples of original tweets from HSS backtranslated using Japanese.

ORIGINAL	BACKTRANSLATED (JAPANESE)
1a) And the Queen will cage your cock and balls! #URL# #URL#	1b) And the Mistress puts your cock and balls in a cage! #URL# #URL#
2a) RT #USER#: I'm confident that all men are inferior to Women.	2b) RT #USER#: I believe that all men are inferior to women.
3a) RT #USER#: Use your man-meat for something meaningful. Let Femocracy Women torture it. Bow & Serve. #URL#	3b) RT #USER#: Use human flesh for something meaningful. Let Femocracy Women torture you. Bow & serve. #URL#
4a) RT #USER#: Sophia is on her usual fine and sadistic form in the new clip at #URL# #HASHTAG#	4b) RT #USER#: Sofia shows her usual feisty sadistic look in new clip on #URL# #HASHTAG#
5a) RT #USER#: A day at the races… nude males competing for the amusement of their female owners. #URL#	5b) RT #USER#: A day in the race… Naked men compete to please their female owners. #URL#
6a) #USER# Ball beatings is one of the most effective methods in order to keep in line the males of the family.	6b) #USER# Ball-hitting is one of the most effective ways to keep the men in your family in line.
7a) Bitches be in relationships and don't even like they bf	7b) Bitch is in a relationship and doesn't like it
8a) Tried to give a bih the world but she wanted the streets	8b) Tried to give the world to a bitch, but she wanted the streets.

Note: Changes in the backtranslated samples are highlighted in yellow.

TABLE 5 Examples of original tweets from ISS backtranslated using German.

ORIGINAL	BACKTRANSLATED (GERMAN)
1a) #USER# #USER# While Pierre's education may not be as elitist as Freeland's, I'd prefer as finance minister someone with his "commerce" education over a Slavic degree.	1b) #USER# #USER# While Pierre's education may not be as elite as Freeland's, as Treasury Secretary I would prefer someone with his "business" background to a Slavic degree.
2a) #USER# #USER# If #USER# wins #HASHTAG# she should consider "coaching" too. She's articulate but needs to shed the "lawyer" blandness. Can't look too meek when debating or Trudeau and media will eat her for breakfast	2b) #USER# #USER# If #USER# #HASHTAG# wins, she should also consider "coaching". She's articulate, but needs to drop the "Lawyer" fade. Can't seem too meek when debating or Trudeau and the media will eat her for breakfast
3a) #USER# #USER# #USER# Counter argument: Back in the '70's, Biden was racist too (different times, let's move on). Other accusations later were "hearsay". They say multi-blackface Trudeau isn't racist either. I think individual perception applies here	3b) #USER# #USER# #USER# Counter argument: In the 70's Biden was also racist (other times, let's move on). Other allegations later were "hearsay". They say multi-blackface Trudeau isn't racist either. I think individual perception counts here

Note: Changes in the backtranslated samples are highlighted in yellow.

In case 1*b*), even if some words have been replaced, the semantics are essentially the same. Furthermore, in the case of the short form *I'd* it is not even appropriate to speak of substitution as it has only been expanded with *I would*. Also in the second case, although the sentence contains many words, only a few of those present in 2*b*) are not present in 2*a*). *Also* achieves the same meaning as *too*, *Lawyer* has simply been replaced with the first capital letter, and *seem* and *look* are generally used interchangeably. Finally, in case 3 only four words are changed and in one case, as before, *also* in place of *too* is added.

This great similarity between the augmented and non-augmented versions of the samples is in fact confirmed by the similarity of the results obtained from the models on the ISS dataset.

6 | CONCLUSION AND FUTURE WORKS

In conclusion, for all three datasets examined, our proposed framework improves the performance if compared to a simplified version of our architecture without the augmentation modules (i.e., backtranslation and expansion). The technique consists of an augmentation model that makes use of **backtranslation** before expanding each sample by concatenating it with the original data. The findings imply that a user's text corpus can be semantically enriched, to enhance the performance of an AP model. The CNN model fared well with the HSS dataset, whereas the RoBERTa model consistently improved with the inclusion of backtranslation and expansion, despite the fact that each model's performance varied among datasets and augmentation combinations.

Differences in performance between the augmented and original models were also tested for statistical significance with a one-tailed *t*-test. Still, the *p*-value is under the threshold of 0.05 only for a few combinations. It is crucial to notice that the low sample size ($N = 5$) can affect the outcome of the test.

We found that enriching samples with their respective backtranslations can lead to performance improvements. The greater the diversity of the backtranslated versions, the more likely can be obtained a performance boost.

In addition, thanks to a qualitative analysis, we have found that backtranslation automatically allows the increasing of information content of a text without feature engineering. One of the most significant things that emerged is that backtranslation toward Japanese allows a relevant increase in performance thanks to a better explanation of the information that can express hatred on social media.

The most impactful shortcoming of our proposed approach is the need of backtranslating each sample before the classification stage. Even if this operation can be performed offline (i.e., backtranslating the dataset in advance), there is still an additional computational request to address on the input data. Furthermore, the new size of each expanded sample slows down the time required for the training and the classification stages. This slowdown is proportional to the number of target languages employed in the process. Even if in future works the backtranslation process could be performed offline, there will be this slowdown due to the increased size of the sample provided to the classifiers. However, this slowdown has been managed without the need of high-performance computing resources during all the experiments that we have performed.

In future works, it would be interesting to investigate this aspect also on other datasets and not only for AP tasks. Furthermore, it could also be of interest to evaluate the impact of other languages used in the backtranslation module, although, as emerged from this study, the inclusion of a larger number of languages does not necessarily lead to an increase in the performance of the classification models employed.

AUTHOR CONTRIBUTIONS

Marco Siino and Francesco Lomonaco were involved in investigation, resources, software, conceptualization, formal analysis, methodology, validation, visualization, writing original draft, review & editing. Paolo Rosso was involved in investigation, methodology, writing original draft, review & editing. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGEMENTS

The work of Paolo Rosso was in the framework of the following research projects: MARTINI — Malicious actors profiling and detection in online social networks through artificial intelligence (Grant PCI2022-135008-2 funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTR) and FairTransNLP-Stereotypes — Fairness and Transparency for equitable NLP applications in social media: Identifying stereotypes and prejudices and developing equitable systems (Grant PID2021-124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe).

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Marco Siino https://orcid.org/0000-0002-4453-5352
Francesco Lomonaco https://orcid.org/0000-0002-2295-1443

ENDNOTES

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge.

[2] https://github.com/marco-siino/DA-BT/tree/main/code.

[3] https://pypi.org/project/googletrans/.

[4] https://pan.webis.de/.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html, accessed on Fri Feb 09 10:52:16 2024.

[6] https://simpleTransformers.ai/about/.

[7] HSS https://pan.webis.de/clef21/pan21-web/author-profiling.html, FNS https://pan.webis.de/clef20/pan20-web/author-profiling.html, ISS https://pan.webis.de/clef22/pan22-web/author-profiling.html.

REFERENCES

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Association for Computational linguistic, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference* (pp. 26–33). Morgan Kaufmann Publishers. https://doi.org/10.3115/1073012.1073017

Bayer, M., Kaufhold, M., & Reuter, C. (2023). A survey on data augmentation for text classification. *ACM Computing Surveys*, *55*(7), 146:1–146:39. https://doi.org/10.1145/3544558

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks Media*, *24*, 100153. https://doi.org/10.1016/J.OSNEM.2021.100153

Behr, D. (2017). Assessing the use of back translation: the shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573–584. https://doi.org/10.1080/13645579.2016.1252188

Body, T., Tao, X., Li, Y., Li, L., & Zhong, N. (2021). Using back-and-forth translation to create artificial augmented textual data for sentiment analysis models. *Expert Systems with Applications*, *178*, 115033. https://doi.org/10.1016/J.ESWA.2021.115033

Bonthu, S., Dayal, A., Lakshmi, M. S., & Rama Sree, S. (2022). Effective text augmentation strategy for NLP models. In *Proceedings of Third International Conference on Sustainable Computing: Suscom 2021* (pp. 521–531). Springer.

Brislin, R. W., & Freimanis, C. (1995). Back-translation: A tool for cross-cultural research. In *An encyclopaedia of translation: Chinese-English, English-Chinese* (Vol. 1, pp. 22–40).

Buda, J., & Bolonyai, F. (2020). An ensemble model using n-grams and statistical features to identify fake news spreaders on twitter. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), (Vol. *2696*) CEUR-WS.org

Chang, C.-C., & Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 1–27. https://doi.org/10.1145/1961189.1961199

Chen, J., Yang, Z., & Yang, D. (2020). Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In D. Jurafsky, J. Chai, N. Schluter, & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020* (pp. 2147–2157). Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.ACL-MAIN.194

Croce, D., Garlisi, D., & Siino, M. (2022). An SVM ensemble approach to detect irony and stereotype spreaders on twitter. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 2426–2432). CEUR-WS. CEUR-WS.org

Cui, L., & Lee, D. (2020). Coaid: Covid-19 healthcare misinformation dataset. arXiv preprint arXiv:2006.00885.

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics. https://doi.org/10.18653/V1/D18-1045

Giachanou, A., Ghanem, B., Ríssola, E. A., Rosso, P., Crestani, F., & Oberski, D. L. (2022). The impact of psycholin-guistic patterns in discriminating between fake news spreaders and fact checkers. *Data & Knowledge Engineering*, *138*, 101960. https://doi.org/10.1016/J.DATAK.2021.101960

Hayashi, T., Watanabe, S., Zhang, Y., Toda, T., Hori, T., Astudillo, R. F., & Takeda, K. (2018). Back-translation-style data augmentation for end-to-end ASR. In *2018 IEEE Spoken Language Technology Workshop, SLT 2018* (pp. 426–433). IEEE. https://doi.org/10.1109/SLT.2018.8639619

Hoang, C. D. V., Koehn, P., Haffari, G., & Cohn, T. (2018). Iterative back-translation for neural machine translation. In A. Birch, A. M. Finch, M. Luong, G. Neubig, & Y. Oda (Eds.), *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 18–24). Association for Computational Linguistics. https://doi.org/10.18653/V1/W18-2703

Joo, T.-M., & Teng, C.-E. (2017). Impacts of social media (Facebook) on human communication and relationships: A view on behavioral change and social unity. *International Journal of Knowledge Content Development & Technology*, *7*(4), 27–50.

Kumar, G., Singh, J. P., & Singh, A. K. (2023). Autoencoder-based feature extraction for identifying hate speech spreaders in social media. *IEEE Transactions on Computational Social Systems*, 1–9. https://doi.org/10.1109/TCSS.2023.3240098

Lee, J., Kim, J., & Kang, P. (2021). Back-translated task adaptive pretraining: Improving accuracy and robustness on text classification. arXiv preprint, arXiv: 2107.10474.

Leonardi, S., Rizzo, G., & Morisio, M. (2021). Automated classification of fake news spreaders to break the misinformation chain. *Infection*, *12*(6), 248. https://doi.org/10.3390/INFO12060248

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint, arXiv:1907.11692.

Lomonaco, F., Donabauer, G., & Siino, M. (2022). COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 573–583). CEUR-WS. CEUR-WS.org

Lomonaco, F., Siino, M., & Tesconi, M. (2023). Text enrichment with Japanese language to profile cryptocurrency influencers. In M. Aliannejadi, G. Faggioli, N. Ferro, & M. Vlachos (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (Vol. 3497, pp. 2708–2716). CEUR-WS. CEUR-WS.org

Mangione, S., Siino, M., & Garbo, G. (2022). Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 2585–2593). CEUR-WS. CEUR-WS.org

Ortega-Bueno, R., Chulvi, B., Rangel, F., Rosso, P., & Fersini, E. (2022). Profiling irony and stereotype spreaders on twitter (irostereo). In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022-Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 2314–2343). CEUR-WS. CEUR-WS.org

Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research – a critique. *Expert Review of Pharmacoeconomics & Outcomes Research*, *20*(1), 69–77. https://doi.org/10.1080/14737167.2020.1734453

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rangel, F., De la Peña Sarracén, G. L., Chulvi, B., Fersini, E., & Rosso, P. (2021). Profiling hate speech spreaders on twitter task at PAN 2021. In G. Faggioli, N. Ferro, A. Joly, M. Maistro, & F. Piroi (Eds.), *Proceedings of the working notes of CLEF 2021–conference and labs of the evaluation forum* (Vol. 2936, pp. 1772–1789). CEUR-WS. CEUR-WS.org

Rangel, F., Giachanou, A., Ghanem, B. H. H., & Rosso, P. (2020). Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In L. Cappellato, C. Eickhoff, N. Ferro, & A. Névéol (Eds.), *Working Notes of CLEF 2020–Conference and Labs of the Evaluation Forum* (Vol. 2696, pp. 1–18). CEUR-WS. CEUR-WS.org

Sánchez-Junquera, J., Chulvi, B., Rosso, P., & Ponzetto, S. P. (2021). How do you speak about immigrants? taxonomy and stereoimmigrants dataset for identifying stereotypes about immigrants. *Applied Sciences*, *11*(8), 3610.

Sánchez-Junquera, J., Rosso, P., Montes-y-Gómez, M., & Chulvi, B. (2021). Masking and bert-based models for stereotype identication. *Proces. del Leng. Natural*, *67*, 83–94.

Sanguinetti, M., Comandini, G., Nuovo, E. D., Frenda, S., Stranisci, M., Bosco, C., Caselli, T., Patti, V., & Russo, I. (2020). Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task. In V. Basile, D. Croce, M. D. Maro, & L. C. Passaro (Eds.), *Proceedings of the Seventh Evaluation Campaign of natural Language Processing and Speech Tools for Italian* (Vol. 2765). CEUR-WS. CEUR-WS.org

Sharmila Devi, V., & Kannimuthu, S. (2023). Author profiling in code-mixed whatsapp messages using stacked convolution networks and contextualized embedding based text augmentation. *Neural Processing Letters*, *55*(1), 589–614. https://doi.org/10.1007/S11063-022-10898-3

Shleifer, S. (2019). Low resource text classification with ulmfit and backtranslation. arXiv preprint, arXiv:1903.09244.

Siino, M., Di Nuovo, E., Tinnirello, I., & La Cascia, M. (2021). Detection of hate speech spreaders using convolutional neural networks. In G. Faggioli, N. Ferro, A. Joly, M. Maistro, & F. Piroi (Eds.), *Proceedings of the Working Notes of CLEF 2021–Conference and Labs of the Evaluation Forum* (Vol. 2936, pp. 2126–2136). CEUR-WS. CEUR-WS.org

Siino, M., Di Nuovo, E., Tinnirello, I., & La Cascia, M. (2022). Fake news spreaders detection: Sometimes attention is not all you need. *Information*, *13*(9), 426. https://doi.org/10.3390/INFO13090426

Siino, M., La Cascia, M., & Tinnirello, I. (2022). Mcrock at Semeval-2022 task 4: Patronizing and condescending language detection using multi-channel CNN, hybrid LSTM, distilbert and XLNET. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, & S. Ratan (Eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation* (pp. 409–417). Association for Computational Linguistics. https://doi.org/10.18653/V1/2022.SEMEVAL-1.55

Siino, M., Tesconi, M., & Tinnirello, I. (2023). Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In M. Aliannejadi, G. Faggioli, N. Ferro, & M. Vlachos (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (Vol. 3497, pp. 2772–2781). CEUR-WS. CEUR-WS.org

Siino, M., & Tinnirello, I. (2023). Xlnet with data augmentation to profile cryptocurrency influencers. In M. Aliannejadi, G. Faggioli, N. Ferro, & M. Vlachos (Eds.), *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)* (Vol. 3497, pp. 2763–2771). CEUR-WS. CEUR-WS.org

Siino, M., Tinnirello, I., & La Cascia, M. (2022). T100: A modern classic ensemble to profile irony and stereotype spreaders. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working notes of CLEF 2022–Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 2666–2674). CEUR-WS. CEUR-WS.org

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? a comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, *121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Subramanian, K. R. (2017). Influence of social media in interpersonal communication. *International Journal of Scientific Progress and Research*, *38*(2), 70–75.

Sulis, E., Farías, D. I. H., Rosso, P., Patti, V., & Ruffo, G. (2016). Figurative messages and affect in twitter: Differences between# irony,# sarcasm and# not. *Knowledge-Based Systems*, *108*, 132–143. https://doi.org/10.1016/J.KNOSYS.2016.05.035

Unsvåg, E. F., & Gambäck, B. (2018). The effects of user features on twitter hate speech detection. In D. Fiser, R. Huang, V. Prabhakaran, R. Voigt, Z. Waseem, & J. Wernimont (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online, alw@emnlp 2018* (pp. 75–85). Association for Computational Linguistics. https://doi.org/10.18653/V1/W18-5110

Wang, Z., & Qu, Z. (2017). Research on web text classification algorithm based on improved CNN and SVM. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)* (pp. 1958–1961). IEEE. https://doi.org/10.1109/ICCT.2017.8359971

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/V1/2020.EMNLP-DEMOS.6

Xie, Q., Dai, Z., Hovy, E. H., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*. Neural information processing systems foundation.

Yu, W., Boenninghoff, B. T., & Kolossa, D. (2022). Bert-based ironic authors profiling. In G. Faggioli, N. Ferro, A. Hanbury, & M. Potthast (Eds.), *Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum* (Vol. 3180, pp. 2720–2733). CEUR-WS. CEUR-WS.org

Zhang, Y., & Wallace, B. C. (2017). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. In G. Kondrak & T. Watanabe (Eds.), *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017* (Vol. 1, pp. 253–263) Asian Federation of Natural Language Processing.

AUTHOR BIOGRAPHIES

Marco Siino received the Bachelor's and the Master's Degree cum Laude in computer engineering from the University of Palermo. He obtained his PhD in Information and Communication Technologies from the University of Palermo in 2023. He is also a freelance full stack developer. His main interests are in Machine Learning, Deep Learning, Natural Language Processing and Recommender Systems. His work involves social-networks-related tasks (e.g., sentiment analysis, hate speech detection, fake news detection). He is currently a post-doc researcher at the University of Bologna.

Francesco Lomonaco is an economist turned data scientist possessing a specialized skill set focused on time series modeling, a proficiency cultivated during the economics studies. He attained a PhD in Data Science from Bicocca University in May 2023. His doctoral research delved into the intricate dynamics of social media, investigating subjects like opinion shifts in polarized settings, threat assessment on social platforms, and the dissemination of visual misinformation.

Paolo Rosso is Full Professor of Computer Science at the Universitat Politècnica de València (UPV), Spain. He is a member of the Pattern Recognition and Human Language Technology (PRHLT) research center and the Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI). He has published 400+ peer-reviewed papers in conferences and journals, and he is among the best computer science scientists in Spain (https://research.com/scientists-rankings/computer-science/es). His current research interests fall in the area of detection of harmful information in social media, both fake news and hate speech. In 2022 he received the UPV Research Award for Excellent Publication in Engineering and Technology on misogyny identification. He is the PI of several related research projects such as FairTransNLP-Stereotypes: Fairness and Transparency for equitable NLP applications in social media – Identifying stereotypes and prejudices and developing equitable systems (Grant PID2021-124361OB-C31), FAKEnHATE-PdC: FAKE news and HATE speech (Grant PDC2022-133118-I00), and XAI-DisInfodemics: eXplainable AI for disinformation and conspiracy detection during infodemics (Grant PLEC2021-007681), funded by MCIN/AEI/10.13039/501100011033 and by European Union NextGenerationEU/PRTRPI.

How to cite this article: Siino, M., Lomonaco, F., & Rosso, P. (2024). Backtranslate what you are saying and I will tell who you are. *Expert Systems*, *41*(8), e13568. https://doi.org/10.1111/exsy.13568