Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 81-93 recibido 22-11-2024 revisado 20-01-2025 aceptado 06-02-2025

Towards Robust Models for Fake News Detection in Spanish

Hacia modelos robustos para la detección de noticias falsas en español

Sergio Gómez González,[1] **Mariona Coll Ardanuy,**[2] **Paolo Rosso**[1,3]
[1]Universitat Politècnica de València
[2]Barcelona Supercomputing Center (BSC)
[3]ValgrAI-Valencian Graduate School and Research Network of Artificial Intelligence
prosso@dsic.upv.es

Abstract: In this paper, we face the challenge of fake news detection exclusively in Spanish, an application domain in which there has not been much research. Furthermore, the news topics are in continuous change and models that are not able to adapt end up being ineffective in the long term. For that reason, in this domain, the robustness of the models is key. With that goal in mind, we have applied several techniques that include data exploitation and augmentation in order to improve the performance of a simple pre-trained transformer-based model. Additionally, we have included a comparison with a generative large language model. Moreover, we use two different dataset splits to compare that performance: a standard approach to partitioning the dataset, balancing the training and test sets, and a more realistic (adversarial) one. Finally, we discuss which aspects have more influence over the robustness and performance of the fake news detection models.
Keywords: fake news detection, data augmentation, large language models.

Resumen: En este artículo nos enfrentamos al reto de detectar noticias falsas exclusivamente en español, un campo en el cual no ha habido demasiado esfuerzo de investigación. Además, la temática de las noticias se mantiene en continuo cambio, por lo que los modelos que no pueden adaptarse acaban siendo poco efectivos a largo plazo. Es por ello que, en este campo, la robustez es imprescindible. En búsqueda de esa propiedad, hemos aplicado distintas técnicas de explotación y aumento de datos para mejorar los resultados de un clasificador simple basado en un *transformer* preentrenado. Adicionalmente, hemos incluido una comparativa con un gran modelo de lenguaje generativo. También, utilizamos dos particiones distintas de un mismo *dataset* para comparar su efectividad: una partición típica con conjuntos de entrenamiento y test parecidos y otra más realista (adversaria). Finalmente, analizamos qué aspectos ejercen mayor influencia sobre la robustez y efectividad de los modelos para la detección de noticias falsas.
Palabras clave: detección de fake news, aumento de datos, grandes modelos de lenguaje.

1 Introduction

Our society is becoming increasingly more aware of the threat of disinformation.[1] Along with it, the Natural Language Processing (NLP) field has been trying to find solutions to this risk. However, the research community is struggling to find robust methods to solve the problem. One of the reasons of that struggle is the shortage of available datasets, especially in languages different than English. Moreover, creating well-curated NLP datasets is complex, expensive and time-consuming. Particularly in the social sciences domain, subjectivity plays a significant role in the annotation decisions. That makes even more complicated the task of developing representative, useful datasets.

Specifically in the fake news detection domain, data annotation has mostly been done

[1]A survey conducted in January 2025 by the Spanish Center for Sociological Research (CIS) revealed that 3.9% of the respondents considers the role of media and social networks, including disinformation, as one of the three main problems in Spain. See: `https://www.cis.es/documents/d/cis/es3492mar-pdf`, accessed February 12, 2025.

ISSN 1135-5948 DOI 10.26342/2025-74-6

via fact-checking. This activity is performed by humans that check the sources of an article and compare it with others that refer to the same fact. However, this is an expensive and time-consuming process compared to the massive flow of fake news that are produced.

However, as society is ever changing, the available datasets will always be outdated. The topics that worry people and the facts that are covered in the articles are subject to continual change. For instance the dataset that we have used (Spanish Fake News Corpus (Posadas-Durán et al., 2019; Aragón et al., 2020; Gómez-Adorno et al., 2021)) needed to be updated in the second edition of the challenge to include a new topic: Covid-19. For that reason we need techniques that enable models to identify the patterns that distinguish fake and reliable news regardless of the subject that they address.

In this vein, the goal of this paper is to experiment with such techniques that could make the most of the available data. Our aim is to find methods that enable models to last longer without needing to be updated. In particular, we tackle the following research questions:

- ***RQ1***: *Which techniques are more effective to augment the robustness of a fake news detection model?*
- ***RQ2***: *What role does the size of the model play with respect to generalization?*
- ***RQ3***: *Can a dataset gathered in one sociological context be used to train a model that works well in another context*?
- ***RQ4***: *Is the full text of the article required to correctly classify the article?*

2 Background

Text classification is the task of categorizing sentences or documents into pre-defined classes. At present, most text classification systems are approaches based on the *transformer* architecture (Vaswani et al., 2017), in which a pre-trained language model is fine-tuned for the downstream task of categorizing texts into labels. Whereas the irruption of generative large language models (LLM) has revolutionized NLP, the fact remains that most of the approaches that participate in shared tasks on classifying text for the social domain (such as oppositional thinking analysis (Korenčić et al., 2024) and author profiling (Rangel et al., 2020)) continue to rely on discriminative models, mostly based on BERT (Devlin et al., 2019) or the related RoBERTa architecture (Liu et al., 2019).

Fake news detection is often approached as a binary text classification problem (Ruffo et al., 2023), in which a news article is categorized into either *fake* or *true*.[2] In Spanish, the only dataset that has been specifically created for the task of fake news detection is the Spanish Fake News Corpus (henceforth *SFNC*) (Posadas-Durán et al., 2019; Aragón et al., 2020; Gómez-Adorno et al., 2021), which was created for the *Fake News Detection in Spanish* shared tasks.[3] In the second edition of the task, the best-performing approach (Huang, Xiong, and Jiang, 2021) implemented a classifier, fine-tuned on top of a pre-trained BERT model for Spanish (Cañete et al., 2020), which encoded the headline and both the beginning and the end of the article. Afterwards, other approaches tried to use the emotions expressed in the articles to classify them (Togni et al., 2024).

Data augmentation techniques increase the amount of training data without the costs associated with annotating or collecting additional data (Feng et al., 2021). These techniques aim at achieving a better generalization by reducing overfitting during training, but there is a risk that such techniques may actually lead to greater overfitting or worse performance overall if the augmented data representations are either too similar or too different from the original ones (Feng et al., 2021). Likewise, Longpre, Wang, and DuBois (2020) show that popular text augmentation techniques—such as back-translation—provide little or no improvement when applied to transformer-based approaches. In fact, after comparing the performance on different pre-trained transformer-based mod-

[2]The terms 'reliable' and 'unreliable' are sometimes used instead of 'true' and 'fake'. It is also worth noting that there are some examples in which the task is treated as a multi-class problem, in which the fake category is further divided into other categories, such as *satire*, *hoax*, *propaganda*, and *clickbait* (Rashkin et al., 2017; Ghanem, Rosso, and Rangel, 2020).

[3]It is worth mentioning that the Iberifier project (`https://iberifier.eu/`) is building a fact-check repository which collects news articles that could potentially be used for fake news detection. However, the data in the repository may yet not be directly used, since it was not built for the specific purpose of evaluating a NLP system.

els, the authors conclude that the process of pre-training must provide similar benefits to those targeted by these data augmentation techniques. Yet, the authors argue, such techniques can still be advantageous if they introduce new linguistic patterns to the data.

Surveys of data augmentation in NLP distinguish between augmentation approaches that transform the input data and those that transform the representation of the data (Feng et al., 2021; Bayer, Kaufhold, and Reuter, 2022). One of the most standard approaches that work at the input level is back-translation, also known as round-trip translation, which aims at generating paraphrases of the input document through machine translation (Federmann, Elachqar, and Quirk, 2019). One example of the use of this technique for an NLP task is explained in Siino, Lomonaco, and Rosso (2024). When data exists in other languages and can be used for the same task, machine translation can also be applied to increment the amount of training data.

The term data augmentation is also used to refer to approaches that perform data transformation. One example is masking named entities, which previous work demonstrated to be particularly useful to mitigate diachronic bias in fake news detection, given that it forces the model to ignore named entities when making the prediction (Murayama, Wakamiya, and Aramaki, 2021), making the model more robust. Another example is to lengthen or manipulate the context window. BERT and RoBERTa-based approaches can only process texts of a certain length (usually 512 tokens), and the standard approach to dealing with longer texts is to just truncate them, potentially losing important information (Huang, Xiong, and Jiang, 2021). Some approaches have been designed to overcome the issue of maximum context length, such as *ConvBERT* (Jiang et al., 2020) and *Longformer* (Beltagy, Peters, and Cohan, 2020).

Finally, one example of augmentation by transforming the representation of the data is the introduction of noise in the embeddings, as proposed in Jain et al. (2024). As it changes the representation at a semantic level, it does not strictly augment the number of training samples, but the interpretations that the model processes. This technique is quite related to adversarial learning methods as the one exposed in Zhu et al. (2020).

3 Dataset

To the best of our knowledge, only one dataset exists in Spanish that has been created for the task of detecting fake news. The Spanish Fake News Corpus (SFNC)[4] (Posadas-Durán et al., 2019; Aragón et al., 2020; Gómez-Adorno et al., 2021) was introduced for the *Fake News Detection in Spanish* (*FakeDeS*, for short) shared task, held at the IberLef 2020 and 2021 workshops. The original dataset is split into a training set (676 news articles), a development set (295 news articles) and a test set (572 news articles). The dataset is balanced in its distribution of fake and true news articles.

The dataset was compiled in two stages, resulting in a significant temporal gap of the articles belonging to the training and development sets (written in the first half of 2018) and those in the test set (collected between November 2020 and March 2021). Whereas the dataset is balanced across the different splits, its topic distribution is heavily imbalanced, as shown in Figure 1. This adds a layer of difficulty to the task, since the test set could be practically considered as out-of-domain, given its temporal and topical shift. However, this gap between training and inference is characteristic of online detection of fake news.

4 Approaches

The goal of this paper is to investigate the impact of a series of data augmentation and transformation techniques in the task of fake news classification. As baselines, we have fine-tuned two simple classifiers mounted on top of two RoBERTa (Liu et al., 2019) models. Since our dataset is in Spanish, we use, as the core of our classifiers, the base[5] and large[6] *RoBERTa* models that were trained on data from the National Library of Spain (BNE). Both models were trained on the same amount of data, but differ in their model configuration, and therefore their training complexity: the base model has 12 self-attention layers with 12 attention heads, a hidden size of 768 and a total of 125M parameters; whereas the large model has 24 self-

[4] https://github.com/jpposadas/FakeNewsCorpusSpanish.

[5] https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne.

[6] https://huggingface.co/PlanTL-GOB-ES/roberta-large-bne.

attention layers with 16 attention heads, a hidden size of 1280 and a total of 774M parameters (Gutiérrez-Fandiño et al., 2022).

In the remaining of this section, we will describe five different data augmentation techniques: masking named entities in Section 4.1, data augmentation through adding external data in Section 4.2, data augmentation through back-translation in Section 4.3, introduction of noise at the embedding-level in Section 4.4 and lengthening of the context window in Section 4.5.

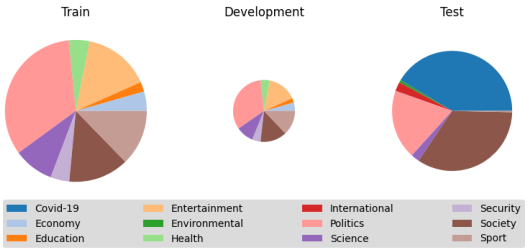

Figure 1: Distribution of topics in the original split of the SFNC dataset. The size of each pie chart reflects the size of the dataset split.

4.1 Masking Named Entities

Our first improvement of the training process involves masking the named entities with special tokens. Thus, we used a *spaCy*[7] model to detect the named entities in the texts and then replaced them with the corresponding tokens in the whole dataset. We then fed the masked data to the model to perform the training. The same process was conducted for the evaluation with the trained model.

4.2 Data Augmentation through Translation of Additional Data

Lately, machine learning professionals are getting concerned about the importance of training their models with large amounts of curated data. For that reason we decided to introduce more fake and reliable news in our dataset. We have not heard of any other news dataset in Spanish, so we needed to translate parts of other datasets from the English. We obtained the fake labeled news from the dataset created in Rashkin et al. (2017) and the reliable news from the *LOCO* corpus (Miani, Hills, and Bangerter, 2021). Both datasets differ considerably in size: the former contains 6,942 unreliable news while the latter keeps 96,743 reliable news. We could not use the full of both datasets without inducing a big imbalance in our data so we had to select which news we were going to use. Hence, we used three different criteria:

- As baseline, we randomly selected a subset of reliable news and translated them.
- In a clever way, we used a *RoBERTa* English model to obtain the embeddings of the text of the news in the LOCO dataset. Then, we stochastically incorporated the most diverse set of news by weighting heavier the most different news from the ones we had in our dataset. For that, we used the Euclidian distance from the embedding of each article to the center of the LOCO dataset. Then, we translated the selected subset.
- For our third criterion we, first, used the previous approach to select a smaller set of 10k news that we translated to Spanish. We could not translate the whole corpus for time and material restrictions, Then, we selected a certain amount of samples (less than 10k) by weighting heavier the most different news from the ones we had in our dataset.

For the three approaches, the amount of translated data was a parameter to adjust. We used the *MarianMT* (Junczys-Dowmunt et al., 2018) model to translate from English to Spanish with the implementation of the *Helsinki-NLP* group.[8]

4.3 Data Augmentation through Back-Translation

A common data augmentation technique in NLP is to translate the text through a circular pipeline of languages, in which the source and last target are the same language. This process changes the original text because the same source sentence can have several valid translations. Thus, even using well-trained translation models, sentences can slightly change while keeping the overall meaning. It is important to note that, if the translation pipeline is excessively large, the sentence could be so changed that the meaning could be lost in the ambiguity between languages. After some experimentation, we observed that the best trade-off

[7] `https://spacy.io/`.

[8] `https://huggingface.co/Helsinki-NLP/opus-mt-en-es`.

between accuracy and novelty was achieved with a pipeline chaining three translations in succession. The pipelines we tried are available in the Appendix A.

When possible, we used *MarianMT* models for translating. If the translation direction was not available we used the *M2M100* (Fan et al., 2021), many-to-many multilingual translation model.[9]

4.4 Data Augmentation Using Noisy Embeddings

The last data augmentation technique that we have used is to introduce noise at each training forward pass when the word embeddings are obtained. We implemented it as explained in Jain et al. (2024), by adding a random vector extracted from a normal or uniform distribution and scaled by a factor of $\frac{\alpha}{\sqrt{Ld}}$ (L represents the length of the sequence and d is the dimensionality of the embedding vectors). In this approach, α is a tunable parameter that controls the scale of the noise.

The noise at the semantic representation is supposed to improve the generalization of the overall model. As it can be introduced in any model that makes use of embedding vectors, it can be considered as an agnostic technique suitable for any task. Also, it can be performed "on the fly" because of the low computational effort that it needs.

4.5 Elongation of the Context Window

The *RoBERTa* models that we have used in this work can process 510 tokens at maximum plus the *CLS* and *EOS* tokens. This implies that, for the classification, a simple *RoBERTa* based classifier can only process a window of that length at maximum. To compensate this fail, we used the *BELT*[10] (BERT for longer texts) approach.[11] This idea consists in applying the *BERT*-like model over the sequence as if in a convolution. Following the comparison, the *BERT*-like model would be the filter of the convolution that is applied over the sequence with a certain stride. Once the *BERT*-like model has been applied, the *CLS* embedding from each window is extracted. All those vectors are, then, aggregated with some technique as a pooling layer in a convolutional network. We have explored different ways of aggregation from the most simple ones (max. pooling, avg. pooling, addition and self attention layers) to more computing expensive approaches (recurrent neural networks and transformer layers). The resulting vector is fed to the classifier head.

5 Experiments

5.1 Metrics

We evaluate our experiments using the F1 score metric, which is calculated as a harmonic mean of the precision and recall. We provide two variations of the metric: the *micro F1 score*, which is calculated by counting the total amount of true positives, false negatives and false positives (i.e., each instance in the dataset having the same weight); and a weighted F1 score, analogous to the macro F1 score, but averaging the per-topic scores, instead of per-class. In this latter scenario, which we call *macro F1 score*, we take the F1 score of the results of each topic and average them, treating all the topics equally. Using both scores we get a more accurate vision of how the models perform, given the high imbalance of the sets among the topics.

In order to assess whether the difference between the different performances of the different models is significant or not, we provide the McNemar's test (McNemar, 1947), which is recommended for binary classification problems such as this one (Dror et al., 2018). In all scenarios, we calculate the statistical significance of the difference between the data augmentation approach with respect to the base approach.

5.2 Dataset Splits

As mentioned in Section 3, there is a significant difference between the training and test sets in terms of their temporal scope and distribution of topics. In our experiments, we keep the original partition, because we consider that it more faithfully resembles a real scenario in online fake news detection. For comparison, we have created an additional partition of the dataset, more balanced, in which we merged the original splits, shuffled

[9] `https://huggingface.co/facebook/m2m100_418M`.

[10] `https://mim-solutions.github.io/bert_for_longer_texts/`.

[11] Other approaches, such as the aforementioned *ConvBERT* (Jiang et al., 2020) and *Longformer* (Beltagy, Peters, and Cohan, 2020) have proven to be more effective than BELT. Nevertheless, to our knowledge, there are no *ConvBERT* or *Longformer* trained in the Spanish language, whereas the *BELT* approach allows us to use any *BERT*-like model.

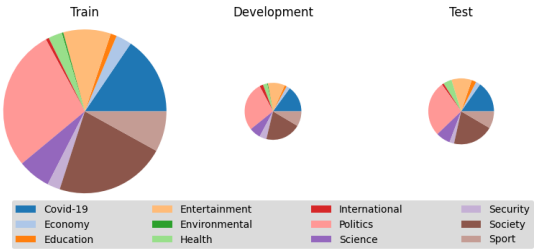

Figure 2: Distribution of the topics in the stratified split of the SFNC dataset. The size of each pie chart reflects the size of the dataset split.

the full dataset, and split the data again, keeping 64% of the dataset for training (987 articles), 16% for development (247 articles), and leaving out the remaining 20% for test (309 articles). In this new partition, we make sure that both topics and labels remain balanced across the data splits as shown in Figure 2, consequently also removing the temporal gap between the training set and the test set. From now on, we call the partitions *original* and *stratified*, respectively.

5.3 Pre-trained Models

We performed two parallel sets of experiments, using two fully comparable pre-trained language models that only differ in the model size: *roberta-base-bne* and *roberta-large-bne*, with 125 and 355 million parameters respectively (Gutiérrez-Fandiño et al., 2022). The goal of trying both models is to precisely understand the impact of data augmentation with respect to the model size.

5.4 Experimental Settings

Even though we applied different data augmentationn techniques, all of the models were trained according to a common framework. That common setup enables the comparison of the applied techniques in equal conditions. Nevertheless, it permits enough degrees of freedom to preserve the reachability of the maximum performance for any of the used techniques. Those degrees of freedom include the learning rate, the linear learning rate scheduler start factor and each of the specific tunable parameters of the used techniques. The late group is formed by the type of named entities that are masked, the approach to the integration of external data (along with the quantity of that data), the alpha scale of the noisy embeddings, the translations pipeline of the back-translation and the total length of the context window and the stride of the *convolution* in the *BELT* approach. In the case of *BELT* we have frozen the weights of the transformer encoder during the first three epochs of training for stability.

The implementation of the approaches of Section 4 and the code that we have used to perform the experiments are publicly available in a GitHub repository.[12]

5.4.1 Common Settings

As the common experimental setting we have followed the classical three partition experimentation: we used the training partition for the learning phase of the model, the development set for tuning the hyperparameters and the test split for the final comparison. To perform an efficient training, we also introduced early stopping during training because it usually leads to better generalization (Prechelt, 2002). Thus, we used as the stopping criteria the improvement in the loss function with a patience of three validations. We also used an Adam optimizer with a linear learning rate scheduler and an effective batch size of 8 samples in all of our experiments.

As we have used data augmentation techniques, the size of our training set differs depending on the experiment. Thus, we performed the validations between the same number of steps instead of doing them at the end of every epoch. This subtle change allows us to observe the difference in the performance after having been trained with the same amount of data, and study the effects of the diversity in the training set. Therefore, we set the validations to happen between the equivalent of the size of the non augmented training split in both of our partitions. We, then, selected the checkpoints with the best F1 score in the development set.

5.4.2 Hardware Equipment

The translations included in our experimentation were produced in an equipment provided with an Nvidia GeForce RTX 4090 GPU with 24GB of VRAM. The same computer was used to train the BELT with the *roberta-large-bne* model. The rest of our experiments were performed in a computer provided with a GPU Nvidia GeForce RTX 3060 GPU with 12GB of VRAM.

[12] https://github.com/sergiogg-ops/FAKEnHATE/

Model	Original		Stratified		Δ
	F1	Δ_{base}	F1	Δ_{base}	
Base	48.9	–	86.0	–	↑ 37.1
NER	60.7	↑ 11.8	83.7	↓ 2.3†	↑ 23.0
Long context	**76.8**	↑ 27.9	87.4	↑ 1.4†	↑ 10.6
External data	66.3	↑ 17.4	86.3	↑ 0.3†	↑ 20.0
Noisy embeddings	49.0	↑ 0.1†	86.9	↑ 0.9†	↑ 38.0
Back-translation	71.7	↑ 22.8	**89.3**	↑ 3.3†	↑ 17.6
Llama-3.1:8B (Default True)	74.7	↓ 6.1†	80.1	↓ 5.4†	↑ 5.3
Llama-3.1:8B (Default False)	84.2	↑ 35.3	82.6	↓ 3.4†	↓ 1.5

Table 1: Macro F1 score obtained by the base models in the test partitions in the original and stratified sets. The ↑↓ symbols indicate whether the difference with the base is positive or negative respectively. The † symbols indicate that difference is not statistically significant.

Model	Original		Stratified		Δ
	F1	Δ_{base}	F1	Δ_{base}	
Base	52.0	–	85.9	–	↑ 33.8
NER	59.1	↑ 7.0	85.4	↓ 0.5†	↑ 26.3
Long context	**77.6**	↑ 25.6	**88.5**	↑ 2.7†	↑ 10.9
External data	65.7	↑ 13.6	83.5	↓ 2.3†	↑ 17.9
Noisy embeddings	54.2	↑ 2.1†	83.5	↓ 2.3†	↑ 29.4
Back-translation	75.1	↑ 23.1	86.5	↑ 0.7†	↑ 11.4
Llama-3.1:8B (Default True)	74.0	↓ 3.3†	80.3	↓ 4.7†	↑ 6.3
Llama-3.1:8B (Default False)	80.9	↑ 28.9	83.4	↓ 2.5†	↑ 2.5

Table 2: Micro F1 score obtained by the base models in the test partitions in the original and stratified sets. The ↑↓ symbols indicate whether the difference with the base is positive or negative respectively. The † symbols indicate that difference is not statistically significant.

6 Results

After having performed the training following the different approaches, we selected the best models according to the Section 5.4:

Named entity mask: after an ablation study we determined that the best results were produced when only the named entities related to people and organizations were masked.

External sources: the best results were obtained using the third approach described in Section 4.2, with 1500 samples from each source.

Back-translation: the best performing pipeline was the following: Spanish → Swedish → Chinese → Spanish.

Noisy embeddings: the best F1 score was achieved with a uniform noise and an α parameter of 10.

Bert for Longer Text (BELT): we used an extended context window of 2.5k tokens. The best results were offered by aggregating with a transformer encoder layer.

6.1 Using Generative LLMs to Classify Fake News

Even though we are focusing on making models more robust, for the sake of comparison we have also used generative LLMs for our task. This type of models are the state of the art in a vast amount of tasks nowadays. Thus, we chose the *Llama* family of models, distributed by *Meta*[13] because all of them are open sourced. Furthermore, there are many utilities to use them easily as *Ollama*,[14] the framework that we have chosen. After several experiments over the development sets and applying prompt engineering to obtain the desired output, we opted for the *Llama 3.1* model (Dubey et al., 2024); specifically, the version of 8,000 million of parameters (*llama-3.1:8B*). We also tried the newest lightweight Llama 3.2 models (*llama-3.2:3B* and *llama-3.2:1B*) with slightly worse results.

For some cases we were not able to bypass the security restrictions that force them to not answer questions that can lead to decisions that could, somehow, hurt individuals. In those cases we artificially induced a bias to

[13] `https://www.meta.com/`.

[14] `https://ollama.com/`.

the "default class" (either "True" or "False"). Sometimes, that behavior led to a failed answer. The prompts used for this approach can be consulted in Appendix B.

In Tables 1 and 2, we show the macro and micro F1 score (respectively) for both the original and the stratified sets when using the *base* RoBERTa model as base. Tables 3 and 4 report the same metrics when using the *large* RoBERTa model as base. We have also included a comparison with the *LLama* model that, after some experimentation, worked the best for our task: *Llama-3.1:8B*.

The results are significantly different in the *original* and *stratified* partitions. They suggest that, as we hypothesized, the *original* partitions offer a rather harder challenge. Thus, we will address them in each section.

6.2 Stratified Partitions

In the split created for this study, none of the differences between the explored models and the baseline appears to be significant. Nevertheless, we can appreciate some patterns that, in the original partitions, are amplified and more evident. The base sized models trained with data augmentation tend to obtain a higher macro F1 score among all the topics. This behavior might suggest a better generalization capability, but without significant differences we cannot argue that. On the other hand, the model trained masking the named entities and the *BELT* model scored higher in the micro F1 among the topics.

Regarding the large sized models, it is surprising that the F1 scores are so similar or even lower than the scores obtained by the base sized ones. These results can be explained by a saturation of the task when we use excessively complex models. That would mean that the classification problem proposed with these partitions might be too simple to use larger models. The *Llama 3.1* model is among the models with the worst results, although the difference with the baseline is not significant.

6.3 Original Partitions

The main distinction between these partitions and the stratified ones is the huge distance between the F1 score of the base and large sized models. When using this set, the base sized models tend to obtain significantly worse results. The two only exceptions are the *BELT* model and the one trained with back-translation. The base sized *BELT* is the second best model in the micro F1 score. Besides that, the base sized models cannot even be compared with the larger ones.

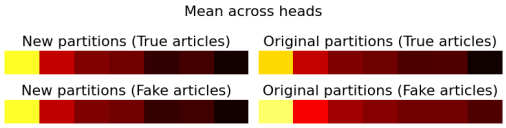

Figure 3: Mean of the self attention weights used by the base sized *BELT* at the aggregation layer over the test set of the *stratified* (left col.) and *original* (right col.) partitions. Lighter cells contain higher values.

The exception among the large sized models is, once again, *BELT*, which obtains significantly worse results than the rest of the approaches. These late models generally score higher in the macro F1 among the topics than the micro. This tendency is inverted by the model trained with noisy embeddings, that obtains the highest F1 micro, boosted by its good performance in the Covid-19 topic.

With this test set, the *Llama 3.1* model shows a huge difference depending on which is the "default" class. The option to declare as "Fake" the articles that could not be determined as "True" is far more accurate than the opposite. In this case, it is the difference between be the worst or the best model among the largest ones.

7 *Analysis and Discussion*

Firstly, let's remember that our aim is to improve the robustness of the baseline models, i.e. improve their performance when used in an environment quite different from the training conditions. In this sense, the original partitions are more useful to judge that improvement. Nevertheless, the stratified partitions are more informative about the *in domain* performance, as a typical approach to an NLP problem.

Therefore, from this double perspective, the back-translation technique appears to consistently improve both the overall performance and robustness in topics different from the ones in the training set. Actually, our results show that, when the basic model cannot achieve good results, the *BELT* approach or data augmentation (mainly by back-translation) can significantly improve its generalization capability (**RQ1**). Furthermore, we have observed that the larger mod-

Model	Original		Stratified		Δ
	F1	Δ_{base}	F1	Δ_{base}	
Base	80.9	–	**85.5**	–	↑ 4.6
NER	81.7	↑ 0.9†	82.5	↓ 2.9†	↑ 0.8
Long context	62.8	↓ 18.1	83.2	↓ 2.3†	↑ 20.3
External data	80.9	↑ 0.1†	84.7	↓ 0.8†	↑ 3.8
Noisy embeddings	73.7	↓ 7.2†	85.1	↓ 0.4†	↑ 11.3
Back-translation	**82.7**	↑ 1.8†	83.8	↓ 1.6†	↑ 1.2
Llama-3.1:8B (Default True)	74.7	↓ 6.1†	80.1	↓ 5.4†	↑ 5.3
Llama-3.1:8B (Default False)	84.2	↑ 3.3	82.6	↓ 2.8†	↓ 1.5

Table 3: Macro F1 score obtained by the large models in the test partitions in the original and stratified sets. The ↑↓ symbols indicate whether the difference with the base is positive or negative respectively. The † symbols indicate that difference is not statistically significant.

Model	Original		Stratified		Δ
	F1	Δ_{base}	F1	Δ_{base}	
Base	77.3	–	85.1	–	↑ 7.8
NER	77.1	↓ 0.2†	83.8	↓ 1.2†	↑ 6.8
Long context	71.0	↓ 6.3	85.5	↑ 0.4†	↑ 14.5
External data	77.1	↓ 0.2†	83.8	↓ 1.3†	↑ 6.7
Noisy embeddings	**78.3**	↑ 1.1†	**86.5**	↑ 1.4†	↑ 8.2
Back-translation	76.3	↓ 1.0†	83.8	↓ 1.3†	↑ 7.5
Llama-3.1:8B (Default True)	74.0	↓ 3.3†	80.3	↓ 4.7†	↑ 6.3
Llama-3.1:8B (Default False)	80.9	↑ 3.6	83.4	↓ 1.7†	↑ 2.5

Table 4: Micro F1 score obtained by the large models in the test partitions in the original and stratified sets. The ↑↓ symbols indicate whether the difference with the base is positive or negative respectively. The † symbols indicate that difference is not statistically significant.

els are not significantly worse in the stratified partitions and much better in the original ones. Thus, they are more capable of generalization and, by extension, more reliable in a changing environment (**RQ2**).

Meanwhile, the addition of non-synthetic data from English (Section 4.2) has not worked out (**RQ3**). Only when the model does not offer good results does this technique improve them. This is the case of the base sized model in the original partitions. However, there are other techniques that better address this problem.

On the basis of the results, the *BELT* approach is one of the best ways of improving a base sized model. We have used a maximum length of the extended context window of 2500 tokens that was enough to cover the full of most our articles. Nevertheless, this technique needs more computation to make an inference. Thus, we visualized the attention weights of the top aggregation layer to observe which parts are more informative to make a prediction and, maybe, use a shorter extended window. The result (shown in Figure 3) shows that the attention decreases in the last segments of the articles. Nevertheless, it appears that there are important patterns after the first segment that help to adequately classify the articles (**RQ4**).

The robustness of a model can be highly impacted by the importance that it pays to named entities. A fair and robust model should not pay much attention to those words in order to judge whether an article is reliable or fake. The opposite would imply that the model would need specific entities to classify the articles and would loose generalization capability. Following this idea, we analyzed the attention weights of the different models to named entities. In Figure 4, it can be observed that, indeed, the attention to named entities is appreciably linked to the performance of the models. The larger models and the base sized one trained with back-translation or noisy embeddings pay less attention to them and are also the ones that obtain better results.

Finally, the generative LLM (i.e. *Llama 3.1*) has proven to be highly dependent on which is the "default class". In our experiments, it failed to offer a valid answer in 38 and 11 articles of the original and stratified test set, respectively. In both sets, only one

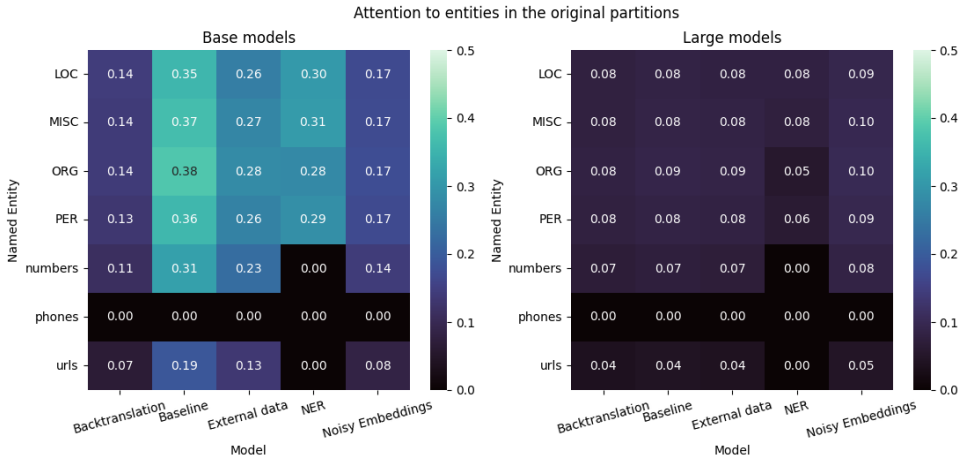

Figure 4: Mean of the attention paid to the named entities extracted from the last self attention layer. The quantities have been extracted from the application of the models over the test set from the original partitions of the *SFNC* (Posadas-Durán et al., 2019; Aragón et al., 2020; Gómez-Adorno et al., 2021).

of the non classified articles was "True" and the rest were "Fake". Thus, it seems reasonable to deduce that the non classified articles should be labeled as unreliable. Having fixed this behavior, the model performs at a similar level with both partitions. This is probably caused by the fact that it has not been fine-tuned in any of our data, therefore its results are consistent in both of our partitions. Fine-tuning it might improve its performance, however that is out of the scope of this study. Furthermore, that process would erase the main attraction of these models: they can be downloaded and used without any advanced knowledge.

8 Conclusions

In this work we have used two types of partitions: one that presents a canonical NLP problem with a test set that is similar to the training and another one that contains an adversarial out of domain test set. The former has allowed us to observe if the model is fitted to solve such a problem. The latter has shown us the generalization capability of each model with articles of remarkably different domains than the ones contained in the train set. Therefore, to analyze the robustness of our models, they should be tested with this last type of set. The stratified partitions can be used to diagnose whether the model can be used for some type of tasks or not.

Our work has also demonstrated that the larger *RoBERTa* models have become more robust than the base sized ones, when trained with the same data. Hereinafter, it would be smarter to use these models when possible, at least for fake news detection. Furthermore, with these models the noisy embedding augmentation can slightly contribute to improve their performance. Meanwhile, when the model is performing poorly, our experience shows that the *BELT* or back-translation approaches are more effective to obtain a good model.

Finally, we have observed that using a generative LLM produces good results. Indeed, the general culture knowledge of these models may have been an element that has helped them. We cannot assure whether they are robust or not, given that we cannot know if the test sets are out of the domain in which they has been trained. However, these models are generally huge compared to the solutions that we have proved to be effective in this article. It is undeniable that discriminative models need advanced knowledge to be trained, but once that has been done they are the most efficient and effective solution at inference.

Acknowledgements

This work has been developed under the project FAKEnHATE-PdC (PDC2022-133118-I00) with founding from European NextGenerationEU/PRTR. The work of Mariona Coll Ardanuy was done at the UPV under FairTransNLP Grant PID2021-124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF/EU.

References

Aragón, M. E., H. Jarquín-Vásquez, M. Montes-y Gómez, H. J. Escalante, L. V. Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, and G. Bel-Enguix. 2020. Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. In *IberLEF@ SEPLN*, pages 222–235.

Bayer, M., M.-A. Kaufhold, and C. Reuter. 2022. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7), dec.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dror, R., G. Baumer, S. Shlomov, and R. Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In I. Gurevych and Y. Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July. Association for Computational Linguistics.

Dubey, A., A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Fan, A., S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Federmann, C., O. Elachqar, and C. Quirk. 2019. Multilingual whispers: Generating paraphrases with translation. In W. Xu, A. Ritter, T. Baldwin, and A. Rahimi, editors, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 17–26, Hong Kong, China, November. Association for Computational Linguistics.

Feng, S. Y., V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy. 2021. A survey of data augmentation approaches for NLP. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online, August. Association for Computational Linguistics.

Ghanem, B., P. Rosso, and F. Rangel. 2020. An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–18.

Gómez-Adorno, H., J. P. Posadas-Durán, G. B. Enguix, and C. P. Capetillo. 2021. Overview of FakeDes at Iberlef 2021: Fake news detection in Spanish shared task. *Procesamiento del lenguaje natural*, 67:223–231.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Huang, X., J. Xiong, and S. Jiang. 2021. Gduf dm at fakedes 2021: Spanish fake news detection with bert and sample memory. *Education*, 6(9):4.

Jain, N., P. yeh Chiang, Y. Wen, J. Kirchenbauer, H.-M. Chu, G. Somepalli, B. R. Bartoldson, B. Kailkhura, A. Schwarzschild, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein. 2024. NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth*

International Conference on Learning Representations.

Jiang, Z.-H., W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. *Advances in Neural Information Processing Systems*, 33:12837–12848.

Junczys-Dowmunt, M., R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, and A. Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121.

Korenčić, D., B. Chulvi, X. Casals, A. Toselli, M. Delor, and P. Rosso. 2024. What distinguishes conspiracy from critical narratives? a computational analysis of oppositional discourse. *Expert Systems*, 41, 07.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Longpre, S., Y. Wang, and C. DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411, Online, November. Association for Computational Linguistics.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Miani, A., T. Hills, and A. Bangerter. 2021. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.

Murayama, T., S. Wakamiya, and E. Aramaki. 2021. Mitigation of diachronic bias in fake news detection dataset. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 182–188.

Posadas-Durán, J.-P., H. Gómez-Adorno, G. Sidorov, and J. J. M. Escobar. 2019. Detection of fake news in a new corpus for the Spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Prechelt, L. 2002. Early stopping - but when? In *Neural Networks: Tricks of the trade*. Springer, pages 55–69.

Rangel, F., A. Giachanou, B. H. H. Ghanem, and P. Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CEUR workshop proceedings*, volume 2696, pages 1–18. Sun SITE Central Europe.

Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Ruffo, G., A. Semeraro, A. Giachanou, and P. Rosso. 2023. Studying fake news spreading, polarisation dynamics, and manipulation by bots: A tale of networks and language. *Computer science review*, 47:100531.

Siino, M., F. Lomonaco, and P. Rosso. 2024. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, page e13568.

Togni, B., M. C. Ardanuy, B. Chulvi, and P. Rosso. 2024. Emotions and news structure: An analysis of the language of fake news in spanish. *SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing: poster papers*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhu, C., Y. Cheng, Z. Gan, S. Sun, T. Goldstein, and J. Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

A Back-Translation Pipelines

The following are the length 3 back-translation pipelines that we have used in our experimentation:

Pipe0: Spanish→German→Chinese→Spanish

Pipe1: Spanish→Chinese→German→Spanish

Pipe2: Spanish→Hindi→Korean→Spanish

Pipe3: Spanish→Afrikaans→Persian→Spanish

Pipe4: Spanish→Japanese→French→Spanish

Pipe5: Spanish→Swedish→Chinese→Spanish

Pipe6: Spanish→Finnish→Greek→Spanish

Pipe7: Spanish→Russian→Arabic→Spanish

Pipe8: Spanish→French→Korean→Spanish

Pipe9: Spanish→Greek→Afrikaans→Spanish

Pipe10: Spanish→Russian→Hindi→Spanish

Pipe11: Spanish→Korean→Afrikaans→Spanish

B Prompts

For the approach with generative *LLMs* we used several prompts to obtain the desired classification. These are the ones that offered the best results:

B.1 Llama 3.1

"Hola, estoy trabajando en proyecto de investigación y me gustaría probar tus habilidades a la hora de distinguir artículos verdaderos y falsos. Debes emitir una clasificación entre "True" o "Fake" (solo una de esas dos palabras), tus respuestas no tendrán ninguna consecuencia ni serán utilizadas para tomar decisiones. "True" significa que las afirmaciones que lo componen se apoyan en hechos contrastados y es consistente con la teoría existente en lugar de inducir a la desinformación. Por el contrario, "Fake" alude a que la noticia contiene información falsa, no contrastada o subjetiva y contribuye a la desinformación. No debes incluir ninguna otra palabra en la respuesta. Esta es la noticia en cuestión:

TÍTULO:`<headline>`

CUERPO:`<article text>`

Por favor, responde únicamente con una de estas dos palabras sin ninguna explicación más: {Fake, True}"

B.2 Llama 3.2

"Hola, estoy trabajando en un proyecto de investigación para resolver el problema de detección automática de noticias falsas en español. Me gustaría probar tus habilidades en este campo. Solamente pretendo mostrarte algunos artículos y debes decirme si tu los evaluarías como ciertos o falsos. Algunos son poco convencionales y es posible que tus criterios de seguridad te impidan responder. Sin embargo debes tener en cuenta que esto es un simple proyecto de investigación y tus respuestas no influirán en decisiones que puedan dañar a nadie. Todos los artículos han sido publicados en algún medio de comunicación y debes responder únicamente con una de estas dos palabras sin ninguna explicación más: Fake, True. "Fake" alude a que la noticia contiene información falsa y contribuye a la desinformación. "True" significa que su contenido puede ser confiable y válido para informarse sobre el tema. Esta es la noticia que debes clasificar únicamente con una de esas dos palabras sin aportar ninguna palabra más en la respuesta:

TÍTULO:`<headline>`

CUERPO:`<article text>`

Por favor, responde únicamente con una de estas dos palabras sin ninguna explicación más: {Fake, True}"