Enhancing Model Performance through Translation-based Data Augmentation in the context of Fake News Detection

Anonymous EACL submission

Abstract

The rapid development of social media in recent years has encouraged the sharing of vast amounts of data, but also the propagation of fake news. This has pushed the scientific community to focus on this phenomenon, particularly those working on natural language processing, by developing detection tools to combat fake news. At the same time, most studies have focused on languages with a high resource content (corpora). The purpose of this paper is to shed light on low-resource languages, in particular the Algerian dialect, through an experimental study with two objectives. The first one is to verify if the automatic translation from Modern Standard Arabic (MSA) to the Algerian dialect can be considered as an approach to increase the resources in the Algerian dialect especially with the rise of large language models (LLMs). The second is to verify the impact of the translation-based data augmentation method on fake news detection by using transformer-based Arabic pre-trained models in different data augmentation configurations. We have discovered that LLMs are capable of generating translations that closely resemble human translations. In this study, we demonstrate that data augmentation can result in a saturation and decline in model performance due to the introduction of noise and variations in writing styles.

1 Introduction

Social media are playing an increasingly important role in our professional and personal lives, particularly through their ability to keep us informed of events reported by our friends and contacts. It has become common for important news to be spread first on social media before being processed by the traditional media. The speed with which information spreads, combined with the number of people who receive it, defines the virality of information. But this virality, a major characteristic of social media, has a downside: users rarely check the veracity of the information they share.

It is therefore common for fake news to circulate in order to manipulate or mislead people. Consequently, the use of natural language processing to automate the detection of fake news, which is formulated in most studies as a classification problem, has received a lot of attention from researchers (Oshikawa et al., 2018).

However, most of the studies have focused on specific languages with high resource content such as English (Faustini and Covoes, 2020), chinese (Du et al., 2021), Modern Standard Arabic (MSA) (Fouad et al., 2022), Spanish (Martínez-Gallego et al., 2021), Korean (Kang et al., 2022), and Russian (Kuzmin et al., 2020). Compared with languages or sub-varieties of languages such as dialects, which are considered to be low-resource languages, where studies are rare or non-existent.

The low-resource languages are turning to the development of models based on machine learning and deep learning techniques such as data augmentation that can be defined as any method for increasing the diversity of training examples without explicitly collecting new data to overcome this lack of resources (Pellicer et al., 2023), (Pellicer et al., 2023).

Since manual data augmentation costs time and effort, most of the work has been oriented toward automatic data augmentation, especially with the rise of LLMs.

The main goal of this study, is to verify a hypothesis aimed to determine if the machine translation from MSA to the Algerian dialect can be considered as an appropriate approach to increase the quantity of data and build efficient models in the context of the detection of fake news.

2 Related Work

2.1 Fake news detection in Algerian dialect

Recently, a few number of studies have focused on Algerian dialectfake news detection. (Righi et al., 2022) conducted their study on a political rumor case involving the health of the Algerian President, which occurred between late 2020 and early 2021, they applied a transfer learning approach. They utilized three different models: mBert, XLM-Roberta, and AraBERT, to analyze a dataset of 3,147 YouTube comments collected via the YouTube API v3. The authors attributed AraBERT's performance to its training on various NLP tasks, including sentiment analysis, as well as its training on a substantial Arabic corpus containing 24GB of text and a vocabulary of 64k words.

(Bousri et al., 2022) focused their study on "infobesity" which refers to the overwhelming volume of news and information received through social media, which has led to the spread of rumors and misinformation in Algerian Arabizi. They propose a method based on rumors and user responses, using attention mechanisms. They used various classification models and textual representations to examine the relationship between rumors and user reactions. Results showed that incorporating associations improved the performance of the LSTM model by over 10% when using Word2vec and n-grams bag representations.

2.2 Data augmentation methods in NLP

Data augmentation is used to create synthetic training data, especially when the original dataset is insufficient. It encompasses a range of techniques, from simple rule-based methods to more advanced, learnable generation-based methods. Recently, many overviews have been done on data augmentation methods. In the context of NLP, these methods can be grouped into 03 categories:

- **Paraphrasing**: This category likely involves methods that generate variations of the text by rephrasing sentences while preserving the same semantics or meaning. Paraphrasing consists of several levels, including lexical paraphrasing, phrase paraphrase, and sentence paraphrase. among these methods: Thesauruses (Coulombe, 2018a) and (Wei and Zou, 2019). Semantic embeddings (Wang and Yang, 2015), (Liu et al., 2020). Language models (Wu et al., 2019), (Lowell et al., 2020). Rules (Coulombe, 2018b), (Regina et al., 2020). Machine translation (Fabbri et al., 2020), (Nishikawa et al., 2020). Model generation (Hou et al., 2018), (Kober et al., 2020).

- **Noising**: the methods aim to introduce random or simulated noise to the text. This noise could include spelling errors, typos, or other types of modifications that maintain the validity of the data while adding diversity. we can mention: Swapping (Longpre et al., 2020), (Dao et al., 2019). Deletion (Rastogi et al., 2020), (Peng et al., 2020), and (Yan et al., 2019). Insertion (Wei and Zou, 2019) and (Yan et al., 2019). Substitution (Song et al., 2021) and (Louvan and Magnini, 2020).

- **Sampling**: methods could involve selecting or generating additional data points from the same distribution as the original data, possibly by using different techniques or sources. We have : Rules (Min et al., 2020), (Shakeel et al., 2020). Non-pre-trained models (Raille et al., 2020), (Yoo et al., 2019). Pre-trained models (Anaby-Tavor et al., 2020), (Kumar et al., 2020). Self-training (Thakur et al., 2020),(Montella et al., 2020). Mixup (Zhang et al., 2017), (Guo et al., 2019).

3 The Arabic Algerian Dialect and Its Challenges

The Algerian dialect, a member of the Maghrebian language family, serves as a vital counterpart to MSA within the Algerian community. It inherits certain characteristics from MSA, such as the absence of diacritics, agglutination, and flexible word order (Saadane and Habash, 2015). Despite being the main way of communication for almost 80% of Algerians (Kerras et al., 2019), it is often disregarded in academic discussions, especially in research on web-based disinformation. This dialect poses unique issues in the field of NLP, which are exacerbated by a lack of resources and specialized datasets designed for it. The dialect's distinctive phonetic, morphological, and orthographic characteristics, which were influenced by a variety of historical influences, including French, Turkish, and Spanish, emphasize its complexity furthermore. Table 1 shows some examples about borrowed words from several languages that affected the Algerian dialect.

Language	Words	English
Turkish	Tabşi	Plate
	Bālāk	Maybe
Spanish	Qmažža	Shirt
	Spardīna	Snickers
Italian	Fat'cha	Face
	Bnine	Delicious
French	Silūn	Cell
	Ravitayma	Provisions

Table 1: Words in Algerian Dialect borrowed from other languages.

An additional noteworthy aspect of the Algerian dialect is the adaptable textual feature. The absence of a tightly regulated orthography in ALGDIA leads to fluidic spellings that emphasize its dynamic nature while possibly making linguistic formalization challenging. As shown in Table 2, this adaptation includes the common practice of code-switching, particularly between Arabic and French, the adoption of the Arabizi script, which blends Latin characters and numerals to represent Arabic phonemes, and the adoption of a white dialect script involves dialect native with replacing words that are known only in the region, or loanwords from other languages, with more MSA words, thereby making it easier for non-locals to understand.

4 Machine Translation Models: Comparative Study

In the field of NLP, translation models take on heightened importance, by developing efficient and precise translation systems. This section examines the suitability and effectiveness of 04 machine translation models through a thorough comparison analysis to translate MSA text to Algerian text. There is an urgent need to comprehend how different models perform in comparison to one another, particularly when tailoring them to certain tasks like data augmentation, especially for low resources languages.

To fully leverage the capabilities of these models, a rigorous assessment is crucial to ensure they align with the diverse needs of different applications. In our research, we assess machine translation models in a zero-shot mechanism, based on the overlap between the predicted and reference sentences (by calculating the blue score), as well as contextually, selecting the most suitable model for each sentence which was done by 3 Algerian experts researchers. In conclusion, we prioritize a model that accurately captures the context of the input sentence. For a sensitive task such as fake news detection, it's imperative to employ a translation model that not only preserves the context and meaning of the sentence but also ensures its coherence throughout. Many machine translation models have developed throughout time, each with a unique method for translating languages. Our research focuses on two well-known large language models and ours which were chosen due to their pertinence to the task of data augmentation:

- **Nllb-200-distilled-600M** (Costa-jussà et al., 2022): A multilingual neural machine translation system that encompasses 200 languages, based on the transformer encoder-decoder framework. For our purposes, we leveraged two versions of this model. One is tailored for translating MSA to Moroccan text, while the other for MSA to Tunisian. This choice was driven by the overlapping characteristics observed among these two dialects and Algerian dialects.

- **GPT-4**: The latest in the GPT series from OpenAI, this model stands out for its capability to comprehend and generate both natural language and code. Pre-trained on a vast corpus encompassing multiple languages and domains, the exact details of its training data and architecture remain unclear. However, it is generally understood to be an autoregressive language model with a foundation in the transformer architecture (Vaswani et al., 2017). In our endeavors to generate an automatic translation using the GPT-4 model, we made several attempts to find the most suitable prompt. We employed the prompt *"Please translate the following to the exact Algerian dialect, please don't confuse it with any Darija dialects such as Moroccan and Tunisian, and try your best please:"* through an API call, with the input being MSA text.

- **Our custom model**: This model was created by optimizing the AraBART model (Eddine et al., 2022) to focus on converting MSA text to Algerian. It was trained using a set of 11,722 parallel sentences from our curated and annotated dataset.

Writing Cases	Algerian Sentence
Arabic Letters	اوبك قالت دوك تطلع دوموند
Latin Letters	OPEC galet douk tetlaa demande
Arabizi Style	OPEC galet douk tetla3 demande
Code-switching Style	demande قالت دوك تطلع OPEC
White dialect style	اوبك قالت دوك يصعد الطلب

Table 2: Different ways of writing the Algerian dialects in social media for the sentence *OPEC said demand will rise* taken from our dataset.

Dataset For a rigorous and unbiased comparison, using a comprehensive and widely recognized dataset is crucial. In this study, we've chosen the "MADAR CORPUS-25" (Bouamor et al., 2018), which consists of 2K sentences for each 25 distinct city dialects. Each sentence is paired with 25 parallel translations, including both the Algerian dialect and the MSA version. After thorough verification and data cleaning, we retained 1,588 parallel sentences. As there is no training phase in this experiment, all 1,588 sentences served as the test set for zero-shot prediction. During testing, each model is tasked with taking an MSA sentence and generating its Algerian translation. Ultimately, to assess the performance of these models, we compute the BLEU score by comparing the predicted sentence to the reference sentence from the dataset. Since the BLEU score might not reflect contextual accuracy, we also manually evaluate context by selecting the best model for each sentence, and aggregating the scores to determine the overall performance of each model.

Model	BLUE score	Scoring
GPT-4	9.42	775
Our Model	5.952	447
NLLB TUN	8.226	412
NLLB MOR	8.94	479

Table 3: Comparative Analysis of Different Models on MSA-AD translation Based on BLUE Score and Scoring out of 1588 sentences.

Based on the findings shown in Table 3, GPT-4 performs the best of the models assessed. Despite Algerian being a low-resource language, GPT-4's robust capacity to produce translations that closely match human-like phrasings and its proficiency in handling the complex structure of the Algerian dialect are reflected in its higher BLEU and raw scoring. Our customized model might not be as fluent and precise as models with higher scores especially with its limited vocabulary. Future iterations might profit from more adjusting or from using larger training datasets that are tailored to the Algerian dialect. The Tunisian dialect-specific NLLB model produced a BLEU score of 8.226 and a raw score of 412 out of 1,588. Given the linguistic similarity of the two dialects, this suggests that it is reasonably proficient in translating the Algerian dialect. Similar to its Tunisian counterpart, the NLLB model for the Moroccan dialect exhibits promise with a BLEU score of 8.94, just below GPT-4. It performed in the middle of the tested models, accurately translating 479 sentences. This indicates that, like the NLLB TUN, it might occasionally lack accuracy even if it offers fluid translations. The outcomes emphasize the difficulties in machine translation for languages with limited resources. When dealing with particular dialects like Algerian, even models like GPT-4, recognized for their huge training data, can run into problems. It is clear that, although BLEU ratings indicate translation fluency, the raw score is essential to comprehend the correctness and practical usefulness of these models.

5 Experiments and Results

5.1 Dataset description

In this section, we provide a comprehensive description of the fake news dataset utilized in our study, including its sources, collection methods, annotations, and pertinent statistics. The foundational source of our dataset is the "Khouja Corpus" (Khouja, 2020), which encompasses a collection of MSA text data related to political content taken and paraphrased from the ANT corpus v1. Each sentence within the dataset was pre-labeled as either "real" or "fake" to facilitate supervised learning tasks. This corpus formed the basis for our research aimed at identifying fake news in the Algerian dialect. A total of 4,429 sentences from this corpus were taken for our analysis. The textual data from the corpus were unbalanced, with 3K real and 1,429 fake sentences; thus, we employed down-sampling. We utilized 1,429 sentences each of fake and real categories, of which 1K sentences per category were allocated for training, while the remaining 429 sentences from each category were designated for testing. Further details about the sentence length in the original, manually translated, and automatically translated datasets can be found in Table 4. The corpus was manually translated into the Algerian dialect using the most known vocabulary words in the Algerian community and social media. Moreover, we employed the state-of-the-art language model, GPT-4, utilizing the version available as of September 25th—for automated translation assistance. A post-processing step was undertaken after generating the translation with GPT-4, which consisted of removing English expressions such as *"The Algerian dialect is:"* and quote marks.

	MSA	**Manual**	**Automatic**
Max tokens	22	23	49
Min tokens	2	2	2
Average tokens	9.12	9.54	9.91

Table 4: Statistics for the various subsets within the Fake news dataset.

Both the manually-translated and the original MSA dataset contained no instances of English, code-switching, or Arabizi writing styles. However, the data generated by GPT-4 (automatic translation) included some Arabizi expressions and French translations.

5.2 Experimental settings

In our experimental framework, we have adopted a suite of metrics, including precision, recall, and F1-score, to rigorously evaluate the performance of the models in the context of binary text classification tasks. This meticulous selection of metrics ensures a well-rounded examination and subsequent analysis of the model's capability to adeptly categorize the textual data into respective binary classes, aligning with the specificity and sensitivity requisites of our study.

As presented in Table 5, we employed three pre-trained transformer-based models for our experiments: AraBERTv02 (Antoun et al., 2020), MARBERTv2 (Abdul-Mageed et al., 2020), and DziriBERT (Abdaoui et al., 2021), each chosen based on their training data and relevance to the dialects under scrutiny in our research. AraBERTv02 was selected for its predominant training on MSA and several dialects. MARBERTv2 was chosen due to its extensive training on the Maghrebi dialect, which encompasses Algerian, Moroccan, and Tunisian dialects, and our findings that indicated GPT-4 exhibited a proclivity towards producing a blend of Algerian and Moroccan dialects. Lastly, DziriBERT was opted for its focus on the Algerian dialect. The comparison of these models provided a strong framework for examining the nuances and effectiveness of each in the particular context of identifying false news within the dialects in question.

Model	**Params.**	**Tokens**	**Vocab.**
AraBERTv02	136M	8.6B	64k
MARBERTv2	163M	6.2B	100k
DziriBERT	124M	20M	50k

Table 5: The selected Arabic pre-trained models in terms of parameters number, size of training data (tokens), and the vocabulary size.

The fine-tuning and testing of models, was executed on the Google Colab platform, utilizing a Tesla T4 - 16GB GPU to harness optimal computational efficiency. Hyperparameters were meticulously fine-tuned leveraging the test set to ensure model efficacy. Specifically, the Adam optimizer (Kingma and Ba, 2014) was employed, with the learning rate varying between 3×10^{-5} and 6×10^{-5}, a batch size varying between 16 and 32, and a seed of 42, across five epochs. Throughout all our experiments, we utilized the Huggingface Transformers library (Wolf et al., 2020).

5.3 Automatic translation evaluation

This section provides results concerning the impact of both manual and automatic translation on model performance in a fake news detection task involving Algerian dialect text. Tables 6 and 7, and Figure 1 present the results for the three models under study.

Model	Pre.	Rec.	F1
AraBERTv02	0.574	0.415	0.481
DziriBERT	0.539	0.9	**0.674**
MARBERTv2	0.685	0.12	0.204

Table 6: Manual translation evaluation results.

Model	Pre.	Rec.	F1
AraBERTv02	0.515	0.81	0.63
DziriBERT	0.573	0.745	**0.647**
MARBERTv2	0.526	0.822	0.641

Table 7: Automatic translation evaluation results.

In the manual translation evaluation table, we can observe that the best performing model is DziriBERT, with the highest F1 score (0.674) predominantly due to its high recall. The worst performing model is MARBERTv2, largely attributed to its extremely low recall, though it has commendable precision. On the other hand, DziriBERT, while still the best-performing model in the automatic translation evaluation, exhibits a slightly reduced F1 score (0.647) compared to its performance with manual translation. The most improved model is MARBERTv2, which displays a marked improvement, especially in recall and F1 score, when using automatic translation. In the case of the largest model (AraBERTv02) in terms of parameters and trained data size, an improved F1 score (0.63) in the automatic translation as compared to manual translation, even though the precision is slightly compromised (0.515) with a notable increase in recall (0.81).

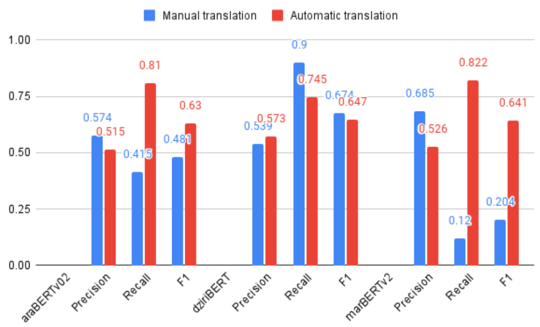

Figure 1: A comparison between manual and automatic translation across the 3 models.

5.4 Data augmentation evaluation

To validate our hypothesis that automatic translation can serve as an effective data augmentation method for context-sensitive tasks, we employed another dataset: the 2nd version of the ANT corpus (Chouigui et al., 2017). The ANT corpus, curated from diverse sources, was designed to reflect the diversity of the Khouja corpus, with a focus on MSA text. From this corpus, we selected 4K pre-annotated sentences from the political domain. Given that there were no explicit indications or features identifying the domain category, we implemented a rule-based algorithm. This algorithm determines whether an MSA sentence is related to the political domain by matching it against a list of 24 political keywords. Subsequently, these sentences were processed through the GPT-4 LLM to generate their automatic translation versions, using the same prompt mentioned in Section 4. Finally, we selected four subsets configurations comprising 500, 1000, 1500, and 2000 sentences, which represent 25%, 50%, 75%, and 100% of the training data from the Khouja dataset, respectively. This was done to analyze the effect of varying data sizes on model performance.

Model	Precision	Recall	F1
Size = 500 (25%)			
AraBERTv02	0.498	0.982	**0.661**
DziriBERT	0.523	0.770	0.623
MARBERTv2	0.514	0.862	0.644
Size = 1000 (50%)			
AraBERTv02	0.5	1.0	**0.667**
DziriBERT	0.507	0.957	0.663
MARBERTv2	0.5	0.965	0.658
Size = 1500 (75%)			
AraBERTv02	0.5	1.0	**0.666**
DziriBERT	0.523	0.855	0.649
MARBERTv2	0.503	0.865	0.636
Size = 2000 (100%)			
AraBERTv02	0.503	1.0	**0.67**
DziriBERT	0.506	0.835	0.63
MARBERTv2	0.497	0.86	0.63

Table 8: Automatic translation evaluation results for different data augmentation configurations.

The table 8, displays the data augmentation results obtained through automatic translation for the three models under consideration. It details the outcomes associated with each data augmentation configuration.

As shown in table 8, AraBERTv02 has consistently high recall, hitting the 1.0 for sizes above 500. Its precision hovers around the 0.5 which leads to capture a significant number of irrelevant features. Its F1 score remains relatively consistent across dataset sizes. For the DziriBERT model,The recall is lower compared to AraBERTv02 but is still high. Notably, its recall decreases slightly as dataset size increases showing a challenge while scaling with more data. Precision remains slightly above 0.5,

which is relatively stable but not significantly different from AraBERTv02. The F1 score, while comparable to AraBERTv02 in smaller dataset sizes (50%), drops a bit for larger dataset sizes. Lastly for MARBERTv2, the recall is generally high but less consistent across the augmentation. It shows some decrease as dataset size increases. Its precision is around the 0.5, similar to the other two models. The F1 score decreases slightly as the dataset size increases, which might be a concern if scalability is a priority.
In terms of precision, all three models have precisions hovering around the 0.5 may be caused by the capturing of noise data. In case of recall, AraBERTv02 has the highest and most consistent recall, while DziriBERT and MARBERTv2 have varying recalls, especially with the increase in dataset size. (75% and 100%). The F1 scores for all models are relatively close, with none surpassing the 0.67 even in the largest dataset configuration. This suggests that while the models can capture relevant data, there's room for improvement in refining their outputs and reducing the occurrence of false positives/negatives. Figure 2 illustrates a comparison of training the DziriBERT model with and without using data augmentation via automatic text translation.

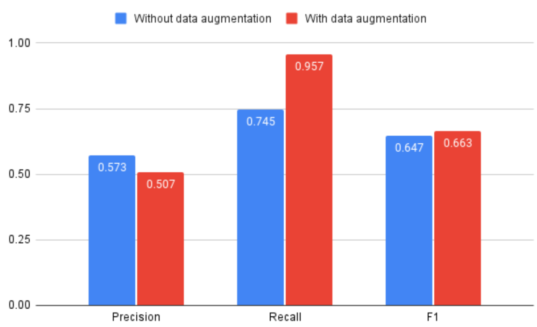

Figure 2: Comparison of DziriBERT model training with and without data augmentation using automatic text translation at 50%(1000 sentences) configuration.

6 Discussion

In this discussion section, we will explore the key findings and implications of these results.

6.1 Impact of Data Augmentation Size

Based on the results presented in table 8, we observed that AraBERTv02 consistently demonstrates a robust performance across all augmentation sizes, achieving an F1 score that notably remains relatively stable despite the increases in data size. This could signify a potential saturation in learning from the additional data. Conversely, DziriBERT and MARBERTv2 illustrate fluctuation in their performance metrics as data augmentation scales. Notably, DziriBERT maintains a competitive edge in precision compared to the other models across different data sizes, even outperforming AraBERTv02 in certain instances. However, its recall and F1 score appear to be more sensitive to changes in data size, signaling a possible challenge in maintaining a balanced precision-recall trade-off with varied data inputs. Although MARBERTv2 often displays sensitivity in precision when dealing with different augmentation sizes, which can be resulted on by the variety of writing styles and new words (noise), it still has some fair consistency in F1 scores. These findings show that data augmentation, especially with larger datasets, is effective in enhancing recall, which is crucial for certain applications such as information retrieval, but it might come at the cost of precision. It is also shown that, at some point, the models exhibit low performance due to the noise and out-of-vocabulary (OOV) words encountered during training with a new distribution dataset.

6.2 Model Performance

Among the models, AraBERTv02 consistently outperforms the other two models across all data augmentation sizes, achieving the highest F1 score in most cases. Not far behind, we find DziriBERT, which, despite its significantly smaller size and data, demonstrates higher precision, showcasing its robustness in handling augmented data. As shown in table 3, the results underscore the importance of model selection for specific tasks and the implications of data augmentation sizes on model performance. While larger datasets might seem beneficial, the performance doesn't scale linearly with dataset size due to the changes in the writing style and the augmentation of OOV words. Each model has its strengths and potential areas for optimization. Depending on the specific requirements (e.g., if high recall is more critical than precision), one might favor one model over the others.

6.3 Hyperparameter Configuration

The choice of hyperparameters plays a crucial role in the models' performance. A comparison of the hyperparameter configurations for the different augmentation sizes suggests that adjustments in learning rate (lr), batch size, and dropout can influence the models' performance. However, finding the optimal hyperparameters is a complex and iterative process, and these configurations provide a starting point for further fine-tuning.

6.4 Manual vs. Automatic translation

In the domain of fake news detection involving translated text, models such as DziriBERT and MARBERTv2 exhibit a notable precision-recall trade-off when evaluated using manual translation, suggesting a potential necessity for threshold adjustments in classification to better balance these metrics. Notably, the effect of automatic translation tends to improve recall across models, possibly by introducing robustness or leniency in recognizing relevant features for fake news detection in the translated text. However, it can quietly degrade precision, suggesting possible noise introduction during the automated process by providing OOV words. Additionally, there are noticeable performance differences, as shown by MARBERTv2, which exhibits a considerable variation in performance amongst translation methods, suggesting a sensitivity to translation quality or style. In general, models yield superior results (reflected in higher F1 scores) with automatic translation. However, DziriBERT constitutes an exception, underperforming a bit slightly compared to its counterpart using manual translation.

6.5 Practical Implications

These results have practical implications for various natural language processing applications. The use of automatic translation as a data augmentation technique is a promising approach to enhance model performance, especially when high recall is essential. However, the choice of model architecture, data size, and hyperparameters should be carefully considered based on the specific requirements of the application.

6.6 Limitations

The choice of model architecture and data size should be tailored to the specific goals of the application, with a focus on achieving the right balance between precision and recall. Additionally, hyperparameter tuning is crucial for maximizing the potential of these models. Prompt engineering for LLM (GPT-4) can also be a limitation in this study, as the translation of LLM may not always be stable. Exploring alternative methods to tune the prompt could potentially yield higher quality translations and greater stability. The lack of standardization in Arabic dialects in general, and the Algerian dialect in particular, influences the behavior of the model. This influence arises from the multitude of writing styles, which can result in alternative meanings for the same sentence or the generation of OOV words. Overall, these findings contribute to our understanding of how to leverage data augmentation using automatic translation for enhancing the capabilities of fake news detection models in low resource languages.

7 Conclusion

In this study, we have delved into the complex nature of data augmentation using automatic translation techniques for improving the performance of fake news detection models in low-resource languages, with a specific focus on the Algerian dialect.

Based on the experiments conducted in this study, we have discovered that the comparison between manual and automatic translation reveals that LLMs like GPT-4 are capable of generating translations that closely resemble human translations and, at times, even surpass them in terms of diversity of writing style, ultimately resulting in improved model performance. However, it is important to note that while larger datasets can enhance recall, they may simultaneously decrease precision. Additionally, the model's performance tends to saturate when integrating new distribution datasets, primarily due to the introduction of noise and OOV words. Moreover, when evaluating different model architectures, we found that the implications of data augmentation sizes on model performance emphasize that larger datasets do not always translate into proportionally better performance, as the changing writing style and OOV word augmentation can introduce complexities.

Nonetheless, our study has certain limitations. The lack of standardization in Arabic dialects, including the Algerian dialect, introduces variability in model behavior, arising from diverse writing styles and alternative meanings for the same sentences. Additionally, prompt engineering for LLM can be a challenge, as translation stability is not guaranteed. While we have achieved promising findings, there is room for further enhancement through more robust training, improved translation models, optimized parameters, and extended training.

References

Amine Abdaoui, Mohamed Berrimi, Mourad Oussalah, and Abdelouahab Moussaoui. 2021. Dziribert: a pre-trained language model for the algerian dialect. *arXiv preprint arXiv:2109.12346*.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mohamed Charafeddine Bousri, Riad Bensalem, Samah Bessa, Zineb Lamri, Chahnez Zakaria, and Nabila Bousbia. 2022. Rumor detection in algerian arabizi based on deep learning and associations. In *International Symposium on Modelling and Implementation of Complex Systems*, pages 165–176. Springer.

Amina Chouigui, Oussama Ben Khiroun, and Bilel Elayeb. 2017. Ant corpus: An arabic news text collection for textual classification. In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pages 135–142.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Claude Coulombe. 2018a. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*.

Claude Coulombe. 2018b. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. 2019. A kernel theory of modern data augmentation. In *International conference on machine learning*, pages 1528–1537. PMLR.

Jiangshu Du, Yingtong Dou, Congying Xia, Limeng Cui, Jing Ma, and S Yu Philip. 2021. Cross-lingual covid-19 fake news detection. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 859–862. IEEE.

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. Arabart: a pretrained arabic sequence-to-sequence model for abstractive summarization. *arXiv preprint arXiv:2203.10945*.

Alexander R Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Pedro Henrique Arruda Faustini and Thiago Ferreira Covoes. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications*, 158:113503.

Khaled M Fouad, Sahar F Sabbeh, and Walaa Medhat. 2022. Arabic fake news detection using deep learning. *Computers, Materials & Continua*, 71(2).

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv preprint arXiv:1807.01554*.

Myunghoon Kang, Jaehyung Seo, Chanjun Park, and Heuiseok Lim. 2022. Utilization strategy of user engagements in korean fake news detection. *IEEE Access*, 10:79516–79525.

Nassima Kerras et al. 2019. Standard arabic and algerian languages: A sociolinguistic approach and a grammatical analysis. *Íkala*, 24(3):521.

Jude Khouja. 2020. Stance prediction and claim verification: An Arabic perspective. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, Seattle, USA. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Thomas Kober, Julie Weeds, Lorenzo Bertolini, and David Weir. 2020. Data augmentation for hypernymy detection. *arXiv preprint arXiv:2005.01854*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Gleb Kuzmin, Daniil Larionov, Dina Pisarevskaya, and Ivan Smirnov. 2020. Fake news detection for the russian language. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 45–57.

Sisi Liu, Kyungmi Lee, and Ickjai Lee. 2020. Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. *Knowledge-Based Systems*, 197:105918.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764*.

Samuel Louvan and Bernardo Magnini. 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. *arXiv preprint arXiv:2009.03695*.

David Lowell, Brian E Howard, Zachary C Lipton, and Byron C Wallace. 2020. Unsupervised data augmentation with naive augmentation and without unlabeled data. *arXiv preprint arXiv:2010.11966*.

Kevin Martínez-Gallego, Andrés M Álvarez-Ortiz, and Julián D Arias-Londoño. 2021. Fake news detection in spanish using deep learning techniques. *arXiv preprint arXiv:2110.06461*.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*.

Sebastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Rojas-Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced rdf verbalization with transformers. *arXiv preprint arXiv:2012.00571*.

Sosuke Nishikawa, Ryokan Ri, and Yoshimasa Tsuruoka. 2020. Data augmentation for learning bilingual word embeddings with unsupervised machine translation. *arXiv preprint arXiv:2006.00262*.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2018. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803.

Baolin Peng, Chenguang Zhu, Michael Zeng, and Jianfeng Gao. 2020. Data augmentation for spoken language understanding via pretrained language models. *arXiv preprint arXiv:2004.13952*.

Guillaume Raille, Sandra Djambazovska, and Claudiu Musat. 2020. Fast cross-domain data augmentation through neural sentence editing. *arXiv preprint arXiv:2003.10254*.

Chetanya Rastogi, Nikka Mofid, and Fang-I Hsiao. 2020. Can we achieve more with less? exploring data augmentation for toxic comment classification. *arXiv preprint arXiv:2007.00875*.

Mehdi Regina, Maxime Meyer, and Sébastien Goutal. 2020. Text data augmentation: Towards better detection of spear-phishing emails. *arXiv preprint arXiv:2007.02033*.

Mohammed El Manar Righi, Djallel Eddine Boussahel, Djamila Mohdeb, Meriem Laifa, and Messaoud Bendiaf. 2022. Rumor stance classification: A case study on the propagation of political rumors on the algerian online social space. In *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, pages 1–6. IEEE.

Houda Saadane and Nizar Habash. 2015. A conventional orthography for algerian arabic. In *the Second Workshop on Arabic Natural Language Processing*, pages 69–79.

Muhammad Haroon Shakeel, Asim Karim, and Imdadullah Khan. 2020. A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. *Information processing & management*, 57(3):102204.

Xiaohui Song, Liangjun Zang, and Songlin Hu. 2021. Data augmentation for copy-mechanism in dialogue state tracking. In *International Conference on Computational Science*, pages 736–749. Springer.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part IV 19*, pages 84–95. Springer.

Ge Yan, Yu Li, Shu Zhang, and Zhenyu Chen. 2019. Data augmentation for deep learning of judgment documents. In *Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9*, pages 232–242. Springer.

Kang Min Yoo, Youhyun Shin, and Sang-goo Lee. 2019. Data augmentation for spoken language understanding via joint variational generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7402–7409.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.