Orthographic Robustness of Persian Named Entity Recognition Models

Henry Gagnier
Pittsford Sutherland High School
Pittsford, New York, USA
henrygagnier9@gmail.com

Sophie Gagnier
Pittsford Sutherland High School
Pittsford, New York, USA
sophieag58@gmail.com

Abstract

Named Entity Recognition (NER) models trained on clean text often fail on real-world data containing orthographic noise. Work on NER for Persian is emerging, but it has not yet explored the orthographic robustness of models to perturbations often exhibited in user-generated content. We evaluate ParsBERT, ParsBERT v2.0, BertNER, and two XLM-r-based models on a subset of Persian-NER-Dataset-500k after applying eleven different perturbations, including simulated typos, code-switching, and segmentation errors. All models were competitive with each other, but XLM-r-large consistently displayed the best robustness to perturbations. Code-switching, typos, similar character swaps, segmentation errors, and noisy text all decreased F1 scores, while Latinized numbers increased F1 scores in ParsBERT. Removing diacritics, zero-width non-joiners, and normalizing Yeh/Kaf all did not have an effect on F1. These findings suggest that Persian NER models require improvement for performance on noisy text, and that the Perso-Arabic script introduces unique factors into NER not present in many high-resource languages, such as code-switching and Eastern Arabic numerals. This work creates a foundation for the development of robust Persian NER models and highlights the necessity of evaluating low-resource NER models under challenging and realistic conditions.

1 Introduction

Named Entity Recognition (NER) is a natural language processing task where important objects, such as a person, location, or organization, are identified from text (Roy, 2021). Low-resource languages often lack tools to be used with existing NER systems, and training data for low-resource systems can be scarce (Murthy et al., 2018; Liu et al., 2021). Persian, also known as Farsi, is a low-resource language primarily used in Iran, Tajikistan, and Afghanistan. Written using a modified version of the Arabic script, Persian presents unique challenges for NER due to its complex orthography, with ZWNJ characters and Yeh/Kaf characters specific to Abjad-based scripts.

Work on evaluating the orthographic robustness of NER models has recently mainly focused on high-resource languages. Srinivasan and Vajjala (2023) evaluate German, Hindi, and English NER models on samples with changed entities and context. Bodapati et al. (2019) evaluates English, Spanish, Dutch, and German NER models on robustness to capitalization errors. Namysl et al. (2020) brings to attention that sequence labeling systems such as NER should work reliably with corrupted inputs, as systems often process user-generated content or error-prone upstream components, and that NER systems are often trained on clean text, making them prone to error in real-world scenarios. They evaluated models on OCR errors and misspellings and found that these errors often significantly decreased model accuracy. This work is important for NER models applied to scenarios with user-generated content and variations, but many robustness findings from Latin-script languages do not transfer to Abjad-based scripts due to segmentation, joining behavior, and numeral variation.

Recently, Persian NER research has been emerging, and various systems have been developed for Persian NER. Mohseni and Tebbifakhr (2019) developed MorphoBERT, an NER model based on morphological analysis. Farahani et al. (2021) created ParsBERT, a monolingual BERT for the Persian language. Datasets for Persian NER have also been recently growing and spanning more domains. Asgari-Bidhendi et al. (2021) constructed ParsNER-Social, a corpus for NER in Persian social media. Shahshahani et al. (2018) create a standardized Persian NER dataset using Persian news websites. While work exists on Persian NER mod-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 110–114

els and corpora for Persian NER, it does not focus on text with real-world orthographic noise.

Work on orthographic robustness for Persian has not yet been performed, but it is necessary. This work makes three contributions: (1) we present the first evaluation of orthographic robustness for Persian NER, (2) we define eleven script-aware perturbations reflecting Abjad-specific noise, and (3) we identify segmentation errors and code-switching as dominant failure modes for current Persian NER models. The purpose of this paper is to (1) construct a benchmark to evaluate Persian NER models on orthographic robustness and (2) analyze how different perturbations affect current model performance. We also aim to improve the inclusion of Abjad-based scripts in NLP.

2 Materials and Methods

2.1 Data

We use the Persian-NER-Dataset-500k (Hamidzadeh, 2024), a comprehensive Persian NER dataset with approximately 500,000 tokens. We use a stratified sample of 5,578 samples to evaluate models in all experiments, ensuring label balance while keeping the computational costs of running three models on a large number of dataset variants manageable. F1 scores on this dataset range from 0.056 to 0.286 without fine-tuning and 0.355-0.551 with fine-tuning (Ghasemi and Salimi Sartakhti, 2025).

2.2 Orthographic Variants

To simulate real-world orthographic noise, we created eleven variants of the original text by applying the following perturbations.

- **No Diacritics:** All Arabic diacritics are removed from tokens to simulate omission in informal text.
- **Normalized Yeh/Kaf:** The Persian characters ي and ك are normalized to their standard forms ی and ک.
- **No ZWNJ:** The zero-width non-joiner (ZWNJ) character is removed to simulate segmentation errors.
- **Typos:** Random character substitutions are introduced at 10% and 20% token-level probabilities based on common confusion pairs in Persian orthography to simulate typos.
- **Similar Character Swaps:** Confusable characters such as ث/س and ذ/ز are swapped.
- **Segmentation Errors:** 10% of consecutive tokens are merged to mimic tokenization mistakes.
- **Code-Switching Names:** Tokens are replaced or mixed with Latin script or foreign names to simulate multilingual text.
- **Noisy Text Variants:** Two composite perturbations combining the above transformations for a more challenging test scenario. Noisy Text 1 was a composite of typos (15%), removal of diacritics, and Yeh/Kaf normalization. Noisy Text 2 was a composite of the removal of diacritics, similar characters swapped, and typos (20%).

2.3 NER Models

Three Persian NER and two multilingual models were evaluated on all orthographic variants: `bert-base-parsbert-ner-uncased` (ParsBERT), `bert-fa-base-uncased-ner-peyma` (ParsBERT v2.0) (Farahani et al., 2021), `bert-fa-zwnj-base-ner` (BertNER), `xlm-roberta-base-wikiann-ner` (XLM-RoBERTa-base), and `xlm-roberta-large-finetuned-conll03-english` (XLM-RoBERTa-large) (Conneau et al., 2020). We evaluated all models without finetuning to the dataset to isolate robustness effects from task-specific adaptation. All experiments were conducted on an NVIDIA Tesla T4 GPU (16GB VRAM).

3 Results

We look at the model performance across all perturbations to see how each perturbation affects NER (Table 1). We also assess the reliability of performance differences using paired two-sided t-tests on per-example F1 scores and use non-parametric confidence intervals with 1,000 resamples and a 95% confidence interval (Table 2). Overall model performance on the original text was similar in monolingual models, with F1 scores ranging from 0.400 to 0.440 throughout all three models. XLM-R-large achieved the highest F1 score of 0.463. These scores are typical of this Persian NER dataset (Ghasemi and Salimi Sartakhti, 2025). Removing diacritics, removing ZWNJ characters, and normalizing Yeh/Kaf all had minimal effects on all mod-

els, displaying the models' robustness to these perturbations. With Latinized numbers, performance was relatively unchanged in most models, but increased by 0.019 in ParsBERT.

When names of people, geopolitical entities, and locations were Romanized (code-switching names), a significant performance drop was exhibited in all models. Models each dropped from 0.047-0.108 points, with XLM-R-large being the most resistant. In the two noisy text samples combining typos with other perturbations, performance decreased significantly, from 0.041 to 0.058 in Noisy Text 1, and 0.118 to 0.188 in Noisy Text 2. In both composite perturbations, the multilingual models were the most robust.

When segmentation errors were simulated, performance decreased significantly to the lowest level of all the perturbations of 0.205 to 0.259, despite only 20% of tokens being affected. Similar characters visually were swapped, imitating OCR errors, resulting in a decrease in F1 scores to 0.286-0.391. Applying typos to 10% and 20% of the tokens, F1 scores decreased but to a lesser extent than other perturbations. Decreases of 0.026 to 0.044 were observed with 10% typos, and decreases of 0.053 to 0.082 were observed with 20% typos, with XLM-R-large having the best performance in both cases.

4 Discussion

We evaluate five NER models on eleven orthographic perturbations, revealing significant drops in accuracy when processing text with certain perturbations. Segmentation errors created the largest performance degradation across perturbations, decreasing model F1 scores to 0.205 to 0.259. This F1 is approximately 45% less than the models' original scores, suggesting that current Persian NER models rely greatly on accurate segmentation. Models were also affected greatly by code-switching, with drops of 0.047 to 0.108 points exhibited. This could affect NER models used in social media or other digital domains where names, brands, and terms may be romanized or in Latin script. Models all showed robust performance when diacritics and ZWNJ characters were removed, and Yeh/Kaf normalization was applied. XLM-R-large, a multilingual NER model, was generally the most robust model across perturbations, having better performance in difficult classification scenarios than the other models.

Surprisingly, F1 increased in ParsBERT by 0.019 when numbers were Latinized. We hypothesize that this occurs because the model was exposed to Latinized numbers during training, or that Eastern Arabic numerals introduce errors in Persian segmentation. The minimal impact of the removal of diacritics, ZWNJ characters, and normalizing Yeh/Kaf was surprising. This robustness may indicate that the models were trained using data with these qualities.

Future work should develop strategies to incorporate perturbed data in training data for Persian NER models to increase model robustness, or explore fine-tuning specifically for orthographic perturbations and real-world data. This benchmark should also be expanded to include additional error types or real-world data rather than synthetic data. Using real-world user-generated data with typos and real-world OCR data with errors would allow the NER model's orthographic robustness to be evaluated for scenarios where it may be used. This benchmark could also be expanded to investigate dialect variations. New models for NER should be evaluated, such as large language models with Persian language support, to see if more training data improves orthographic robustness, and multilingual NER models should be used due to XLM-R-large's strong performance (Litake et al., 2023). Persian error correction should continue to be a focus of research, as errors greatly impact NER, and correction would mitigate errors and greatly improve model accuracy.

These findings have great implications for Persian NER models in production environments with error-prone text. Current NER models, such as ParsBERT and BertNER, work for processing clean and well-formatted text, but social media data, data with code-switching, and other user-generated data likely require correction in NER models. We recommend that researchers design NER models for orthographic robustness and evaluate models on noisy test sets as well as standard benchmarks.

5 Conclusion

This study presents the first evaluation of Persian NER models for orthographic robustness. We create eleven perturbation types to reflect real-world noise and evaluate ParsBERT, ParsBERT v2.0, BertNER, XLM-R-base, and XLM-R-large, five publicly available NER models, on our benchmark.

Variant	ParsBERT	ParsBERT v2.0	BertNER	XLM-R-base	XLM-R-large
Original Text	0.440	0.436	0.400	0.365	**0.463**
No Diacritics	0.440	0.436	0.400	0.365	**0.463**
Normalized Yeh/Kaf	0.440	0.437	0.400	0.366	**0.463**
No ZWNJ	0.440	0.436	0.394	0.366	**0.464**
Latinized Numbers	0.459	0.436	0.399	0.365	**0.462**
Typos (10%)	0.405	0.406	0.356	0.339	**0.433**
Typos (20%)	0.369	0.378	0.318	0.312	**0.407**
Similar Character Swaps	0.302	0.318	0.287	0.286	**0.391**
Segmentation Errors	0.243	0.243	0.216	0.205	**0.259**
Code-Switching Names	0.366	0.376	0.322	0.257	**0.417**
Noisy Text 1	0.382	0.391	0.342	0.324	**0.422**
Noisy Text 2	0.252	0.270	0.230	0.242	**0.345**

Table 1: F1 scores of Persian NER models under orthographic perturbations.

	ParsBERT		ParsBERT v2.0		BertNER		XLM-R-base		XLM-R-large	
Perturbation	**ΔF1**	**Sig.**	**ΔF1**	**Sig.**	**ΔF1**	**Sig.**	**ΔF1**	**Sig.**	**ΔF1**	**Sig.**
No Diacritics	0.000	ns	0.000	ns	0.000	ns	0.000	ns	0.000	ns
Normalized Yeh/Kaf	0.000	ns	+0.001	ns	0.000	ns	+0.001	ns	0.000	ns
No ZWNJ	0.000	ns	0.000	ns	-0.006	ns	+0.001	ns	+0.001	ns
Latinized Numbers	+0.019	***	0.000	ns	-0.001	ns	0.000	ns	-0.001	ns
Typos (10%)	-0.035	***	-0.030	***	-0.044	***	-0.026	***	-0.030	***
Typos (20%)	-0.071	***	-0.058	***	-0.082	***	-0.053	***	-0.056	***
Similar Char Swaps	-0.138	***	-0.118	***	-0.113	***	-0.079	***	-0.072	***
Segmentation Errors	-0.197	***	-0.193	***	-0.184	***	-0.160	***	-0.204	***
Code-Switching Names	-0.074	***	-0.060	***	-0.078	***	-0.108	***	-0.047	***
Noisy Text 1	-0.058	***	-0.045	***	-0.058	***	-0.041	***	-0.041	***
Noisy Text 2	-0.188	***	-0.166	***	-0.170	***	-0.123	***	-0.118	***

Table 2: Statistical significance of changes in F1 from orthographic perturbations. ns = not significant ($p > 0.05$); *** = $p < 0.001$.

We find that XLM-R-large, a multilingual model, has the best performance and orthographic robustness of the models tested, but all models are competitive. While competitive with each other, all models require improvement for NER on noisy text. Simulated typos, code-switching, segmentation errors, character swaps, and noisy text all created large decreases in model accuracy, but models were robust to Latinized numbers, no diacritics or ZWNJ, and normalized Yeh/Kaf.

This research creates a foundation for more robust Persian NER models and shows the need for evaluating low-resource language NER models on noisy data, simulating user-generated data. These findings contribute to advancing low-resource and Persian NLP research, creating reliable models in real-world scenarios, and advancing the inclusion of Abjad-based scripts in NLP.

Limitations

There are multiple limitations that should be considered in this study. First, this study only focuses on five BERT-based Persian monolingual NER models. Other model architectures may perform better for Persian NER using perturbed data. Second, the perturbations that we used may not completely reflect user-generated data, and real-world errors stemming from processes typically upstream from NER. Third, the dataset used may not completely reflect scenarios where NER is used, such as in legal and scientific documents and in scenarios where local dialects are used. Finally, findings from Persian may not generalize to other similar languages using similar scripts.

References

M. Asgari-Bidhendi, B. Janfada, O. R. Roshani Talab, and B. Minaei-Bidgoli. 2021. Parsner-social: A corpus for named entity recognition in persian social media texts. *Journal of AI and Data Mining*, 9(2):181–192.

Sravan Bodapati, Hyokun Yun, and Yaser Al-Onaizan. 2019. Robustness to capitalization errors in named entity recognition. *Preprint*, arXiv:1911.05241.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Ali Reza Ghasemi and Javad Salimi Sartakhti. 2025. Multilingual language models in persian nlp tasks: A performance comparison of fine-tuning techniques. *Journal of AI and Data Mining*, 13(1).

Mansoor Hamidzadeh. 2024. Persian-ner-dataset-500k.

Onkar Litake, Maithili Sabane, Parth Patil, Aparna Ranade, and Raviraj Joshi. 2023. Mono versus multilingual bert: A case study in hindi and marathi named entity recognition. In *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, pages 607–618, Singapore. Springer Nature Singapore.

Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Ner-bert: A pre-trained model for low-resource entity tagging. *Preprint*, arXiv:2112.00405.

Mahdi Mohseni and Amirhossein Tebbifakhr. 2019. MorphoBERT: a Persian NER system with BERT and morphological analysis. In *Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019 - Short Papers*, pages 23–30, Trento, Italy. Association for Computational Linguistics.

Rudra Murthy, Mitesh M. Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(2).

Marcin Namysl, Sven Behnke, and Joachim Köhler. 2020. NAT: Noise-aware training for robust neural sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1501–1517, Online. Association for Computational Linguistics.

Arya Roy. 2021. Recent trends in named entity recognition (ner). *Preprint*, arXiv:2101.11420.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2018. Peyma: A tagged corpus for persian named entities. *Preprint*, arXiv:1801.09936.

Akshay Srinivasan and Sowmya Vajjala. 2023. A multilingual evaluation of ner robustness to adversarial inputs. *Preprint*, arXiv:2305.18933.