HarfoSokhan: A Comprehensive Parallel Dataset for Transitions between Persian Colloquial and Formal Variations

Hamid Jahad Sarvestani[1†] **Vida Ramezanian**[1†] **Saee Saadat**[2†] **Neda Taghizadeh Serajeh**[2†]
Maryam S. Razavi Taheri[1] Shohreh Kasaei[1] Mohammad A. Fazli[1] Ehsaneddin Asgari[3]

[1]Department of Computer Engineering, Sharif University of Technology
[2]Institute of Computer Science, University of Bonn
[3]Qatar Computing Research Institute, HBKU, Doha, Qatar

easgari@hbku.edu.qa

Abstract

A wide array of NLP/NLU models have been developed for the Persian language and have shown promising results. However, the performance of such models drops significantly when applied to the colloquial form of Persian. This challenge arises from the substantial differences between colloquial and formal Persian and the lack of parallel data facilitating the robustness of the model to the colloquial data or to transform the data to formal Persian. In addressing this gap, our research is dedicated to the development of the HarfoSokhan dataset, a large-scale colloquial to formal Persian parallel dataset of 6M sentence pairs. Our proposed dataset is a critical resource for training models that can effectively bridge the linguistic variations between colloquial and formal Persian. To illustrate the utility of our dataset, we used it to train a GPT2 model, which exhibited remarkable proficiency in colloquial to formal text style transfer, outperforming both OpenAI's GPT-3.5-turbo model and a leading rule-based system in this task. This conclusion is supported by our proposed ranking-based human evaluation. The results underscore the significance of the HarfoSokhan dataset in enhancing the performance of natural language processing models in the challenging task of colloquial to formal Persian conversion. Resources are available at huggingface. [1] [2]

1 Introduction

Persian is one of the most spoken languages of the world (ranked 24[3]) and it is among the top-ten languages contributing to the internet content [4] (Internet Society Foundation, 2023). The Persian language has two main forms: formal (standard or written style) and colloquial (informal or conversational form). These two forms differ in vocabulary, grammar, and style. The colloquial Persian is spoken in everyday conversations and reflects the dynamic form of the language that adapts to social and cultural adjustments (Tajalli et al., 2023). The high daily internet and social media usage has led to increased colloquial textual content generated, which are not adhering to standard grammars and structures. The coexistence of these two forms of Persian language in written form proposes a challenge in computational linguistics. Understanding these styles requires linguistic rules and vocabulary in both formal and informal aspects. To address this, we introduce HarfoSokhan, a large-scale parallel dataset of 6M sentence pairs. This dataset is constructed using a combination of a core, high-quality 12K sentence-pair corpus created and verified by native speakers, and a larger, machine-generated corpus created via back-translation

2 Related Work

The task of converting Persian colloquial text to its formal equivalent is an instance of Formality Style Transfer (FST) (Wang et al., 2020), a well-established sub-task within the broader field of Text Style Transfer (TST) (Jin et al., 2022). The fundamental goal of TST is to modify specific stylistic attributes of a text, such as its formality, sentiment, or politeness, while preserving its core semantic content (Jin et al., 2022). A central challenge that has shaped the trajectory of TST research is the pervasive scarcity of large-scale, high-quality parallel corpora, which consist of sentence pairs that share the same meaning but differ in style (Yin et al., 2019). This section first reviews the dominant architectural paradigms in the broader TST field, then discusses the specific challenges and prior work in the Persian context.

[1]https://huggingface.co/datasets/llm-lab/HarfoSokhan
[2]This work is licensed under Creative Commons NonCommercial license.
[3]https://www.ethnologue.com/
[4]Persian comes in 10^{th} place with 1.8%

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7380–7392

2.1 Paradigms in Text Style Transfer

Text Style Transfer aims to modify stylistic attributes of text while preserving its semantic content. The primary challenge is the style-content tradeoff, where separating "what" is said from "how" it is said is often intractable (Jin et al., 2022). The availability of parallel data largely shapes methodologies in TST.

Unsupervised and Disentanglement-Based Methods: In the common scenario where parallel data is unavailable, a dominant paradigm has been unsupervised learning based on disentangled representations. Using architectures like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), these methods attempt to learn a latent space where style and content are captured in separate, independent vectors. Style transfer is then achieved by swapping the style vector and decoding from the original content vector (Shen et al., 2017).

Back-Translation for Style Removal: An alternative paradigm, particularly relevant to our work, reframes TST by leveraging Neural Machine Translation (NMT) for style obfuscation. (Prabhumoye et al., 2018) pioneered the use of back-translation (translating a sentence to a pivot language and back) as an effective method to "wash out" stylistic features while preserving core semantics. The hidden state from the back-translation model's encoder is then treated as a style-agnostic content representation, which can be fed to a style-specific decoder. Our work adopts this core concept, applying it at a large scale to the specific problem of formal-to-colloquial Persian style transfer, using English as the pivot language.

Supervised Sequence-to-Sequence Models: When a parallel corpus is available (containing sentence pairs with the same content but different styles), TST can be framed as a supervised machine translation task. Sequence-to-sequence (seq2seq) models, typically based on the Transformer architecture, are trained to directly "translate" from a source style to a target style (Rao and Tetreault, 2018). The primary contribution of our work is the creation of a massive-scale parallel corpus that enables the effective training of such models for Persian.

2.2 Formality and Style Transfer in Persian

Applying TST to Persian is uniquely challenging due to the profound linguistic gap between its formal written register and its informal, colloquial form ("Shekaste-nevisi"). This is not merely a lexical difference but involves deep syntactic, morphological, and phonological changes (Tajalli et al., 2023). This complexity, coupled with Persian's status as a low-resource language, has historically constrained progress.

Despite the availability of numerous colloquial language resources in Persian, there is a shortage of parallel datasets of formal and colloquial data. Previous research has primarily focused on developing translation models from colloquial to formal dialect, often targeting only a subset of the colloquial dialect rules without the support of a parallel corpus. Alternatively, some studies have solely presented datasets consisting of either colloquial or formal phrases (MUT Deep Group, 2022; Kabiri et al., 2022; Khojasteh et al., 2020). In general, approaches can be divided into (i) rule-based systems and (ii) machine learning approaches.

Naeimi (Naemi et al., 2021) presents a rule-based approach that involves converting colloquial words into their formal counterparts. The approach involves extracting rules that convert colloquial words into formal ones by analyzing the changing processes, mostly simplification for the conversational form, in Persian words. Khojasteh (Khojasteh et al., 2020) proposes a large-scale colloquial Persian (LSCP) dataset that involves 120M colloquial sentences from Persian Twitter that translate to five different languages. These sentences were gathered from 27M tweets annotated with parsing trees, part-of-speech tags, sentiment polarity, and translation. Kabiri et al. (2022) presents an Informal Persian Universal Dependency Treebank, showing differences between informal and formal Persian language from phonological, morphological, and syntactic views. In fact, while Kabiri et al. (2022) and Khojasteh et al. (2020) focused on analyzing the colloquial style of the Persian language, they did not provide parallel corpora contrasting formal and conversational Persian. Additionally, Tajalli et al. (2023) developed an informal-formal Persian parallel dataset of 50,000 sentences, which is not publicly available.

3 Persian Linguistics Background

The Persian language is primarily spoken in Iran, Afghanistan, Tajikistan, and other countries in the Middle East and Central Asia, with more than 100 million native speakers (University of Texas

at Austin, Department of Middle Eastern Studies, 2024). Formal Persian is characterized by rigid grammar rules, formal vocabulary, and a more structured literary style, typically following the "subject-object-verb" structure. It is commonly used in written documents, academic publications, legal texts, and formal speeches (Tajalli et al., 2023). In contrast, colloquial Persian is known for shorter sentence lengths, recurrent use of idioms, and less compliance with the "subject-object-verb" structure; many formal expressions and even the dictation of some phrases change into their colloquial form.

Studies show that changes between formal and colloquial language occur through deletion, substitution, insertion, and interchange (Martins and Silva, 2004). Below, we explain each process with examples:

- Deletion: deletion of one or more letters, an example is "کتاب ها": *ketāb-hā* (books), which changes into "کتابا": *ketābā*.
- Substitution: Change of one letter to another. For example, "باران": *bārān* (rain) becomes "بارون": *bārūn*.
- Insertion: addition of one or more letters. As an example, "آن": *ān* (that) changes into "اون": *ūn*.
- Interchange: the reverse ordering of two adjacent letters of a word. For example, "قفل": *qofl* (lock) becomes "قلف": *qolf*.

Understanding the complexities and differences is crucial to successfully developing an automatic system to convert colloquial Persian to formal Persian.

Formal and informal Persian differ primarily in usage, vocabulary, and structure. Formal Persian, used in official, literary, and academic contexts, follows strict grammar, employs complex vocabulary and emphasizes honorifics and politeness markers. In contrast, informal Persian, prevalent in casual conversations and social media, features simplified vocabulary, colloquial expressions, and relaxed pronunciation, often omitting formal structures and using direct language. Here is a comprehensive overview of these differences (Mazdeh, 2013):

- **Verb Conjugation in Colloquial Persian**: One of the most significant differences between formal and colloquial Persian is in verb conjugation. Colloquial Persian modifies three key verb suffixes.
  - The suffix of "ید" changes to "ین". For example, "شما دارید": *Shomā dārid* changes into "شما دارین": *Shomā dārin*, which means *You have*.
  - The suffix of "ند" is replaced by "ن". For example, "آنها می روند": *Ānhā mi-ravand* changes into "آنها می رن": *Ānhā mi-ran*, which means *They go*.
  - The suffix of "د" changes to "ه". For example, "او دارد": *Oo dārad* changes into "او داره": *Oo dāre*, which means *She/He has*.
- **Stem Changes in Colloquial Persian**: Many verbs have a shorter stem form in colloquial Persian, particularly affecting present stems. For example, the formal present stem of "رفتن" (to go) is "رو", while the colloquial present stem is "ر". Therefore, "من می روم" (I will go) becomes "من می رم" in colloquial Persian.
- **General Pronunciation Shift Rules**: Certain combinations of sounds undergo consistent changes in colloquial Persian, such as:
  - The combination "ان" changes to "ون" in most cases, especially in single syllables.
- **Irregular Pronunciation Differences**
  - Pronouns and certain other words have irregular pronunciation differences in colloquial Persian, for example, the third person singular pronoun "او" becomes "اون", and the third person plural pronoun "آنها" becomes "اونا". The plural marker suffix "ها" changes to "ا" unless the noun ends in the vowels "ه" or "ا". Additionally, the consonant "ه" at the beginning of the word "هم" (meaning "too" or "even") is not pronounced.
  - Prepositions and similar words often lose their 'ezafe' (a grammatical particle in Persian) and any intervening consonant when their last sound is a vowel in colloquial Persian. Here are two examples:
    * "روی میز": *ru-ye miz* (on the table) changes to "رو میز": *ru miz*.

* "برای شما": *barā-ye shomā* (for you) changes to "برا شما": *bra shomā*.

- **Differences in Vocabulary**: There are specific words unique to colloquial Persian or that undergo unpredictable pronunciation changes For example:
  - "سر" *sar* changes to "کله" *kalleh*, meaning "head".
  - "بزرگ" *bozorg* changes to "گنده" *gondeh*, meaning "big".
  - "باز" *bāz* changes to "وا" *vā*, meaning "open".

4 Building Parallel Corpus

This section describes the construction of a large-scale colloquial-to-formal Persian dataset, HarfoSokhan, consisting of both manually and automatically generated parallel data. The design of our parallel corpus is directly informed by the linguistic contrasts between formal and colloquial Persian, as outlined in Section 3. To capture the full spectrum of these variations, we selected data sources rich in conversational phenomena.

4.1 Manual Dataset

The manual dataset consists of 12K colloquial and formal Persian pairs. In the development process, we sampled 3K lines from the DegarBayan subtitle dataset (MUT Deep Group, 2022) randomly and crawled 9K entries from Persian content on e-commerce platforms and social media. To generate the corresponding formal sentences, Native speakers with formal education translated informal sentences to their formal equivalents, with each translation manually verified by a second reviewer.

4.2 Large-scale Machine Generated Dataset

To create the large-scale dataset, we used Persian movie subtitles from the Opensubtitle project (Tiedemann, 2012) and the DegarBayan dataset, collecting 6M colloquial Persian sentences.

While we acknowledge that some subtitles can be formal, they predominantly reflect spoken, conversational language, making them a rich source for the colloquial phenomena we aim to model.

Recognizing the time-consuming nature of manual data annotation, we opted for a back-translation method to generate the formal component of our large-scale dataset. We selected the Google Translate API for this task. The decision was validated through a dedicated quality assessment, the results of which are detailed in 6.5. In this assessment, we compared Google Translate's output against translations from two independent human annotators. Pairwise METEOR scores showed that the quality of Google Translate's output was comparable to that of human annotators. This gave us confidence in its ability to generate high-quality formal Persian for our large-scale corpus creation. The basic statistics of the HarfoSokhan dataset are shown in Table 1.

Table 1: Statistics of HarfoSokhan Dataset.

	Formal	**Colloquial**
# Sentences	5,995,304	5,995,304
# Tokens	67,224,318	62,815,506
# Unique tokens	191,058	533,773
Average length	48.37	47.18

4.3 Downstream Tasks Dataset

To perform downstream task evaluation, we sampled 150 entries from the Persian Sentiment Analysis dataset (Hooshvare Lab, 2021b), which consists of three classes. Similarly, for news classification, we sampled 150 entries from the Persian News Dataset (Hooshvare Lab, 2021a), which includes seven classes.

5 Models

We utilize both decoder and encoder-decoder architectures for the informal-formal conversion. (i) GPT2 (Radford et al., 2019) is a transformer-based decoder language model designed for various natural language processing tasks. In this work, we use a pretrained Farsi version of GPT2, named ParsGPT2 (Hooshvare Team, 2021), which retains the GPT2 architecture with 117M parameters. The training data source for ParsGPT2 has not been disclosed. (ii) "Text-to-Text Transfer Transformer" (T5) (Raffel et al., 2020) is an encoder-decoder model built on transformer architecture. In this work, we utilize the Farsi version of T5, named ParsT5 (Pouramini, 2021), with 275M parameters. ParsT5 is trained on the uncleaned Farsi portion of OSCAR (Abadji et al., 2022).

Our choice of these architectures was motivated by their ability to model the complex transformations outlined in Section 3. Unlike simple rule-based systems, the attention mechanisms in these transformer models can capture the non-local de-

pendencies required for challenging conversions. By using ParsGPT2 and ParsT5 as base models, we leverage their pre-training on diverse Persian corpora, which provides a foundational understanding of the language's lexical and grammatical nuances.

We conducted fine-tuning on both ParsT5 and ParsGPT2 models to generate formal correspondences for our dataset of colloquial Persian sentences. Each model was fine-tuned separately on the HarfoSokhan dataset, which consists of large-scale machine-generated and manually annotated corpora, and we named them T5-HarfoSokhan and GPT2-HarfoSokhan. Additionally, we fine-tuned the models specifically on the manually annotated part of the HarfoSokhan dataset, naming them T5-Manual and GPT2-Manual.

6 Evaluation

To demonstrate the effectiveness of the HarfoSokhan dataset, we implemented a comprehensive, three-pronged evaluation approach. We recognize that converting colloquial text to formal text is a nuanced task of style transfer, not just translation. Metrics like BLEU, which rely on n-gram overlap, can be misleading; a model might achieve a high BLEU score by making only minor, superficial changes, failing to capture true formality. Therefore, our evaluation rests on three pillars: (i) human evaluation, which we consider the gold standard for this subjective task; (ii) LLM-as-a-Judge, to assess contextual and semantic appropriateness in a scalable way; and (iii) BLEU score, included as a traditional baseline to provide a complete picture of the models' performance. This multi-faceted approach allows for a more holistic assessment of model quality. The evaluation approaches are discussed in more detail in this section.

6.1 Test Set Creation

To ensure a robust and realistic evaluation, we deliberately created a test set from sources external to our training data, preventing any potential data leakage. This approach tests the models' ability to generalize to new, unseen data distributions. Our test set includes 500 sentences from our manual dataset (held out from all training) and 500 colloquial sentences from the LSCP dataset (Khojasteh et al., 2020). The LSCP data is particularly valuable as it consists of informal Persian tweets, which often lack formal structure and contain spelling errors, providing a challenging real-world scenario to test model robustness.

In order to generate formal correspondences, we utilized six different models on our test dataset of 1,000 sentences: (i) GPT2-Manual, (ii) T5-Manual, (iii) T5-HarfoSokhan, (iv) Farisyar, (v) GPT-3.5-turbo and (vi) GPT2-HarfoSokhan.

We generated formal correspondences using the ParsGPT2 and ParsT5 models, both fine-tuned on the HarfoSokhan (GPT2-HarfoSokhan and T5-HarfoSokhan respectively) and manual dataset (GPT2-Manual and T5-Manual respectively). Moreover, we introduced additional formal sentences generated by OpenAI's GPT-3.5-turbo model (ChatGPT) (OpenAI, 2022) and the FarsiYar rule-based system (FarsiYar Text Mining Group, 2020). These additional sentences were evaluated alongside those from our dataset, enhancing the completeness of our evaluation framework by presenting a wider array of formal sentence structures.

6.2 Human Evaluation

We devised an approach that incorporates human judgment into our evaluation setup. To effectively compare the model's results, we developed a web application equipped with a scoring system that allows users to rank formal sentences based on their colloquial counterparts. This method enables human evaluators to consider both linguistic and contextual nuances, facilitating a more comprehensive evaluation.

The scoring system presented six different formal outputs for a single colloquial sentence, with these outputs randomly ordered and the sources concealed.

We asked ten native Persian speakers to evaluate the quality of 200 sentences using the scoring system mentioned above. Each of the 1,000 test sentences was reviewed twice by different evaluators. The evaluation score was based on the number of times the output of each model was ranked first, second, or third according to the participants' opinion.

Subsequently, we calculated the cumulative sum of ranks and determined the percentage of occurrences for each model being ranked in the first, second, and third positions and named it the top-rank frequency score. The overall results of the ranking system are shown in Table 2. The GPT2 model, trained on the entire HarfoSokhan dataset (GPT2-HarfoSokhan), outperformed OpenAI's GPT-3.5-turbo, a large language model (LLM), and Farsi-

Table 2: Model performance comparison based on overall scores. top@1, top@2, and top@3 present the top-rank frequency score for the top 1, 2, and 3 predictions for each model. Higher values indicate better performance.

Model	top@1	top@2	top@3
GPT2-Manual	8.3	25.6	45.4
T5-Manual	4.4	12.6	23
T5-HarfoSokhan	5.4	19.1	38.4
Farsiyar	18	43.5	65.4
GPT-3.5-turbo	20.5	37.5	53.9
GPT2-HarfoSokhan	**43.0**	**61.4**	**73.5**

Table 3: The BLEU score for each model based on its reference translations

Model	BLEU score
GPT2-Manual	0.185
T5-Manual	0.038
T5-HarfoSokhan	0.164
Farsiyar	0.697
GPT-3.5-turbo	0.243
GPT2-HarfoSokhan	0.338

Yar, a rule-based system, in colloquial to formal conversion.

Moreover, when fine-tuned on the entire dataset, both T5 and GPT2 models (T5-HarfoSokhan and GPT2-HarfoSokhan, respectively) demonstrated better results than those fine-tuned only on the manually annotated dataset. This highlights the effectiveness of the back-translating method for generating the HarfoSokhan dataset in providing a source for colloquial to formal data. In order to evaluate the strengths and weaknesses of our model, we divided the test dataset based on the length of the sentences into three subgroups: short (less than 50 characters), medium (between 50 and 150 characters), and long (more than 150 characters) sentences. The number of sentences in each group is 739, 661, and 607, respectively. The results are shown in Table 5. Since our models were fine-tuned on relatively short sentences, our leading model, GPT2-HarfoSokhan, ranked first in evaluating short and medium-length sentences. This means that, according to the participants, our model had better output for short and medium-length sentences.

6.3 Machine Translation Evaluation

We calculated the BLEU scores (Papineni et al., 2002) for each model, as detailed in Table 3. We used the highest-rated translations from two reviewers as reference points for evaluating sentence quality. Notably, our model achieved a second-place ranking in this evaluation.

Although Farsiyar and GPT2-HarfoSokhan placed first and second, respectively, in BLEU score, upon manual comparison, we observed that Farsiyar sentences are often less formal than GPT2-HarfoSokhan sentences, a distinction not captured by the BLEU metric. To investigate this, two native Persian speakers analyzed 100 randomly sampled outputs from both models. The analysis revealed that Farsiyar frequently applies superficial, rule-based substitutions while preserving informal structures. For instance:

Colloquial input: او داره میره (He is going).

Farsiyar output: او داره می رود (A partial formalization that retains the colloquial verb structure).

GPT2-HarfoSokhan output: او در حال رفتن است (A fully formal, restructured sentence).

This qualitative analysis aligns with the human evaluation rankings (Table 2), where GPT2-HarfoSokhan was preferred for its deeper, more contextually-aware formalizations.

We hypothesized that the problem lies in how BLEU measures performance. The exact matches in Farsiyar are more frequent than in GPT2-HarfoSokhan, as only a few informal words and expressions are converted to their formal correspondences in Farsiyar, leading to a higher BLEU score. To further evaluate our model performance against Farsiyar, we decided to add a third evaluation method.

6.4 LLM-as-a-Judge Evaluation

Unlike machine translation metrics such as BLEU, which evaluate sentences based on exact matching or word overlap, the strength of large language models (LLMs) lies in understanding context. To leverage this capability, we employ a reference-free metric approach referred to as LLM-as-a-Judge, as introduced in (Zheng et al., 2023). This approach uses the point-wise scoring structure for single answer grading. To mitigate position bias in LLM-as-a-Judge, we test multiple orderings of the same input, which is crucial for assessing the model's consistency and impartiality. Detailed results of this evaluation are presented in Table 4. In

	GPT2-HarfoSokhan						Farsiyar					
	GPT2-HarfoSokhan		Farsiyar		Neutral		GPT2-HarfoSokhan		Farsiyar		Neutral	
	Outperf.(%)	Avg. Score	Outperf.(%)	Avg. Score	Outperf.(%)	Avg. Score	Outperf.(%)	Avg. Score	Outperf.(%)	Avg. Score	Outperf. (%)	Avg. Score
Prompt 1	**69.18**	**6.09**	29.65	5.05	1.16	-	**77.32**	**6.38**	30.34	4.5	2.32	-
Prompt 2	**66.27**	**6.68**	30.23	6.1	3.4	-	**56.39**	**6.96**	41.27	6.66	2.3	-

Table 4: LLM-as-a-Judge evaluation report comparing GPT2-HarfoSokhan, and Farsiyar with different input orderings to examine position bias. i) Outperformance rate describes the percentage of times a model has outperformed another one. ii) Avg. Score is the average score assigned to each model by LLM-as-a-Judge based on its performance. Prompts are explained in detail in the Appendix E.

Table 5: Performance comparison of GPT2-HarfoSokhan on evaluation instances in different settings. top@1, top@2, and top@3 present the top-rank frequency score for the top 1, 2, and 3 predictions for each model. The results in **bold** indicate the best-performing metrics in each category.

Model	**Short**			**Medium**			**Long**		
	top@1	**top@2**	**top@3**	**top@1**	**top@2**	**top@3**	**top@1**	**top@2**	**top@3**
GPT2-Manual	10.15	30.85	47.90	7.413	24.357	45.537	7.084	20.593	42.339
T5-Manual	5.548	15.426	25.440	3.933	9.985	20.121	3.789	12.191	23.394
T5-HarfoSokhan	7.713	26.387	46.955	5.144	18.760	35.703	2.965	10.873	31.137
Farsiyar	13.802	36.401	61.434	17.852	43.116	64.902	23.558	52.883	71.005
GPT-3.5-turbo	12.720	22.598	38.024	16.944	36.762	56.581	**34.102**	**56.507**	**70.511**
GPT2-HarfoSokhan	**50.068**	**68.336**	**80.2436**	**48.714**	**67.020**	**77.156**	28.501	46.952	61.614

our evaluation using the LLM-as-a-Judge approach, GPT2-HarfoSokhan outperformed Farsi-Yar.

6.5 Annotation Quality Evaluation

To assess the quality of human annotations and the performance of Google Translate, we conducted a systematic evaluation. Two annotators independently translated a sample of informal sentences, and we included Google Translate outputs for comparison. Additionally, a third annotator provided a reference baseline to measure consistency across translations. We then computed pairwise METEOR (Banerjee and Lavie, 2005) scores, which account for word order and semantic equivalence, to quantify the similarity among all translations, as shown in Table 6. Variances across scores were consistently below 0.003, indicating strong agreement among annotators despite the task's complexity. On average, human annotations closely matched Google Translate quality relative to the reference dataset.

Table 6: Pairwise METEOR scores comparing human annotations and Google Translate (GT) outputs. "Ref" represents the reference annotations provided by the third annotator.

Comparison	**AVG**	**VAR**
Annt 1 vs. Annt 2	0.461	0.002
Annt 1 vs. GT	0.453	0.001
Annt 2 vs. GT	0.487	0.0001
Ref vs. GT	0.447	0.002
Ref vs. Annt 1	0.421	0.00008
Ref vs. Annt 2	0.455	0.003

6.6 Downstream Tasks Evaluation

To assess the effectiveness and practical utility of the HarfoSokhan dataset, we conducted experiments on two widely used downstream tasks: (i) news classification and (ii) sentiment analysis. These tasks serve two primary goals: (1) Quality Validation: To quantitatively measure if the formal text generated by GPT2-HarfoSokhan improves performance in standard NLP pipelines compared to the original colloquial text or text generated by other models like ChatGPT. (2) Practical Utility: To demonstrate that our model can be used as an effective pre-processing tool to formalize user-generated content, thereby enhancing the robustness of downstream applications like sentiment analysis or content classification. In each experiment, we compared a small sample of data produced by our GPT2-HarfoSokhan model against data generated by ChatGPT (GPT-4o). Our primary objective was to evaluate whether GPT2-HarfoSokhan could generate data of comparable or superior quality compared to a well-established baseline.

For evaluation, we utilized the datasets described in Section 4.3. The sentiment analysis task involved three classes, while the news classification task encompassed seven classes. The original datasets contained informal text, which was first

transformed into a more formal style using either GPT2-HarfoSokhan or ChatGPT. The transformed texts were then evaluated using ParsBERT (Farahani et al., 2021), a transformer-based model for Persian language understanding that had been separately fine-tuned for sentiment and news classification.

In the news classification task, models tested on GPT2-HarfoSokhan-generated data achieved an accuracy of 74.60%, surpassing the accuracy of models tested on ChatGPT-generated data (71.43%). Similarly, in the sentiment analysis task, models trained on GPT2-HarfoSokhan-generated data attained an accuracy of 85.09%, exceeding the 83.33% achieved with ChatGPT-generated data.

These results indicate that GPT2-HarfoSokhan is capable of generating high-quality data that matches or even surpasses the performance of data derived from ChatGPT for both news classification and sentiment analysis.

7 Conclusion

In this paper, we introduce HarfoSokhan, the first large-scale Persian colloquial to formal parallel dataset. The dataset contains 6 million pairs of colloquial and formal Persian sentences, which consist of manually annotated and large-scale machine-generated corpora. We employed the back-translation method to create the machine-generated corpus by translating the colloquial dataset into English and then retranslating it into formal Persian. We trained two models on the HarfoSokhan dataset and used a novel scoring system to evaluate their performance. To ensure accuracy, we enlisted the help of native Persian speakers to rank formal sentences generated by these models against those produced by other methods. Our evaluation results demonstrate the effectiveness of this dataset in bridging the gap between colloquial and formal Persian languages, highlighting its importance as a valuable resource in this domain. Our evaluation framework, which prioritizes human judgment as the primary measure of success, demonstrates the effectiveness of the HarfoSokhan dataset. While the large-scale dataset relies on machine generation, its quality was validated against human translators, and the final model outputs were primarily assessed by native Persian speakers. This human-centric approach ensures our findings are grounded in real-world language use and not merely an artifact of a closed generative loop. The HarfoSokhan dataset can be utilized in various downstream tasks within the NLP domain. It can help build models that classify text as colloquial or formal, which is beneficial for content moderation. In addition, it can enhance NLU systems by improving their ability to interpret colloquial language accurately and provide formal translations and responses.

8 Limitations

Despite the model's strong performance, a detailed error analysis reveals several challenges, many of which align with the linguistic categories identified in Section 3. Our leading model, GPT2-HarfoSokhan, occasionally struggles with specific transformations, highlighting areas for future improvement. Below we categorize the most common error types:

- **Verb Conjugation and Stems:** While often successful, the model sometimes fails to convert colloquial verb stems to their formal equivalents, or it retains informal suffixes, especially in complex sentences. As noted in Table 9, it might partially formalize a verb but keep the colloquial stem.
- **Irregular Pronunciation and Vocabulary:** The model can struggle with rare or highly context-dependent slang. Words with irregular colloquial forms, such as the pronoun اونا *unā* (them) for آنها *ānhā* (them), are sometimes only partially converted (e.g., to اونها *unhā* (them)), indicating an incomplete transformation.
- **Proper Nouns and Named Entities:** As noted in Table 9, the model sometimes incorrectly alters proper nouns (e.g., فرشچیان *Farshchiyān* to فرشکانی). This suggests a difficulty in distinguishing between a common word requiring transformation and a name that should be preserved.
- **Sentence Length Dependency:** As shown in our evaluation (Table 5), model performance degrades on longer sentences (>150 characters). The top@1 score for GPT2-HarfoSokhan drops from 50.0% on short sentences to 28.5% on long sentences, suggesting that maintaining context and applying consistent formalisms across longer text spans is a significant challenge.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. 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*, 54:365–383.

FarsiYar Text Mining Group. 2020. FarsiYar text mining tools. https://github.com/Text-Mining?language=go.

Hooshvare Lab. 2021a. Persian news dataset. https://hooshvare.github.io/docs/datasets/tc.

Hooshvare Lab. 2021b. Persian sentiment analysis dataset. https://hooshvare.github.io/docs/datasets/sa.

Hooshvare Team. 2021. ParsGPT-2, a Persian version of GPT-2. https://github.com/hooshvare/parsgpt.

Internet Society Foundation. 2023. What are the most used languages on the internet? https://www.isocfoundation.org/2023/05/what-are-the-most-used-languages-on-the-internet/ Accessed: 2024-09-02.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Roya Kabiri, Simin Karimi, and Mihai Surdeanu. 2022. Informal Persian universal dependency treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 7096–7105, Marseille, France. European Language Resources Association.

Hadi Abdi Khojasteh, Ebrahim Ansari, and Mahdi Bohlouli. 2020. LSCP: Enhanced large scale colloquial Persian language understanding. *arXiv preprint arXiv:2003.06499*.

Bruno Martins and Mário J. Silva. 2004. Spelling correction for search engine queries. In *Proceedings of the 4th International Conference on Advances in Natural Language Processing (EsTAL 2004)*, pages 372–383, Alicante, Spain. Springer.

Mohsen Mahdavi Mazdeh. 2013. Colloquial Persian lessons. https://persiandee.com/view/colloquial/currentLesson/0.

MUT Deep Group. 2022. Degarbayan-SC: A colloquial paraphrase Farsi subtitles dataset. https://github.com/mut-deep/Degarbayan-SC.

Amin Naemi, Marjan Mansourvar, Mostafa Naemi, Bahman Damirchilu, Ali Ebrahimi, and Uffe Kock Wiil. 2021. Informal-to-formal word conversion for Persian language using natural language processing techniques. In *Proceedings of the 2nd International Conference on Computing, Networks and Internet of Things*, pages 1–7. ACM.

OpenAI. 2022. ChatGPT. https://openai.com/blog/chatgpt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ahmad Pouramini. 2021. parsT5-base. https://huggingface.co/Ahmad/parsT5-base.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI Blog. Version 1.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6833–6844, Long Beach, CA, USA. Curran Associates Inc.

Vahide Tajalli, Fateme Kalantari, and Mehrnoush Shamsfard. 2023. Developing an informal-formal Persian corpus. *arXiv preprint arXiv:2308.05336*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

University of Texas at Austin, Department of Middle Eastern Studies. 2024. Persian language program. https://liberalarts.utexas.edu/languages/persian.html. Accessed: 2024-09-02.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wen-Han Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2236–2249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Di Yin, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2019. Utilizing non-parallel text for style transfer by making partial comparisons. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 5379–5386, Macao, China. International Joint Conferences on Artificial Intelligence Organization.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36.

A Data Collection Protocol

The annotators, aged between 22 and 40, were from Iran and had formal education. Participation in the project was voluntary. We discussed the terms of data usage, ensuring that the annotators were informed that the data would be made public. The data collection protocol was reviewed, and consent was obtained from the participants.

To generate the corresponding formal sentences, native speakers with formal education translated informal sentences into their formal equivalents. Each translation was manually verified by a second reviewer following a predefined review protocol. This protocol provided specific instructions: (1) Preserve the original meaning of the sentence without omission or addition. (2) Convert all colloquial verbs, pronouns, and suffixes to their standard written (formal) equivalents as outlined in Section 3. (3) Correct any spelling errors or informal abbreviations. (4) Ensure the final sentence is grammatically correct and fluent in formal Persian. This two-step process of translation and verification was designed to ensure high-quality and consistent annotations.

B Packages and Implementation

Hazm and NLTK were used for text preprocessing. NLTK, METEOR, and Spacy were used for evaluation, with default parameter settings unless otherwise specified.

C Training Setup

The ParsGPT2 model training process is optimized with the set of hyperparameters provided in table 7. Also, table 8 provides the set of hyperparameters for ParsT5 model training setup.

Table 7: ParsGPT2 hyperparameters.

Parameter	Value
Batch size	48
Number of attention heads	12
Embedding dimension	768
Number of decoder layers	12
Loss	Cross-entropy
Optimizer	AdamW
Learning rate	5e-4

Table 8: ParsT5 hyperparameters.

Parameter	Value
Batch size	4
Number of attention heads	12
Embedding dimension	768
Number of encoder layers	12
Number of decoder layers	12
Loss	Cross-entropy
Optimizer	AdamW
Learning rate	3e-4

D Scoring System

A ranked sample of the scoring system is shown in Figure 1. The figure illustrates six different formal outputs generated for a single colloquial sentence. The outputs are randomly ordered and the sources are hidden to ensure an unbiased evaluation.

E LLM-as-a-Judge Evaluation Setup

There are different types of LLM-as-a-Judge variations (Zheng et al., 2023). For free-reference evalu-

Table 9: Examples of Model Errors in Transforming Persian Colloquial to Formal Language

Colloquial Sentence	Model Output	Problem
فقط باید شاخه هاش و تنظیم کنید.	فقط باید شاخه ها را تنظیم و تنظیم کنید.	Word repetition
یک جعبه کامل برای کسانی که فنی هستند	یک جعبه جعبه برای کسانی که فنی هستند	Word repetition
فرشچیان نقاش خوبیه.	فرشکانی نقاش خوبی است.	Change special noun
فوتسال به بازی فوتبال درون سالن می گن.	فوتبال در سالن فوتبال را فوتبال می گویند.	Change special noun
محمد موسوی ورزشکار خوبیه	محمد رهنما ورزشکار خوبی است.	Change special noun
به تلخی میزد. راضی نیستم .	به سختی ضربه می زد. من راضی نیستم .	Context misunderstanding

Figure 1: Example of a ranked sample in the scoring system.

ation, we use single-answer grading. Two prompts are used for evaluation: First, we request a score for each formal sentence generated on a scale of 1 to 10, as shown in Figure 2. Second, we include the informal sentence along with two formal sentences in the evaluation process, as shown in Figure 3. Additionally, we provide examples of different input ordering. In the first example in Figure 4, sentences are provided by GPT2-HarfoSokhan and Farsiyar, respectively. In the second example in Figure 5, the order is reversed.

Prompt 1

You are tasked with evaluating the formality of two Persian sentences in comparison to a given informal sentence. Using your understanding of Persian language and formality conventions, score sentence from 1 (least formal) to 10 (most formal) based on how they maintain the context adjusting the formality level.

Consider the following two sentences:
First sentence: sentence #1
Second sentence: sentence #2
Evaluate the formality of each sentence and provide a score for each:
- For the first sentence, provide a score from 1 to 10.
- For the second sentence, provide a score from 1 to 10.

Your final output should be in the format: "X, Y" where X is the formality score of the first sentence and Y is the formality score of the second sentence.

Figure 2: Prompt for pointwise scoring of formal sentences on a scale of 1 to 10.

Prompt 2

You are tasked with evaluating the formality of two Persian sentences in comparison to a given informal sentence. Using your understanding of Persian language and formality conventions, score sentence from 1 (least formal) to 10 (most formal) based on how they maintain the context adjusting the formality level.
Here is the reference informal sentence: {informal}

Consider the following two sentences:
First sentence: sentence #1
Second sentence: sentence #2
Evaluate the formality of each sentence and provide a score for each:
- For the first sentence, provide a score from 1 to 10.
- For the second sentence, provide a score from 1 to 10.

Your final output should be in the format: "X, Y" where X is the formality score of the first sentence and Y is the formality score of the second sentence.

Figure 3: Prompt for pointwise scoring that includes an informal sentence alongside two formal sentences.

Example 1

You are tasked with evaluating the formality of two Persian sentences in comparison to a given informal sentence. Using your understanding of Persian language and formality conventions, score sentence from 1 (least formal) to 10 (most formal) based on how they maintain the context adjusting the formality level.
Here is the reference informal sentence:
ولی بازم فکر میکنم باید به قلب ادمها نگاه کرد اینکه چه مسیری رو پشت سر گذاشته

Consider the following two sentences:
First sentence: ولی باز هم فکر می کنم باید به قلب آدم ها نگاه کرد که چه مسیری را پشت سر گذاشته است
Second sentence: ولی باز هم فکر می کنم باید به قلب ادمها نگاه کرد اینکه چه مسیری را پشت سر گذاشته
Evaluate the formality of each sentence and provide a score for each:
- For the first sentence, provide a score from 1 to 10.
- For the second sentence, provide a score from 1 to 10.

Your final output should be in the format: "X, Y" where X is the formality score of the first sentence and Y is the formality score of the second sentence.

Output: 8, 5

Figure 4: Evaluation prompt with the order GPT2-HarfoSokhan then Farsiyar.

Example 2

You are tasked with evaluating the formality of two Persian sentences in comparison to a given informal sentence. Using your understanding of Persian language and formality conventions, score sentence from 1 (least formal) to 10 (most formal) based on how they maintain the context adjusting the formality level.
Here is the reference informal sentence:
ولی بازم فکر میکنم باید به قلب ادمها نگاه کرد اینکه چه مسیری رو پشت سر گذاشته

Consider the following two sentences:
First sentence: ولی باز هم فکر می کنم باید به قلب ادمها نگاه کرد اینکه چه مسیری را پشت سر گذاشته
Second sentence: ولی باز هم فکر می کنم باید به قلب آدم ها نگاه کرد که چه مسیری را پشت سر گذاشته است
Evaluate the formality of each sentence and provide a score for each:
- For the first sentence, provide a score from 1 to 10.
- For the second sentence, provide a score from 1 to 10.

Your final output should be in the format: "X, Y" where X is the formality score of the first sentence and Y is the formality score of the second sentence.

Output: 7, 9

Figure 5: Evaluation prompt with the order Farsiyar then GPT2-HarfoSokhan.