*Natural Language Engineering* (2024), **30**, pp. 1065–1090
doi:10.1017/S1351324923000426

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

# PGST: A Persian gender style transfer method

Reza Khanmohammadi and Seyed Abolghasem Mirroshandel

Computer Engineering Department, Faculty of Engineering, University of Guilan, Rasht, Iran
**Corresponding author:** Seyed Abolghasem Mirroshandel; Email: mirroshandel@guilan.ac.ir

(Received 30 March 2021; revised 26 June 2023; accepted 22 July 2023; first published online 15 August 2023)

**Abstract**

Recent developments in text style transfer have led this field to be more highlighted than ever. There are many challenges associated with transferring the style of input text such as fluency and content preservation that need to be addressed. In this research, we present PGST, a novel Persian text style transfer approach in the gender domain, composed of different constituent elements. Established on the significance of parts of speech tags, our method is the first that successfully transfers the gendered linguistic style of Persian text. We have proceeded with a pre-trained word embedding for token replacement purposes, a character-based token classifier for gender exchange purposes, and a beam search algorithm for extracting the most fluent combination. Since different approaches are introduced in our research, we determine a trade-off value for evaluating different models' success in faking our gender identification model with transferred text. Our research focuses primarily on Persian, but since there is no Persian baseline available, we applied our method to a highly studied gender-tagged English corpus and compared it to state-of-the-art English variants to demonstrate its applicability. Our final approach successfully defeated English and Persian gender identification models by 45.6% and 39.2%, respectively.

**Keywords:** Text classification; gender difference; text style transfer; word embedding; beam search

## 1. Introduction

Transfer of expressive style has been addressed less in natural language processing (NLP) than computer vision (Gatys, Ecker, and Bethge 2015; Gatys *et al.* 2016; Luan *et al.* 2017), primarily due to the absence of a reliable evaluation metric and a shortage of parallel corpora (Fu *et al.* 2018). But recent robust pre-trained language models for language generation and both manual and automatic evaluation metrics have facilitated research on text style transfer, resulting in its growing significance among NLP applications. Low-resource natural languages, however, have not yet witnessed this rising trend. Even with the scarcity of resources, Persian natural language processing has recently seen impressive advancements. However, to the best of our knowledge, no prior study has yet attempted to investigate text style transfer in Persian. Lack of corpora, ambiguous semantics, and exacting pragmatic are among the most substantial challenges of processing this natural language.

Besides six middle eastern countries that partly speak it, Persian/Farsi is the official language of Iran, Afghanistan and Tajikistan. Despite the Indo-European family of languages, Persian is one of the most important members of the Indo-Iranian branch. Persian has undergone significant changes as one of the most ancient languages. Due to the adjacency of Persian and Arabic speakers, a plethora of Arabic loaned words have been injected into Persian. Since Persian speakers suggest replacements for such words and that there is no specified boundary on which Arabic words are officially accepted in Persian, the ratio of out-of-vocabulary (OOV) words gets

unintentionally even higher compared to English. As Shamsfard (2019) suggests, the following is a list of challenges that need to be handled to process this language:

1. Generally, and most specifically at the lexical level, there is a vast gap between how the colloquial and formal Persian is spoken and written. Due to the change of grammar in colloquial, most resources can only process the formal literature. This issue resulted in using a formal corpus containing texts that are harder to distinguish based on gender as opposed to informal and colloquial texts.
2. To some extent, Persian has turned out to be free of word order. Meaning sentences are still expressive even if specific part-of-speech tags are relocated. Due to these characteristics, the language is difficult to process computationally, making the objective of Natural Language Understanding a challenging goal to attain.
3. There exist many scripts and types of writing a Persian letter.
4. Unlike English, German or French, Persian has no definite article. This makes gender style transfer more complex since there exists no gender-distinguishing information in the text.
5. Uncountable nouns are probable to appear in plural form.
6. Persian adjectives are likely to be used in place of nouns. This causes many semantic and structural ambiguities among noun phrases. This issue was the most challenging impediment to overcome for our beam search algorithm.

Despite previously delineating the majority of the critical complexities intrinsic to the Persian language, we further direct readers to the comprehensive study conducted by Shamsfard (2019). This work delves profoundly into the subject matter and provides additional insights.

To transfer the style of a text from one gender to another, first, we must understand their essential differences. Based on sociolinguistic studies, men and women have been shown to have distinct, deeply rooted variations in their language (Wallentin 2020). Such linguistic phenomena may have been caused due to non-identical social and psychological circumstances that men and women undergo throughout their lives (Jin-yu 2014). Such studies paved the way for us to address style transfer with far more excellent knowledge and base our approach on such differences, which emerge mostly in specific parts of speech tags. On the other hand, Li *et al.* (2018) simply overcame the task of attribute transfer by deleting identified attribute markers of text and replacing them with their equivalent retrieved target attributes. This work highly motivated us to distinguish gender-dependent words for gender style transfer purposes.

In this work, we introduce the first Persian instance of a text style transfer method called **PGST**, a **P**ersian **G**ender **S**tyle **T**ransfer method, which mainly revolves around transferring the style of a sentence by the gender of its author. The objective here is to change the linguistic style of an input text from its source style ($S_s$) to a target style ($S_t$). As an example, in our case, if the input to the transfer method is written by a male author ($S_s$), the method is expected to produce text that appears to have been written by a female author ($S_t$) while preserving content. In order to achieve this, we have developed a methodology that executes stylistic modifications at the granular level of individual words. This way, input (styled as $S_s$) is first split into words, then a set of similar $S_t$ styled words are suggested per each word, and finally, a fluent combination of these suggestions is extracted as the final output. More specifically, a pre-trained word embedding, a character-based token classifier and an exclusive beam search decoder are used in this research as building blocks for our final style transfer method. The pre-trained word embedding is used to represent words in an n-dimensional space, so any constituent word of the input can be linked to a list of its most similar words. The token classifier then distinguishes members of these lists by how male- or female-like they members are. Finally, the beam search algorithm is used to extract a favourable combination from $S_t$ styled list members.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

When transferring an input from its source style to its transferred style ($S_t$), preserving an input's content and fluency are among the two most critical challenges that a style transfer method faces. In our introduced method, given an input sample, content's resistance to change is being handled by suggesting replacements from each token's embedding space. The fluency aspect is under control by running a beam search algorithm on all suggestions, ranking suggestions continuously by our proposed scorer function. By following the preceding approach, we will have fluent transferred sentences with the same contextual information. Furthermore, style transfer has just recently reached out to text data, making it a laborious and very time-consuming task to evaluate. Hence, we evaluated our approach using automatic, statistical and human judgement-based scores to highlight its success, which will be discussed in great detail in the evaluation section of this paper. Lastly, to achieve acceptable performance, it is imperative to leverage language-specific knowledge for a low-resource and enigmatic natural language, such as Persian.

The contributions of this paper are as follows:

- In accordance with the disparities evident in gender-specific text, we have formulated a Persian text style transfer method that relies on linguistic evidence. Our pioneering approach, PGST, represents the inaugural instance of a text style transfer method being introduced to Persian, a natural language that is low in resources.
- We have established a benchmark for the Persian text style transfer research trajectory. This benchmark involves the comparison of various models that utilize either word or character embeddings.
- In order to showcase the potential applicability of our method, we extended its implementation to English text and contrasted it with the existing state-of-the-art methodologies. We conducted a comprehensive series of evaluations, including statistical, human and automated tests, the results of which are discussed in the subsequent results section.

The remainder of this paper is structured as follows. In Section 2, we take a look at some related work, previously done in the scope of gender difference, text classification and text style transfer. Section 3 is dedicated to giving an account of our proposed method. In Sections 4 and 5, we share our experiments and discuss our method's outcome and finally, we conclude our paper in Section 6.

## 2. Related work

Our proposed method's central concept is structured on gender differences in written text, a well-studied sub-field of sociolinguistics. Trudgill (1972) and Eckert (1989) carried out some preliminary work by focusing on male and female text's lexical and phonological differences. About a decade later, Bucholtz (2002) laid the groundwork for the term "Sex Differences" and how the study of gender variations evolves based on theories as a complex and context-specific system. In the following decades, the field witnessed an overabundant domain-specific growth of gender-driven language studies in education, social networks and science (Li and Kirkup 2007; Cavas 2010; Metin *et al.* 2011; Zare 2013; Nahavandi and Mukundan 2014; Serizel and Giuliani 2017). For instance, Kayaoğlu (2012) studied the language learning strategies of male and female students in five different categories (memory, compensation, cognitive, metacognitive and social strategy). In contrast, the most recent corpus-based research endeavours have meticulously examined gender disparities at the word level. They have called attention to the role of part-of-speech tags in gender language (Pearce 2008; Baker 2014; Norberg 2016; Hoyle *et al.* 2019), specifically by studying the aftermath of different adjective choices on nouns. We drew inspiration from these long-studied sociolinguistics-based perspectives and led word-level gender differences to underlie our proposed text style transfer method.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

Among text classification models introduced in Persian, the topic-model approach (Ahmadi, Tabandeh, and Gholampour 2016) overcame the problems of dealing with bag of words, which considered each token as a feature, thus dealing with a vast number of elements and features inside a document. Besides, (Moradi and Bahrani 2016) narrowed the task of text classification down to gender domain, where different statistical models such as Naïve Bayes, alternating decision tree and support vector machine were evaluated. Although a very small number of researches have successfully addressed author gender identification in Persian, a large number of previous studies have met this challenge in English under various terms (Cheng, Chandramouli, and Subbalakshmi 2011; Soler and Wanner 2016; Bsir and Zrigui 2018; Fatima *et al.* 2018; Martinc and Pollak 2018, Yildiz 2019; Sotelo *et al.* 2020; Sotelo *et al.* 2020). We believe our developed gender identification model, introduced as our automatic evaluator, takes a step towards further task advancements in Persian.

In terms of style transfer, English NLP has experienced rapid growth where several approaches are recently introduced using the latest architectures and algorithms. Having disentangled image features such as colour (Chen *et al.* 2016), Hu *et al.* (2017) focused on controlled text generation by learning disentangle latent representations and made a relation between an input's style and content. Such models are trained hardly in an adversarial manner, which generates poor-quality sentences as a result. Following Hu's disentangling latent representation method, John *et al.* (2019) recently proposed a simple yet efficient approach to approximate content information of bag-of-words features. Other researches, either directly or indirectly, have played several different parts in contributing to text style transfer. In particular, several models are introduced based on attention weights (Feng *et al.* 2018), neural machine translation (Subramanian *et al.* 2018) and deep reinforcement learning (Gong *et al.* 2019). Nevertheless, they came up short by misunderstanding input content, requiring an immense number of samples from the target style and generating low-quality output, respectively. However, besides their flaws, these three approaches had a chief facet of contribution in common: using some of the most important deep learning algorithms in their proposed methods. Furthermore, leveraging pre-trained language models as discriminators (Yang *et al.* 2018) in generative-adversarial-network-based systems was another instance of how practical can an approach be, provided that it is built upon a pre-trained language model. However, this unsupervised model overcame the problem of the discriminator's unstable error signal and should not be taken for granted. The same scenario happened in the generative style transformer (GST) model (Sudhakar, Upadhyay, and Maheswaran 2019). It filled the quality loss gap caused by its delete, retrieve, generate framework (Li *et al.* 2018) by powering up with a language model that made outputs' quality loss no more a debilitating concern. However, despite overcoming numerous challenges, the dearth of available target-style data remained a persistent issue. This problem was addressed by the domain adaptive model introduced by Li *et al.* (2019). Their approach was notable for its concurrent focus on relevant attributes and content within the target domain. More recently, Kumar *et al.* (2021) introduced a controlled text generation algorithm named MuCoCO. Based on conditional pre-trained language models, their decoding algorithm uses multiple differentiable constraints and formulates the challenge of text style transfer as an optimization problem.

Jin *et al.* (2022) surveyed the evolution of Text Style Transfer (TST) techniques, highlighting advancements from traditional linguistic methods to neural network-based approaches. Challenges include limited parallel data, evaluation metrics, content preservation and ethical considerations. Luo *et al.* (2023) introduced prompt-based style transfer, improving performance. Lai *et al.* (2023) assessed ChatGPT (OpenAI 2023) as a comprehensive evaluator for TST, comparing it with existing metrics. Lastly, Shibaev *et al.* (2023) proposed an information-theoretical framework for assessing information decomposition in style transfer models, providing a faster alternative to empirical experiments.

Akin to our domain-specific text style transfer approach, where we concentrated primarily on the gender domain, recent task developments have metamorphosed into domain-specific

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

methodologies as well. As a case in point, Rao and Tetreault (2018) created a large corpus for bench-marking style transfer approaches and depicted machine translation's strength as a strong baseline, specifically in sentiment transfer. Moreover, based on back-translation and sentiment analysis, Pant *et al.* (2020) introduced SentiInc to facilitate the task of sentiment-to-sentiment transfer by integrating sentiment-specific loss. Needless to mention that context plays a pivotal role in such tasks and has per se opened the doors of debate. Besides introducing two new datasets (Enron-Context and Reddit-Context), Cheng *et al.* (2020) developed CAST, a context-aware style transfer model, in which they allowed for the context being jointly considered alongside the style translation process by designing a sentence encoder and a context encoder in both presence and absence of parallel data settings. Based on their sequence-to-sequence architecture (Sutskever, Vinyals, and Le 2014), they strengthened both content preservation and context coherence and effectively took the first steps in modelling contextual details in text style transfer. Additionally, Zhou *et al.* (2020) utilized a neural style component with an attention-based sequence-to-sequence model to measure contextual word-level style relevance in an unsupervised setting. Recently, Madaan *et al.* (2020) introduced the politeness dataset containing more than 1.39 million automatically labelled samples and motivated the challenge of politeness transfer as another domain-specific text style transfer problem. They also designed the *tag* and *generate* method to tackle this problem. Later, Lai *et al.* (2021) reached a new state-of-the-art in the formality style transfer task through leveraging GPT-2 and BART pre-trained models. In terms of multilingual style transfer, Briakou *et al.* (2021) initially tackled multilingual formality style transfer by introducing a multilingual benchmark in which Brazilian, Portuguese, French, and Italian were investigated.

Assessing text style transfer methodologies is intricate due to the elusive nature of style definition and the selection of suitable evaluation metrics (Mir *et al.* 2019). Fu *et al.* (2018) proposed metrics such as Transfer Strength, evaluated by a classifier, and Content Preservation, evaluated through the cosine similarity of relative embeddings. These metrics have been shown to significantly align with human judgements. Likewise, Mir *et al.* (2019) presented Style Transfer Intensity, Content Preservation [utilizing BLEU (Papineni *et al.* 2002)] and Naturalness. To evaluate fluency and content preservation, Hu *et al.* (2022) introduced metrics such as perplexity score (PPL), style transfer accuracy (ACC), Word Overlap (WO) and self-BLEU, further summarizing these with Geometric Mean (*G*-Score) and Harmonic Mean (*H*-Score). Conversely, Yamshchikov *et al.* (2021) argued against the usage of fastText and Word2Vec embeddings for the evaluation of content preservation in text style transfer.

## 3. Our approach

In this section, we detail our proposed method for gender style transfer in multilingual contexts, primarily focusing on the English and Persian languages. Our approach, as illustrated in Figure 1, can be broken down into five key stages: Identifying gender style representatives (Section 3.2), extracting similar tokens as replacement candidates (Section 3.3), selecting target-styled candidates, returning a fluent combination and running experiments for evaluation. First, we utilize advanced language processing toolkits to identify and tag parts of speech within our chosen corpora that are representative of a particular gender style. Next, we utilize pre-trained word embeddings to suggest replacement tokens for each identified gender style representative. Following this, we deploy a character-based token classifier (Section 3.4) to select replacements that are more representative of the target gender style. Then, to maintain the coherence and fluency of the original text, we implement an optimized algorithm based on a beam search (Section 3.5) to choose the most appropriate combination of replacement tokens. Finally, we evaluate our model through a series of experiments involving human, statistical and automatic evaluations. The goal of our approach is not only to effectively transfer style from one gender to

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

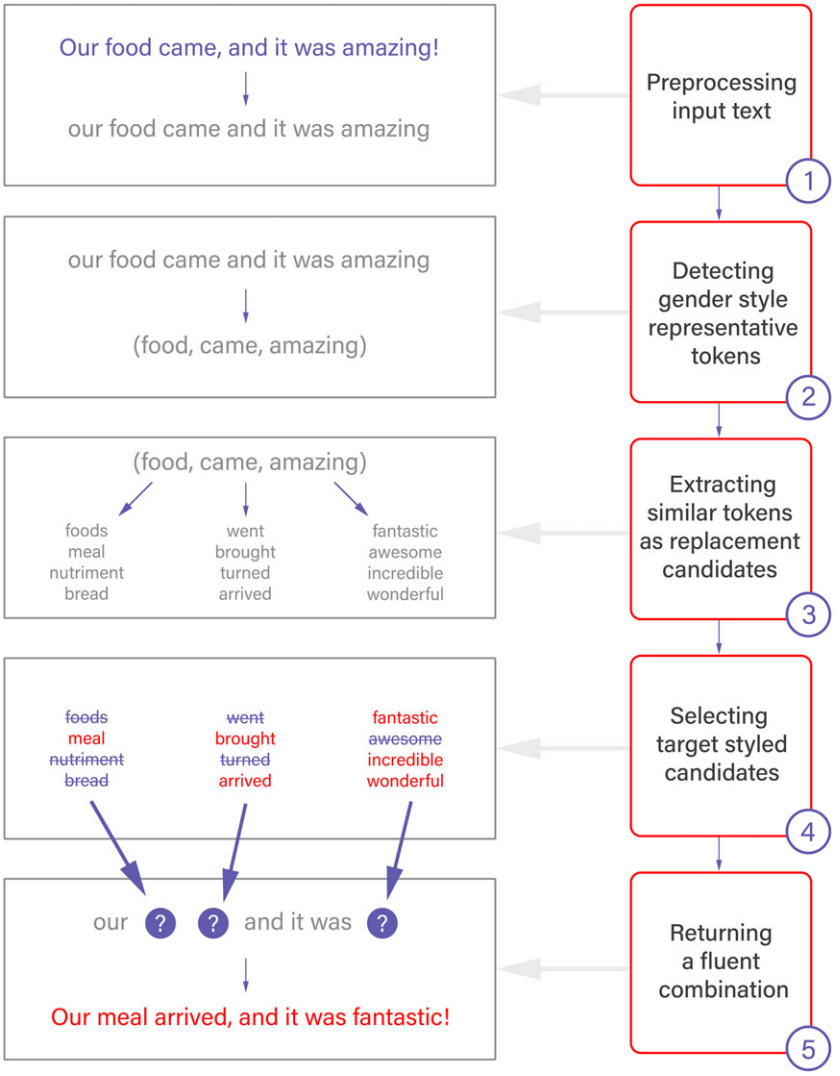


**Figure 1.** An illustration of the various stages of our proposed method. Text rendered in blue/red denotes a male/female stylistic adaptation, respectively.

another but also to ensure that the transferred text retains its original meaning and structure. All the symbols used in this paper have been collected and are listed in Table 1.

### *3.1. Baseline gender classifier*

Convolutional neural networks (CNN) have performed considerably better in terms of training time than other networks by peaking a better validation accuracy for small datasets with more consistency. We considered a 3-channel CNN layer architecture with long-short-term memory (LSTM) layers on top for our specified gender classification task. The purpose of using a multi-channel CNN architecture is to proceed with sentences in different resolutions (or $n$-grams) at each time step by defining different kernel sizes for each channel's convolutional layer. Although CNNs are generally used in computer vision, they have performed exceptionally well in capturing

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 1.** The list of symbols/notations used in this paper

| Notation | Description |
|---|---|
| sim | Cosine similarity |
| $S_s$ | Source style |
| $S_t$ | Target style |
| $t_i$ | The $i$th token of text |
| $r_{ij}$ | The $j$th replacement for the $t_i$ |
| fg | 4-gram |
| tg | Trigram |
| bg | Bigram |
| ug | Unigram |
| $f$ | The set of test samples that faked the gender classifier |
| $h$ | The set of test samples that unintentionally helped the gender classifier |
| $n$ | Size of the style transfer test set |
| $s_D$ | Standard deviation |
| $R$ | The set of samples that annotators agreed on |
| $I$ | The set of samples that annotators disagreed on |
| $K$ | Agreement percentage |
| Adj | Adjective |
| Adv | Adverb |
| $V$ | Verb |
| $N$ | Noun |
| $D_a$ | Gender classifier's test set in $S_s$ style |
| $D_r$ | The subset of $D_a$ samples predicted correctly by the classifier |
| $D_w$ | The subset of $D_a$ samples predicted incorrectly by the classifier |
| $T_a$ | $D_a$ set in $S_t$ style |
| $T_r$ | $D_r$ set in $S_t$ style |
| $T_w$ | $D_w$ set in $S_t$ style |
| $X_D$ | Pair average |
| $p$ | Number of dependent pairs in t-test |
| $M_0$ | Hypothesized mean |
| FM | Female to male transfer |
| MF | Male to female transfer |

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

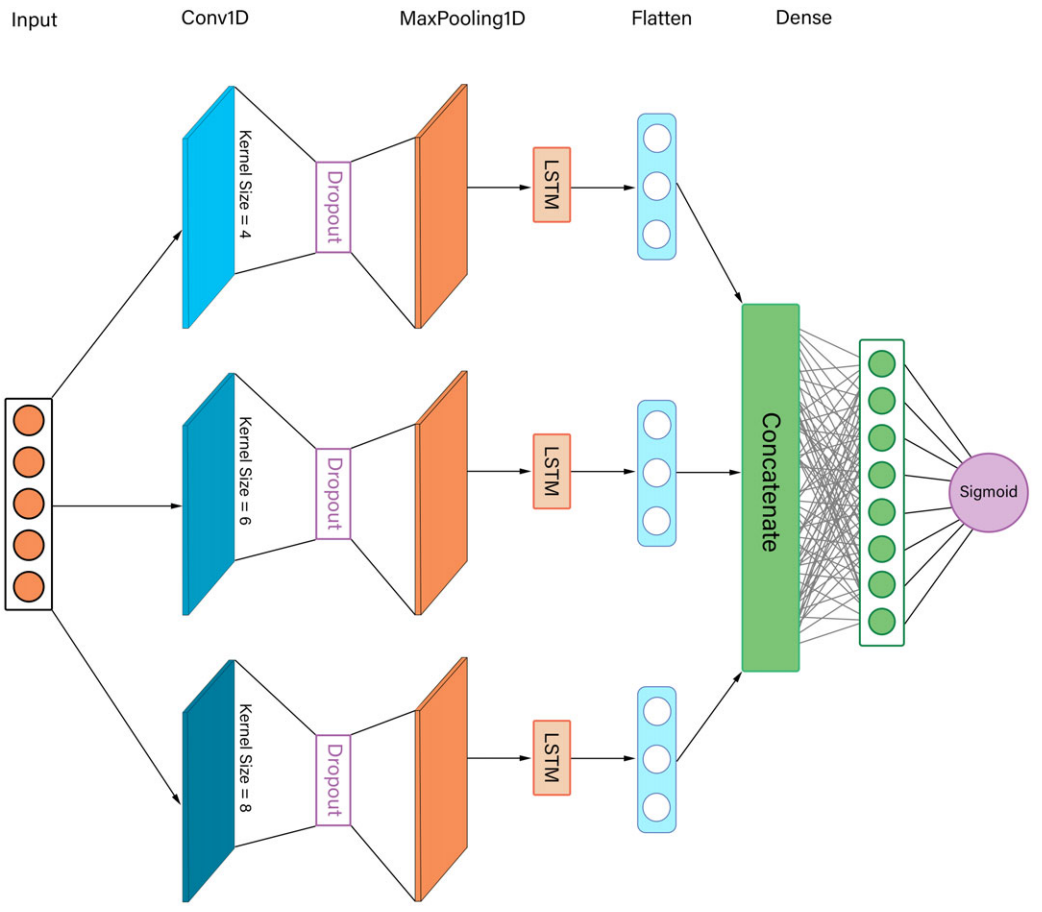


**Figure 2.** Proposed baseline gender classifier's neural network architecture.

input text patterns. The necessity of LSTM layers' presence in our architecture is that the model needs to memorize these extracted patterns. Therefore, by locating LSTM layers on top of convolutional layers, LSTM fulfils such demand (model visualized in Figure 2). Since our experiments demonstrated significantly minor improvements in Persian gender style transfer when accounting for punctuation and stop words, we remove them in the preprocessing phase before training the aforementioned 3-channel model on gender-labelled text. Referring to a sentence's source and target styles as $S_s$ and $S_t$, we transfer $S_s$ to $S_t$ in Sections 3.2–3.5.

### *3.2. Detecting gender style representative tokens*

Many previous studies have underscored the role of specific part-of-speech tags as vital indicators in the determination of an author's gender (Ishikawa 2015; Bozic Lenard 2016; Slavova, Atanasov, and Andonov 2016; Garimella *et al.* 2019). In many languages, notably in our Persian corpus, each gender-tagged sentence contains a specific set of words that play the most prominent stylistic roles. These words are typically classified as either adjectives or adverbs. The distinguishing factor between the two genders largely lies in the author's choice of words within these particular part-of-speech tags. Adjectives, recognized as the most decisive part-of-speech tag, consist of different types that vary based on the language in question. Here, we provide a list of the types of adjectives as found in English and Persian languages.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

- **English:** Adjectives are grouped either as **Descriptive** (e.g., I have a fast metabolism) or **Limiting** (e.g., I saw four cars). Although the former describes the quality of the noun, the latter limits it. Descriptive adjectives bifurcate in terms of where they are located. If the adjective appears directly beside the noun it describes, it is called an attributive adjective (e.g., *The restaurant has a* ***remarkable*** *view!*) and if connected to the noun with a linking verb, they are called predicate adjectives (e.g., *The pizza was too* ***salty****!*). Besides definite & indefinite articles, limiting adjectives are categorized as 8 different subgroups of possessive, demonstrative, indefinite, interrogative, cardinal, ordinal, proper, and nouns used as adjectives.
- **Persian:** Similar to English and most other languages, Persian adjectives also consist of descriptive adjectives. The only exception is that unlike English, attributive adjectives come after the noun. Other Persian adjective types are cardinal, ordinal, interrogative, indefinite and exclamatory adjectives.

In addition to adjectives, we also include verbs and nouns, since, as shown by Garimella *et al.* (2019) and Slavova *et al.* (2016), gender differences can also exist between these parts of speech. Thus, by identifying adjectives, adverbs, verbs and nouns present in an input text, we create a set of gender representative tokens that we will replace with those of the opposite gender and make progress towards the goal of gender style transfer.

To detect the specified tags, we used a part-of-speech tagger that would tag each token of a sentence to search for an elegant replacement token from its opposite gender in the subsequent step. We used Parsivar (Mohtaj *et al.* 2018) and Spacy (Honnibal and Johnson 2015) language processing toolkits' part-of-speech taggers to do so on our Persian and English corpora, respectively. The Parsivar toolkit is reported to be 95% accurate.

### 3.3. Extracting similar tokens as replacement candidates

By detecting gender style representatives of an input text, we look after replacements from which we may bear down on style transfer purposes. We used fastText word vectors as a pre-trained word embedding that was trained using continuous bag of words with character n-grams of length 5, a window of size 5 and 300 in dimension, specified to return the top$_n$ most similar words of a given token.

$$\text{sim} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{A}_i^2} \sqrt{\sum_{i=1}^{n} \mathbf{B}_i^2}} \tag{1}$$

Given $\mathbf{A}$ and $\mathbf{B}$ as two predetermined words' word vectors and $\mathbf{A}_i$ and $\mathbf{B}_i$ as their vector components, the cosine of the angle between word vectors is calculated using Equation (1). By considering $A$ as a gender style representative's word vector and $B$ as for all other available word vectors in the fastText's vocabulary, we apply this computation on all $(A, B)$ pairs. Altogether, whenever a token is categorized as either adjective or adverb (or any other specified tag), we pass the token to the *most_similar* built-in function of fastText to obtain a set of suggested replacement tokens with their specific similarity rates to choose between. However, to transfer a sentence's style, we have to select replacement tokens from the opposite gender suggested ones. This is where our character-based token classifier indicates the gender from which the suggested tokens are derived.

### 3.4. Selecting target styled candidates

Besides classifying sample sentences, we need to train a new model to classify tokens as either male or female. This opportunity allows us to waive those suggested replacement tokens with the same style as $S_s$ and leave the set only with the $S_t$ styled ones. We used a sequential neural network with

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

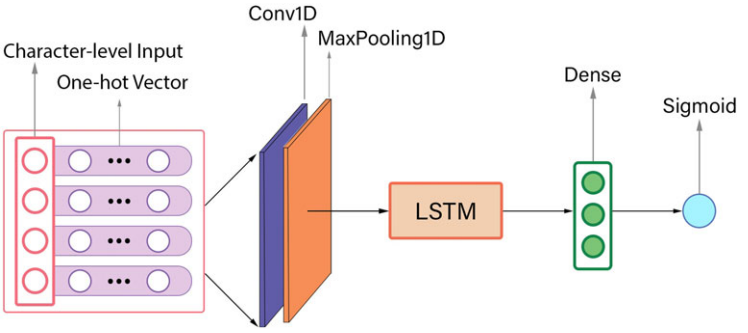


**Figure 3.** Proposed character-based token classifier's neural network architecture.

a convolutional layer and a LSTM layer on top. Each input token is represented using character-based representation, where each character is encoded using a one-hot vector. A visualization of this model is shown in Figure 3.

The reason behind using a character-based model when classifying a single token is to handle the unfortunate probability that a token is out of embedding's vocabulary (OOV) or is misspelt (Nguyen, Ngo, and Chen 2020). In either case, as long as the model is character-based, it will automatically determine the best pattern to digest the token and represent it as a vector. In addition, character-based models are better at capturing underlying emotions from an input than word-based models (Zhang, Zhao, and LeCun 2015). As Jandl *et al.* (2017) suggest, characters can convey emotions such as passion, envy and nervousness. Hence, processing documents at the character level can potentially lead the system towards obtaining a deeper understanding of input text. Furthermore, word-based models for token classification are much more likely to become biased towards a specific output label.

Findings on gender differences in Persian have revealed how male and female writers adhere to non-identical linguistic forms, resulting in the selection of words with varying frequencies (Rasekh and Saeb 2015; Jouya, Sayadian, and Naeimi 2018). This phenomenon in part is due to the nature of the Persian word, which, as opposed to the English, focuses much more directly on the gender of the subject than is seen in English. For example, the term "mother-in-law" refers to the mother of one's spouse and can be used by both genders. While in Persian, there exist two versions of this term where one addresses the mother of one's husband (مادرشوهر) or the mother of one's wife (مادرخانم). The gender-differentiated usage of Persian words thus opens up the possibility of learning their hidden gender patterns based on usage frequency. Hence, we label words according to the gender of the authors that used them more frequently. Next, we feed the aforementioned character-based model with all extracted adjectives, adverbs, verbs and nouns from both male/female sentences. As a result, the model would categorize the word embedding's suggested tokens as either male or female. A visualized result of applying this model on a word embedding space is shown in Figure 4.

It is worth to mention that extra filters are applied on specific part-of-speech tags' suggested lists. For instance, we step through each replacement candidate of a verb and remove those with the same stem or lemma. This allows us to only focus on candidates with different origins. Additionally, we transform each candidate to comply with both tense and form of the original token for which it is suggested as replacement. Hence, the method avoids replacing tokens with candidates that will damage the flow and fluency of the original text.

### *3.5. Returning a fluent combination*

As of now, we have a set of target-styled tokens for each word, and our only concern is to choose the most fluent combination. The search for such a fluent sentence heavily requires an optimized

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

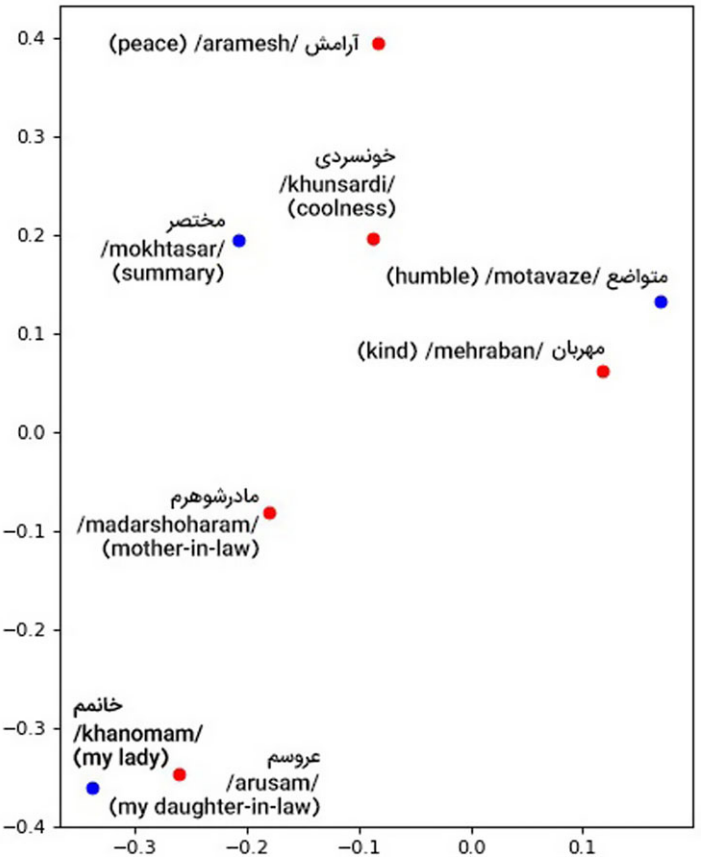


**Figure 4.** An example of fastText word embedding space that has been projected with PCA. Note: each blue/red scatter represents male/female classified. (Note: Persian pronunciations are shown between slashes, and English translations are included between parenthesis.)

algorithm and heuristic. Our developed algorithm is based on Langkilde and Knight (1998) as well as the beam search algorithm, two of the algorithms heavily studied in machine translation.

To keep up the fluency of a given sample, we keep track of all unigram, bigram, trigram and 4-gram counts in our baseline gender classifier's train set. To do so, we define a dictionary and iterate over all sentences and extract their specified n-gram counts and assign them as keys and their counts as values. This dictionary will further be used in the scorer function below:

$$\text{BeamScore} = \frac{4 \times \text{fg} + 3 \times \text{tg} + 2 \times \text{bg} + 1 \times \text{ug}}{4 \times 10 \times (1 - \text{sim})} \tag{2}$$

Given fg, tg, bg, ug and sim as the 4-gram, trigram, bigram, unigram and the similarity rate (calculated by word embedding for each of the suggested tokens), we calculate BeamScore (Equation 2) which is the mean of standardized n-gram counts and divides it by the dissimilarity of the tokens. We designate the efficacy of 40%, 30%, 20% and 10% to each of the 4 to 1-gram counts, respectively. However, the dilemma here is to score the first and the last words of a sentence and examine a suggested token's score to start or end a sentence with. Therefore, the two tags of <START>and <END> are added as tokens to each sentence's beginning and end to overcome the problem. Having defined the root as <START> and replacement tokens as nodes, the

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Algorithm 1.** Our Proposed beam search

**Data:** suggestions, BW
**Result:** BW most probable decoded texts
Add $\{< END >\}$ to *suggestions*;
$beams = \{\{< START >\}\}$;
$scores(\{< START >\}) = 0$;
**for** $row \in suggestions$ **do**
  $candidates = \{\}$;
  **for** $b \in beams$ **do**
    **for** $t \in row$ **do**
      $c' = b + \{t\}$;
      $s = BeamScore(b.top(1), b.top(2), b.top(3), t)$;
      Add $c'$ to *candidates*;
      $scores(b) += s$;
    **end**
  **end**
  $beams = bestBeams(candidates, BW)$;
**end**

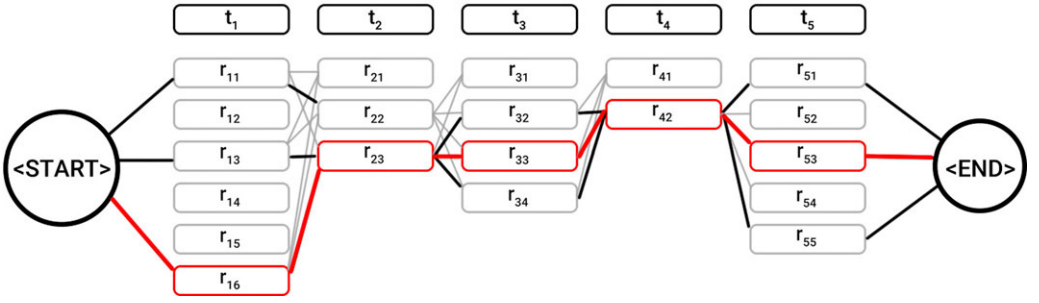


**Figure 5.** An overview of what our model's approach for transferring an input's style from $S_s$ to $S_t$ does. (Note: for an input with five tokens, we have $t_{1..5}$ and each $t_i$ has a set of $r_{ij}$ with j as the number of opposite gender predicted set of the token classifier between the top$_n$=10 word embedding's most similar suggested words.)

algorithm calculates at each step the beam score based on the traversed path's last three nodes and the visiting node until it reaches the <END> node. By the end, the algorithm returns a fluent combination of tokens since we extracted the most probable sentence and transferred an input sample from $S_s$ to $S_t$ meanwhile. The pseudocode of our implemented beam search is given in Algorithm 1.

By transferring the test set and passing it again to our gender classifier model, we measure our model's accuracy loss on $S_t$ text, which it formerly predicted in their $S_s$ style. An overview of our style transfer approach is visualized in Figure 5. We have made all implementations and models of this paper publicly available at its GitHub repository.[a]

## 4. Experiments

This section begins with a breakdown of our employed datasets and previously mentioned models by the hyperparameter choices. We then demonstrated our results in great detail and conducted

[a]https://github.com/Ledengary/PGST

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 2.** Dataset comparison

| Dataset | Style | Train | Dev | Test | Overall |
|---|---|---|---|---|---|
| Persian | Male | 105,928 | 13,241 | 13,241 | 132,410 |
| | Female | 260,164 | 32,520 | 32,521 | 325,205 |
| English | Male | 2,062,289 | 257,787 | 257,786 | 1,288,931 |
| | Female | 2,062,289 | 257,787 | 257,786 | 1,288,931 |

human, statistical, and automatic evaluations to gain a comprehensive understanding of our approach.

### *4.1. Experimental setup*

**Datasets:** Despite the focus of this study being Persian, we also apply our model to an English dataset to compare its performance with state-of-the-art English gender style transfer models. The following two datasets of Formal Gender-Tagged Persian Corpus by (Moradi and Bahrani 2016) and Gender (English) by (Reddy and Knight 2016) are used for our experiments. The Persian dataset contains 132,410 and 325,205 sentence-level samples from male and female authors, respectively. In contrast, the Gender-tagged Yelp dataset is constructed out of sentence-level reviews of different food businesses, each categorized as either male or female. A comparison between the two datasets is shown in Table 2.

**Hyperparameters:** There is an embedding, a convolutional layer and an LSTM layer prepared in each of our baseline gender classifier channels 's with the output dimension of 100 for embedding layers, 32 filters, a dropout rate of 0.5 and a maximum pool size of 2 for convolutional layers and 256 hidden units with a recurrent dropout rate of 0.2 for LSTM layers. To process texts at different N-gram levels, each channel's convolutional layer has different kernel sizes of 4, 6 and 8. The model is trained for 10 epochs with these hyperparameter choices.

In our character-based token classifier, we sequenced the same setup in a single channel as the previous model, a convolutional and LSTM layer is stacked up, with 32 filters, 8 in kernel size, max pool size of 2 for the convolutional layer and 125 hidden units for the LSTM layer. This model was trained for the same number of epochs as our baseline model with no dropouts.

With its learning rate set to 0.001, the Adam optimizer (Kingma and Ba 2015) was used in both models, as it is better at handling sparse gradients.

**Evaluation Metrics:** Assessing methodologies for text style transfer has proven to be a complex problem (Mir *et al.* 2019). The concept of style is hard to define, and this has resulted in the use of various evaluation metrics, aspects and methods. Accordingly, the justification of the style attribute while maintaining content comes with the complexity of choosing the right evaluation setup to enable comparison. Therefore, to overcome this challenge, different evaluation metrics have been presented. Given $x$ and $x'$ as the original and transferred text, Fu *et al.* (2018) introduced the following metrics which correlate considerably with human judgements:

1. **Transfer Strength:** Motivated by Shen *et al.* (2017), they used a classifier based on Keras examples, which measures transfer accuracy.
2. **Content Preservation:** To evaluate the similarity between $x$ and $x'$, they calculated the cosine similarity of their relative embeddings.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 3.** Model comparison

| | Accuracy | |
|---|---|---|
| Model | Persian | English |
| Naïve Bayes | 69% | 73% |
| Logistic regression | 62% | 70% |
| Multilingual BERT | 65% | 75% |
| SVM (Moradi and Bahrani, 2016) | 72% | – |
| CNN | 83% | 76% |
| **CNN + LSTM NN** | **90%** | **81%** |

Similarly, Mir *et al.* (2019) introduced the following:

1. **Style Transfer Intensity:** Having mapped $x$ and $x'$ to their style distributions, it quantifies the style difference of their distributions to alleviate evaluation.
2. **Content Preservation:** Unlike Fu *et al.* (2018), they utilize BLEU (Papineni *et al.* 2002) to measure the similarity between $x$ and $x'$.
3. **Naturalness:** having passed $x'$ to its function, it quantifies to what extent the transferred text is human-like.

Hu *et al.* (2022) used the perplexity score (PPL) and the style transfer accuracy (ACC) to measure the fluency and transfer strength of $x'$. Additionally, Word Overlap (WO), Cosine Similarity and self-BLEU were used to measure content preservation. Finally, they summed up all metrics in the following two:

1. **Geometric Mean (***G*-Score**):** Which is equal to the mean of 1/PPL, WO, ACC and self-BLEU. (Note that they calculated the inverse of PPL since lower PPL is preferred and that the cosine similarity metric is not included in the mean due to its insensitivity)
2. **Harmonic Mean (***H*-Score**):** The Harmonic Mean of the above sub-metrics is calculated to highlight different priorities when evaluating.

It is worth mentioning that Yamshchikov *et al.* (2021) proved in a recent study that fastText and Word2Vec pre-trained embedding vectors should not be used to evaluate text style transfer approaches in terms of content preservation. They demonstrated how such evaluation pipelines suffer from inaccurate content prediction, in analogy to similar human judgements. However, in this study, we have used only cosine similarity as one of our proposed method's crucial components to handle content preservation, but, in terms of evaluation, we left decisions to human annotators. Additionally, we utilize ACC, BLEU, PPL and a metric similar to naturalness in our automatic and human judgements as well. The details of our evaluation metrics are broken down in detail in Section 4.

### *4.2. Model evaluation*

To acquire the highest accuracy possible, we have stepped through different models and architectures. The process of choosing our gender classification model was based on a comparison between Naïve Bayes, Logistic Regression, Multilingual BERT, SVM, CNN and CNN + LSTM as baseline classifier. As shown in Table 3, our proposed baseline classifier architecture peaked the

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 4.** A comparison of the effects of the developed style transfer approaches in defeating the gender classifier model

| | Persian accuracy | | | English accuracy | | |
|---|---|---|---|---|---|---|
| Approach | $T_a$ | $T_r$ | $T_w$ | $T_a$ | $T_r$ | $T_w$ |
| 1 Word-based + (Adj, Adv) | 86% | 92% | 15% | 77% | 95% | 16% |
| 2 Word-based + (Adj, Adv, *V*) | 83% | 90% | 20% | 70% | 88% | 22% |
| 3 Word-based + (Adj, Adv, *V*, *N*) | 69% | 62% | 33% | 68% | 79% | 31% |
| 4 **Character-based + (Adj, Adv, *V*, *N*)** | **65%** | **56%** | **36%** | **34%** | **61%** | **37%** |

highest accuracy for our classification problem on Persian text with 90% and 81% in English. We finalize our model here since an efficient model in both languages is our main concern.

Due to the relatively small amount of data in our Persian corpora, a probabilistic model like Naïve Bayes performs poorly with very low precision and recall, as long as the frequency-based probability estimate becomes zero for a value with no occurrences of a class label.

Pre-trained language models like bi-directional encoder representations from transformer (Devlin *et al.* 2019) or BERT have had a significant rise due to their success in topping state-of-the-art natural language processing tasks. However, at the time of our research, there has not been any Persian-specific pre-trained transformer language model introduced. However, among proposed pre-trained models of BERT, multilingual cased contains the top 100 languages with the largest Wikipedias including Persian (Farsi). But, as shown, fine-tuning it did not have the same durability as training a classifier from the scratch.

A logistic regression model is a generalized linear model that could be reminded of a neural network with no hidden layers. Evidently, a neural network model with such hidden layers as convolutional and LSTM carries more advantages in solving our problem. convolutional layers perform outstandingly in pointing out tokens that are good indicators of an input's class, and LSTM layers in associating both short and long-term memory with the model, thus resulting in a better accuracy score in an analogy to support vector machine (SVM) algorithm (Moradi and Bahrani, 2016).

When designing a token classifier, casting each token to its 300-dimensional embedding space representation as model inputs results in contextual information loss, making a word-based model an inappropriate choice for classifying tokens as male/female. On the other side, designing a model on a character level fills the gap and distributes tokens by stylish criteria in a better way.

Achieved experimental results by repassing the test set to our gender classifier depicts our research's success the most. It demonstrates how have different approaches resulted, using different models and architectures on our set. We call our gender classifier's $S_s$ test set as $D_a$ and divide it into two different subsets: (1) $D_r$, which includes all samples that have been predicted correctly by the model, and (2) $D_w$, which consists of all samples that have been mispredicted by the model. Besides, by transferring all $D_a$ samples to $S_t$, we name the transferred set as $T_a$, $D_r$ as $T_r$ and $D_w$ as $T_w$. At inference, each of the three $D_a$, $D_r$ and $D_w$ sets have their own accuracy scores when solely passed to the classifier. In Persian, the model yielded 90% in $D_a$, 100 % in $D_r$ and 0% in $D_w$. Additionally, in English, the model performed 90% in $D_a$, 100 % in $D_r$ and 0% in $D_w$ (when the model is solely evaluated on $D_r$ and $D_w$ sets, it achieves 100% and 0% mainly because they are subsets of the $D_a$ containing samples that the classifier correctly and incorrectly predicts). In Table 4, each approach's name contains two vital information: (a) what its token classifier model was based on [word or character (Septiandri 2017)] (b) what tags are supposed to be identified by our part-of-speech tagger to be changed into their opposite gender. Our goal is to defeat the gender

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 5.** A comparison of positive and negative effects of applying different approaches

| | Persian | | | English | | |
|---|---|---|---|---|---|---|
| Approach | *f* | *h* | Trade-off | *f* | *h* | Trade-off |
| 1 Word-based + (Adj, Adv) | 3295 | 686 | 5.70 | 20,881 | 15,673 | 1.01 |
| 2 Word-based + (Adj, Adv, *V*) | 4119 | 915 | 7.77 | 51,114 | 21,550 | 5.54 |
| 3 Word-based + (Adj, Adv, *V*, *N*) | 15,651 | 1510 | 34.33 | 87,699 | 30,366 | 11.12 |
| 4 **Character-based + (Adj, Adv, *V*, *N*)** | **18,122** | **1647** | **36** | **162,870** | **36,244** | **24.56** |

identification model by transferring a sentence's style to its opposite gender, thus diminishing the gender identification model's accuracy. This means that the gender classifier mistakenly predicts male authors as female and female authors as male, which demonstrates the success of our style transfer approach. As shown in the table, it has been clearly demonstrated how the primary approach has been elevated by changing its different components. The more robust our token classifier and the more varied our part-of-speech tags' scope gets, the weaker the gender identification model's performance gets. In a task similar to ours, Septiandri (2017) demonstrated that character embeddings perform better at classifying full names and first names according to gender. Table 4 demonstrates the same effect, with the character-level model outperforming word-level models.

As mentioned, after transferring the $D$ sets, we obtain $T_a$, $T_r$ and $T_w$. The success of a style transfer approach is expected to result in a decrease in $T_r$ accuracy and remain the $T_r$ accuracy unchanged. The former shows that the transfer approach has successfully faked the gender classifier and makes it mistakenly predict samples it used to correctly classify their authors before their transfer. The latter, ideally, should remain unchanged (0%) and not be increased. However, an increase in the $T_w$ accuracy means that the transfer has failed and unintentionally aided the classifier. Accordingly, the first approach resulted in a 4% decrease overall, 8% on faking correct predictions, but unintentionally helping it correctly predict those it had mistaken before. Meaning the approach has helped the model instead of faking it. Given the number of samples that faked the model as $f$, the number of samples that unintentionally helped the model as $h$ and the size of the test set as $n$, we define the trade-off value in Equation (3). This formula is designed to capture the effects of both $f$ and $h$ to ease comparison between approaches (Table 5). With this formula, those which fake the model more and help it less, receive higher scores.

$$\text{Trade-off} = \frac{f - h}{n} \times 100 \tag{3}$$

As shown in Table 6, Persian $D_a$ contains 45,762 test samples in which 41,185 samples (90%) were correctly guessed and belong to set $D_r$ and 4577 samples (10%) guessed incorrectly by the gender identification model, which belongs to set $D_w$. As shown in Table 5, there was an 8% decrease in $D_r$'s accuracy (3295 samples) and a 15% rise in $D_w$ (686 samples), resulting in a trade-off value of 5.70, which demonstrates its lack of ability in defeating gender identification. But as we go along testing approaches 2, 3 and 4, we get back the trade-off values of 7.77, 34.33 and 36 in Persian. The major leap between the second and third models' trade-off values represents the pivotal role that a bigger scope of part-of-speech tags plays. In addition, switching the token classifier's processing level from word to character in the fourth approach also shows how it affects efficiency. The same evaluations have been made in English where the final approach yielded the results of 34%, 61% and 37% for $T_a$, $T_r$ and $T_w$, and the trade-off value of 24.56. Specific contingencies of applying our finalized approach [i.e., Character-based + (Adj, Adv, $V$, $N$)] on our defined sets of $D_a$ and $T_a$ in both languages are shown in Table 5, which demonstrates the number of samples that were predicted either correctly or incorrectly by our baseline gender classifier.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 6.** A contingency table on our finalized style transfer approach [i.e., character-based + (Adj, Adv, *V*, *N*)] in Persian and English

| | Persian | | English | |
|---|---|---|---|---|
| Set | Correctly | Incorrectly | Correctly | Incorrectly |
| $D_a$ (all docs) | 41,185 | 4577 | 417,613 | 97,959 |
| $D_a$ (male docs) | 11,253 | 2011 | 137,783 | 58,789 |
| $D_a$ (female docs) | 29,932 | 2566 | 279,830 | 39,170 |
| $T_a$ (all docs) | 29,745 | 16,017 | 175,294 | 340,278 |
| $T_a$ (male docs) | 12,418 | 9390 | 106,247 | 222,647 |
| $T_a$ (female docs) | 17,327 | 6627 | 69,047 | 117,631 |

In order to prevent the transfer approach from unintentionally helping the classifier to predict the samples correctly, amplifying our character-based token classifier is the most rational alternative, since converting a sample's content to its target style is its primary essence, and content is what the classifier is obligated to detect.

### 4.3. Statistical evaluation

To determine whether there is a significant difference between the means of the two gender labels, we use a statistical hypothesis testing tool called t-test (Kim 2015) to assure that there is not an unknown variance and that labels are all distributed normally.

In order to assure if the two gender labels come from the same population, by taking samples from each of the two labelled sets, t-test hypothesizes that the two means are equal. By calculating certain values and comparing them with the standard values afterwards, t-test decides whether inputs are strong and not accidental or that they are weak and probably due to chance, resulting in rejection and acceptance of the hypothesis, respectively.

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{p}} \tag{4}$$

We use inputs in both original and transferred style in Equation (4) to measure the t statistic value for dependent paired samples in which $\overline{X}_D$ and $s_D$ are each pair's average and standard deviation of their difference, $p$ as the number of pairs, and $\mu_0$ as hypothesized mean, which we assign to zero when testing the average of the difference.

The p-value is the probability of obtaining an equal or more extreme result than the one obtained when the hypothesis is true. *Significance level* (or alpha) is a threshold value which is the eligible probability of making a wrong decision (rejecting the hypothesis). We have calculated $p$-values in both languages by considering $n - 1$ degrees of freedom and assigning 0.01 to alpha. As demonstrated in Table 7, test results are significant in both languages with their acquired $p$-values.

### 4.4. Human evaluation

A proper assessment of the generated text is necessary to prove its correctness. We considered facets like *fluency* and *semantic* that each sample had to be assessed based on. The former facet determines whether a given text is reasonably close to legible human language or that it is presented in an indecipherable manner. As the name implies, the latter is employed to

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 7.** *P*-values for paired samples in our corpora (alpha = 0.01, degree of freedom = $n - 1$)

| Language | *p*-values |
|---|---|
| Persian | 1.64813366054e-10 |
| English | 4.04649941132e-3 |

evaluate inputs depending on conceptual meanings when interpreted. We additionally added an *adulteration* facet to determine whether the given text seemed adulterated or not. Most importantly, annotators were also asked to assign 1 to samples where they believed the author was male and 0 to samples where they believed it was female which is indicated as the *gender* facet. We randomly selected 300 inputs that included 75 $S_s$ and 75 $S_t$ styled samples in both English and Persian. By shuffling inputs and dividing them by three, we assigned each 100 samples to an annotation group with three different annotators. Annotators were then asked to rank each sample based on our set of criteria in binary form. The reason for using $S_s$ texts from the English and Persian corpora in human evaluation is to draw a comparison between how ground-truth samples are overall ranked. This then makes it more comprehensive to understand human judgements among $S_t$ samples.

#### 4.4.1. Inter-annotator agreement

Before particularizing annotations, we test inter-rater reliability with kappa (Viera and Garrett 2005), a standard measure of inter-annotator agreement which aims to compare the amount of agreement that we are actually getting between judges to the amount of agreement that we would get purely by chance.

By letting $N$ be the number of samples and defining $R$ and $I$ as two sets of agreed and disagreed samples for each of the three annotators in a specific group, $A$ would be the set of samples where all three annotators agreed on and $P(A)$ and $P(E)$ as fractions of real and accidental agreements.

$$A = (R_1 \cap R_2 \cap R_3) \cup (I_1 \cap I_2 \cap I_3) \tag{5}$$

$$P(A) = \frac{|A|}{N} \tag{6}$$

$$P(E) = \left(\frac{R_1}{N}\right)\left(\frac{R_2}{N}\right)\left(\frac{R_3}{N}\right) + \left(\frac{I_1}{N}\right)\left(\frac{I_2}{N}\right)\left(\frac{I_3}{N}\right) \tag{7}$$

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{8}$$

$K$ value is calculated for each facet in every group. Finally, their average score is stated in Table 8 which demonstrates the stability of annotations since such obtained results are counted as *substantial* ones in Kappa inter-annotator agreement's jargon.

#### 4.4.2. Quality assessment

Since our annotator groups each consists of three different annotators, when classifying each sample's facet, we consider the two most agreed on opinion as the sample's final class (e.g., if at least two out of three annotators classified a sample as 1 in fluency, we call that a fluent sample). Table 9

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 8.** Results of Kappa inter-annotator agreement

| Criteria | *K* | Agreement level |
|---|---|---|
| Fluency | 74.1% | Substantial |
| Semantic | 75.32% | Substantial |
| Adulteration | 69.96% | Substantial |
| Gender | 68.43% | Substantial |

**Table 9.** Quality assessment of annotated samples

| Set | Fluency | Semantic | Adulteration | Gender |
|---|---|---|---|---|
| $S_s$ Persian | 97.33% | 97.33% | 21.33% | 84.10% |
| $S_t$ Persian | 77.03% | 63.51% | 40.54% | 80.53% |
| $S_s$ English | 98.67% | 90.67% | 26.67% | 85.50% |
| $S_t$ English | 75.0% | 68.42% | 28.95% | 68.95% |

demonstrates random samples' quality assessment based on their language and style. The important thing to note in Table 8 is that the source samples in both languages are not identical as "ground-truth style transferred". Meaning we do not aim to approach those accuracies. They are simply raw text from the dataset (not transferred) that have been given to annotators to be assessed alongside the transferred samples. The reason we have also asked the annotators to assess raw dataset text is to provide them with an unbiased annotation scheme towards transferred-only outputs and demonstrate the evaluation confidence of annotators.

**Fluency:** As mentioned in Algorithm 1, when choosing the right combination among all suggestions, we prioritize tokens with the highest frequency in different n-gram scopes (Equation 2) when choosing among replacements. This strategy will lead us towards a fluent $S_t$ sample which is clearly proved by annotations shown in the table, with high accuracies of 77% and 75% in Persian and English.

**Semantic:** Our results have marginally lower semantic accuracy compared with the samples' overall fluency, which is probably due to the literary nature of the Persian and the informality of the English corpora. In spite of the difficulty of replacing intended tokens in such non-identically styled corpora in the two languages, our approach demonstrates its compatibility with such diverse texts.

**Adulteration:** The rationale for using $S_s$ samples in annotating process was to see if $S_t$ samples seemed evidently adulterated among the others, or that they obeyed of a similar format, which surprisingly, nearly the same amount of $S_t$ English samples was detected as adulterated (28.95%) as the $S_s$ English ones (26.67%), which heralds of low difference among them. Since our acquired pre-trained Persian word embedding is not as well-trained as pre-trained English word vectors, $S_t$ Persian samples appear in greater frequency (40.54%) than $S_s$ Persian s (21.33%). Additionally, the cumbersome structure of Persian literary texts makes it even more difficult to deal with during the transfer process.

**Gender:** Having asked the annotators to label each sample as either male (1) or female (0), we then compared their answers to their true labels to see how accurate human judgement is in determining the gender of authors. As indicated, the gender can be more accurately predicted in Persian rather than English. This is potentially due to the substantial linguistic differences between

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 10.** Comparison of automatic evaluation results of different models in English

| Model | BLEU | Perplexity | Accuracy loss |
|---|---|---|---|
| Source | 100 | 183.4 | 18.9 |
| BT | 46.0 | 196.2 | 52.9 |
| G-GST | 78.5 | 252.0 | 49.0 |
| B-GST | **82.5** | **189.2** | **57.9** |
| PGST | 68.4 | 198.9 | 45.6 |

Persian and English which made our model more efficient in the former (80.53%), for which it was exclusively developed, than in the latter (68.95%).

### *4.5. Automatic evaluation*

To indicate our proposed method's correctness, we assess our method via automatic evaluation measurements in different aspects of fluency, content preservation and transfer strength. Previous works measured transfer strength with a style classifier, explicitly trained for evaluation purposes. But since proposing such a classifier was one of our key contributions, our model has already been particularized in previous sections. This helps us measure the loss of accuracy (the division of the number of correctly guessed authors' genders by size of the test set) caused by the style transfer approach. The higher the loss gets, the better the approach has performed. We also employed BLEU (Papineni *et al.* 2002) and OpenAI GPT-2 language model, respectively, to measure content preservation and fluency of our transferred set. Although evaluating with such automatic metrics like BLEU is inadequate if being applied single-handedly (Sulem, Abend, and Rappoport 2018), it benefits us with a general understanding of how preserved a $S_t$ sample's content is. But in fluency terms, we calculated the perplexity of those samples using GPT-2 language model.

As a comparison with previous work, we intend to compare our method (**PGST**) on the analogy of three other previously proposed methods. Prabhumoye *et al.* (2018) came up with the idea of adversarial mechanism and Back-Translation (**BT**) and Sudhakar *et al.* (2019) proposed **B-GST** and **G-GST** which respectively were blinded and guided towards particular desired $S_t$ style attributes using transformers (Vaswani *et al.* 2017) which at the time of writing this paper and to our best of knowledge, is the state-of-the-art on our mutual English gender-tagged dataset.

When making an analogy, the key factor is not to consider each evaluation metric separately but to contemporaneously assess them all together. As shown in Table 10, in terms of target-style accuracy, the **BT** model performs admissibly well, but its generated text does not preserve much content, thus resulting in a low BLEU score, whereas in $S_t$ style matter our method almost obtained the same result as state-of-the-art method collateral, **G-GST**, but with much lower perplexity and higher BLEU score comparing to the prior state-of-the-art method, **BT**. All in all, it can be concluded that besides other monolingual methods, even with a much simpler foundation to a multilingual extent, our proposed method has achieved reasonable success in English, whereas our main focus was on devoting such a method in Persian.

## 5. Discussion

The aim of this section is to discuss the advantages and disadvantages of our proposed method and to determine under what circumstances it succeeds or fails to overcome the challenges it may

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 11.** Test samples of transferring Persian text using the PGST method

| Examples | $S_s$ style | $S_t$ style |
|---|---|---|
| FM@1 | از لحظه ورودش به شیراز، عطر شیرین بهار نارنج را حس می‌کرد. | از بدو ورودش به شیراز، بوی خوش بهار نارنج را حس می‌کرد. |
| FM@2 | ستاره‌هایی که درپوست آسمان فرو رفته بودند، هریک گونه‌ای می‌درخشیدند | اخترهایی که درپوست سما داخل شده بودند، هریک گونه‌ای می‌درخشیدند |
| FM@3 | پسرک عزیزم بیا، تو با آن پیراهن نازکت سرما خواهی خورد | دخترک دلبندم بیا، تو با آن لباس نازکت سرما خواهی خورد |
| FM@4 | رخسار مادرم از شدت عصبانیت سرخ شده بود و به کبودی می‌زد. | رخسار پدرم از شدت غضب قرمز شده بود و به تیرگی می‌زد. |
| FM@5 | او با دست آزاد چادر سفید گلدارش را مرتب کرد و به دنبال مادر راهی اتاق شد. | او با دست آزاد لباس سفید طرحدارش را منظم کرد و به دنبال پدر راهی اتاق شد. |
| Examples | $S_s$ style | $S_t$ style |
| MF@1 | آب از لبش می‌ریخت روی گوش‌ماهی‌ها و از پشت دستش می‌ریخت روی زمین و با بقیه‌ی قطره‌های شیشه‌ای باران جمع می‌شد. | آب از لبش می‌چکید روی گوش‌ماهی‌ها و از پشت دستش می‌چکید روی زمین و با دیگر چکه‌های بلورین باران انباشته می‌شد. |
| MF@2 | آبدارچی که بیرون رفت استکان چای را هی جلوی چشمش و در مقابل نوری که از پنجره به اتاق می‌آمد بالا و پایین برد. | خدمتکار که بیرون رفت ظرف چای را هی جلوی دیدش و در مقابل نوری که از پنجره به اتاق می‌آمد بالا و پایین کرد. |
| MF@3 | تو اماکن عمومی خیلی سریع خجالتش عود می‌کرد و مثل کودک مدرسه‌ای ها سرخ می‌شد. | در مکان‌ها عمومی خیلی زود خجالتش عود می‌کرد و مثل کودک مدرسه‌ای ها گلگون می‌شد. |
| MF@4 | از سخن نغز او خوشم می‌آید | از کلام شیرین او خوشم می‌آید |
| MF@5 | او همیشه یك روزنه امید در سختی‌ها پیدا می‌کرد. | او پیوسته یك دریچه امید در مشکلات پیدا می‌کرد. |

FM stands for Persian Female to Male transfer. MF stands for Persian Male to Female transfer.

encounter. Generally, the most significant difference between our work and the ones mentioned above is the underlying intention behind how words are replaced. More specifically, in contrast to previously proposed token-replacing techniques, our work hypothesized that there are potential gender-differentiating patterns among certain parts of speech tags. This claim is backed up with linguistic and sociolinguistic claims of prior studies, making the development of a more straightforward yet effective Persian text style transfer approach more feasible in the absence of Persian variants of state-of-the-art transformer models.

In our study, the Persian dataset presents additional noteworthy computational challenges. Unlike English, the dataset contains a variety of stories and books which makes it a formal corpus as a whole. Nevertheless, we have also spotted informal text in instances where the author wishes to quote someone. Altogether, the Persian dataset comprises both formal and informal texts each posing unique challenges. Furthermore, the vocabulary consists of OOV words, mostly due to the dataset's poetic or ancient literature. However, in the English dataset, almost no slang, contractions or abbreviations are seen.

In Table 11, we provide ten Persian test samples (a translation of the content is provided in Table 12) where one-half of the samples pertains to **F**emale to **M**ale (FM) transformation and the other to **M**ale to **F**emale (MF) transformation. In FM@2, the *BeamSearch* function's effect can be seen where a Persian expression has been successfully decoded as final text. Besides, the MF@1 sample indicates that Persian complex verbs are also being handled by the method. In specific terms, the method fails to ignore artificial replacements such as in FM@4 and FM@1. Interestingly, even though Persian has no gender-distinguished nouns, "dress" and "clothes" are handled in the third example (FM@3). Moreover, stemming and lemmatizing candidates may not necessarily aid us in selecting better replacements when solely utilized. Hence, we considered embedding suggestions of both the filtered and the raw formats of the words. For instance, MF@1 is an example that backs this claim up.

Despite the applicability of our PGST method to an English corpus for comparative purposes, it's imperative to highlight that PGST has been primarily devised and optimized considering the unique nuances of the Persian language, which inherently restricts its direct applicability to

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Table 12.** Translation of Table 11's test samples

| Translations | $S_s$ style | $S_t$ style |
|---|---|---|
| FM@1 | From the moment he arrived in Shiraz, he could feel the sweet aroma of orange blossom. | Upon his arrival in Shiraz, he smelled the pleasant aroma of orange blossom. |
| FM@2 | The stars that had plunged into the skin of the sky, each shone in a way. | The stars that had entered the skin of the sky, each shone somehow. |
| FM@3 | Come on my dear son, you will catch a cold with that thin shirt of yours. | Come on my beloved daughter, you will catch a cold with that thin dress of yours. |
| FM@4 | My mother's face was flushed red hot with anger and it got almost bruised-like. | My mother's face was flushed red with rage, and it got almost dark. |
| FM@5 | She organized her white veil With her free hand and followed her mother into the room. | He organized his white clothes with his free hand and followed his father into the room. |
| Translations | $S_s$ style | $S_t$ style |
| MF@1 | Water poured from her lips on the scallops and poured from the back of her hand on the ground and got added to the rest of the glassy raindrops. | Water dripped from her lips on the scallops and dripped from the back of her hand on the ground and got accumulated with the rest of the crystalline raindrops. |
| MF@2 | When the butler went out, she raised and lowered the cup of tea in front of her eyes and in front of the light that was coming from the window. | When the butler went out, she raised and lowered the cup of tea in front of her sight and in front of the light that was coming from the window. |
| MF@3 | In public, she was quickly embarrassed and blushed like a schoolgirl. | In public, she was quickly embarrassed and reddened like a schoolboy. |
| MF@4 | I like her bon mot. | I like her sweet words. |
| MF@5 | She always found a glimmer of hope in hardships. | She constantly found a break of hope in difficulties. |

The translations are not necessarily valid as English-transferred samples and are only provided to aid non-Persian readers in understanding samples indicated in Table 11.

other languages, including English. Rooted in the PGST approach is an in-depth comprehension of Persian's distinct linguistic subtleties, its semantic constructs and the gender-based lexical variations that this language uniquely offers. PGST capitalizes on exclusive Persian features like its flexible word order, higher OOV word ratio and different scripts for the same letter, among others. Additionally, the absence of definite articles in Persian and a common occurrence of adjectives replacing nouns pose complexities which our method is specifically designed to handle. Conversely, English and many other languages follow distinct linguistic rules and patterns, starkly different from Persian. English, for instance, utilizes definite articles, maintains a stricter word order and employs a different mechanism for gendered language. Adapting our approach to English or any other language would mandate substantial alterations to accommodate these unique linguistic disparities. While PGST lays out a robust groundwork for text style transfer, it's vital to understand its success is inherently tied to its bespoke design for Persian. Applying it to other languages is an intriguing proposition, but would require developing new models, with the unique linguistic characteristics of the target language taken into account.

## 6. Conclusion and future work

In this study, we have tackled the challenging task of text style transfer between genders in Persian. Our work contributes to the existing body of research by introducing the first instance of a text

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

style transfer method specifically designed for this low-resource natural language, while also showcasing its versatility through application to English texts. The novelty of our approach lies in its unique ability to navigate linguistic nuances across different languages and genders, providing a more inclusive and comprehensive solution to style transfer. Our experimental results, evaluated using statistical, automatic, and human metrics, demonstrate the efficacy of our approach. We've found that our method consistently produces high-quality style transfers without compromising the original meaning or structure of the texts. Not only do these results stand favourably in comparison to current leading methods in English style transfer, but they also set a new baseline for future work in Persian text style transfer. Furthermore, by closely analysing the patterns in our model's style transfers, we have uncovered new information about gender-specific language use in both English and Persian from a sociolinguistic perspective. This highlights the broader implications of our work, suggesting that our method could be used as a tool for sociolinguistic research in addition to its primary function as a style transfer model. By introducing a novel approach that is both linguistically versatile and sociolinguistically insightful, we believe we have advanced our understanding of gendered language use and opened new avenues for future research in this field. Looking forward, while text style transfer has become a well-developed task in highly resourced languages, we anticipate seeing a more focused research effort for low-resource languages. At the time of writing this paper, attention-based models and transformers are yet to be developed for Persian, and their implementation could significantly contribute to advancements in this language.

## References

**Ahmadi P.**, **Tabandeh M. and Gholampour I.** (2016). Persian text classification based on topic models. In *2016 24th Iranian Conference on Electrical Engineering (ICEE)*, pp. 86–91.

**Baker P.** (2014). *Using Corpora to Analyze Gender*, vol. 19. London: Bloomsbury Publishing. ISBN: 9781441108777.

**Bozic Lenard D.** (2016). Gender differences in the personal pronouns usage on the corpus of congressional speeches. *Journal of Research Design and Statistics in Linguistics and Communication Science* **3**(2), 161–188.

**Briakou E.**, **Lu D.**, **Zhang K. and Tetreault J.** (2021). Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online. Association for Computational Linguistics, pp. 3199–3216.

**Bsir B. and Zrigui M.** (2018). Enhancing deep learning gender identification with gated recurrent units architecture in social text. *Computacion y Sistemas* **22**(3), 757–766.

**Bucholtz M.** (2002). *From 'sex Differences' to Gender Variation in Sociolinguistics*. University of Pennsylvania Working Papers in Linguistics.

**Cavas B.** (2010). A study on pre-service science, class and mathematics teachers' learning styles in turkey. *Science Education International* **21**, 47–61.

**Chen X.**, **Duan Y.**, **Houthooft R.**, **Schulman J.**, **Sutskever I. and Abbeel P.** (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16*. Red Hook, NY: Curran Associates Inc, pp. 2180–2188.

**Cheng N.**, **Chandramouli R. and Subbalakshmi K.** (2011). Author gender identification from text. *Digital Investigation* **8**(1), 78–88.

**Cheng Y.**, **Gan Z.**, **Zhang Y.**, **Elachqar O.**, **Li D. and Liu J.** (2020). Contextual text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online. Association for Computational Linguistics, pp. 2915–2924.

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

**Eckert P.** (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change* **1**(3), 245–267.

**Fatima M.**, **Anwar S.**, **Naveed A.**, **Arshad W.**, **Nawab R. M. A.**, **Iqbal M. and Masood A.** (2018). Multilingual sms-based author profiling: Data and methods. *Natural Language Engineering* **24**(5), 695–724.

**Feng S.**, **Wallace E.**, **Grissom II A. I.**, **Iyyer M.**, **Rodriguez P. and Boyd-Graber J.** (2018). Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 3719–3728.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Fu Z.**, **Tan X.**, **Peng N.**, **Zhao D. and Yan R.** (2018). Style transfer in text: Exploration and evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**(1), 663–670.

**Garimella A.**, **Banea C.**, **Hovy D. and Mihalcea R.** (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 3493–3498.

**Gatys L. A.**, **Bethge M.**, **Hertzmann A. and Shechtman E.** (2016). Preserving color in neural artistic style transfer. *ArXiv*, abs/1606.05897.

**Gatys L. A.**, **Ecker A. S. and Bethge M.** (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:* 1508.06576.

**Gong H.**, **Bhat S.**, **Wu L.**, **Xiong J. and Hwu W.-m.** (2019). Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics, pp. 3168–3180.

**Honnibal M. and Johnson M.** (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics, pp. 1373–1378.

**Hoyle A. M.**, **Wolf-Sonkin L.**, **Wallach H.**, **Augenstein I. and Cotterell R.** (2019). Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 1706–1716.

**Hu Z.**, **Lee R. K.-W.**, **Aggarwal C. C. and Zhang A.** (2022). Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter* **24**(1), 14–45.

**Hu Z.**, **Yang Z.**, **Liang X.**, **Salakhutdinov R. and Xing E. P.** (2017). Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML'17*, vol. 70, pp. 1587–1596. JMLR.org.

**Ishikawa Y.** (2015). Gender differences in vocabulary use in essay writing by university students. *Procedia - Social and Behavioral Sciences* **192**, 593–600. The Proceedings of 2nd Global Conference on Conference on Linguistics and Foreign Language Teaching.

**Jandl I.**, **Knaller S. and Schönfellner S.** (2017). *Writing Emotions: Theoretical Concepts and Selected Case Studies in Literature*. Bielefeld: Transcript Publishing. ISBN: 9783837637939.

**Jin D.**, **Jin Z.**, **Hu Z.**, **Vechtomova O. and Mihalcea R.** (2022). Deep learning for text style transfer: A survey. *Computational Linguistics* **48**(1), 155–205.

**Jin-yu D.** (2014). Study on gender differences in language under the sociolinguistics. *Canadian Social Science* **10**, 92–96.

**John V.**, **Mou L.**, **Bahuleyan H. and Vechtomova O.** (2019). Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, pp. 424–434.

**Jouya M.**, **Sayadian S. and Naeimi A.** (2018). The role of gender in persian translations of a thousand splendid suns based on waddington's model. *Jostarnameh Journal of Comparative Literature Studies* **2**(3), 113–136.

**Kayaoğlu M.** (2012). Gender-based differences in language learning strategies of science students. *Journal of Turkish Science Education* **9**(2), 12–24. ISSN: 1304-6020.

**Kim T.** (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology* **68**(6), 540.

**Kingma D. and Ba J.** (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, pp. 1–13.

**Kumar S.**, **Malmi E.**, **Severyn A. and Tsvetkov Y.** (2021). Controlled text generation as continuous optimization with multiple constraints. In *Advances in Neural Information Processing Systems (NourIPS)*, vol. 34. Available at https://openreview.net/forum?id=kTy7bbm-4I4

**Lai H.**, **Toral A. and Nissim M.** (2021). Thank you BART! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online. Association for Computational Linguistics, pp. 484–494.

**Lai H.**, **Toral A. and Nissim M.** (2023). Multidimensional evaluation for text style transfer using ChatGPT. Chicago. arXiv preprint. arXiv:2304.13462.

**Langkilde I. and Knight K.** (1998). Generation that exploits corpus-based statistical knowledge. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, vol. 1. Montreal: Association for Computational Linguistics, pp. 704–710.

**Li D.**, **Zhang Y.**, **Gan Z.**, **Cheng Y.**, **Brockett C.**, **Dolan B. and Sun M.-T.** (2019). Domain adaptive text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 3304–3313.

**Li D.**, **Zhang Y.**, **Gan Z.**, **Cheng Y.**, **Brockett C.**, **Sun M.-T. and Dolan B.** (2018). Delete, retrieve, generate: A simple approach to sentiment and 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)*. New Orleans: Association for Computational Linguistics, pp. 1865–1874.

**Li N. and Kirkup G.** (2007). Gender and cultural differences in internet use: A study of china and the uk. *Computers & Education* **48**(2), 301–317.

**Luan F., Paris S., Shechtman E. and Bala K.** (2017). Deep photo style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6997–7005.

**Luo G., Han Y. T., Mou L. and Firdaus M.** (2023). Prompt-based editing for text style transfer. arXiv preprint. arXiv:2301.11997.

**Madaan A., Setlur A., Parekh T., Poczos B., Neubig G., Yang Y., Salakhutdinov R., Black A. W. and Prabhumoye S.** (2020). Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 1869–1881.

**Martinc M. and Pollak S.** (2018). Reusable workflows for gender prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).

**Martinc M. and Pollak S.** (2019). Combining n-grams and deep convolutional features for language variety classification. *Natural Language Engineering* **25**(5), 607–632.

**Metin M., Yılmaz G. K., Birişçi S. and Coşkun K.** (2011). The investigating pre-service teachers' learning styles with respect to the gender and grade level variables. *Procedia - Social and Behavioral Sciences* **15**, 2728–2732. 3rd World Conference on Educational Sciences - 2011.

**Mir R., Felbo B., Obradovich N. and Rahwan I.** (2019). Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, MN: Association for Computational Linguistics, pp. 495–504.

**Mohtaj S., Roshanfekr B., Zafarian A. and Asghari H.** (2018). Parsivar: A language processing toolkit for Persian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki: European Language Resources Association (ELRA).

**Moradi M. and Bahrani M.** (2016). Automatic gender identification in persian text. *Signal and Data Processing* **12**(4), 83–94.

**Nahavandi N. and Mukundan J.** (2014). Language learning strategy use among Iranian engineering EFL learners. *Advances in Language and Literary Studies* **5**, 34–45.

**Nguyen M., Ngo G. H. and Chen N. F.** (2020). Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **28**, 461–473.

**Norberg C.** (2016). Naughty boys and sexy girls: The representation of young individuals in a web-based corpus of english. *Journal of English Linguistics* **44**(4), 291–317.

**OpenAI** (2023). *Chatgpt* (accessed 2 June 2023).

**Pant K., Verma Y. and Mamidi R.** (2020). SentiInc: Incorporating sentiment information into sentiment transfer without parallel data. In *European Conference on Information Retrieval*. Cham: Springer International Publishing, pp. 312–319.

**Papineni K., Roukos S., Ward T. and Zhu W.-J.** (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*. Association for Computational Linguistics, pp. 311–318.

**Pearce M.** (2008). Investigating the collocational behaviour of man and woman in the bnc using sketch engine. *Corpora* **3**(1), 1–29.

**Prabhumoye S., Tsvetkov Y., Salakhutdinov R. and Black A. W.** (2018). Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne: Association for Computational Linguistics, pp. 866–876.

**Rao S. and Tetreault J.** (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)*. New Orleans: Association for Computational Linguistics, pp. 129–140.

**Rasekh A. E. and Saeb F.** (2015). Gender differences in the use of intensifiers in persian. *International Journal of Applied Linguistics and English Literature* **4**(4), 200–204.

**Reddy S. and Knight K.** (2016). Obfuscating gender in social media writing. In *Workshop on Natural Language Processing and Computational Social Science*, pp. 17–26.

**Septiandri A. A.** (2017). Predicting the gender of indonesian names. *ArXiv*, abs/1707.07129.

**Serizel R. and Giuliani D.** (2017). Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering* **23**(3), 325–350.

**Shamsfard M.** (2019). Challenges and opportunities in processing low resource languages: A study on persian. In *International Conference Language Technologies for All (LT4All)*.

https://doi.org/10.1017/S1351324923000426 Published online by Cambridge University Press

**Shen T.**, **Lei T.**, **Barzilay R. and Jaakkola T.** (2017). Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*. Red Hook, NY: Curran Associates Inc., pp. 6833–6844.

**Shibaev V.**, **Olbrich E.**, **Jost J. and Yamshchikov I. P.** (2023). Quick estimate of information decomposition for text style transfer. *Entropy (Basel)* **25**(2), 322.

**Slavova V.**, **Atanasov D. and Andonov F.** (2016). Gender differences in the use of noun concept categories – a statistical study based on data from child language acquisition. *International Journal "Information Content and Processing"* **3**(2), 103–116.

**Soler J. and Wanner L.** (2016). A semi-supervised approach for gender identification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož: European Language Resources Association (ELRA), pp. 1282–1287.

**Sotelo A. F.**, **Gómez-Adorno H.**, **Esquivel-Flores O. and Bel-Enguix G.** (2020). Gender identification in social media using transfer learning. *Pattern Recognition* **12088**, 293–303.

**Subramanian S.**, **Lample G.**, **Smith E. M.**, **Denoyer L.**, **Ranzato M. and Boureau Y.-L.** (2018). Multiple-attribute text style transfer. *ArXiv*, abs/1811.00552.

**Sudhakar A.**, **Upadhyay B. and Maheswaran A.** (2019). "Transforming" delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, pp. 3269–3279.

**Sulem E.**, **Abend O. and Rappoport A.** (2018). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, pp. 738–744.

**Sutskever I.**, **Vinyals O. and Le Q. V.** (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*. Cambridge, MA: MIT Press, pp. 3104–3112.

**Trudgill P.** (1972). Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in Society* **1**(2), 179–195.

**Vaswani A.**, **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A. N.**, **Kaiser L. and Polosukhin I.** (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30. Long Beach, CA: Curran Associates, Inc.

**Viera A. and Garrett J.** (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine* **37**(5), 360–363.

**Wallentin M.** (2020). Gender differences in language are small but matter for disorders. In *Sex Differences in Neurology and Psychiatry*, Handbook of Clinical Neurology, vol. 175. New York: Elsevier, pp. 81–102.

**Yamshchikov I. P.**, **Shibaev V.**, **Khlebnikov N. and Tikhonov A.** (2021). Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(16), 14213–14220. AAI Technical Track on Speech and Natural Language Processing III.

**Yang Z.**, **Hu Z.**, **Dyer C.**, **Xing E. P. and Berg-Kirkpatrick T.** (2018). Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, vol. 31. Long Beach, CA: Curran Associates, Inc.

**Yildiz T.** (2019). A comparative study of author gender identification. *Turkish Journal of Electrical Engineering & Computer Sciences* **27**, 1052–1064.

**Zare P.** (2013). Exploring reading strategy use and reading comprehension success among EFL learners. *World Applied Sciences Journal* **22**, 1566–1571.

**Zhang X.**, **Zhao J. and LeCun Y.** (2015). Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15*. Cambridge, MA: MIT Press, pp. 649–657.

**Zhou C.**, **Chen L.**, **Liu J.**, **Xiao X.**, **Su J.**, **Guo S. and Wu H.** (2020). Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online. Association for Computational Linguistics, pp. 7135–7144.

**Cite this article:** Khanmohammadi R and Mirroshandel SA (2024). PGST: A Persian gender style transfer method. *Natural Language Engineering* **30**, 1065–1090. https://doi.org/10.1017/S1351324923000426