A Survey on Text Style Transfer

Siddharth Raybharam NQR5356@PSU.EDU
Pennsylvania State University

Abstract

Style is an essential component of a sentence. People have a variety of ways of expressing themselves; nevertheless, they adapt the manner in which they talk and write depending on the social setting, the audience, the interlocutor, or the level of formality of the event. Text style transfer is the process of modifying and/or changing the stylistic manner in which a sentence is written while maintaining the meaning of the sentence that was originally written. This process is defined as the task of "text style transfer." Natural language generation uses text style transfer to regulate traits like politeness, emotion, humor, and others. Deep neural models have revived its popularity in natural language processing.

In this paper, we give a comprehensive analysis and assessment of the text style transfer approaches that use deep learning. We implement[1] 3 different models from the studies we have chosen to analyze the performance. We highlight the technological developments in deep neural networks that have been the primary impetus behind recent achievements in the fields of natural language comprehension and generation. This paper is structured around two phases in the text style, namely, style classification and style transfer. The discussion focuses on the similarities and differences among the proposed solutions, as well as the challenges and opportunities that are anticipated to guide and stimulate future research in the field.

Keywords: Text Style Transfer, Deep Learning, Natural Language Processing, Natural Language Generation, Neural Networks

1. Introduction

Culture, individual characteristics, and social context have been attributed to linguistic variations in spoken and written language [Eckert and Rickford, 2001, Coupland, 2007]. An individual's gender, age, education level, personality, and emotional state are some of the personal characteristics that are represented in their writing style. Different people have different ways of expressing themselves, and these characteristics are mirrored in their writing style [Labov, 1972]. However, style evolves over time [Eckert and Rickford, 2001] and we adapt to a given social context, audience, communicator, and/or occasion formality [Kiesling and Schilling-Estes, 1998]. While direct mapping of sociolinguistic categories is not always possible, natural language understanding and generation research has classified stylistic properties along multiple dimensions.

Text style transfer refers to the process of changing the style of a sentence by rewriting the original sentence in a new style while retaining its semantic content. The range of language types is mirrored in the field's research interests. Some scholars defined style transfer as the ability to change the emotional content of a written document, while others associated it with formality or politeness. Changing the sentiment polarity of a sentence can alter the meaning of a text or transform the message it conveys, but the ability to change

1. Implementation available at https://github.com/maitreya2954/comp597-nlp

the emotional content in a written text should be viewed more in terms of adjusting the tone of a message to be more appropriate, emphatic, and less severe or offensive to the audience or conversational partner. Other researchers have focused their efforts on developing a far more sound understanding of style as a genre, or the language style of a person or a certain social group.

In present and future intelligent interaction systems that comprehend, process, or generate speech or text, language style should be given specific care. When used on discussion forums and comment-based communities, automatic text style adjustment may become even more crucial for users to develop their communication skills (e.g., learning to write formal messages, being more polite), as well as toning down negative sentiment and neutralizing offensiveness.

Over the course of a decade or two, the work on the subject evolved from a few articles to an active research area. The majority of text style transfer approaches are based on deep neural networks. The research that address the automatic adjustment of a written text's style are crucial to this review. We begin by introducing numerous text styles that have been the subject of text style transfer during the last decade.

The paper is organized as follows. After the introductory section, Section 2 provides a description of various text styles that have been in the focus of the selected research papers along with the breif explanation of style transfer. Section 3 provides more details on the initial attempts to separate style from the context using statistical methods. Later, in Section 4 the paper gives a simple overview of Recurrent Neural Networks which are now the current state of text style transfer. Finally, we give more details about the recent research work in the domain of Text Style Transfer over the past few years in the Section 5. The discussion of a set of measures, which have been proposed as meaningful criteria for evaluating style transfer models, is also presented. Section 6 concludes the paper.

2. Text Style and Transfer

2.1 Text style

Individual, societal, and situational factors affect language nuance. Emerging research on automatic style transfer of written text agrees that style is an important aspect of a sentence, signaled by word choice [Argamon and Koppel, 2010]. We introduce text styles studied in automatic style transfer research. Individual style, and text formality, politeness, offensiveness, and sentiment are briefly described.

2.1.1 Personal Style

Words indicate personality, gender, and age. Several studies imply that language usage differs between demographic groups and that specific terms can be used to identify the author's gender and/or age with 80% accuracy [Koppel et al., 2001]. Female users use more emoticons and terms with positive emotional connotations [Rao et al., 2010]. However, the study shows that younger people use chat-specific e-language and refer to themselves more often, whereas older people use more complex sectences and include more links and hashtags [Nguyen et al., 2013].

2.1.2 Formality

Language style is typically linked to register, or formality, of a text. Formal language is not defined, yet the difference between formal and informal language is clear. Academic writings are more formal than social media. Longer and passive-voiced texts are more formal. Detachment, precision, objectivity, rigidity, and increased cognitive burden characterize formal writing [Sheikha and Inkpen, 2010] . On the other hand, short words, contractions, and abbreviations are informal [Sheikha and Inkpen, 2010]. Informal writing is more subjective, less accurate, less informative, and lighter in structure. Automatic formality detection study considers slang, grammatical errors, social distance, and shared knowledge between the writer and the audience. Writing aides can automatically formalize a text.

2.1.3 Politeness

The social distance between the writer and the listener influences the politeness of the language. The amount of politeness is vital for keeping a positive face in social interactions with others, and it has a substantial impact on the whole communication experience. Polite and impolite are on opposing ends of the spectrum, yet varying levels of politeness may be employed. Systems for automated politeness adjustment might protect online writing, particularly when someone (unintentionally) produces an unfriendly text that will be received and viewed by others.

2.1.4 Offensiveness

The negative repercussions of malevolent online conduct such as hate speech, trolling, and the use of inappropriate language continue to be a recurring issue for practically every social media site. The public, governments, and institutions all want systems and interaction mediators that will automatically recognize, delete, and/or label posts containing objectionable language and hate speech.

Detecting offensive language is a broad research field that focuses on detecting whether a sentence is offensive or notc [Pavlopoulos et al., 2019], or determining the audience (group or person) that is targeted by a message. According to research, the use of particular terms may be associated with inflammatory language. Words like "killed," "fool," and "ignorant" are frequently associated with unpleasant language. Many social media and comment-based news communities applaud the potential advantage of a style transfer mechanism to neutralize harsh statements before they are posted.

2.1.5 Sentiment

Emotions affect human behavior, and language frequently reflects them. In many cases, reworking a line with toned-down negative emotions may be preferable. Predictive analytics uses sentiment polarity to determine if a sentence is favorable or negative. Online postings may be used to "sense the mood of a community", public opinion on events, news headline emotions, and political attitude. Sentiment polarity has been used to predict book sales, sales performance, product rankings based on user reviews, stock market predictions based on Twitter moods [Bollen et al., 2011], website popularity, and more.

Task	Input Text	Output Text
Sentiment style transfer	*Great food, but horrible staff and very very rude workers! (negative)*	*Great food, awesome staff, very personable and very efficient atmosphere! (positive)*
Politeness transfer	*Send me the data (non-polite)*	*Could you please send me the data (polite)*
Formality transfer	*Gotta see both sides of the story. (informal)*	*You have to consider both sides of the story. (formal)*
Transferring offensive to non-offensive text	*I hope they pay out the ***, fraudulent or no. (offensive)*	*I hope they pay out the state, fraudulent or no. (non-offensive)*
Personal style transfer	*My lord, the queen would speak with you, and presently. (shakespearean english)*	*My lord, the queen wants to speak with you right away. (contemporary english)*

Table 1: Illustrative examples of selected style transfer tasks.

2.2 Transfer

Text style transfer refers to the process of rewriting a sentence in a new style, which involves generating a new (output) sentence with the same explicit meaning as the original (input sentence) but deviating stylistically from the original. Style transfer is used to change, modify, or adapt the way a sentence is written. Table 1 provides instructive examples for each of the style transfer tasks studied in the literature.

3. Style Classification

Due to lack of advances in deep learning neural networks, previous techniques to processing text style were primarily statistical study of language semantics. These statistical methods were used to categorize the documents into different styles. The papers included for this review are concerned with the concept of formality in language and its applications in natural language processing.

Formality is a complex and multidimensional concept that can vary depending on context and audience. Lahiri et al. [Lahiri et al., 2011] concentrate on one component of formality, namely how closely a sentence adheres to normal grammar and vocabulary rules. They build four datasets, each including 100 documents, from blog postings, news items, academic papers, and internet threads. The authors discovered that the F-score drops across the datasets in the sequence of academic papers, news articles, blog posts, and online threads using the F-score distribution on human annotated sentences and applying the sentence level F-score distribution on each of these datasets. F-score conveys profound formality, which is an essential discovery. The document's *deep formality* grows as additional context is provided. However, adding context necessitates the addition of new words, which makes reading more challenging. Using F-score distribution on an annotated dataset, the

authors were able to link F-score with human perception of sentence-level formality, and suggested that some correlation between F-score and readability is expected.

Sheikha et al. [Sheikha and Inkpen, 2010] go further describing the contrasts between formal and informal texts. Using 1000 texts from several corpora, the authors were able to extract elements that characterize formal and informal writings, hypothesizing to be a good signal to discern between both styles. The characteristics include a list of formal and informal terms, pronouns, contractions, abbreviations, voice, phrasal verbs, and so on. Using these attributes, the authors tested decision trees, Support Vector Machines (SVM), and the Naive Bayes Algorithm and discovered that these classifiers could predict the classes of new texts with high accuracy of 98%.

Ashok et al. [Ashok et al., 2013] use natural language processing techniques to analyze the writing style of a large corpus of novels and then attempt to predict their commercial success based on various stylistic features. The authors use a dataset of over 100,000 novels and their sales rankings to investigate the relationship between writing style and commercial success. They extract various stylistic features such as sentence length, word length, frequency of certain words and phrases, and syntactic complexity. They then use machine learning algorithms such as random forest and support vector machines to predict the commercial success of novels based on these stylistic features. The results of their experiments show that there is indeed a significant relationship between writing style and commercial success. They find that novels with shorter sentences, simpler words, and less syntactic complexity tend to be more commercially successful. Additionally, they find that certain genres such as romance and mystery tend to have more predictive stylistic features than others.

Preotiuc-Pietro et al., [Preotiuc-Pietro et al., 2016] examine whether there are differences in the way males and females, or younger and older people, write text based on three user attributes: gender, occupational class, and age. They use the Pearson coefficient to determine the level of correlation between the various paraphrases. The authors discovered that the scores are most correlated between users under the age of 25 and users from low occupational classes, as younger users are more likely to have a job with a lower skill level. Intriguingly, the next highest correlation is between low occupational class and female gender. In addition, to predict user attributes, the authors employ the Naïve Bayes Classifier and examine ROC AUC and correlation. Age was the feature that showed the highest paraphrase differences in the analysis. Incorporating paraphrase features into gender prediction yields similar outcomes. In the case of occupational class predictions, adding these features actually reduces predictive performance. Age is also the easiest to classify in terms of user trait prediction performance (.901 ROC AUC), followed by occupational class (.870 ROC AUC) and gender (.784 ROC AUC). The authors demonstrated that there are significant changes at the phrase choice level that are both predictive of user traits and intuitive to human annotators.

Lahiri et al., Sheikha et al., Ashok et al. and Preotiuc-Pietro et al. demonstrate the diverse applications of formality and style analysis in natural language processing. By analyzing the linguistic features of text, researchers can develop methods for automatically classifying documents, predicting the success of novels, judging the formality of individual sentences and even predicting user attributes. These techniques have the potential to be

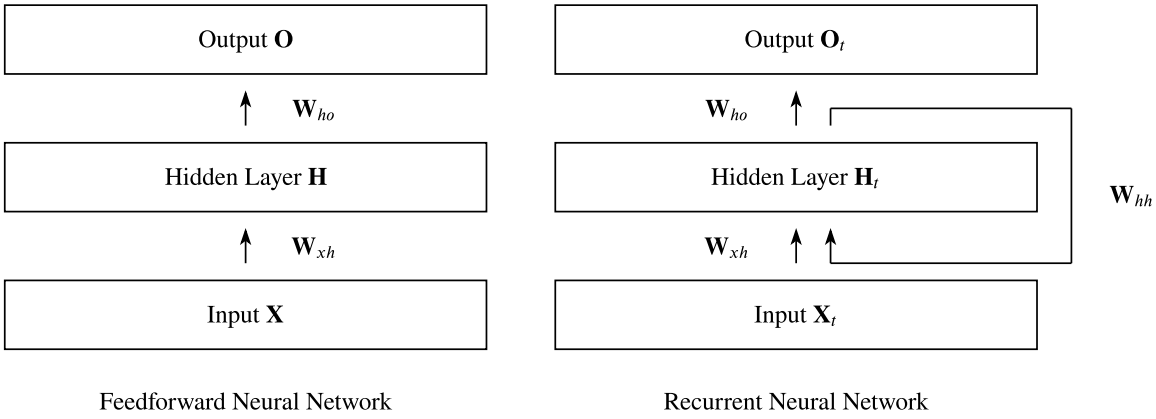

Figure 1: Visualisation of differences between Feedfoward NNs und Recurrent NNs [Schmidt, 2019]

used in a wide range of fields, from information retrieval and document management to literary analysis and the publishing industry.

4. RNN

In next section, we will study the use of machine learning and deep learning techniques in order to transfer or learn the style of the text. The majority of the research articles that are utilized for machine translation make use of some form of Recurrent Neural Networks (RNNs) [Schmidt, 2019].

RNNs are well suited for machine translation because machine translation involves analyzing words in one language and creating words in another language that convey the original meaning. RNNs can capture long-term dependencies and process sequences of any length, making them ideal for this purpose. In specific the “Encoder-decoder RNNs” are employed in machine translation. Encoder RNNs convert sentences into fixed-length vectors. The decoder RNN uses this vector, called the ”context vector,” to create the output sentence word by word. The decoder RNN generates each word in the output sentence using the context vector and relevant input sentence information.

The manner in which information is transported through the network distinguishes Recurrent Neural Networks from Feedforward Neural Networks, commonly known as Multi-Layer Perceptrons (MLPs). While Feedforward Networks send data through the network without using cycles, RNNs use cycles and send data back to themselves. This allows them to enhance the capability of Feedforward Networks to include prior inputs $\mathbf{X}_t$. This difference is visualised on a high level in Figure 1. Note, that here the option of having multiple hidden layers is aggregated to one Hidden Layer block $\mathbf{H}$. This block can obviously be extended to multiple hidden layers.

We may use the mathematical notation proposed in [Zhang et al., 2021] to explain the process of transmitting information from the previous iteration to the hidden layer. For this, we designate the hidden state and the input at each step t as $\mathbf{H}_t \in \mathbb{R}^{n \times h}$ and $\mathbf{X}_t \in \mathbb{R}^{n \times d}$

where n is number of samples, d is the number of inputs of each sample and h is the number of hidden units. Further, we use a weight matrix $\mathbf{W}_{xh} \in \mathbb{R}^{d \times h}$, hidden-state-to-hidden-state matrix $\mathbf{W}_{hh} \in \mathbb{R}^{h \times h}$ and a bias parameter $\mathbf{b}_h \in \mathbb{R}^{1 \times h}$. Finally, all of these variables are sent to an activation function ϕ, which is typically a logistic sigmoid or tanh function, to prepare the gradients for use in backpropagation. When all of these notations are combined, we get Equation 1 as the hidden variable and Equation 2 as the output variable.

$$\mathbf{H}_t = \phi_h \left(\mathbf{X}_t \mathbf{W}_{xh} + \mathbf{H}_{t-1} \mathbf{W}_{hh} + \mathbf{b}_h \right) \tag{1}$$

$$\mathbf{O}_t = \phi_o \left(\mathbf{H}_t \mathbf{W}_{ho} + \mathbf{b}_o \right) \tag{2}$$

5. Style Transfer

Cho et al. [Cho et al., 2014] use RNN to learn phrase representation to capture meaning of entire phrases instead of individual words by encoding a variable length input sequence and then decode the output sequences. The authors also propose a novel hidden unit which is motivated by LSTM unit but is much simpler to compute and implement. This proposed RNN is trained on a large bilingual corpora including Europarl (61M words), news commentary (5.5M) and UN (421M) to learn the translation probability of an english phrase to a corresponding french phrase. This model is then used as part of phrase-based statistical machine translation (SMT) framework [Koehn, 2005, Marcu and Wong, 2002] by scoring each phrase pair in the phrase table. The authors observe that using CSLM [Schwenk, 2007] and word penality where we penalizes the number of unknown words to neural networks along with RNN has the highest performance in terms of BLEU scores [Papineni et al., 2002] among different configurations that were tested to translate phrases from English to French.

Gan et al. [Gan et al., 2017] propose a novel framework named StyleNet to ad- dress the task of generating attractive captions for images and videos with different styles. They devise a novel model component, named factored LSTM, which automatically distills the style factors in the monolingual text corpus. Then at runtime, they explicitly control the style in the caption generation process so as to produce attractive visual captions with the desired style. A factual image/video-caption paired data, and a stylized monolingual text data (e.g., romantic and humorous sentences) are leveraged in order to achieve the goal. StyleNet has a generator and style encoder. Style encoders generate style embeddings from sentences. This embedding conditions the generator. The caption generator takes a noise vector and style embedding as input. A discriminator separates genuine and produced captions after generation. Figure 2 illustrates the learning of Style using the text corpora for different styles. The model is same all different styles except for the styling factor which is trained using mutli-task training on a particular style text corpora.

StyleNet training uses GAN's adversarial loss and CVAE's reconstruction loss. StyleNet was tested on the COCO dataset, which contains many images with captions. StyleNet captions were compared to vanilla GAN, vanilla CVAE, and a combination of GAN and CVAE without the style encoder. StyleNet outperforms baseline techniques in quantitative and qualitative criteria, the data show.

Prabhumoye et al. [Prabhumoye et al., 2018b] propose style transfer of text using back-translation. Back-translation is used to learn two inverse mappings: one from the style of a

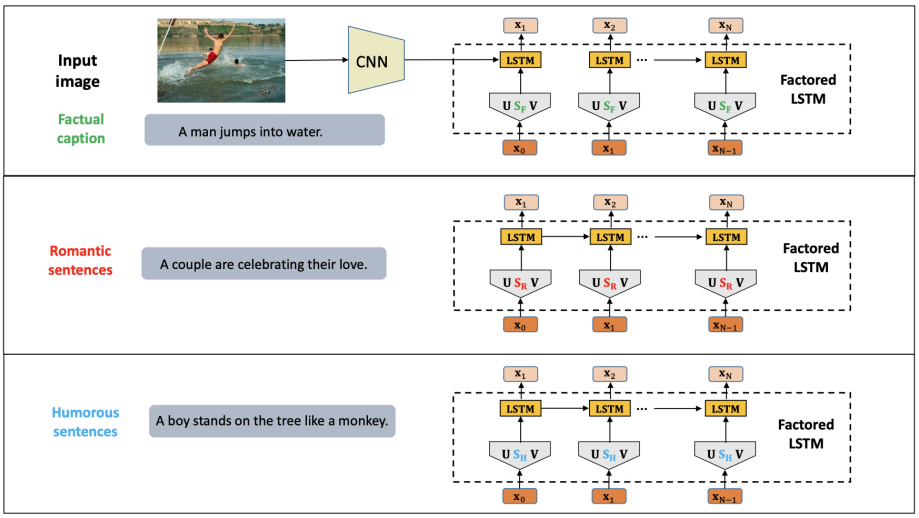

Figure 2: The framework of StyleNet.

target sentence to the content of a source sentence, and the other from the content to the original source phrase. By matching the source sentence's content, these inverse mappings increase style transfer. The authors first transfer a sentence to one pivot language and then utilize the pivot language's encoding to train the generative models that correspond to the two styles. They also employ feedback from a pre-trained classifier to direct the generators toward the desired style. This model is called Back-Translated Style Transfer (BST).

Later in another study [Prabhumoye et al., 2018a] in 2018, Prabhumoye et al. propose an extension to BST to improve the its state-of-the-art performance. They also use feedback mechanism is employed to iteratively refine the style transfer output by allowing users to provide feedback on the generated sentences. The authors analyze the performance of the improved model, MBST+F, along with baseline and BST model on gender transfer, political slant transfer and sentiment modification tasks and observes that the MBST+F performs much better compared to other models. Table 2 along with Figure 3 shows that the results of this model with different prompts during evaluation.

Dataset	**Model**	**Prompt**	**Accuracy (%)**
sst-2	roberta-large	AgentMediaGradeOfficials Grade	94.2
sst-5	roberta-large	iciticititableually immediately	45.2
agnews	roberta-large	Alert Blog Dialogue Diary Accountability	82.0
dbpedia	roberta-large	CommonExamplesSenate Similar comparable	86.1
subj	roberta-large	BufferActionDialogDialog downright	84.6
yahoo	roberta-large	AlertSource mentioning Besidesadays	49.7
trec	roberta-large	DonaldTrump	66.8
Yelp[2]	roberta-large	Absolutely	90.4

Table 2: Comparison of accuracy for different prompts and different datasets.

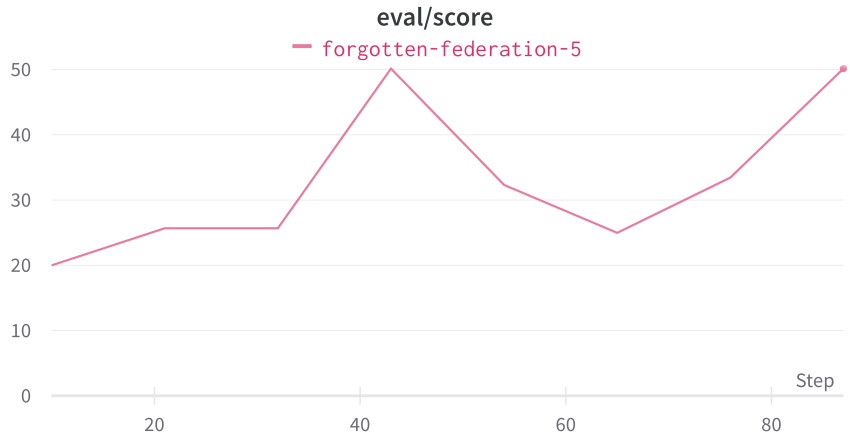

Figure 3: The evaluation accuracy of the model over 200 epochs.

Cao et al. [Cao et al., 2020] reimplement five state-of-the-art models from prior text simplification (TS) and style transfer (ST) studies on both MSD and SimpWiki datasets. The models used are OpenNMT+PT [Klein et al., 2017], UNTS [Sheang and Saggion, 2021], ControlledGen [Hu et al., 2018], DeleteAndRetrieve [Li et al., 2018] and StyleTransformer [Dai et al., 2019]. The results demonstrate that the expertise style transfer task is difficult and requires more advanced methodologies than typical style transfer tasks. The basic model outperforms other methods, although it can be improved. The authors then perform a human evaluation research to examine the quality of the generated sentences and discover that the transferred sentences are more layman-like while keeping the content of the expert-written words.

Kashyap et al. [Kashyap et al., 2022] improve the unsupervised attribute transfer by enforcing invariances via explicit constraints. They notice the lack of mechanisms in current methods in text attribute transfer to enforce such constraints between source and the transferred sentence. The authors propose a seq2seq model to encode and decode the text sequences using Adversarially Regularized Auto-encoder (ARAE) [Zhao et al., 2018] which are the auto-encoder variant of Generative Adversarial Networks (GAN). ARAE's have been widely used in text generation and text modeling. The authors test their approach on sentiment, politeness, and formality style transfer tasks. Their model outperforms baseline methods in content preservation and style transfer. The authors further analyze their model's learnt representations, showing that it learns a style-agnostic representation of input sentences independent of the target sentence's style. This analysis illuminates their technique's mechanics and its versatility.

We run the model for 1000 epochs to reduce the training time of the model and compared the results with those of the authors who have run the model for 12000 epochs in the Table 3. We can see that the implementation scores are comparatively lower than the authors' run scores. However, the model is really good in detecting the style are transferring them to the texts. Figure 4 indicates the improvement in the accuracy of the model over 1000 epochs.

2. Own implementation which was run for 200 epochs

	Implementation	Authors
Content	74.5	33.4
Style	94.3	96.0
Fluency	89.1	94.4
Joint	62.2	28.6
GM	85.5	67.1
BLEU	25.7	8.7
BERTScore	60.1	32.4
PPL	33.4	38.8

Table 3: Comparison of different scores of styles

Figure 4: The evaluation accuracy of the model over 1000 epochs.

Laugier et al. [Laugier et al., 2021] want to reduce the negative effects of toxic language in online groups by rephrasing toxic communications in a more respectful and civil way. The suggested method is based on a language transformer model that has already been trained and improved using a large dataset of toxic and civil language pairings. The writers use a contrastive learning goal to get the model to come up with rephrases that are both polite and have the same meaning as the offensive original message. The results of the experiments show that the proposed method is more natural and polite than some other ways of rephrasing that are used today. The authors also do an evaluation study on people, which shows that the reworded mean words are seen as more polite and respectful. Overall, the suggested method is a promising way to reduce the negative effects of toxic language in online communities by automatically rephrasing toxic messages in respectful and civil language. This could make the Internet a more welcoming place for everyone. The results showed that the proposed model achieved state-of-the-art performance in terms of fluency and civility while maintaining competitive effectiveness in terms of toxicity reduction. Specifically, the proposed model achieved a 26.6% reduction in toxicity on the Civil Comments dataset while maintaining a fluency score of 3.56 out of 4, which is the highest

among the compared models. Additionally, the authors conducted a human evaluation, which showed that the proposed model generated more civil and polite rephrases compared to the baseline models.

6. Conclusion

The main reason for doing the study in this work was the recent progress made in text style transfer using deep learning. A systematic review and implementation of the research spanning across the previous decade shows the trends that seemed to hold true across studies as well as the differences and variations in how deep learning is used to spread styles. This review focuses on how encoder-decoder-based designs still rule the field, even though Generative Adversarial Networks have been used to move toward adversarial learning. Even though it seems like choosing one deep neural network over another has nothing to do with style, researchers always have to find a way to balance the trade-offs between the complexity of a model and the predicted performance gains added by extra parts (like a classifier or a discriminator). The review is based on the key stages of the style transfer process and the different research methods that have been used at each step. Transfer learning and multitask learning studies are the opportunities that could make further progress possible. Interpretability is a recurring challenge that is shared across respective fields – a better understanding of what stylistic indicators are captured and learned by neural models might elucidate the nature of stylistic variations in language.

References

Shlomo Argamon and Moshe Koppel. The rest of the story: Finding meaning in stylistic variation. In *The Structure of Style*, pages 79–112. Springer, 2010.

Vikas Ashok, S. Feng, and Y. Choi. Success with style: Using writing style to predict the success of novels. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1753–1764, 01 2013.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. Expertise style transfer: A new task towards better communication between experts and laymen. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1061–1071. Association for Computational Linguistics, 2020.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN Encoder–Decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014.

Nikolas Coupland. *Style: Language variation and identity*. Cambridge University Press, 2007.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1601. URL https://aclanthology.org/P19-1601.

Penelope Eckert and John R Rickford. *Style and sociolinguistic variation*. Cambridge University Press, 2001.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. StyleNet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text, 2018.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. So different yet so alike! constrained unsupervised text style transfer, 2022.

Scott F Kiesling and Natalie Schilling-Estes. Language style as identity construction: A footing and framing approach. 1998.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-4012.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

M Koppel, S Argamon, and AR Shimoni. Automatically determining the gender of a text's author. *Bar-Ilan University Technical Report BIU-TR-01-32*, 2001.

William Labov. *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press, 1972.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. Informality judgment at sentence level and experiments with formality score. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 446–457. Springer, 2011.

Leo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. Civil rephrases of toxic texts with self-supervised transformers, 2021.

Juncen Li, Robin Jia, He He, and Percy Liang. 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), pages 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1169. URL https://aclanthology.org/N18-1169.

Daniel Marcu and Daniel Wong. A phrase-based,joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118711. URL https://aclanthology.org/W02-1018.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. "How old do you think I am?" A study of language and age in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576, 2019.

Shrimai Prabhumoye, Yulia Tsvetkov, Alan W Black, and Ruslan Salakhutdinov. Style transfer through multilingual and feedback-based back-translation. *CoRR*, 2018a.

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

Daniel Preotiuc-Pietro, Wei Xu, and Lyle Ungar. Discovering user attribute stylistic differences via paraphrasing. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44, 2010.

Robin M. Schmidt. Recurrent neural networks (rnns): A gentle introduction and overview, 2019.

Holger Schwenk. Continuous space language models. *Computer Speech and Language*, 21 (3):492–518, 2007. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2006.09.003. URL https://www.sciencedirect.com/science/article/pii/S0885230806000325.

Kim Cheng Sheang and Horacio Saggion. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK, August

2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.inlg-1.38.

Fadi Abu Sheikha and Diana Inkpen. Automatic classification of documents by formality. In *Proceedings of the 6th international conference on natural language processing and knowledge engineering (nlpke-2010)*, pages 1–5. IEEE, 2010.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning. *CoRR*, abs/2106.11342, 2021. URL https://arxiv.org/abs/2106.11342.

Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders, 2018.