

A Review of Text Style Transfer using Deep Learning

Martina Toshevska and Sonja Gievska

Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University - Skopje,
North Macedonia

Email: {martina.toshevska, sonja.gievska}@finki.ukim.mk

Abstract—Style is an integral component of a sentence indicated by the choice of words a person makes. Different people have different ways of expressing themselves, however, they adjust their speaking and writing style to a social context, an audience, an interlocutor or the formality of an occasion. Text style transfer is defined as a task of adapting and/or changing the stylistic manner in which a sentence is written, while preserving the meaning of the original sentence.

A systematic review of text style transfer methodologies using deep learning is presented in this paper. We point out the technological advances in deep neural networks that have been the driving force behind current successes in the fields of natural language understanding and generation. The review is structured around two key stages in the text style transfer process, namely, representation learning and sentence generation in a new style. The discussion highlights the commonalities and differences between proposed solutions as well as challenges and opportunities that are expected to direct and foster further research in the field.

Impact Statement—Motivated by recent advancements in the field, we have carried out a systematic review of state-of-the-art research to highlight the trends, commonalities and differences across style transfer methodologies using deep learning. The discussion is organized around key stages of the process, namely, representation learning of style and content of a given sentence, and generation of the sentence in a new style. A comprehensive view of methodologies, available datasets and evaluation metrics is compiled to foster further research in the field.

Index Terms—Text Style Transfer, Deep Learning, Natural Language Processing, Natural Language Generation, Neural Networks

I. INTRODUCTION

Naturally occurring linguistic variations in spoken and written language have been contributed to culture, personal attributes and social context [1, 2]. The underlying factors contributing to linguistic variations in spoken language have been extensively studied in the field of variationist sociolinguistics. The adjustments of one’s individual style to match or shift away [3] from the style of the interlocutor, the audience or social context are prominent in the work of the American linguist, William Labov [4, 5, 6]. Different people have different ways of expressing themselves [4] and personal attributes, such as gender, age, education, personality, emotional state [7] are reflected in their writing style. However, style changes over time [1] and we adjust to a social context, an audience we address, a person we communicate with [8], and/or the formality of an occasion [3]. While direct mapping of sociolinguistics categories is not always possible, stylistic

properties have been classified along several dimensions in the research on natural language understanding and generation.

Adjusting the style of a sentence by rewriting the original sentence in a new style, while preserving its semantic content, is referred to as text style transfer. The diversity of linguistic styles is matched by the diversity in research interests in the field. Some researchers viewed style transfer as an ability to adjust the emotional content in a written text; others equated the concept with formality or politeness. Changing the sentiment polarity of a sentence might change the meaning of a text or transform the message it conveys, although the ability to change the emotional content in a written text should be viewed more along the lines of adjusting the tone of a message that is more appropriate, emphatic and less severe or offensive to the audience or the conversational partner. Other researchers have directed their efforts towards much more sound conceptualization of a style as a genre, or linguistic style of a person, or a particular social group.

Language style should be a special consideration in current and future intelligent interaction systems [9] that understand, process, or generate speech or text. Automatically adjusting the text style could help users improve their communication skills (e.g., being more polite, learning to write formal messages), and could become even more important, when employed in future prosocial interaction mediators on discussion platforms and comment-based communities (e.g., toning down negative sentiment, neutralizing offensiveness).

In a decade or so, the work on the topic expanded from a few articles to an active research area. Most of the methods for text style transfer are based on deep neural networks. The success of deep learning in other areas has provided fruitful directions to be followed. Inspired by the success of the encoder-decoder models in other fields, including machine translation (MT) [10, 11], text summarization [12] and dialogue generation [13], a number of style transfer models are built upon this end-to-end model of learning [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]. New advances directed toward adversarial learning have also inspired more recent works on text style transfer [32, 33, 34, 35].

Pivotal in this review are the studies that address the automatic adjustment of the style of a written text. At the onset of our paper, we introduce the reader to various text styles that have been in the focus of the selected research papers. We have compiled a list of publicly available datasets that we discuss in terms of their suitability for a particular style transfer task(s).

The evaluation of how successful a particular model is on the task of style transfer has two objectives: to measure how well the semantic content in the generated sentence was preserved and to assess the quality of rewriting the sentence in a new (target) style. Evaluation of the performance of style transfer models is of special importance for future research in the area.

The discussion of the specifics of the proposed approaches to style transfer is organized to allow readers to follow the advances in deep learning and their impact on style transfer tasks. The discussion follows the two key stages in the text style transfer process: 1) representation learning of the style and content of a given sentence and 2) generation of the sentence that has the same meaning as the input sentence, but is expressed in a different style. Auxiliary elements, such as style embeddings [14, 15, 17, 18, 19, 20, 23, 29, 31, 32, 34, 35], style classifiers [16, 17] and/or adversarial discriminators [32, 33] are also discussed. We discuss the critical stage of the process, the output sentence generation by categorizing the approaches among three groups. Namely, models that use a simple approach to generation by reconstructing the input sentence, models that incorporate additional style classifier in their encoder-decoder architectures, and models that adopt adversarial learning.

The paper is organized as follows. After the introductory section, Section II provides a description of various text styles that have been in the focus of the selected research papers. Section III gives a formalization of text style transfer and discusses the publicly available datasets suitable for the task at hand. The discussion of a set of measures, which have been proposed as meaningful criteria for evaluating style transfer models, is also presented. Beginning with a brief introduction of several deep neural networks in Section IV, the discussion of state-of-the-art style transfer methodologies using deep learning is presented in Section V. Section VI reflects on the challenges style transfer faces and casts light on potential research directions that are expected to further advance the field. Section VII concludes the paper.

II. TEXT STYLE

The nuance and subtlety of language variations are functions of individual, social as well as situational differences. Emerging research on automatic style transfer of written text converges toward a common view that style is an integral part of a sentence, indicated by the choices of words a person makes [36]. We provide an introduction of various text styles that have been in the focus of the research on the automatic style transfer. In particular, a short description of the following linguistic styles is given: individual style, genre, as well as the formality, politeness, offensiveness, and sentiment that is conveyed by a given text.

A. Personal Style

The words people use reveal a lot about themselves, such as their personality, gender, or age [7]. There are several studies examining language variations across different gender and age groups found in formal texts [37], social media [38, 39], and blog posts [40, 41]. The findings suggest that differences

in language usage between various demographic groups do exist and the identification of the author's gender and/or age could be done with an accuracy of 80% on the basis of usage of specific words [42]. For instance, female users tend to use more emoticons [43] and choose words with positive emotional connotation [44]. On the other hand, the study of online language highlights the differences between various age groups - younger people use chat-specific e-language and refer to themselves more frequently, while older people use more complex sentences and include more links and hashtags [45].

These findings could be fruitfully applied in human-computer interaction [9] as important interaction features that develop user trust and satisfaction. A key challenge in designing believable virtual assistants is endowing them with dialog capabilities that are not only responsive to the user's need, but have a style of their own that matches the user's language style.

Shakespearean writing style¹ has been recognized as a specific writing style. An interesting research task of rewriting sentences in Shakespearean style have been reported in [46] that could be potentially used for edutainment purposes. Research on generating image caption used a rather unorthodox approach to generating caption in a style that was learned from romance novels and Taylor Swift's song lyrics² [47, 48].

B. Formality

Language style is very often associated with register i.e., formality of a given text. There is no unified definition of what formal language is and yet, the distinction between the language used in formal and informal settings is well recognized. For example, the language in academic papers is considered more formal than the language used in social media. Longer texts as well as texts containing passive voice tend to be perceived as more formal [49]. The formal style of writing is usually characterized by detachment, precision, objectivity, rigidity, and higher cognitive load [50]. On the other side, texts that contain short words, contractions, and abbreviations are considered informal [49]. The informal style is more subjective, less accurate, less informative, and with a much lighter form [50]. Indicators of formality considered in the research on automatic formality detection include the use of slang and grammatically incorrect words [51], social distance, and shared knowledge between the writer and the audience [52]. Automatically improving the level of formality of a written text is a useful feature incorporated in writing assistants [53].

C. Politeness

The politeness of the language we use is affected by the social distance between the writer and the audience [54, 55, 56]. The level of politeness is important for "maintaining a positive face" in social interaction with others [57] and it plays a significant role in the overall experience of communication [58].

¹https://en.wikipedia.org/wiki/Shakespeare's_writing_style, last visited: 09.03.2021

²<https://medium.com/@samim/generating-stories-about-images-d163ba41e4ed>, last visited: 01.04.2021

Polite and impolite are located on the opposite sides of the spectrum, although different levels of politeness might be used. A study presented in [59], shows that high frustration is correlated with a writing style that is less polite and less formal. Systems for automatic adjustment of politeness could safeguard online writing, especially in a situation when someone (unintentionally) writes an impolite text that will be received and read by others.

D. Offensiveness

The damaging consequences of malicious online behavior in the form of hate speech, trolling, and use of offensive language remain a recurrent problem for almost any social media platform. Devising systems and establishing interaction mediators that will automatically identify, remove, and/or label posts with offensive language and hate speech is demanded by public, governments, and institutions.

Detecting offensive language is a widespread research area that focuses on determining whether a sentence is offensive or not [60], or determining the audience that is targeted by a message (group or individual) [61]. Studies show that usage of specific words might correlate with offensive language. For example, words, such as "killed", "fool", "ignorant" are often correlated with offensive language [62]. The potential benefit of a style transfer system to neutralize offensive remarks before they are posted is welcomed by many social media and comment-based news communities.

E. Genre

Genre of a document is determined on the basis of some external criteria [63], such as purpose and target audience [64]. News, advertisements, and technical reports are some of the genres text documents are categorized into. Identifying the genre of a document could potentially improve Information Retrieval systems by search results that match or are relevant to a particular user's search. For example, when one intends to buy something, advertisements might be more relevant than scientific reports [65].

A document written in a style that matches the language style used by a particular group of people is expected to be more understandable by the target audience. For instance, medical reports are often difficult to understand by non-experts in the field. Automatically transforming a medical report into a document in layman terms, might improve its readability by a wider audience.

F. Sentiment

Emotions play a crucial role in human behavior [66] and one's emotional state is often reflected in one's spoken or written language. While emotional connotation carried by a sentence may not be a typical stylistic variation of language, rewriting a sentence with toned down negative emotions might be desired in many applications. Several categorical and dimensional models for emotions have been proposed [67, 68, 69, 70]. Detecting sentiment polarity of a text i.e. whether the overall sentiment of a particular text is

positive or negative [71, 72, 73, 74, 75, 76] have been used in predictive analytics. Being able to detect emotions expressed in online posts have been used to "sense the mood of a community" [77, 78, 79, 80], opinion of the public about specific events [81], the emotions embedded in news headlines [82], or political sentiment [83]. Sentiment polarity has been used for predicting the impact of users' reviews on book sales [84], sales performance prediction [85], ranking products based on user reviews [86], stock market prediction according to Twitter moods [87], website popularity prediction [88], etc.

III. STYLE TRANSFER FOR TEXT

A. Style Transfer Tasks

Text style transfer refers to the process of rewriting a sentence in a new style, which involves generating a new (output) sentence that has the same explicit meaning as the original (input sentence), while stylistically differing from the original one. Style transfer has been applied to adjust, modify or adapt the manner in which a sentence is written. The term style has been used rather broadly and encompasses properties, such as: *register (formality)*, *politeness*, *offensiveness*, *genre according to purpose*, *genre according to the target audience*, *sentiment* or the *individual style of the author or the social group they belong to*. Table I presents illustrative examples for each of the style transfer tasks that have been given attention in research literature.

The objective of each style transfer task is to adjust the style of a sentence with respect to particular style properties. For example, adjusting the emotions conveyed in a sentence is referred to as *sentiment style transfer*. Adjusting the politeness or the formality of a sentence is associated with *politeness* and *formality transfer*, respectively. Removing the offensiveness and substituting it with a neutral style has been the objective in the task of *transferring offensive to non-offensive text*. Rewriting a text that stylistically adheres to the personal writing style of an author (e.g., Shakespeare writing style, Taylor Swift's lyrics) or a social group (e.g., masculine vs feminine language style, democrat vs republican language) is referred to as *personal style transfer*. *Genre style transfer* could be related to a purpose (i.e. advertisement or news articles are written in a different style) or the intended audience (e.g., content written in expert language vs layman language).

B. Datasets

A number of datasets have been used in the research on style transfer of text. The list of publicly available datasets targeted by the research on style transfer offered in this paper is presented in Table II. Each dataset has been described with the following attributes: the year when a dataset has been published, whether the dataset is composed of parallel text sample pairs or not, type of text (e.g., emails, reviews, tweets, posts, documents), the number of text data samples, labels for the style used, as well as references to studies that have previously used the dataset.

A short description of the datasets, divided into parallel and non-parallel is presented below. The number of parallel datasets suitable for style transfer is limited since creating

Task	Input sentence (style 1)	Output sentence (style 2)
Sentiment style transfer	<i>Great food, but horrible staff and very very rude workers!</i> (negative)	<i>Great food, awesome staff, very personable and very efficient atmosphere!</i> (positive)
Politeness transfer	<i>Send me the data.</i> (non-polite)	<i>Could you please send me the data?</i> (polite)
Formality transfer	<i>Gotta see both sides of the story.</i> (informal)	<i>You have to consider both sides of the story.</i> (formal)
Transferring offensive to non-offensive text	<i>I hope they pay out the ***, fraudulent or no.</i> (offensive)	<i>I hope they pay out the state, fraudulent or no.</i> (non-offensive)
Personal style transfer (Shakespearean)	<i>My lord, the queen would speak with you, and presently.</i> (shakespearean english)	<i>My lord, the queen wants to speak with you right away.</i> (contemporary english)
Genre based on audience (expert/layman)	<i>Many cause dyspnea, pleuritic chest pain, or both.</i> (expert)	<i>The most common symptoms, regardless of the type of fluid in the pleural space or its cause, are shortness of breath and chest pain.</i> (layman)

TABLE I
ILLUSTRATIVE EXAMPLES OF SELECTED STYLE TRANSFER TASKS.

a large number of pairs of text (e.g., sentences, paragraphs, documents) containing sentences that express the same meaning in a different manner requires a lot of human work. In non-parallel datasets there are no paired data to learn from. The number of these datasets is larger because most of them are subsets or adapted versions of datasets that have been previously created for other tasks, such as sentiment analysis, author profiling, genre classification, etc.

1) *Parallel Datasets*: **Shakespeare**³ dataset [89, 90] contains 21,075 sentence pairs from 16 Shakespeare’s plays and their line-by-line paraphrases in contemporary English. A style transfer task on this dataset has been defined as a transformation of a sentence written in contemporary English into a sentence written in Shakespeare’s language style.

GYAFC⁴ dataset (Grammarly’s Yahoo Answers Formality Corpus) [53] is a parallel dataset of formal and informal sentence pairs. A subset of informal sentences is selected from the Yahoo Answers L6 corpus⁵. For each sentence, a formal version is written by people recruited through Amazon Mechanical Turk (AMT). The final dataset contains 112,975 pairs of informal-formal sentences and 111,266 pairs of formal-informal sentences.

Cheng et al. [18] have created a dataset containing informal and formal versions of 600,000 email messages from the **Enron corpus** [91]. The AMT annotators were asked to identify informal sentences in each email and rewrite them in a formal style.

Captions [92] is composed of 7,000 image captions that were classified as factual, romantic, or humorous. For each image, a caption in all three styles is created, making the dataset appropriate for a stylistic transformation of a sentence among the three alternate styles.

2) *Non-parallel Datasets*: **Yelp**⁶ dataset is a collection of 8.6 million business reviews that are classified as positive or negative according to their 5-star rating system, making the dataset suitable for sentiment style transfer. The dataset was often used to train systems for changing the polarity of a given text.

Gender [93] is a subset of Yelp reviews annotated with gen-

der labels (male and female) that were assigned by inferring the gender of the review’s author by his or her first name. This subset is suitable for rewriting text written by a female in a masculine writing style (and vice versa).

Amazon⁷ dataset [94] of 1 million product reviews, **SST**⁸ (Stanford Sentiment Treebank) [95] consisting of 9,613 movie reviews, and **IMDB** [96] dataset composed of 350,000 movie reviews have been labeled with sentiment polarity making them often used for sentiment transfer.

Paper-News Titles dataset [19] contains 200,000 titles categorized into two groups: titles of scientific articles and headlines of news articles. The news are collected from the UC Irvine Machine Learning Repository and the papers were compiled from publishing websites, such as ACM Digital Library, Springer, Nature, Science Direct, arXiv, and others.

Gigaword dataset [97, 98] is composed of 4 million news articles from seven news media publishers. The headlines of the news articles have been labeled according to their publisher.

Political slant [99] is a dataset composed of 540,000 comments on Facebook posts from 412 members of the United States Senate and House of Representatives. Every comment is categorized as either democratic or republican on the basis of the political affiliation of a member.

dos Santos et al. [20] have created two datasets, **Twitter** containing 2 million tweets and **Reddit** dataset of 7.5 million sentences. The datasets have been used in research on style transfer i.e. adjusting or removing the offensiveness in a sentence.

Madaan et al. [100] use the Enron corpus [91] to create **Politeness** dataset⁹ of 270,000 emails, labeled as polite or impolite. The potential use of this dataset would be to convert a neutral sentence into a polite sentence.

Cao et al. [101] have created the **Expertise** dataset¹⁰ based on the Mericks Manuals that is suitable for transfer of style between expert and layman medical language style. The dataset is composed of sentences from the domain of medical science: 130,349 sentences were written in medical expert style and

³<https://github.com/cocoxu/Shakespeare>, last visited: 28.08.2020

⁴<https://github.com/raosudha89/GYAFC-corpus>, last visited: 28.08.2020

⁵<https://webscope.sandbox.yahoo.com/catalog.php?datatype=l>, last visited: 28.08.2020

⁶<https://www.yelp.com/dataset>, last visited: 13.03.2020

⁷<http://jmcauley.ucsd.edu/data/amazon/>, last visited: 13.03.2020

⁸<https://nlp.stanford.edu/sentiment/treebank.html>, last visited: 27.08.2020

⁹<https://github.com/tag-and-generate/politeness-dataset>, last visited: 28.08.2020

¹⁰<https://srhthu.github.io/expertise-style-transfer/#disclaimer>, last visited: 26.02.2021

114,674 sentences in layman style. A small subset of the dataset contains aligned sentence pairs in both styles.

C. Evaluation of Automatic Style Transfer

Evaluations of text style transfer face the longstanding challenges in the field of natural language generation (NLG) [108]. In regard to text style transfer, the objective of the evaluation is two-fold: 1) to measure how well the meaning of the original sentence was preserved in the output (generated sentence) and 2) to evaluate the quality of the style. Ideally, the goal is to create a model that successfully modifies the style of a text, while its meaning is preserved.

A different set of metrics have been used for evaluating both aspects. The quality of content preservation is evaluated using evaluation metrics that measure the extent to which the generated sentence matches human output, which is used in other NLG tasks, including summarization [12], image captioning [109] and machine translation [110]. A new set of metrics, specifically tailored to measure the style strength, are proposed for measuring the quality of generating a text in the target style.

1) *Evaluation of the Quality of Semantic Content Preservation*: Despite the criticism of using metrics based on language modeling and similarity measures, a number of well-established metrics have been adopted for measuring the quality of text generation.

Word overlap based metrics **METEOR** [111] and **BLEU** [112], were introduced for the evaluation of machine translation, by computing a score that indicates the similarity between the system output and one or more human-written reference texts. **METEOR** [111] evaluates the generated sentence by aligning it to one or more reference sentences. Alignments are based on exact, stem, synonym, and paraphrase match between words and phrases. METEOR is calculated as a harmonic mean of unigram precision and recall, with recall being weighted higher. **BLEU** [112] measures how close a candidate sentence is to a reference sentence based on matches of n-grams of a sentence to a reference one. **NIST** [113] is a version of BLEU metric that values the less frequent n-grams more. **BERTScore** [114] computes the cosine similarity between contextualized BERT [115] word embeddings of the sentence being evaluated and a set of reference sentences.

ROUGE-L [116] is a recall-oriented metric established for the evaluation of text summarization that applies the concept of the Longest Common Subsequence (LCS). The intuition behind the LCS concept is that the longer the LCS between two sentences, the more similar they are. The score is 1 when the two sentences are equal, and 0 when there is nothing in common between them.

SARI [117] is a metric for text simplification that considers the number of additions and deletions. It measures the goodness of words that are added, deleted, and kept by the system. SARI first calculates precision and recall for each operation (addition, keep, and deletion). The final value is an average of these scores.

PINC [118] is a measure originally developed for evaluating paraphrasing. It evaluates how much a generated sentence

resembles a reference sentence i.e. how many n-grams differ between the sentences. The final score is the percentage of n-grams that appear in the generated sentence but not in the reference. The novelty of paraphrases is greater as the value increases.

2) *Quality of a Style*: To evaluate the quality of generating a sentence in a target (output) style, various researchers calculate accuracy with a pre-trained classifier [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 31, 32, 33, 34, 35]. A style quality is calculated as a percentage of generated sentences that were labeled with the target style by the classifier. Higher value indicates better style quality. Precision, recall, and F1-measure are also appropriate for the evaluation of the quality of a style.

IV. DEEP NEURAL NETWORKS FOR TEXT GENERATION

A. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) [119] are a class of deep learning networks designed for modeling sequential data. RNNs process the input sequence from beginning to end (forward direction). Bidirectional Recurrent Neural Networks (BiRNNs) [120] are composed of two unidirectional RNNs operating in both directions (forward and backward).

Long Short-Term Memory Networks (LSTM) [121] are a specific type of RNN, designed to learn long-range dependencies as well as to overcome the problem of vanishing and exploding gradients. Gated Recurrent Units (GRU) [122] have the same purpose as LSTM, but are known to be simpler and faster to train.

B. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) [123] are a class of deep neural networks developed for representing spatial features. Each neuron in CNNs is connected with a local region of neurons in the previous layer. The parameters are known as kernels that operate in two dimensions (2D convolution) over the spatial data thus producing two-dimensional feature maps of input. CNNs are suitable and are most commonly applied for image processing [124, 125]. Recently, CNNs have been explored for modeling sequential data [126, 127, 128, 129, 130, 131], where the convolutional kernel operates in one dimension (1D convolution).

C. Attention Mechanism

Attention is a deep learning technique inspired by human cognitive attention that was introduced by Bahdanau et al. [110] to improve machine translation. Attention is a technique, which computes a weighted sum of attention scores assigned to the elements of the input sequence that help the decoder to attend to certain elements of the input. There are various types of attention: dot-product attention, multiplicative attention, additive attention, self attention [132, 133, 134].

Self-attention is used for representation of a sequence while giving attention to relevant parts of the same sequence. Performing self-attention multiple times in parallel is defined as multi-head self-attention. Multi-head self-attention combines

Dataset Name	Year	Parallel (Y/N)	Type of Text Samples	Number of Samples	Labels for Style	References
Enron corpus [18, 91]	2004	Y	emails	600K	informal formal	[18]
Shakespeare [89, 90]	2012	Y	sentences	21K	shakespearean english contemporary english	[33]
Gigaword [97, 98]	2012	N	news articles	4M	seven publishers	[30]
SST [95]	2013	N	reviews	9.6K	positive negative	[31]
IMDB [96]	2014	N	reviews	16K	positive negative	[17], [31], [34], [102] [14], [15], [19]
Amazon [94]	2016	N	reviews	1M	positive negative	[22], [26], [27] [29], [103], [104]
Gender [93]	2016	N	reviews	*	female male	[15], [21], [24]
Captions [92]	2017	Y	image captions	7K	factual romantic humorous	[14], [15], [27]
GYAFC [53]	2018	Y	sentences	110K	informal formal	[16], [18], [28], [105]
Paper-News Titles [19]	2018	N	titles	200K	paper news	[19], [106]
Political slant [99]	2018	N	posts	540K	democratic republican	[15], [21], [24], [25]
Twitter [20]	2018	N	tweets	2M	offensive non-offensive	[20]
Reddit [20]	2018	N	sentences	7.5M	offensive non-offensive	[18], [20] [14], [15], [17], [21] [22], [23], [24], [25] [26], [27], [28], [29] [32], [33], [34], [35] [103], [105], [107] [102], [104], [106]
Yelp	2020	N	reviews	8.6M	positive negative neutral	
Politeness [100]	2020	N	emails	270K	polite	[100]
Expertise [101]	2020	N	documents	200K	expertise laymen	[101]

* The dataset is composed of reviews written by 432M users, however the number of reviews is not specified by the authors.

TABLE II
A LIST OF PUBLICLY AVAILABLE DATASETS FOR TEXT STYLE TRANSFER.

information from different representation subspaces. Transformer [134], a deep neural architecture proposed for language modeling by multi-head self-attention, allows significantly more parallelization than RNN. Various models have been built following the Transformer architecture: BERT [115], DistilBERT [135], RoBERTa [136], GPT [137], GPT-2 [138], GPT-3 [139], etc.

Pointer network [140, 141] is a mechanism for "pointing out" to relevant parts of the input. In a pointer network, attention is used as a pointer for selecting parts of the input sequence as members of the output sequence.

D. Encoder-decoder

Encoder-decoder (also referred to as sequence-to-sequence network) is a deep neural network architecture for text generation [142]. The encoder learns to generate a latent fixed-length vector representation of the input sentence. The decoder learns to generate an output sentence by decoding the fixed-length representation of the input sentence. The encoder and the decoder could be RNNs, CNNs, MLPs, attention-based networks, or a combination.

Autoencoder (AE) [143] and Variational Autoencoder (VAE) [144] are deep learning architectures intended for learn-

ing an internal representation of the input. As in the encoder-decoder architecture, the encoder is a neural network that produces fixed-length representation. The decoder learns to reconstruct the input sentence based on the encoded representation. VAE is a generative model that learns the distribution of the data with a stochastic variational and learning algorithm. AE and VAE could be viewed as a specific type of encoder-decoder architecture where the goal is to generate an encoded representation of the input data.

Encoder-decoder network has been applied in natural language generation for a number of tasks: machine translation [10, 11, 110, 145], text summarization [146], question answering [147], etc.

E. Generative Adversarial Networks

Generative Adversarial Network (GAN) [148] is a deep neural network architecture comprised of two networks – generator and discriminator. These two networks are trained simultaneously in a two-player minimax game. The generator network aims to learn the distribution of the training data and to generate samples from the learned distribution. The discriminator network determines whether a sample is

from the data distribution or from the model distribution, by maximizing the probability of assigning the correct label to samples from the training data as well as from the generated data. The objective of the generator is to generate samples that are indistinguishable from the training samples, by maximizing the opposite objective of the discriminator. GANs have been applied in many natural language generation tasks [149, 150, 151] including machine translation [152], text summarization [153] and question answering [154].

V. METHODS FOR STYLE TRANSFER

In this section, the discussion of deep learning (DL) models that were used in the research on text style transfer is presented. The vast majority of the models are built upon the encoder-decoder architecture [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 103, 104], while in another line of research, adversarial learning with GANs have been put forward [32, 33, 34, 35, 102, 105, 106, 107].

A general encoder-decoder-based architecture for style transfer is depicted in Figure 1. The encoder creates a latent representation of the input sentence i.e., encodes the input sentence. The decoder generates the output sentence conditioned on the latent representation, while the classifier determines the style label of the output sentence. The classifier is an optional component in style transfer models. If the style transfer model is based on GAN, the encoder is referred to as generator and the classifier component as a discriminator. Style embeddings have been introduced to assist in the encoding of input sentence and/or generation of output sentence. Style embedding could be added as input to the encoder [29, 31], decoder [14, 15, 18, 19, 23, 34, 103, 104], or both [17, 20, 32, 35].

A. Representation Learning

An encoder is applied to create a latent vector representation of the content of a sentence. It transforms a sentence while preserving its semantic and syntactic properties. The initial building block, the encoder in style transfer models is usually a type of RNN. Two types of RNNs, LSTM [23, 26, 31, 35, 102, 103, 107] and GRU [14, 19, 20, 22, 25, 28, 32, 34], have been employed to learn the representation of input sentences.

In the literature on text style transfer, two types of encoders can be identified: shared and private encoders. When a shared encoder is used, the parameters are shared across the sentences of the entire dataset, so the encoder learns the style characteristics of the entire dataset. A private encoder is used to learn style-specific characteristics since the parameters are shared only across sentences of a specific style. Most of the style transfer models have opted for a shared encoder. Zhang et al. [30] have introduced a system that uses both, private and shared encoders. Each sentence is passed through two GRU encoders, one private encoder for a particular style and one encoder shared across all styles. Zhao et al. [33] proposed decomposing each sentence into two latent representations by using two GRU encoders, one for style representation, and the other for creating content representation. In a model called StyIns [105], instead of learning style embeddings

from a single sentence, the generative flow techniques [155] have been used to learn the stylistic properties from a set of sentences sharing the same style i.e., style instances. Coupled with an attention-based decoder, StyIns model yielded higher style accuracy while preserving the content of the original sentence when evaluated on three style transfer tasks.

Motivated by the findings that architectures for machine translation preserve the semantic meaning of a sentence, but not its stylistic properties [156], Prabhumoye et al. [21, 24] have incorporated a machine translation-based model in the first stage of style transfer i.e., for representation learning of input sentences. While deep learning architectures for machine translation have reached state-of-the-art performances for many languages, their integration into deep learning pipelines for other tasks still face challenges.

Various models for text style transfer employ variants of Transformer architecture [16, 17, 18, 27, 29, 104], to benefit from the self-attention mechanisms when learning the representation of the input sentence.

An additional step in the process of style transfer, related to detection and removal of style markers from the input sentence, has been included in a number of studies [14, 15, 23, 29]. Style markers are words that have the most discriminative power for determining the style of a sentence. Models proposed by Sudhakar et al. [15] do not use an encoder to create a latent representation of the sentence. Instead, the input sentence is reduced by removing the style markers and then it is passed directly to the decoder.

1) *Detecting Style Markers*: In most models for style transfer, the entire sentence is fed into the model [16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35]. Li et al. [14], Sudhakar et al. [15], Zhang et al. [23] and Lee [29] argue that style transfer can be accomplished by changing a few style markers. Based on this idea, the input sentence is preprocessed in a way that style markers are removed from it. The preprocessed sentence is then fed into the model as an input sentence. It should be noted that detecting and removing style markers has been proposed and evaluated only on the style transfer task of sentiment modification.

Several approaches for detecting style markers have been proposed. Li et al. [14] have used n-gram salience measure for identifying style markers. It calculates the relative frequency of n-grams in sentences with a specific style. An n-gram is considered to be style marker if its salience is above a specific threshold. Zhang et al. [23] used attention weights [157] to detect style markers. A word is defined as a style marker, if its attention weight is greater than the average attention value. Sudhakar et al. [15] introduced an importance score for each token in the input sentence, based on attention scores of BERT [115] style classifier. Tokens with the highest importance score represent style markers. Lee [29] proposed to identify style markers by monitoring the change in probabilities of a style classifier. Important Score (IS) of a token is defined as a difference between the probability of the style conditioned on the entire sentence and the probability of the style conditioned on a sentence without the specific token. Token is a style marker if it has largest IS.

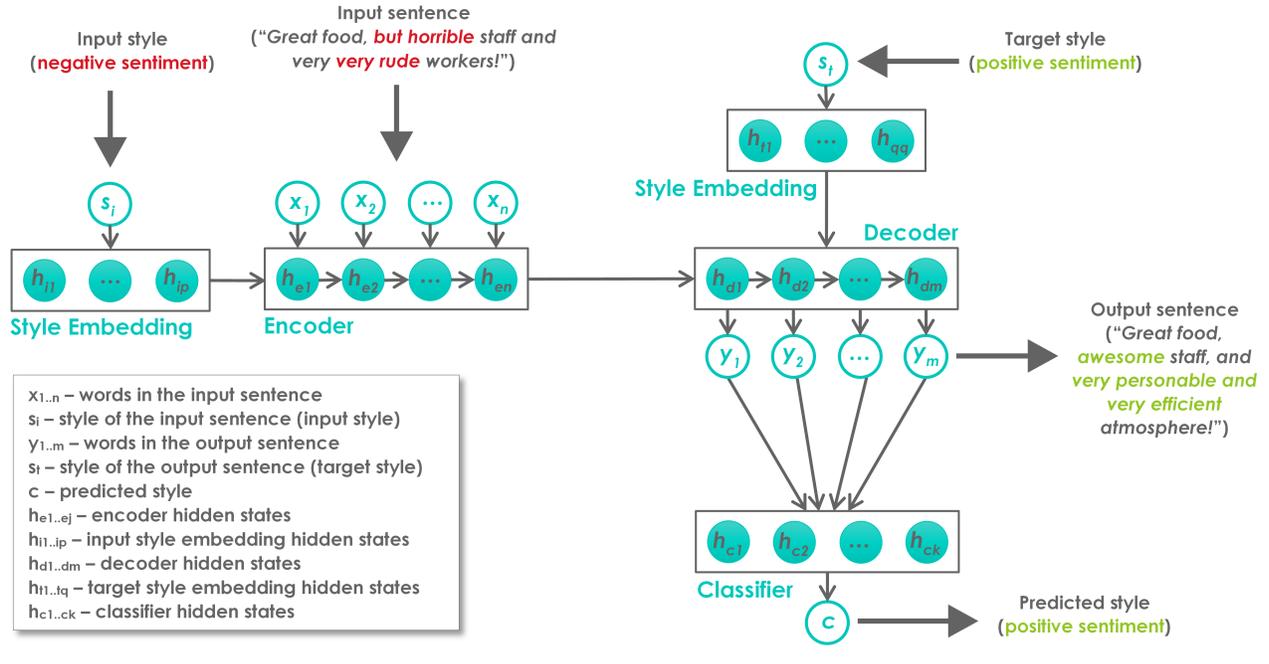


Fig. 1. General architecture of deep neural style transfer model.

B. Sentence Generation

A crucial component of deep style transfer models is the part that generates the output sentence based on the representation of the content of the original sentence and the corresponding styles. We start our discussion with two models for sentiment modification, proposed by Li et al. [14], that are often used as baseline methods other research is compared to. The two models are not based on deep learning, but they use rather simplistic methods of retrieval or word swapping using two corpora of sentences with positive and negative sentiment. The **RetrieveOnly** model simply outputs the retrieved sentence from the corpus that is most similar to the input sentence. A grammatically correct output sentence is expected, although the content of the original sentence might not be preserved.

The **TemplateBased** model removes the words identified as style markers from the input sentence and replaces them with the style markers from the retrieved corpus sentence that is most similar to the original. This is a naïve method of word swapping based on the assumption that original style markers could be replaced with words with the opposite sentiment if they appear in a similar context (retrieved sentence). It is not surprising that very often the generated output sentence appears to be grammatically incorrect.

The first deep learning models developed for text style transfer were inspired by sequence-to-sequence systems for machine translation [110] and paraphrasing [158, 159, 160]. An MT-based sequence-to-sequence model has been proposed for transforming text from modern English to Shakespearean English [46]. The output probability distribution is produced by a two-part decoder, consisting of an attention-based LSTM decoder and a pointer network. The pointer network was added to facilitate direct copying of the words from the input into the

output sequence. Pre-trained word embeddings using external sources, such as dictionaries have been used to mitigate the problem of a limited amount of parallel data. An interesting approach of harnessing rules for formality style transfer using pre-trained GPT-2 Transformers was incorporated in several models proposed by Wang et al. [161] to overcome the problem of small parallel datasets.

We group the models for style transfer into three groups according to the architectural blocks for generating the output sentence: *simple reconstruction models*, *models with style classifiers*, and *adversarial models*. Simple reconstruction models are trained to simply reconstruct the sentences based on the reconstruction loss (called self-reconstruction). Models in the second group, incorporate additional style classifier(s) to assist in the generation of the output sentences, while models in the adversarial group are built upon GAN architecture.

1) *Simple Reconstruction Models*: Most of the style transfer models generate the output sentence as a simple “reconstruction” of the content of the input sentence by maximizing the probability distribution of the next word in the output sequence conditioned on the latent representation of the words in the input sentence. Style transfer research that belongs to the group of simple reconstruction models are presented in Table III.

Two of the pioneering encoder-decoder models using simple reconstruction, **DeleteOnly** and **DeleteAndRetrieve** [14] have been proposed by the authors advocating style markers removal for the sentiment modification task. The decoder of the **DeleteOnly** model generates the output sentence based on the encodings of the input sentence and the target style (encoded by a separate style encoder). The style of the retrieved corpus sentence that is most similar to the input sentence is encoded in

the **DeleteAndRetrieve** model to condition the output sentence generation.

Sudhakar et al. [15] proposed two models, Blind Generative Style Transformer (**B-GST**) and Guided Generative Style Transformer (**G-GST**), that follow the same modeling approach as **DeleteOnly** and **DeleteAndRetrieve** [14], respectively. However, both models incorporate Transformer-based decoders.

Lample et al. [103] have suggested using a back-translation technique instead of adversarial training in their model **MultipleAttrTransfer**. By doing this, the generated output by the model is also used as a training input example fed into the encoder during “back-translation”. In terms of the objective function being optimized, a so-called cycle reconstruction is added to the original denoising auto-encoder. In addition, by using a temporal max-pooling layer on top of the encoder i.e. latent representation pooling, the decoder has better control over content preservation.

Shared-Private Encoder-Decoder (**SHAPED**) [30] is composed of multiple GRU encoders and multiple GRU decoders to learn both general and style-specific characteristics. One encoder and one decoder are shared across all styles in addition to the private encoders and decoders for each style. The outputs of both, private and shared encoder-decoders are concatenated and processed with a multi-layer feed-forward network to generate the output sentence. When the style of the input sentence is unknown, the sentence is fed into all private encoders. Their outputs are concatenated and fed into a style classifier that determines the style of the input sentence.

Zhang et al. [23] take on another approach to the task of sentiment modification in their Sentiment-Memory based auto-encoder (**SMAE**). A sentiment classifier with self-attention mechanism is utilized to separate sentiment from non-sentiment words, creating “sentiment memories” i.e. weighted matrices of positive and negative sentiment word embeddings. During decoding, the context of the input sentence is used to extract closely-related sentiment entries from the sentiment memory matrix to condition the output sentence generation

2) *Models with Style Classifier*: A large group of models incorporate a style classifier to facilitate the generation of the output sentence. The group of models shown in Table IV incorporate a style classifier in the process of generating the output sentence.

ControllableAttrTransfer [27] first embeds the input sentence with a Transformer encoder, and then generates the output sentence with a Transformer decoder. The encoded representation is additionally fed into a two-layer linear classifier to provide a direction for editing the latent representation, so that it conforms to the target style.

Back-translation for Style Transfer (**BST**) [21] is a back-translation based model that learns to rephrase the sentence by reducing the effects of the original style using English-to-French machine language translation model. A style classifier is used to identify the style of the latent representation of the back-translation model that later guides the generation of the output in style-specific generators. **BST** includes multiple BiLSTM style-specific decoders, one for each style. A style classifier is used to identify the style of the latent repre-

sentation of the back-translation model that later guides the generation of the output sentence by the style-specific decoders. Multi-lingual Back-translated Style Transfer (**M-BST**) and Multi-lingual Back-translated Style Transfer + Feedback (**M-BST+F**) [24] are extensions of **BST**. **M-BST** creates latent sentence representation with multilingual MT model, while **BST** exploits monolingual MT model. In **M-BST+F**, a feedback-based loss function was introduced to guide the decoder.

Neural Text Style Transfer (**NTST**) [20] and **StableStyle-Transformer** [29] utilize a CNN classifier to classify the style of the generated sentence. Unlike **BST**, these models are composed of a single decoder and therefore the inclusion of a style embedding as input to the decoder is needed to generate a sentence in the desired style.

StyleTransformer [17] model uses a discriminator network as another Transformer encoder to distinguish the styles of sentences. The authors have experimented with two types of discriminator networks: a conditional discriminator to confirm or not the input style, and a multi-class discriminator that classifies a given sentence to one of the K style classes. The training algorithm goes through two phases: one for training the discriminator and the other is for training of the StyleTransformer network. For better preservation of the content of the original style, a cycle reconstruction loss is used as an objective function when training the model to generate the original input sentence if the generated output sentence is fed into the network.

Context-aware Style Transfer (**CAST**) [18] trains two separate decoders for each sentence to ensure coherence with the adjacent context i.e., sentences in the same paragraph. By doing this, the model was able to preserve the style-independent content of the input sentence, while maintaining its consistency with the surrounding text.

To make use of parallel dataset when available, a bidirectional translation loss was introduced in **HybridST** [16], which is a combination of: 1) the loss of generating the output sentence given the input sentence and 2) the loss of generating an input sentence given an output sentence. In Fine-Grained Controlled Text Generation (**FineGrainedCTGen**) [26] additional Bag of Words (BOW) component was added to enhance generation of specific words and to preserve the content by minimizing the negative log probability of generating BOW features for the output sentence.

In **POS-LM** [25], two additional components were added: Part of Speech (POS) tagger and Language Model (LM). POS tagger assists in generating previously determined nouns in the output sentence, while LM controls the perplexity of the generated sentence. Zhou et al. [28] point out that training a model with reconstruction and classification loss might result in extremely short sentences that might match the target style, but would fail to preserve the original meaning. They proposed the **CP-LM** model that incorporates two losses: 1) content preservation loss to force the word embedding representation of the input and output sentences to be close, by minimizing the difference of their embedding representations and 2) fluency modeling loss to ensure that the output sentences are fluent by minimizing the negative log probability of generated

Model	Year	DL Architecture	Style Transfer Task(s)	Dataset(s)
RetrieveOnly [14]	2018	Baseline ML model	Sentiment Style Transfer	Yelp, Amazon, Captions
TemplateBased [14]	2018	Baseline ML model	Sentiment Style Transfer	Yelp, Amazon, Captions
DeleteOnly [14]	2018	GRU encoder GRU decoder	Sentiment Style Transfer	Yelp, Amazon, Captions
DeleteAndRetrieve [14]	2018	GRU encoder GRU decoder	Sentiment Style Transfer	Yelp, Amazon, Captions
SHAPED [30]	2018	multiple GRU encoders multiple attentive GRU decoders	Genre Transfer	Gigaword
SMAE [23]	2018	LSTM encoder LSTM decoder	Sentiment Style Transfer	Yelp
B-GST [15]	2019	BERT decoder	Sentiment Style Transfer Personal Style Transfer	Yelp, Amazon, Captions Political slant, Gender
G-GST [15]	2019	BERT decoder LSTM encoder	Sentiment Style Transfer Personal Style Transfer	Yelp, Amazon, Captions Political slant, Gender
MultipleAttrTransfer [103]	2019	attentive LSTM decoder	Sentiment Style Transfer	Yelp, Amazon

TABLE III
SELECTION OF STYLE TRANSFER RESEARCH USING SIMPLE RECONSTRUCTION OF THE INPUT SENTENCES IN A NEW STYLE.

words, similar to the bidirectional translation loss used in **HybridST**.

Kim and Sohn [104] argue that performing sentence reconstruction and style control in a single task increases the complexity of the model. Their proposed model **AdaptiveStyleEmbedding** consists of two modules, one for each task: 1) a style module that learns the style embeddings using a style classifier, and 2) an autoencoder that generates the output sentence conditioned on the combined vector of latent representation of the input sentence and the learned style embedding.

3) *Adversarial Models*: Several style transfer models incorporate style discriminators, which have a similar role as the discriminator in the GAN architecture. The research studies using the adversarial framework for generating the output sentence are listed in Table V.

AttrControl [34] is an encoder-decoder based model that employs a Projection Discriminator [162] to generate realistic and style compatible sentences. The discriminator determines whether the generated sentence is real or fake, based on a style embedding and output sentence obtained by a GRU decoder. Controlled Text Generation (**CTGen**) [31] is built upon VAE architecture. Additional CNN discriminators were included to assist the generation process. The generator and the discriminators provide feedback to each other in a collaborative manner with the wake-sleep procedure [163].

Aligned Autoencoder (**AAE**) [32] incorporates a feed-forward discriminator to align both, posterior probability distributions learned with encoders for each style (input and target style). Cross-Aligned Autoencoder (**CAAE**) [32] incorporates two CNN discriminators for the same purpose. Assuming the transfer is between two styles (style s_i and style s_t), one discriminator learns to distinguish between a real sentence with style s_i and generated sentence with style s_t , while the other discriminator learns to distinguish between a real sentence with style s_t and generated sentence with style s_i .

MultiDecoder and **StyleEmbedding** [19], incorporate two multi-layer classifiers to classify the style of the input sentence given the representation learned by the GRU encoder, by 1) maximizing the probability of correctly predicting style labels, and 2) maximizing the entropy of the predicted style labels.

StyleEmbedding uses an additional embedded representation of the target style to the GRU decoder to generate an output sentence, while **MultiDecoder** is composed of multiple GRU decoders (one for each style) to generate a sentence in the target style.

The research study presented by John et al. [22] tackled on somewhat divisive topic of the feasibility of disentangling the content from the style in the latent space. Deterministic AutoEncoder (**DAE**) and Variational AutoEncoder (**VAE**) have been used in their models for the task of sentiment style transfer. A number of content-oriented and style-oriented reconstruction and adversarial losses have been proposed to afford the separation of the latent spaces. The content and style information have been approximated by the bag-of-words (BOW) features. Two classifiers have been used: one for detecting the style and the other over the BOW content vocabulary.

Cycle-consistent Adversarial Autoencoder (**CAE**) [107] is a three-component network consisting of LSTM autoencoder for representing sentences in different styles, adversarial style transfer network, and a novel cycle-consistent constraint. The cycle-consistent reconstruction imposes constraint on the latent representation collectively learned by the LSTM autoencoder and adversarial style network. The results of the conducted ablation study show that the cycle-constraint was instrumental in content preservation during sentiment style transfer. In a similar way, Dual-Generator Network for Text Style Transfer (**DGST**) [102] learns to generate sentences in a target style in a cyclic process. However, this model does not rely upon discriminators. Instead, it applies neighborhood sampling to introduce noise to each sentence. **FM-GAN** [35] is trained with Feature Mover’s Distance instead of traditional loss for adversarial learning.

Zhao et al. [33] point out that the generated sentence may not necessarily capture the target style by training with an objective function that includes only reconstruction and adversarial loss. Their model **StyleDiscrepancy** incorporates an additional discriminator to determine whether a given sentence has the target style with a loss function called style discrepancy. They also apply cycle consistency as in **Style-**

Model	Year	Architecture	Style Transfer Task(s)	Dataset(s)
BST [21]	2018	MT encoder	Sentiment Style Transfer	Yelp, Gender, Political slant
		multiple BiLSTM decoders	Personal Style Transfer	
M-BST [24]	2018	CNN classifier	Sentiment Style Transfer	Yelp, Gender, Political slant
		multilingual MT encoder	Personal Style Transfer	
M-BST+F [24]	2018	multiple BiLSTM decoders	Sentiment Style Transfer	Yelp, Gender, Political slant
		CNN classifier	Personal Style Transfer	
NTST [20]	2018	GRU encoder	Transferring Offensive	Reddit, Twitter
		attentive GRU decoder	to Non-offensive Text	
POS-LM [25]	2018	CNN classifier	Sentiment Style Transfer	Yelp, Political slant
		GRU encoder	Personal Style Transfer	
HybridST [16]	2019	attentive GRU decoder	Formality Transfer	GYAFC
		CNN classifier		
StyleTransformer [17]	2019	Transformer encoder	Sentiment Style Transfer	Yelp
		Transformer decoder		
ControllableAttrTransfer [27]	2019	Transformer discriminator	Sentiment Style Transfer	Yelp, Amazon, Captions
		Transformer encoder		
CP-LM [28]	2020	Transformer decoder	Sentiment Style Transfer	Yelp, GYAFC
		MLP classifier	Formality Transfer	
CAST [18]	2020	GRU encoder	Formality Transfer	GYAFC, Enron corpus, Reddit
		attentive GRU decoder	Transferring Offensive	
StableStyleTransformer [29]	2020	CNN classifier	to Non-offensive Text	Yelp, Amazon
		Transformer encoder		
FineGrainedCTGen [26]	2020	Transformer decoder	Sentiment Style Transfer	Yelp, Amazon
		CNN classifier		
AdaptiveStyleEmbedding [104]	2020	LSTM encoder	Sentiment Style Transfer	Yelp, Amazon
		LSTM decoder		
		MLP classifier	Sentiment Style Transfer	Yelp, Amazon

TABLE IV
SELECTION OF STYLE TRANSFER RESEARCH USING A CLASSIFIER TO ASSIST IN THE GENERATION OF SENTENCES.

Transformer [17] and **CAST** [18] models. **StyIns** [105] model incorporates adversarial style loss to ensure better style supervision during generation. Similar to **StyleTransformer** [17], a multi-class discriminator is applied to determine the style of the generated sentence.

A new framework named Pre-train and Plug-in Variational Autoencoder (**PPVAE**) was proposed by Duan et al. [106] with a realistic system in mind that can mitigate the problem of starting from scratch whenever we need to learn a new style. The framework PPVAE is composed of two variational autoencoders: the PretrainVAE, which learns to represent and reconstruct a sentence in its original style and the PluginVAE, which learns the conditional latent space for each style. The role of PluginVAE as a lightweight easily-trained network is to transform the conditional “style-specific” latent space into the global latent space learned by PretrainVAE and vice versa.

VI. CHALLENGES AND DIRECTIONS FOR FUTURE RESEARCH

Before concluding this paper, it is important to emphasize the challenges the task of text style transfer faces. Research performed thus far offers a promising perspective but also points to unexplored avenues that require further attention.

A. Datasets

Advancing the state-of-the-art systems for style transfer is dependent on the quality and quantity of the available datasets. They provide the necessary support for advocating the use of deep learning techniques. Due to the diversity of style categories, the ambiguity of the stylistic properties being modeled and the costs of labeling, creation of benchmarking datasets is still non-trivial. Publicly available datasets for style transfer vary depending on the type of style they contain (e.g., sentiment, formality, politeness) they contain, the format of text samples, and the procedure used for content validation and labeling (e.g., crowdsourcing, experts). While parallel corpus datasets would be desirable for style transfer tasks, it is often unrealistic to obtain and label large datasets, so the rise in unsupervised deep style transfer methods is not surprising.

B. Deep Learning

Heralded for their capabilities for automated feature learning and complex pattern recognition from vast quantities of big data, deep learning techniques have been at the frontier of innovations for decades now. The aim of this survey was to highlight the importance and to demonstrate the suitability of deep learning for the task at hand.

Model	Year	Architecture	Style Transfer Task(s)	Dataset(s)
AAE [32]	2017	GRU encoder GRU generator	Sentiment Style Transfer	Yelp
		MLP discriminator		
CAAE [32]	2017	GRU encoder GRU generator	Sentiment Style Transfer	Yelp
		CNN discriminator		
CTGen [31]	2017	LSTM encoder LSTM generator	Sentiment Style Transfer	SST, IMDB
		CNN discriminator		
AttrControl [34]	2018	GRU encoder GRU decoder	Sentiment Style Transfer	Yelp, IMDB
		Projection discriminator		
StyleDiscrepancy [33]	2018	two GRU encoders GRU generator	Sentiment Style Transfer	Yelp, Shakespeare
		CNN discriminator	Shakespearean Style Transfer	
FM-GAN [35]	2018	LSTM encoder LSTM generator	Sentiment Style Transfer	Yelp
		MLP classifier		
StyleEmbedding [19]	2018	GRU encoder GRU decoder	Genre Transfer	Paper-News Titles, Amazon
		MLP classifier	Sentiment Style Transfer	
MultiDecoder [19]	2018	GRU encoder multiple GRU decoders	Genre Transfer	Paper-News Titles, Amazon
		MLP classifier	Sentiment Style Transfer	
DAE [22]	2019	GRU encoder GRU decoder	Sentiment Style Transfer	Yelp, Amazon
		CNN classifier		
VAE [22]	2019	GRU encoder GRU decoder	Sentiment Style Transfer	Yelp, Amazon
		CNN classifier		
StyIns [105]	2020	BiLSTM encoder attentive RNN decoder	Sentiment Style Transfer	Yelp, GYAFC
		CNN discriminator	Formality Style Transfer	
CAE [107]	2020	LSTM encoders LSTM generators	Sentiment Style Transfer	Yelp
		MLP discriminators		
DGST [102]	2020	BiLSTM encoders BiLSTM generators	Sentiment Style Transfer	Yelp, IMDB
		BiGRU and MLP encoders		
PPVAE [106]	2020	Transformer and MLP decoders	Genre Transfer	Yelp, News Titles
		MLP discriminator	Sentiment Style Transfer	

TABLE V

SELECTION OF STYLE TRANSFER RESEARCH USING ADVERSARIAL LEARNING FOR GENERATING SENTENCES IN A NEW STYLE.

The focus of the research has shifted from feature extraction to model-free machine learning. The knowledge is unearthed directly from abundant data without the need for domain expertise, hand-crafted feature extraction, or data labeling. Deep learning is currently being preferred choice for text style transfer and have proved to be more scalable, robust, and superior in performance on various style transfer tasks. Existing deep neural architectures have been adapted for both stages in the process: the representation learning of the input sentence whose style needs to be changed and the generation of the output sentence in a new style.

The technological solutions for output sentence generation were the criteria for clustering the style transfer research discussed in the survey into three groups. There are strengths and limitations associated with deep learning. Their dependency on large quantities of data and the complexity of the neural models are related to the problems of overfitting from training data and inability to generalize well.

C. Deep Style Transfer

Points of particular interest for future directions in style transfer are expected in the following areas:

1) *Style-content Disentanglement*: The discussions that extend across several studies are the challenges in disentanglement of content and style in text. Style is indeed inseparably woven into spoken and written language. Some of the past research studies have advocated that the success of the style transfer task depends on the clear separation between the semantic content from the stylistic properties. Approaches being proposed include: partitioning of the latent space into content and style subspaces [19, 22, 31, 32, 33, 105], removal of style markers [14, 15, 23, 29], and use of back-translation for reducing the effects of the style of the input sentence [21, 24].

Across recent studies, the majority of researchers clearly reject the necessity for disentangling content from style; a target that is difficult to reach even by adversarial training. A number of methods have been suggested that are much more effective on style transfer tasks without the need for disentanglement. The use of back-translation technique [17, 18, 20, 103], latent representation editing [27], independent modules for style control and sentence reconstruction [104], and a cycle-consistent reconstruction [107] are some of the approaches put forward. Given the promising success of adversarial-based method for style transfer task, there is evidence that further advancing the

proposed adversarial architectures is worth exploring. Surely, these divisive views need to be further explored to consolidate the views across different style transfer tasks.

2) *Content Preservation vs. Style Strength Trade-off*: A key challenge for the methods for style transfer in text is identifying strategies that can effectively balance the trade-off between preserving the original content and changing the style of a given sentence [14, 18, 26, 29, 102, 103, 104, 105]. The balance is problematic because the precise nature of the interdependence between style-free and style-dependent content is not clearly defined. The unavoidable trade-off between content preservation and style strength in the proposed models is shared among researchers in the field. Furthermore, previous studies suggest that some technological remedies are at odds with one another [17, 102, 107].

3) *Interpretability*: Deep neural networks are sometimes criticized for being black-box models, their structure and output not intelligible enough to associate causes with effects. In the field of natural language processing, we would want to interpret the model in a way that we can identify the useful patterns and features that contribute more to a better understanding and generation of text. Making attempts to understand how and how well different deep network components have been done in machine vision by dissecting GANs [164].

In the realm of natural language understanding, a number of perspectives has been fitted under the umbrella of style, from genre and formality to personal style and sentiment. This is an area where interdisciplinary endeavor of theoretical and empirical research should complement each other. The topic of stylistic language variations has a rich history in sociolinguistic theory [4, 5, 6], although reciprocal contributions from both sides are expected to shed light on the multifaceted nature of style, provide guidance to the modeling efforts and improve the interpretation of the empirical results.

4) *Transfer Learning*: Training models is a computationally intensive and time-consuming process. In the field of computer vision and natural language understanding, transfer learning has been used to both, to speed up the process and improve model performance. General semantic patterns are learned during pre-training and can be "transferred" to new tasks. By fine-tuning, such additional semantic information can be "transferred" into the learned representation. Currently, the number of research studies using transfer learning for style transfer is still limited. However, as popularity and attention to the topic increases, the idea of pre-training, multi-task training, and then fine-tuning can be more promising.

5) *Ethical Considerations*: The far-reaching consequences of any research should engage the community in ethical discussions on potential malicious abuse as well as the impact a technology has on transforming many aspects of human life. The need for reflection is further amplified by the increased reliance of current machine learning technologies on abundant data that is scraped from social networks.

6) *Deep Reinforcement Learning*: The new directions towards human-like understanding, such as few-shot learning [165], inductive learning [166] or lifelong learning [167] are promising areas of machine learning. A vision of designing machines that learn like humans from experience

based on a limited number of training examples is yet to be reached [168]. Notably, deep reinforcement learning has recently been revisited by researchers in many fields including our own [169, 170, 171].

The advances reviewed in this survey should be of interest to researchers not only limited to style transfer, but also to other related fields, such as language generation, summarization, question answering and dialogue.

VII. CONCLUSION

Recent advancements in text style transfer using deep learning have been the primary motivation to carry out the survey presented in this paper. A systematic review of state-of-the-art research highlights the trends that appeared to extend across research studies as well as differences and variations in style transfer methodologies using deep learning. In particular, this review examines how encoder-decoder-based architectures are still dominating the field, with a more recent move toward adversarial learning using Generative Adversarial Networks. While it appears that a choice of one deep neural network over another is style-independent, balancing the trade-offs between the complexity of a model and the expected performance gains added by auxiliary components (e.g., classifier, discriminator) are consistent challenges faced by researchers. The review is structured around the key stages in style transfer process and the methodological differences adopted by researchers for each stage.

It is our hope that the review would serve as a guideline for future studies that are built on the best practices of past research as well as the new direction that can enrich the field. Notably, successful studies of generalizing results across style transfer tasks are rarely reported. 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

- [1] P. Eckert and J. R. Rickford, *Style and sociolinguistic variation*. Cambridge University Press, 2001.
- [2] N. Coupland, *Style: Language variation and identity*. Cambridge University Press, 2007.
- [3] S. F. Kiesling and N. Schilling-Estes, "Language style as identity construction: A footing and framing approach," 1998.
- [4] W. Labov, *Sociolinguistic patterns*. University of Pennsylvania Press, 1972, no. 4.
- [5] W. Labov *et al.*, "Field methods of the project on linguistic change and variation," 1981.
- [6] W. Labov, "Some principles of linguistic methodology." *Language in Society*, vol. 1, no. 1, pp. 97–120, 1972.
- [7] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology*, vol. 54, no. 1, pp. 547–577, 2003.

- [8] A. Bell, “Language style as audience design,” *Language in society*, vol. 13, no. 2, pp. 145–204, 1984.
- [9] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, “A persona-based neural conversation model,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 994–1003.
- [10] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–Decoder approaches,” in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 103–111.
- [11] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” 2016.
- [12] S. Chopra, M. Auli, and A. M. Rush, “Abstractive sentence summarization with attentive recurrent neural networks,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 93–98.
- [13] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, “A hierarchical latent variable encoder-decoder model for generating dialogues,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 3295–3301.
- [14] J. Li, R. Jia, H. He, and P. 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)*, 2018, pp. 1865–1874.
- [15] A. Sudhakar, B. Upadhyay, and A. Maheswaran, ““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)*, 2019, pp. 3260–3270.
- [16] R. Xu, T. Ge, and F. Wei, “Formality style transfer with hybrid textual annotations,” *CoRR*, 2019.
- [17] N. Dai, J. Liang, X. Qiu, and X.-J. Huang, “Style Transformer: Unpaired text style transfer without disentangled latent representation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5997–6007.
- [18] Y. Cheng, Z. Gan, Y. Zhang, O. Elachgar, D. Li, and J. Liu, “Contextual text style transfer,” *CoRR*, 2020.
- [19] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, “Style transfer in text: Exploration and evaluation,” in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [20] C. dos Santos, I. Melnyk, and I. Padhi, “Fighting offensive language on social media with unsupervised text style transfer,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 189–194.
- [21] S. Prabhunoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black, “Style transfer through back-translation,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 866–876.
- [22] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, “Disentangled representation learning for non-parallel text style transfer,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 424–434.
- [23] Y. Zhang, J. Xu, P. Yang, and X. Sun, “Learning sentiment memories for sentiment modification without parallel data,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1103–1108.
- [24] S. Prabhunoye, Y. Tsvetkov, A. W. Black, and R. Salakhutdinov, “Style transfer through multilingual and feedback-based back-translation,” *CoRR*, 2018.
- [25] Y. Tian, Z. Hu, and Z. Yu, “Structured content preservation for unsupervised text style transfer,” *CoRR*, 2018.
- [26] D. Liu, J. Fu, Y. Zhang, C. Pal, and J. Lv, “Revision in continuous space: Unsupervised text style transfer without adversarial learning,” in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 8376–8383.
- [27] K. Wang, H. Hua, and X. Wan, “Controllable unsupervised text attribute transfer via editing entangled latent representation,” in *Advances in Neural Information Processing Systems*, 2019, pp. 11 036–11 046.
- [28] C. Zhou, L. Chen, J. Liu, X. Xiao, J. Su, S. Guo, and H. Wu, “Exploring contextual word-level style relevance for unsupervised style transfer,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schlueter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 7135–7144.
- [29] J. Lee, “Stable Style Transformer: Delete and Generate approach with Encoder-Decoder for text style transfer,” *CoRR*, 2020.
- [30] Y. Zhang, N. Ding, and R. Soricut, “SHAPED: Shared-Private Encoder-Decoder for text style adaptation,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1528–1538.
- [31] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, “Toward controlled generation of text,” in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1587–1596.
- [32] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, “Style transfer from non-parallel text by cross-alignment,” in

- Advances in neural information processing systems*, 2017, pp. 6830–6841.
- [33] Y. Zhao, W. Bi, D. Cai, X. Liu, K. Tu, and S. Shi, “Language style transfer from sentences with arbitrary unknown styles,” *CoRR*, 2018.
- [34] L. Logeswaran, H. Lee, and S. Bengio, “Content preserving text generation with attribute controls,” in *Advances in Neural Information Processing Systems*, 2018, pp. 5103–5113.
- [35] L. Chen, S. Dai, C. Tao, H. Zhang, Z. Gan, D. Shen, Y. Zhang, G. Wang, R. Zhang, and L. Carin, “Adversarial text generation via feature-mover’s distance,” in *Advances in Neural Information Processing Systems*, 2018, pp. 4666–4677.
- [36] S. Argamon and M. Koppel, “The rest of the story: Finding meaning in stylistic variation,” in *The Structure of Style*. Springer, 2010, pp. 79–112.
- [37] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” *Text & Talk*, vol. 23, no. 3, pp. 321–346, 2003.
- [38] C. Peersman, W. Daelemans, R. Vandekerckhove, B. Vandekerckhove, and L. Van Vaerenbergh, “The effects of age, gender and region on non-standard linguistic variation in online social networks.”
- [39] D. Bamman, J. Eisenstein, and T. Schnoebelen, “Gender identity and lexical variation in social media,” *Journal of Sociolinguistics*, vol. 18, no. 2, pp. 135–160, 2014.
- [40] M. Koppel, S. Argamon, and A. R. Shimoni, “Automatically categorizing written texts by author gender,” *Literary and linguistic computing*, vol. 17, no. 4, pp. 401–412, 2002.
- [41] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, “Effects of age and gender on blogging,” in *AAAI spring symposium: Computational approaches to analyzing weblogs*, vol. 6, 2006, pp. 199–205.
- [42] M. Koppel, S. Argamon, and A. Shimoni, “Automatically determining the gender of a text’s author,” *Bar-Ilan University Technical Report BIU-TR-01-32*, 2001.
- [43] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in twitter,” in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 2010, pp. 37–44.
- [44] D. Preotiuc-Pietro, W. Xu, and L. Ungar, “Discovering user attribute stylistic differences via paraphrasing,” in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [45] D. Nguyen, R. Gravel, D. Trieschnigg, and T. 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.
- [46] H. Jhamtani, V. Gangal, E. Hovy, and E. Nyberg, “Shakespearizing modern language using copy-enriched sequence to sequence models,” in *Proceedings of the Workshop on Stylistic Variation*, 2017, pp. 10–19.
- [47] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, “Skip-thought vectors,” *arXiv preprint arXiv:1506.06726*, 2015.
- [48] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books,” *arXiv preprint arXiv:1506.06724*, 2015.
- [49] F. A. Sheikha and D. Inkpen, “Automatic classification of documents by formality,” in *Proceedings of the 6th international conference on natural language processing and knowledge engineering (nlpke-2010)*. IEEE, 2010, pp. 1–5.
- [50] F. Heylighen and J.-M. Dewaele, “Formality of language: Definition, measurement and behavioral determinants,” *Interneter Bericht, Center “Leo Apostel”, Vrije Universiteit Brussel*, vol. 4, 1999.
- [51] K. Peterson, M. Hohensee, and F. Xia, “Email formality in the workplace: A case study on the enron corpus,” in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 86–95.
- [52] S. Lahiri, P. Mitra, and X. Lu, “Informality judgment at sentence level and experiments with formality score,” in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2011, pp. 446–457.
- [53] S. Rao and J. Tetreault, “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)*, 2018, pp. 129–140.
- [54] P. Brown, S. C. Levinson, and S. C. Levinson, *Politeness: Some universals in language usage*. Cambridge university press, 1987, vol. 4.
- [55] P. Chilton, “Politeness, politics and diplomacy,” *Discourse & Society*, vol. 1, no. 2, pp. 201–224, 1990.
- [56] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 250–259.
- [57] L. Coppock, “Politeness strategies in conversation closings,” *Unpublished manuscript: Stanford University*, 2005.
- [58] L. M. Andersson and C. M. Pearson, “Tit for tat? The spiraling effect of incivility in the workplace,” *Academy of management review*, vol. 24, no. 3, pp. 452–471, 1999.
- [59] N. Chhaya, K. Chawla, T. Goyal, P. Chanda, and J. Singh, “Frustrated, polite, or formal: Quantifying feelings and tone in email,” in *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, 2018, pp. 76–86.
- [60] J. Pavlopoulos, N. Thain, L. Dixon, and I. 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*, 2019, pp.

- 571–576.
- [61] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Predicting the type and target of offensive posts in social media,” 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)*, 2019, pp. 1415–1420.
- [62] J. Risch, R. Ruff, and R. Krestel, “Offensive language detection explained,” in *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, 2020, pp. 137–143.
- [63] D. Y. Lee, “Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the bnc jungle,” 2001.
- [64] D. Biber, *Variation across speech and writing*. Cambridge University Press, 1991.
- [65] N. Dewdney, C. Van Ess-Dykema, and R. MacMillan, “The form is the substance: Classification of genres in text,” in *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, 2001.
- [66] R. F. Baumeister, K. D. Vohs, C. Nathan DeWall, and L. Zhang, “How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation,” *Personality and social psychology review*, vol. 11, no. 2, pp. 167–203, 2007.
- [67] J. A. Russell, “A circumplex model of affect.” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.
- [68] R. Plutchik, “Emotions: A general psychoevolutionary theory,” *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.
- [69] P. Ekman, “An argument for basic emotions,” *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
- [70] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, “The hourglass model revisited,” *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.
- [71] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment classification using machine learning techniques,” in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002, pp. 79–86.
- [72] S.-M. Kim and E. Hovy, “Determining the sentiment of opinions,” in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, 2004, pp. 1367–1373.
- [73] M. S. Akhtar, A. Ekbal, and E. Cambria, “How intense are you? Predicting intensities of emotions and sentiments using stacked ensemble [application notes],” *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64–75, 2020.
- [74] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, “A transformer-based joint-encoding for emotion recognition and sentiment analysis,” in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020, pp. 1–7.
- [75] S. Javdan, B. Minaei-Bidgoli *et al.*, “Applying transformers and aspect-based sentiment analysis approaches on sarcasm detection,” in *Proceedings of the Second Workshop on Figurative Language Processing*, 2020, pp. 67–71.
- [76] A. Shenoy and A. Sardana, “Multilogue-Net: A context-aware rnn for multi-modal emotion detection and sentiment analysis in conversation,” in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020, pp. 19–28.
- [77] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Seventh international AAAI conference on weblogs and social media*, 2013.
- [78] M. Hasan, E. Rundensteiner, and E. Agu, “Emotex: Detecting emotions in twitter messages,” 2014.
- [79] Y. Ma, H. Peng, and E. Cambria, “Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [80] M. Kale and M. Rikters, “Fragmented and valuable: Following sentiment changes in food tweets,” *CoRR*, 2021.
- [81] B.-K. H. Vo and N. Collier, “Twitter emotion analysis in earthquake situations,” *International Journal of Computational Linguistics and Applications*, vol. 4, no. 1, pp. 159–173, 2013.
- [82] C. Strapparava and R. Mihalcea, “Learning to identify emotions in text,” in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.
- [83] A. Khatua, A. Khatua, and E. Cambria, “Predicting political sentiments of voters from Twitter in multi-party contexts,” *Applied Soft Computing*, vol. 97, p. 106743, 2020.
- [84] J. A. Chevalier and D. Mayzlin, “The effect of word of mouth on sales: Online book reviews,” *Journal of marketing research*, vol. 43, no. 3, pp. 345–354, 2006.
- [85] Y. Liu, X. Huang, A. An, and X. Yu, “ARSA: A sentiment-aware model for predicting sales performance using blogs,” in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 607–614.
- [86] M. McGlohon, N. Gance, and Z. Reiter, “Star quality: Aggregating reviews to rank products and merchants,” in *Fourth international AAAI conference on weblogs and social media*, 2010.
- [87] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.
- [88] Y. Li, S. Wang, Y. Ma, Q. Pan, and E. Cambria, “Popularity prediction on vacation rental websites,” *Neurocomputing*, vol. 412, pp. 372–380, 2020.
- [89] W. Xu, A. Ritter, B. Dolan, R. Grishman, and C. Cherry, “Paraphrasing for style,” in *Proceedings of COLING 2012*, 2012, pp. 2899–2914.
- [90] W. Xu, “Data-driven approaches for paraphrasing across

- language variations,” Ph.D. dissertation, New York University, 2014.
- [91] B. Klimt and Y. Yang, “Introducing the enron corpus.” in *CEAS*, 2004.
- [92] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, “StyleNet: Generating attractive visual captions with styles,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3137–3146.
- [93] S. Reddy and K. Knight, “Obfuscating gender in social media writing,” in *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016, pp. 17–26.
- [94] R. He and J. McAuley, “Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering,” in *proceedings of the 25th international conference on world wide web*, 2016, pp. 507–517.
- [95] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [96] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang, “Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS),” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 193–202.
- [97] D. Graff, J. Kong, K. Chen, and K. Maeda, “English gigaword,” *Linguistic Data Consortium, Philadelphia*, vol. 4, no. 1, p. 34, 2003.
- [98] C. Napoles, M. R. Gormley, and B. Van Durme, “Annotated gigaword,” in *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction (AKBC-WEKEX)*, 2012, pp. 95–100.
- [99] R. Voigt, D. Jurgens, V. Prabhakaran, D. Jurafsky, and Y. Tsvetkov, “RtGender: A corpus for studying differential responses to gender,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
- [100] A. Madaan, A. Setlur, T. Parekh, B. Póczos, G. Neubig, Y. Yang, R. Salakhutdinov, A. W. Black, and S. Prabhume, “Politeness transfer: A tag and generate approach,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 1869–1881.
- [101] Y. Cao, R. Shui, L. Pan, M. Kan, Z. Liu, and T. Chua, “Expertise style transfer: A new task towards better communication between experts and laymen,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Association for Computational Linguistics, 2020, pp. 1061–1071.
- [102] X. Li, G. Chen, C. Lin, and R. Li, “DGST: a dual-generator network for text style transfer,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, 2020, pp. 7131–7136.
- [103] G. Lample, S. Subramanian, E. Smith, L. Denoyer, M. Ranzato, and Y.-L. Boureau, “Multiple-attribute text rewriting,” in *International Conference on Learning Representations*, 2018.
- [104] H. Kim and K.-A. Sohn, “How positive are you: Text style transfer using adaptive style embedding,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2115–2125.
- [105] X. Yi, Z. Liu, W. Li, and M. Sun, “Text style transfer via learning style instance supported latent space.” *IJCAI*, 2020.
- [106] Y. Duan, C. Xu, J. Pei, J. Han, and C. Li, “Pre-train and Plug-in: Flexible conditional text generation with variational auto-encoders,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 253–262.
- [107] Y. Huang, W. Zhu, D. Xiong, Y. Zhang, C. Hu, and F. Xu, “Cycle-consistent adversarial autoencoders for unsupervised text style transfer,” in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 2213–2223.
- [108] A. Gatt and E. Krahmer, “Survey of the state of the art in natural language generation: Core tasks, applications and evaluation,” *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.
- [109] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [110] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.
- [111] M. Denkowski and A. Lavie, “METEOR universal: Language specific translation evaluation for any target language,” in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 376–380.
- [112] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.
- [113] G. Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” in *Proceedings of the second international conference on Human Language Technology Research*, 2002, pp. 138–145.
- [114] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia*,

- April 26-30, 2020. OpenReview.net, 2020.
- [115] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “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)*, 2019, pp. 4171–4186.
- [116] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” *Text Summarization Branches Out*, 2004.
- [117] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, “Optimizing statistical machine translation for text simplification,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.
- [118] D. L. Chen and W. B. Dolan, “Collecting highly parallel data for paraphrase evaluation,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 190–200.
- [119] A. Graves, “Generating sequences with recurrent neural networks,” *CoRR*, 2013.
- [120] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [121] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [122] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *CoRR*, 2014.
- [123] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *CoRR*, 2015.
- [124] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [125] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, 2014.
- [126] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 655–665.
- [127] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.
- [128] Y. Zhang and B. C. Wallace, “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 253–263.
- [129] Y. Zha, R. Li, and H. Lin, “Gated convolutional bidirectional attention-based model for off-topic spoken response detection,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 600–608.
- [130] J. Wang and X. Hu, “Convolutional neural networks with gated recurrent connections,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [131] S. Adams, D. Melanson, and M. De Cock, “Private text classification with convolutional neural networks,” in *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, 2021, pp. 53–58.
- [132] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.
- [133] D. Britz, A. Goldie, M.-T. Luong, and Q. Le, “Massive exploration of neural machine translation architectures,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1442–1451.
- [134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [135] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,” *CoRR*, 2019.
- [136] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” *CoRR*, 2019.
- [137] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018.
- [138] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [139] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *CoRR*, 2020.
- [140] O. Vinyals, M. Fortunato, and N. Jaitly, “Pointer networks,” in *Advances in neural information processing systems*, 2015, pp. 2692–2700.
- [141] S. Merity, C. Xiong, J. Bradbury, and R. Socher, “Pointer sentinel mixture models,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [142] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [143] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” California Univ San Diego La Jolla Inst for

- Cognitive Science, Tech. Rep., 1985.
- [144] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.
- [145] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. 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)*, 2014, pp. 1724–1734.
- [146] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, “Neural abstractive text summarization with sequence-to-sequence models,” *ACM Transactions on Data Science*, vol. 2, no. 1, pp. 1–37, 2021.
- [147] B. Liu, “Neural question generation based on Seq2Seq,” in *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence*, 2020, pp. 119–123.
- [148] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [149] M. A. Haidar and M. Rezagholizadeh, “TextKD-GAN: Text generation using knowledge distillation and generative adversarial networks,” in *Canadian Conference on Artificial Intelligence*. Springer, 2019, pp. 107–118.
- [150] A. Ahamad, “Generating text through adversarial training using skip-thought vectors,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 53–60.
- [151] V.-P. Berges, I. Bush, and P. Rosello, “Text generation using generative adversarial networks.”
- [152] L. Wu, Y. Xia, F. Tian, L. Zhao, T. Qin, J. Lai, and T.-Y. Liu, “Adversarial neural machine translation,” in *Asian Conference on Machine Learning*. PMLR, 2018, pp. 534–549.
- [153] L. Liu, Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, “Generative adversarial network for abstractive text summarization,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [154] S. Rao and H. Daumé III, “Answer-based adversarial training for generating clarification questions,” 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)*, 2019, pp. 143–155.
- [155] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *International Conference on Machine Learning*. PMLR, 2015, pp. 1530–1538.
- [156] E. Rabinovich, R. N. Patel, S. Mirkin, L. Specia, and S. Wintner, “Personalized machine translation: Preserving original author traits,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 1074–1084.
- [157] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, “Structured attention networks,” in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.
- [158] A. Prakash, S. A. Hasan, K. Lee, V. Datla, A. Qadir, J. Liu, and O. Farri, “Neural paraphrase generation with stacked residual lstm networks,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2923–2934.
- [159] Z. Cao, C. Luo, W. Li, and S. Li, “Joint copying and restricted generation for paraphrase,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 3152–3158.
- [160] A. Gupta, A. Agarwal, P. Singh, and P. Rai, “A deep generative framework for paraphrase generation,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, S. A. McIlraith and K. Q. Weinberger, Eds. AAAI Press, 2018, pp. 5149–5156.
- [161] Y. Wang, Y. Wu, L. Mou, Z. Li, and W. Chao, “Harnessing pre-trained neural networks with rules for formality 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)*, 2019, pp. 3564–3569.
- [162] T. Miyato and M. Koyama, “cGANs with projection discriminator,” in *International Conference on Learning Representations*, 2018.
- [163] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, “The “wake-sleep” algorithm for unsupervised neural networks,” *Science*, vol. 268, no. 5214, pp. 1158–1161, 1995.
- [164] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks,” in *International Conference on Learning Representations*, 2018.
- [165] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, “Generalizing from a few examples: A survey on few-shot learning,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020.
- [166] S. Murty, P. W. Koh, and P. Liang, “ExpBERT: Representation engineering with natural language explanations,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2106–2113.
- [167] Z. Chen and B. Liu, “Lifelong machine learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 12, no. 3, pp. 1–207, 2018.
- [168] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and brain sciences*, vol. 40, 2017.

- [169] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, “Deep reinforcement learning for dialogue generation,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192–1202.
- [170] Z. Li, X. Jiang, L. Shang, and H. Li, “Paraphrase generation with deep reinforcement learning,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3865–3878.
- [171] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.