Challenges in Emotion Style Transfer: An Exploration with a Lexical Substitution Pipeline

David Helbig, Enrica Troiano and **Roman Klinger**
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart, Germany
{david.helbig,enrica.troiano,roman.klinger}@ims.uni-stuttgart.de

Abstract

We propose the task of emotion style transfer, which is particularly challenging, as emotions (here: anger, disgust, fear, joy, sadness, surprise) are on the fence between content and style. To understand the particular difficulties of this task, we design a transparent emotion style transfer pipeline based on three steps: (1) select the words that are promising to be substituted to change the emotion (with a brute-force approach and selection based on the attention mechanism of an emotion classifier), (2) find sets of words as candidates for substituting the words (based on lexical and distributional semantics), and (3) select the most promising combination of substitutions with an objective function which consists of components for content (based on BERT sentence embeddings), emotion (based on an emotion classifier), and fluency (based on a neural language model). This comparably straightforward setup enables us to explore the task and understand in what cases lexical substitution can vary the emotional load of texts, how changes in content and style interact and if they are at odds. We further evaluate our pipeline quantitatively in an automated and an annotation study based on Tweets and find, indeed, that simultaneous adjustments of content and emotion are conflicting objectives: as we show in a qualitative analysis motivated by Scherer's emotion component model, this is particularly the case for implicit emotion expressions based on cognitive appraisal or descriptions of bodily reactions.

1 Introduction

Humans are capable of saying the same thing in many ways. Careful lexical choices can re-shape a concept in different modes of presentation, giving it a humourous tone, for example, or some degree of formality, or a rap vibe. This type of linguistic creativity has recently been mirrored in the task of textual style transfer, where a stylistic variation is induced on an existing piece of text. The core idea is that texts have a content and a style, and that it is possible to keep the one while changing the other.

Past work on style transfer has targeted attributes (or styles) like sentiment (Dai et al., 2019) and tense (Hu et al., 2017), producing a rich literature on deep generative models that disentangle the content and the style of an input text, and subsequently condition generation towards a desired style (Fu et al., 2018; Shen et al., 2017; Prabhumoye et al., 2018). With this paper, we propose a non-binary style transfer setting, namely emotion style transfer, in which the target corresponds to one emotion (following Ekman's fundamental emotions of anger, fear, joy, surprise, sadness, and disgust). Further, this setting is particularly challenging as emotions are on the fence between content and style. To the best of our knowledge, this type of attribute has been explored only to some degree by the unpublished work by Smith et al. (2019), who transfer text towards 20 affect-related styles. Emotions received more attention in conditioned text generation (Ghosh et al., 2017; Huang et al., 2018; Song et al., 2019).

To explore the challenges of emotion style transfer (for which we depict an example in Figure 1), we develop a transparent pipeline based on lexical substitution (in contrast to a black-box neural encoder/decoder approach), in which we first (1) select those words that are promising to be changed to adapt the target style, (2) find candidates that may substitute these words, (3) select the best combination regarding content similarity to original

In (Anger):	This **soul-crushing drudgery plagues** him
Out (Joy):	This **fulfilling job motivates** him

Figure 1: An example of emotion transfer performed with lexical substitution.

input, target style, and fluency. As we will see, this straight-forward approach is promising while it still enables to understand the changes to the text and their function.

Emotions are not only interesting from the point of view that they contribute to content and style. They are also a comparably well-investigated phenomenon with a rich literature in psychology. For instance, Scherer (2005) states that emotions consist of different components, namely a cognitive appraisal, bodily symptoms, a subjective feeling, expression, and action tendencies. Descriptions of all these components can be realized in natural language to communicate a specific private emotional state. We argue (and analyze based on examples later) that a report of a feeling ("*I am happy*") might be challenging in a different way than descriptions of bodily reactions (*"I am sweating"*) or events (*"My dog was overrun by a car"*).

With our white-box approach of style transfer and the evaluation on the novel task of emotion transfer, we address the following research questions: To what extent can lexical substitution modulate the emotional leaning of text? What is its limitation (e.g., by changing the emotion "style", does content change as well)? Our results show that the success of this approach, both in terms of style change and content preservation, depends on the strategies used for selection and substitution, and that emotion transfer is a viable task to address. Further, we see in a qualitative analysis that what an emotion classification model bases its decisions on might not be sufficient to guide a style transfer method. This becomes evident when we compare how transfer is realized across types of emotion expressions, corresponding to specific components of Scherer's model.

Our implementation is available at http://www.ims.uni-stuttgart.de/data/lexicalemotiontransfer.

2 Related Work

2.1 Emotion Analysis

In the field of psychology, the two main emotion traditions are categorical models and the strand that focuses on the continuous nature of humans' affect (Scherer, 2005). Emotions are grouped into categories corresponding to emotion terms, some of which are prototypical experiences shared across cultures. For Ekman (1992), they are anger, joy, surprise, disgust, fear and sadness; on top of these, Plutchik (2001) adds anticipation and trust. Posner et al. (2005), instead locates emotions along interval scales of affect components (valence, arousal, dominance).

These studies have also influenced computational approaches to emotions, whose preliminary requirement is to follow a specific conceptualization coming from psychology, in order to determine the number and type of emotion classes to research in language. Emotion analysis in natural language processing has mainly established itself as a classification task, aimed at assigning a text to the emotion it expresses (Alm et al., 2005). It has been conducted on a variety of corpora that encompass different types of annotations[1], based on one of the established emotion models mentioned above. Such studies also differ with respect to the textual genres they consider, ranging from from tweets (Mohammad et al., 2017; Klinger et al., 2018) to literary texts (Kim et al., 2017).

While emotion classification approaches have been used to guide controlled generation of text (Ghosh et al., 2017; Huang et al., 2018; Song et al., 2019), computationally modelling emotions has not yet been applied to style transfer. After describing a method to address such task, we analyse its performance by leveraging Scherer's component model: emotions are underlied by various dimensions of cognitive appraisal, which can be differently expressed in text and may pose different challenges for style transfer.

2.2 Style Transfer

Most of the recently published approaches to style transfer make use of artificial neural network architectures, in which some latent semantic representation is the backbone of the system. For instance, Prabhumoye et al. (2018) use neural back-translation to encode the content of text while reducing its stylistic properties, and later decoding it with a specific target style. Gong et al. (2019) evaluate paraphrases regarding their fluency, similarity to the input text and expression of a desired target style, and use this as feedback in a reinforcement learning approach. Li et al. (2018) combine rules with neural methods to explicitly encode attribute markers of the target style.

Such transfer methods have been applied to a variety of styles, including sentiment (Shen et al., 2017; Fu et al., 2018; Xu et al., 2018) and a num-

[1] A comprehensive list of available emotion datasets and annotation schemes can be found in Bostan and Klinger (2018).

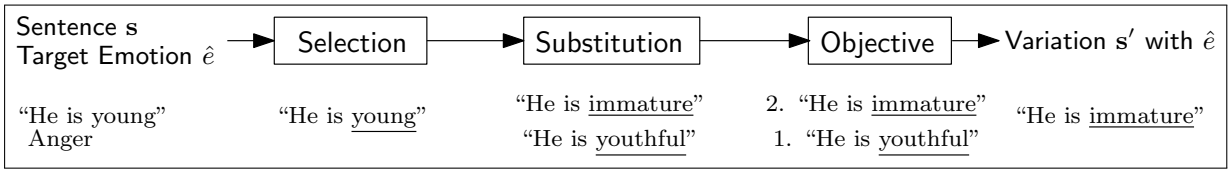


Figure 2: Pipeline model architecture. The selection module marks tokens to substitute, the substitution module retrieves candidates and perform substitution. The objective ranks and scores variations.

ber of affect-related variables (Smith et al., 2019). Other examples include text genres (Lee et al., 2019; Jhamtani et al., 2017), romanticism (Li et al., 2018), politeness/offensiveness and formality (Sennrich et al., 2016; Nogueira dos Santos et al., 2018; Wang et al., 2019).

One of the earliest methods that targets sentiment is proposed by Guerini et al. (2008), who change, add and delete sentiment-related words in a lexical substitution framework. Their strategy to retrieve candidate substitutes is informed by a thesaurus and an emotion dictionary: the first facilitates the extraction of substitutes standing in a specific semantic relation to the input words, the other allows to pick those words that have the desired valence score. Following this approach, Whitehead and Cavedon (2010) filter out ungrammatical expressions resulting from lexical substitution.

Like some works mentioned above, we adopt the view that emotions can be transfered by focusing on specific words, we use WordNet as a source of lexical substitutes, and we consider the three objectives of fluency, similarity and the presence of the target style. Moreover, we opt for a more interpretable solution than neural strategies, as we aim at pointing out what leads to a successful transfer, and what, on the contrary, prevents it.

2.3 Paraphrase Generation through Lexical Substitution

Lexical substitution received some attention independent of style transfer, as it is useful for a range of applications, like paraphrase generation and text summarisation (Dagan et al., 2006). This task, which was formulated by McCarthy and Navigli (2007) and implemented as part of the SemEval-2007 workshop, consists in finding lexical substitutes close in meaning to the original word, given its context within a sentence. The task has mainly been addressed using handcrafted and crowdsourced thesauri, such as WordNet, in order to retrieve lexical substitutes (Martinez et al., 2007; Sinha and Mihalcea, 2014; Kremer et al., 2014; Biemann, 2013). Moreover, it has been approached with distributional spaces, where the embeddings of the candidate substitutes of a target word can be found, and they can be ranked according to their similarity to the target embedding (Zhao et al., 2007; Hassan et al., 2007), as well as the similarity of their contextual information (Melamud et al., 2015)[2].

In the present paper, we follow a similar progression: we retrieve candidates for lexical substitution in WordNet; then, in our more advanced systems, we switch to embedding-based retrieval models.

3 Methods

Emotion transfer can be seen as a task in which a sentence $\mathbf{s}$ is paraphrased, and the result of this operation exhibits a different emotion than $\mathbf{s}$, specifically, a target emotion. We address emotion transfer with a pipeline in which each unit contributes to the creation of emotionally loaded paraphrases. The pipeline is shown in Figure 2. First is a *selection* component, which identifies the tokens in $\mathbf{s}$ that are to be changed. Then, the *substitution* component takes care of the actual substitution. It is responsible for finding candidate substitutes for the tokens that have been selected, producing paraphrases of the input sentence. Importantly, paraphrases are over-generated: at this stage of the pipeline, the output is likely to include sentences that do not express the target emotion. Paraphrases are then scored and re-ranked in the last, objective component, which picks up the "best" output.

3.1 Selection

This component identifies those tokens from a sentence $\mathbf{s} = t_1, \ldots t_n$ that will be substituted later, and groups them into *selections* $\mathcal{S} = \{S_i\}$, where each S_i consists of tokens, $S_i = \{t_i, \ldots, t_j\}$ $(1 \geq i, j \leq n)$. We experiment with two selection strategies, in which the maximal number of tokens

[2] A comparison of different context-aware models for lexical substitution can be found in Soler et al. (2019).

in one selection is p and the maximal number of selections is q ($p, q \in \mathbb{N}$).

Brute-Force. This baseline selection strategy picks each token separately, therefore, we obtain n selections, one for each token, i.e., $\mathcal{S} = \{\{t_1\}, \ldots, \{t_n\}\}$ ($p = 1, q = n$).

Attention-based. To pick words that are likely to influence the (current and target) emotion of a sentence, we exploit an emotion classification model to inform the selection strategy. We train a biLSTM with a self-attention mechanism (Baziotis et al., 2018) and then select those words with a high attention weight to be in the set of selections. To avoid a combinatorial explosion, we consider the k tokens with highest attention weights and add all possible combinations of up to p tokens. Therefore, $q = |\mathcal{S}| = \sum_{i=1}^{k} \binom{p}{i}$. As an example, possible selections in the sentence from Figure 1 for $k = 3, p = 2$ would be $S = \{\{\text{soul-crushing}\}, \{\text{drudgery}\}, \{\text{plagues}\}, \{\text{soul-crushing, drudgery}\}, \{\text{soul-crushing, plagues}\}, \{\text{drudgery, plagues}\}\}$.

3.2 Substitution

The selections $\mathcal{S}$ are then passed to the substitution model together with part-of-speech information. Two tasks are fulfilled by this component: substitution candidates are found for the tokens of each S_i, and the substitution is done by replacing those candidate tokens at position $i, \ldots, j$ in the input sentence $\mathbf{s}$. The next paragraphs detail our strategies for candidate retrieval. We compare a lexical semantics and two distributional semantics-based methods.

WordNet Retrieval. In the WordNet-based method (Fellbaum, 1998), we retrieve the synsets for the respective selected token with the assigned part of speech. Candidates for substitution are the neighboring synsets with the hyponym and hypernym relation (for verbs and nouns) and antonym and synonym relation (for adjectives).

Note that we do not perform word-sense disambiguation prior to retrieving the base synsets. Accordingly, the sense of the selected token in the context of the source sentence and the sense of some retrieved candidates may be different. This is in line with the design of the pipeline and we expect irrelevant forms to be penalised in the objective component.

Distributional Retrieval – Uninformed. In the "Distributional Retrieval – Uninformed" setting, we retrieve u substitution candidates based on the cosine similarity in a vector space. To build the vector space, we employ pre-trained word embeddings.[3] They are the same that are used for training the emotion classifier responsible for retrieving attention scores in the selection stage.

Distributional Retrieval – Informed. A disadvantage of the uniformed method mentioned before might be that the selected u substitutions for each token might not contain words with the targeted emotional orientation. In this approach, we slightly change the substitution selection process by first retrieving a list of u most similar tokens from the vector space. Based on this list, which is presumably of sufficient similarity to the selected token, we select those v relevant for the target emotion.

Let E be the set of emotion categories and $\hat{e} \in E$ the target emotion (with vector representation $\hat{\mathbf{e}}$). Further, let $\bar{\mathbf{e}}$ be the centroid of concepts associated with the respective emotion, as retrieved from the NRC emotion dictionary (Mohammad and Turney, 2013). From the list of semantically similar u candidates c for one token to be substituted, we select the v top scoring ones via

$$\text{score}(c, \hat{e}) = \cos(\hat{\mathbf{e}}, \mathbf{c}) - \frac{1}{|E| - 1} \sum_{\bar{e} \in E \setminus \hat{e}} \cos(\bar{\mathbf{e}}, \mathbf{c}) \,.$$

3.3 Objective

The set of candidate paraphrases produced at substitution time, based on the selections, are an overgeneration which might not be fluent, diverge from the original meaning, and might not contain the target emotion. To select those paraphrases which do not have such unwanted properties, we subselect those with the desired properties based on an objective function $f(\cdot)$ which consists of three components for fluency of the paraphrase $\mathbf{s}'$, semantic similarity between the original sentence $\mathbf{s}$ and the paraphrase $\mathbf{s}'$, and the target emotion $\hat{e}$ of the paraphrase, therefore

$$\text{f}(\mathbf{s}, \mathbf{s}', \hat{e}) = \lambda_1 \cdot \text{emo}(\mathbf{s}', \hat{e}) + \lambda_2 \cdot \text{sim}(\mathbf{s}, \mathbf{s}') + \lambda_3 \cdot \text{flu}(\mathbf{s}').$$

The paraphrase with the highest final score is selected as the result of the emotion transfer process ($\sum_i \lambda_i = 1$).

[3] 300 dimensional embeddings, available at https://github.com/cbaziotis/ntua-slp-semeval2018

Emotion Score. To obtain a score for the target emotion $\hat{e}$ we use an emotion classification model (the same as for the attention selection procedure) in which the last layer is a fully connected layer of size $|E|$ and the output layer is a softmax. Let g represent the classification model that takes a sequence of tokens $\mathbf{s}$ and an emotion e as inputs and produces the activation for e in the final layer. Therefore,

$$\text{emo}(\mathbf{s}', \hat{e}) = \frac{\exp(g(\mathbf{s}', \hat{e}))}{\sum_{e \in E} \exp(g(\mathbf{s}', e))}.$$

Similarity Score. To keep the semantic similarity as much as possible between the input sentence $\mathbf{s}$ and the candidate paraphrase $\mathbf{s}'$, we calculate the cosine similarity between the respective sentence embeddings, based on the pre-trained BERT model (Devlin et al., 2019), in the implementation provided by Wolf et al. (2019). We conceptualize BERT as a mapping function that takes a sequence of tokens $\mathbf{s}$ as input and produces a hidden vector representation for each token. The sentence embeddings $\mathbf{r}$ are obtained by averaging over all hidden vectors.[4] Therefore,

$$\text{sim}(\mathbf{s}, \mathbf{s}') = \cos(\mathbf{r}, \mathbf{r}').$$

Fluency Score. To avoid that tokens are substituted with words which do not fit in the context, we include a language model which scores the paraphrase $\mathbf{s}'$ (similar to Zhao et al., 2018). This model assesses the fluency by perplexity using GPT (Radford et al., 2018), an autoregressive neural language model based on the transformer architecture, which allows us to read the probability of the next token in a sentence given its history. We use a pretrained version of the model provided by Wolf et al. (2019). The perplexity as the average negative log probability over the tokens of our variation sentence $\mathbf{s}'$ is

$$\text{perplexity}(\mathbf{s}') = \frac{1}{n-1} \sum_{i}^{n-1} -\log(P(t_{i+1}|t_1, \ldots, t_i)).$$

Since we are dealing with negative log values, a low perplexity score indicates high probability and therefore high fluency. In order to obtain our final fluency score, we normalize the perplexity to the range $[0, 1]$ and reverse the polarity. To this end, we use the highest perplexity score ($\text{perplexity}_{\text{max}}$) and lowest perplexity score ($\text{perplexity}_{\text{min}}$) that we retrieve among all variation sentences created for our input sentence as scaling factors:

$$\text{flu}(\mathbf{s}') = \frac{\text{perplexity}(\mathbf{s}') - \text{perplexity}_{\text{max}}}{\text{perplexity}_{\text{min}} - \text{perplexity}_{\text{max}}}$$

4 Experiments

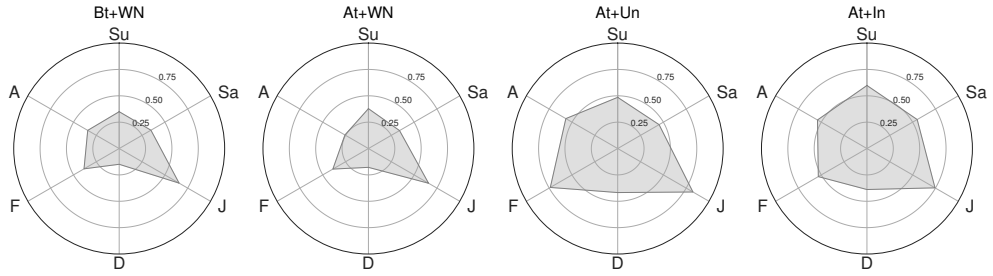
Having established the general pipeline, we move on to the question whether our strategies for selection and substitution actually produce variations with the desired emotion (RQ1). In addition, we examine the interaction between the emotion connotation of the paraphrases and their similarity to the inputs (RQ2). These questions are answered in an automatic and a human evaluation.

4.1 Setting

We instantiate and compare four model configurations for lexical substitution with different combinations of selection and substitution components. These are designed such that we can compare the selection procedure separately from the substitution component.

- **Bf+WN**: We select isolated words in the brute-force configuration and substitute those with the WordNet-based approach.
- **At+WN**: To compare if the attention mechanism is more powerful in finding relevant words to be substituted, we change the brute force selection to the attention-based method. Here, we consider the tokens with the $k = 2$ highest attention scores and combine them to selections with a maximum of $p = 2$ tokens in each selection.
- **At+Un**: We keep the attention mechanism for selection with $k = 2$ and $p = 2$, but vary the substitution component to select $u = 150$ candidates based on semantic similarity. As embedding space, we employ the same pre-trained embeddings we use for training the emotion classifier responsible for retrieving attention weights and calculating emotion scores. The number of variations created amounts to $\sum_{i=1}^{p} \binom{k}{i} u^i = 2 \cdot 150 + 1 \cdot 150^2 = 22800$.
- **At+In**: While the model configuration At+Un generates many possibly irrelevant variations, this model makes informed decisions on how to substitute: we keep the selection as in At+Un, but exchange the substitution method with the informed

[4] As recommended in the documentation of the implementation by Wolf et al. (2019) (https://huggingface.co/transformers/model_doc/bert.html, accessed on March 27, 2020), we do not use the reserved classification token [CLS] as a sentence embedding.



$m = 0.3717$ $m = 0.3668$ $m = 0.5807$ $m = 0.5591$

Figure 3: Automated evaluation results. Each radar plot shows the average emotion scores achieved by transferring 1,000 tweets to anger (A), disgust (D), fear (F), joy (J), sadness (Sa) and surprise (Su); m is the average over all emotions.

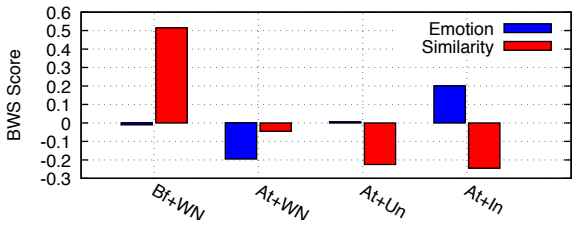


Figure 4: Results for the two human annotation trials, combined by model configuration.

strategy. Specifically, $u = 100$ candidates are found based on their semantic similarity to the token to be substituted, and among those, $v = 25$ tokens are subselected based on their emotion-informed score, leading to $\sum_{i=1}^{p} \binom{k}{i} v^i = 3 \cdot 25 + 3 \cdot 25^2 = 1950$ variations (with $k = 3$, $p = 2$). To inform this method about emotion in the embedding space, we use the NRC emotion dictionary (Mohammad and Turney, 2013).

Automatic Evaluation. The main goal of the automatic evaluation is to compare the potential of increasing the probability that the paraphrase contains the target emotion. To achieve that, we compare the four pipeline configurations, but only use the emotion score as the objective function to pick the best candidate. We use 1000 uniformly sampled Tweets from the corpus TEC (Mohammad, 2012). The emotion classification model used for scoring is trained on the same corpus using pre-trained Twitter embeddings provided by Baziotis et al. (2018).[5]. We use the attention scores obtained from this model for our attention-based selection method. As embedding space for the At+Un and At+In models, we use the same embeddings. As we transfer to the six emotions annotated in TEC, we obtain 6,000 paraphrases with At+Un and At+In and 5,904 with Bf+WN and At+WN (the latter due to non-English words which are not found in WordNet).

Human Evaluation. The goal of the human evaluation is to verify the automatic results (the potential of the selection and substitution components). Further, we compare the association of the paraphrase with the target emotion. To compare a basic setup and the most promising setup, we use $\text{emo}(\mathbf{s}', \hat{e})$ and $\text{sim}(\mathbf{s}, \mathbf{s}')$ for Bf+WN, At+WN, and At+Un and $\text{flu}(\mathbf{s}')$ in addition for At+In. This evaluation is based on 100 randomly sampled Tweets for which we ensure that they are single sentences from TEC. The annotation of emotion connotation and similarity to the original text is then setup as a best-worst-scaling experiment (Louviere et al., 2015), in which each of our two annotators is presented with one paraphrase for each of the four configurations, all for the same emotion (randomly chosen as well). Note that in contrast to best-worst scaling used for annotation as, e.g., in emotion intensity corpus creation (Mohammad et al., 2018), where textual instances are scored, here the instances change from quadruple to quadruple, but the originating configurations remain the same and receive the score. The agreement calculated with Spearman correlation of both annotators is $\rho = 1$ for the emotion connotation and $\rho = 0.8$ for semantic similarity.

4.2 Results

RQ1: Whats is the potential of emotion transfer with lexical substitution? We answer RQ1 by inspecting how likely the paraphrases are to contain the desired emotion and first turn to the automatic evaluation. Figure 3 shows the results. Each radar plot indicates the extent to which the paraphrases of each configuration express the tar-

[5] https://github.com/cbaziotis/ntua-slp-semeval2018

Text	Target
Input	surprises are great when the person is surprised !
Output for Sadness	*depresses* are great when the person is *disappointed* !
Input	love watching my daughter be so excited around christmas
Output for Anger	*detest* watching my daughter be so *annoyed* around christmas

Table 1: Examples of paraphrases produced with At+Inf for different target emotions, using all three components of the objective function.

get emotions. The average probability of the target emotion in the best paraphrases of Bf+WN is 0.3717, indicating that this method has a slightly higher potential than At+WN (0.3668); still, the shape of their plots is comparable. When we compare the substitution method while keeping the selection fixed (At+WN, At+Un, At+In), we see that the distributional methods show a clear increase (0.5807 and 0.5591 average target emotion probability).

In the manual evaluation, we see in Figure 4 (in blue) that the results are in line with the automatic evaluation. Instances originating from At+In are most often chosen as the best results, followed by At+Un and Bf+WN. At+WN scores the worst in human evaluation. Note that the best-worst-scaling results cannot directly be compared to automatic evaluation measures obtained with an automatic text classifier.

RQ2: Is semantic content preserved when changing the emotional orientation? We answer this research question based on the human annotation experiment, with the results in Figure 4. Contrary to the results on the transfer potential, Bf is judged as the most efficient selection strategy for content preservation, while At configurations are dispreferred. The ones based on distributional substitution appear to be worse compared to solutions leveraging WordNet. This shows that Bf provides a lower degree of freedom to the substitution component. The attention mechanism finds the relevant words to be substituted, but the annotators perceive these changes also as a change to the content.

To sum up, highest transfer potential is reached with a combination of attention-based selection, and distributional substitution. The fact that the latter surpasses WordNet-based retrieval may be traced back to the richness of embedding spaces, where substitution candidates can be found which have a higher semantic variability than those found in the thesaurus, and hence, have more varied emotional connotations. In addition, the distributional strategy performing better is the emotion-informed one (0.2 in Figure 4). This suggests that accessing emotion information during substitution is beneficial. The performance of this configuration is exemplified in Table 1, and further discussed in the qualitative analysis.

By comparing the two human trials, it emerges that no configuration excels in both emotion transfer and meaning preservation. In the second case, Attention-based configurations are largely downplayed by Bf+WN. Therefore, to tackle RQ2, the more a system changes emotions, the less it preserves content.

5 Analysis

We now turn to a more qualitative analysis of the results. Due to space restrictions, we show examples for the four pipeline configurations, all with the same objective function $\text{emo}(\cdot)+\text{sim}(\cdot)+\text{flu}(\cdot)$ and a comparison of the At+In model with different objective functions in supplementary material upon acceptance of this paper. Here, in Figure 2, we focus on a discussion of those cases which we consider particularly difficult, though common in everyday communication of emotions. In the selection of these examples, we follow the emotion component model of Scherer (2005) and use two examples, which correspond to a direct (explicit) communication of a subjective feeling (Ex, ID 1, 2), the description of a bodily reaction (BR, ID 3, 4), and a description of an event for which an emotion is developed based on a cognitive appraisal (Ap, ID 5, 6).

The examples which communicate an emotion directly are challenging because there is no other content available than the emotion that is described (ID 1, 2). The model has the choice to exchange two out of three words, and in nearly all cases, it choses to keep "i" and replaces the verb and the emotion word. While the latter is replaced appropriately, the verb is in most cases not substituted in a grammatically correct way. We see here that the emotion classification component in the objective function outrules the language model. This illustrates one fundamental issue with presumably all existing affect-related style transfer method: the original emotion is turned into the target emotion,

ID	Text	Type	Target
1	**I am happy**	Ex	
	i *fuck annoyed*		A
	i *dislike crabby*		D
	i *regret king*		F
	and am *happy*		J
	i am *bummed*		Sa
	i am *surprise*		Su
2	**I am sad**	Ex	
	i am *angrier*		A
	i *embarrassed disgusting*		D
	i *must lies*		F
	finally am *tiring*		J
	i *depressed* sad		Sa
	i *came realise*		Su
3	**Tears are running over my face**	BR	
	rage fuck running over my face		A
	puke are *puking* over my face		D
	shadows are *creeping* over my face		F
	gladness are running over my face		J
	depressed are *leaving* over my face		Sa
	squealed came running over my face		Su

ID	Text	Type	Target
4	**I was tembling**	BR	
	fuck irked trembling		A
	fatass reeks trembling		D
	i *hallucinated trembling*		F
	finally finally trembling		J
	bummed was *trembling*		Sa
	mom showed trembling		Su
5	**My son was standing close to the street**	Ap	
	my *fuck* was standing *annoyed* to the street		A
	my *molest* was *peeing* close to the street		D
	my *coward* was *creeping* close to the street		F
	my *yeshua* was *soaking* close to the street		J
	my *funeral* was *leaving* close to the street		Sa
	my *son* was standing *surprise* to the street		Su
6	**My grandmother died**	Ap	
	fckin grandmother *punched*		A
	ugh grandmother *farted*		D
	my *voldemort attack*		F
	my *family rededicated*		J
	cried grandmother died		Sa
	my *mama showed*		Su

Table 2: Challenging cases for different ways to communicate an internal emotion state. Inputs are in bold; all paraphrases are produced with At+Inf and all three components of the objective function. Ex: Explicit emotion mention, BR: Bodily reaction, Ap: Event appraisal.

but their intensities do not correspond.

In the examples which describe a bodily reaction (ID 3, 4), we see that the attention mechanism does not allow the words "over my face" or "trembling" to change. Instead, it finds the other words more likely to be substituted – the classifier is not informed about the meaning of "trembling" and "over my face". The substituted words make sense, but content and fluency are sacrificed again for the maximal emotion intensity available.

Similarly, the emotion classifier and therefore the associated attention mechanism do not find "close to the street" to be relevant to develop an emotion (ID 5). Instead, other words are exchanged to introduce the target emotion. These issues are mostly due to issues in the emotion classification module. Further, we see that the substitution and selection elements might have a higher chance to perform well if they considered phrases instead of isolated words.

We observe a lack of fluency in many of our output sentences, which we attribute to a dominance of the emotion classifier score. Adapting the weights of the scores in the objective might have potential, however, our findings might suggest that content, emotion and fluency are in conflict with each other – and that obtaining a particular emotion is only possible by sacrificing content similarity. Not doing so seems to lead to non-realistic utterances.

6 Conclusion & Future Work

With this paper, we introduced the task of emotion style transfer, which we have seen to be particularly difficult, on the one side due to being on the fence between content and style, and on the other side due to being a non-binary problem. Our quantitative analyses have shown that there is indeed a trade-of between content preservation and obtaining a target style and that emotion transfer is especially challenging when the text consists of descriptions of emotions in which the separation between content and style is not linguistically clear (as in "I am happy that X happened"). We propose that such test sentences based on descriptions of bodily reactions and event appraisal will be part of future test suits for emotion style transfer, in order to ensure that this task does not work well only on particular expressions of emotions.

We identified the challenge to find the right trade-of between fluency, target emotion, and content preservation. This is particularly challenging, as it would be desirable to separate the emotion intensity from our objective function. We therefore propose that intensity is handled as a fourth component in future work. This could be combined with a decoder as suggested by (Li et al., 2018). Finally, a larger-scale human evaluation should be carried out to clarify the contribution of each component.

References

Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *HLT-EMNLP*.

Christos Baziotis, Athanasiou Nikolaos, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan, and Alexandros Potamianos. 2018. NTUA-SLP at SemEval-2018 task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning. In *SemEval*.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Laura Ana Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *COLING*.

Ido Dagan, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *ACL-COLING*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. Language, speech, and communication. MIT Press, Cambridge, Mass.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A neural language model for customizable affective text generation. In *ACL*.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *NAACL-HLT*, pages 3168–3180.

Marco Guerini, Oliviero Stock, and Carlo Strapparava. 2008. Valentino: A tool for valence shifting of natural language texts. In *LREC*.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. 2007. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *SemEval*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

Chenyang Huang, Osmar Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *NAACL*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*.

Evgeny Kim, Sebastian Padó, and Roman Klinger. 2017. Investigating the relationship between literary genres and emotional plot development. In *LaTeCH-CLfL*.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. IEST: WASSA-2018 implicit emotions shared task. In *WASSA*.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us - analysis of an "all-words" lexical substitution corpus. In *EACL*.

Joseph Lee, Ziang Xie, Cindy Wang, Max Drach, Dan Jurafsky, and Andrew Ng. 2019. Neural text style transfer via denoising and reranking. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 74–81.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: a simple approach to sentiment and style transfer. In *NAACL-HLT*.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-worst scaling: theory, methods and applications*. Cambridge University Press, Cambridge, United Kingdom.

David Martinez, Su Nam Kim, and Timothy Baldwin. 2007. MELB-MKB: Lexical substitution system based on relatives in context. In *SemEval*.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *SemEval*.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*.

Saif Mohammad. 2012. #emotional tweets. In **SEM*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *SemEval*.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology*, 17(3):715–734.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *ACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understandingby generative pre-training. Preprint. Retrieved from https://openai.com/blog/language-unsupervised/ [accessed on December 12, 2019].

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *ACL*.

Klaus R. Scherer. 2005. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *NAACL-HLT*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NIPS*.

Ravi Sinha and Rada Mihalcea. 2014. Explorations in lexical sample and all-words lexical substitution. *Natural Language Engineering*, 20(1):99–129.

Eric Michael Smith, Diana Gonzalez-Rico, Emily Dinan, and Y-Lan Boureau. 2019. Zero-shot fine-grained style transfer: Leveraging distributed continuous style representations to transfer to unseen styles. *arXiv preprint arXiv:1911.03914*.

Aina Garí Soler, Anne Cocos, Marianna Apidianaki, and Chris Callison-Burch. 2019. A comparison of context-sensitive models for lexical substitution. In *ICCS*, pages 271–282. Association for Computational Linguistics.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *ACL*.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *EMNLP-IJCNLP*.

Simon Whitehead and Lawrence Cavedon. 2010. Generating Shifting Sentiment for a Conversational Agent. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In *ACL*.

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

Shiqi Zhao, Lin Zhao, Yu Zhang, Ting Liu, and Sheng Li. 2007. HIT: Web based scoring method for English lexical substitution. In *SemEval*.