# Responsibility Perspective Transfer for Italian Femicide News

**Gosse Minnema[a]*, Huiyuan Lai[a]*, Benedetta Muscato[b] and Malvina Nissim[a]**

[a]University of Groningen, The Netherlands
[b]University of Catania, Italy
{g.f.minnema,h.lai,m.nissim}@rug.nl

## Abstract

Different ways of linguistically expressing the same real-world event can lead to different perceptions of what happened. Previous work has shown that different descriptions of gender-based violence (GBV) influence the reader's perception of who is to blame for the violence, possibly reinforcing stereotypes which see the victim as partly responsible, too. As a contribution to raise awareness on perspective-based writing, and to facilitate access to alternative perspectives, we introduce the novel task of automatically rewriting GBV descriptions as a means to alter the perceived level of responsibility on the perpetrator. We present a quasi-parallel dataset of sentences with low and high perceived responsibility levels for the perpetrator, and experiment with unsupervised (mBART-based), zero-shot and few-shot (GPT3-based) methods for rewriting sentences. We evaluate our models using a questionnaire study and a suite of automatic metrics.

## 1 Introduction

"A terrible incident involving husband and wife", "Husband kills wife", "Her love for him became fatal": these different phrasings can all be used to describe the same violent event, in this case a *femicide*, but they won't trigger the same perceptions in the reader. Perceptions vary from person to person, of course, but also depend substantially and systematically on the different ways the same event is framed (Iyengar, 1994). Especially in the context of gender-based violence (GBV), this has important consequences on how readers will attribute responsibility: victims of femicides are often depicted, and thus perceived, as (co-)responsible for the violence they suffer.[1]

**Responsibility Perspective Transfer**

**Definition:** given a sentence $S$ that references an act of violence, write a sentence $S'$ that describes the same facts as $S$ but increases the perceived level of responsibility on the perpetrator of the violence.

**Examples:** *"A fatal stabbing"* $\mapsto$ *"Someone stabbed another person to death"*
*"Woman murdered by husband"* $\mapsto$ *"Man murders wife"*

Box 1: Task definition

There is indeed evidence from the linguistic literature (Pinelli and Zanchi, 2021; Meluzzi et al., 2021) that people perceive responsibility differently according to how femicides are reported (more blame on the perpetrator in "Husband kills wife", more focus on the victim in "Her love for him became fatal"). In general, linguistic strategies that background perpetrators have been shown to favour victim blaming (Huttenlocher et al., 1968; Henley et al., 1995; Bohner, 2002; Gray and Wegner, 2009; Hart and Fuoli, 2020; Zhou et al., 2021). This way of reporting contributes to reinforcing such social stereotypes.

If we want social stereotypes to be challenged, the language we use to describe GBV is thus an excellent place to start, also from a Natural Language Processing (NLP) perspective. Recent work has shown that perspectives on femicides and their triggered perceptions can be modelled automatically (Minnema et al., 2022b,a). In this paper, as shown in Box 1, we explore the challenge of *rewriting* descriptions of GBV with the aim to increase the perceived level of blame on the perpetrator, casting it as a style transfer task (Xu et al., 2012; Jin et al., 2022). In this novel *responsibility perspec-*

*Shared first co-authorship.

[1]A report on femicides from November 2018 by two Italian research institutes points out that the stereotype of a shared responsibility between the victim and its perpetrator is still widespread among young generations: "56.8% of boys and 38.8% of girls believe that the female is at least partly responsible for the violence she has suffered" (Laboratorio Adolescenza and Istituto IARD, 2018).



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 7907–7918
July 9-14, 2023 

*tive transfer* task, a given sentence from femicide news reports gets rewritten in a way that puts more responsibility on the perpetrator, while preserving the original content.

**Contributions** We create an evaluation set containing semi-aligned pairs with "low" and "high" sentences expressing similar information relative to an event, from an existing dataset of Italian news (§2.1). In absence of parallel training data, we follow previous work (Lample et al., 2019; Luo et al., 2019; Lai et al., 2021) to train an unsupervised style transfer model using mBART (Liu et al., 2020) on non-parallel data (with style labels) with a zero-shot and a few-shot approach using GPT-3 (Brown et al., 2020) to perform rewriting (§2.2). We run both human-based and automatic evaluations to assess the impact of rewriting on the perceived blame, comparing original and rephrased texts to find that models can achieve detectable perspective shifts (§3). By introducing the novel task of responsibility perspective transfer, providing an evaluation dataset, a battery of trained models, and evidence of a successful methodology, we hope to foster further research and application developments on this and other perspective rewriting tasks that are relevant to society.[2]

## 2 Experimental Settings

### 2.1 Datasets

Our work makes use of the *RAI femicide corpus* (Belluati, 2021), a dataset of 2,734 news articles covering 582 confirmed femicide cases and 198 other GBV-related cases[3] in Italy between 2012-2017. Of these, 182 cases (comprising 178 femicides and 4 other cases) are linked to a set of news articles from the period 2015-2017 that report on these cases. This dataset is augmented with perspective annotations from Minnema et al. (2022a). Gold annotations (averaged z-scored perception values from 240 participants) are available for 400 sentences, and silver annotations (annotated with the best-scoring model from Minnema et al. 2022a) are available for 7,754 further sentences. Using event metadata, we automatically extracted pairs of sentences $\langle L, H \rangle$, where $L$ and $H$ both reference the same GBV case, but respectively have a below-average ($L$) or above-average ($H$) level of perceived perpetrator blame. Next, for a subset of 1,120 sentences from the combined gold-silver perspective dataset, we performed manual filtering to ensure that for pair, $L$ and $H$ reference not only the same *case*, but also show substantial overlap in terms of the specific *events* within this case that they describe (e.g. the violence itself, the police investigation, conviction of a suspect, etc.). This yielded a set of 2,571 pairs (or 304 pairs if each sentence is used only once).

### 2.2 Models

Due to the limited availability of parallel data, we experiment with several existing text generation methods known to work in low-data settings.

**Unsupervised mBART** We train an unsupervised model with iterative back-translation (Hoang et al., 2018): two mBART-based models, one for each transfer direction, where outputs of one direction with source sentences are used to supervise the model in the opposite direction. All experiments are implemented atop Transformers (Wolf et al., 2020) using mBART-50 (Tang et al., 2021). We use the Adam optimizer with a polynomial learning rate decay, and a linear warmup of 100 steps for a maximum learning rate of 1e-4. We limit the maximum token length to 150. To alleviate computational costs and catastrophic forgetting, we only update the parameters of the decoder, freezing the other parameters.

**mBART + meta-information** A unique feature of our dataset is the availability of detailed meta-information about the events. We made a selection of the properties likely to be most relevant for characterizing the event and assigning responsibility (names of the victim and perpetrator, type of victim-perpetrator relationship, murder weapon and location) and concatenated this meta-information to the corresponding source sentence as input during training. We tried two order settings: ***source-meta*** and ***meta-source***. Preliminary experiments showed that concatenating only the event properties themselves, without including property names, produced the most promising results. For example: *"Trapani, Donna di 60 anni uccisa dall'ex marito — Anna Manuguerra, Antonino Madone, ex coniuge, arma da taglio, Nubio, casa"* ("Trapani: 60-year old woman killed by ex-husband — [victim name], [perpetrator name], ex-spouse, cutting weapon, [town name], at home"). We use the same training setup as for the previous model.

[2] Data and code are available at https://github.com/gossminn/responsibility-perspective-transfer.

[3] Including cases of non-lethal violence, suspected femicide, and suicide.

| Perspective Model *Dimension* | $R^2$ | Source | Target *(avg)* | mBART *base* | mBART *src-meta* | mBART *meta-src* | GPT-3 *na-zero* | GPT-3 *na-few* | GPT-3 *iter-1* | GPT-3 *iter-2* |
|---|---|---|---|---|---|---|---|---|---|---|
| **"blames the murderer"** | *0.61* | -0.511 | 0.445 | -0.250 | -0.188 | 0.284 | -0.157 | -0.375 | 0.109 | -0.116 |
| **"caused by a human"** | *0.60* | -0.228 | 0.362 | -0.037 | 0.005 | 0.371 | 0.042 | -0.095 | 0.278 | 0.076 |
| **"focuses on the murderer"** | *0.65* | -0.518 | 0.597 | -0.184 | -0.108 | 0.567 | 0.033 | -0.349 | 0.179 | -0.104 |

Table 1: Automatic evaluation of perspective using the BERTino-based model from Minnema et al. (2022a). Scores are z-normalized (i.e., a score -1 or 1 means "one standard deviation below/above average"). Target scores are averaged across different target sentences.

**GPT-3: Naive implementation** We also experimented with using the *text-davinci-002* version of GPT-3 (Brown et al., 2020) in a range of zero-shot and few-shot setups. Our ***naive-zero*** setup uses a simple prompt telling the model to rewrite the sentence with more focus on the perpetrator.[4] Next, ***naive-few*** uses a similarly simple prompt[5] along with a set of ten low-high sentence pairs randomly sampled from the gold annotations.

**GPT-3: Iterative few-shot** A challenging factor for our naive few-shot approach is that the the 'natural' source-target pairs from our annotated data are not perfect minimal pairs, as they differ in perspective but also have some content differences. In an effort to use maximally informative pairs as few-shot examples, we designed an iterative process for compiling small curated sets of examples. First, we designed an improved zero-shot prompt by giving a set of source-target pairs sampled from the gold annotations to the model and prompting it to explain the differences between the pairs. We discovered by accident that this yields a very plausible and concise task definition, and we reasoned that a definition generated by the model on the basis of real examples might be more informative as a prompt than a manually designed one. We then provided two annotators[6] with the resulting definition[7], as well as with five more source sentences sampled from the corpus. Each of the annotators then adapted the definition into a zero-shot prompt, used that prompt to generate target sentences for each of the source sentences, and selected the best candidate from these to create a set of pairs with maximal perspective contrast and content overlap, to be used in a few-shot prompt. We kept both versions of the few-shot prompt, ***iter-1*** and ***iter-2*** in order to measure the combined effects of small difference in prompt, randomness in the generated candidates, and judgement differences in the selection of the best candidate.

### 2.3 Evaluation Methods

The main goal of responsibility perspective transfer is to generate a sentence with the desired perspective ("style strength" in classic style transfer tasks) that still has the same semantic content as the source sentence. We assess the performance of different models using standard metrics commonly employed in text style transfer (Mir et al., 2019; Briakou et al., 2021; Lai et al., 2022; Jin et al., 2022), and custom automatic metrics; we also run a questionnaire study with human participants.

**Automatic Evaluation** For estimating perspective quality, we used the best-performing perspective regressor from Minnema et al. (2022a) which is based on an Italian monolingual DistilBERT model (*BERTino*; Muffo and Bertino, 2020).

For content preservation, we use three popular text generation metrics: $n$-gram-based *BLEU* (Papineni et al., 2002) and *ROUGE* (Lin, 2004), as well as a neural-based model *COMET* (Rei et al., 2020).

**Human Evaluation** Participants were given an online survey with 50 blocks, each corresponding to one source sentence sampled from the dataset. In each block, participants rated: 1) the level of perceived agent responsibility in each of the seven *target* candidates; 2) the level of *content preservation* of each target relative to the source. We also designed a separate, smaller questionnaire that

[4] *"Riscrivi la frase concentrandoti sul colpevole"* ("Rewrite the sentence and concentrate on the culprit")

[5] *"Riscrivi le seguenti frasi da low ad high. Per high si intende che la colpa è attribuita interamente al killer. Ecco alcuni esempi: [...] Riscrivi la seguente frase:"* ("Rewrite the following sentences from low to high. 'High' means that the blame is entirely put on the killer. Here are some examples: [...] Rewrite the following sentence:")

[6] The annotators were authors G.M. and M.B.

[7] The definition (slightly edited for grammar) is: *"Le frasi precedute dall'etichetta "Low:" tendono ad essere più brevi e non danno la colpa esplicita all'assassino, mentre le frasi precedute dall'etichetta "High:" tendono ad essere più dirette e a dare la colpa all'assassino."* ("The sentences preceded by "Low:" tend to be shorter and don't explicitly blame the murderer, while the sentences preceded by "High:" tend to be more direct and blame the murderer.")

| Metric | ↔ | Source | Target (avg) | mBART base | mBART src-meta | mBART meta-src | GPT3 na-zero | GPT3 na-few | GPT3 iter-1 | GPT3 iter-2 |
|---|---|---|---|---|---|---|---|---|---|---|
| **BLEU** | src | - | 0.015 | 0.725 | 0.612 | 0.236 | 0.303 | 0.435 | 0.489 | 0.285 |
| **ROUGE** | src | - | 0.100 | 0.808 | 0.701 | 0.351 | 0.551 | 0.638 | 0.659 | 0.450 |
| **COMET** | src | - | -1.216 | 0.540 | 0.257 | -0.591 | 0.103 | 0.538 | 0.379 | -0.058 |
| **BLEU** | tgt | 0.015 | - | 0.014 | 0.016 | 0.024 | 0.010 | 0.013 | 0.014 | 0.009 |
| **ROUGE** | tgt | 0.100 | - | 0.110 | 0.104 | 0.132 | 0.088 | 0.094 | 0.098 | 0.090 |
| **COMET** | tgt | -1.175 | - | -1.194 | -1.178 | -1.002 | -1.090 | -1.045 | -1.057 | -1.059 |

Table 2: Automatic content preservation metrics (BLEU, ROUGE, COMET), comparing generated sentences against source and gold target sentences.

| | | Perspective | Similarity | HM |
|---|---|---|---|---|
| **mBART** | *base* | 2.14 | 7.72 | 3.34 |
| | *src-meta* | 2.50 | 6.78 | 3.65 |
| | *meta-src* | 4.50 | 3.62 | 4.01 |
| **GPT-3** | *na-zero* | 2.77 | 6.52 | 3.89 |
| | *na-few* | 2.08 | 8.17 | 3.31 |
| | *iter-1* | 3.57 | 7.97 | 4.98 |
| | *iter-2* | 3.84 | 6.60 | 4.85 |
| **Examples** | *for iter-1* | 5.20 | 6.93 | 5.94 |
| | *for iter-2* | 3.87 | 5.27 | 4.46 |

Table 3: Human evaluation results on model outputs and examples for few-shot. HM is the harmonic mean of perspective and similarity scores

asked the same questions about the few-shot examples used in ***iter-1*** and ***iter-2***.

The pool of invited participants was a group of people with mixed genders and backgrounds from the personal network of the authors. No remuneration was offered. Four invitees responded to the main questionnaire, and three invitees responded to the few-shot example questionnaire (all female, mean age: 46). The participants have different levels of education (from middle school to university) and live in different regions of Italy.

Our evaluation study should be seen as a pilot, and larger-scale, more representative studies are planned for the future. The main aim of the pilot was to have a small-scale validation of our automatic metrics (taken from previous work and developed on the basis of a large-scale human study) and to test our evaluation setup (which questions to ask, etc.). The questionnaire was designed and distributed using Qualtrics.[8]

## 3 Results

### 3.1 Automatic Results

**Perspective Evaluation** Following Minnema et al. (2022a), we distinguish between several perceptual dimensions using a perception regression model, as shown in Table 1. Our main dimension of interest (highlighted in blue) is *blame on murderer*, but we also look at the two closely related dimensions of *cause* and *focus on murderer*. As shown by the $R^2$ scores, regression quality is decent for all of these dimensions. We observe that the source and target sentences have lower and higher blame scores respectively, which are also consistent on the two related dimensions, affirming that our testing data is of good quality in terms perspective aspect.

For all models, the perception scores of the predicted sentences are higher than those of the source sentences, with mBART/*meta-src* achieving the highest scores. This suggests that all models alter perceptions of responsibility to some extent. However, in virtually all cases, perception scores stay well below the target, and in many cases below the average level (zero). For mBART-based results, models with meta-information perform better than the baseline, with *meta-src* reaching particularly high scores. Within the GPT-3 settings, zero-shot (*na-zero*), surprisingly, performs better than few-shot; (*na-few*), and *iter-1* yields the highest scores.

**Content Preservation** When taking source sentences as the reference, three metrics show that the outputs have higher similarities to them than the target sentences. mBART/*base* has the highest scores, which (combined with the low perception scores of this model) suggests that the model tends to copy from the source sentence. Within the GPT-3 settings, *iter-1* has the highest scores. Using instead the target sentences as reference, we see that all scores are very close, with mBART/*meta-src*) reaching the best performance, followed by GPT-3/*na-few* and GPT-3/*iter-1*.

### 3.2 Human-based Results

Table 3 reports the results of our human evaluation study. We find that mBART/*meta-src* is the best overall model on perspective, but has poor

[8] https://www.qualtrics.com/

similarity. Meanwhile, GPT3/*na-few* achieves the highest score on similarity but the lowest score in terms of perspective, and its overall performance is lower than that of GPT3/*na-zero*. GPT3/*iter-1* has the best overall performance with an HM of 4.98. We found reasonably high levels of inter-annotator agreement (Spearman's rank correlation between pairs of annotators). Correlations ranged between 0.3-0.6 (blame) and 0.4-0.6 (similarity) with high levels of significance ($p < 0.0001$). The examples for few-shot are of higher quality overall as they were picked by the authors.

### 3.3 Case Study

Box 2 shows two sets of example outputs generated by mBART and GPT-3.[9] While hand-picked, these examples show that both models are capable of generating sentences that increase responsibility while trying to preserve content. However, they also highlight a key challenge: what if the source sentence lacks details about the event? The mBART model has access to event metadata and uses this effectively in Example 1 to produce a sentence that stays close to the source but with details from the metadata filled in (though with rather clunky sentence structure). In Example 2, instead, it produces a sentence that is factually correct but also loses most of the information from the source sentence. On the other hand, GPT-3, without access to metadata, often 'invents' missing information. This is evident in the second example, in which it faithfully preserves the source sentence and increases the level of blame by adding plausible but (partially) incorrect information about the crime.

**Prediction Examples**

**(Ex. 1)** *Provaglio d'Iseo, donna trovata morta in casa : si sospetta il compagno* "[town name], woman found dead at home: [her] partner is suspected"

**Target:** *Donna massacrata in casa nel Bresciano, il convivente fermato per omicidio volontario* "Woman massacred at home in [area name]: [her] cohabitant [partner] arrested for voluntary homocide"

**mBART (meta-src):** *Un uomo, Elio Cadei, è stato fermato per l'omicidio di Simona Simonini, la donna trovata morta in casa: si sospetta il compagno di lei.* "A man, Elio Cadei, has been arrested for the murder of Simona Simonini, the woman found dead at home: the partner of her is suspected"

**GPT-3 (iter-1):** *Provaglio d'Iseo, donna trovata morta in casa: il compagno è il principale sospettato* "[town name], woman found dead at home: [her] partner is the principal suspect"

---

**(Ex. 2)** *«Come vuole che ci si senta, quando ci sono stati due morti?»* "«How do you want us to feel, when there have been two deaths?»"

**Target:** *Albenga, uccide l'ex compagna a coltellate e poi si toglie la vita Loredana e l'incubo-stalker foto* "[town name], [someone] kills [his] ex-partner by stabbing and then takes his life[.] Loredana and the nightmare stalker[, see] foto"

**mBART (meta-src):** *Un uomo, Mohamed Aziz El Mountassir, di 43 anni, ha ucciso la moglie, Loredana Colucci, di 44, con una coltellata al collo.* "A man, Mohamed Aziz El Mountassir, 43 years old, has killed his wife, Loredana Colucci, 44 years old, with a knife to [her] neck."

**GPT-3 (iter-1):** *«Come vuole che ci si senta, quando ci sono stati due morti causati da un uomo che ha ucciso prima la sua ex moglie e poi la sua nuova compagna?* "How do you want us to feel, when there have been two deaths caused by a man who has first killed his ex-wife and then his new partner?"

Box 2: Prediction examples

## 4 Discussion & Conclusion

We proposed responsibility perspective transfer as a new task and introduced a dataset and models for applying this task to Italian news reporting about femicides. Our dataset contains a limited amount of quasi-aligned pairs that proved useful for evaluation and few-shot learning. We experimented with two modeling approaches: unsupervised mBART (with or without enriching the input with metadata) and zero-shot/few-shot learning with GPT-3.

Our human and automatic evaluations suggest *GPT-3/iter-1* as the best overall model, with a relatively high level of responsibility placed on the perpetrator and a good degree of content preservation. For the latter, most models score at least 6/10 on average on the human survey. The perspective change itself has also been achieved by our models, with substantially increased levels of perceived perpetrator blame compared to the source, but there is still much room for improvement: none of the models comes close to having the same level of blame as the target sentences do, and in the human evaluation survey no model achieves a 'blame score' of more than 4.5/10. The main obstacle for future improvements seems to lie with the lack of truly parallel data; however, our GPT-3-based iterative approach of creating minimal pairs seems to have worked quite well, and might be further exploited on a larger scale.

[9] Due to lack of space, we include only generations from the overall best-performing model from each category.

## 5 Limitations

This paper introduced the new task of responsibility perspective transfer and provided initial data collection and modeling for a specific domain (news about gender-based violence) and language (Italian). The main limitation of our work is that the (mBART) models that we trained and the prompts (for GPT-3) that we designed are specific to this domain and language and cannot be applied 'out-of-the-box' in other contexts. However, all of our modeling setups require no or limited training data and make use of readily available existing models, so we believe the general approach to be easily transferrable to other domains.

Another limitation comes from the fact that we used GPT-3: the model is closed-source and can only be accessed with a paid subscription to the OpenAI API (https://beta.openai.com/). This has consequences for reproducibility for several reasons. First of all, we do not have access to the exact technical specifications of the model or to the training data that was used. The GPT-3 models are regularly updated (at the time of our experiments, *text-davinci-002* was the most recent available version), but limited information is available about what distinguishes each version from the previous ones or from the original model introduced in Brown et al. (2020). Moreover, access to the API is controlled by OpenAI and could be closed at any time at the company's discretion; the API is currently quite accessible with no waiting list and a reasonably generous free trial, but the rates (paid in USD) might not be affordable for researchers outside of institutions in high-income countries, and not all researchers might be comfortable agreeing to the company's terms and conditions. Finally, the generative process involves a degree of randomness, and through the API it is not possible to fixate the model's random seed, meaning that the model produces different predictions every time it is called, even when using exactly the same prompt.

## 6 Ethics Statement

We see three important ethical considerations around our paper. The first consideration is related to the use of large proprietory language models (GPT-3). Apart from the reproducibility limitations resulting from the use of GPT-3 discussed above, there are more general ethical questions surrounding the use of GPT-3 and similar models, for example the high energy usage and resulting carbon emissions, and societal questions around the oligopoly on state-of-the-art language models that is currently in the hands of a handful of large US-based companies.

The second consideration relates to the task that we introduce: while we see perspective transfer models as a valuable tool for studying how language 'frames' (social) reality that could also have practical applications, for example in journalism, we strongly believe that any such applications must be approached with extreme care. The models that we introduce are scientific analysis tools that could be used to suggest alternative viewpoints on an event, but we believe that generations should *not* be seen as necessarily reflecting a 'true' or 'better' perspective, and should not used in a prescriptive way (i.e. used to tell someone how to write). We believe that the authors (journalists or others) of any text ultimately bear exclusive responsibility for the views, perspectives and (implicit) values expressed in it, and should be careful in making use of texts (re-)written by computers, such as the ones produced by our proposed models.

Finally, we are aware that our task domain (femicide/gender-based violence) is a societally and emotionally loaded topic, and that the texts contained in our dataset and produced by our models might be disturbing. In particular, in some cases, models may produce graphic descriptions of violence and/or produce questionable moral judgements (e.g., we have occasionally seen statements such as "the perpetrator of this horrible crime does not have the right to live" spontaneously produced by some of the models), and potential users of applications of the model should be aware of this. For the purposes of this paper, the only people external to the research team who have been extensively exposed to model outputs were the annotators in our human evaluation study. In the introduction page of our online questionnaire, annotators were warned about the sensitive nature of the topic and advised that they could stop their participation at any time if they felt uncomfortable and could contact the authors with any questions. Prior to running the online questionnaire we have requested and obtained ethical approval by the Ethical Review Committee of our research institution.

## Author contributions

Authors G.M. and H.L. share first co-authorship (marked with '*'). G.M. had primary responsibility for data collection and preparation, setting up the GPT-3 experiments and running the human evaluation survey. H.L. had primary responsibility for the mBART experiments and the automatic evaluation. B.M. annotated data (pair alignment) and contributed to prompt engineering and the design of the evaluation questionnaire. M.N. coordinated and supervised the overall project.

## Acknowledgements

Authors G.M. and M.N. were supported by the Dutch National Science organisation (NWO) through the project *Framing situations in the Dutch language*, VC.GW17.083/6215. Author H.L. was supported by the China Scholarship Council (CSC).

We would like to thank the annotators for helping us evaluate the models' outputs. We also thank the ACL anonymous reviewers for their useful comments. Finally, we thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

## References

M. Belluati. 2021. *Femminicidio. Una lettura tra realtà e interpretazione*. Biblioteca di testi e studi. Carocci.

Gerd Bohner. 2002. Writing about rape: Use of the passive voice and other distancing features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology*, 40:515–529.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1321–1336, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Kurt Gray and Daniel M. Wegner. 2009. Moral typecasting: divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96:505–520.

Christopher Hart and Matteo Fuoli. 2020. Objectification strategies outperform subjectification strategies in military interventionist discourses. *Journal of Pragmatics*, 162:17–28.

Nancy M Henley, Michelle Miller, and Jo Anne Beazley. 1995. Syntax, semantics, and sexual violence: Agency and the passive voice. *Journal of Language and Social Psychology*, 14(1-2):60–84.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Janellen Huttenlocher, Karen Eisenberg, and Susan Strauss. 1968. Comprehension: Relation between perceived actor and logical subject. *Journal of Verbal Learning and Verbal Behavior*, 7:527–530.

Shanto Iyengar. 1994. *Is anyone responsible? How television frames political issues*. University of Chicago Press.

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

Huiyuan Lai, Jiali Mao, Antonio Toral, and Malvina Nissim. 2022. Human judgement as a compass to navigate automatic metrics for formality transfer. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 102–115, Dublin, Ireland. Association for Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. Generic resources are what you need: Style transfer tasks without task-specific parallel training data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4241–4254, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *Proceedings of Seventh International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122.

Chiara Meluzzi, Erica Pinelli, Elena Valvason, and Chiara Zanchi. 2021. Responsibility attribution in gender-based domestic violence: A study bridging corpus-assisted discourse analysis and readers' perception. *Journal of pragmatics*, 185:73–92.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022a. Dead or murdered? Predicting responsibility perception in femicide news reports. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1078–1090, Online only. Association for Computational Linguistics.

Gosse Minnema, Sara Gemelli, Chiara Zanchi, Tommaso Caselli, and Malvina Nissim. 2022b. SocioFillmore: A tool for discovering perspectives. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 240–250, Dublin, Ireland. Association for Computational Linguistics.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

Matteo Muffo and Enrico Bertino. 2020. BERTino: An Italian DistilBERT model. In *CLiC-it 2020: 7th Italian Conference on Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meetings of the ACL*, pages 311–318.

Erica Pinelli and Chiara Zanchi. 2021. Gender-based violence in Italian local newspapers: How argument structure constructions can diminish a perpetrator's responsibility. *Discourse Processes between Reason and Emotion: A Post-disciplinary Perspective*, page 117.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

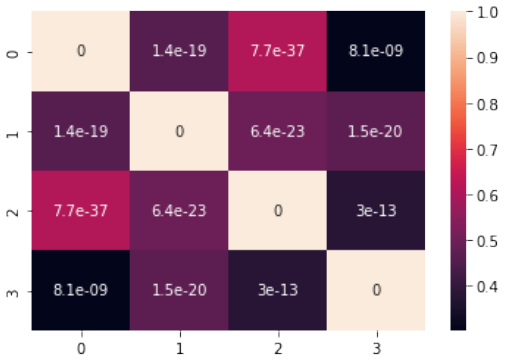
Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.

Karen Zhou, Ana Smith, and Lillian Lee. 2021. Assessing cognitive linguistic influences in the assignment of blame. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 61–69, Online. Association for Computational Linguistics.

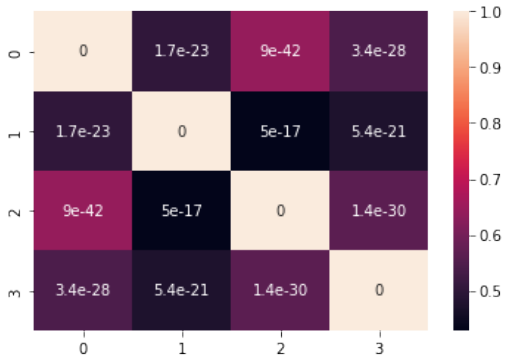
# A Annotation Statistics

## A.1 Inter-annotator agreement

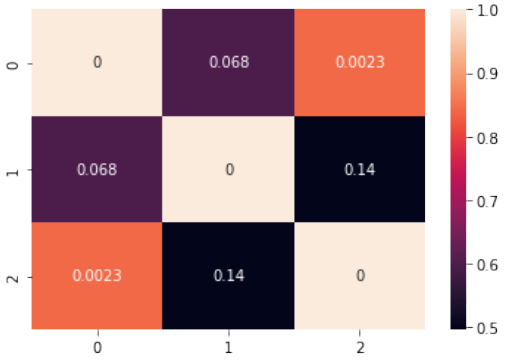
Figures A.1 give inter-annotator agreement scores for the human evaluation. Columns and rows represent individual annotators; colors represent Spearman correlations; numbers in cells are p-values.



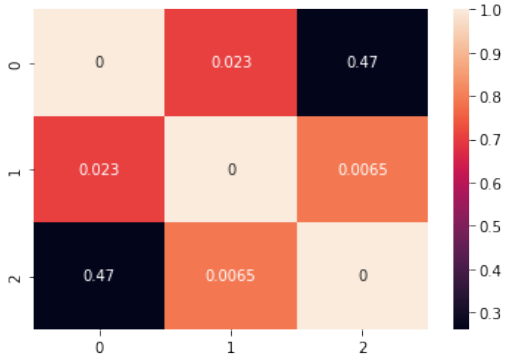

(a) Inter-annotator agreement for blame scores in the main human evaluation study.




(b) Inter-annotator agreement for content preservation scores in the main human evaluation study.




(c) Inter-annotator agreement for blame scores in the few-shot prompt evaluation study.




(d) Inter-annotator agreement for content preservation scores in the few-shot prompt evaluation study.

Figure A.1: Inter-annotator agreement

## B Questionnaire Materials

Mockups from the online survey are given in Figures B.2 and B.3.


Figure B.2: Qualtrics mockup: "speedometer" for rating agentivity

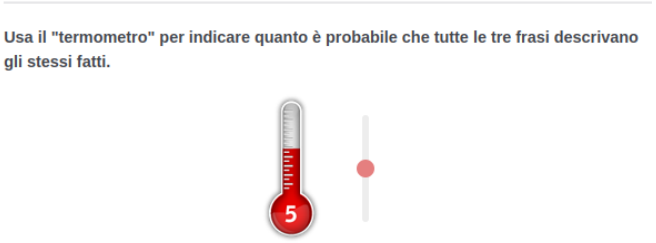



Figure B.3: Qualtrics mockup: "thermometer" for rating content preservation

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*§5 ("Limitations")*

☑ A2. Did you discuss any potential risks of your work?
*§6 ("Ethics Statement")*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*§1 ("Introduction")*

☒ A4. Have you used AI writing assistants when working on this paper?
*Used GPT-3 as a model in our experiments but not for writing any part of the paper itself*

### B ☑ Did you use or create scientific artifacts?

*§2 ("Experimental Settings") - §2.1 "Datasets", §2.2 "Models", §2.3 "Evaluation Methods"*

☑ B1. Did you cite the creators of artifacts you used?
*§2 ("Experimental Settings") - §2.1 "Datasets", §2.2 "Models", §2.3 "Evaluation Methods"*

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Left blank.*

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Left blank.*

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*The data that we collected was an anonymous online survey with only multiple choice questions, no chance for leaking identifiable personal information*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*§2 ("Experimental Settings") - §2.3 "Evaluation Methods"*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*§2 ("Experimental Settings") - §2.1 "Datasets", §2.2 "Models", §2.3 "Evaluation Methods"*

### C ☑ Did you run computational experiments?

*§2 ("Experimental Settings"), §3 ("Results")*

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*We fine-tuned mBART (on own machine) and few/zero-shot GPT-3 (via API), both have low computational footprint on our side*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*§2 ("Experimental Settings"), did prompt engineering and data augmentation but no hyper-parameter search*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*§3 ("Results")*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*§2 ("Experimental Settings") (mentioned specific model & API for GPT-3)*

**D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?**
*§2.1 ("Datasets"), §2.3 ("Evaluation Methods")*

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Appendix A*

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*§2.3*

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*§2.3 (participants recruited through personal network of authors)*

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*§2.3*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*§2.3*